Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species

Andrew P. Jackson[1]*, Andrew Berry[1], Martin Aslett[1], Harriet C. Allison[2], Peter Burton[3], Jana Vavrova-Anderson[3], Robert Brown[4], Hilary Browne[1], Nicola Corton[1], Heidi Hauser[1], John Gamble[1], Ruth Gilderthorp[1], Jacqueline McQuillan[1], Thomas D. Otto[1], Michael A. Quail[1], Mandy Sanders[1], Andries Van Tonder[1], Michael L. Ginger[4], Mark C. Field[2], J. David Barry[3], Christiane Hertz-Fowler[1,5], Matthew Berriman[1]

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA. U.K.
[2] Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, U.K.
[3] Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, 120 University Place, Glasgow, G12 8TA. U.K.
[4] School of Health and Medicine, Division of Biomedical and Life Sciences, Lancaster University, Lancaster, LA1 4YQ. U.K.
[5] Centre for Genomic Research, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool, L69 7ZB. U.K.

*Email andrew.jackson@sanger.ac.uk

Abstract

Antigenic variation enables pathogens to avoid the host immune response by continual switching of surface proteins. The protozoan blood parasite *Trypanosoma brucei* is a model system for antigenic variation, and survives by periodically replacing a monolayer of variant surface glycoproteins (VSG) covering its cell surface. We compared the genome of *T. brucei* with two closely related parasites *T. congolense* and *T. vivax*, to reveal how the variant antigen repertoire evolved, and how this might affect contemporary antigenic diversity. Here we show that *VSG* in each species have distinct patterns of sequence variation and phylogenetic diversity, due to the divergent evolutionary trajectories each has followed, and reflecting fundamental differences in the scale and mechanism of recombination.

Antigenic variation enables pathogens to evade immune responses by continual switching of surface proteins[1,2]. The African trypanosome *Trypanosoma brucei* is a protozoan blood parasite that causes human African Trypanosomiasis ('sleeping sickness') across sub-Saharan Africa. It survives in the host by periodically replacing a monolayer of variant surface glycoproteins (VSG[3]) that shield its cell surface[4-5]; the mechanisms for expression and dynamic replacement of VSG are a model system for antigenic variation[4]. We compared the genomes of *T. brucei* and two closely related parasites *T. congolense* and *T. vivax*, to better understand how *VSG* repertoire has evolved and how this affects contemporary antigenic variability. The *T. brucei* genome includes many hundreds of *VSG*[6] but each cell expresses just a single gene from a specialized telomeric expression site at any time[4-5]. The parasite population collectively express multiple *VSG*; when the host becomes immune to the dominant type, clones expressing alternative copies proliferate in a frequency-dependent manner, maintaining the infection and resulting in characteristic 'waves of parasitaemia'. To survive long-term, *T. brucei* must generate novel *VSG* sequences through recombination; mechanisms may include domain shuffling[7], and *in situ* gene conversion, possibly within the expression site[8-9]. Functional variant antigens in *T. brucei* consist of a- and b-type *VSG* (hereafter a-*VSG* and b-*VSG*), which share the cysteine-rich carboxy-terminal domain (CTD) but are otherwise distantly related[10-12]. It has been suggested that these *VSG* are a source of novel genes. Two gene families, the Expression Site-Associated Genes (*ESAG6*/7) encoding transferrin receptors and the *VSG*-related (*VR*) genes, are thought to have evolved from a-*VSG*[13-15] and b-*VSG*[9,12] respectively.

Results

The *VSG* gene repertoires of *T. congolense* and *T. vivax*

We have produced high-quality draft genome sequences for *Trypanosoma congolense* IL3000, a sister species of *T. brucei*, and *Trypanosoma vivax* Y486, a third species that branches close to the root of the African trypanosome lineage[16]. These genome sequences are described in supplementary information (see Supplementary Table 1) and are accessible through GeneDB (www.genedb.org) or TritrypDB (www.tritryp.org). Comparative analysis including the existing *T. brucei* 927 genome sequence shows that the principal differences in genome content relate to cell surface architecture (see Supplementary Table 2-4). To define *VSG* repertoires, gene sequences with predicted cell surface roles were extracted from all three genomes and were sorted using BLASTx, resulting in 81 gene families (see Methods and Supplementary Table 5). Phylogenies of these families were estimated that we collectively termed the 'cell-surface phylome', (www.genedb.org/Page/trypanosoma_surface_phylome). The phylome contains *VSG* and related families already known in *T. brucei* but it also defines new families that we believe encode the *VSG* repertoires of *T. congolense* and *T. vivax*.

The *T. congolense VSG* repertoire differs from that of *T. brucei* in three ways. First, there is no a-*VSG* subfamily of variant antigens; second, there are two b-*VSG* subfamilies, termed Fam13 (n = 302) and Fam16 (n = 512) by their phylome designations; and third, unlike *T. brucei VSG*, which all share a relatively uniform CTD, *T. congolense VSG* have 15-20 different CTD types, each associated with a specific subset of Fam13 or 16, and none homologous to the *T. brucei* CTD. Hence, *T. congolense* b-*VSG* are more structurally heterogeneous than *T. brucei* b-*VSG*. We know that both Fam13 and 16 contain functional variant antigens because each family includes examples of published *T. congolense VSG* and *VSG* expressed sequence tags (EST). While there is no a-*VSG* variant antigen, there are homologs of the a-*VSG*-like transferrin receptor genes of *T. brucei*, i.e. Procyclin-Associated Genes (*PAG*) (Fam14, n = 22) and *ESAG6/7* (Fam15, n = 43).

VSG structural diversity is even greater in *T. vivax*. We have identified four *VSG* subfamilies (Fam23-26) that each possess definitive patterns of conserved cysteine residues (see supplementary information). Fam23 (n=540) and Fam24 (n=279) members possess sequence motifs homologous to a-*VSG* and b-*VSG* respectively. Fam25 (n = 227) and Fam26 (n = 87) are two subfamilies unique to *T. vivax*, but with low (~20%) protein sequence similarity to known VSG (see Supplementary Fig. 1). These may have evolved in *T. vivax*, or may represent ancestral lineages not inherited by *T. brucei* and *T. congolense.* Transcriptomic data shows that multiple members of all four families are transcribed in bloodstream-stage cells (see Supplementary Table 6). We find no orthologs to the transferrin receptor-like genes of *T. brucei* and *T. congolense* among *T. vivax VSG*-like genes or indeed the numerous, novel *T. vivax*-specific gene families.

Amino acid sequence homology with *T. brucei VSG* alone does not guarantee that putative *T. vivax VSG* function as variant antigens. To date, only one *T. vivax* VSG (ILDat 2.1[17]) has been characterized, albeit from a dissimilar strain, and is most closely related to Fam26. Therefore, we identified an expressed VSG in the genome strain Y486 by mass-spectrometry analysis of a protein specific to a relapsed infection population, peptide fragments of which are 100% identical to a predicted protein in Fam23 (TvY486_0027060; see Supplementary Fig. 2). Therefore, at least one a-*VSG*-like (i.e. Fam23) gene in *T. vivax* encodes a functional variant antigen.

The phylogeny of *VSG* diversification

In total, the three genome sequences yielded 1083 a-*VSG*-like and 1537 b-*VSG*-like full-length genes (see Supplementary Table 7). We estimated Bayesian and Maximum Likelihood phylogenies from amino acid sequence alignments (see Supplementary Figs. 3-4) but, given the large number of sequences, and to enable global visualization, we also estimated a similarity network from pair-wise maximum likelihood protein distances that delivered a clearer picture of relationships within the a- and b-*VSG* lineages. The distance network includes examples of all *VSG* subfamilies and represents individual genes as spheres connected to others sharing identity above a threshold (see Methods). The network and phylogenies are fully consistent. Figure 1 shows the similarity network from two angles (a supplementary video displays the network in three-dimensions); four principal features emerge.

