**UDP-glycosyltransferase genes in trypanosomatid genomes have diversified independently to meet the distinct developmental needs of parasite lineages.**

Sara Silva Pereira, Andrew P. Jackson

Department of Infection Biology

Institute of Infection and Global Health

University of Liverpool

Liverpool Science Park Ic2

146 Brownlow Hill

Liverpool, L3 5RF

United Kingdom

**Abstract**

**Background**: Trypanosomatid parasites such as *Trypanosoma* spp. and *Leishmania* spp. are a major source of infectious disease in humans and domestic animals worldwide. Fundamental to the host-parasite interactions of these potent pathogens are their cell surfaces, which are highly decorated with glycosylated proteins and other macromolecules. Trypanosomatid genomes contain large multi-copy gene families encoding UDP-glycosyltransferases (UGTs), the primary role of which is cell-surface protein decoration, necessary for antigenic variation and cellular adhesion. Here we report a phylogenetic analysis of UGTs from diverse trypanosomatid genomes, the aim of which was to understand the origin and evolution of their diversity.

**Results**: By combining phylogenetics with analyses of recombination, and selection, we compared UGT repertoire, genomic context and sequence evolution across 19 trypanosomatids. We identified a UGT lineage present in Stercorarian trypanosomes and a free-living Kinetoplastid *Bodo saltans* that likely represents the ancestral state of this gene family. The phylogeny of parasite-specific genes shows that UGTs repertoire in *Leishmaniinae* and Salivarian trypanosomes has expanded independently and with distinct evolutionary dynamics. In the former, the ancestral UGT repertoire was organised in a tandem array from which sporadic transpositions to telomeric regions occurred, allowing expansion most likely through telomeric exchange. In the latter, the ancestral UGT repertoire was comprised of seven subtelomeric lineages, two of which have greatly expanded potentially by gene transposition between these dynamic regions of the genome.

**Conclusions**: The phylogeny of UGTs confirms that they represent a substantial parasite-specific innovation, which has diversified independently in the distinct trypanosomatid lineages. Nonetheless, developmental regulation has been a strong driver of UGTs diversification in both African trypanosomes and *Leishmania*.

**Keywords**: UDP-glycosyltransferases, trypanosomatids, glycosylation, sugar transferases,

**Background**

Trypanosomatid parasites are the causes of several neglected tropical diseases worldwide that put 500 million people and over 60 million cattle at risk of infection [1]. They include Stercorarian trypanosomes such as *Trypanosoma cruzi*, the cause of Chagas disease in central and south America; Salivarian trypanosomes such as *Trypanosoma brucei*, the cause of African trypanosomiasis in Humans and animals, (as well as *T. vivax* and *T. congolense* that cause disease exclusively in animals); and they include *Leishmania* spp., which cause various kinds of leishmaniasis. Collectively, these vector-borne diseases have a significant impact on human and animal health, and are a profound constraint on the socio-economic development of low and middle-income countries.

The life cycles of trypanosomatids may be monoxenic or dixenic. All human and animal parasites are dixenic, cycling between a vertebrate host and an invertebrate vector. African trypanosomes alternate between a procyclic stage in the tsetse fly (*Glossina* spp.) and an extracellular bloodstream-form in a mammalian host. *T. cruzi* infects a wide range of mammals and is transmitted by the bite of triatomine bugs. *Leishmania* spp. alternate between a motile, promastigote form in a sand-fly vector, and an intracellular amastigote form in their mammalian host. Besides these, and many other dixenic parasites, there are multiple genera of monoxenic trypanosomatids that parasitize insects and are transmitted through the faecal-oral route, such as *Crithidia*, *Leptomonas* and *Lotmaria* [2–4]. Regardless of whether they have one or multiple hosts, all trypanosomatids have a complex development and are able to adopt multiple cell morphologies depending on the precise host environment they inhabit [5,6]. Associated with these different cell morphologies, are characteristic cell-surface architectures that are typically parasite-specific and substituted during transmission between hosts [7–9].

The assembly of these enigmatic cell surface coats in trypanosomatids is dependent on the expression of proteins and other macromolecules on the parasite surface via a glycophosphatidylinositol (GPI) anchor. UDP-glycosyltransferases (UGTs) catalyse the transfer of the sugar molecule N-Acetylglucosamine (GlcNAc) from UDP-GlcNAc to phosphatidylinositol [10,11]. They catalyse the first steps in the synthesis of GPI anchors, but also play a crucial role in the synthesis of glycans of various functions, contributing to the extraordinary collection of glycoconjugates that decorate the surface of trypanosomatids [12].

UGTs are part of a superfamily of glycosyltransferases (GT) present in all organisms, which typically play a role in detoxification and homeostatic processes [13]. Three types of GTs have been characterised (A-C): GT-A share a catalytic domain, the DXD motif, whose carboxylated side chains coordinate enzymatic activity; GT-B are very diverse; and GT-C have only recently been described from iterative sequence searches with a single 3-D structure not supporting the presence of a common active site [14]. Trypanosomatid UGTs belong to the inverting GT-A family 31 (GT31 in CAZY nomenclature), a family present in eukaryotes and prokaryotes. In plants, GT31 includes enzymes involved in proteoglycan and N-glycan synthesis [15]; in mammals it includes chondroitin synthases, responsible for the synthesis of glycosaminoglycan  chains that regulate homeostatic processes, such as cell proliferation and extracellular matrix deposition [16], and fringe proteins which modulate the notch signalling pathway [17]. In bacteria, GT31 enzymes also play an important role in epitope synthesis, such as the catalysis of the final steps in formation of the O antigen repeating unit in pathogenic *E. coli*, through the glycosylation of the nonreducing end of oligosaccharides [18]. In trypanosomatids, this family has expanded greatly compared to other eukaryotes and its function closely relates to surface protein decoration [12].

In *Leishmania*, UGTs synthesise the galactose-rich glycoconjugates that compose the parasite dense glycocalyx, using UDP-galactose as the glycosyl donor [19]. Simultaneously, a subset of UGTs belonging to the side chain galactose-related gene families (SCG, SCGL, SCGR) catalyse the attachment of Gal(ß1,3) side chains to the phosphoglycan (PG) polymer repeating units of the lipophosphoglycan (LPG) coat. These are required for parasite survival in the sandfly midgut, where parasite differentiation to the replicating procyclic promastigote stage occurs [20]. Whilst most microbial adhesins are proteins that interact with various molecules in host epithelial receptors, *Leishmania* stage-specific adhesion potential is provided by LPG, a glycoconjugate interacting with lectin receptors in the epithelium of the sandfly midgut [21]. The galactose side chains permit binding and adhesion to lectins in the midgut epithelium during the digestion process, so the parasite can avoid excretion with the peritrophic matrix [22].

