Supporting Information

Jackson et al. 10.1073/pnas.1117313109

SI Results

Draft Genome Sequences for *Trypanosoma congolense* IL3000 and *Trypanosoma vivax* Y486. Summary statistics for the genome sequences are shown in Table S1; these demonstrate all three species are similar with respect to properties such as GC content and gene length, although the *T. vivax* genome may be less gene-dense and, consequently, have more noncoding material. However, it is important to note that these genomes are drafts; there are numerous gaps, and annotation is only preliminary. Therefore, the number, placement, and order of coding regions are prone to change. Furthermore, mapping sequence contigs on to the *Trypanosoma brucei* 927 reference genome showed core chromosomal regions mapped fairly easily, producing 11 pseudochromosomes for both species. Subtelomeric and strand-switch regions did not map onto the reference genome, suggesting that these regions are more dynamic and likely to contain species-specific genes, as has been observed in trypanosomatids previously (1). It is notable that most species-specific genes currently identified are found on unassembled contigs.

Using the *T. brucei* genome as a guide for assembly may conceal evidence of karyotypic differences between species. So we took all *T. vivax* contigs (assembled de novo) that mapped to core chromosomal regions in *T. brucei* and assembled these into scaffolds based on read-pair information, independent of *T. brucei*. A total of 1,828 contigs were assembled into 280 scaffolds with an average length of 655 kb, which showed the genuine gene order in *T. vivax*. A total of 413 putative transposition events were identified from instances where a *T. vivax* scaffold contained genes with orthologs on multiple *T. brucei* chromosomes. These concerned on average 1.16 genes, i.e., most disruptions to colinearity involved a single gene. Because the *T. vivax* genome did not assemble further, we cannot say more about its karyotype, and it remains possible that major genomic rearrangements could have happened. Although the data do not allow these 280 *T. vivax* scaffolds to be arranged further, into 11 chromosomes or otherwise, it is true nonetheless that, when considering *T. vivax* contigs assembled independently of *T. brucei*, rearrangements to microsynteny are few and relatively small; therefore, gene order is well conserved in all three African trypanosomes.

Comparison of Genome Content. Within core chromosomal regions (i.e., not subtelomeres) and where fully contiguous sequences were available for all three species, the African trypanosome genomes are widely conserved in both gene order and content. Despite the gaps that remain in the *T. congolense* and *T. vivax* drafts, we believe that this observation will not change with improvement of the assemblies. Fully 97.5% of genes within chromosomal cores were positionally conserved; that is, for a given locus, the genes flanking it were orthologous in all species. After discounting variant surface glycoproteins (*VSG*), 72.1% of all genes have single orthologs in all species. Most other genes are conserved but do not display such simple orthology because they are members of multigene families, for example, there is variation in the number of amino acid transporter genes but not the phylogenetic diversity of this family. The genome sequences were essentially very similar; breaks in chromosomal colinearity and gene losses identified by manually comparing core regions are listed in Table S2. In 90% of cases where genes were not present, no reciprocal top hit was found in unassembled sequence reads, confirming the absence. This essential similarity comes across in Table S3, which lists the differences in gene content with respect to particular cell functions. For 86% of these apparent deletions, no reciprocal top hit was found in unassembled sequence reads, indicating that the absence was genuine. Thus, the three genomes contain largely the same capacity with regard to metabolism (only 11 genes present in *T. brucei* and absent in *T. vivax* and 5 genes present in *T. vivax* and absent in *T. brucei*) as well as motility (of 326 proteins associated with the flagellum, 11 are absent from *T. vivax*) and similarly for intracellular transport, protein phosphorylation, and glycosylation.