First, the common CTD of *T. brucei VSG* must have evolved through horizontal transfer from one subfamily to the other. In Figure 1, sequences cluster by lineage (a or b) rather than by species; for instance, *T. vivax* a-*VSG* (Fam23) is more similar to a-*VSG*-like subfamilies in *T. brucei* and *T. congolense* than to *T. vivax* b-*VSG* (Fam24). This demonstrates that *VSG* lineages are older than the genomes they occupy, indeed, they were present in the common ancestor of all African trypanosomes. The only above-threshold sequence connections occurring between a- and b-*VSG* subfamilies (point *i*) concern *T. brucei VSG* and, in particular, their common CTD. This is a unique feature of *T. brucei VSG* and an exception that proves the rule: despite belonging to ancient lineages separated in the ancestral trypanosome, a- and b-*VSG* in *T. brucei* share a CTD that is species-specific. This can only be explained if the CTD evolved in one subfamily and was transposed to the other.

Second, b-*VSG* in *T. brucei* are derived from a single ancestral lineage while *T. congolense* b-*VSG* are drawn from many lineages, which suggests that *T. brucei* b-*VSG* have passed through a 'bottleneck'. In Fig. 1, all b-*VSG* in *T. brucei* (dark blue) form a cluster to the exclusion of other subfamilies. Hence, they share a recent common ancestor that evolved after the split from *T. congolense*. In contrast, *T. congolense* b-*VSG* comprise two lineages (Fam13 and 16) that originated in the *T. brucei/congolense* ancestor and form separate clusters in the network (point *ii*). We know that these lineages did not originate in *T. congolense* because their closest relatives are *VSG*-like genes in *T. brucei* (see below). In fact, Fam13 and 16 themselves split into multiple clusters in Fig. 1 (point *ii*), emphasizing the phylogenetic diversity of *T. congolense VSG* and relative homogeneity in *T. brucei*.

Third, *VSG* have repeatedly been a source of functional novelty on the cell surface. We know that *VSG* can be co-opted from variant antigen functions to novel roles, for example, the serum-resistance antigen (*SRA*[18]) and *TgsGP*[19] proteins in *T. b. rhodesiense* and *T. b. gambiense* respectively. However, these represent secondary loss of function in contemporary *VSG*. Figure 1 shows that *ESAG2*, a gene family associated with the polycistronic *VSG* expression site in *T. brucei*, is a b-*VSG*-like gene, nested among *T. congolense* b-*VSG* (Fam13, point *iii*). Similarly, *VR* genes (purple in Fig. 1), rather than being derived from b-*VSG* in *T. brucei*, have an ancestral-type structure, more akin to Fam16 in *T. congolense*. We have also identified another *T. brucei*-specific family (Fam1; pink in Fig. 1), which encode proteins homologous to b-VSG, with a predicted GPI-anchor, but also a highly modified CTD. Fam1 (i.e. Tb927.6.1310) is preferentially expressed in bloodstream forms and localizes to the flagellar pocket and endosomal membranes (see Supplementary Fig. 5). Phylogenetic analysis clearly demonstrates that both *ESAG2* and *VR* gene subfamilies, for which the evidence is against a variant antigen function[9,12], are not recent derivations from *T. brucei VSG*, (like *SRA* and *TgsGP*), but belong to ancestral *VSG* lineages with representatives in *T. congolense* that still encode functional variant antigens (see Supplementary Fig. 4). Hence, some of the ancestral lineages in *T. congolense* identified above remain in *T. brucei* but have been co-opted to novel roles.

Lastly, the network indicates that the transferrin receptor evolved from an a-*VSG* gene as suggested previously[9,20-21]. However, this did not occur within the *T. brucei VSG* expression site but instead in the *T. brucei/congolense* ancestor. A tight cluster of transferrin receptor-like genes (i.e. *ESAG6/7* and *PAG*) from *T. brucei* as well as Fam14 and 15 sequences from *T. congolense* is distinct from other a-*VSG* subfamilies in Fig. 1 (point *iv*). This reflects their phylogeny, which shows that Fam14 and 15 are sister lineages to *PAG* and *ESAG6/7* respectively, and their primary structures, which show that amino acid residues crucial for transferrin-binding[15] are conserved in both species (see Supplementary Fig. 6). Given the absence of this family from *T. vivax*, we conclude that the transferrin-receptor genes evolved prior to the separation of *T. brucei* and *T. congolense* but after their split from *T. vivax*. This does not preclude other *T. vivax*-specific proteins performing a transferrin-binding function in that species.

These results are summarized in a model of *VSG* evolution (Fig. 2). The ancestral African trypanosome possessed a- and b-*VSG* type genes; which probably formed multi-gene families or functioned as variant antigens. Both *VSG* types were inherited by *T. vivax*, the a-*VSG* family of which includes functional variant antigens. The *T. brucei-congolense* ancestor inherited both a- and b-*VSG* lineages and at this point one a-*VSG* gene was co-opted to a transferrin-binding role differentiated between insect and vertebrate life stages, founding a lineage that was inherited by both daughter species. Another a-*VSG* lineage retained its variant antigen function in *T. brucei*, but was lost from *T. congolense* (see supplementary information). Of the ancestral b-*VSG* repertoire, two different lineages have been inherited by both species. The first has produced *ESAG2* and Fam13 in *T. brucei* and *T. congolense* respectively; while the second has produced b-*VSG* and *VR* in *T. brucei* and Fam16 in *T. congolense.* There is no step in this deduced scheme where a trypanosome lacks variant antigen. Clearly, these two species have adapted their common legacy differently. *T. congolense VSG* are drawn from multiple ancestral lineages, whereas *T. brucei* has relegated corresponding genes (*VR*, *ESAG2*, and perhaps Fam1) to novel roles, and derives its variant antigens from single lineages, derived after speciation. This difference in the phylogenetic diversity of *VSG* repertoires is important because it could affect the ability of the parasites to present novel antigens to their hosts, and therefore maintain infection.

Tree shape and the distribution of *VSG* sequence variation

We examined the phylogenies of *VSG* subfamilies within species for evidence that their distinct evolutionary legacies have affected contemporary sequence evolution. Figure 3 demonstrates how these trees have distinct topologies. This is due to variation in the ratio of internal to terminal branches, (described by 'treeness'[22], *T*), which is low for *T. brucei* (T = 0.282 and 0.275), higher for *T. congolense* (T = 0.376 and 0.412) and highest for *T. vivax* (T = 0.681 and 0.763). *T. congolense* and *T. vivax* trees are more 'tree-like' because they retain information about the past in basal nodes and internal branches, while the *T. brucei* tree consists mostly of long, terminal branches. Figure 3 also compares the distribution of *VSG* sequence variation, showing that *T. brucei* distances have much narrower distributions than either *T. congolense* or *T. vivax VSG* because both short, terminal branches and long, basal internodes are rare. Importantly, these patterns are genome-specific rather than lineage-specific effects, i.e. a- and b-*VSG* in *T. brucei* display the same dynamic despite having greater identity with subfamilies in other species. They confirm that the mechanisms for antigenic variability vary between species now and likewise in the past.

Recombination is a principal evolutionary pressure affecting *T. brucei VSG*[5,9], and exchange of the unique *VSG* C-terminal domain is well recorded[7,12]. Recombination is also the mechanism through which *VSG* are transposed from subtelomeric loci into the telomeric expression site[4,5,8,9]. *T. brucei VSG* phylogenies in Figure 3 are consistent with frequent recombination but the cladistic structure of *T. congolense* and *T. vivax VSG* phylogenies could only persist if recombination between clades is rare. Furthermore, the incidence of pseudogenes, which are thought to result from gene conversion between *VSG* genes[5], is much lower in *T. congolense*, (where only 21.1% of Fam13 and 29.7% of Fam16 are predicted pseudogenes), and *T. vivax* (15.5% and 27.2% of Fam23 and Fam24 respectively), than in *T. brucei*, (69.2% of a-*VSG* and 72.2% of b-*VSG*)[6]. Therefore, we suspected that recombination frequency might account for species differences in sequence variation.