In African trypanosomes, UGTs are involved in the synthesis of complex poly-N-acetyllactosamie-containing type N-linked and GPI-linked glycans. N-linked glycans can have various functions: on VSG, they are predicted to assist the protection of invariant surface antigens by filling the spaces between VSGs [23]; on the transferrin receptor, they ensure enough space is left at the flagellar pocket to allow efficient binding of the receptor to transferrin [24]; and on the lysosome-associated membrane protein p67, N-linked glycans might function as internalisation signals for endocytosis [25]. GPI-linked glycans play a role in tsetse fly colonisation as procyclin and, in the mammal, as VSG GPI-anchor side chains [26,27]. Since UDP-Gal-dependent glycosylation pathways are essential for the survival of *T. brucei* in both insect and mammal forms [28,29], UGTs make logical targets to understand parasite-host interactions.

The publication of genomes for most trypanosomatid species [2–4,30–37] together with transcriptomic and proteomic studies [38–42] demonstrated that trypanosomatids possess large repertoires of UGT isoforms encoded by multi-gene families often found in irregular tandem gene arrays. The recent publication of a genome sequence for the free-living kinetoplastid *Bodo saltans* [43] provides an out-group for a comparative analysis of trypanosomatid UGT genes, able to answer fundamental questions about their diversity.

Three main reasons make UGTs sensible study targets: i) Despite being a multi-copy gene family with distinct repertoires across species and important roles in pathogenesis, their diversity across the genus is poorly understood; ii) The understanding of its diversity may elucidate phenotypic differences in disease progression; and iii) Through genomic comparison we can identify shared and species-specific loci, as well as stage-specific isoforms, to expedite the search for suitable drug and transmission targets.

Here we describe the phylogeny and comparative genomics of UGT genes in trypanosomatids and *Bodo saltans* with particular emphasis on African trypanosomes and *Leishmania*. We aim to identify monophyletic free-living (*B. saltans)* and parasitic (trypanosomatid) UGTs to understand more about their ancestral form and the origin of family expansion; investigate orthology across parasites to know whether UGT expansion was independent in distinct parasites; understand the role of recombination among paralogs and that of selection in gene divergence during expansion; and finally interpret those results in the context of available gene expression and functional studies to search for evidence of functional differentiation, since non-redundant paralogs under strong negative selection could offer targets for functional studies and interventions.

**Methods**

**Data Collection and Nomenclature**

Annotated UGT sequences were obtained from genome sequences of *Trypanosoma cruzi* CL Brenner Esmeraldo-like, *T. rangeli* SC58, *T. grayi* ANR4, *T. brucei* TREU927, *T. congolense* IL3000, *T. vivax* Y486, *Leishmania major* Friedlin, *L. infantum* JPCM5, *L. mexicana* MHOM/GT/2001/U1103, *L. tarantolae* Parrot-Tarll, *L. enriettii* LEM3045, *L. braziliensis* MHOM/BR/75/M2904, *Leptomonas pyrrhocoris* H10, and *Crithidia fasciculata* Cf-Cl hosted by TritrypDB v.28 (<http://tritrypdb.org>) [44]; *Bodo saltans* hosted by the GeneDB website (<http://genedb.org>) [45]*;* and *Angomonas deanei and Strigomonas culicis* hosted by Ensembl Protists v.31 (<http://protists.ensembl.org>). Additionally, a sequence similarity search with tBLASTn using *T. brucei*, *L. major* and *B. saltans* UGTs as query was performed to identify relevant genes annotated as hypothetical.

To expand the sample repertoire of monoxenic species, the genome sequences from *Crithidia acanthocephali* and *Lotmaria passim* unannotated genomes were retrieved from NCBI (<http://www.ncbi.nlm.nih.gov/genome>). These were inspected for UGTs by sequence similarity search with tBLASTn using its closest relative, *C. fasciculata*, UGTs as the query. Identified putative UGTs were named *L.passim*1-4 and C.*acanthocephali*1-4.

The presence of the conserved UDP catalytic domain previously described (DXD) in the sequences was a requirement for the inclusion in this study [46].

**Multiple Sequence Alignment**

Translated nucleotide sequences were aligned with ClustalW (Larkin et al. 2007) using BioEdit 7.2.5 (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>), producing a nucleotide alignment of 1005 nucleotides and a protein alignment of 335 amino acids around the catalytic domain, after trimming non-conserved regions. This corresponded to 23-93% of the full glycosyltransferase proteins sequence, due to *Leishmania* spp. having a large specific insertion. When analysed separately, the African trypanosome alignment was 305 amino acids long, while the *Leishmaniinae* alignment was 824 amino acids long.

**Phylogenetic Estimation**

The UGT phylogeny was estimated from protein sequence alignments with maximum likelihood (ML) under a WAG+ substitution model [47] using PHYML v3.0 [48] and Minimum Evolution (ME) under a Tamura-Nei substitution model using MEGA7 [49]. Robustness was assessed with 500 bootstrap replicates. We also attempted to estimate a phylogeny using Bayesian inference (BI) but the analysis failed to converge on stable parameter values and therefore was not pursued.

The UGT phylogenies of African trypanosomes and *Leishmaniinae* were estimated from both nucleotide and protein sequence alignments with three methods: ML using PHYML v3.0 [48], BI using MrBayes v3.1.2. [50,51], and ME using MEGA7 [49].

Optimal substitution models for ML trees were found with the Smart Model Selection option in PHYML, using the Akaike Information Criterion (AICc). Protein trees were estimated with WAG+ (African trypanosomes) or LG+G model (*Leishmaniinae*). Nucleotide trees were estimated with the GTR+model with 500 bootstrap replicates.

The BI trees were estimated with gamma rates function in MrBayes and four Markov chain Monte Carlo chains run in parallel over 2,500,000 generations, with a burnin of 5000. The nucleotide BI trees were estimated with default parameters whereas the protein BI trees were estimated with a fixed WAG+G model. Posterior probabilities of each node were used to assess accuracy of BI trees.

Nucleotide ME trees were estimated using logdet genetic distances to correct base composition bias [52]. Protein ME trees were estimated with JTT substitution model using MEGA7 [49] and robustness was assessed with 500 bootstrap replicates.

**Tests for recombination**

Evidence for recombination was investigated in *L. major*, *L. infantum*, *L. mexicana*, *T. brucei*, and *T. congolense*. For *Leishmania*, SCG and SCGR subfamilies were separately analysed (Figure 2). For African trypanosomes, each lineage was separately analysed. Four sequences were randomly selected for each species and subject to different tests. In *Leishmania*, a negative control comprised of four sequences known not to recombine (one SCG, one SCGR, one SCGL, and one SCGR gene phylogenetically closer to SCGL) were included. In African trypanosomes, the negative control was comprised of all genes from lineages 2-5 (Figure 3).

