Biological Sciences (Evolution)

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

Andrew P. Jackson1\*, Andrew Berry1, Martin Aslett1, Harriet C. Allison2, Peter Burton3, Jana Vavrova-Anderson3, Robert Brown4, Hilary Browne1, Nicola Corton1, Heidi Hauser1, John Gamble1, Ruth Gilderthorp1, Lucio Marcello3, Jacqueline McQuillan1, Thomas D. Otto1, Michael A. Quail1, Mandy Sanders1, Andries Van Tonder1, Michael L. Ginger4, Mark C. Field2, J. David Barry3, Christiane Hertz-Fowler1,5, Matthew Berriman1

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](mailto: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* causes human African Trypanosomiasis (‘sleeping sickness’) across sub-Saharan Africa and is a model system for antigenic variation, surviving by periodically replacing a monolayer of variant surface glycoproteins (VSG) that cover 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 has evolved, and how this might affect contemporary antigenic diversity. For the first time, we reconstruct *VSG* diversification showing that *T. congolense* employs variant antigens derived from multiple ancestral *VSG* lineages, while in *T. brucei* *VSG* have recent origins, and ancestral gene lineages have been repeatedly co-opted to novel functions. These historical differences are reflected in fundamental differences between species in the scale and mechanism of recombination. Using phylogenetic incompatibility as a metric for genetic exchange, we show that the frequency of recombination is comparable between *T. congolense* and *T. brucei*, but much lower in *T. vivax*. Furthermore, in showing that the C-terminal domain of *T. brucei VSG* plays a crucial role in facilitating exchange, we reveal substantial species differences in the mechanism of *VSG* diversification. Our results demonstrate how past *VSG* evolution indirectly determines the ability of contemporary parasites to generate novel variant antigens through recombination, and suggest that the current model for antigenic variation in *T. brucei* is only one means by which these parasites maintain chronic infections.

**Introduction**

Antigenic variation enables pathogens to evade immune responses by continual switching of surface proteins1,2. The African trypanosomes (*Trypanosoma* spp*.*)are vector-borne, protozoan blood parasites that survive in their hosts by antigenic variation, periodically replacing a monolayer of variant surface glycoproteins (VSG3) that shield the cell surface from immune effectors4-5. *Trypanosoma brucei* is the cause of Human African Trypanosomiasis (or ‘sleeping sickness’) and the mechanisms for expression and dynamic replacement of VSG in this species are a model system for antigenic variation4, as well as a classic example of adaptive evolution at the host-pathogen interface. Two related veterinary parasites, *T. congolense* and *T. vivax*, also employ antigenic variation to cause devastating diseases in domesticated animals. Through their detrimental effects on livestock productivity, these species arguably represent greater threats to socio-economic well-being than *T. brucei* in the agrarian societies in which they are endemic. Our understanding of how antigenic diversity is organized in *T. brucei* was greatly improved by the *T. brucei* 927 reference genome sequence6. In this study we present draft genome sequences for *T. congolense* and *T. vivax*;we define their global *VSG* repertoires in a three-way comparative analysis with *T. brucei*, revealing how antigenic diversity evolved in trypanosome genomes past and present.

The *T. brucei* genome includes many hundreds of *VSG* that comprise a transcriptionally-silent reservoir of variant antigens6 and each cell expresses just a single gene from a specialized telomeric expression site at any time4-5. 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 shuffling7 and gene conversion among silent, subtelomeric gene copies or possibly *in situ* within the expression site8-9. 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 related10-12. While *VSG* are known to occur in *T. congolense* and *T. vivax*13-17, the repertoire of variant antigens in these species is uncharacterized. Consequently, the evolutionary diversification of the *VSG* gene family has not been examined, although it has been suggested that *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*18-20 and b-*VSG*9,12 respectively.

The antigenic variation phenotype is observed in all African trypanosomes and it is assumed that this reflects a common physiological model, which has been defined in *T. brucei*. The aim of this study is to identify the evolutionary processes that have created contemporary *VSG* diversity, and reveal any significant differences in how trypanosome species generate variant antigens. Despite their shared phenotype, our results show that species differ in the organization of antigenic diversity at the genome level, and they provide a basis to better understand disease progression, pathology and host range in all African trypanosomes.

