Transcriptional and genomic parallels between the monoxenous parasite *Herpetomonas muscarum* and *Leishmania*.

Megan A. Sloan[1], Karen Brooks[2], Thomas D. Otto[2,3], Mandy J. Sanders[2], James A. Cotton[2*] and Petros Ligoxygakis[1*]

1. Department of Biochemistry, University of Oxford, South Parks Rd OX1 3QU Oxford UK.
2. The Wellcome Sanger Institute, Wellcome Genome Campus, Hixton, Cambridgeshire CB10 1SA UK
3. Present address: Centre of Immunobiology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK.

*Equal Corresponding Authors:

jc17@sanger.ac.uk (James Cotton)

petros.ligoxygakis@bioch.ox.ac.uk (Petros Ligoxygakis)

Abstract

Trypanosomatid parasites are causative agents of important human and animal diseases such as sleeping sickness and leishmaniasis. Most trypanosomatids are transmitted to their mammalian hosts by insects, often belonging to Diptera (or true flies). These are called dixenous trypanosomatids since they infect two different hosts, in contrast to those that infect just insects (monoxenous). However, it is still unclear whether dixenous and monoxenous trypanosomatids interact similarly with their insect host, as fly-monoxenous trypanosomatid interaction systems are rarely reported and under-studied – despite being common in nature. Here we present the genome of monoxenous trypanosomatid *Herpetomonas muscarum* and discuss its transcriptome during *in vitro* culture and during infection of its natural insect host *Drosophila melanogaster*. The *H. muscarum* genome is broadly syntenic with that of human parasite *Leishmania major*. We also found strong similarities between the *H. muscarum* transcriptome during fruit fly infection, and those of *Leishmania* during sand fly infections. Overall this suggests *Drosophila-Herpetomonas* is a suitable model for less accessible insect-trypanosomatid host-parasite systems such as sand fly-*Leishmania*.

Author Summary

Trypanosomes and *Leishmania* are parasites that cause serious Neglected Tropical Diseases (NTDs) in the world's poorest people. Both of these are dixenous trypanosomatids, transmitted to humans and other mammals by biting flies. They are called dixenous as they can establish infections in two different types of hosts – insect vectors and mammals. In contrast, monoxenous trypanosomatids usually only infect insects. Despite establishment in the insect's midgut being key to transmission of NTDs, events during early establishment inside the insect are still unclear in both dixenous and monoxenous parasites. Here, we study the interaction between a model insect – the fruit fly *Drosophila melanogaster* – and its natural monoxenous trypanosomatid parasite *Herpetomonas muscarum*. We show that both the genome of this parasite, and gene regulation at early stages of infection have strong parallels with *Leishmania*. This work has begun to identify evolutionarily conserved aspects of the process by which trypanosomatids establish in insects, thus potentially highlighting key checkpoints necessary for transmission of dixenous parasites. In turn, this might inform new strategies to control trypanosomatid NTDs.

Introduction

The family Trypanosomatidae belong to the order Kinetoplastida, a group characterized by the presence a mitochondrial organelle rich in DNA (kDNA) called the kinetoplast. This family includes parasitic flagellates that undergo cyclical development in both vertebrate and invertebrate hosts (and are therefore dixenous). These parasites are best known as agents of important diseases in humans, domestic animals and plants. However, several genera of this order such as *Crithidia*, *Herpetomonas*, *Blastocrithia* and *Leptomonas* are restricted to a single host (monoxenous), usually an insect from the orders Diptera, Hemiptera or Siphonaptera (Wallace 1966, Vickerman 1976, 1994). Although such monoxenous or "lower" trypanosomatids seem to have their lifecycle essentially confined to insect hosts, they have also been reported in plants (Rowton and McGee, 1983) and immunocompromised humans (Pacheco *et al.,* 1998).

There is an increasing interest in monoxenous trypanosomatids as a model for understanding the evolution and ecology of trypanosomatids (e.g. Zidkova *et al.*, 2010), as well as how they may modify their insect host (Lange and Lord, 2012). It is now clear that monoxenous trypanosomatids are ubiquitous parasites of a wide range of insect groups and have numerous effects on the physiology of the insect host (reviewed in Lange and Lord, 2011). These effects include alterations in fertility and reproduction, modified food intake, delayed development and reduction in lifespan (reviewed in Vega and Kaya 2012). In projections of total animal biodiversity, insects represent more than 60% of all animals (Erwin, 1982). Knowledge of insect physiology and what can influence it is therefore essential for maintaining a species-rich environment especially when longitudinal population data show a sharp decline in flying insect biomass (Hallmann *et al.,* 2017). Thus,

studies of trypanosomatid-insect interactions will provide vital insights into the ecology of crucial insect species (e.g. pollinators).

To this end, a number of monoxenous trypanosomatid genomes and transcriptomes are being investigated (Motta *et al.,* 2013); including bee parasites from the genus *Crithidia* (Schmid-Hempel *et al.,* 2018), *Lotmaria passim* (the honey bee parasite, Runckel, DeRisi and Flenniken, 2014) and *Leptomonas pyrrhocoris* (Flegontov *et al.,* 2016) a globally disseminated parasite isolated from fire bugs. These studies, and earlier work on the molecular biology of trypanosomatids, have revealed that monoxenous parasites share many distinctive genome features with their better-studied dixenous relatives (Teixeira *et al.,* 2012).

The genomic DNA is arranged into 'polycistronic' (multi-gene) transcriptional units of functionally unrelated genes, which lack introns. Given this gene arrangement, the cells do not control an individual gene's expression by varying its transcription level, instead expression is controlled by RNA-binding proteins (Zoltner *et al.,* 2018) and other post-transcriptional processes such as RNA editing (Stuart *et al.,* 1997). RNA editing processes include trans-splicing where 39 nucleotides, called a splice leader sequence, are added to the 5' end of mRNAs (reviewed in Liang *et al.,* 2003). The splice leaders (also called mini exons) are encoded in tandem repeats in a different genomic locus to the gene.

Trypanosomatid kDNA is arranged in interlocking 'maxi-circles' (Chen *et al.,* 1995; Lukeš *et al.*, 2002; Borghesan *et al.,* 2013). The kDNA maxicircle is homologous to mitochondrial genomes in other systems but the sequence encoding many of typical mitochondrial proteins is scrambled, relying on post-transcriptional mRNA editing to reconstitute the correct coding sequence (Simpson and Thiemann, 1995). The kinetoplast

also contains thousands of associated 'mini-circles' which encode guide RNAs involved in this editing process (Lukeš *et al.,* 2002).

In addition to ecological insights, studies of monoxenous trypanosomatids may help us gain new perspective on interactions of more medically important parasites and their insect vectors, which mediate neglected tropical diseases such as Leishmaniasis (vectored by phlebotomine sand flies) and sleeping sickness (tsetse flies) (see Wang *et al.,* 2019). To inform, and accelerate, research in these experimentally challenging dipteran-parasite relationships, we have developed the study of the model dipteran *Drosophila melanogaster* and its natural trypanosomatid *Herpetomonas muscarum* (Wang *et al.*, 2019). We have established that a network of signalling in the intestine of the host was important for clearance as well as for maintaining fecundity. This network involved NF-κB and STAT-mediated transcription, which regulate intestinal stem cell proliferation that the parasite attempts to suppress (Wang *et al.*, 2019). Here, we turn our attention to the parasite. We report the genome of *Herpetomonas muscarum* isolated from a wild population of *Drosophila melanogaster* in Oxfordshire, UK. We also report the transcriptomes of this *H. muscarum* isolate from *in vitro* culture and during the course of infection in *D. melanogaster*. The similarities with *Leishmania major* both at the genome level as well as transcriptome regulation were striking. This was especially the case in the early phases of host infection when the parasite needs to overcome the barrier of the insect midgut and establish infection. Given the resistance mechanisms to parasite establishment (and therefore onward transmission) reside in the dipteran midgut (Van den Abbeele and Rotureau, 2013; Knöckel *et al.*, 2013), the *Drosophila-Herpetomonas* model may allow researchers to take advantage of the extensive toolkit of genetic approaches available for *Drosophila* to uncover mechanistic details of evolutionary conserved aspects of the

relationship between trypanosomatids and dipteran vectors like sand flies, where the tool-box for functional studies is not yet fully developed.

Results and Discussion

The Herpetomonas muscarum genome

Assembly. PacBio and Illumina sequence reads were generated from an axenic culture of *H. muscarum* promastigotes as described in Materials and Methods. The reads were assembled into a genome of 41.7 Mbp in 264 scaffolds with the largest 1,793,442 bp in length (N50 = 707,495 bp). We observed a median read coverage of 114x with populations of scaffolds coverage at approximately 50x and 160x which may represent monosomic and trisomic scaffolds (Figure 1, predicting 37-39 chromosomes). Kmer analysis of the sequencing reads estimated the haploid genome length to be approximately 35.2 Mbp with a read error rate of less than 1% (Figure S1, Vurture *et al.,* 2017). While the GenomeScope model does not fit the aneuploid nature of trypanosomatid genomes (see below), we believe this suggests our assembly is approximately the correct size.

Annotation. Gene model annotation was generated with Companion (Steinbiss *et al.,* 2016) using evidence from RNA-seq data (described below) and the proteomes of *Leishmania major*, *L. braziliensis* and *T. brucei* as described in Materials and Methods. The final *H. muscarum* v1 annotation contains 12,687 genes, of which 12,162 are inferred to be protein-coding (Table 1). All unique open reading frames produced by the gene models were kept, even in cases where the gene prediction was not strongly supported by RNA-sequencing evidence, in an attempt to not 'miss' genes. It is therefore likely that this annotation contains a higher number of genes than the 'true annotation'. However, the number of reported genes is close to that reported for other trypanosomatid species e.g. *T. brucei* TREU927 strain contains 11,567 genes (Berriman *et al.,* 2005).

Conserved features of trypanosomatid genomes

Genome structure and large scale synteny. As seen in other trypanosomatid genomes, open reading frames were found on both strands on many scaffolds. Genes are (mostly) arranged in large groups of genes present on the same strand and in the same direction, which is indicative of the polycistronic transcripts typical in trypanosomatid genomes. The regions between polycistrons, commonly referred to as strand switch regions (SSRs), are thought to contain the transcriptional start sites for transcription of each group of genes. We used the SSRs to define and estimate the number of polycistrons. Here we defined SSRs to begin and end at genes where the downstream open reading frame is on the opposing strand of the same scaffold. This highlighted 386 genes from 112 different scaffolds. These putative strand switches were manually inspected and could be grouped into different three situations. There were 128 *bona fide* strand switches which were either divergent (72 cases) or convergent (56 cases) (Table S1). There were 166 cases where a single gene (or small group of < 5 genes) had become inverted within a polycistron (166 cases). Small genes (< 350bp) encoding hypothetical proteins and tRNAs were commonly found in these cases, though other larger genes were also found in these groups e.g. HMUS00935500.1 an putative trans-sialidase. Finally, there were 92 cases where a strand switch does occur, but the precise locus was unclear. These cases tended to be at where a single gene at the end of a scaffold was on the opposing strand to all other genes on the scaffold – as such it was unclear if this represented a *bona fide* strand switch or a single gene inversion. Overall, this indicated there are at least 128 polycistrons in the *H. muscarum* genome, though this is likely to be an underestimate given the ambiguity of some strand switch regions.

Comparisons with other trypanosomatids genomes also suggest this figure is an underestimate, e.g. *L. major* is predicted to have 184 polycistrons (Thomas *et al.,* 2009) and *T. brucei* is predicted to have 150 (Daniels, Gull and Wickstead 2010), both of which have smaller genomes and fewer predicted chromosomes than *H. muscarum*.

Despite diverging before the existence of mammals (El-Sayed *et al.,* 2005), trypanosomatids show high gene order conservation across the genome. As expected, the *H. muscarum* scaffold showed synteny with other trypanosomatid genomes (Figure 2A-E). *Herpetomonas* was most highly syntenic with *L. major* despite being considered phylogenetically closer to *Phytomonas* and *Leptomonas*. To quantify this, we took non-overlapping windows of adjacent *H. muscarum* genes with single-copy orthologs in three comparator genomes: *Leishmania major*, *T. brucei* and *Leptomonas seymouri.* For each window size, we count for how many windows have all orthologs on the same scaffold in the comparator (syntenic windows), and for how many of those all the genes are in the same relative order as their *H. muscarum* orthologs (colinear windows). Almost 96% of 3-gene windows of single-copy orthologs between *H. muscarum* and *L. major* (1845/1926) are syntenic, and 53% of these are colinear (985/1845). This conserved genome structure is shared, to a slightly lesser extent across the trypanosomatids (91.7% or 1386/1511 syntenic with *T. brucei brucei*, 55% or 766/1386 colinear, 80.9% or 1643/2030 syntenic with *L. seymouri*, 46% or 761/1643 colinear). This relationship holds across window sizes (Figure 2F). The values for synteny with *Leptomonas seymouri* are likely to be biased downwards by the fragmentary assembly available for that species, and this analysis does not capture rearrangements, expansions or contractions of multi-gene families, for which one—to-one orthology is unlikely to be clear.

Splice leader sequence. In trypanosomatids, each mRNA is capped, via trans-splicing (reviewed in Liang *et al.,* 2003), with a conserved 39bp sequence called the splice leader (SL). The SL is encoded by the mini-exon genes which are found throughout the genome in tandem arrays. Each mini-exon has two components; the highly conserved 39bp sequence trans-spliced on to mRNAs (the exon) and a less well conserved intronic sequence. Between each mini-exon gene there is a variable spacer region which is not transcribed. To find the splice leader sequence for our *H. muscarum* isolate, we searched for the conserved 39bp SL sequence from *Phytomonas serpens* (L42381.1) in the *H. muscarum* scaffolds. This gave 259 hits over 24 scaffolds, which we used to identify 19 clusters of mini-exon gene repeats (over 15 scaffolds) containing 3-43 copies of the mini exon gene (see Table S2). The first 111bp of the gene are common to all copies of the mini-exon gene and contain a 40bp splice leader sequence and what we predict to be the intron.

The splice leader sequence (1-40bp) and the putative intronic region (41-111bp) were then aligned with mini-exon sequences of several other trypanosomatids in the Leishmaniinae clade - including 9 other *Herpetomonas* isolated from heteropterans in the neotropics (Yurchenko *et al.,* 2009). Whilst the splice leader sequence is well-conserved across the clade (Table 2), we observe variability in the A/T-rich region between bases 11-19bp which appears genus specific, with the exception of the *Herpetomonas* sequences. *H. rotimani and H. nabiculae* have identical sequence across the 11-19bp region. However, the *H. muscarum* and *H. nabiculae* differ from each other, and the other *Herpetomonas* sequences over this variable region. Additionally, compared to other trypanosomatids, the *Herpetomonas* sequences have an 'additional' adenosine between bases 10 and 11. The intronic region from *H. muscarum* shows high similarity to that of previously reported *Herpetomonas* sequences. The first 15bp of the intronic sequence appear to be conserved in

other species from the Leishmaniiae clade, however the sequence becomes more variable thereafter in both in terms of base content and length.

Tubulin loci. The architecture of the tubulin arrays have been described in a number of trypanosomatids (Jackson, Vaughan and Gull, 2006), with two mutually exclusive formats being defined – monotypic and alternating. Monotypic tubulin arrays consist of either alpha-tubulin or beta-tubulin. Alternating arrays contain both alpha-tubulin and beta-tubulin genes which alternate along the array. The *H. muscarum* orthologues of *Trypanosoma brucei* alpha and beta tubulin genes were found using Orthofinder and used to locate the tubulin arrays.

We identified three genomic loci containing *H. muscarum* tubulin genes (Figure 3). Two of these loci consist of beta-alpha alternating arrays and the third locus consists of four copies of a beta tubulin genes. The alternating beta-alpha arrays are consistent with previous findings (reported as *Herpetomonas megaseliae* in Jackson *et al.,* 2006) and suggested that, like *Trypanosoma brucei*, *H. muscarum* genome has the alternating tubulin array configuration. However, the presence of a monotypic beta tubulin array in addition to the alternating arrays contrasts the established model in which each species has either alternating or monotypic arrays, but not both.

The genes surrounding the monotypic beta tubulin locus shared some synteny with regions of chromosome 4 of *T. brucei* and chromosome 8 of *L. major* (gene numbers Tb927.5.970 – Tb927.927.5.3090 and Lmj.08.1090-Lmj.08.11140). Interestingly this region of *L. major* chromosome 8 is one of two singleton beta-tubulin loci in the species. As such, the tubulin configuration of *H. muscarum* was an intermediate between the tubulin array configurations of *T. brucei* and *L. major*.

The predicted *Herptomonas muscarum* Proteome

Orthofinder (Emms and Kelly 2015) was used to identify orthologous proteins from other trypanosomatids in the predicted proteome of *H. muscarum*. For the analysis, protein coding genes from the following species were used: 9 *Trypanosoma* species/subspecies (*Trypanosoma brucei brucei*, *Trypanosoma brucei gambiense*, *Trypanosoma congolense*, *Trypanosoma cruzi*, *Trypanosoma evansi*, *Trypanosoma grayi*, *Trypanosoma rangeli*, *Trypanosoma theileri* and *Trypanosoma vivax*), 4 *Leishmania* species (*Leishmania braziliensis*, *Leishmania donovani*, *Leishmania infantum* and *Leishmania major*); 6 additional monoxenous trypanosomatids along with our *Herpetomonas muscarum* predictions (*Angomonas deanei*, *Leptomonas pyrrhocoris*, *Leptomonas seymori*, *Crithidia bombi*, *Crithidia expoeki*, *Crithidia fasciculata*). Finally, we included a free-living, non-trypanosomatid kinetoplastid, *Bodo saltans*, as an outgroup. From these 21 species 87.5% of genes were assigned to 12,701 orthogroups (for summary see Table 3, full orthogroups table Table S3). We found 7,265 of these orthogroups contained *H. muscarum* genes. There were 45 orthogroups containing only *H. muscarum* genes, these groups contain 215 genes. Overall, 90.7% of *H. muscarum* predicted proteins were assigned to an orthogroup.

Orthofinder also produced a phylogenetic tree based on protein sequences from proteins in orthogroups which contained a single gene from every species used in the analysis (Figure 4A). This tree is consistent with others published for the trypanosomatids (Maslov *et al.,* 2013). Unsurprisingly *H. muscarum* shares more orthogroups with *L. major* (6,607) than *T. brucei* (5,893) – which is more distantly related (Figure 4B). However, *H. muscarum* had slightly more orthogroups in common (6754) with the two *Leptomonas sp.* used in the analysis (Figure 4C). Finally, within the Leishmaniinae clade *H. muscarum* and two species of

'old world' *Leishmania*, *L. major* and *L. donovani*, shared 81.2% of their orthogroups (Figure 4D). A global examination of the patterns of gene family sharing between *H. muscarum*, and other trypanosomatid groups confirmed these patterns (Figure 5A). Most gene families, including most genes, are present in all of the groups, and another significant set of families is shared by all the trypanosomatid groups but missing from the outgroup, the free-living kinetoplastid *Bodo saltans*. These trypanosomatidae-specific gene families tend to be quite large, while many smaller gene families are specific to genera *Crithidia* and *Trypanosoma*, perhaps because of the more extensive taxon sampling of these lineages. There are exceptions, including some strikingly large gene families unique to trypanosomes, *Leishmania* and a number of other taxonomic groups (Figure 5B). Monoxenous trypanosomatids share many more genes families with *Leishmania* than *Trypanosoma*, and there are strikingly few families specific to the *Leishmania* lineage or any of the monoxenous parasites except *Crithidia*, explaining the strikingly similar predicted proteomes of *Leishmania* and *H. muscarum*.