Recombination probability was detected with the pair-wise homoplasy index (PHI) [53] as part of the SplitsTree package [54]. Breakpoints were predicted with the Genetic Algorithm for Recombination Detection (GARD) [55], run using the REV model, under the AICc information criterion. The KH test was applied to test for rate heterogenicity to prevent false positives arising from significant topological incongruences rather than recombination. These tests informed on the likelihood of recombination affecting sequence evolution. The breakpoint(s) identified with GARD were used to split the sequences into non-recombinant parts before subsequent analyses of selection to prevent false positives due to recombination.

**Positive selection tests**

To evaluate whether positive selection was affecting sequence evolution, full sequences where recombination was unlikely and non-recombinant partial sequences were subject to six site-level selection tests: Single Likelihood Ancestor Counting (SLAC) to perform ancestral reconstruction; Fixed Effects Likelihood (FEL) to directly estimate dN/dS ratios [56]; Random Effects Likelihood (REL) to infer selection pressures using an empirical Bayes approach and model dN/dS ratios at individual sites based on a pre-defined distribution; Partitioning Approach for Robust Inference of Selection (PARRIS) [57] to test for alignment-wide evidence of selection taking into account recombination and synonymous rate variation; Fast Unbiased Bayesian Approximation (FUBAR) to estimate the dN/dS ratio based on Bayesian Inference using a MCMC routine [58]; and the standalone package Phylogenetic Analysis Using Maximum Likelihood (PAMLx) to construct likelihood ratio tests [59].

Significance thresholds for recombination were *p*-value<0.05 and posterior probability >0.9. For sites to be considered under positive selection, support by 4 out of 5 tests was required. Unless specified, all programs were hosted at the DataMonkey server (<http://datamonkey.org>).

**Results**

We estimated a UGT maximum likelihood phylogeny (Figure 1) from a 335 residue multiple alignment of 236 protein sequences. Among these sequences were 2 UGT sequences from the non-parasitic out-group *B. saltans*, 2 from *Angomonas deanei*, 1 from *Strigomonas culicis*, 82 from Stercorarian trypanosomes (3 species), 71 from African trypanosomes (3 species), and 70 from *Leishmaniinae* (10 species). It is immediately clear from this phylogeny that trypanosomatids have greatly expanded their UGT repertoires relative to the free-living *B. saltans* (e.g. ratios of 1:13 for *T. brucei*; 2:13 for *L. major*; 1:23 for *T. cruzi*). Furthermore, UGT repertoires have been elaborated to different extents in trypanosomatid genera, e.g. *T. brucei*: *L. major* ratio of 2:1).

The tree topology broadly reflects the major trypanosomatid lineages and contains four main features, (numbered 1-4 in Figure 1), that will be examined further: a clade comprising *B. saltans* sequences and rare orthologs from Stercorarian species (‘the ancestral lineage’) (1); a clade of *Leishmaniinae* sequences (2); two clades of African trypanosomes (3). Most Stercorarian trypanosome sequences clustered together (although some *T. grayi* sequences were ambiguous) but without adequate node support (4). The lack of species diversity hampers orthology analysis and thus we have not examined Stercorarian sequences further in this study.

**An ancestral UGT lineage shared by Stercorarian trypanosomes and *B. saltans***

The ancestral lineage is composed of four genes retaining orthology: two *B. saltans* (BSAL\_27930 and BSAL\_69925), one *T. cruzi* (TcCLB.503487.50), and one *T. rangeli* (TRSC58\_00816), all close in length (352 to 495 amino acids). The *B. saltans* sequences share 31% overall identity between each other and 34-36% identity with *T. cruzi*, *T. rangeli* and *T. grayi* (Tgr.1587.1000). The latter was not included in the phylogeny due to its short length (90 amino acids). The absence of this lineage of UGTs in *Leishmaniinae* and African trypanosomes suggests post-speciation gene loss. Transcriptomic data from genomic microarrays show TcCLB.503487.50 is constitutively expressed, being the most abundant in amastigotes and the least in epimastigotes [60]. The genomic locus of these genes could not be investigated due to the current quality of the assemblies of *T. rangeli* and *T. grayi* genomes.

A search for similar sequences in *Euglena gracilis* transcriptome and *Trypanoplasma borreli* and *Naegleria* *fowleri* genomes did not produce any relevant matches.

***Leishmania* UGT repertoire derives from ancestral tandem array**

The UGT phylogeny raised specific questions about the different gene lineages in the *Leishmaniinae* subfamily so we investigated it further by building a *Leishmaniinae*-only phylogeny based on a longer multiple sequence alignment (Figure 2), comparing genomic loci and looking at available gene expression data.

All *Leishmaniinae* species have multiple UGT genes organised in a tandem array in chromosome 2 (*L. major*) in a conserved genomic locus flanked by a putative phosphatidylinositol kinase related protein (LmjF.02.0120) and small GTP binding protein rab6-like protein (LmjF.02.0260) (Figure 3). Genes belonging to this array have been identified as side chain galactose receptors (SCGR) [46,61]. SCGR genes cluster by position between closely related species (e.g. LmjF.02.0200 and its neighbors), but by species between distant relatives (e.g. Lbr.02.0250 and its relatives), with instances of extensive gene duplication in one lineage of the monoxenic species (e.g. CFAC1\_160011900) and *L. braziliensis* (e.g. Lbr.02.0250). This suggests slow but on-going concerted evolution arising from rapid gene duplication and resulting in the loss of orthology over time.

One possible exception to this prevailing pattern is Lbr.02.0260 and its orthologs. Although orthology between *L. major* and *L. braziliensis* is mostly absent within the array, there is an example of a lineage present in all *Leishmania* species (LmjF.02.0230), whose sequences show a unique change in the catalytic domain from DDD to YDD, hinting functional differentiation.

Orthology is also conserved in the *Leishmania*-specific single-copy lineage located in chromosome 14, which has been previously identified in *L. major* as side-chain galactose ligand (SCGL) [61]. The phylogeny suggests it derives from a single transposition event from the array to chromosome 14 in the *Leishmania* ancestor. The absence of a gene at this particular locus in *L. braziliensis* and *L. mexicana* indicates loss in these species (Supplementary Figure 1).

Unlike the two previous lineages, the last lineage of UGTs in the *Leishmaniinae* sub-family, which comprises the *Leishmania*-specific side-chain galactose (SCG) genes [61], has a dynamic of concerted evolution. These locate at the telomeres of multiple chromosomes, but although the genomic loci are structurally conserved, these genes do not retain orthology between the different species. Supplementary Figure 1a shows an example of this at the distal telomere of chromosome 25. This scenario suggests that this gene lineage transposed to telomeres in the *Leishmania* ancestor and has since expanded to other chromosomes perhaps by telomeric exchange, providing strong evidence for concerted evolution.