**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 lineage21. These genome sequences are described in supplementary information (see Table S1) 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 Table S2-4). 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 Table S5). 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 encompasses both published *T. congolense VSG* 13-15 and *VSG* expressed sequence tags (EST)16. While there is no a-*VSG* variant antigen, there are homologs of the a-*VSG*-like transferrin receptor genes of *T. brucei*, i.e. *ESAG6/7* (Fam15, n = 43), as well as the Procyclin-Associated Genes (*PAG*) (Fam14, n = 22), which also have an a-*VSG*-like structure22.

*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 Fig. S1). These may have evolved in *T. vivax*, or may represent ancestral lineages not inherited by *T. brucei* and *T. congolense.* Transcriptomic data show that multiple members of all four families are transcribed in bloodstream-stage cells (see Table S6). We find no orthologs to the transferrin receptor-like genes of *T. brucei* and *T. congolense* among *T. vivax VSG*-like genes or indeed thenumerous, 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 2.117) 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 Fig. S2). 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 Table S7). We estimated Bayesian and Maximum Likelihood phylogenies from amino acid sequence alignments (see Figs. S3-4) 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 (the supplementary video contains an animation of the three-dimensional network); 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 Fig. 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 presents a interesting anomaly: 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*23) and *TgsGP*24 proteins in *T. b. rhodesiense* and *T. b. gambiense* respectively. However, these represent secondary loss of function in contemporary *VSG*. Fig. 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 Fig. S5). Phylogenetic analysis clearly demonstrates that both *ESAG2* and *VR* gene subfamilies, for which the evidence is against a variant antigen function9,12, 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 Fig. S4). 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 previously9,25-26. 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-binding15 are conserved in both species (see Fig. S6). 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. Fig. 3 demonstrates how these treeshave distinct topologies. This is due to variation in the ratio of internal to terminal branches, (described by ‘treeness’27, *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 recorded7,12. Recombination is also the mechanism through which *VSG* are transposed from subtelomeric loci into the telomeric expression site4,5,8,9. *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 result from, at least partly, gene conversion between *VSG* genes5, 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)28-29, taking random samples of each alignment set and observing the proportion showing significant PI (*Ppi*.; see Table S8). Fig. 4 shows that *Ppi* (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 *Ppi* 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 identity30-31 and when we increased sequence identity within alignments by sampling only within crown clades, *Ppi* 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 *Ppi* for *T. brucei* a-*VSG* (0.152, p < 0.0001) and b-*VSG* (0.234, p < 0.0001), but in *T. congolense* *Ppi* 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 *Ppi* 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 Fig. 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 *T. brucei* infections are an additional source of diversity among expressed *VSG*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*. Opportunities for allelic recombination among *VSG* may be affected by species differences in mating system; recent work indicates the sexual reproduction is frequent in *T. brucei* and *T. congolense* populations32-34, while *T. vivax* population structure is consistent with clonal reproduction35. However, the importance of allelic recombination to *VSG* diversity is debatable; since the subtelomeres of homologous chromosomes in *T. brucei* are frequently dissimilar, subtelomeric *VSG* are effectively hemizygous and so most recombination between *VSG* must be ectopic. This study shows that 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. This emphasizes the role of gene structure in promoting ectopic recombination, rather than the frequency of allelic recombination during meiosis.

Differences in the role of the CTD indicate that, in addition to scale, the mechanism of recombination varies between species. The CTD is exchanged between *T. brucei VSG*, but is not exposed to antibodies and therefore, may not directly contribute to antigenic diversity36. However, the CTD-encoding sequence may have a role in the duplicative transposition of *VSG*, which promotes the diversification of the silent archive12, and is of paramount importance to antigenic variationbecause it is required to move *VSG* from silent, subtelomeric loci into the telomeric expression site4,8,9. It is thought that *VSG* transposition is facilitated by an upstream 70bp repeat region and the conserved CTD9, with the transposed region terminating around the 3’ end of the coding sequence37. Hence, in 37 recent such duplicative events in the *T. brucei* 927 genome12, 16 terminated within the CTD region. In showing that the majority of PI in *T. brucei VSG* alignments concerns the CTD, our data demonstrate that a major recombination breakpoint occurs between the N- and C-terminal domains, supporting a principal role for the CTD in promoting *VSG* switching. 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 among *T. brucei* *VSG* 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* mayhave evolved a distinct mechanism for the *VSG* transpositions that are key for diversification and antigen switching.

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.

**Materials and methods**

**Genome sequencing and annotation.** *Trypanosoma congolense* IL3000 and *Trypanosoma vivax* Y486 genomes were capillary sequenced using a whole genome shotgun strategy as described previously6.