We could not look in detail at all of the homology relationships between genes in this extensive comparison. We used a more focused OrthoFinder analysis to investigate specific groups of orthologues between *H. muscarum* and *T. brucei* genes of interest e.g. metabolic pathway genes, as *T. brucei* is the best-studied kinetoplastid at the molecular and cellular level. We summarise our findings in Table 4 (for full data see Tables S4-S15) and discuss some of the orthologues of interest, including surprisingly 'missing' orthologues, below.

Metabolism. *H. muscarum* is missing sphingolipid (SL) biosynthesis genes SLS1-4, including the inositol phosphorylceramide synthase and two choline phosphorylceramide synthases.

These genes are part of the same orthogroup from our analysis. Most of the *Trypanosoma* have 4 genes assigned to this orthogroup (with the exception of *T. cruzi* (2) and *T. vivax* (0)). However, other species used in this analysis had only 1 gene assigned to this orthogroup. Given that SLs are thought to be essential to eukaryotic membranes (Sutterwala *et al.,* 2008), this seemed surprising. However, *Leishmania major* promastigotes do not require *de novo* SL synthesis and a mutant devoid of SLs was viable and replicated as log-phase promastigotes (Zhang *et al.,* 2003 and 2007). However, the SL-free mutant was unable to differentiate into a metacyclic stage *in vitro* and showed severe defects in vesicular trafficking. As such, like *L. major*, *H. muscarum* and the other species without a complete SLS pathway may rely on scavenging sphingolipids from the environment.

H. muscarum did not have orthologues for the carnitine O-acetyltransferase (CAT) (Tb927.11.2230) and L-threonine 3-dehydrogenase (Tb927.6.2790) genes of the acetate metabolism pathway. We were also unable to find an orthologue to these genes in other species from the Leishmaniinae clade used in the analysis. As such these genes may have been lost sometime after the group diverged from *Trypanosoma*.

Additionally, three *T. brucei* respiratory chain genes did not appear to have orthologues in *H. muscarum,* including mitochondrial NADH-ubiquinone oxidoreductase flavoprotein 2 (Tb927.7.6350), which had orthologues in all species used in the analysis apart from *H. muscarum*. Similarly, the only genomes in the analysis without an orthologue for the cytochrome c oxidase assembly protein (Tb927.10.3120) were *H. muscarum* and *Phytomonas EM1*. Given the importance of these genes, this likely indicates an important gap in the *H. muscarum* annotation. Finally, no orthologue was identified for the *T. brucei* alternative oxidase (AOX) (Tb927.10.7090) which is found in *Trypanosoma* and is upregulated in bloodstream forms. This oxidase is thought to enhance organisms ability to

cope with stress associated with temperature change, infections and oxidative stress (Vanlerberghe and McIntosh, 1997).

We also note that for several *T. brucei* genes there were multiple *H. muscarum* orthologues. Two of the most extreme examples of this being the high-affinity arginine transporter AAT13 (Jackson 2007, Shaked-Mishan *et al.,* 2006) and the endo-/lysosome-associated membrane-bound phosphatase 2 (MBAP2) which have 38 and 18 orthologous genes in *H. muscarum* respectively. The increased copy number of these genes hints at their importance, though the reason for their high-copy number in *H. muscarum* is as yet unclear. AAT13 and MBAP2 have been shown to be highly upregulated in *Leishmania* during sand flies and in conditions of nutrient starvation (Martin *et al.,* 2014; Inbar *et al.,* 2017). Speculatively, the increased copy number of these genes may reflect the nutrient availability in *Herpetomonas*' environment/host(s).

Differentiation. RNA-binding proteins (RBPs) have emerged as key modulators of gene expression in trypanosomatids - particularly in the context of trypanosome development and differentiation (Kolev *et al.,* 2014). Orthologues were found for 72/75 *T. brucei* RNA-binding proteins. RNA-binding proteins with no orthologues found in *H. muscarum* were: chromatin-remodelling-associated RRM2 (Tb927.6.2550, (Naguleswaran *et al.,* 2015), the pre-RNA processing protein RBSR1 (Tb927.9.6870, Wippel *et al.,* 2019) and a hypothetical RBP (Tb927.10.14950).

We have not observed differentiation in *H. muscarum* using 'classical' temperature/pH manipulations *in vitro* or during *D. melanogaster* infections. As such the 'completeness' of the *H. muscarum* RBP repertoire, relative to *T. brucei* which has multiple discrete forms, is of interest. Several of these proteins had multiple orthologues in *H. muscarum* including RBP10

(4 orthologues, Tb927.8.2780). RBP10 is known to be highly expressed in bloodstream forms of *T. brucei* and its overexpression in procyclics led to an increase of many bloodstream-form specific mRNAs, as well as transcripts associated with sugar transport, the flagellum and cytoskeleton (Wurst *et al.,* 2012). The role for this protein in *H. muscarum* is unclear, as it does not appear to have a *bona fide* vertebrate host, however given this proteins links to sugar transport, it may play a more general role in metabolism in *H. muscarum*. Comparisons of *H. muscarum* RBP expression levels/timings with other trypanosomatids may shed more light on their role in the cell and potentially why we do not observe differentiated forms for this species.

In addition to the RBPs, we were unable to find any orthologues for the hydrophilic acylated surface proteins (HASPs) or small hydrophilic endoplasmic reticulum-associated proteins (SHERPs) which are associated with metacyclogenesis in *Leishmania*. We also note that the repressor of differentiation kinase 1 (RDK1, Tb927.11.14070) has 6 orthologues in *H. muscarum*. In *T. brucei*, RDK1 acts with the PTP1/PIP39 phosphatase cascade to prevent uncontrolled differentiation from bloodstream to procyclic form (Jones *et al.,* 2014). Given that *H. muscarum* is thought to be confined to insects, the presence of multiple copies of this gene which assists in maintaining a 'vertebrate' cell form in *T. brucei* is intriguing. It may be that this protein has an alternative role in *H. muscarum*.

Surface proteins. No orthologues were found for the EP procyclins which are known to be expressed highly *T. brucei* procyclic whilst in the tsetse vectors and are thought to provide protection from the digestive enzymes in the insect midgut (Acosta-Serrano *et al.,* 2001; Haines *et al.,* 2010). As such *H. muscarum* likely relies on other surface proteins for protection in the insect midgut (see the transcriptomic data below).

Additionally, Orthofinder was unable to find an orthologue to the major surface proteins of salivary gland forms of *T. brucei* - BARPs (bloodstream alanine-rich proteins). These GPI-anchored proteins required for tsetse salivary gland colonisation (Urwyler *et al.,* 2007, Fragoso *et al.,* 2009). Additionally, we do not find orthologues for the *T. brucei* metacyclic invariant surface proteins (MISPs) which are found extending above the VSG coat in salivary gland metacyclic forms (Casas-Sánchez *et al.,* 2018). Given the proteins are crucial for salivary gland colonisation, the lack of copies in the *H. muscarum* genome may partially explain the inability of *H. muscarum* to colonise the salivary glands of *D. melanogaster*, instead infections are confined to the insect crop and gut (Wang, Sloan and Ligoxygakis 2019).

Finally, the 13 *T. brucei* GP63 genes were grouped with 28 *H. muscarum* genes. GP63 is a major surface protease in *L. major* promastigotes. The comparatively high copy number of GP63 in *H. muscarum* may highlight its importance. Furthermore, GP63 has been implicated in *Leishmania* virulence (Pereira. *et al.,* 2010), and as such these will be of interest in future studies.

Nuclear proteome. Kinetochore interacting protein 3 (KKIP3, Tb927.10.6700) and SR protein (Tb927.9.6870) had no orthologues in *H. muscarum* or other species from the Leishmaniiae clade used in the analysis and as such they appear to be *Trypanosoma* specific. RNAi of KKIP3 in *T. brucei* resulted in defects in DNA segregation and reduced population growth (D'Archivio and Wickstead, 2017).

Additionally, *T. brucei's* kinetochore interacting protein 1 (KKIP1), PHF5-like protein (Tb927.10.7390) and U1 small nuclear ribonucleoprotein 24 kDa (Tb927.3.1090) had orthologues in all species used in the analysis apart from *H. muscarum*. Similar to KKIP3,

RNAi knock down of KKIP1 caused defects in DNA replication, though in the case of KKIP these defects were more severe – resulting in the loss of entire chromosomes (D'Archivio and Wickstead, 2017). It is unclear if these genes have been lost in *H. muscarum* or this indicates a gap in the current annotation. Based on the importance of KKIP1 and the fact these genes have orthologues in all other species analysed, it is likely to be the latter.

Finally, *H. muscarum* appears to have a 'full set' of the *T. brucei* RNA interference pathway genes including an orthologue for TbARGO1 (Tb927.10.10850). Genes from this well-conserved (in metazoans) pathway have been lost in several trypanosomatids including: *L. major, L. donovani* and *T. cruzi* (Robinson and Beverley 2003, DaRocha *et al.* 2004). The loss of this pathway in these organisms has been linked to *Leishmania* RNA virus perturbation (Beverley 2003, Lye *et al.* 2010) - though this has not been explicitly demonstrated. Further investigations to look for evidence of viruses akin to the LRVs in *H. muscarum* could test the link between RNAi and virus infection in trypanosomatids. The presence of a functional RNAi pathway has also been linked to transposon activity in *Leishmania* – with RNA-negative species lacking active transposable elements (TEs), and RNAi competent *L. braziliensis* harbouring several classes of active TEs (Peacock *et al.* 2007, Lye *et al.* 2010). Given this, it is possible that the loss/lack of active TEs in *L. major* and *L. donovani* have lifted the requirement of the RNAi pathway to protect against TE-associated genomic perturbations. We did observe transcripts corresponding to the telomere associated mobile elements (TATEs) in all *H. muscarum* transcriptomes (see below). As such, there may also be an important link between RNAi and transposon activity in trypanosomatids.

The *H. muscarum* transcriptome in *in vitro* culture: log vs. stationary phase

We first analysed the transcriptome of *H. muscarum* during *in vitro* axenic culture, specifically to compare log-phase and stationary phase cultures. Knowledge of the log-phase transcriptome was especially important as this was the 'pre-infection' transcriptome in our *Drosophila* infection model. By comparing the log-phase *H. muscarum* transcriptome with that of *H. muscarum* in flies we sought to identify genes important in the establishment of infection (see section below). The principal component analysis (PCA) plot (Figure S2) shows that the first principal component is mostly capturing variation between distinct clusters of samples from log and stationary phase and explains 68% of the variance in these data. As expected, we found extensive differential expression between log-phase and stationary phase, with 4044 genes significantly differentially regulated (p-adjusted <0.05) (Table S16). This is approximately a third of the genome but most changes in expression were modest, with only 264 genes upregulated ≥ 2-fold in stationary phase cells and 811 downregulated ≥ 2-fold which we will discuss further below. GO enrichment analysis, using Ontologizer (Vingron *et al.,* 2007), did not identify any significantly enriched GO terms associated with differentially regulated genes. However only 62% of *H. muscarum* genes have associated GO terms. As such, we looked for enrichment in Pfam domains. There were 26 Pfam domains significantly enriched in the genes upregulated in stationary phase and 73 Pfam domains significantly enriched among downregulated transcripts (Table S17), which we discuss further below.

Cell cycle associated proteins. The Pfam domain associated with cyclins was significantly enriched in genes upregulated in stationary phase cells. From this, we investigated the expression profiles of the cyclins, and their associated kinases. Eleven were found to be differentially regulated between the two cell populations (Table 5). There was significant

downregulation of the mitosis-associated cyclin 8, CRK3 and several mitochondrial DNA polymerase subunits in stationary phase cells. Knockdown of CRK3 in *T. brucei* is associated with a reduction in cell growth (Tu and Wang, 2004). Furthermore, there was upregulation of the G_1-associated cyclins 7, 4 and 11. These observations reflect the observed reductions in cell replication at higher cell densities. Consistent with this, and with a reduction in cell growth, there were also significant reductions in transcripts for α- and β-tubulins, DNA polymerases and several protein synthesis-related genes including: 40S ribosomal subunits, 28S rRNAs and five putative elongation factor 2 genes. However, there was also upregulation of mitosis-associated cyclin 2 in the stationary phase cells. Cyclin 2 has two roles in *T. brucei* procyclics: cell cycle progression through G1 and the maintenance of correct cell morphology at the posterior end of the cell (Hammarton, Engstler and Mottram, 2004). The CRKs 10 and 12, which were also upregulated in stationary phase cells, have been shown to interact with cyclin 2 and their knock-down results in growth defects (Liu, Hu and Li, 2013). CRK12 is also essential to survival of *T. brucei* in mice and its depletion by RNAi lead to defects in endocytosis, an enlarged flagellar pocket and abnormal kinetoplast localisation (Monnerat *et al.,* 2013). Given the relative abundance of many transcripts associated with reduced replication in stationary phase cells, the upregulation of cyclin 2 and its associated CRKs (10 and 12) may be more relevant to the maintenance of correct cell morphology than mitosis.

Stress and Metabolism. Stationary phase (of growth) is associated with build-up of toxic waste products and fewer nutrients available per cell. It was therefore unsurprising that we observed transcriptional changes indicating metabolic change and nutrient starvation. Genes containing the Pfam domain associated with major autophagy marker ATG8 were

significantly enriched in stationary phase transcripts (33 in total). Autophagy is a vital process for survival in nutrient poor environments and involves the segregation of the cell components to be recycled into double membrane-bound vesicles called autophagosomes. The requirement for increased amounts of membrane in autophagy, may partially explain the upregulation of fatty-acid synthesis related genes in stationary phase, as fatty acids are crucial components of cell membranes. Three lipases, two putative lipase precursor-like proteins, fatty-acyl-CoA Synthase 1 and putative fatty acid elongase (ELO) protein were upregulated upon entry into stationary phase. This is consistent with observations of *Trypanosoma cruzi* cultures (Lee *et al.,* 2006).

Whilst the upregulation of autophagy-related genes is an indicator of cell stress, we also observed the downregulation of several genes with domains associated with responding to oxidative stress including: thioredoxin, glutathione S-transferase and alkyl hydroperoxide reductase (AhpC) and thiol specific antioxidant (TSA). As such, cells do not appear to be under significant oxidative stress. Other forms of stress, such as reduced nutrient availability or pH changes, may be driving the predicted increases in autophagy. Additionally, transcripts bearing the heat shock protein 60 HSP60 domain (PF00118) were also significantly enriched in the downregulated transcripts, which is another indicator of cell stress.

Cell surface proteins. Proteins sharing a domain (cl28643) with the variant surface protein (VSP) proteins of the *Giardia lamblia*, a flagellated intestinal pathogen, were highly represented among genes upregulated in stationary phase *H. muscarum*. In *G. lamblia*, these VSPs are integral membrane proteins rich in cysteine residues, often in CxxC repeats. They have a highly conserved C-terminal membrane spanning region which has a

hydrophilic cytoplasmic tail with a conserved five amino acid CRGKA signature sequence, and an extended polyadenylation signal (Svärd *et al.,* 1998; Adam, 2001). One VSP, of hundreds in the *Giardia* genome, is expressed per *Giardia* cell and they are thought to protect the cells from proteolysis (Nash *et al.,* 1988). A similar strategy of surface protein expression is utilised by blood stage *T. brucei* cells (Aitcheson *et al.,* 2005). This method of antigen switching plays a major role in immune system avoidance and survival in vertebrate hosts. In *H. muscarum* the VSP domain-containing genes are predicted, by Phobius (Käll, Krogh and Sonnhammer, 2004), to encode proteins with 8-9% cysteine residues, and a single predicted transmembrane domain predicted at the C-terminus. Notably there were also ten VSP domain containing proteins down-regulated upon entry into stationary phase.

In addition to the VSP domain containing genes, several other putative surface proteins were differentially regulated upon entry to stationary phase; two putative amastin genes were highly upregulated, and eight transcripts which encode for proteins with the cytomegalovirus UL20A protein domain (PF05984), were downregulated in stationary phase *H. muscarum* cells. The functions of proteins with UL20a domains, including the domains namesake, are largely unknown. Deletion of UL20a from the human cytomegalovirus genome resulted in reduced viral production in infected fibroblasts (Van Damme and Van Loock, 2014). Further study will be required to elucidate the role of these proteins in trypanosomatids.

Transcription. The bias towards downregulated transcripts in the stationary phase cells as compared to log phase suggests a reduction of transcription and translation during stationary phase. Furthermore, five tRNA-synthase Pfam domains (PF00133.22, PF00749.21, PF00152.20, PF00587.25, PF01411.19) were significantly enriched in downregulated

transcripts (chi-squared, $p < 0.05$) and RNA polymerase III subunits were also downregulated. Overall, transcriptomic changes associated with cell surface remodelling, autophagy and reductions in transcription were observed in cells entering stationary phase. Cyclin expression patterns appear to suggest a bias in cells at G1 phase, as reported for *in vitro* culture of *T. brucei* procyclics (Matthews, 2005).

Transcriptome of *H. muscarum* inside *Drosophila melanogaster* compared to *in vitro* culture

To identify potentially important *H. muscarum* genes during the infection of *D. melanogaster* we sought to analyse the transcriptome of the trypanosomatid over the course of infection by RNA-sequencing analysis. RNA was purified from infected flies at 6, 12, 18, and 54 hours post-ingestion of *H. muscarum*. The resulting RNAs were sequenced and mapped to the concatenated genomes of *D. melanogaster* and *H. muscarum*. Reads were later resolved to the corresponding species. Here we will discuss the resulting transcriptome of *H. muscarum*: the transcriptome of *D. melanogaster* after ingestion of *H. muscarum* in the same experiment was discussed elsewhere (Wang *et al.,* 2019).

The number of reads which mapped to the *H. muscarum* genome ranged from 6949 to approx. 16.2 million reads per sample. At 6 hours post ingestion 40% of the total mapped reads were shown to map to *H. muscarum* (average of 3 biological replicates). This decreased to 20% in samples from 12 hours and 9% at 18 hours post ingestion. This correlates with the observed decrease in *H. muscarum* numbers as the parasite was cleared by *D. melanogaster* 18-54 hours post ingestion (Wang *et al.,* 2019). For differential expression analysis, only data up to 18 hours post infection was used as at 54 hours the

number of sequencing reads mapping to the *H. muscarum* genome dropped below 1% of the total number of mapped reads (Figure 6).

Principal component analysis (PCA) shows that the first two principal components of variation in mRNAs between *H. muscarum* from in vitro culture and *H. muscarum* after ingestion by *D. melanogaster* explained 58% and 10% of the variance in these data (Figure 6B). The PCA plot shows a high degree of difference between the *in vitro* samples and samples isolated from infected flies. The level of change in expression was much higher than between the two *in vitro* conditions discussed above.