We have examined existing evidence for protein expression of SCG genes for *L. infantum* [62] and *L. major* [38] (Figure 2). Available data for *L. infantum* [62] reveal three of four SCG genes being differentially expressed in the amastigote stage, as opposed to all SCGR genes being constitutively expressed. The SCGL gene LinJ.14.1500 was not detected in the study. Proteomic analysis in *L. major* showed differential expression at the amastigote stage of LmjF.02.0230 only, but all seven SCG genes and LmjF.02.0190 seem to be more abundant in the amastigote stage (Figure 3). The remaining SCGR and the SCGL genes do not show developmental regulation [38].

In summary, SCG genes seem generally more abundant in the amastigote stage of *Leishmania* species; SCGR generally constitutively expressed; and SCGL present in very low abundances. This suggests that developmental regulation accounts for some degree of gene differentiation.

Prior to selection testing, evidence for recombination was investigated. Both recombination tests suggest *L. major* SCG genes to be under recombination, with GARD identifying one significant breakpoint at nucleotide 489. Trees inferred from GARD were fed into six tests for selection. Only PAMLx and FUBAR found evidence for positive selection, but not significant compared to the negative control. Selection tests for sequences where GARD did not predict significant breakpoints were not consistent, but no sites under positive selection were identified in any of the sequence collections by more than 3 out of 6 tests (Supplementary Table 1). Hence, there is little evidence for *Leishmania* UGTs to be under positive selection.

**Seven lineages underline the UGT repertoire in African trypanosomes**

To further understand the different lineages of UGTs in African trypanosomes, we estimated a phylogeny of these species, with *B. saltans* as the out-group (Figure 4). Furthermore, we investigated the genomic loci of the distinct lineages and interpreted them in the context of gene expression.

The phylogeny of UGTs in African trypanosomes shows seven lineages present in the common ancestor (numbered 1-7 in Figure 4) that retain orthology or co-orthology between species. Lineages 2-6 remain mostly single-copy orthologs. Evidence for conservation of genomic synteny is sporadic due to the quality of current genome assemblies of *T. congolense* and *T. vivax*. For example, in *T. brucei* and *T. congolense*, lineage 4 locus is conserved, being flanked by a leucine-rich repeat protein (Tb927.7.290) and a thioestherase-like superfamily protein (Tb927.7.330), but the *T. vivax* contig containing the former does not span the UGT gene. Similarly, lineage 6 locus seems conserved in all three species, being delimited by a methyltransferase domain containing protein (Tb927.10.12270) and a helicase-like protein (Tb927.10.310), although sequence gaps in *T. congolense* assembly preclude a final decision.

The pattern of orthologs among the seven lineages is disrupted on occasions. Lineage 2 was lost from *T. brucei* and *T. congolense*, while lineage 7 has been lost in *T. vivax*, but vastly expanded in the remaining species. Within *T. congolense* and *T. brucei*, concerted evolution of paralogs occurs, with genes arranged by species in lineages 1 and 7 and conservation of subtelomeric locations, suggesting expansion is arising from transposition of UGTs between these dynamic regions of the genome.

The analysis of the available expression data at the proteomic level reveals some developmental regulation of *T. brucei* genes, with lineages 3 (Tb927.3.5660) and 4 (Tb927.7.300) being differentially expressed in the bloodstream form, and lineages 5 (Tb927.10.12290) and 7 (Tb927.2.3370 and Tb927.4.5240 to Tb927.4.5290) being preferentially expressed in the procyclic form of the life cycle [39,40].

At the transcriptomic level, the higher abundance in bloodstream forms of Tb927.3.5660 and Tb927.7.300 is already significant, but not of Tb927.10.12290 in procyclics. Transcriptomic data also shows Tb927.5.2760 as differentially regulated in bloodstream forms. Tb927.2.3370 and Tb927.4.5240 seem to be constitutively transcribed, whilst Tb927.9.800 and Tb927.4.5790 are preferentially transcribed in BSF [41].

Available ribosomal profiling studies agree with proteomic data results and suggest higher abundance at the bloodstream form stage of Tb927.8.8090, Tb927.8.8100, Tb927.4.4290, Tb927.4.4250 and Tb927.4.4270 (Jensen et al., 2014) (Figure 4).

Expression data for *T. congolense* is not available, but the *T. vivax* expression study revealed higher protein abundance of TvY486\_0403910, TvY486\_0200980 and TvY486\_0305070 (corresponding to lineages 2 and 3) in bloodstream forms compared to metacyclics (maximum fold change of 1.42), as well as of TvY486\_0403900 (lineage 4) when compared to epimastigotes (maximum fold change of 11.02) [42]. Transcriptomic data suggest differential transcription of TvY486\_0403900 between bloodstream forms, epimastigotes and metacyclics (fold change of 2.66 and 3.54, respectively) [42].

In summary, the UGTs repertoire of African trypanosomes seems to be under strong developmental regulation, corroborating the hypothesis of functional differentiation within the family.

To test the contribution of selection to UGT expansion in African trypanosomes, we first searched for evidence of recombination and subsequently performed six tests of site-level selection. Three tests found evidence for recombination among *T. congolense* genes with three significant breakpoints identified by GARD taking into account rate variation. The six tests for positive selection performed did not show evidence for positive selected sites; only PAML identified one site under positive selection at nucleotide 257 of the alignment.

Selection tests for sequences where GARD did not predict significant breakpoints did not find any evidence for positive selection at the site level, but rather negative selection in lineages 2-6, suggesting UGT family expansion is not driven by positive selection or gene conversion.

**Discussion**

All trypanosomatids sampled, with the exception of *Angomonas deanei* and *Strigomonas culicis,* have a broad repertoire of UGT, which suggests these enzymes have important roles for parasite survival. We have identified the ancestral lineage of UGT still present in *B. saltans* and Stercorarian trypanosomes from which lineage expansion has occurred in different ways in trypanosomes and leishmanias. The trypanosomatid UGT phylogeny lacks support in the Stercorarian trypanosomes and *T. grayi* nodes, which could potentially be improved through the introduction of sequences from related trypanosomes, such as *T. theileri* or *T. avium.* These would strengthen robustness of *T. grayi* nodes and help deciphering the relative phylogenetic distance between *T. grayi* UGTs and the remaining trypanosomes.

**The ancestral lineage**

The UGT ancestral lineage retained in *B. saltans* and Stercorarian trypanosomes indicates that the UGT repertoire of the ancestral organism was considerably smaller, with fewer loci, supporting the theory that UGT expansion in trypanosomatids is a parasitic-specific innovation. The fact that UGT expansion is occurring under different dynamics in *Leishmania* spp. and trypanosomes strengthens the argument of UGTs having evolved to perform specific, essential roles in the life cycles of these parasites, making them pertinent targets for functional characterisation. The identification of the ancestor lineage of trypanosomatid UGTs, which we term the ‘protolog’, is an important asset for further comparative studies of parasite-specific sequences. For instance, if the protolog is shown to have a distinct function, it could offer an insight on the role of parasite-specific UGTs in the origin of parasitism. In *T. cruzi*, the gene belonging to the ancestral lineage is constitutively transcribed, but more abundant in amastigotes, the intracellular stage in the mammal host. This contrasts with the transcriptomic data available for other UGT genes (TcCLB.511339.30; TcCLB.508673.20; TcCLB.511395.120; TcCLB.508605.20; TcCLB.510553.50; TcCLB.510071.30; TcCLB.504557.20; TcCLB.508975.30), since mostly seem to be more abundant in trypomastigotes, the bloodstream stage of the parasite [60].