**Annotation of *VSG* genes.***T. congolense* and *T. vivax* *VSG* were identified by BLASTp-based homology searches and Hidden Markov Models. The boundaries of all *VSG* open reading frames were manually checked against global sequence alignments.

**Comparison of gene content.**OrthoMCL38 was used to examine putative gene gains and losses. 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 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 and RNA was isolated from the pellet.

**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.

**Transfection and Fam1 protein localization**.Ectopic expression of haemagglutinin epitope-tagged Tb927.6.1310 at the N-terminus (following the predicted signal peptide sequence) was carried out using constitutive and inducible expression vectors.

**VSG purification and sequencing***. T. vivax* cell 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.

**Cell-surface phylome**.Homologs to each *T. brucei* ‘surface’ gene were identified among all *T. brucei, T. congolense, T. vivax* and *T. cruzi* predicted genes using wuBLAST. Gene family phylogenies were estimated using Maximum Likelihood and Bayesian methods.

**Recombination analysis**.Phylogeneticincompatibility (PI) within VSG sequence alignments was used as a measure of recombination. The pair-wise homoplasy index (PHI29) returns a single probability value for PI and this was applied to amino acid sequence alignments for seven *VSG* sub-families (see Table S8), sub-sampled 1000 times. The proportion of sub-alignments with significant PI, termed *Ppi*, was compared between species.