For the infections, log phase *H. muscarum* cultures were used to feed the flies. In order to identify transcriptomic changes in *H. muscarum* associated with being ingested by the fly, we compared the transcriptome of *H. muscarum* cells from log phase *in vitro* culture to the in-fly transcriptomes. Over a third of the genome, 4,633 genes, was significantly differentially regulated (Wald test, adjusted p-value < 0.05) between log phase axenic culture samples and samples from infected flies (Table S18). Comparisons of gene expression between sequential time points over the course of infection revealed that there was a large initial transcriptomic change upon ingestion with 4662 genes differentially regulated between log phase culture and six hours post ingestion. This large initial transcriptomic shift was followed by more subtle transcriptomic changes between 6-12 (204 genes) and 12-18 hours (25 genes) (adjusted p-values < 0.05). Here we describe some of the changes in gene expression observed after ingestion and how these compare with other published transcriptome studies of trypanosomatids in their insect vectors including notable work by Inbar *et al.,* 2017 on genes expression of four morphologically distinct *Leishmania major* stages in a sand fly vector and Savage *et al.,* 2016 on *Trypanosoma brucei* in three tsetse fly tissues.

Herpetomonas muscarum* genes differentially regulated at six hours post-ingestion by *Drosophila melanogaster

Approximately a third of the *H. muscarum* genome was found to be significantly differentially expressed between log phase axenic culture and six hours post ingestion by *D. melanogaster* ($p < 0.05$) (Table S19). Of this subset, 640 genes had a fold change of ≥ 4 between the time points – highlighting the magnitude of the trypanosomatids response to ingestion. GO enrichment analysis, using Ontologizer (Vingron *et al.,* 2007), identified two significantly enriched GO terms in the 346 transcripts comparatively enriched at six hours post ingestion; OG0000045 (autophagosome assembly, $p = 0.0014$) and OG0003333 (amino acid transporters, $p= 0.0002$). Given the aforementioned lack of annotated GO terms in *H. muscarum*, we also looked at Pfam enrichment in the *H. muscarum* genes significantly upregulated upon ingestion by the fly. The top 15 represented Pfam domains in genes upregulated $\geq$ 4-fold at six hours post-ingestion are all significantly enriched compared to the full gene set (Table S20). Additionally, there were several Pfam domains enriched in the downregulated transcripts, which we discuss further below.

Leucine-rich repeat proteins. The most represented Pfam domain in genes upregulated at 6 hours post ingestion were the leucine-rich repeat (LRR) domains. LRRs are primarily known to be involved in protein-protein and protein-glycolipid interactions and are the major domain of the *Leishmania* protein surface antigens (PSAs), which are known virulence factors. Ten of the upregulated LRR-containing genes encode orthologues of the *Leishmania* PSAs (Figure 7A). The predicted protein structures for 8/10 of these transcripts consists of a single transmembrane domain at the N-terminus, with the majority of the protein predicted

to be on the external face of the cell (Table S21). One transcript encodes a protein with no predicted transmembrane domains and could therefore be a secreted protein. The remaining transcript encodes a protein with two predicted transmembrane domains, with the region between these domains on the external face of the cell. Other upregulated LRR-containing transcripts are putative adenylate cyclases. These proteins also feature prominently in the *T. brucei* genes which are differentially regulated upon ingestion by tsetse (Savage *et al.,* 2016). These signalling proteins likely assist in the coordination of the trypanosomatids' responses to the environment with its vector.

Cell surface genes. Seven of the top fifteen genes, 21/346 overall, upregulated in *H. muscarum* at six hours post ingestion by *D. melanogaster* contained the *Giardia* variant-specific surface protein (VSP) domain (PF03302.13). These genes are members of three distinct orthogroups. A heatmap showing the normalised read counts for these genes across all samples is shown in Figure 7B. Transmembrane domain prediction tools (Krogh *et al.,* 2001; Käll, Krogh and Sonnhammer, 2004) predict a single transmembrane domain at the N-terminus in the majority of predicted protein sequence for these genes. However, there were also eight transcripts without predicted transmembrane domains, which are predicted to be secreted proteins. The majority of these putative surface antigens are 769-781 amino acids in length, have a single predicted transmembrane helix at residues 7-29 (Table S21). As previously mentioned, many of these proteins are also upregulated by the cells upon entry into stationary phase, though not to the same levels. Additionally, several transcripts for VSP-containing proteins are downregulated in *H. muscarum* upon entry into the fly. These thirteen proteins are generally smaller than those upregulated at the same time point (95-501 amino acids) and tended to be part of orthogroup 11.

Thirty amastins, from 11 different orthogroups, were differentially regulated in *H. muscarum* at 6 hours post ingestion (Figure 7C). The majority (21) were upregulated upon entry into the fly, though 14 transcripts were also upregulated during stationary phase *in vitro* culture. Each orthogroup represented contained both up- and down-regulated genes. The function of this family of glycoproteins, are not well understood. In *Leishmania*, amastins are more commonly associated with macrophage-dwelling amastigote forms, where they are known to be important to both survival and virulence (de Pavia *et al.,* 2015). However, it has also been shown that β-amastins are upregulated during the insect stages of the life cycle in *T. cruzi* (Kangussu-Marcolino *et al.,* 2013). The *H. muscarum* amastins from orthogroup 18 share only 25-30% identity (across the whole sequence) to the two pairs of *T. cruzi* β-amastin alleles highlighted in this study. This may initially seem to be quite low, however the β-amastins have been shown to be highly divergent (18-25% identity) between *T. cruzi* strains (Kangussu-Marcolino *et al.,* 2013). Therefore, based on sequence alone, it is unclear which proteins may have parallel roles in the two trypanosomatid species.

Several other classes of surface protein genes were differentially expressed between log-phase axenic culture and six hours post-ingestion. Transcripts for proteins containing the Cytomegalovirus UL20A protein domain (PF05984) were significantly down regulated upon ingestion. Five of these genes were from orthogroup 11 – the same group as many of the down regulated VSP domain containing genes. Finally, sixteen (of the twenty-eight in the genome) *H. muscarum* orthologues to known *Leishmania* virulence factor, GP63, were significantly differentially regulated in the first six hours post ingestion by the fly. All but one of the differentially regulated GP63 orthologues were predicted to be GPI-anchored at the cell surface (GPI-SOM online tool, Fankhauser and Mäser 2005). The exception,

HMUS00892600.1, is predicted (THTMM v2.0, Krogh *et al.,* 2001) to have a single transmembrane domain and for the majority of the protein to be cytosolic. Most GP63 transcripts were upregulated in *H. muscarum* after ingestion (log2 foldchanges 0.29-2.73), however two putative GP63 genes, HMUS01311000 and HMUS01311200, were downregulated with log2 foldchanges of -1.94 and -1.58 respectively.

Stress-related genes. The insect gut is a hostile environment. The presence of digestive enzymes, changes in pH and the insect's gut microbiota make surviving a difficult challenge for any invading organisms. In correlation with this, a number of stress-associated genes and pathways are upregulated in *H. muscarum* upon entry into the fly. As previously mentioned, autophagy is an important process for survival in stressful conditions where fewer nutrients are available - such as in the midgut of an insect. Similar to observed in stationary phase axenic culture, twenty-six putative ATG8 genes were upregulated in *H. muscarum* at six hours post ingestion compared to log-phase axenic culture – suggesting extensive protein recycling is occurring in the cells. Additionally, 40 heat shock protein 83 genes were shown to be upregulated at six hours after ingestion. Heat shock proteins act as molecular chaperones which stabilise other proteins, help them to fold correctly and be regulated after damage in stressful conditions. The upregulation of these genes provides further evidence that these cells are in a stressed state.

Metabolism. There was significant enrichment of putative amino acid, pteridine and sugar transporters in the upregulated transcripts. These included the amino acid transporters (AATs) orthologous to the *Leishmania* amino acid permease 3 (AAP3), AAT11, AAT12 and AAT20. AAP3 has been shown to be arginine specific and is linked to virulence in *Leishmania donovani* infections in humans (Darlyuk *et al.,* 2009). AAT11 is upregulated in during stress

responses associated to purine starvation (Marchese *et al.,* 2018). In *Leishmania major*, AAP3 and AAT20 were strongly upregulated in the motile, gut-dwelling nectonomad forms (Inbar *et al.,* 2017). These transporters have been shown to transport neutral amino acids across the cell membrane, notably proline and alanine, which can be used as alternative carbon sources by trypanosomatids and are abundant in insect vectors haemolymph.

Six putative pteridine transporters were also upregulated in *H. muscarum* at 6 hours post ingestion. Pteridines are needed by trypanosomatids to produce enzyme cofactors such as biopterin. *Leishmania* parasites are unable to synthesize their own pteridines (Cunningham and Beverly 2001) and as such must scavenge them from their environment. It is not currently known if *H. muscarum* is also a pteridine auxotroph, however like the *Leishmania* species, the cells appear to scavenge from the environment upon entry into the fly.

Several transcripts putatively involved in lipid metabolism were downregulated in *H. muscarum* following ingestion by *D. melanogaster*, including triglyceride lipases and members of the biotin/lipoate protein ligase (BLPL) family. This contrasts what has been observed in *L. major* in the midgut of sand flies where genes from these families were upregulated (Inbar *et al.* 2017). Therefore, whilst upregulation of pteridine and amino acid transporters appears to be a conserved trypanosomatid response to being ingested by insects, lipid metabolism during insect infection may differ between trypanosomatid genera.

Gene expression-related transcripts. Consistent with the differential expression of many genes upon entry into the fly, and therefore a predicted increase in chromatin remodelling and translation activity, there was upregulation of histones (2A, 3 and 4), RNA polymerase subunits 1 and 2, putative 40S/60S ribosomal proteins and putative 28S beta rRNAs in H.

muscarum after ingestion by the fly. This result is consistent with what has been reported in *T. brucei* where the 40S and 60S ribosomal subunits were amongst the most highly upregulated genes in cells isolated from the midgut and proventriculus of *G. morsitans* (Savage *et al.,* 2016).

Cell cycle. Upon ingestion by the fly there was strong upregulation of putative G1-associated cyclins 4, 7 and 11 as well as the G1 associated cyclin-related kinase 1 (CRK1) (Tu and Wang, 2004, 2005). Cyclin 6, cyclin 8 and CRK9, which are associated with the G2/mitosis transition (Hammarton *et al.,* 2003; Gourguechon and Wang, 2009), were slightly downregulated suggesting a reduction in cell replication at six hours post ingestion (Table 6). Consistent with this there was also downregulation of putative DNA polymerase kappa, the theta DNA polymerase subunit and mitochondrial DNA polymerase subunits. Furthermore meiosis-associated genes NBS1, Rad50 and SPO11 were also downregulated.

Given the apparent reduction replication rate in *H. muscarum* cells at six hours after ingestion, the upregulation of nine tubulin genes (3 alpha- and 6 beta-tubulins) is likely to accommodate the changes in cell morphology, rather than to produce new daughter cells. Tubulin upregulation is also observed in *T. brucei* isolated from the midgut and proventriculus of *Glossina morsitans* (Savage *et al.,* 2016), though these cells are replicative – as such the 'motivation' for increased tubulin gene expression may be different.

Differentiation and RNA-binding proteins. It is well documented that (human) disease-causing trypanosomatids have several life cycle stages within their respective vectors. Coordinated differentiation between these discrete stages requires a suite of RNA-binding proteins (RBPs) which regulate parasite gene expression (Kolev *et al.,* 2014). Despite the lack

of observed differentiated forms in infections of *D. melanogaster*, several differentiation associated-RBPs are differentially regulated in the trypanosomatid after infection including RBP10 and hnRNP F/H. These proteins have been shown to regulate gene expression in *T. brucei* blood-stream forms (Wurst *et al.,* 2012, Gupta *et al.,* 2013). RNAi knockdown of RBP10 in bloodstream trypanosomes resulted in the down-regulation of a large number of bloodstream form mRNAs (Wurst *et al.,* 2012). The same study showed that overexpression of the protein in procyclics led to an increase of many bloodstream-form specific mRNAs, including genes involved in sugar transport. This is likely owing to the fact blood is a glucose-rich environment and the cell will attempt to utilize this ready carbon source (Smith *et al.,* 2017). Three out of the four orthologues of *Tb*RBP10 were strongly (> 4-fold) upregulated in *H. muscarum* cells after ingestion by *D. melanogaster*. During feeding experiments sucrose is added to the *H. muscarum* culture media to encourage the flies to feed. As such these genes may be unregulated in response to increased sugars available in the environment.

However, several other cell-cycle regulating RBPs associated with blood-stream form trypanosomes were also upregulated in *H. muscarum* after ingestion by the fly, including zinc-finger domain-containing RBPs ZC3H11 and ZC3H18. The former is essential in bloodstream-form trypanosomes and is involved in protection from heat shock, whilst depletion of ZC3H18 delayed blood stream form-to-procyclic differentiation in *T. brucei* (Benz *et al.,* 2011, Droll *et al.,* 2013). As such the situation may be more complex than solely metabolism-driven expression changes.

In addition to parallels with blood-stream form trypanosomes, transcripts for ALBA3/4 proteins (named for their 'acetylation lowers binding affinity' domain) were significantly downregulated in *H. muscarum* upon entry into the fly. In *T. brucei*, these proteins are expressed in all stages, except those found in the tsetse proventriculus. RNAi knockdown of

these proteins in *T. brucei* axenic procyclics resulted in elongation of the cell body and repositioning of the nucleus and the kinetoplast to resemble the epimastigote cell-stage (Subota *et al.,* 2011). As such the reduction in ALBA3/4 transcripts suggests there may be parallels between trypanosomes during the latter stages of tsetse infection and *H. muscarum* during *D. melanogaster* infection.

Other differentially regulated RNA-binding proteins with as yet unclear roles in differentiation included: the essential gene expression regulation protein RBP42 and ZC3H12, a protein associated with differentiation (Kolev *et al.,* 2014).

Herpetomonas muscarum genes differentially regulated between six- and twelve-hours post-ingestion by _Drosophila melanogaster_

There were 204 genes which were differentially regulated between six- and twelve-hours post ingestion (p-adjusted < 0.05), 161 of these had a fold change of ≥ 2 with just 31 genes upregulated at the latter timepoint (Table S22). Hypothetical proteins lacking functional information dominated the highly upregulated genes. The most enriched transcript at 12 hours post ingestion encodes a putative surface protein, the top blastp hit for which was the *Giardia* variant-specific surface protein VSP136-4. This suggests VSP domain-containing proteins continue to be important throughout infection of the fly. Two DNA replication and repair associated transcripts were also upregulated at 12 hours post ingestion: an orthologue of *T. brucei* cell division cycle protein 45 (CDC45), and tyrosyl-DNA

phosphodiesterase-like protein. CDC45 is part of the CMG (Cdc45·Mcm2–7·GINS) complex which functions as a helicase during DNA replication (Dang and Li, 2011) and may also play a role in DNA repair (DeBrot, Lancaster and Bjornsti, 2016). Furthermore, Tyrosyl-DNA phosphodiesterases are involved in the repair of topoisomerase-related DNA damage (Kawale and Povirk, 2018). These observations indicate that *H. muscarum* cells are under genotoxic stress after ingestion by *D. melanogaster*.

Herpetomonas muscarum genes differentially regulated between twelve- and eighteen-hours post-ingestion by _Drosophila melanogaster_

In the 23 genes found to be upregulated at 18 hours post ingestion (compared to at 12 hours, Table S23) genes involved in binding to damaged DNA (OG00033330) were significantly enriched. Only two of these transcripts were able to be assigned putative functions: eukaryotic replication factor A and a structure-specific endonuclease. This observation provides further evidence of genotoxic stress in *H. muscarum* after ingestion, as indicated by other upregulated DNA repair genes at 12 hours post-ingestion.

The most highly upregulated transcript at 18 hours post ingestion was an orthologue of the *Leishmania major* UDP-galactose transporter LPG5B. This protein allows import of UDP-galactose into the golgi body where they are used to synthesize phosphoglycans. Capul *et al.,* (2007) showed that, in *L. major*, loss of LPG5b resulted in cells with defects in proteophosphoglycans (PPG) (Capul *et al.,* 2007). PPGs are known virulence factors and are found in membrane bound, filamentous and secreted forms. The viscous secreted PPG is thought to protect the *L. major* in the gut and may also force the fly to regurgitate the infective *Leishmania* cells in to the bite wounds of vertebrates.

Herpetomonas muscarum genes differentially regulated between stationary phase in vitro culture and in-fly samples

Comparisons between stationary phase in vitro culture and in-fly samples revealed 5102 differentially expressed genes (adjusted p-value < 0.05). Approximately 55% of the genes differentially regulated between *in vitro* and in-fly samples were the same for log phase vs in-fly and stationary phase vs. in-fly comparisons (Figure 8). However, 1639 genes were only significantly differentially regulated in stationary phase vs in-fly comparisons (Table S24). Genes differentially regulated between log phase in vitro culture and in-fly samples have already been discussed, we will now outline the genes only differentially regulated when the transcriptomes of stationary phase in vitro samples of *H. muscarum* are compared with those after ingestion by *D. melanogaster*. Of the 1639 genes, 750 had a fold change of ≥ 2 – approximately a third of which were upregulated in *H. muscarum* after ingestion by *D. melanogaster*.

Half of the top ten in-fly enriched transcripts were TATE (telomere associated mobile elements) DNA transposons and among the most represented Pfam hits in the fly-enriched transcripts were reverse transcriptase (PF00078.27) and phage integrase (PF00589.22) domains (Table S25). Though TATE DNA transposons comprise 1.32% of the *Leishmania major* genome, very little is known about these transposable elements, other than that they contain a tyrosine recombinase (Pita *et al.,* 2019). It is possible that these transposable elements are more mobile in *H. muscarum* cells ingested by the fly. However, we predict that the overall level of transcription of cells in stationary phase cultures are reduced (vs. log phase, see above). As such, the comparative increase in TATE transposon transcription

between stationary phase cells and *H. muscarum* from *Drosophila* may not be specifically a result of ingestion, but a reflection of general transcription levels in the two groups of cells.

As previously discussed, transcripts for several proteins containing a *Giardia* VSP domain are enriched in stationary phase compared to log phase in axenic culture. However, five were shown to be even more abundant in the *H. muscarum* cells ingested by *D. melanogaster*. Two other putative surface antigens were also enriched in ingested *H. muscarum* which contained a domain similar to Cytomegalovirus UL20A glycoprotein and the domain of unknown function DUF4148.

Transcripts encoding for putative anti-oxidant proteins were significantly enriched in *H. muscarum* after ingestion by the fly. Enriched Pfam domains in the upregulated gene set included thioredoxin, glutathione S-transferase and alkyl hydroperoxide reductase (AhpC)/thiol specific antioxidant (TSA) domains. Our previous work showed that the *D. melanogaster* response to *H. muscarum* ingestion included the production of reactive oxygen species (Wang, Sloan and Ligoxygakis, 2019), as such the upregulation of these anti-oxidant proteins is likely an attempt to cope with this insect immune response.