***UGTs subfamilies are conserved across Leishmania***

UGTs in *Leishmania* are divided into SCG, SCGR and SCGL families, which is consistent with the phylogeny in Figure 2. These families have been previously described in *L. major* [46,61], but little is known for the remaining species.

SCGR genes are arranged in a tandem array with members of the arabinosyltransferase family. This array is conserved across the *Leishmaniinae* subfamily with striking amino acid conservation, particularly in the surroundings of the “DXD motif” catalytic domain. This domain, composed of three aspartic acid residues, is conserved across most GT-A proteins, but is slightly modified in SCGR1 and 4 (LmjF.02.0230 and Lmj.02.0190) in all *Leishmania* species (i.e. DDD to YDD). In the phylogeny, these genes cluster closer to SCGL genes than to the other array members. When these genes were described in *L. major*, expression analysis by Western Blot suggested higher abundance in metacyclics [61], while proteomic studies revealed LmjF.02.0230 to be differentially expressed in amastigotes and LmjF.02.0190 to be constitutively expressed with higher peptide abundance in amastigotes [38]. In both studies all the remaining genes of the array are predicted to be more abundant in promastigotes, which strengthens the argument of developmental regulation for functional differentiation within the tandem array and in this particular lineage.

SCGL is a gene lineage that likely arose from a transposition event from the SCGR array in chromosome 2 to chromosome 14. Members of this family are found in *Leishmania* genus only in a paraphyletic clade with single gene copies in *Crithidia* and *Lotmaria*. When first described, LmjF.14.1400 was detected at low levels in all life cycle stages, compared to the high expression of SCGR and SCG members [61], which was corroborated by proteomics in *L. major* and *L. infantum*, where the gene was either not detected [62] or constitutively expressed at low abundance [38]. These data combined suggest that localization in the tandem array is essential for high protein expression and that transposition from the array may have resulted in functional differentiation and expression constraints.

SCG genes are *Leishmania-*specific and located at the subtelomeres of several chromosomes. In *L. major*, these genes have been shown to encode functional proteins, which are expressed in the parasite [61]. Most likely, the ancestor of this genus also possessed several copies of these glycosyltransferases, although their trace has been lost due to their highly unstable genomic location. Early investigation of developmental regulation revealed LmjF.07.1170 to be more abundant in promastigotes, but LmjF.31.3170 and LmjF.35.0010 in metacyclics and amastigotes [61]. However, this contrasts with proteomic studies in the same strain, which suggest higher protein abundance in amastigotes for all SCG genes [38]. Similarly, proteomic studies in *L. infantum* revealed differential expression at the amastigote level in 3 of the 4 homologs [62]. The evidence for preferential expression in the intracellular amastigote stage is also consistent with the absence of these genes in monoxenic trypanosomatids.

Recombination seems to be happening particularly between *L. major* sequences, and although evidence of positive selection to be acting upon this clade, as previously suggested [61], could not be found, it is possible that a combination of relaxation of negative selection and subtelomeric localization are aiding coincident evolution of SCG genes in most *Leishmania* species.

In general, findings in this study agree with the literature, but indicate the need for standardization of expression studies and suggest the change in catalytic domain in some SCGR genes an interesting target for functional characterization and developmental regulation studies.

**African trypanosomes retain orthology during expansion of UGTs**

The UGTs lineages in African trypanosomes retain orthology between *T. vivax*, *T. congolense* and *T. brucei*. Extensive duplication occurred in both *T. brucei* and *T. congolense* at the subtelomeres, but not in *T. vivax*.

Lineages 3-6 seem to be under strong purifying selection, which potentially reflects functional differences. In fact, in *T. brucei*, three are differentially expressed in the BSF, and only one in PF. Of the three more abundant in BSF, Tb927.3.5660 previously described as TbGT11, was shown to be an acetylglucosaminyltransferase through gene knockout and *in vitro* activity assay, even though its sequence is highly divergent from all other eukaryotic homologs. It was experimentally demonstrated that its correspondent gene in *T. brucei* strain Lister 427 encodes a functional UDP-GlcNAc: α3-D- mannoside ß1-2-N-acetylglycosaminyltransferase I activity (EC 2.4.1.101) part of the N-glycans biosynthetic pathway involved in VSG glycosylation [63]. This enzyme was shown to have evolved specifically in *T. brucei* both at the amino acid and the substrate specificity levels. A separate study suggested that African trypanosomes UDP-sugar-dependent GT all belong to a single family evolved from a common ancestor of the ß3-glycosyltransferase (Izquierdo et al. 2009), but have the ability to catalyse distinct linkages to account for the parasite extensive glycoconjugate repertoire, which is consistent with the findings to date.

Apart from the possible association between Tb927.5.2760 and suramin efficacy and potential resistance together with 27 other genes, some of which shared N-acetylglucosamine biosynthesis activity [65], not much is known about the other two genes. An early study suggested that distinct COOH-termini in VSG impose distinct steric constrains on GPI-modifying galactosyltransferases activity [12]. Thus, it is possible that a and b VSGs require different UGTs for optimal activity, which would explain the evolutionary pressure not to change Tb927.7.300 and Tb927.5.2760 and their orthologs in *T. congolense* and *T.vivax*. Further functional studies are required to investigate the affinity of these transferases for the distinct steric conformations, if any, displayed by the different VSG families of African trypanosomes.

The only single-copy gene preferentially expressed in PF, Tb927.10.12290, is also under negative selection in all three African trypanosomes, suggesting a non-redundant function. Functional characterization has revealed *T. brucei* PF null mutants have smaller procyclins, resulting from modified GPI-anchor side chains (Izquierdo et al. 2009), suggesting Tb927.10.12290 encodes a GPI side-chain UDP-GlcNAc: βGal β1-3 GlcNAc- transferase. Furthermore, the authors also suggested involvement in the N-linked poly LacNAc chain synthesis in BSF. The latter is interesting since it would explain why this gene has been conserved in *T. vivax* and potentially even duplicated to TvY486\_0038690, since this parasite does not have a procyclic life stage. This gene has also been linked to Tb927.2.3370. Tb927.2.3370 has recently been functionally described through biochemical characterization of conditional null mutants under nonpermissive conditions [66]. This study revealed that the product of this gene is non-essential for the survival of *T. brucei* in culture, likely acting downstream the product of Tb927.10.12290, as a GPI side chain modifying UDP-Gal : βGlcNAc β1-3 Gal-transferase.