The contribution of recombination to antigenic diversity

We examined *VSG* alignments for evidence of recombination, in the form of phylogenetic incompatibility (PI)[23-24], taking random samples of each alignment set and observing the proportion showing significant PI ($P_{pi.}$; see Supplementary Table 8). Figure 4 shows that P_{pi} (color lines) was greatest for *T. brucei* a-*VSG* (0.392) and b-*VSG* (0.450) and Fam16 (0.433), and lower for Fam13 (0.125) and *T. vivax* Fam23 (0.138) and Fam24 (0.126). In all cases, observed P_{pi} was significantly greater than a null distribution (black lines), confirming that PI was not solely due to other homoplastic effects, such as rate heterogeneity (see methods). Recombination frequency is known to be proportional to sequence identity[25-26] and when we increased sequence identity within alignments by sampling only within crown clades, P_{pi} increased significantly (dashed lines) for *T. brucei* a-*VSG* (0.681) and b-*VSG* (0.642), and for *T. congolense* Fam13 (0.466) and Fam16 (0.823), but not for *T. vivax*. Finally, as the CTD is known to recombine in *T. brucei*[7,12], we removed the CTD from *T. brucei* and *T. congolense* alignments; this resulted in a significant decrease in P_{pi} for *T. brucei* a-*VSG* (0.152, $p < 0.0001$) and b-*VSG* (0.234, $p < 0.0001$), but in *T. congolense* P_{pi} actually increased.

Therefore, in *T. brucei* and *T. congolense* the evidence for recombination is greatest among closest related *VSG*, but was seldom observed in *T. vivax*, even when sampling within clusters of highly related sequences. While the frequency of PI is similar for *T. brucei VSG* and Fam16, if we compare P_{pi} in a global alignment of *T congolense* b-*VSG* (0.163) with the corresponding value for *T. brucei* (0.450), it is clear that PI is prevalent throughout the *T. brucei* repertoire but only within *T. congolense VSG* clades. This is a sampling effect caused by their divergent evolutionary histories. Given that *T. congolense VSG* are phylogenetically diverse and have a wider distribution of sequence variation, they have proportionally more distant relationships and so more structural barriers to genetic exchange. In short, there are cohorts of *T. congolense VSG* that never recombine, as the topological differences in Figure 3 suggest.

Discussion

The past and present evolution of *VSG* can now be brought together. We have shown that the composition of contemporary *VSG* repertoires is determined by how each species has modified the common inheritance. *T. vivax* has the most structurally-diverse repertoire comprising a-*VSG*, b-*VSG* and two additional types absent elsewhere; *T. congolense* combines multiple, ancestral b-*VSG* lineages each with a distinct CTD, while *T. brucei* a- and b-*VSG* are recently derived, single lineages with a common CTD. It is worth remembering that sequence mosaics generated in late *T. brucei* infections have the potential to further increase *VSG* diversity[8,12]; it is not known if this dynamic assortment of *VSG* sequences occurs in other species. Nevertheless, as a result of compositional differences, the scale of recombination varies between species, being more frequent among *T. brucei* and *T. congolense VSG* than in *T. vivax*, and more prevalent among *T. brucei VSG* than in *T. congolense*. However, PI in *T. brucei VSG* is due in large part to the CTD promoting exchange throughout the repertoire, whereas the conservative CTDs of *T. congolense VSG* actually reduce the scale of PI and illustrate the lack of recombination between clades.

Differences in the role of the CTD indicate that, in addition to scale, the mechanism of recombination also varies between species. The CTD is exchanged between *T. brucei VSG*, but does not solicit an immune response and therefore, may not directly contribute to antigenic diversity[27]. However, it has been speculated that the CTD may have a role in the transposition of *VSG*, which is of paramount importance to antigenic variation[4,5,8,9]. *VSG* genes are frequently transposed around the *T. brucei* genome through gene conversion, and this is required to move *VSG* from silent, subtelomeric loci into the telomeric expression site, from where a single *VSG* is transcribed[4,8,9,12]. It has been suggested that transposition of the antigenic N-terminal domain, (i.e. the major part of the *VSG* exposed to the host), is facilitated by the 70bp repeat region, (which precedes telomeric and subtelomeric *VSG*), and the CTD, which provide conserved annealing points up- and downstream respectively[9,12]. Our observation that the majority of PI in *T. brucei VSG* alignments concerns the CTD confirms the prediction of this model that a recombination breakpoint should occur between the N- and C-terminal domains, which are essentially decoupled. Immediately, we can see that this mechanism cannot operate in *T. congolense*, where the CTDs are heterogeneous and have no role in promoting exchange. Hence, we propose that the pre-eminence of the CTD in PI reflects the frequent transposition of N-terminal domains, and through its solitary CTD type, which originated uniquely through horizontal transfer between *VSG* lineages, *T. brucei* may have evolved a distinct mechanism for the movement of *VSG* between genomic loci and into the telomeric expression site.

Antigenic variation is central to the host-trypanosome relationship, intimately linked to the course and severity of disease, to parasite transmission and host range, and therefore to disease epidemiology. All African trypanosomes display antigenic variation and although the current *T. brucei*-based model might adequately describe the general phenomenon, this study shows that the genomic basis for antigenic variation has diverged among trypanosomes in a manner consistent with distinct mechanisms for generating antigenic variability. Consequently, we now have reason to expect substantial species differences beneath the general phenotype, a framework to dissect this variation, and so a basis for understanding how the enigmatic *VSG* connects with the wider disease.

Acknowledgements

We thank Phelix Majiwa for sample preparation and Sara Melville for advice and advocacy during project planning. Barbara Marchetti (University of Glasgow) prepared RNA for transcriptomic analysis. Richard Burchmore (University of Glasgow) performed 2D-electrophoresis. Douglas Lamont (University of Dundee) performed LC/MSMS analysis. Arnab Pain and Chris Newbold (WTSI) gave help and advice during manuscript preparation. This work was funded by the Wellcome Trust [grant number WT 085775/Z/08/Z]. We thank our colleagues in the sequencing and informatics groups at the Wellcome Trust Sanger Institute.

Figure legends

Figure 1. A sequence similarity network of *VSG*-like sequences from African trypanosome genomes, shown from 0º and 270º angles. A 3-D rendering of the network is provided as a supplementary video. The network represents pair-wise maximum likelihood protein sequences, generated in PHYLIP[28] using a WAG+Γ model[29] from multiple alignments of selected a-*VSG*-like (a-*VSG*, Fam23, *TFR*-like and *PAG*-like proteins; n = 174) and b-*VSG*-like (b-*VSG*, Fam13, Fam16, Fam24, *VR*, *ESAG2* and Fam1 proteins; n = 339) protein sequences, which are representative of global diversity. Spheres represent individual sequences shaded according to subfamily. The network was created with BioLayout Express 3D v2.0[30], which optimizes the placement of each sphere in three-dimensional space to minimize the size of the graph, such that highly related sequences cluster together. It was necessary to apply a lower threshold on pair-wise distances to reduce noise (i.e. weak connections between very distantly related sequences; see Methods). A dashed line separates a-*VSG*-like subfamilies (above) and b-*VSG* subfamilies (below). Four significant features identified in the text are labeled: i) sequence similarity between a- and b-lineages due to the shared CTD of *T. brucei VSG*; ii) the position of *ESAG2* nested within Fam13; iii) the position of Fam1, a *T. brucei*-specific b-*VSG*-like gene family; and iv) tight cluster of transferrin receptor-like genes from both *T. brucei* and *T. congolense*.