It follows that core chromosomal regions contained few species-specific features. Hence, for core regions, we estimate that only ~4% of genes are species-specific. This is a difficult quantity to determine from genome sequences alone because, by their nature, species-specific genes are not easily annotated. Therefore, many small- to medium-sized ORFs cannot be validated without through transcriptomic and proteomic data. The draft annotations for *T. congolense* and *T. vivax* will doubtless contain some spurious coding sequences while omitting others that are genuine, and we have removed any gene encoding a hypothetical protein <150 aa in length from the analyses to reduce this effect. Table S4 lists species-specific gene families derived from a comparison of all genes above the 150-aa threshold by using OrthoMCL (2, 3). *T. brucei*-specific clusters are dominated by familiar surface-expressed genes: invariant surface glycoprotein (*ISG*), *T. brucei*-alanine-rich protein (*BARP*), *VSG*, expression site-associated genes (*ESAG*), procyclin, nucleoside transporters, retrotransposon hot spot (RHS) protein, and hypothetical proteins. A total of 48 of 68 *T. congolense* clusters and 143 of 179 *T. vivax* clusters contained putatively surface-expressed genes (i.e., encoding putative signal peptides, transmembrane helices, or GPI anchor).

Several observations indicate that these draft genome assemblies are largely accurate and contain most genes: the high proportion of orthologs found in all species, the general colinearity of core chromosomal gene order, and, where gene losses do occur, the low proportion of apparent deletions subsequently identified in unassembled reads. Improvement of the genome sequence will not, we believe, introduce further genes into core regions. It would close gaps in strand-switch regions and subtelomeres where sequence gaps currently exist, and this would place largely species-specific genes that are currently on unassembled contigs (but still included in comparisons of gene content). By emphasizing the gross similarity of these genome sequences, and by stressing the importance of change on the cell surface, we do not say that there are no differences in genes expressed within the cell. There are differences, often substantial, but we cannot at present relate these to phenotypic differences in life cycle, host range, and cellular morphology (4, 5). However, the differences listed in Tables S2 and S3 may yet affect the genomic basis to core cellular physiology.

Cell-Surface Phylome. The surface phylome was created by taking every *T. brucei*, *T. congolense*, and *T. vivax* gene >450 bp in length that included one or more transmembrane helices, a predicted signal peptide, and/or a predicted GPI anchor. Each gene was compared with all genes by using BLASTp to identify homologs in each species (if any). The gene and its homologs were then removed from the analysis before moving on to the next case. Phylogenetic analysis is not practical with fewer than four taxa, so a family was defined as a group of homologous with at least four copies in one or more species, resulting in 4,238 BLAST analyses and 291 putative families. The list was checked to remove families without credible evidence for cell-surface expression and those associated with intracellular and mitochondrial membranes and to combine families split by BLAST

comparison but that were still homologous. Multiple sequence alignments were also scrutinized to remove cases of dubious annotation. This produced 81 families that are described in Table S5. Families in the cell-surface phylome will not contain surface-expressed genes with fewer than three homologs in at least one species. It is also conceivable that surface-expressed families have been omitted because they do not possess signal peptides, GPI anchors, or transmembrane helices that can be recognized by current methods. Equally, spurious recognition of these domains in hypothetical proteins (mostly *T. vivax* families) cannot be ruled out. The phylome is accessible through GeneDB (http://www.genedb.org/Page/trypanosoma_surface_phylome).

Fam1 Expression. Most of the species-specific genes identified in the surface phylome encode hypothetical proteins with no matches to existing databases. Therefore, we sought to validate in principle our description of these genes as both protein-coding and surface-expressed by using Fam1 (a *T. brucei*-specific *VSG*-like gene with predicted signal peptide and GPI anchor). An eightfold increase in mRNA levels for bloodstream-form cells grown in vivo compared with those grown in vitro can be seen, suggesting that a host–pathogen interaction is required to stimulate transcription of Fam1 to its full extent. In addition, short stumpy in vivo and procyclic form in vitro relative mRNA quantities are also significantly down-regulated in comparison with bloodstream form in vivo (Fig. S5*A*), confirming the expression profiles produced in genome-wide microarray experiments (6).