The lineage that duplicated extensively in *T. brucei* and *T. congolense* but is absent in *T. vivax* (7) is comprised of subtelomerically located genes in *T. brucei*. Subtelomeres are unstable genomic locations, where genes under neutral evolution may be transposed or expressed due to their proximity to other genes under positive selection. This may explain the unusual length of the branches particularly in *T. brucei*, even though these genes do not seem to be recombining or under positive selection. Eight of seventeen *T. brucei* genes in this clade are preferentially expressed in the procyclic stage at the proteomic level, which explains the absence of co-orthology in *T. vivax*, since this parasite does not have a fly midgut stage. Time-point proteomics analysis of stumpy to procyclic form differentiation identified five genes (Tb927.4.5260, Tb927.4.5270, Tb927.4.5280, Tb927.4.5290, and Tb11.v5.0880) in this clade to be up regulated only 12h after differentiation induction, continuing until established procyclic stage [67]. This suggests these genes are involved in the late stages of stumpy form to procyclic differentiation.

What remains to be explained is why *T. brucei* retain so many procyclic-specific UGTs, and whether members of this lineage are all non-essential or redundant as Tb927.2.3370. If so, it could be hypothesized that the family expansion and its subtelomeric localization may be advantageous for expression.

The lack of gene expansion in *T. vivax* is intriguing and might be explained by crucial differences in parasite surface coating, such as a lower requirement of GPI-anchored proteins to be secreted.

***African trypanosomes and Leishmaniinae UGTs have different evolution patterns***

Although the UGT repertoire expansion is common to most trypanosomatids, the approach to lineage evolution is remarkably distinct. Whilst African trypanosomes UGTs retain orthology throughout lineage expansion, in *Leishmaniinae* they generally do not. SCG and SCGR genes appear to be under concerted evolution although at different rates. Orthology is retained only on the single copy SCGL genes, which have been physically moved from the tandem array, and those SCGR genes where a change in the catalytic domain occurred. The differences in lineage evolution may be linked to key differences in function of these proteins. While UGTs in African trypanosomes are mostly involved GPI-anchor side chains glycosylation for structural conservation of the parasite surface coat [12,27,64,66], *Leishmaniinae* proteins catalyse the modification of phosphoglycan repeats in the LPG coat, which allows stage-specific adhesion mechanisms. Differences and combinations in modifications have been linked to transmissibility in the sandfly vector and parasite fitness in the mammalian host [46]. The identification of SCG genes as *Leishmania*-specific and more abundantly expressed in the amastigote stage may suggest them as interesting targets for therapeutics, whilst the general SCGR genes may be useful for transmission disruption research.

**Conclusion**

Multiple copies, a need for greater decoration of cell surfaces, evolved in parallel.

Phylogenies betray different paths to multiplicity, determined by the different starting points.

Ancestral african trypanosome evolved seven loci not shared with other tryps, subsequent adaptation has been through duplication of two loci (in brongo), through gene duplication at sub-telomeric loci, a bi-product of the evolution of sub-telomeres for VSG.

Ancestral leish evolved a tandem gene array containing at least three distinct lineages. Adaptation has proceeded through transposition away from the array or functional differentiation within it.

So their origins are different, which is consistent with the view that their dixenic origins were independent, so apparently are the evolutionary forces incident upon them. Dev reg clearly plays a major role in driving diversification, yet the neutral evolution of AT duplicates suggests a dosage function not observed in leishmania.

This study has identified a protolog that is the most basal branching lineage in this family, and which may retain a plesiomprhic character that makes it a good model for the ancestral UGT in the non-parasitic kinetoplastid. The presence of this lineage in both Bodo and T. cruzi offers an opportunity to test this hypothesis by comparing the function and regulation of the protolog in comparison with parasite-specific UGT.

Like the cell surface molecules they modify, UGT repertoires have diversified in tryps to meet the demands of transmission and immune evasion through a multi-stage life cycle. That they have done so independently in Trypanosoma and Leishmania shows that novel adaptations of the cell surface have continued to innovate after the origin of parasitism to meet the dynamic needs of a host-parasite interaction.

The UGT phylogeny reflects the wider theme of genome evolution being driven by host-parasite interactions that are unique to each parasite lineage as revealed in the parallel evolution from an ancestral state of low diversity to higher diversity, following the need for glycosylation of cell surface, but stressing the very specific nature of the decoration required. The infection of different hosts and vectors drove the need for the development of specific interaction mechanisms. The key role of UGTs in those interactions expedited genus- and species-specific adaptation of these enzymes to fit precise roles required by the colonization of different environments.

**Figure legends**

**Figure 1** **Consensus maximum likelihood phylogeny of UDP-glycosyltransferases protein sequences from diverse trypanosomatids and *B. saltans*.** The phylogeny was estimated with PHYML using a maximum likelihood method with a WAG+Γ model and 500 bootstrap replicates. Terminal nodes are named with Genedb [45] and Tritypdb [44] identifiers; internal nodes are labelled with bootstrap percentages for maximum likelihood and neighbor joining methods. Tips are labelled according to key. The tree is rooted with two *Bodo saltans* sequences as the outgroup.

**Figure 2 Consensus maximum likelihood phylogeny of UDP-glycosyltransferases nucleotide sequences from *Leishmaniinae*.** The phylogeny was estimated with PHYML with a GTR+ model and 500 bootstrap replicates. Terminal nodes are named with Tritypdb [44] identifiers; internal nodes are labeled with bootstrap percentages for maximum likelihood (ML), posterior probabilities (BI), and neighbor joining (NJ) methods for nucleotides and protein sequences. Tree is rooted with *Angomonas deanei* and *Strigomonas culicis* as outgroup.The gray star indicates a gene loss in *L. mexicana* and *L. braziliensis* after the gene transposition event to chromosome 14. Tips are labelled according to key. Clades are identified as SCG, SCGR, and SCGL according to previous annotation in L. major (Dobson *et al.*, 2006). Available expression data is represented as log2 fold change of amastigote (AMA), constitutive (C), or promastigotes (PRO) in a heat map according to study reference. Asterisk indicates data is transcriptomic.

**Figure 3 Structural conservation of genomic loci containing UDP-glycosyltransferases among *Leishmaniinae* species.** Conserved genomic locus in *L. major* Friedlin, *L. braziliensis* M2904, *Crithidia fasciculata* CfC1, *Leptomonas pyrrhocoris* H10, and *T. cruzi* Esmeraldo. The UDP-glycosyltransferase genes are shaded yellow, flanking genes are shaded orange, arabinosyltransferase genes are shaded in green, and other hypothetical proteins in the array are shaded in black; sequence homology is illustrated by grey vertical bars. Gene terminology is according to Tritrypdb [44] identifiers. Comparisons were obtained with Artemis Comparison Tool (ACT) [68].