Figure 2. A model of *VSG* gene family evolution in African trypanosomes. This cartoon depicts the elaboration of *VSG* subfamilies in contemporary and ancestral genomes. Uncertain origins are indicated by dashed lines. An asterisk * indicates that a subfamily includes a proven variant antigen, although other variant antigens may occur in unmarked subfamilies. The presence of a-*VSG* and b-*VSG*-like structures in all three trypanosome species indicates that contemporary *VSG* are representatives of a- and b-lineages that were present in their common ancestor. Each species has modified this shared inheritance differently. *T. vivax* has additional subfamilies that may have been present in the ancestor, and subsequently lost by the *T. brucei/congolense* ancestor, or could represent *T. vivax*-specific developments. Close relationships between *T. brucei* and *T. congolense VSG*-like genes, for instance *ESAG2* and Fam13, shows that these lineages had already evolved in the *T. brucei/congolense* ancestor, and suggest that distinct functions have evolved in one or both daughter species. A red arrow indicates that the CTD is uniquely shared between a- and b-*VSG* in *T. brucei* and has been donated from one subfamily to the other in either direction.

Figure 3. Comparisons of phylogenetic tree topologies for *VSG*-like subfamilies. Bayesian phylogenies were estimated for six *VSG* subfamilies from *T. brucei* 927 (in blue, at left), *T. congolense* IL3000 (in green, centre) and *T. vivax* Y486 (in red, at right) with MrBayes 3.2.1.[31] using a WAG+Γ model. Default settings were applied, except for: Ngen=5000000, Nruns=4, samplefreq=500, burnin=1000-2500 (as required to achieve convergence). These trees contain all full-length protein sequences available (n) and include both intact genes and predicted pseudogenes. All trees are drawn to the same scale. The 'treeness' statistic (*T*) describes the proportion of tree length taken up for internal branches[20], and is a measure of the phylogenetic signal/noise ratio. Below each tree a histogram describes the distribution of pair-wise genetic distances (grouped into bins; x-axis) plotted against frequency (y-axis); mean average (μ) and standard deviation (σ) are provided.

Figure 4. Prevalence of significant phylogenetic incompatibility within *VSG*-like sequence alignments. Phylogenetic incompatibility (PI) describes the presence of multiple, conflicting phylogenetic signals within a single data set. Typically, PI is caused by recombination but can also result from heterogeneity in substitution rate or other molecular homoplasy[23]. Protein sequence alignments for six *VSG* subfamilies were examined for PI using the Pairwise Homoplasy Index (PHI[24]). Each alignment was randomly sampled 100 times and the proportion of samples displaying PI was counted (P_{pi}). A distribution for P_{pi} was generated by creating 100 bootstrapped alignments in each case (solid, coloured line).

To generate a null distribution, 100 alignments were simulated using the observed Bayesian phylogeny with a maximum likelihood substitution model (WAG+Γ) that corrected for rate heterogeneity but did not consider recombination (black lines). Finally, to demonstrate the effect of genetic distance on PI, the analysis was repeated on smaller alignments of closely related sequences taken from crown clades (dashed lines; see Methods). Mean average values, followed by standard deviations, are provided for observed (μ_{obs}), simulated (μ_{sim}) and within-clade sampling (μ_{within}) distributions.

Methods

Genome sequencing and annotation. *Trypanosoma congolense* IL3000 and *Trypanosoma vivax* Y486 were propagated as described previously[32-33]. High molecular weight DNA was extracted in late log phase by phenol-chloroform extraction and purified by gel electrophoresis. Genomic DNA was capillary sequenced using a whole genome shotgun strategy as described previously[6]. Sequence reads were assembled using Phrap (www.phrap.org). Automated in-house software (Auto-Prefinish) was used to identify primers and clones for additional sequencing to close gaps by oligo-walking and manual base checking. Repetitive regions or others with an unexpected read depth were manually inspected. The assembled contigs were iteratively ordered and orientated against the *T. brucei* 927 genome sequence[6] (TritrypDB Version 1.0) using ABACAS[34]. The manually curated genome annotation of *T. brucei* was transferred to the *T. congolense* and *T. vivax* assemblies using custom perl scripts, based on sequence and positional homology, and manually edited where appropriate using Artemis[35]. Ordering contigs against the *T. brucei* reference creates pseudo-chromosomes that suit comparative genomics, but these may be misleading if it enforces spurious similarity. *T. congolense* is the closest relative of *T. brucei* and both species have 11 megabase chromosomes[36]. However, *T. vivax* is more distantly related with an uncertain karyotype[37]. Therefore, in addition to producing pseudo-chromosomes, we manually assembled scaffolds from *T. vivax* contigs using read-pair information.

Annotation of *VSG* genes. *VSG* structures are highly mutable, and therefore annotation transfer and BLAST-based sequence homology with *T. brucei VSG* may not adequately annotate variant antigens in other species. Therefore, Hidden Markov Models (HMM) built using HMMER v3.0 (http://hmmer.janelia.org/) from *T. brucei* a- and b-*VSG* sequence alignments initially, and then native *T. congolense* and *T. vivax VSG*, were used to identify additional VSG candidates. This process increased the size of *T. congolense* and *T. vivax VSG* families by 10-37%. HMM searching also showed that many gene models were partial. Failure to annotate complete coding regions might under-estimated the frequency of pseudogenes, so the boundaries of all putative *VSG* open reading frames in *T. congolense* and *T. vivax* were manually checked against the HMM-defined boundaries to ensure that they began with a conserved signal peptide and terminated in a GPI anchor signal. Finally, each sequence was compared with relevant *VSG* sequence alignments to confirm completeness.

Data accessibility. Draft genome sequences have been submitted to EMBLBank: *T. congolense* accession numbers HE575314 to HE575324 and CAEQ01000352-CAEQ01002824; *T. vivax* accession numbers HE573017-HE573027 and CAEX01000001-CAEX01008277. The data can be examined via GeneDB (http://www.genedb.org) and TritrypDB (http://tritrypdb.org). *T. vivax* transcriptome data have been submitted to the European Bioinformatics Institute Array Express Archive (accession number E-MTAB-475). Sequence alignments and phylogenetic trees comprising the cell surface phylome are contained in GeneDB (http://www.genedb.org/Page/trypanosoma_surface_phylome).

Comparison of gene content. We used OrthoMCL[37-38] to examine species-specific genes and gene families, as well as conserved families with interspecific disparities in copy number. To check and expand on these putative gains and losses, we manually compared each *T. brucei* chromosome with *T. congolense* and *T. vivax* pseudo-chromosomes using the Artemis Comparison Tool (ACT[39]). Disruptions to co-linear gene order were identified but, since sequence gaps occasionally prevented a three-way comparison, we only considered disruptions that occurred within contigs (i.e. were not adjacent to gaps). The orthoMCL analysis shows that the principal differences in genomic complement concerned surface-expressed genes. To confirm that other areas of cell function were conserved, we manually inspected the locations of genes involved in the *T. brucei* flagellar proteome[40], intracellular transport[41], glycosyl transfer[42], ribosomal structure, phosphorylation, as well as a range of genes involved in metabolism. All putative losses were confirmed by examining expected genomic position and by searching unassembled sequence reads for reciprocal sequence matches by tBLASTn/BLASTx.

T. vivax **transcriptome**. *T. vivax* Y486 was grown from stabilate in BALB/c mice immunosuppressed with cyclophosphamide (250 mg.kg^{-1}) and was amplified at patent parasitaemia in three immunosuppressed mice, from which whole blood was collected. The blood was treated with the erythrocyte lysis buffer (EL; QIAGEN), following the manufacturer's instructions, and RNA was isolated from the pellet using the RNeasy mini kit protocol (QIAGEN).