**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.

**References**

1. Zambrano-Villa S, Rosales-Borjas D, Carrero JC, Ortiz-Ortiz L (2002) How protozoan parasites evade the immune response. *Trends Parasitol* **18**, 272-278.
2. Deitsch KW, Lukehart SA, Stringer JR (2007) Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol* **7**, 493-503.
3. Borst P, Cross GA (1982) Molecular basis for trypanosome antigenic variation. *Cell* **29**, 291-303.
4. Pays E (2005) Regulation of antigen gene expression in *Trypanosoma brucei*. *Trends Parasitol* **21**, 517-520.
5. Morrison LJ, Marcello L, McCulloch R (2009) Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. *Cell Microbiol* **11**, 1724-1734.
6. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H *et al*. (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416-422.
7. Hutchinson OC, Smith W, Jones NG, Chattopadhyay A, Welburn SC *et al*. (2003) VSG structure: similar N -terminal domains can form functional VSGs with different types of C-terminal domain. *Mol Biochem Parasitol* **130**, 127-131.
8. Taylor JE, Rudenko G (2006) Switching trypanosome coats: what's in the wardrobe? *Trends Genet* **22**, 614-620.
9. Pays E, Salmon D, Morrison LJ, Marcello L, Barry JD in *Trypanosomes after the genome*,eds Barry JD, McCulloch R, Mottram J, Acosta-Serrano A (Horizon Bioscience, 2007), pp. 339-372.
10. Carrington M, Miller N, Blum M, Roditi I, Wiley D (1991) 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.
11. Blum ML, Down JA, Gurnett AM, Carrington M, Turner MJ *et al*. (1993) A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature* **362**, 603-609.
12. Marcello L, Barry JD (2007) 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.
13. Strickler JE, Binder DA, L’Italien JJ, Shimamoto GT, Wait SW *et al*. (1987) *Trypanosoma congolense*: structure and molecular organization of the surface glycoproteins of two early bloodstream variants. *Biochemistry* **26:** 796-805.
14. Rausch S, Shayan P, Salnikoff J, Reinwald E (1994) Sequence determination of three variable surface glycoproteins from *Trypanosoma congolense*. Conserved sequence and structural motifs. *Eur J Biochem* **223:** 813-821.
15. Eshita Y, Urukawa T, Hirumi H, Fish WR, Majiwa PA (1992) Metacyclic form-specific variable surface glycoprotein-encoding genes of *Trypanosoma* (*Nannomonas*) *congolense*. *Gene* **113:** 139-148.
16. Helm JR, Hertz-Fowler C, Aslett M, Berriman M, Sanders M *et al.* (2009) Analysis of expressed sequence tags from the four main developmental stages of *Trypanosoma congolense*. *Mol Biochem Parasitol* **168**: 34-42.
17. Gardiner PR, Nene V, Barry MM, Thatthi R, Burleigh B *et al*. (1996) Characterization of a small variable surface glycoprotein from *Trypanosoma vivax*. *Mol Biochem Parasitol*. **82**, 1-11.
18. Hobbs MR, Boothroyd JC (1990) 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.
19. Schell D, Evers R, Preis D, Ziegelbauer K, Kiefer H *et al*. (1991) 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.
20. Salmon D, Geuskens M, Hanocq F, Hanocq-Quertier J, Nolan D *et al*. (1994) A novel heterodimeric transferrin receptor encoded by a pair of *VSG* expression site-associated genes in *T. brucei*. *Cell* **78**, 75-86.
21. Adams ER, Hamilton PB & Gibson WC (2010) African trypanosomes: celebrating diversity. *Trends Parasitol* **26**: 324-328.
22. Koenig-Martin E, Yamage M, Roditi I (1992) A procyclin-associated gene in *Trypanosoma brucei* encodes a polypeptide related to ESAG 6 and 7 proteins. *Mol Biochem Parasitol*. **55**: 135-145.
23. de Greef C, Hamers R (1994) The *serum resistance-associated* (*SRA*) gene of *Trypanosoma brucei* *rhodesiense* encodes a variant surface glycoprotein-like protein. *Mol Biochem Parasitol* **68**, 277–284.
24. Berberof M, Pérez-Morga D, Pays EA (2001) A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense*. *Mol. Biochem. Parasitol* **113**, 127-138.
25. Carrington M, Boothroyd JC (1996) Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol Biochem Parasitol* **81**, 119-126.
26. Borst P, Fairlamb AH (1998) Surface receptors and transporters of *Trypanosoma brucei*. *Annual Rev Microbiol* **52**, 745-778.
27. White WT, Hills SF, Gaddam R, Holland BR, Penny D (2007) Treeness triangles: visualizing the loss of phylogenetic signal. *Mol Biol Evol* **24**, 2029-2039.
28. Weiller GF (2008) Detecting genetic recombination. *Methods Mol Biol* **452**, 471-483.
29. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665-2681.
30. Deng CX, Capecchi MR (1992) Re-examination 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.
31. Bell JS, McCulloch R (2003) Mismatch repair regulates homologous recombination, but has little influence on antigenic variation, in *Trypanosoma brucei*. *J Biol Chem* **278**: 45182-45188.
32. Morrison LJ, Tweedie A, Black A, Pinchbeck GL, Christley RM, *et al*. (2009) Discovery of mating in the major African livestock pathogen *Trypanosoma congolense*. *PLoS ONE* **4**: e5564. doi:10.1371/journal.pone.0005564B.
33. Peacock L, Ferris V, Bailey M, Gibson W (2009) Intraclonal mating occurs during tsetse transmission of *Trypanosoma brucei*. *Parasites and Vectors* **2**: 43. doi:10.1186/1756-3305-2-43
34. Holzmuller P, Herder S, Cuny G, De Meeus T (2010) From clonal to sexual: a step in *T. congolense* evolution? *Parasitology Today* **26**: doi:10.1016/j.pt.2009.11.006.
35. Duffy CW, Morrison LJ, Black A, Pinchbeck GL, Christley RM, et al. (2009) *Trypanosoma vivax* displays a clonal population structure. *International Journal for Parasitology* **39**: 1475–1483.
36. Schwede A, Jones N, Engstler M, Carrington M (2011) The VSG C-terminal domain is inaccessible to antibodies on live trypanosomes. *Mol Biochem Parasitol* **175**: 201-204.
37. Michels PA, Liu AY, Bernards A, Sloof P, Van der Bijl MM, *et al*. (1983) Activation of the genes for variant surface glycoproteins 117 and 118 in *Trypanosoma brucei*. *J Mol Biol* **166**: 537-556.
38. Li L, Stoeckert Jr CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.
39. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.

**Figure legends**

**Fig. 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 distances, generated 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 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) diverse clusters of *T. congolense* b-*VSG* belonging to Fam13 and Fam16; iii) the position of *ESAG2* nested within Fam13; and iv) tight cluster of transferrin receptor-like genes from both *T. brucei* and *T. congolense*.

**Fig. 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.

**Fig. 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 3.2.1.39 using a WAG+ 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 branches27, 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 () and standard deviation () are provided.

**Fig. 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 homoplasy28. Protein sequence alignments for six *VSG* subfamilies were examined for PI using the Pairwise Homoplasy Index (PHI29). Each alignment was randomly sampled 100 times and the proportion of samples displaying PI was counted (*Ppi*). A distribution for *Ppi* 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+) 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 (obs), simulated (sim) and within-clade sampling (within) distributions.