**Figure 4 Consensus maximum likelihood phylogeny of UDP-glycosyltransferases nucleotide sequences from African trypanosomes.** The phylogeny was estimated with PHYML a GTR+Γ model and 500 bootstrap replicates. Terminal nodes are named with Genedb [45] and Tritrypdb [44] identifiers, internal nodes are labelled with bootstrap percentages for maximum likelihood (ML), posterior probabilities (BI), and neighbor joining (NJ) methods for nucleotides and protein sequences. The tree is rooted with *Bodo saltans* as outgroup. Tips are labelled according to key. Available expression data is represented as log2 fold change as procyclic (PF), constitutive (C), or bloodstream form (BF) in a heat map according to study reference. Asterisk indicates data is transcriptomic.

**References**

1. Stuart K, Brun R, Croft S, Fairlamb A, Gürtler RE, Mckerrow J, et al. Review series Kinetoplastids : related protozoan pathogens , different diseases. 2008;118:1301–10.

2. Alves JMP, Klein CC, da Silva FM, Costa-Martins AG, Serrano MG, Buck GA, et al. Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. BMC Evol. Biol. [Internet]. 2013;13:190. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3846528&tool=pmcentrez&rendertype=abstract

3. Runckel C, DeRisi J, Flenniken ML. A draft genome of the honey bee trypanosomatid parasite crithidia mellificae. PLoS One. 2014;9.

4. Flegontov P, Butenko A, Firsov S, Kraeva N, Eliáš M, Field MC, et al. Genome of Leptomonas pyrrhocoris: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of Leishmania. Sci. Rep. [Internet]. Nature Publishing Group; 2016;6:23704. Available from: http://www.nature.com/articles/srep23704

5. H. Lopes A. Trypanosomatids: Odd Organisms, Devastating Diseases. Open Parasitol. J. 2010;4:30–59.

6. Wheeler RJ, Gluenz E, Gull K. The limits on trypanosomatid morphological diversity. PLoS One. 2013;8.

7. Snary D. The Cell Surface of Trypanosoma cruzi. In: Hudson L, editor. Biol. Trypanos. [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1985. p. 75–92. Available from: http://dx.doi.org/10.1007/978-3-642-70538-0\_4

8. Ilgoutz SC, McConville MJ. Function and assembly of the Leishmania surface coat. Int. J. Parasitol. 2001;31:899–908.

9. Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M. A Cell-surface Phylome for African Trypanosomes. PLoS Negl. Trop. Dis. [Internet]. 2013 [cited 2014 Sep 11];7:e2121. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605285&tool=pmcentrez&rendertype=abstract

10. Doering TL, Masterson WJ, Englund PT, Hart GW. Biosynthesis of the Glycosyl Phosphatidylinositol Membrane Anchor of the Trypanosome Variant Surface Glycoprotein t t t. J. Biol. Chem. 1989;264:11168–73.

11. Hirose S, Ravi L, Prince GM, Rosenfeld MG, Silber R, Andresen SW, et al. Synthesis of mannosylglucosaminylinositol phospholipids in normal but not paroxysmal nocturnal hemoglobinuria cells. Proc. Natl. Acad. Sci. U. S. A. 1992;89:6025–9.

12. Ferguson MAJ, Brimacombe JS, Brown JR, Crossman A, Dix A, Field RA, et al. The GPI biosynthetic pathway as a therapeutic target for African sleeping sickness. Biochim. Biophys. Acta - Mol. Basis Dis. 1999;1455:327–40.

13. Bock KW. The UDP-glycosyltransferase (UGT) superfamily expressed in humans, insects and plants: Animal-plant arms-race and co-evolution. Biochem. Pharmacol. [Internet]. Elsevier Inc.; 2015;99:11–7. Available from: http://dx.doi.org/10.1016/j.bcp.2015.10.001

14. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: Structures, Functions, and Mechanisms. Annu. Rev. Biochem. 2008;77:521–55.

15. Egelund J, Ellis M, Doblin M, Qu Y, Bacic A. Genes and Enzymes of the GT31 Family: Towards Unravelling the Function(s) of the Plant Glycosyltransferase Family Members. Annu. Plant Rev. 2010. p. 213–34.

16. Kitagawa H, Uyama T, Sugahara K. Molecular cloning and expression of a human chondroitin synthase. J. Biol. Chem. 2001;276:38721–6.

17. LeBon L, Lee T V., Sprinzak D, Jafar-Nejad H, Elowitz MB. Fringe proteins modulate Notch-ligand cis and trans interactions to specify signaling states. Elife. 2014;3:e02950.

18. Wang S, Czuchry D, Liu B, Vinnikova AN, Gao Y, Vlahakis JZ, et al. Characterization of two UDP-Gal: GalNAc-diphosphate-lipid β1,3-galactosyltransferases WbwC from Escherichia coli serotypes O104 and O5. J. Bacteriol. 2014;196:3122–33.

19. Damerow S, Lamerz AC, Haselhorst T, Führing J, Zarnovican P, von Itsztein M, et al. Leishmania UDP-sugar pyrophosphorylase: The missing link in galactose salvage? J. Biol. Chem. 2010;285:878–87.

20. Sacks D, Kamhawi S. Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis. Annu. Rev. Microbiol. [Internet]. 2001;55:453–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11544364

21. Dillon RJ, Lane RP. Detection of Leishmania lipophosphoglycan binding proteins in the gut of the sandfly vector. Parasitology [Internet]. 1999;118:27–32. Available from: http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=24326\nhttp://journals.cambridge.org/abstract\_S0031182098003588\nhttp://www.ncbi.nlm.nih.gov/pubmed/10070658

22. Pimenta P, Turco S, McConville M, Lawyer P, Perkins P, Sacks D. Stage-specific adhesion of Leishmania promastigotes to the sandfly midgut. Science (80-. ). [Internet]. 1992;256:1812–5. Available from: https://www.researchgate.net/publication/21541059\_Pimenta\_P\_F\_et\_al\_Stage-specific\_adhesion\_of\_Leishmania\_promastigotes\_to\_the\_sandfly\_midgut\_Science\_256\_1812-1815

23. Mehlert A, Bond CS, Ferguson M a J. The glycoforms of a Trypanosoma brucei variant surface glycoprotein and molecular modeling of a glycosylated surface coat. Glycobiology [Internet]. 2002;12:607–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12244073

24. Mehlert A, Wormald MR, Ferguson MAJ. Modeling of the N-glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of trypanosoma brucei. PLoS Pathog. 2012;8:1–11.

25. Nolan DP, Geuskens M, Pays E. N-linked glycans containing linear poly-N-acetyllactosamine as sorting signals in endocytosis in Trypanosoma brucei. Curr. Biol. 1999;9:1169–72.