Conclusion. Here we have described the genome and predicted proteome of the monoxenous trypanosomatid *Herpetomonas muscarum* and characterised the transcriptome of the parasite both in culture and inside the gut of its natural host *D. melanogaster*. *H. muscarum* shows similarity in both genome structure and content to *Leishmania*, with significant synteny to *Leishmania major* and sharing 80% of orthogroups with other members of the subfamily Leishmaniinae. While most *Herpetomonas* genes have orthologs in other trypanosomatids, a number of genes found elsewhere appear to have been lost in *Herpetomonas*, in particular genes associated with the specialised life stages of

dixenous trypanosomatids. We might expect loss of some mammal-stage specific genes, such as HASPs, HERPs and sphingolipid synthesis genes important in metacyclic *Leishmania* cells, but more surprising might be the loss of genes expressed in insect stages such as BARPs and procyclins.

The transcriptome of *Herpetomonas* inside its insect host also showed strong parallels with the responses of *Leishmania* promastigotes inside the sand fly gut, in particular both parasites showing significant upregulation of PSAs and GP63 (this study; Inbar *et al.,* 2017). These proteins have been shown to be associated with virulence in *Leishmania* and are important for establishment of parasite infection in the midgut, and so for transmission. The extensive changes in transcript abundance of genes likely to be expressed on the cell surface during insect infections includes a number of gene families not known to be important in dixenous trypanosomatids (e.g. related to *Giardia* variant surface protein) implies that a dynamic cell surface may be a shared feature of trypanosomatid life cycles beyond dixenous groups (Jackson 2016), and that even more diversity of surface proteins may be present in the monoxenous trypanosomatids, supporting findings from free-living kinetoplastids (Jackson *et al.,* 2016). We also note that the majority of the genes showing changes in expression later in insect infections are hypothetical, including many hypothetical genes conserved with other trypanosomatids. This reflects similar findings in better-studied dixenous parasites (Inbar *et al.,* 2017; Savage *et al.,* 2016) and highlights how much we still have to learn about the interactions between trypanosomatids and their insect host.

In the wild, there is little data pertaining to the percentage of sand flies with established *Leishmania* infection in endemic regions. In this context, the parallel to the more accessible *Drosophila-Herpetomonas* system is important, as the genetic component

of the parasite that influences midgut establishment is easier to determine. However, more work is needed to ascertain whether genes upregulated in *Leishmania* and those in *Herpetomonas* are truly functionally related. The limitation is the difference between the lifestyles of these insects. Most strikingly, female sand flies become infected with *Leishmania* during blood feeding, while Drosophila is never haematophagous. Nevertheless, sand flies are also plant feeders, so there is some overlap in the ecological niche as well as in their basic biology. The presence of trypanosomatids is another shared feature of the midgut landscape of these flies, and our data suggest that at least some aspects of the molecular interaction between flies and trypanosomatids may also be conserved.

Materials and Methods

Herpetomonas muscarum culture. *H. muscarum* were cultured in supplemented BHI (3% brain heart infusion broth, 2.5mg/ml haemin, 1% FCS) and incubated at 28°C. For most experiments, cells were maintained in a log phase of growth by splitting every 3 days.

Infection of *D. melanogaster* (see Wang *et al.*, 2018). For each independent infection of a group of 20-30 flies, 10^7 *H. muscarum* cells were harvested from a 3 days-old culture (which showed the highest infectivity rate from our experience) and resuspended in 500ul 1% sucrose. The parasite solution was then transferred to a 21mm Whatman Grade GF/C glass microfibre filter circle (Fisher Scientific). Circles containing the parasite cells were placed into standard *Drosophila* small culture vial without any food. The flies used in the infections were 4-5 days old before they were starved overnight. After starvation, the flies were

transferred to food vials that contained the Whatman circles with the parasite cells. After 6h of feeding, flies were moved and reared on standard yeast/molasses medium. At different time points post oral infection, infected flies were collected for downstream experiments and frozen at -80°C for molecular analyses.

DNA extraction for genome sequencing. Genomic DNA was extracted from 100 million *H. muscarum* cells from log phase cells from *in vitro* culture using the Norgen Biotek Genomic DNA extraction kit according to the manufacturer's instructions.

RNA extraction for RNA-seq. 8ml of *H. muscarum* promastigote culture at a density of 9.25 $x10^6$ cells per ml (measured by haemocytometer) was diluted 1:40 in supplemented BHI and divided between 4 tissue culture flasks. The immediate post-dilution density was 6.5 $x10^5$ cells per ml. The following day the cell density was measured to be 1.18 x 10^6 cells per ml. 45ml was taken from each flask and the cells pelleted by centrifugation for 10 mins at 1000xg. The supernatant was discarded and the Norgen Biotek RNA Purification kit was (according to manufacturer's instructions) used to purify RNA from the cell pellet. This process was repeated for 5.3ml of the remaining culture three days later when the cell density was $1.21x10^7$ cells per ml. The resulting RNA was eluted at concentrations 97-170 ng per µl with a 260/230 absorbance 1.86-2.19.

Reference Genome To produce the reference genome Illumina and Pacific Biosciences sequencing platforms were used. For Illumina sequencing 1ug of genomic DNA was sheared into 300–500 base pair (bp) fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology, AFA Inc., Woburn, USA). An amplification-free Illumina library was

prepared (Kozarewa *et al.,* 2009) and 150 bp paired-end reads were generated on an Illumina MiSeq following the manufacturer's standard sequencing protocols (Bronner *et al.,* 2014). For the Pacific Biosciences SMRT technology, 8 µg of genomic DNA was sheared to 20-25kb by passing through a 25mm blunt ended needle. A SMRT bell template library was generated using the Pacific Biosciences issued protocol (20 kb Template Preparation Using BluePippin(tm) Size-Selection System). After a greater than 7kb size-selection using the BluePippin(tm) Size-Selection System (Sage Science, Beverly, MA) the library was sequenced using P6 polymerase and chemistry version 4 (P6C4) on 6 single-molecule real-time (SMRT) cells (Eid *et al.,* 2009).

The Pacific Bioscience reads were assembly with HGAP3 (Chin *et al.,* 2013), with genome size parameter set to 25Mb, to produce 285 contigs. The obtained assembly was then corrected with ICORN2 (Otto *et al.,* 2010), for five iterations. Using the Argus Optical Mapping System from OpGen, an optical map was generated from high molecular weight genomic DNA captured in agarose plugs and the restriction enzymes KpnI and BamHI. The data was analysed with associated MapManager and MapSolver software tools (http://www.opgen.com/products-services/argus-system). The optical map consisted of 37-38 chromosomes with approximately half being contiguous. With the information obtained from the optical map and REAPR (Hunt *et al.,* 2013), manual genome improvement was performed on the PacBio assembly to produce a final genome assembly of 181 contigs. Analysis of the frequency distribution of Kmers was performed using GenomeScope version 1.0 (Vurture *et al.,* 2017) with the kmer frequencies estimated using Jellyfish (Marçais and Kingsford 2011) using the default parameters suggested in the GenomeScope manual.

Transcriptomic libraries Poly-A mRNA was purified from total RNA using oligodT magnetic beads and strand-specific indexed libraries were prepared using the KAPA Stranded RNA-Seq kit followed by ten cycles of amplification using KAPA HiFi DNA polymerase (KAPA Biosystems). Libraries were quantified and pooled based on a post-PCR Agilent Bioanalyzer and 75 bp paired-end reads were generated on the Illumina HiSeq v4 following the manufacturer's standard sequencing protocols (as above).

Data Release All sequencing data was submitted to the European Nucleotide Archive (ENA) under accession number ERP008869.

Genome annotation. CRAM output files containing RNA sequencing reads from both *H. muscarum* in vitro culture and infected D. melanogaster were converted to fastq format and then mapped to the genome sequence using the next generation sequencing reads alignment package HISAT2 version 2.1.0 (Kim, Langmead and Salzberg, 2015). The mapped reads from each sample were assembled into transcripts with the Cufflinks package version 2.2.1 (Trapnell *et al.,* 2012) and merged to form a single transcripts set for all reads. The Companion annotation tool [ref:Steinbiss *et al.,* 2016] was then used to generate several genome annotation files based on the RNA sequencing transcriptomic evidence and pre-existing gene models from three other trypanosomatids – *Leishmania braziliensis*, *L. major* and *Trypanosoma brucei* (individual annotation statistics Table S26).

Orthofinder proteome analysis. The following proteomes were inputted into the Orthofinder script; *Trypanosoma brucei brucei* 927 v5.1 (Berriman *et al., 2005*), *Trypanosoma brucei gambiense* DAL972 v3 (Jackson *et al.,* 2010), *Trypanosoma congolense*

IL3000 (Jackson *et al.,* 2012), *Trypanosoma cruzi* (CL Brener) (El-Sayed *et al.,* 2005), *Trypanosoma evansi* STIB805 (Carnes *et al.,* 2015), *Trypanosoma grayi* ANR4 v1 (Kelly *et al.,* 2015), *Trypanosoma rangeli* SC_58 v1 (Stoco *et al.,* 2014), *Trypanosoma theileri* Edinburgh (Kelly *et al.,* 2017), *Trypanosoma vivax* Y486 (Jackson *et al.,* 2012), *Leishmania braziliensis* M2903 (Peacock *et al.,* 2007), *Leishmania donovani* BPK282 v1 (Downing *et al.,* 2011), *Leishmania infantum* JPCM5 (Peacock *et al.,* 2007), *Leishmania major* Friedlin v6 (Ivens *et al.,* 2005), *Leptomonas pyrrhocoris* ASM129339v1 (Flegontov *et al.,* 2016), *Leptomonas seymori* ASM129953v1 (Kraeva *et al.,* 2015), *Crithidia bombi* (Schmid-Hempel *et al.,* 2018), *Crithidia expoeki* (Schmid-Hempel *et al.,* 2018), *Crithidia fasciculata* v14.0 (Runkel *et al.,* 2014), *Angomonas deanei* (Motta *et al.,* 2013), *Phytomonas* EM1 (Porcel *et al.,* 2014) and *Bodo saltans* v3 (Jackson *et al.,* 2008). Where possible the above sequences were obtained from TriTrypDB v41 (Aslett *et al.,* 2010).

RNAseq Analysis *in vitro* culture. CRAM output files were converted to fastq format and then mapped to the concatenated *D. melanogaster* and *H. muscarum* genome sequences using the hisat2 (Kim *et al.,* 2015) mapper. Mapped reads were then counted using HTseq-count (v. 0.10.0, Anders *et al.,* 2014) and differential expression analysed using the DESeq2 package in R (Love *et al.,* 2014).

RNAseq analysis samples from whole flies. Total RNA of 8-10 flies at 6h, 12h, 18h post *H. muscarum* oral infection was extracted with total RNA purification kit from Norgen Biotek following the manufacturer's instruction. Each time point was repeated in three independent experiments. cDNA libraries were prepared with the Illumina TruSeq RNA Sample Prep Kit v2. All sequencing was performed on the Illumina HiSeq 2000 plaftform

using TruSeq v3 chemistry (Oxford Gene Technology, OGT). All sequence was paired-end and performed over 100 cycles. Read files (Fastq) were generated and then mapped to the concatenated *D. melanogaster* and *H. muscarum* genome sequences using the hisat2 (Kim *et al.,* 2015) mapper. Mapped reads were then counted using HTseq-count (v. 0.10.0, Anders *et al.,* 2014) and differential expression analysed using the DESeq2 package in R (Love *et al.,* 2014).

Acknowledgements

KD, TDO, MJS and JAC were supported by Wellcome via their core support for the Wellcome Sanger Institute (WSI) through grant 206194. Work in Oxford was supported by a Consolidator grant from the European Research Council (310912 Droso-Parasite, to PL), project grant BB/K003569 from the BBSRC (to PL) and a Wellcome Trust doctoral scholarship (to MAS). We thank the staff of the DNA pipelines at WSI for sequencing and generating sequencing libraries.

References

Acosta-Serrano, A., Vassella, E., Liniger, M., Renggli, C., Brun, R., Roditi, I. and Englund, P. 2001. The surface coat of procyclic *Trypanosoma brucei*: Programmed expression and proteolytic cleavage of procyclin in the tsetse fly. *Proceedings of the National Academy of Sciences*. 98(4), pp. 1513–1518.

Agami R. and Shapira M. 1992. Nucleotide sequence of the spliced leader RNA gene from *Leishmania mexicana* amazonensis. *Nucleic Acids Research*. 20(7):1804

Anders, S., Pyl, P. T. and Huber, W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2): pp. 166–169.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Garjria, B., Gao, X., *et al.,* 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research.* 38(37): D457-D462

Barclay, M. R. and Postell, F. J. 2007. Axenic Cultivation of *Phytomonas davidi* Lafont (Trypanosomatidae), a Symbiote of Laticiferous Plants (Euphorbiaceae)*. *The Journal of Protozoology* 23(2): pp. 238–241.

Bates, P. A. 2008. *Leishmania* sand fly interaction: progress and challenges. *Curr Opin Microbiol* 11(4): 340-344.

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(5733): 416 LP-422.

Beverley, S.M., 2003. Protozomics: trypanosomatid parasite genetics comes of age. *Nature Reviews Genetics,* 4(1): pp.11

Borghesan, T. C., Ferreira, R. C., Takata, C. S., Campaner, M., Borda, C. C., Paiva, F., Milder, R. V., Teixeira, M. M., Camargo, E. P. 2013. Molecular Phylogenetic Redefinition of *Herpetomonas* (Kinetoplastea, Trypanosomatidae), a Genus of Insect Parasites Associated with Flies. *Protist*. Urban & Fischer, 164(1): pp. 129–152.

Carnes, J., Anupama, A., Balmer, O., Jackson, A., Lewis, M., Brown, R., Cestari, I., Desquesnes, M., Gendrin, C., Hertz-Fowler, C. 2015. Genome and Phylogenetic Analyses of *Trypanosoma evansi* Reveal Extensive Similarity to *T. brucei* and Multiple Independent Origins for Dyskinetoplasty. *PLOS Neglected Tropical Diseases*. Public Library of Science 9(1): e3404.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569

Cunningham, M. L. and Beverley, S. M. 2001. 'Pteridine salvage throughout the Leishmania infectious cycle: implications for antifolate chemotherapy. *Molecular and Biochemical Parasitology*. Elsevier, 113(2): pp. 199–213.

Chen, J., Rauch, C. A., White, J. H., Englund, P. T., Cozzarelli, N. R. 1995. The topology of the kinetoplast DNA network. *Cell*, 80(1): pp. 61–69.

Chicharro, C. and Alvar J. 2013. Lower trypanosomatids in HIV/AIDS patients. *Annals of Tropical Medicine & Parasitology,* 97:sup1, 75-78.

Daniels, J. P., Gull, K., & Wickstead, B. 2010. Cell biology of the trypanosome genome. *Microbiology and molecular biology reviews: MMBR*, *74*(4): pp. 552–569.

DaRocha, W.D., Otsu, K. and Teixeira, S.M., 2004. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracyclineinducible T7 promoter system in *Trypanosome cruzi*. *Mol Biochem Parasitol*, 133(2): pp.175-186

de Paiva, R. M. C., Grazielle-Silva, V., Cardoso, M. S., Nakagaki, B. N., Mendonça-Neto, R. P., Monte, A., Canavaci, C., Melo, S. M., Martinelli, P. M., Fernandes, A. P., da Rocha, W. D. and Teixeira, S. M. 2015. Amastin Knockdown in *Leishmania braziliensis* Affects Parasite-Macrophage Interaction and Results in Impaired Viability of Intracellular Amastigotes. *PLOS Pathogens*, 11(12): e1005296

Dhiman, R. C. and Yadav, R. S. 2016. Insecticide resistance in phlebotomine sandflies in Southeast Asia with emphasis on the Indian subcontinent. *Infectious Diseases of Poverty*. London: BioMed Central, 5: pp. 106.

Downing, T., Imamura, H., Decuypere, S., Clark, T. G., Coombs, G. H., Cotton, J. A., Hilley, J. D., de Doncker, S., Maes, I., Mottram, J. C., *et al.* 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res*, 21(12): pp. 2143-56.

Eid, J. L., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., *et al.* 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323: pp. 133-138

Ellis, M., Sharma, D. K., Hilley, J. D., Coombs, G. H., Mottram, J. C. 2002. Processing and Trafficking of *Leishmania mexicana* GP63: Analysis using GP18 mutants deficient in glycosylphosphatidylinositol protein anchoring. *J. Biol. Chem.,* 277: pp. 27968–27974

El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A-N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., et al. 2005. The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease. *Science,* 309(5733): 409 LP-415.

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C. *et al.,* 2005. Comparative Genomics of Trypanosomatid Parasitic Protozoa. *Science*, 309(5733): pp. 404 LP-409.

Emms, D. and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology.* 16(157).

Fawaz, E. Y., Zayed A. B., Fahmy, N. T., Villinsk, J. T. , Hoel, D. F., Diclaro, J. W. 2016. Pyrethroid insecticide resistance mechanisms in the adult *Phlebotomus papatasi* (diptera: Psychodidae). *Journal of Medical Entomology*, 53(3): pp. 620–628.

Fernandes, M. C., Dillon, L. A. L., Belew, A. T., Bravo, H. C., Mosser, D.M., El-Sayed, N. M. 2016. Dual Transcriptome Profiling of *Leishmania*-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. *mBio*, 7(3):e00027-16.

Fiorini, J. E., Takata, C. S., Teofilo, V. M., Nascimento, L. C., Faria-e-Silva, P. M., Soares, M. J., Teixeira, M. M., De Souza, W. 2001. Morphological, biochemical and molecular characterization of *Herpetomonas samuelpessoai* camargoi n. subsp., a Trypanosomatid isolated from the flower of the squash *Cucurbita moschata*. *Journal of Eukaryotic Microbiology*, 48(1): pp. 62–69.

Flegontov, P., Butenko, A., Firsov, S., Kraeva, N., Eliáš, M., Field, M. C., Filatov D., Flegontova, O., Gerasimov, E. S., Hlaváčová**, J.** *et al.,* 2016. Genome of *Leptomonas pyrrhocoris*: A high-quality reference for monoxenous trypanosomatids and new insights into evolution of Leishmania. *Scientific Reports*, 6.

Fragoso, C. M., Schumann Burkard, G., Oberle, M., Renggli, C. K., Hilzinger, K., Roditi, I. 2009. PSSA-2, a Membrane-Spanning Phosphoprotein of *Trypanosoma brucei*, Is Required for Efficient Maturation of Infection. *PLoS ONE*, 4(9):e7074.

Hart, D.T. and Coombs, G. H. 1982. *Leishmania mexicana*: energy metabolism of amastigotes and promastigotes. *Exp. Parasitol*, 54: pp. 397–409.

Hassan, M. M., Widaa, S. O., Osman, O. M., Numiary, M. S., Ibrahim, M. A., Abushama, H. M. 2012. Insecticide resistance in the sand fly, *Phlebotomus papatasi* from Khartoum State, Sudan. *Parasites & Vectors*, 5(1): pp. 46.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T. D. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biol,* 14: R47

Inbar, E., Hughitt, V. K., Dillon, L. A. L., Ghosh, K., El-Sayed, N. M. and Sacks, D. L. 2017. The Transcriptome of *Leishmania major* Developmental Stages in Their Natural Sand Fly Vector, *mBio*, 8(2). doi: 10.1128/mBio.00029-17.