A Fam1 protein (Tb927.6.1310) was expressed with a constitutive expression system (pXS5) and tagged at the N terminus of the mature protein with an HA-9 epitope. The HA epitope was placed two residues C-terminal to the predicted mature N terminus of the mature protein after signal sequence processing. By Western blot analysis, a single band was detected migrating at ∼45 kDa in whole-cell lysates. However, the predicted molecular mass of the protein is ∼39 kDa, suggesting the addition of N-linked carbohydrate. Cells were stained with a monoclonal antibody against HA and counterstained with FITC–concanavalin A (FITC-ConA) (Fig. S5*C*). At 4 °C, Tb927.6.1310 clearly co-localized with ConA, conditions that block endocytosis and so retain the lectin exclusively within the flagellar pocket, a subdomain of the plasma membrane, therefore demonstrating access of Tb927.6.1310 to the cell surface. When cells were permeabilized with detergent, it was clear that Tb927.6.1310 was also present in additional internal compartments and, based on partial overlap of ConA at 12 °C (which retains ConA in the flagellar pocket and Rab5-positive early endosomes) and Tb927.6.1310 signals, these structures likely correspond to early and recycling endosomes. Hence, we conclude that Tb927.6.1310 is present at the parasite surface; may be restricted to the flagellar pocket, which is frequently observed for low-abundance GPI-anchored proteins in this organism; and is also present within the endosomal apparatus.

***T. vivax* Transcriptome.** The *T. vivax* genome sequence contains hundreds of putative coding sequences without any matches to known protein domains or families; most of these belong to gene families with predicted cell-surface expression, as described in Table S5. To support the protein-coding prediction for these unique ORFs, we generated RNAseq data for bloodstream-stage *T. vivax* Y486 in mice: 81% of the data were of mouse origin, and the residual sequence data amounted to only 316,728 Illumina reads, so the transcriptome did not achieve genome-wide coverage. However, Table S6 lists those genes that were represented among the transcriptomic data and shows the percentage coverage of the coding sequence. Transcriptomic data are sufficient to confirm that 12 of the 23 *T. vivax* gene families in the phylome are transcribed. The remaining 11 families may be expressed in other life stages or may be noncoding. It is perhaps significant that the four *VSG*-like gene families are among the most abundant transcripts. For instance, transcripts cover 99% and 95% of TvY486_0900440 and TvY486_0017810 (Fam23), respectively, as well as 99% of TvY486_0028795 (Fam24), 94% of TvY486_0026030 (Fam25), and 93% of TvY486_0002910 (Fam26).

***VSG* Sequence Properties.** The first indication of substantial species differences in *VSG* came from their physical sequence properties. These properties are described in Table S7. Intact *T. brucei VSG* genes (mean length = 498 ± 29 aa) are significantly longer than their counterparts in either *T. congolense* or *T. vivax* (mean length = 388 ± 30 and 394 ± 95, respectively). The difference in length of predicted proteins is due to the expansion of the "hypervariable" domain toward the N terminus and the C-terminal domain (CTD), which is longer than repetitive CTDs in other species. There are marked differences in the predicted chemical properties such as hydropathicity and aromaticity. *T. vivax* VSG are more hydropathic and less aromatic because of the composition bias in the amino acids encoded. Hence, codon use is lowest in *T. vivax VSG*, with other species using a more balanced range of amino acids. This codon use bias is likely to be related to a significant difference in base composition; *T. brucei* and *T. congolense VSG* have similar GC content (pGC = 0.488 ± 0.016 and 0.481 ± 0.032, respectively) but *T. vivax VSG* have a pGC of 0.599 ± 0.019. Although pGC is higher genome-wide in *T. vivax* (Table S1), protein-coding genes generally do not display this disparity in base composition; indeed, it may be this *VSG*-specific effect that explains the higher average for all genes in *T. vivax*.