Analysis of Fam1 gene expression. To determine mRNA expression levels of Fam1 family members quantitative real-time polymerase chain reaction (qRT-PCR) was carried out on total RNA extracted using RNeasy Mini Kit (QIAGEN). cDNA was generated using SuperScript II Reverse Transcriptase according to the manufacturer's instructions. qRT-PCR was carried out using three different isolated mRNA samples from four life-cycle stages [*in vitro* cultured bloodstream-stage and procyclic forms; *in vitro* cultured short stumpy bloodstream-stage; and *in vivo* cultured *T. brucei* bloodstream-stage]. *T. brucei* Rab11 was used as a control to determine relative quantity of mRNA. The relative abundance of specific RNA was subsequently determined.

Transfection and protein localization. A Fam1 gene (Tb927.6.1310) was synthesized by Eurogentec. *T. brucei* single marker bloodstream line cells were cultured in HMI-9 medium as described previously[43]. Ectopic expression of haemagglutinin (HA) epitope-tagged Tb927.6.1310 at the N-terminus (following the predicted signal peptide sequence) was carried out using pXS5/pDEX-577[44] constitutive and inducible expression vectors respectively. For protein extraction, proteins were transferred onto Immobilon polyvinyildene fluoride membrane and incubated with primary mouse anti-HA antibody (1:8,000) and subsequently with secondary rabbit anti-mouse peroxidase conjugate antibody (1:10,000, Sigma). Immunofluorescence microscopy was carried out on permeabilised and non-permeabilised transfected cells harvested at log phase.

VSG purification and sequencing. *T. vivax* Y486, grown from stabilate as described above, was injected into a mouse with intact immune system, inducing a relapsing parasitaemia. After 14 days, trypanosomes were purified from the blood by Percoll gradient fractionation, as described[33]. Trypanosomes were lysed in sample buffer and the extract was fractionated by 2D-electrophoresis according to the manufacturer's instructions (Amersham). Comparison of the day 14 with the initiating population, prepared in the same way, revealed significant differences in both dimensions in a ~40 kDa spot group, which is consistent with *VSG* switching. Both extracts were run in one-dimensional SDS-PAGE and three bands in the estimated size range were extracted from each, trypsinized and subjected to liquid chromatography/tandem mass spectrometry analysis. The major band in the day 14 population revealed Mascot hits with putative *VSG* contigs; the five other bands were 'housekeeping' proteins. For cDNA cloning, total RNA from purified *T. vivax* was primed with oligo[dT] and cDNA was generated using a primer specific to the 5' spliced leader[44] and an anchored oligo[dT] primer. A dominant ~1.3 kb band was gel extracted and was cloned into the TOPO plasmid (Invitrogen), and clone inserts were sequenced.

Cell-surface phylome. The African trypanosome cell surface phylome is a collection of phylogenies for gene families with predicted cell surface expression. All *T. brucei* genes with cell surface motifs, (i.e. a predicted signal peptide, a predicted GPI anchor or a transmembrane helix) were extracted. Genes annotated as 'unlikely' or <150 codons were removed. Homologs to each *T. brucei* 'surface' gene were identified among all *T. brucei, T. congolense, T. vivax* and *T. cruzi* predicted genes (the latter included as an outgroup) using wuBLAST. Where at least four homologs occurred in at least one species, this constituted a 'family' amenable to phylogenetic analysis. After removing genes already identified as homologous to *T. brucei* genes, this exercise was repeated for *T. congolense* and *T. vivax* genes, for which signal peptides were predicted using Signal P[46], GPI anchors were predicted using Fraganchor[47] and transmembrane helices were predicted using TMHMM[48]. A total of 291 'surface expressed' families was reduced to 81 by removing cases of poor alignment (i.e. spurious homology), obvious non-coding sequence (i.e. mis-annotation), and cases with fewer than four unique sequences (i.e. duplicated sequence), by combining families with overlapping homology, and by removing known mitochondrial and lysosomal genes or other families expressed in internal membranes.

Phylogenetic analysis. Amino acid sequences for each family were aligned in ClustalW[49]; all multiple alignments were then manually edited.. In most cases the amino acid sequence alignment was used, but nucleotide sequences were examined in cases of low sequence divergence. Bayesian phylogenies were estimated using MrBayes v3.2.1[31,50] (Nruns=2, Ngen=10000000, samplefreq=1000 and default prior distribution). Nucleotide sequence alignments were analyzed using a GTR+Γ model. Maximum likelihood phylogenies were estimated using PHYML v3.0[51] under an LG+Γ model[29] for amino acid sequences or a GTR+Γ model for nucleotide sequences. Node support was assessed using 100 non-parametric bootstrap replicates[52]. The trees were rooted using *T. cruzi* sequences, or otherwise mid-point rooted. Bayesian *VSG* phylogenies were estimated using alignments of selected, full-length sequences representative of global diversity (Nruns=1, Ngen=1000000, samplefreq=100 and default prior distribution). 'Treeness' was calculated for each tree topology using TreeStat v1.2 (http://tree.bio.ed.ac.uk/software/treestat/); this is defined as the proportion of total tree length taken up by internal branches and measures the noise to signal ratio in a phylogenetic data set[22].

Recombination analysis. Recombination results in sequence alignments with multiple phylogenetic signals[23], otherwise known as phylogenetic incompatibility (PI). The pair-wise homoplasy index (PHI[24]) returns a single probability value for PI and this was applied to amino acid sequence alignments for seven *VSG* sub-families (see Supplementary Table 8). For each alignment, 1000 sub-alignments of 10 sequences were prepared by selecting sequences at random. The proportion of sub-alignments with significant PI, termed P_{pi}, was compared between species. Confidence intervals on P_{pi} were obtained by repeating the analysis on 100 non-parametric bootstraps of each alignment, generated using SEQBOOT [http://evolution.genetics.washington.edu/phylip/doc/seqboot.html]. To confirm that significant PI was not due simply to rate heterogeneity or other forms of homoplasy, a null distribution for P_{pi} was obtained from simulated alignments generated with SEQGEN [http://tree.bio.ed.ac.uk/software/seqgen/], using maximum likelihood branch lengths and a WAG+Γ model that incorporated corrections for rate heterogeneity, but not recombination. To assess the effect of sequence identity on P_{pi} the analysis was repeated using alignments of sequences belonging to individual crown clades only as defined by *VSG* subfamily phylogenies; this is referred to as ‘intensive sampling’. To assess the effect of the CTD on P_{pi}, the analysis was repeated using *T. brucei* and *T. congolense* alignments with the CTD removed, (curtailed to the 3’-most universally conserved cysteine residue). This was not done for the *T. vivax* alignments since there is no obvious CTD.