Iraad, F., Bronner, M. A., Quail, D. J., Turner, H. S. 2014. Improved Protocols for Illumina Sequencing. *Curr. Protoc. Hum. Genet.,* 80:18(2): pp. 1-42.

Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M. A., Adlem, E., Aert, R., *et al.* 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, 309(5733): pp. 436-42.

Jackson A. P. 2007. Origins of amino acid transporter loci in trypanosomatid parasites. *BMC evolutionary biology*, *7*, 26.

Jackson, A. P., Quail, M. A., Berriman, M. 2008. Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). *BMC Genomics*. 9(9):594.

Jackson, A. P., Sanders, M., Berry, A., McQuillan, J., Aslett, M. A., Quail, M. A., Chukualim, B., Capewell, P., MacLeod, A., Melville, S. E., Gibson, W., Barry, J. D., Berriman, M., Hertz-Fowler, C. 2010. The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African trypanosomiasis. *PLoS Negl. Trop. Dis*., 13(4):e658.

Jackson, A. P., Berry, A., Aslett, M., Allison, H. C., Burton, P., Vavrova-Anderson, J., Brown, R., Browne, H., Corton, N., Hauser, H. *et al.* 2012. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. PNAS, 109: pp. 3416–3421.

Jackson, A. P., Otto, T. D., Aslett, M., Armstrong, S. D., Bringaud, F., Schlacht, A., Hartley, C., Sanders, M., Wastling, J. M., Dacks, J. B., Acosta-Serrano, A., Field, M. C., Ginger, M. L., Berriman, M. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Curr Biol*., 26(2): pp. 161-172.

Jackson, A. P. 2016. Gene family phylogeny and the evolution of parasite cell surfaces. *Molecular and Biochemical Parasitology,* 209(1): pp. 64–75.

Käll, L., Krogh, A., Sonnhammer, E. L. L. 2004. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, 338(5): pp. 1027-1036.

Kamhawi, S., Ramalho-Ortigao, M., Pham, V. M., Kumar, S., Lawyer, P. G., Turco, S. J., Barillas-Mury, C., Sacks, D. L., Valenzuela, J. G. 2004. A role for insect galectins in parasite survival. *Cell*. 119: pp. 329–341.

Kangussu-Marcolino, M. M., de Paiva, R. M. C., Araújo, P. R., de Mendonça-Neto, R. P., Lemos, L., Bartholomeu, D. C., Mortara, R. A., da Rocha, W. D., Teixeira, S. M. R. 2013. Distinct genomic organization, mRNA expression and cellular localization of members of two amastin sub-families present in *Trypanosoma cruzi*. *BMC Microbiology*, 13(1): pp. 10.

Kelly, S., Ivens, A., Manna, P. T., Gibson, W., Field, M. C. 2014. A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. *Sci Data*. 5(1):140024.

Kelly, S., Ivens, A., Mott, G. A., O'Neill, E., Emms, D., Macleod, O., Voorheis, P., Tyler, K., Clark, M., Matthews, J., Matthews, K., Carrington M. 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 Biol Evol*., 9(8): pp. 2093-2109.

Kim, D., Langmead, B. and Salzberg, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12, pp. 357.

Kraeva, N., Butenko, A., Hlaváčová, J., Kostygov, A., Myškova, J., Grybchuk, D., Leštinová, T., Votýpka, J., Volf, P., Opperdoes, F., Flegontov, P., Lukeš, J., Yurchenko, V. 2015. *Leptomonas seymouri*: Adaptations to the Dixenous Life Cycle Analyzed by Genome Sequencing,

Transcriptome Profiling and Co-infection with *Leishmania donovani*. *PLoS Pathog*., 11(8):e1005127.

Krieger, S., Schwarz, W., Ariyanayagam, M. R., Fairlamb, A. H., Krauth-Siegel, R. L. and Clayton, C. 2000. Trypanosomes lacking trypanothione reductase are avirulent and show increased sensitivity to oxidative stress. *Molecular Microbiology*, 35: pp. 542-552.

Krogh. A., Larsson, B., von Heijne G., and Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305(3): pp. 567-580.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., Turner, D. J. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods,* 6: pp. 291–295.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol.,* 5(2): R12.

Lamontagne, J. and Papadopoulou, B. 1999. Developmental Regulation of Spliced Leader RNA Gene in *Leishmania donovani* Amastigotes Is Mediated by Specific Polyadenylation. *Journal of Biological Chemistry,* 274(10): pp. 6602–6609.

Lang, T., Warburg, A., Sacks, D. L., Croft, S. L., Lane, R. P. 1991. Transmission and scanning EM-immunogold labeling of *Leishmania major* lipophosphoglycan in the sandfly *Phlebotomus papatasi*. *Eur. J. Cell Biol.*, 55: pp. 362–372.

Lee, S. H., Stephens, J. L., Paul, K. S., Englund, P. T. 2006. Fatty Acid Synthesis by Elongases in Trypanosomes. *Cell*, 126(4): pp. 691–699.

Liang, X. Haritan, A., Uliel, S., Michaeli, S. 2003. trans and cis Splicing in Trypanosomatids: Mechanism, Factors, and Regulation. *Eukaryotic Cell*, 2(5): pp. 830–840.

Love, M. I., Huber, W., Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15: pp. 550.

Lukeš, J., Guilbride, D. L., Votýpka, J., Zíková, A., Benne, R., Englund, P. T. 2002. Kinetoplast DNA Network: Evolution of an Improbable Structure. *Eukaryotic Cell*, 1(4): pp. 495–502.

Lye L. F., Owens, K., Shi, H., Murta, S. M. F., Vieira, A. C., Turco, S. J., Tschudi, C., Ullu, E., Beverley. 2010. Retention and Loss of RNA Interference Pathways in Trypanosomatid Protozoans. *PLOS Pathogens*, 6(10): e1001161.

Marçais, G. and Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* 27(6): pp. 764–770.

Marín, C., Fabre, S., Sánchez-Moreno, M., Dollet, M. 2007. *Herpetomonas spp.* isolated from tomato fruits (*Lycopersicon esculentum*) in southern Spain'. *Experimental Parasitology*, 116(1): pp. 88–90.

Maslov, D. A., Votýpka, J., Yurchenko, V., Lukeš, J. 2013. Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends in Parasitology*, 29(1): pp. 43–52.

Mcconville, M. J., Turco, S. J., Ferguson, M. A. J., Sacks, D, L. 1992. Developmental modification of lipophosphoglycan during the differentiation of *Leishmania major* promastigotes to an infectious stage. *EMBO J.*, 11: pp. 3593–3600.

Morio, F., Reynes, J., Dollet, M., Pratlong, F., Dedet, J-P., Ravel, C. 2008. Isolation of a protozoan parasite genetically related to the insect trypanosomatid *Herpetomonas samuelpessoai* from a human immunodeficiency virus-positive patient. *Journal of Clinical Microbiology*, 46(11): pp. 3845–3847.

Motta, M. C. M., Martins, A. C., de Souza, S. S., Catta-Preta, C. M., Silva, R., Klein, C. C., de Almeida, L. G., de Lima Cunha, O., Ciapina, L. P., Brocchi, M. 2013. Predicting the Proteins of

Angomonas deanei, *Strigomonas culicis* and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family. *PLoS ONE*. 8(4): e60209.

Mungube, E. O., Vitouley, H. S., Allegye-Cudjoe, E., Diall, O., Boucoum, Z., Diarra, B., Sanogo, Y., Randolph, T., Bauer, B., Zessin, K. H., Clausen, P. H. 2012. Detection of multiple drug-resistant *Trypanosoma congolense* populations in village cattle of south-east Mali. *Parasites & Vectors*, 5: pp. 155.

Opperdoes, F. R. and Coombs, G. H. 2007. Metabolism of *Leishmania*: proven and predicted. *Trends Parasitol.*, 23: pp. 149–158.

Otto, T. D., Sanders, M., Berriman, M., Newbold, C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinforma. Oxf. Engl.* 26: pp. 1704–1707.

Pacheco, **R. S.**, Marzochi, M. C. A., Pires, M. Q., Brito, C. M. M., Madeira, M. F., Barbosa-Santos, E. G. O. 1998. Parasite genotypically related to a monoxenous trypanosomatid of dog's flea causing opportunistic infection in an HIV positive patient. *Mem. Inst. Oswaldo. Cruz*.

Peacock. C. S., Seeger, K., Harris, D., Murphy, L., Ruiz, J. C., Quail, M. A., Peters, N., Adlem, E., Tivey, A., Aslett, M., Kerhornou, A., *et al.* 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet*. 39(7): pp. 839-47.

Pereira, F., Santos-Mallet, J. R., Branquinha, M. H., d'Avila-Levy, C. M., Santos, A. L. 2010. Influence of leishmanolysin-like molecules of *Herpetomonas samuelpessoai* on the interaction with macrophages. *Microbes and infection,* doi: 10.1016/j.micinf.2010.07.010

Pimenta, P. F. P., Turco, S. J., Mcconville, M. J., Lawyer, P. G., Perkins, P. V., Sacks, D. L. 1992. Stage-specific adhesion of *Leishmania* promastigotes to the sandfly midgut. *Science*, 256: pp. 1812–1815.

Pimenta, P. F. P., Saraiva, E. M. B., Rowton, E., Modi, G. B., Garraway, L. A., Beverley, S. M., Turco, S. J., Sacks, D. L. 1994. Evidence that the vectorial competence of phlebotomine sand flies for different species of Leishmania is controlled by structural polymorphisms in the surface lipophosphoglycan. *Proc. Natl. Acad. Sci. USA*, 91: pp. 9155–9159

Podlipaev, S., Votýpka, J., Jirků, M., Svobodová, M., Lukes J. 2004. *Herpetomonas ztiplika* n. sp. (Kinetoplastida: Trypanosomatidae): a parasite of the blood-sucking biting midge *Culicoides kibunensis* Tokunaga, 1937 (Diptera: Ceratopogonidae). *The Journal of Parasitology*, 90(2): pp. 342–347.

Ponte-Sucre, A., Gamarro, F., Dujardin, J-C., Barrett, M. P., López-Vélez, R., García-Hernández, R., Pountain, A. W., Mwenechanya, R., Papadopoulou, B. 2017. Drug resistance and treatment failure in leishmaniasis: A 21st century challenge. *PLOS Neglected tropical diseases*, 11(12).

Robinson, K.A. and Beverley, S.M., 2003. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite *Leishmania*. *Molecular and biochemical parasitology,* 128(2): pp.217-228.

Runckel, C., DeRisi, J., Flenniken, M. L. 2014. A draft genome of the honey bee trypanosomatid parasite *Crithidia mellificae*. *PLoS ONE*, 9(4).

Porcel, B. M., Denoeud, F., Opperdoes, F., Noel, B., Madoui, M-A., Hammarton, T. C., Field, M. C., Da Silva, C., Couloux, A., Poulain, J., *et al.* 2014. The streamlined genome of *Phytomonas spp.* relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS genetics.* e1004007.

Savage, A. F. *et al.* 2016. Transcriptome Profiling of *Trypanosoma brucei* Development in the Tsetse Fly Vector *Glossina morsitans*. *PloS ONE,* 11(12): pp. e0168877–e0168877.

Schmid-Hempel, P., Aebi, M., Barribeau, S., Kitajima, T., du Plessis, L., Schmid-Hempel, R., Zoller, S. 2018. The genomes of *Crithidia bombi* and *C. expoeki*, common parasites of bumblebees. *PLoS ONE*, 13(1).

Shaked-Mishan, **P.** , Suter-Grotemeyer, M. , Yoel-Almagor, T. , Holland, N. , Zilberstein, D. and Rentsch, D. 2006. A novel high-affinity arginine transporter from the human parasitic protozoan *Leishmania donovani*. *Molecular Microbiology*, 60: pp. 30-38.

Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S., Cross, G. A. M. 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Research*. 38(15): pp. 4946-4957.

Simpson, L. and Thiemann, O. H. 1995. Sense from nonsense: RNA editing in mitochondria of kinetoplastid protozoa and slime molds. *Cell*, 81: pp. 837–840.

Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., Otto, T. D. 2016. Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Research*, 44: pp. W29-W34.

Stoco, P. H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., *et al.* 2014. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. *PLOS Neglected Tropical Diseases*, 8(9): e3176.

Stuart, K., Allen, T. E., Heidmann, S., Seiwert, S. D. 1997. RNA editing in kinetoplastid protozoa. *Microbiology and Molecular Biology Reviews*, 61(1): pp. 105–120.

Sutterwala, S. S., Hsu, F., Sevova, E. S., Schwartz, K. J., Zhang, K., Key, P., Turk, J., Beverley, S. M., Bangs, J. D. 2008. Developmentally regulated sphingolipid synthesis in African trypanosomes. *Molecular Microbiology*, 70: pp. 281-296.

Teixeira, S. M., de Paiva, R. M., Kangussu-Marcolino, M. M., Darocha, W. D. 2012. Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. *Genetics and Molecular Biology*, 35(1): pp. 1–17.

Thomas, S., A. Green, N. R. Sturm, D. A. Campbell, and P. J. Myler. 2009. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. *BMC Genomics*, 10: pp. 152.

Tovar, J., Wilkinson, S., Mottram, J. C., Fairlamb, A. H. 1998. Evidence that trypanothione reductase is an essential enzyme in Leishmania by targeted replacement of the *tryA* gene locus. *Molecular Microbiology*, 29: pp. 653-660.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7: pp. 562.

Urwyler, S., Studler, E., Renggli, C. K., Roditi, I. 2007. A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of *Trypanosoma brucei*. *Mol Microbiol.,* 63: pp. 218–228

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., Schatz, M. C. 2017. GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, 33(14): pp. 2202–2204.

Zhang, K., Showalter, M., Revollo, J., Hsu, F. F., Turk, J., Beverley, S. M. 2003. Sphingolipids are essential for differentiation but not growth in Leishmania. *EMBO J.,* 22: pp. 6016–6026.

Wang, L., Sloan, M., Ligoxygakis, P. 2018. Intestinal NF-κB and STAT signalling is important for uptake and clearance in a *Drosophila-Herpetomonas* interaction model. *PLoS Genetics.*

WHO, E. C. on V. B. 1980. Resistance of vectors of disease to pesticides: fifth report of the WHO Expert Committee on Vector Biology and Control. [meeting held in Geneva from 3 to 9 June 1980].

WHO, E. C. on V. B. 1986. Resistance of vectors and reservoirs of disease to pesticides. Tenth report of the WHO Expert Committee on Vector Biology and Control. *World Health Organization - Technical Report Series*, 737: pp. 1–87.

Yurchenko, V., Kostygov, A., Havlová, J., Grybchuk-Ieremenko, A., Ševčíková, T., Lukeš, J., Ševčík, J., Votýpka, J. 2016. Diversity of Trypanosomatids in Cockroaches and the Description of *Herpetomonas tarakana sp. n.*', *Journal of Eukaryotic Microbiology*. 63(2): pp. 198–209.

Zídková, L., Cepicka, I., Votypka, J., Svobodová, M. 2010. *Herpetomonas trimorpha sp. nov.* (Trypanosomatidae, Kinetoplastida), a parasite of the biting midge *Culicoides truncorum* (Ceratopogonidae, Diptera). *International Journal of Systematic and Evolutionary Microbiology*. 60(9): pp. 2236–2246.

Figure legends:

Figure 1 - Average coverage depth of *H. muscarum* scaffolds > 100 kb. The solid line shows the global median read coverage. The dashed line shows 1.5x and the dotted line shows 2x the global median read coverage respectively. In blue are scaffolds which were mapped, by PROmer (Kurtz *et al.,* 2004), to the *L. major* chromosome 31 - sequences of 300bp which map with > 70% identity. The shade of blue represents the proportion of the scaffold which was mapped.

Figures 2A-F - Synteny and colinearity between *H. muscarum* and other trypanosomatids. As an example, this plot shows co-linearity between *H. muscarum* genes (genes highlighted in blue) on scaffold 40 and: **A.** *L. major* chromosome 1 (genes highlighted in red). **B.** *Phytomonas EM1* scaffolds HF955082, HF955140 and HF955140 (genes highlighted in green) **C.** *Leptomonas pyrrhocoris* scaffolds LpyrH10_33 and LpyrH10_41 (genes highlighted in pink) **D.** *Crithidia bombi* scaffolds (genes highlighted in yellow) OESO01000125 and OESO01000148. **E.** *Trypanosoma brucei* chromosomes 9 and 11 (genes highlighted in purple). Scaffold/Chromosome labels show length in bp. This data was produced using Promer alignments (Delcher *et al.,* 2002). Ribbons between scaffolds show windows of >100 amino acid (translated) align with at least 50% identity. This data was visualised using Circos (Krzywinski et al. 2009). To quantify these relationships, we investigated all windows of consecutive genes with single-copy orthology in *H. muscarum* in comparison to *L. major*, *T. brucei brucei* and *Leptomonas seymouri*. **F.** Shows the proportions of these windows for which all genes occurred on a single scaffold in the comparison genome (syntenic windows), and the proportion of those for which all gene occurred in the same order as in *H.*

muscarum (colinear windows) for a range of window sizes from 3 to 60 genes. Numbers of windows included in the comparisons varies from 1926 windows of 3 single-copy orthologs with *L. major* to 49 windows of 60 adjacent genes with single-copy orthologs in *T. brucei*. Note that synteny values are also affected by the degree of continuity of the comparison species genome for *Leptomonas seymouri*.

Figure 3 A. Alternating tubulin arrays in *H. muscarum*. Scaffolds 22 and 67 were found to have two loci containing alternating putative alpha (red) and beta (blue) tubulin genes. Several of these genes we predict to be tubulin pseudogenes (alpha - pink, beta - light blue) as they contain tubulin domains but also contain sequence consistent with non-LTR transposons. **B. A monotypic beta tubulin locus in *H. muscarum*.** Four copies of a putative beta tubulin (blue) were found in tandem on *H. muscarum* scaffold 20. This locus appears similar to the single copy beta tubulin locus on *L. major* chromosome 8 as the order of adjacent genes (grey) is conserved. We also see synteny with a locus in *T. brucei* on chromosome 5, however the beta tubulin gene is absent. Dotted lines indicate orthologous genes. Blue lines indicate orthologous beta tubulin genes.

Figure 4 - Relationship between *H. muscarum* and other trypanosomatids. A. Phylogeny based on all orthogroups containing a single gene from each species. Other panels show Venn-Euler diagrams in which the areas of each eliptical section are approximately proportional to the number of orthogroups shared by each of (**B)** *H. muscarum*, *L. major* and *T. brucei brucei*; **(C)** *H. muscarum*, *Leptomonas pyrrhocoris* and *L. seymouri* and **(D)** *H. muscarum*, *Leishmania donovani* and *L. major*. Diagram layouts were generated by EulerApe v2.0.3 (Micallef and Rodgers 2014).