VSG genomic sequences in *T. congolense* or *T. vivax* also differ in the incidence of pseudogenes and gene fragments. It is a characteristic feature of the *T. brucei* 927 and *Trypanosoma brucei gambiense* 972 genomes that most *VSG* are predicted pseudogenes (i.e., they have internal stop codons or frameshifts) or gene fragments (7, 8). More recent data are confirming that this is typical of the entire species. However, in *T. congolense*, only 21.2% of Fam13 and 29.7% of Fam16 genes were predicted to be pseudogenes, and we observed no sequence fragments. Similarly, in *T. vivax*, we find that predicted pseudogenes are a minority in Fam23 (15.5%), Fam24 (27.2%), Fam25 (26.5%), and Fam26 (36.9%), and only seven genuine *VSG* fragments were identified (i.e., fragments not adjacent to sequence gaps). The draft *T. vivax* contains many *VSG* fragments next to gaps, e.g., 112 Fam23 members and 52 Fam24 members, but these are otherwise intact.

***VSG* Distribution and Structure.** In *T. brucei*, *VSG* are arranged in irregular tandem arrays at the subtelomeres of some chromosomes (9). We find that *T. congolense* and *T. vivax* also arrange their *VSG* in subtelomeres and not among the core chromosomal loci. We cannot at this stage relate subtelomeric contigs to the core or each other, so it is not possible to confirm that the *VSG* are also associated with certain chromosomes or the minichromosomes; although we have identified unassembled *T. vivax* contigs displaying the conserved minichromosomal repeat motif (10) and putative *VSG* are present on these contigs (e.g., TvY486 bin contig 2898). In *T. congolense*, the principal differences are that *VSG* are mostly intact rather than pseudogenic (see above) and they are intermixed with genes of other large families, such as transferrin receptor (*TFR*)-like genes (Fam14/15), invariant surface glycoprotein (Fam49), and Fam22, a *T. congolense*-specific hypothetical gene found immediately downstream of some *VSG* (see the cell-surface phylome website for more detail). *T. vivax VSG* are also mostly intact and often intermixed with other species-specific families. However, here we observe tandem duplication patterns that suggest *T. vivax VSG* proliferate through this mechanism and that explain the clusters of almost identical *VSG* described elsewhere in this paper.

VSG are telomerically expressed in *T. brucei* from a dedicated expression site that, for bloodstream-stage expression, also includes various *ESAG* (*ESAG1–12*) (11). The canonical structure of the expression site is conserved between telomeres, although the presence or absence of individual components varies (12). Although we cannot directly assess the presence of telomeric VSG and dedicated VSG expression sites from the draft *T. congolense* and *T. vivax* genome sequences, we can say that the components of the *T. brucei* bloodstream expression site are largely absent from other species. Thus, the surface phylome demonstrates that some *ESAG* are unique to *T. brucei* (e.g., *ESAG1*, *-9*, and *-11*), whereas others are *T. brucei*-specific members of more widespread families (e.g., *ESAG2–5* and *-10*). Only the *TFR* genes *ESAG6/7* have orthologous lineages beyond *T. brucei* (i.e., in *T. congolense*). Other features of the canonical expression site, such as the 70-bp repeat region, were also not observed outside of *T. brucei*, indicating that the regulation of *VSG* expression may have important species differences, which specific sequencing of *T. congolense* and *T. vivax* telomeres will examine.

The primary structures of *T. congolense* and *T. vivax* VSG contain the patterns of conserved glycine, tryptophan, and, particularly, cysteine residues that have been documented among *T. brucei* VSG (9, 13, 14). Multiple sequence alignments of both a- and b-type *VSG* (a-*VSG* and b-*VSG*) from all three species are available from the cell-surface phylome website. These demonstrate that familiar motifs in *T. brucei VSG*, such as the "GRIDE" motif of a-VSG and TFRs (13, 15) and the CxC motif of b-VSG (14), are widespread; these and several other conserved residues provide clear justification of the structural and evolutionary unity of VSG-like proteins from all African trypanosomes. And although the *T. vivax*-specific genes that we have designated *VSG*-like (i.e., Fam25 and Fam26), which do encode proteins with the precise motifs familiar from a-*VSG* and b-VSG, they do display topologically similar patterns of conserved cysteine residues, which ultimately are responsible for the low BLAST matches with recognized VSG. Finally, although in *T. brucei* the CTD is considered to be relatively invariant (for very species-specific reasons), it should be noted that across all *VSG* subfamilies, the CTD is typically the most evolutionarily labile part of the molecule.