26. Nagamune K, Acosta-Serrano A, Uemura H, Brun R, Kunz-Renggli C, Maeda Y, et al. Surface sialic acids taken from the host allow trypanosome survival in tsetse fly vectors. J. Exp. Med. 2004;199:1445–50.

27. Izquierdo L, Nakanishi M, Mehlert A, Machray G, Barton GJ, Ferguson MA. Identification of a glycosylphosphatidylinositol anchor-modifying beta1-3 N-acetylglucosaminyl transferase in Trypanosoma brucei. Mol Microbiol [Internet]. 2009;71:478–91. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\_uids=19040631

28. Roper JR, Guther MLS, Milne KG, Ferguson MAJ. Galactose metabolism is essential for the African sleeping sickness parasite Trypanosoma brucei. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2002;99:5884–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=122871&tool=pmcentrez&rendertype=abstract

29. Roper JR, G??ther MLS, MacRae JI, Prescott AR, Hallyburton I, Acosta-Serrano A, et al. The suppression of galactose metabolism in procylic form Trypanosoma brucei causes cessation of cell growth and alters procyclin glycoprotein structure and copy number. J. Biol. Chem. 2005;280:19728–36.

30. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science. 2005;309:416–22.

31. El Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science (80-. ). [Internet]. 2005;309:409–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16020725

32. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science (80-. ). [Internet]. 2005;309:436–42. Available from: http://www.sciencemag.org/content/309/5733/436.full.pdf

33. Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, et al. The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human African Trypanosomiasis. PLoS Negl. Trop. Dis. 2010;4.

34. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton J a., et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. 2011;2143–56. Available from: http://dx.doi.org/10.1101/gr.123430.111

35. Jackson AP, Barry JD. The Evolution of Antigenic Variation in African Trypanosomes. In: Sibley LD, Howlett BJ, Heitman J, editors. Evol. Virulence Eukaryot. Microbes. Wiley-Blackwell; 2012. p. 324–37.

36. Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome Trypanosoma grayi. Sci. data [Internet]. 2014;1:140024. Available from: http://www.nature.com/articles/sdata201424

37. Kraeva N, Butenko A, Hlaváčová J, Kostygov A, Myškova J, Grybchuk D, et al. Leptomonas seymouri: Adaptations to the Dixenous Life Cycle Analyzed by Genome Sequencing, Transcriptome Profiling and Co-infection with Leishmania donovani. PLoS Pathog. 2015;11.

38. Akopyants NS, Kruvand E, Wong I, Beverley SM. Manuscript in preparation. 2010;

39. Butter F, Bucerius F, Michel M, Cicova Z, Mann M, Janzen CJ. Comparative proteomics of two life cycle stages of stable isotope-labeled Trypanosoma brucei reveals novel components of the parasite’s host adaptation machinery. Mol. Cell. Proteomics. 2012;172–9.

40. Urbaniak MD, Guther MLS, Ferguson MAJ. Comparative SILAC proteomic analysis of trypanosoma brucei bloodstream and procyclic lifecycle stages. PLoS One. 2012;7.

41. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GAM. Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. Nucleic Acids Res. 2010;38:4946–57.

42. Jackson AP, Goyard S, Xia D, Foth BJ, Sanders M, Wastling JM, et al. Global gene expression profiling through the complete life cycle of Trypanosoma vivax. PLoS Negl. Trop. Dis. 2015;9:1–29.

43. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Curr. Biol. 2016;26:161–72.

44. Aslett M, Aurrecoechea C, Berriman M, Al. E. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. [Internet]. 2010;38:D457-62. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808979&tool=pmcentrez&rendertype=abstract

45. Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, et al. GeneDB-an annotation database for pathogens. Nucleic Acids Res. 2012;40.

46. Dobson DE, Scholtes LD, Valdez KE, Sullivan DR, Mengeling BJ, Cilmi S, et al. Functional identification of galactosyltransferases (SCGs) required for species-specific modifications of the lipophosphoglycan adhesin controlling Leishmania major-sand fly interactions. J. Biol. Chem. 2003;278:15523–31.

47. Whelan S, Liò P, Goldman N. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 2001. p. 262–72.

48. Guindon S, Gascuel O. A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood. Syst. Biol. 2003;52:696–704.

49. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. [Internet]. 2016;msw054. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27004904

50. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogeny. Bioinformatics. 2001;17:754–5.

51. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4.

52. Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 1994;11:605–12.

53. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172:2665–81.

54. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 2006. p. 254–67.

55. Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 2006;23:1891–901.

56. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 2005;22:1208–22.

57. Scheffler K, Martin DP, Seoighe C. Robust inference of positive selection from recombining coding sequences. Bioinformatics. 2006;22:2493–9.

58. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. Mol. Biol. Evol. 2013;30:1196–205.

59. Xu B, Yang Z. PamlX: A graphical user interface for PAML. Mol. Biol. Evol. 2013;30:2723–4.

60. Minning T a, Weatherly DB, Atwood J, Orlando R, Tarleton RL. The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. BMC Genomics [Internet]. 2009;10:370. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2907688&tool=pmcentrez&rendertype=abstract

61. Dobson DE, Scholtes LD, Myler PJ, Turco SJ, Beverley SM. Genomic organization and expression of the expanded SCG/L/R gene family of Leishmania major: Internal clusters and telomeric localization of SCGs mediating species-specific LPG modifications. Mol. Biochem. Parasitol. 2006;146:231–41.

62. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. Mol. Biochem. Parasitol. 2009;165:32–47.

63. Damerow M, Rodrigues J a, Wu D, Güther MLS, Mehlert A, Ferguson M a J. Identification and Functional Characterization of a Highly Divergent N-Acetylglucosaminyltransferase I (TbGnTI) in Trypanosoma brucei. J. Biol. Chem. [Internet]. 2014;289:9328–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24550396

64. Izquierdo L, Atrih A, Rodrigues JA, Jones DC, Ferguson MAJ. Trypanosoma brucei UDP-glucose:glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress. Eukaryot. Cell. 2009;8:230–40.

65. Alsford S, Eckert S, Baker N, Glover L, Sanchez-Flores A, Leung KF, et al. High-throughput decoding of anti-trypanosomal drug efficacy and resistance. Nature. 2012;482:232–6.

66. Izquierdo L, Acosta-Serrano A, Mehlert A, Ferguson MA. Identification of a glycosylphosphatidylinositol anchor-modifying ??1-3 galactosyltransferase in Trypanosoma brucei. Glycobiology. 2015;25:438–47.

67. Dejung M, Subota I, Bucerius F, Dindar G, Freiwald A, Engstler M, et al. Quantitative Proteomics Uncovers Novel Factors Involved in Developmental Differentiation of Trypanosoma brucei. PLoS Pathog. 2016;12:1–20.

68. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: The Artemis comparison tool. Bioinformatics. 2005;21:3422–3.