Figure 5 - A global view of gene family sharing between trypanosomatids. (a) The numbers of gene families (orthogroups; pink bars; values on left-hand y-axis) and the numbers of genes in those groups (blue bars; values on right-hand y-axis) with particular patterns of sharing between high-level groups in our Orthofinder data. Shading in the lower panel from pink to blue represents how widespread each set of families are, with pink representing families specific to one group and dark blue those families present in all groups. (b) Scatterplot of gene family size against the number of species a family is present in, with each point representing a single gene family (families with less than 3 genes in total are excluded), and points coloured according to the number of higher-level taxonomic groups they are shared between, as in the lower part of panel (a). [code to draw this diagram is a modified version of UpSetR]

Figure 6 - A. RNA-seq reads extracted from infected flies (whole) which mapped to *H. muscarum* genome. Error bars show the standard error of the mean. **B. Principal component analysis of differentially expressed *H. muscarum* genes in log phase culture vs. samples isolated from infection flies at 6, 12 and 18 hours post ingestion.** There are two clear sample groupings (circled) which correspond to RNA from l in in vitro culture log phase cells and RNA isolated from infected flies. Different shades of blue indicate the sample origin (n=3 per condition).

Figure 7 - Heat map of normalised, log transformed counts for differentially expressed *Herpetomonas muscarum* surface proteins. A. *Herpetomonas muscarum* orthologues to the *Leishmania* promastigote surface antigens. **B.** Transcripts encoding proteins with a *Giardia*

variant surface protein (PF03302.13) domain. The black bar indicates the genes from orthogroup 11 which are mostly downregulated upon ingestion of *H. muscarum* by the fly. **C.** Differentially regulated ***Herpetomonas muscarum*** **amastin genes.** Log = log phase axenic culture samples, Stat = stationary phase axenic culture samples. 6h = six hours post ingestion by *D. melanogaster*, 12h = twelve hours post ingestion by *D. melanogaster*, 18h = eighteen hours post ingestion by *D. melanogaster.*

Figure 8 - Venn diagram showing the numbers of genes differentially expressed in *Herpetomonas muscarum* between two *in vitro* culture conditions and after ingestion by *Drosophila melanogaster*.

Table legends:

Table 1 - Summary statistics of the *H. muscarum* genome

Table 2 - Alignment of highly conserved splice leader sequences (bases 1-40 of mini-exon gene) of *H. muscarum* and other species from the Leishmaniiae clade. The variable AT-rich region (positions 11-19) is shown coloured by genus. *Herpetomonas sp.* appear to have an additional A or T residue, dependant on species at position 11.

Table 3 - Summary of Orthofinder analysis of 13 trypanosomatid genomes. (*Trypanosoma rangeli, Trypanosoma grayi, Trypanosoma brucei brucei, Trypanosoma brucei gambiense,*

Trypanosoma vivax, Trypanosoma congolense, Leishmania donovani, Leishmania major, Leishmania mexicana, Leptomonas pyrococcus, Leptomonas seymori, Crithidia fasciculata and *Bodo saltans).*

Table 4 - Summary of *H. musccarum* proteins orthologous to important *T. brucei* proteins.

Table 5 – Significantly differentially regulated cyclins and cyclin-related kinases between stationary and log phase *Herpetomonas muscarum*.

Table 6 – Cell cycle-associated proteins differentially expressed in *H. muscarum* upon ingestion by *D. melanogaster*

Supplementary figure and table legends:

Figure S1 – GenomeScope kmer profile and model for *H. muscarum* genome

Figure S2 - Principal component analysis of differentially expressed *H. muscarum* genes in log phase culture vs. stationary phase culture. There are two clear sample groupings (circled) which correspond to RNA each condition (n=3 per condition). Dark blue = log phase samples and light blue = stationary phase samples.

Table S1 – Coordinates of putative strand switch regions in the *H. muscarum* genome

Table S2 - BLAST hits for the *Phytomonas serpens* spliced leader sequence in the *H. muscarum* genome.

Table S3 - Alignment of intronic region of the mini-exon gene from several trypanosomatids of the Leishmaniiae clade. The first 15bp of the intron sequences appear to be conserved across the clade with the sequence becoming more variable thereafter.

Table S4 - Protein orthogroups from Orthofinder analysis (in full)

Tables S4-S15 - *H. musccarum* proteins orthologous to important *T. brucei* proteins (in full). Tables: **4** - Metabolism, **5** - Differentiation and RNA, **6** - RNAi, **7** - Phosphatases, **8** - Protein kinases, **9** - GP63, **10** - Mucins, **11** - *H. muscarum* unique proteins, **12** - Kinetochore, **13** - Spliceosome, **14** - Exosome and **15** - Nuclear proteins.

Table S16 - *H. muscarum* genes differentially expressed between log and stationary phase *H. muscarum in vitro* culture

Table S17 – Pfam domains significantly enriched in differentially regulated *H. muscarum* genes upon entry into stationary phase during axenic culture (vs. log phase).

Table S18 – *H. muscarum* genes differentially expressed between log phase in vitro culture and after ingestion by *D. melanogaster* (all samples).

Table S19 - *H. muscarum* genes differentially expressed between log phase in vitro culture and 6 hours after ingestion by *D. melanogaster*

Table S20 - Significantly enriched Pfam domains in differentially regulated *Herpetomonas muscarum* genes at six hours post ingestion by *Drosophila melanogaster* (vs log-phase axenic culture). The table shows the top 10 represented Pfam domains in the significantly up- and downregulated genes. Chi-squared tests were performed to test for statistically significant enrichment of the Pfams frequency in upregulated genes vs. the Pfams in the whole genome.

Table S21 –Structural predictions for differentially expressed H. muscarum surface proteins. Structural predictions were acquired using the TMHMM1.0 online tool (Krogh et al., 2001).

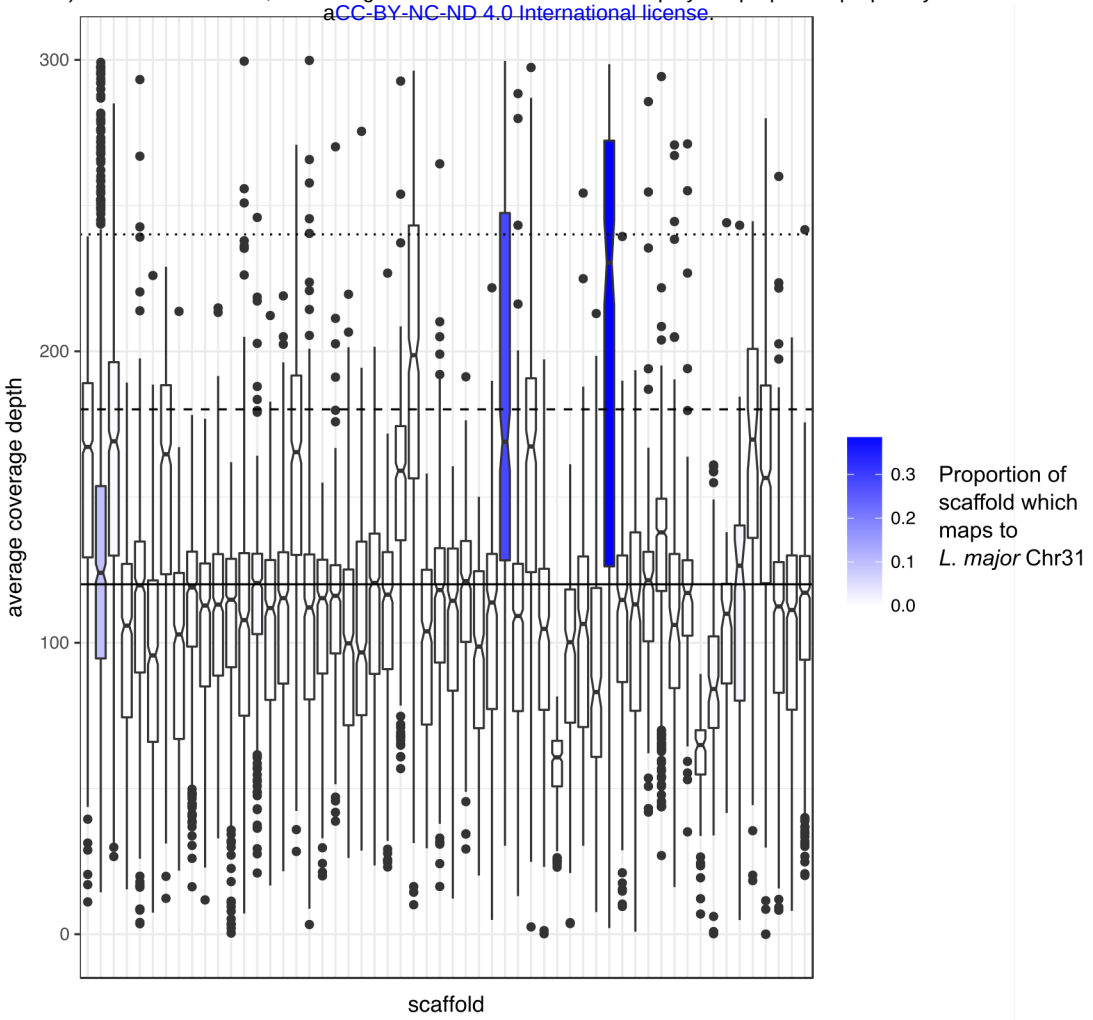
Table S22 - *H. muscarum* genes differentially expressed between 6 and 12 hours after ingestion by *D. melanogaster*

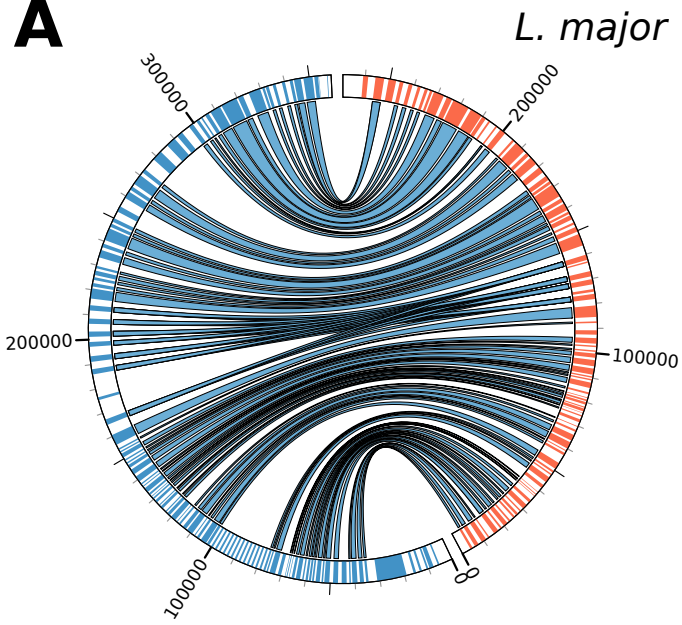
Table S23 - *H. muscarum* genes differentially expressed between 12 and 18 hours after ingestion by *D. melanogaster*

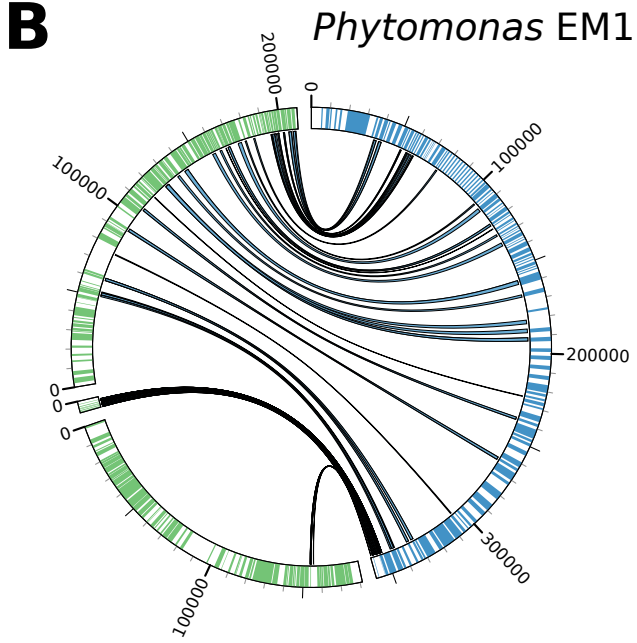
Table S24 - Table of genes differentially regulated between *Herpetomonas muscarum* after ingestion by *Drosophila melanogaster* vs. stationary phase axenic culture and not log phase axenic culture, p-adjusted < 0.05.

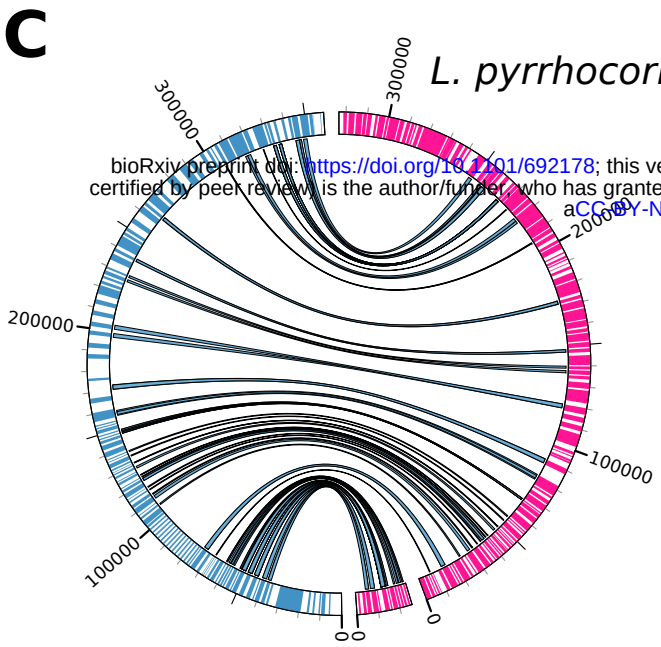
Table S25 - The top 10 represented Pfam domains in *Herpetomonas muscarum* ingested by *Drosophila melanogaster* vs. stationary phase axenic culture

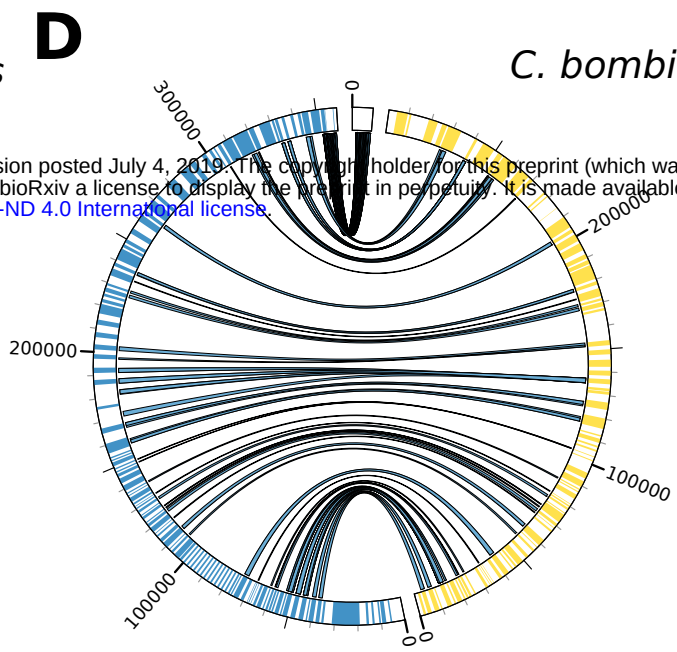
Table S26 - Summary statistics of three *H. muscarum* genome annotations using gene models from *L. major*, *L. braziliensis* and *T. brucei* and a maximal annotation which combines all three annotations.