Accuracy of *VSG* Repertoires. We have compared *VSG* repertoires from two draft genome sequences with that of the finished *T. brucei* 927 genome sequence. Inevitably, this comparison is not exhaustive because, first, gaps in the draft sequences may contain more *VSG* and, second, because it was sequenced from single-chromosome shotgun libraries, the *T. brucei* genome does not include the numerous *VSG* located on intermediate and minichromosomes or fully contiguated sequences for all subtelomeric regions. Despite this, the *VSG* repertoires are representative of the whole. Although annotated *VSG* in *T. brucei* 927 may represent only half of the total number (9), the remaining genes are unlikely to add further diversity. We know this because published minichromosomal *VSG* are not structurally distinct (16) and cluster among the known repertoire in phylogenies (see a-*VSG* phylogeny on the cell-surface phylome website). Similarly, telomeric *VSG* from *T. brucei* 927 and 427 (12) are not structurally different from the nontelomeric repertoire, although we have included them for completeness. Finally, we have previously compared the *T. brucei* 927 *VSG* repertoire with that of *T. b. gambiense* 972, the genome of which was sequenced by using a whole-genome shotgun approach (17). In the previous study, despite the underrepresentation of VSG in *T. brucei* 927, the two repertoires overlapped significantly (17), and there were no *T. b. gambiense*-specific groups of *VSG* that would have suggested that the *T. brucei* 927 repertoire did not accurately represent global diversity. In short, those *VSG* missing from the *T. brucei* do not compromise our ability to use the repertoire in our comparisons because they are simply "more of the same."

Similarly, we believe that any *T. congolense* or *T. vivax* VSG not annotated here will not add significantly to genome-wide *VSG* diversity as we have defined it. Published *VSG* sequences from *T. congolense* (18–21) cluster throughout the Fam13 and Fam16 phylogenies (see phylome pages), showing that the observed repertoire encompasses all known variant antigens. And although there are ∼170 partial *T. vivax VSG* in the present assembly that are curtailed by sequence gaps, these are members of established clades, not novel sequence types.

If our *VSG* repertoires do not underrepresent diversity, they could perhaps exaggerate it. In *T. brucei*, the experimental literature makes it clear that *ESAG2*, *VSG*-related (*VR*), and other *VSG*-like genes do not contribute to the pool of variant antigens. However, if analogous gene families were present in *T. congolense* or *T. vivax*, we do not know of them, and, given the structural divergence from *T. brucei*, we could not identify them. Although we have shown that known *T. congolense VSG* fall within both Fam13 and Fam16, and that Fam23 and Fam26 contain proven *T. vivax VSG*, it remains possible (perhaps even probable) that *VSG*-like genes with derived functions exist among the many other cases.

Phylogenetic Position of *VR* Genes. It is known that *VR* genes differ from canonical *VSG* in two ways. First, they are located on the chromosomal cores (principally at strand-switch regions) rather than in subtelomeres, and these positions are conserved between subspecies. Second, they lack the diagnostic N-terminal domains and CTDs thought to be functionally important in antigenic variation. Although occasionally observed in expression sites (12), expression of *VR* genes as functional variant antigens is not observed. *VR* genes have been considered to be secondarily reduced *VSG*, transposed to the chromosomal core, that have lost their variant antigen function (7). If so, in phylogenetic trees and networks, the *VR* genes would cluster within the b-*VSG* of *T. brucei* in one or more locations, depending on how many times secondary loss had occurred. However, Fig. 1 and Fig. S4 show that *VR* genes cluster together (suggesting one or two origins) and apart from *T. brucei* b-*VSG*. In fact, because they lack the diagnostic features of canonical *T. brucei* b-*VSG*, i.e., the 5′ hypervariable region and CTD, they have greater sequence identity with *T. congolense VSG* (Fam16), and, in BLAST searches of *T. brucei* proteins using *T. congolense* VSG, the top hits are to *VR* products. These relationships reject secondary reduction as an origin for *VR* and instead promote an older origin in the *T. brucei*/*T. congolense* ancestor.