References

1. Zambrano-Villa, S., Rosales-Borjas, D., Carrero, J.C. & Ortiz-Ortiz, L. How protozoan parasites evade the immune response. *Trends Parasitol*. **18**, 272-278 (2002).
2. Deitsch, K.W., Lukehart, S.A. & Stringer, J.R. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol*. **7**, 493-503 (2007).
3. Borst, P. & Cross, G.A. Molecular basis for trypanosome antigenic variation. *Cell* **29**, 291-303 (1982).
4. Pays, E. Regulation of antigen gene expression in *Trypanosoma brucei*. *Trends Parasitol*. **21**, 517-520 (2005).
5. Morrison, L.J., Marcello, L. & McCulloch, R. Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. *Cell Microbiol*. **11**, 1724-1734 (2009).
6. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H. *et al*. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416-422 (2005).
7. Hutchinson, O.C., Smith, W., Jones, N.G., Chattopadhyay, A., Welburn, S.C. *et al*. VSG structure: similar N -terminal domains can form functional VSGs with different types of C-terminal domain. *Mol Biochem Parasitol*. **130**, 127-131 (2003).
8. Taylor, J.E. & Rudenko, G. Switching trypanosome coats: what's in the wardrobe? *Trends Genet.* **22**, 614-620 (2006).
9. Pays, E., Salmon, D., Morrison, L.J., Marcello, L. & Barry, J.D. in: *Trypanosomes after the genome*, J.D. Barry, R. McCulloch, J. Mottram, A. Acosta-Serrano, Eds. (Horizon Bioscience, 2007), pp. 339-372.
10. Carrington, M., Miller, N., Blum, M., Roditi, I. & Wiley, D. Variant specific glycoprotein of *Trypanosoma brucei* consists of two domains each having an independently conserved pattern of cysteine residues. *J Mol Biol*. **221**, 823-835 (1991).
11. Blum, M.L., Down, J.A., Gurnett, A.M., Carrington, M., Turner, M.J. *et al*. A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature* **362**, 603-609 (1993).
12. Marcello, L. & Barry, J.D. Analysis of the *VSG* gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res*. **17**, 1344-1352 (2007).
13. Hobbs, M.R. & Boothroyd, J.C. An expression-site-associated gene family of trypanosomes is expressed *in vivo* and shows homology to a variant surface glycoprotein gene. *Mol Biochem Parasitol*. **43,** 1-16 (1990).
14. Schell, D., Evers, R., Preis, D., Ziegelbauer, K., Kiefer, H., *et al*. A transferrin-binding protein of *Trypanosoma brucei* is encoded by one of the genes in the variant surface glycoprotein gene expression site. *EMBO J*. **10**, 1061-1066 (1991).
15. Salmon, D., Geuskens, M., Hanocq, F., Hanocq-Quertier, J., Nolan, D. *et al*. A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*. *Cell* **78**, 75-86 (1994).
16. Adams, E.R., Hamilton, P.B. & Gibson, W.C. African trypanosomes: celebrating diversity. *Trends Parasitol* **26**: 324-328 (2010).

17. Gardiner, P.R., Nene, V., Barry, M.M., Thatthi, R., Burleigh, B. *et al*. Characterization of a small variable surface glycoprotein from *Trypanosoma vivax*. *Mol Biochem Parasitol*. **82**, 1-11 (1996).
18. de Greef, C. & Hamers, R. The *serum resistance-associated* (*SRA*) gene of *Trypanosoma brucei rhodesiense* encodes a variant surface glycoprotein-like protein. *Mol Biochem Parasitol*. **68**, 277–284 (1994).
19. Berberof, M., Pérez-Morga, D. & Pays, E.A. A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense*. *Mol. Biochem. Parasitol*. **113**, 127-138 (2001).
20. Carrington, M. & Boothroyd, J.C. Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol Biochem Parasitol*. **81**, 119-126 (1996).
21. Borst, P. & Fairlamb, A.H. Surface receptors and transporters of *Trypanosoma brucei*. *Annual Rev Microbiol.* **52**, 745-778 (1998).
22. White, W.T., Hills, S.F., Gaddam, R., Holland, B.R. & Penny, D. Treeness triangles: visualizing the loss of phylogenetic signal. *Mol Biol Evol*. **24**, 2029-2039 (2007).
23. Weiller, G.F. Detecting genetic recombination. *Methods Mol Biol*. **452**, 471-483 (2008).
24. Bruen, T.C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665-2681 (2006).
25. Deng, C.X. & Capecchi, M.R. Reexamination of gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. *Mol Cell Biol* **12**: 3365-3371 (1992).
26. Bell J. S. & McCulloch, R. Mismatch repair regulates homologous recombination, but has little influence on antigenic variation, in *Trypanosoma brucei*. *J Biol Chem* **278**: 45182-45188.
27. Schwede, A., Jones, N., Engstler, M., & Carrington, M. The VSG C-terminal domain is inaccessible to antibodies on live trypanosomes. Mol Biochem Parasitol **175**: 201-204 (2011).
28. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2005).
29. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol*. **25**: 1307-1320 (2008).
30. Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S. & Mazière, P. *et al*. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol.* **3**, 2032-2042 (2007).
31. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754-755 (2001).
32. Hirumi, H. & Hirumi, K. *In vitro* cultivation of *Trypanosoma congolense* bloodstream forms in the absence of feeder cell layers. *Parasitology* **102**, S225–236 (1991).
33. Ndao, M., Magnus, E., Buscher, P., & Geerts, S. *Trypanosoma vivax*: a simplified protocol for *in vivo* growth, isolation and cryopreservation. *Parasite* **11**, 103 (2004).
34. Cutler, D.J., Zwick, M.E., Carrasquillo, M.M., Yohn, C.T., & Tobin, K.P. *et al*. High-Throughput Variation Detection and Genotyping Using Microarrays. *Genome Res* **11**, 1913-1925 (2001).
35. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T. & Rice, P. *et al*. Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945 (2000).

36. Van der Ploegi, L.H.T., Cornelissen, A.W.C.A., Barry, J.D. & Borst, P. Chromosomes of Kinetoplastida. *EMBO Journal* **3**, 3109-3115 (1984).
37. Li, L., Stoeckert Jr, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. **13**, 2178-2189 (2003).
38. Chen, F., Mackey, A.J., Stoeckert Jr, C.J. & Roos, D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*. **34**, D363-D368 (2006).
39. Carver, T., Berriman, M., Tivey, A., Patel, C. & Böhme, U. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672-2676 (2008).
40. Broadhead, R., Dawe, H.R., Farr, H., Griffiths, S. & Hart, S.R. *et al*. Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature* **440**, 224-227 (2006).
41. Koumandou, V.L., Natesan, S.K., Sergeenko, T. & Field, M.C. The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. *BMC Genomics* **9**, 298 (2008).
42. Izquierdo, L., Nakanishi, M., Mehlert, A., Machray, G., Barton, G.J. & Ferguson, M.A. Identification of a glycosylphosphatidylinositol anchor-modifying beta1-3 N-acetylglucosaminyl transferase in *Trypanosoma brucei*. *Mol Microbiol*. **71**, 478-491 (2009).
43. Wirtz, E., Leal, S., Ochatt, C. & Cross, G.A. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol Biochem Parasitol*. **99**, 89-101 (1999).
44. Kelly, S., Reed, J., Kramer, S., Ellis, L. & Webb, H. *et al*. Functional genomics in *Trypanosoma brucei*: a collection of vectors for the expression of tagged proteins from endogenous and ectopic gene loci. *Mol Biochem Parasitol*. **154**, 103-109 (2007).
45. De Lange, T., Berkvens, T.M., Veerman, H.J., Frasch, A.C., Barry J.D. *et al*. Comparison of the genes coding for the common 5′ terminal sequence of messenger RNAs in three trypanosome species. *Nucl Acids Res*. **12**, 4431-4443 (1984).
46. Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat Protocols* **2**, 953-971 (2007).
47. Poisson, G., Chauve, C., Chen, X. & Bergeron, A. FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. *Genomics Proteomics Bioinformatics* **5**, 121-130 (2007).
48. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L.L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol. Biol*. **305**, 567-580 (2001).
49. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R. & McGettigan, P.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
50. Ronquist , F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
51. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* **52**, 696-704 (2003).
52. Felsenstein, J. Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution* **39**, 783-791 (1985).

Figure 1.