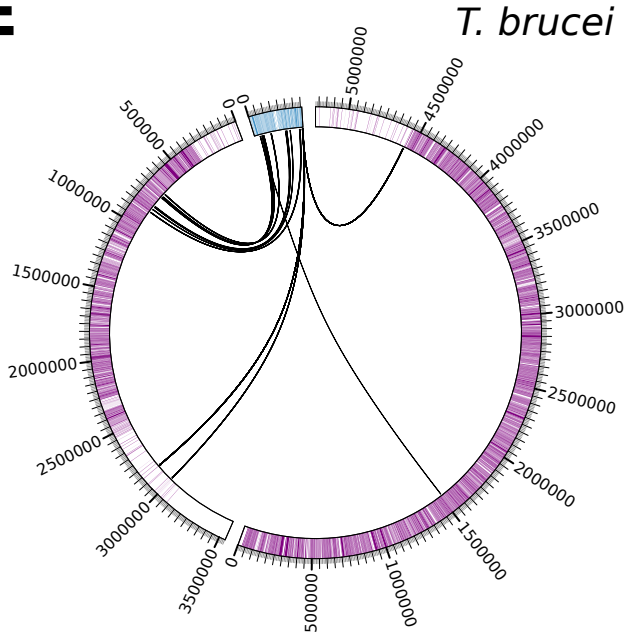


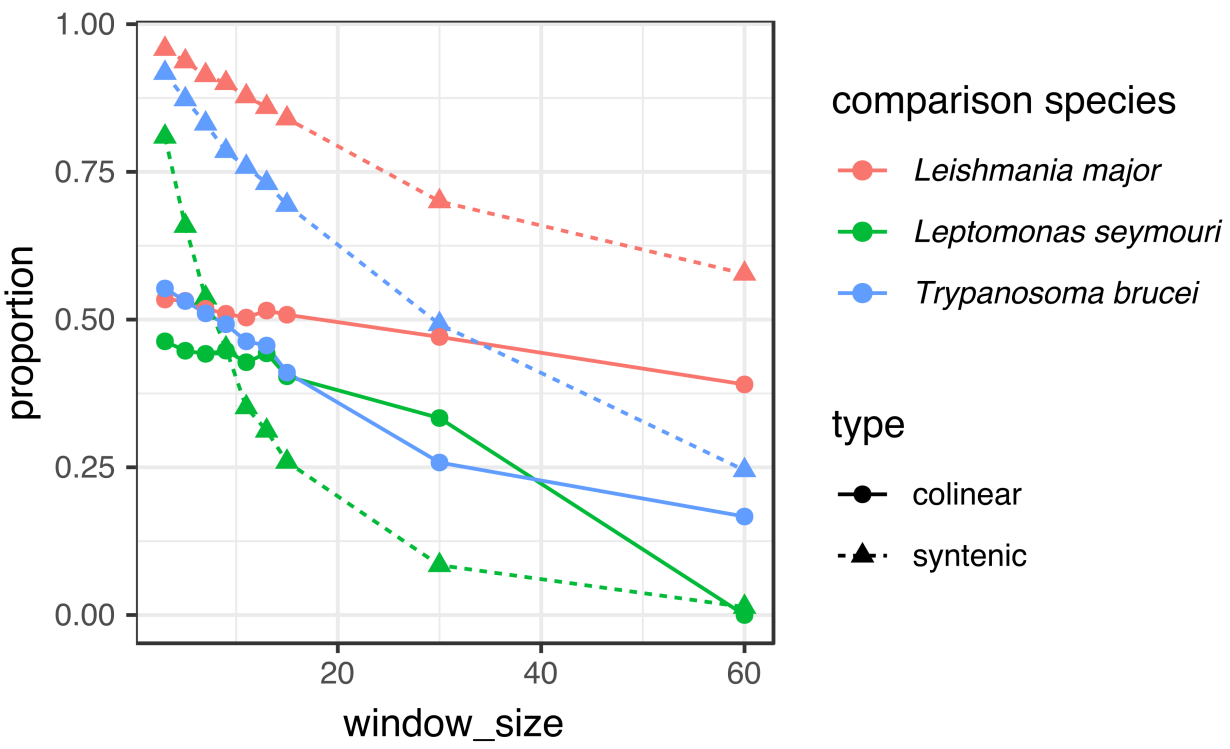


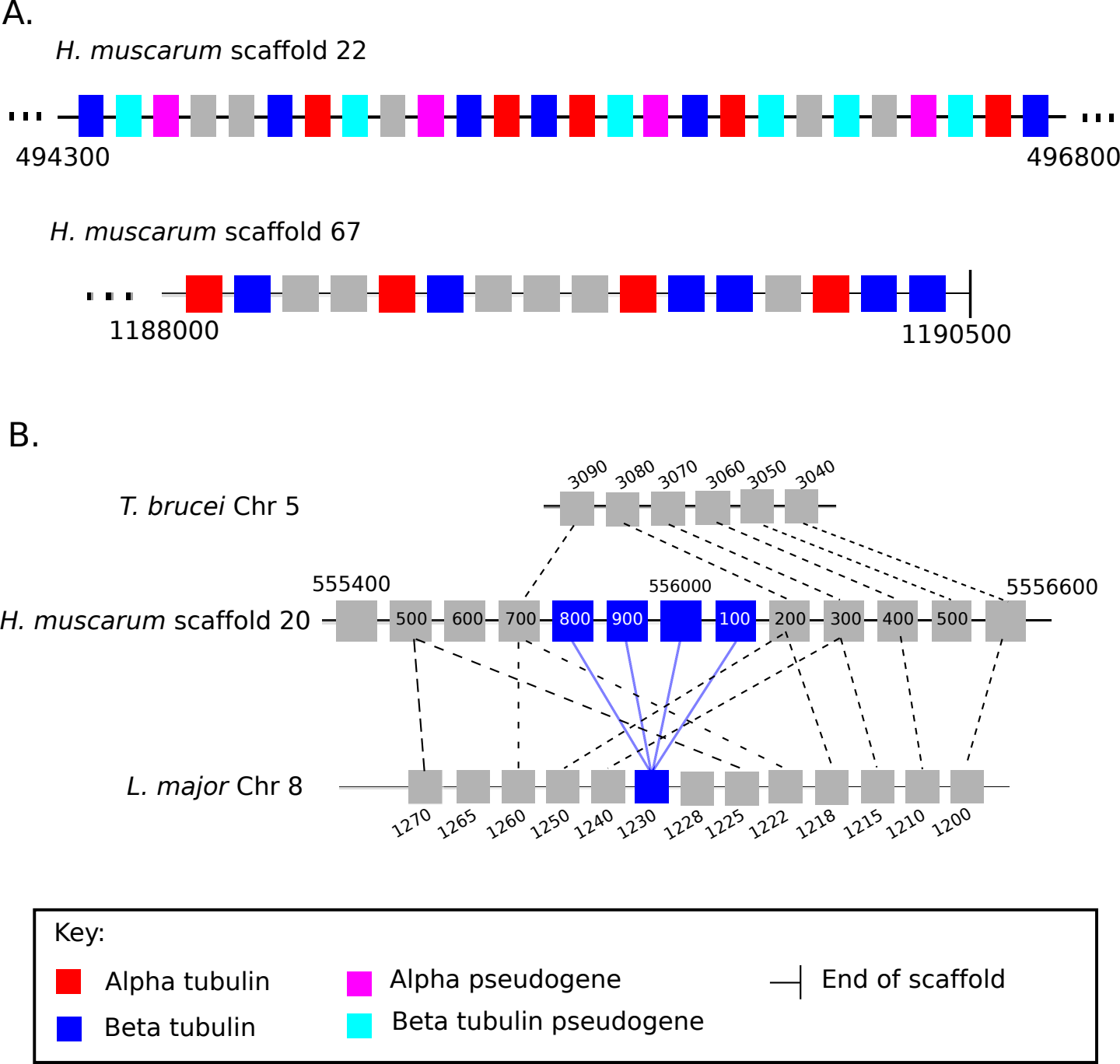


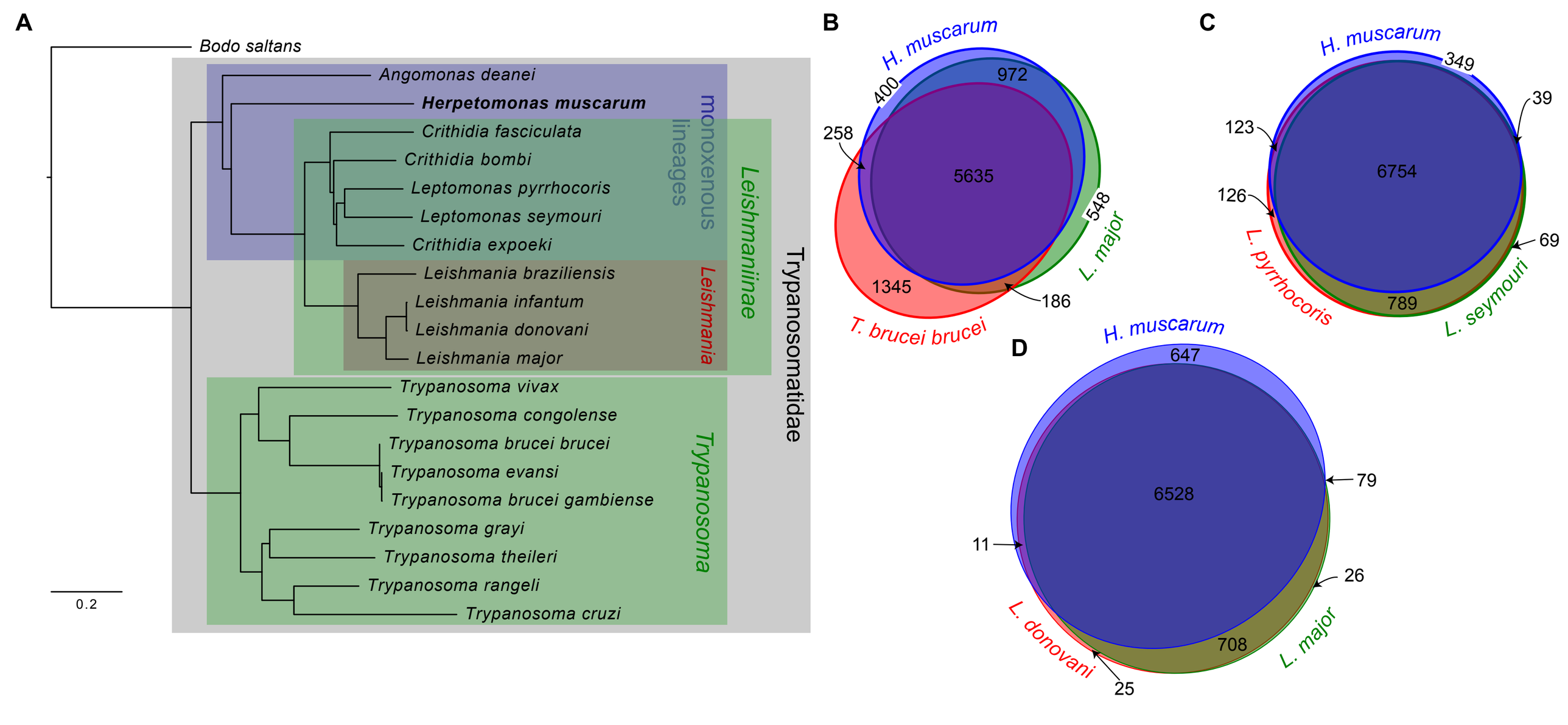


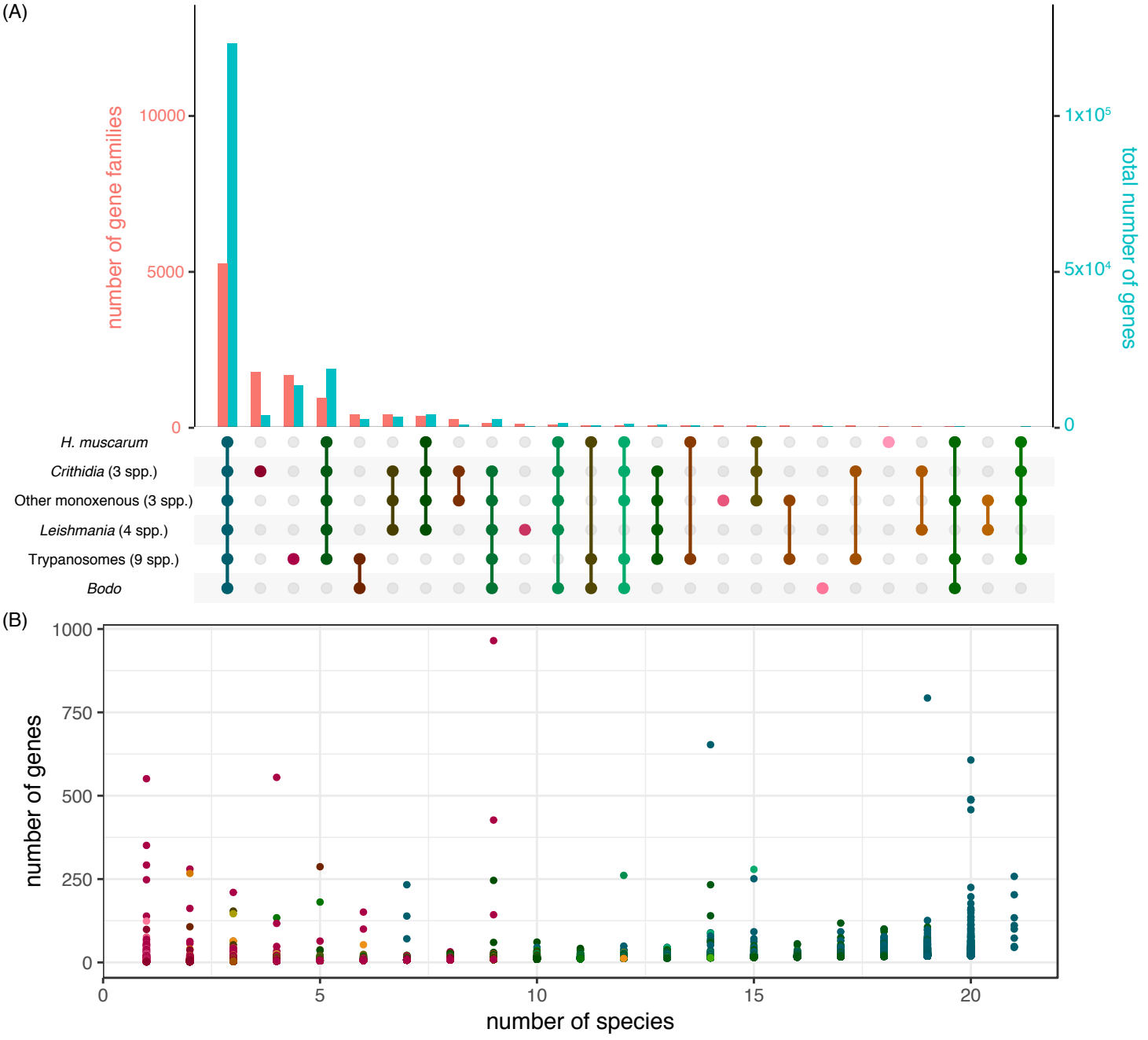




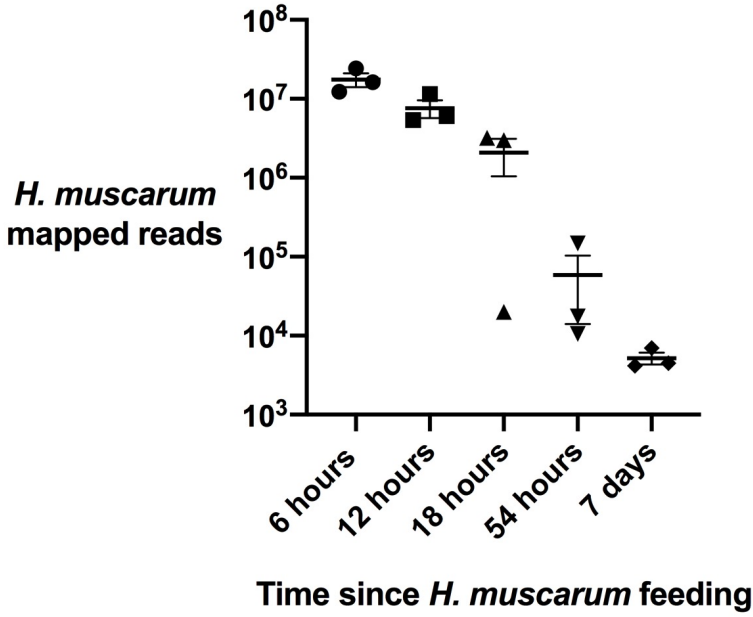



A.
H. muscarum scaffold 22
494300
496800
H. muscarum scaffold 67
1188000
1190500
B.
T. brucei Chr 5
3090
3080
3070
3060
3050
3040
555400
556000
5556600
H. muscarum scaffold 20
500
600
700
800
900
100
200
300
400
500
L. major Chr 8
1270
1265
1260
1250
1240
1230
1228
1225
1222
1218
1215
1210
1200
Key:
Alpha tubulin
Beta tubulin
Alpha pseudogene
Beta tubulin pseudogene
End of scaffold

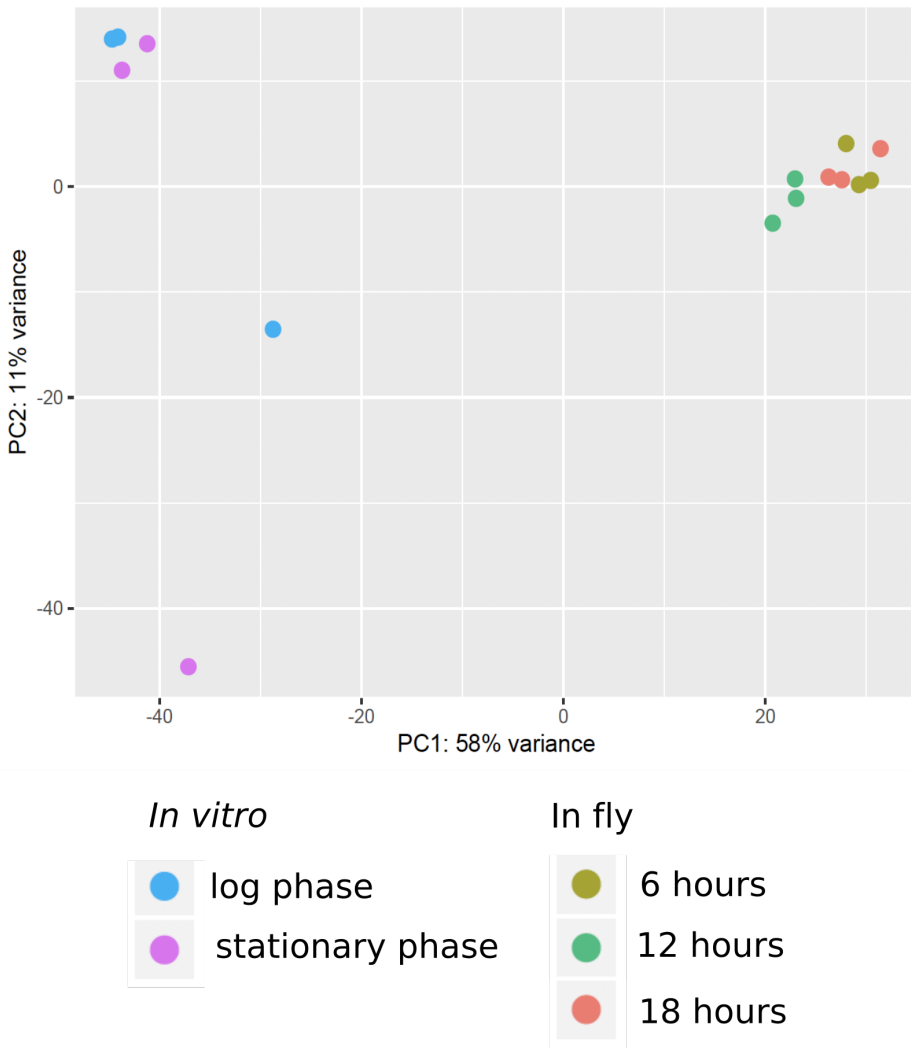

A
Bodo saltans
Angomonas deanei
Herpetomonas muscarum
Crithidia fasciculata
Crithidia bombi
Leptomonas pyrrhocoris
Leptomonas seymouri
Crithidia expoeki
Leishmania braziliensis
Leishmania infantum
Leishmania donovani
Leishmania major
Trypanosoma vivax
Trypanosoma congolense
Trypanosoma brucei brucei
Trypanosoma evansi
Trypanosoma brucei gambiense
Trypanosoma grayi
Trypanosoma theileri
Trypanosoma rangeli
Trypanosoma cruzi
monoxenous lineages
Leishmaniinae
Leishmania
Trypanosoma
Trypanosomatidae
0.2
B
H. muscarum
400
972
258
5635
548
L. major
1345
186
T. brucei brucei
C
H. muscarum
349
39
123
6754
126
69
L. pyrrhocoris
789
L. seymouri
D
H. muscarum
647
79
6528
11
26
L. donovani
708
L. major
25