Evolution of the TFR. *T. brucei TFR* genes are expression site-associated (22, 23) and are the sister clade to *procyclin-associated genes 1*, *-2*, *-4*, and *-5* (*PAG1*, *-2*, *-4*, and *-5*), which are associated with the procyclin expression site (24). *T. congolense* homologs are distributed throughout the subtelomeres and split into *ESAG6*-like and *PAG*-like genes clearly in a phylogeny, corresponding to Fam14 and Fam15 (see a-*VSG* phylogeny in Fig. S3). These *TFR*-like genes are not among the known variant antigens in either species and are not thought to have this function. Their phylogeny indicates that the functional differentiation observed in *T. brucei* is conserved in *T. congolense* (but not in *T. vivax*) because both organisms have two distinct clades of *TFR*-like genes, one of which is GPI-anchored whereas the other is not (see Fam15 on the cell-surface phylome website for more details). In fact, this heterodimeric pattern may also be conserved among *PAG* genes (i.e., Fam14) in both species.

Closer examination of the TFR-like protein sequences in *T. congolense* suggests that the secondary structure of the TFR constituents in *T. brucei* is conserved (Fig. S6), as are amino acid residues required for effective transferrin binding (as defined in

ref. 15). These include the GRLEE motif (GRLDE in *T. congolense*), cysteine residues and peptidic turns that are the most conserved features in *T. brucei*, as well as heptad repeat units and glycine positions within variable regions that are crucial to protein function (15).

All *TFR* genes from both organisms are monophyletic, showing that they descend from the *T. brucei*/*T. congolense* ancestor. Had the *TFR* been co-opted from *VSG* in *T. brucei* (15), *ESAG6*/*7* would nest within *T. brucei* a-*VSG*, closest to their presumed progenitor. However, the undoubted homology between all *TFR* [*ESAG6*/*7*, *PAG*, and *T. congolense* homologs (25)] means that a *TFR*/a-*VSG* ancestor existed. Our observation that *T. vivax* employs an a-*VSG* protein as a variant antigen indicates that this ancestor was a variant antigen primarily and that TFR evolved from it in the *T. brucei*/*T. congolense* ancestor. Clearly, we would like to know what *T. vivax* uses to bind transferrin: some or all a-VSG or a nonhomologous receptor? Hence, the evolution of *TFR* genes from the a-*VSG* lineage constitutes an ancient functional innovation, which we predict is shared by *T. brucei* and *T. congolense* but absent from *T. vivax*.

Absence of a-VSG Variant Antigens in *T. congolense*. Roughly half of *T. brucei VSG* genes are a-*VSG*, and *T. vivax* possesses a family of over 500 a-*VSG*–like genes, including a proven variant antigen. It is surprising therefore that *T. congolense* possesses no a-*VSG*–like family other than *TFR*-like genes. Published *T. congolense* VSG and expressed sequence tags do not include any a-*VSG*–like sequences or, indeed, any of the *TFR*-like genes. The absence of a-*VSG* variant antigens in *T. congolense* can be explained through either evolutionary loss (i.e., deletion from the *T. congolense* genome) or gain (i.e., secondary gain by the *T. brucei* genome), and it is unclear which scenario is correct.

In the "loss" hypothesis, a-*VSG* were present in the *T. brucei*/*T. congolense* ancestor and included two distinct lineages encoding variant antigens and TFR, respectively. The *TFR* lineage was clearly inherited by both daughter species, but the variant antigen lineage was retained by *T. brucei* alone. This hypothesis is simple but it requires the deletion of a large number of genes, the mechanism for which is doubtful. Assuming that the *T. brucei*/*T. congolense* ancestor possessed an a-*VSG* repertoire like *T. brucei* and *T. vivax*, which is reasonable if it were to function in antigenic variation, hundreds of genes must have been deleted after speciation to produce a *T. congolense* genome lacking even the vestiges of orthologous a-*VSG*.