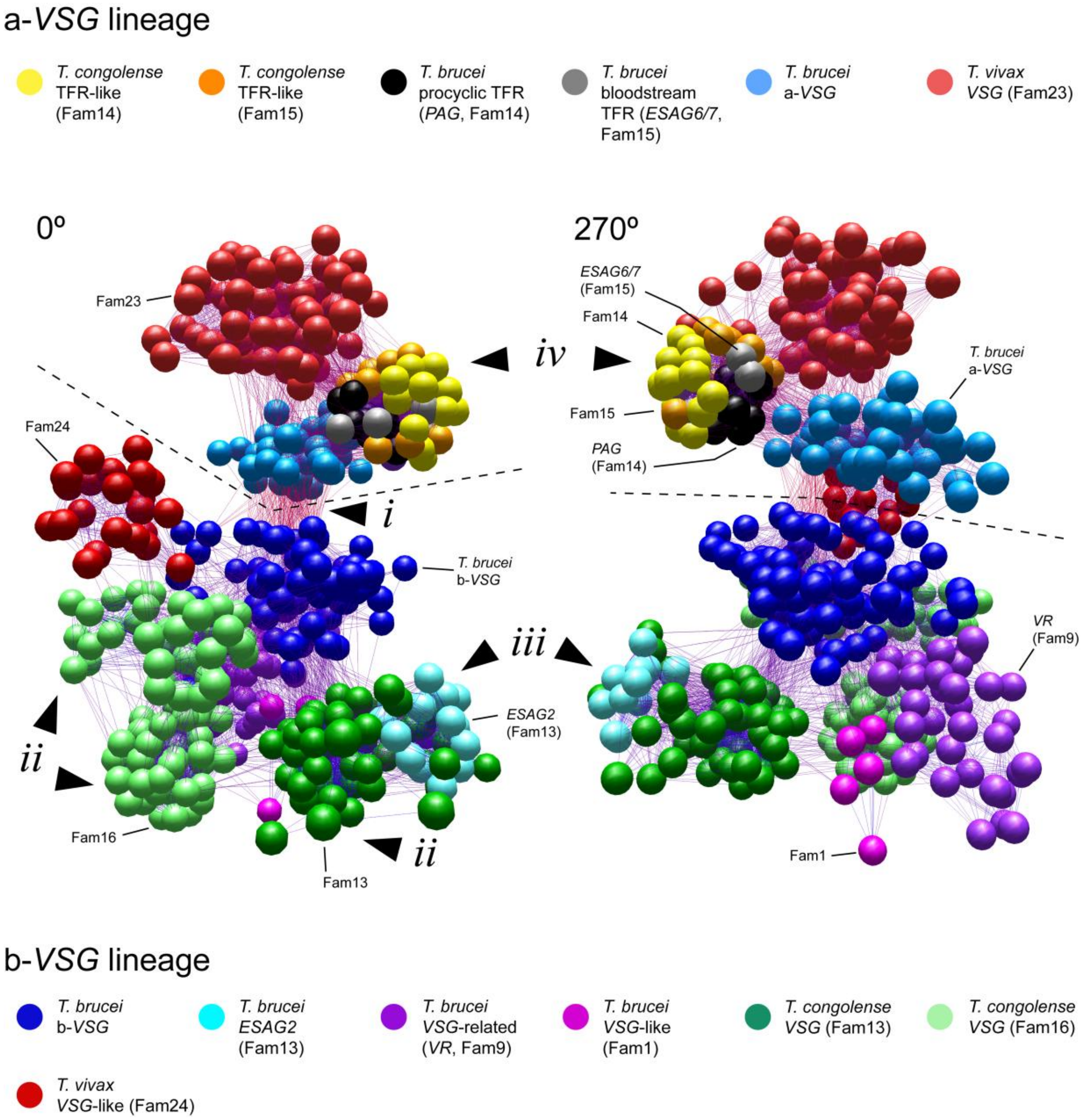

a-VSG lineage
T. congolense TFR-like (Fam14)
T. congolense TFR-like (Fam15)
T. brucei procyclic TFR (PAG, Fam14)
T. brucei bloodstream TFR (ESAG6/7, Fam15)
T. brucei a-VSG
T. vivax VSG (Fam23)
0º
270º
Fam23
Fam24
iv
i
ii
iii
T. brucei b-VSG
ESAG2 (Fam13)
Fam16
Fam13
ESAG6/7 (Fam15)
Fam14
Fam15
PAG (Fam14)
T. brucei a-VSG
VR (Fam9)
Fam1
b-VSG lineage
T. brucei b-VSG
T. brucei ESAG2 (Fam13)
T. brucei VSG-related (VR, Fam9)
T. brucei VSG-like (Fam1)
T. congolense VSG (Fam13)
T. congolense VSG (Fam16)
T. vivax VSG-like (Fam24)

Figure 2.

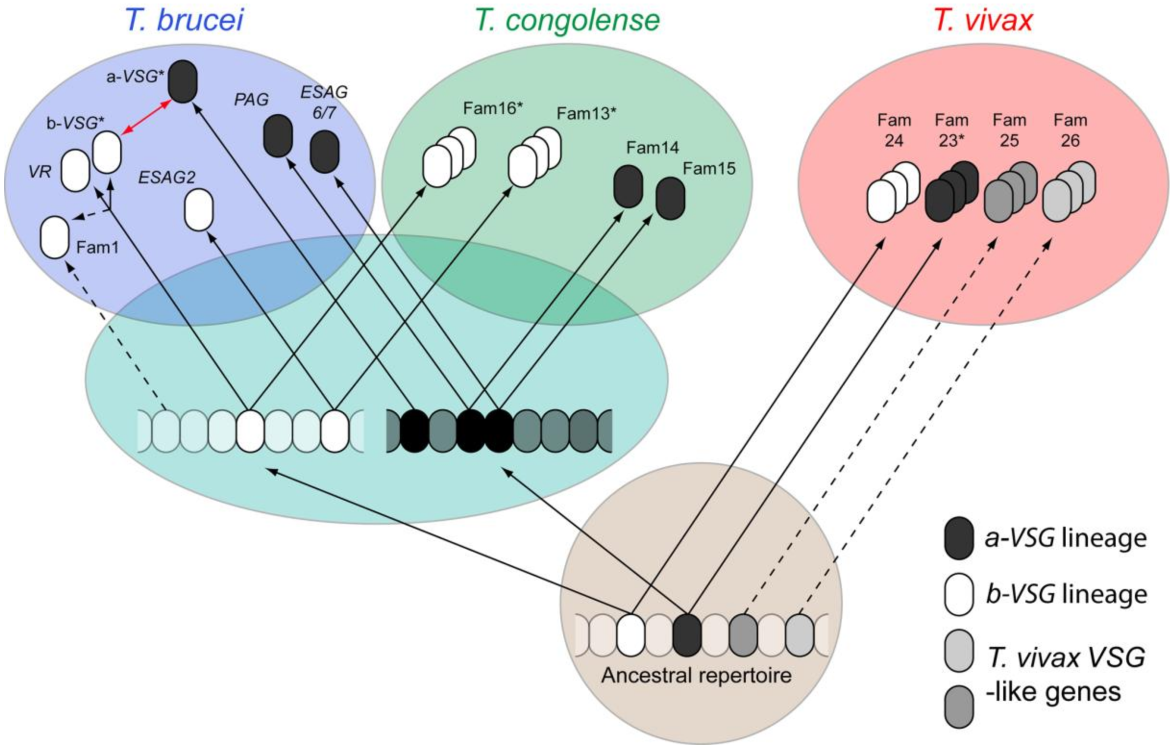

T. brucei
T. congolense
T. vivax
a-VSG*
b-VSG*
PAG
ESAG 6/7
VR
ESAG2
Fam1
Fam16*
Fam13*
Fam14
Fam15
Fam 24
Fam 23*
Fam 25
Fam 26
Ancestral repertoire
a-VSG lineage
b-VSG lineage
T. vivax VSG -like genes

Figure 3.