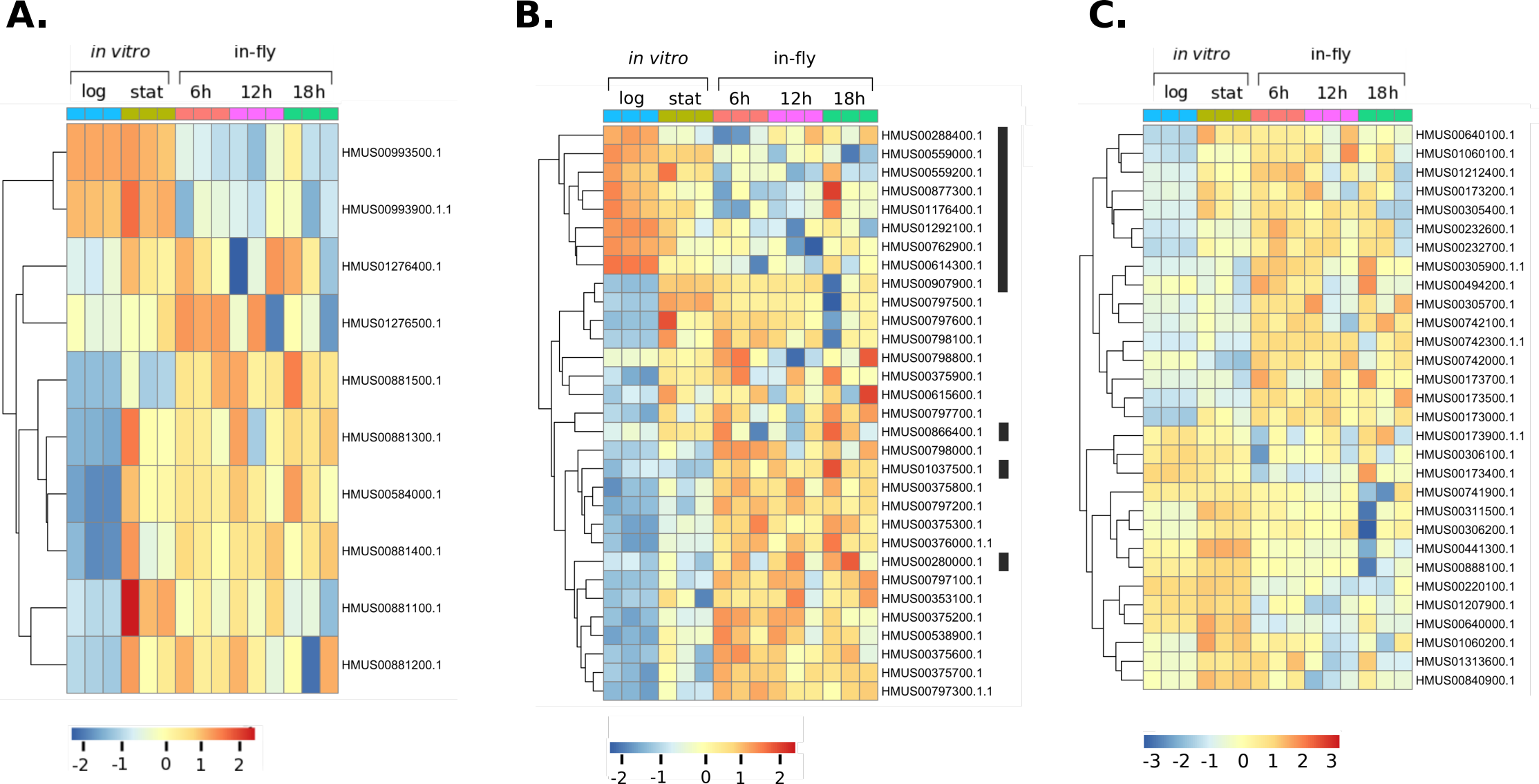

(A)
number of gene families
10000
5000
0
1x10^5
5x10^4
0
total number of genes
H. muscarum
Crithidia (3 spp.)
Other monoxenous (3 spp.)
Leishmania (4 spp.)
Trypanosomes (9 spp.)
Bodo
(B)
number of genes
1000
750
500
250
0
0
5
10
15
20
number of species

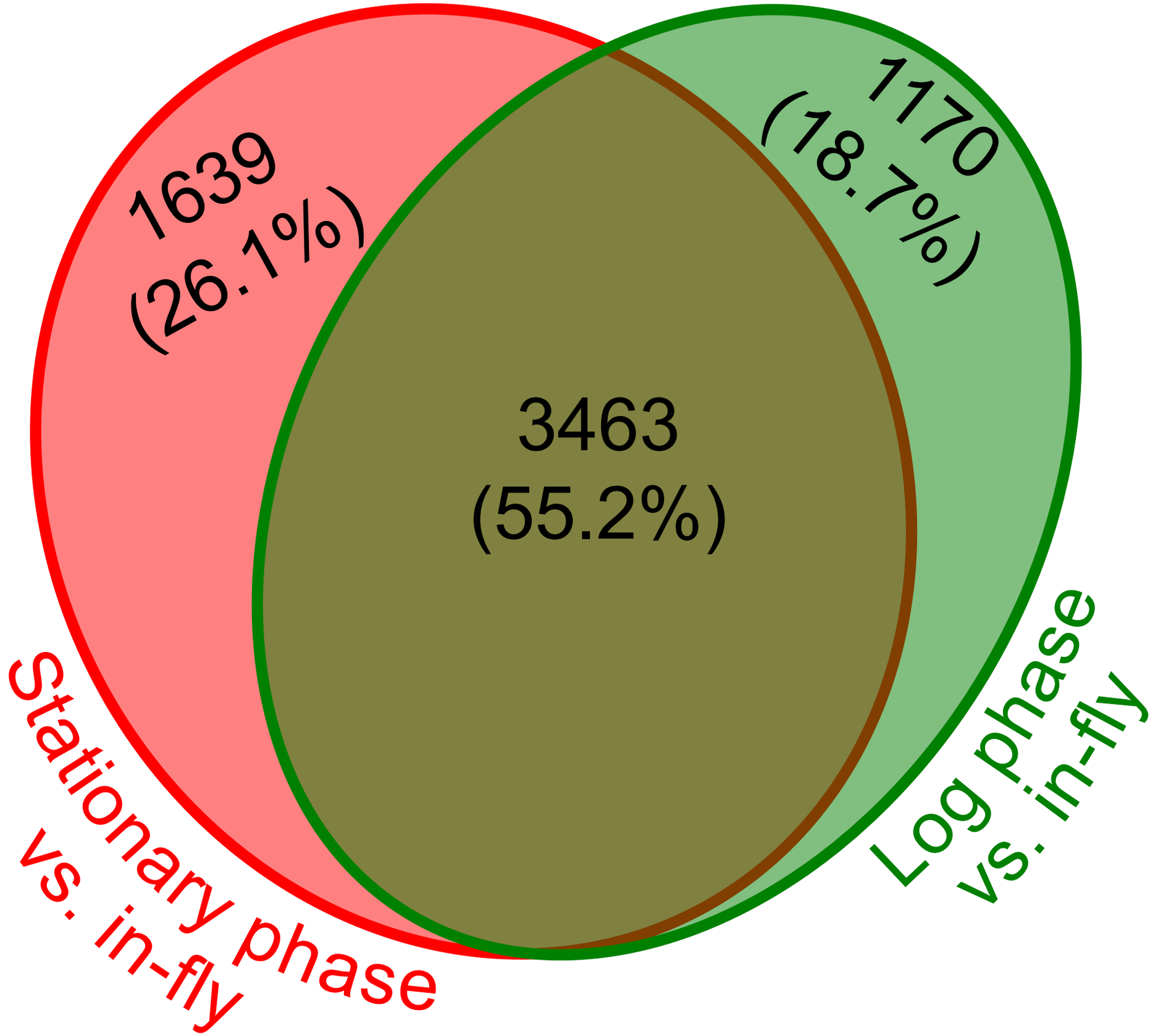
A.


H. muscarum mapped reads
10^8
10^7
10^6
10^5
10^4
10^3
6 hours
12 hours
18 hours
54 hours
7 days
Time since H. muscarum feeding

B.


PC2: 11% variance
0
-20
-40
-40
-20
0
20
PC1: 58% variance
In vitro
log phase
stationary phase
In fly
6 hours
12 hours
18 hours


A.
in vitro
in-fly
log
stat
6h
12h
18h
HMUS00993500.1
HMUS00993900.1.1
HMUS01276400.1
HMUS01276500.1
HMUS00881500.1
HMUS00881300.1
HMUS00584000.1
HMUS00881400.1
HMUS00881100.1
HMUS00881200.1
-2
-1
0
1
2
B.
in vitro
in-fly
log
stat
6h
12h
18h
HMUS00288400.1
HMUS00559000.1
HMUS00559200.1
HMUS00877300.1
HMUS01176400.1
HMUS01292100.1
HMUS00762900.1
HMUS00614300.1
HMUS00907900.1
HMUS00797500.1
HMUS00797600.1
HMUS00798100.1
HMUS00798800.1
HMUS00375900.1
HMUS00615600.1
HMUS00797700.1
HMUS00866400.1
HMUS00798000.1
HMUS01037500.1
HMUS00375800.1
HMUS00797200.1
HMUS00375300.1
HMUS00376000.1.1
HMUS00280000.1
HMUS00797100.1
HMUS00353100.1
HMUS00375200.1
HMUS00538900.1
HMUS00375600.1
HMUS00375700.1
HMUS00797300.1.1
-2
-1
0
1
2
C.
in vitro
in-fly
log
stat
6h
12h
18h
HMUS00640100.1
HMUS01060100.1
HMUS01212400.1
HMUS00173200.1
HMUS00305400.1
HMUS00232600.1
HMUS00232700.1
HMUS00305900.1.1
HMUS00494200.1
HMUS00305700.1
HMUS00742100.1
HMUS00742300.1.1
HMUS00742000.1
HMUS00173700.1
HMUS00173500.1
HMUS00173000.1
HMUS00173900.1.1
HMUS00306100.1
HMUS00173400.1
HMUS00741900.1
HMUS00311500.1
HMUS00306200.1
HMUS00441300.1
HMUS00888100.1
HMUS00220100.1
HMUS01207900.1
HMUS00640000.1
HMUS01060200.1
HMUS01313600.1
HMUS00840900.1
-3
-2
-1
0
1
2
3


1639
(26.1%)
1170
(18.7%)
3463
(55.2%)
Stationary phase
vs. in-fly
Log phase
vs. in-fly

Feature	*H. muscarum* v1.0
genes	12687
mRNAs	12162
CDSs	12175
polypeptides	12934
pseudogenes	772
rRNAs	168
snRNAs	3
snoRNAs	181
tRNAs	173

Table 1 - Herpetomonas muscarum genome annotation summary.

Species	Accession #
Herpetomonas muscarum	
Herpetomonas muscarum	EU095982.1*, EU095980.1*,
Herpetomonas sp. TCC263	EU095976.1
Herpetomonas sp. TCC263	EU095977.1
Herpetomonas roitmani	EU095978.1
Herpetomonas nabiculae	KF054153.1
Phytomonas EM1	X87138.1
Phytomonas serpens	L42381.1, L42378.1,
Phytomonas sp. Mar8	AF250993.1
Phytomonas sp. Alp1	AF250967.1
Leishmania braziliensis	MG010484.1
Leishmania tarentolae	AY100201.1
Leishmania hoogstraali	AY100197.1, AY100200.1
Leishmania gymnodactyli	AY100195.1, AY100196.1
Leishmania adleri	AY100199.1, AY100194.1
Leishmania major	XR_002460055.1
Leishmania mexicana	Agami and Shapira 1992
Leishmania donovani	CP022617.1
Leishmania infantum	AF097653.1
Blastocrithidia culicis	DQ860204.1
Blastocrithidia culicis	DQ860203.1

Splice leader sequence (bases 1-40)
AACTAACGCT**AAAAATTGTT**ACAGTTTCTGTACATTATTG
AACTAACGCT**AAAAATTGTT**ACAGTTTCTGTACTATATTG
AACTAAAGCA**TTATATAGAT**ACAGTTTCTGTACTATATTG
AACTAAAGCA**TTATATAGAT**ACAGTTTCTGTACTATATTG
AACTAAAGCA**TTATATAGAT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**AT-TATTGTT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**-ATTCTAGAT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**-ATTCTAGAT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**-ATTCTAGAT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**-ATTCTAGAT**ACAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT-**ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATAAGTA**TCAGTTTCTGTACTTTATTG
AACTAACGCT**-ATATTTGTT**ACAGTTTCTGTACTATATTG
AACTAACGCT**-ATATTTGTT**ACAGTTTCTGTACTTTATTG

Total number of genes	212,664
Number of genes in orthogroups	186,070
Number of unassigned genes	26,594
Percentage of genes in orthogroups	87.50%
Percentage of unassigned genes	12.50%
Number of orthogroups	12,701
Number of species-specific orthogroups	313
Number of genes in species-specific orthogroups	4,212
Percentage of genes in species-specific orthogroups	2.0%
Mean orthogroup size	14.7
Median orthogroup size	14
Number of orthogroups with all species present	9
Number of single-copy orthogroups	0

Table 3 – A summary of orthogroup assignments of representative trypanosomatid proteomes.

	Number of
Glycolysis	44/45
Gluconeogenesis	2/2
Pentose-phosopate pathways	12/13
NADPH metabolism	4/4
Acetate Metabolism	14/17
TCA Cycle	17/17
Mitochondrial carriers	24/25
Respiratory Chain	79/82
Amino acid transporters	31/31
Lipid metabolism	9/11
Leu-Isoleu-Val Degradation	22/23
Fatty Acid Biosynthesis	14/14
Sphingolipid biosynthesis	7/11
Glycerophospholipid biosynthesis	16/16
GPI-N-glycosylation biosynthesis	47/49
Quorum sensing	32/35
Bloodstream to procyclic form differentiation	10/12
Epimastigote meiosis	4/5
RNA regulators of the life cycle	18/18
Proteins with RNA-binding annotation	54/57
RNAi machinary	5/5
PROTEIN KINASES	147/169
PHOSPHATASES	86/93
Nuclear pores	27/27
Exosome	12/12
Spliceosome	56/59
Kinetochore	30/34
GP63	14/15
Mucins	8/11

T. brucei genes without orthologues in *L. muscorum*
METABOLISM
fructose- 2,6- bisphosphatase-like protein (Tb927.10.4520)
n/a
sedoheptulose-1,7-bisphosphatase (Tb927.2.5800)
n/a
carnitine O-acetyltransferase (CAT) (Tb927.11.2230)
n/a
mitochondrial carrier protein 1 (Tb927.9.12140)
NADH-ubiquinone oxidoreductase mitochondrial, (Tb927.7.6350),
n/a
NAD/NADP dependent oxidoreductase (Tb927.10.11930),
hypothetical protein (Tb927.4.2700)
n/a
inositol phosphorylceramide synthase (Tb927.9.9410),
n/a
hypothetical protein (Tb927.4.4200),
DIFFERENTIATION AND RNA
protein phosphatase 1 adjacent gene (Tb927.4.3650),
EP1 (Tb927.10.10260) or PSSA2 (Tb927.10.11220)
BARP protein (Tb927.9.15510)
n/a
Tb927.10.14950, Tb927.6.2550, Tb927.9.6870
n/a
Tb11.v5.0564, Tb11.v5.0644, Tb927.1.3130, Tb927.10.12480,
Tb927.07.v5.1, Tb07.30D13.60, Tb927.10.4930, Tb927.11.11740,
NUCLEAR PROTEOME
n/a
n/a
SF3b(SAP)14b (Tb927.10.7390), SR protein (Tb927.9.6870), U1-
KKIP3 (Tb927.10.6330), KKT7 (Tb927.11.1030), KKIP1
OTHER PROTEINS OF INTEREST
Tb927.11.7610
mucin-2 precursor (Tb927.8.7190), mucin (TcMUCII,

T.brucei genes found with	Comments
Tb927.4.1360, Tb927.10.5620,	The three phosphoglycerate kinase genes: PGKA
n/a	
Tb927.11.8970	The top blastp results for Tb927.11.8970 was
n/a	Two malic enzymes, MEc (Tb927.11.5440) and MEm
Tb927.11.2690, Tb927.8.2520	
Tb927.11.5050, Tb927.10.2560,	Two isocitrate dehydrogenase genes, the mitochondrial
Tb927.9.10310, Tb927.2.2970	Three mitochondrial carrier proteins, MCP5a, b and c, were
Tb927.8.5120, Tb927.5.1710,	The three *T. brucei* ATPase subunit 9 genes (Tb927.10.1570,
Tb927.7.2440, Tb927.1.3950,	Two copies of AAT-17 (Tb927.11.15950 and
Tb927.11.10060, Tb927.3.4850	
Tb927.3.4850, Tb927.10.13560,	
Tb927.8.2520, Tb927.7.4160	*T . brucei*'s fatty acid CoA synthetase genes ASC1-5 were
Tb927.8.7730, Tb927.4.4740	*T. brucei*'s dihydroceramide synthase (Tb927.8.7730) and
Tb927.11.15150, Tb927.4.3160,	
Tb927.3.4020, Tb927.9.12700,	
n/a	The three *T. brucei* protein phosphatase 1 genes
Tb927.10.6690, Tb927.6.3490,	*T. brucei MSP-B (Tb927.8.1610) has has 28 H. muscarum*
n/a	This list of proteins was taken from Kolev et al; Cell
n/a	
Tb927.9.11030, Tb927.11.7010,	*T. brucei* protein kinases Tb11.v5.0534 and Tb927.1.1530
Tb927.4.4510, Tb927.8.1710,	Two *T. brucei* acidocalcisomal pyrophosphatase genes
Tb927.11.11080, Tb927.11.11080,	This list of proteins was taken from Goos et al. 2017: PLoS
Tb927.11.16600	This list of proteins was taken from Goos et al. 2017: PLoS
Tb927.11.11150	This list of proteins was taken from Goos et al. 2017: PLoS
Tb927.11.12410, Tb927.11.12420,	This list of proteins was taken from Goos et al. 2017: PLoS
n/a	A single orthologue was found for Tb927.11.7410. However,
n/a	

Gene Name	***H. muscarum* orthologue ID**	**log2FoldChange**	**adjusted p-value**
CRK4	HMUS00195900.1	1.3	8.89E-10
cyclin 11	HMUS01322900.1	1.2	4.32E-05
cyclin 2	HMUS00751100.1	1.2	2.72E-19
cyclin 4	HMUS00787500.1	1.1	1.53E-17
cyclin 7	HMUS00475100.1	0.8	2.41E-14
CRK10	HMUS01143000.1	0.7	6.49E-09
cyclin 5	HMUS00580100.1	0.7	2.02E-12
cyclin 10	HMUS01323000.1	0.5	0.001
CRK12	HMUS00986000.1	0.3	0.015
DNA-directed RNA polymerase III subunit, putative	HMUS00638800.1	-0.3	0.032
mitochondrial DNA polymerase I protein C	HMUS00828800.1	-0.5	0.006
mitochondrial DNA polymerase I protein D	HMUS00617400.1	-0.5	0.018
mitochondrial DNA polymerase I protein B,	HMUS01100200.1	-0.6	0.007
DNA polymerase alpha/epsilon subunit B	HMUS00740000.1	-0.7	0.004
DNA polymerase delta catalytic subunit	HMUS00566500.1	-0.7	0.006
CRK3	HMUS00914500.1	-1.0	1.06E-40
cyclin 8	HMUS00524500.1	-1.0	1.40E-39

Table 5 – Significantly differentially regulated cyclins and cyclin-related kinases between stationary, and log phase *H. muscarum*.

Gene Name	***H. muscarum* orthologue ID**	**log2foldchange**	**adjusted p-value**
cyclin 11	HMUS01322900.1	-3.31	6.54E-27
cyclin 4	HMUS00787500.1	-1.15	3.62E-13
CRK4	HMUS00195900.1	-0.95	3.46E-03
CRK1	HMUS01116400.1	-0.84	9.95E-08
CRK8	HMUS00385600.1	-0.49	2.32E-02
cyclin 7	HMUS00475100.1	-0.44	2.21E-02
cyclin 8	HMUS00524500.1	0.36	1.94E-02
mitochondrial DNA polymerase I protein D	HMUS00617400.1	0.57	9.16E-03
cyclin 6	HMUS00719100.1	0.74	2.14E-02
cyclin 5	HMUS00580100.1	0.85	9.34E-05
CRK9	HMUS01274200.1	0.87	1.45E-03
DNA polymerase theta catalytic subunit	HMUS00097200.1	1.15	1.51E-07
mitochondrial DNA polymerase I protein C	HMUS00828800.1	1.25	7.27E-09
DNA polymerase kappa	HMUS01207400.1	1.36	4.93E-03
CRK11	HMUS00452900.1	1.46	8.20E-04
CRK12	HMUS00986000.1	2.07	6.68E-18

Table 6 – Cell cycle-associated proteins differentially expressed in *H. muscarum* upon ingestion by *D. melanogaster*. Fold changes shown are at 6 hours post ingestion compared to log phase axenic culture.