In the "gain" hypothesis, a-*VSG* variant antigens were not present in the *T. brucei*/*T. congolense* ancestor, but the *TFR* lineage was present. a-*VSG* in *T. brucei* evolved through the donation of the CTD from b-*VSG* to a *TFR* gene, forming a chimera that could function as a variant antigen. This hypothesis does not require mass deletion, accounts for the unique (and anomalous) CTD common to all *T. brucei VSG*, and explains the sequence similarity between a-*VSG* and *TFR*. However, close examination of the phylogenetic evidence indicates that *T. brucei* a-*VSG* are not significantly more related to *T. brucei TFR* genes than to their *T. congolense* counterparts, which is inconsistent with a gain hypothesis.

Therefore, until further data are generated from other African trypanosome genomes, for instance, from an organism intermediate between *T. brucei* and *T. congolense* (26), we favor the simpler loss hypothesis that posits an a-*VSG* variant antigen family conserved since the ancestral African trypanosome genome.

SI Materials and Methods

Genome Sequencing and Annotation. *T. congolense* IL3000 and *T. vivax* Y486 were propagated as described previously (27, 28). High molecular-weight DNA was extracted in late log phase by phenol-chloroform extraction and purified by gel electrophoresis. Genomic DNA was capillary-sequenced by using a whole-genome shotgun strategy as described previously (9). Sequence reads were assembled with Phrap (http://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 (9) (TriTrypDB v1.0) by using ABACAS (29). The manually curated genome annotation of *T. brucei* was transferred to the *T. congolense* and *T. vivax* assemblies with custom perl scripts, based on sequence and positional homology, and manually edited where appropriate with Artemis (30). Ordering contigs against the *T. brucei* reference creates pseudochromosomes 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-Mb chromosomes (31). However, *T. vivax* is more distantly related with an uncertain karyotype (31). Therefore, in addition to producing pseudochromosomes, we manually assembled scaffolds from *T. vivax* contigs by using read-pair information.

Annotation of *VSG* Genes. *VSG* structures are highly mutable, and therefore annotation transfer and sequence homology with *T. brucei VSG* may not adequately annotate variant antigens in other species. *T. congolense* and *T. vivax VSG* were identified by searching with hidden Markov models built using HMMER v3.0 (http://hmmer.janelia.org/) from *T. brucei* a-*VSG* and b-*VSG* sequence alignments and then, once identified, native *T. congolense* and *T. vivax VSG*. This process revealed few *VSG* in addition to those identified by BLASTp-based homology searches. However, it did show that many gene models were partial, which is particularly relevant to the quantification of pseudogenes, which might be underestimated if full-length coding regions were not marked up. So the boundaries of all ORFs identified by hidden Markov models as being homologous to *VSG* in *T. congolense* and *T. vivax* were manually checked 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 deposited in the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (EMBL-Bank): *T. congolense* accession nos. HE575314–HE575324 and CAEQ01000352–CAEQ01002824 and *T. vivax* accession nos. 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 no. 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 (2, 3) to examine species-specific genes, gene families, and 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* pseudochromosomes with the Artemis Comparison Tool [ACT (32)]. Disruptions to colinear gene order were identified; however, because sequence gaps occasionally prevented a three-way comparison, we only considered disruptions that occurred within contigs (i.e., that 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 (33), intracellular transport (34), glycosyl transfer (35),

ribosomal structure, phosphorylation, and a range of genes involved in metabolism. All putative losses were confirmed by examining expected genomic position and searching unassembled sequence reads for reciprocal sequence matches with tBLASTn/BLASTx.