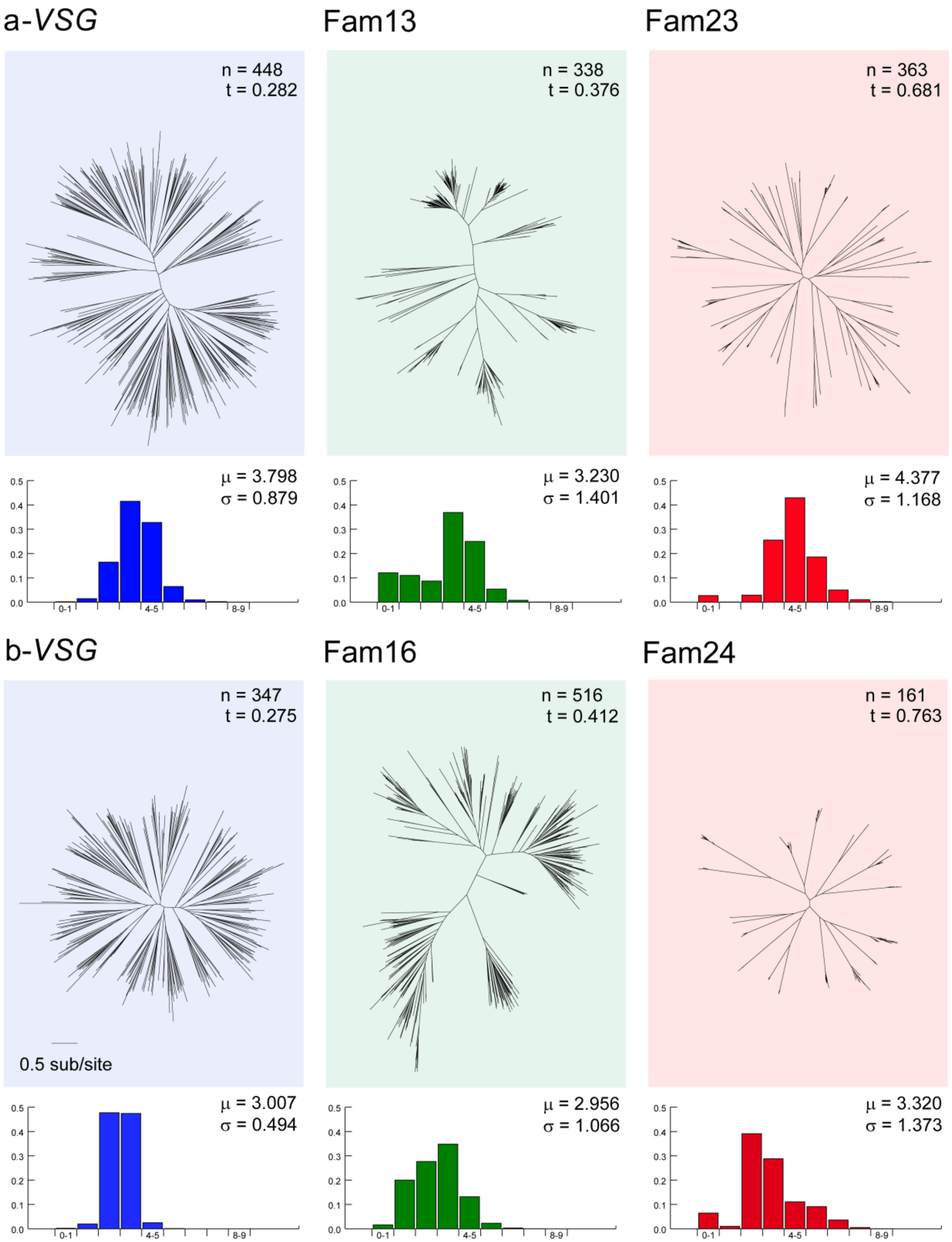
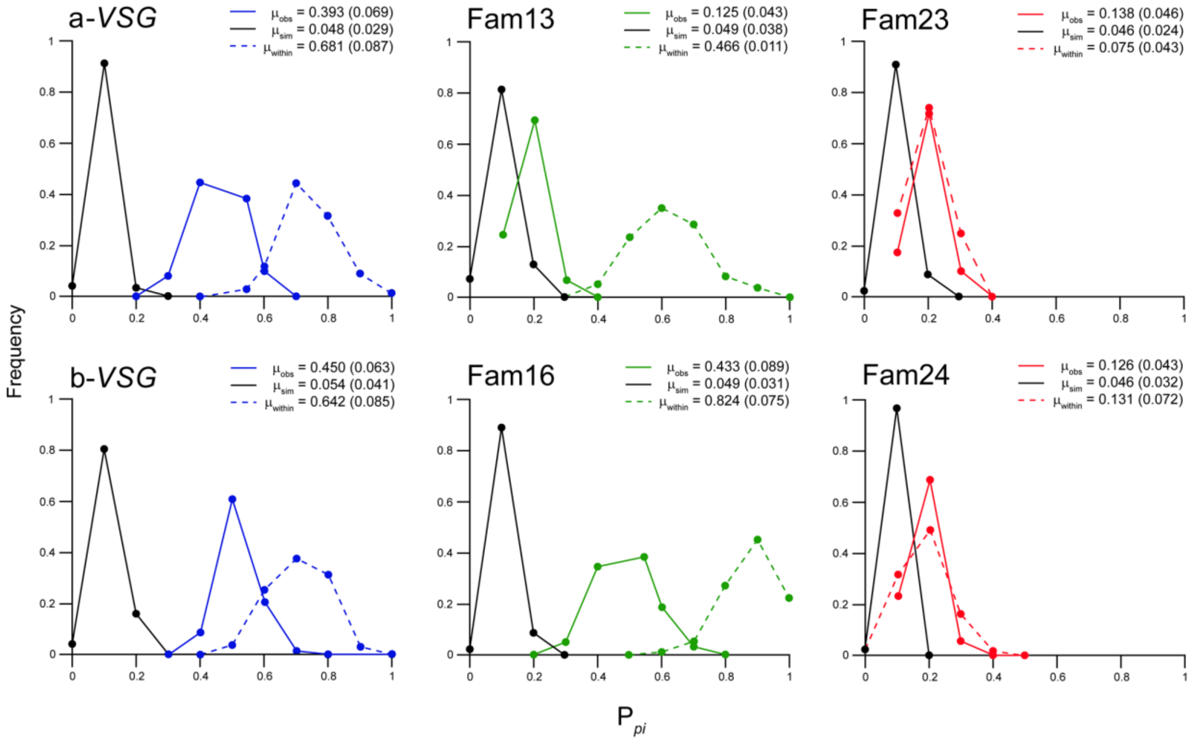

a-*VSG*
n = 448
t = 0.282
μ = 3.798
σ = 0.879
Fam13
n = 338
t = 0.376
μ = 3.230
σ = 1.401
Fam23
n = 363
t = 0.681
μ = 4.377
σ = 1.168
b-*VSG*
n = 347
t = 0.275
0.5 sub/site
μ = 3.007
σ = 0.494
Fam16
n = 516
t = 0.412
μ = 2.956
σ = 1.066
Fam24
n = 161
t = 0.763
μ = 3.320
σ = 1.373
0.5
0.4
0.3
0.2
0.1
0.0
0-1
4-5
8-9

Figure 4.


a-*VSG*
μobs = 0.393 (0.069)
μsim = 0.048 (0.029)
μwithin = 0.681 (0.087)
Fam13
μobs = 0.125 (0.043)
μsim = 0.049 (0.038)
μwithin = 0.466 (0.011)
Fam23
μobs = 0.138 (0.046)
μsim = 0.046 (0.024)
μwithin = 0.075 (0.043)
b-*VSG*
μobs = 0.450 (0.063)
μsim = 0.054 (0.041)
μwithin = 0.642 (0.085)
Fam16
μobs = 0.433 (0.089)
μsim = 0.049 (0.031)
μwithin = 0.824 (0.075)
Fam24
μobs = 0.126 (0.043)
μsim = 0.046 (0.032)
μwithin = 0.131 (0.072)
Frequency
0
0.2
0.4
0.6
0.8
1
P_{pi}