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

Analysis of Fam1 Gene Expression. To determine mRNA expression levels of Fam1 family members, quantitative real-time PCR (qRT-PCR) was carried out on total RNA extracted with the RNeasy Mini Kit (QIAGEN). cDNA was generated with SuperScript II Reverse Transcriptase according to the manufacturer's instructions. qRT-PCR was carried out with 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 (36). Ectopic expression of HA epitope-tagged Tb927.6.1310 at the N terminus (after the predicted signal peptide sequence) was carried out by using pXS5/pDEX-577 (37) constitutive and inducible expression vectors, respectively. For protein extraction, proteins were transferred onto Immobilon polyvinylidene fluoride membrane and incubated with primary mouse anti-HA antibody (1:8,000) and subsequently with secondary rabbit anti-mouse peroxidase-conjugated antibody (1:10,000; Sigma). Immunofluorescence microscopy was carried out on permeabilized and nonpermeabilized 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 an intact immune system, inducing a relapsing parasitaemia. After 14 d, trypanosomes were purified from the blood by Percoll gradient fractionation, as described in ref. 28. 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 population with the initiating population prepared in the same way revealed significant differences in both dimensions in an ∼40-kDa spot group, which is consistent with *VSG* switching. Both extracts were run in 1D SDS/PAGE, and three bands in the estimated size range were extracted from each, trypsinized, and subjected to liquid chromatography/tandem MS 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 by using a primer specific to the 5′ spliced leader (38) and an anchored oligo(dT) primer. A dominant ∼1.3-kb band was gel-extracted and 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 *Trypanosoma cruzi* predicted genes (the latter was included as an outgroup) with wuBLAST. At least four homologs occurring in at least one species 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 with Signal P (39). GPI anchors were predicted with FragAnchor (40), and transmembrane helices were predicted with TMHMM (41). The total of 291 surface-expressed families was reduced to 81 by removing cases of poor alignment (i.e., spurious homology), obvious noncoding sequence (i.e., misannotation), 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 (42); 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 with MrBayes v3.2.1 (43, 44) (Nruns = 2, Ngen = 10,000,000, samplefreq = 1,000, and default prior distribution). Nucleotide sequence alignments were analyzed with a GTR+Γ model. Maximum likelihood phylogenies were estimated with PHYML v3.0 (45) under an LG+Γ model (46) for amino acid sequences or a GTR+Γ model for nucleotide sequences. Node support was assessed by using 100 nonparametric bootstrap replicates (47). The trees were rooted with *T. cruzi* sequences or otherwise midpoint rooted. Bayesian *VSG* phylogenies were estimated by using alignments of selected full-length sequences representative of global diversity (Nruns = 1, Ngen = 1,000,000, samplefreq = 100, and default prior distribution). Treeness (48) was calculated for each tree topology by using TreeStat v1.2 (http://tree.bio.ed.ac.uk/software/treestat/); treeness is defined as the proportion of total tree length taken up by internal branches and measures the signal-to-noise ratio in a phylogenetic dataset (49).

Recombination Analysis. Recombination results in sequence alignments with multiple phylogenetic signals (50), otherwise known as phylogenetic incompatibility (PI). The pairwise homoplasy index (51) returns a single probability value for PI, which was applied to amino acid sequence alignments for seven *VSG* subfamilies (Table S8). For each alignment, 1,000 subalignments of 10 sequences were prepared by selecting sequences at random. The proportion of subalignments with significant PI, termed P_{pi}, was compared between species. Confidence intervals on P_{pi} were obtained by repeating the analysis on 100 nonparametric bootstraps of each alignment generated with Seqboot (http://evolution.genetics.washington.edu/phylip/doc/seqboot.html). To confirm that significant PI was not simply caused by rate heterogeneity or other forms of homoplasy, a null distribution for P_{pi} was obtained from simulated alignments generated with Seq-Gen (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 with alignments of sequences belonging to individual crown clades only as defined by *VSG* subfamily phylogenies; we refer to this process 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 because there is no obvious CTD.

1. El-Sayed NM, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–409.
2. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.
3. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS (2006) OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34(Database issue):D363–D368.
4. Maxie MG, Losos GJ, Tabel H (1979) Experimental bovine trypanosomiasis (*Trypanosoma vivax* and *T. congolense*). I. Sumptomatology and clinical pathology. *Tropenmed Parasitol* 30(3):274–282.
5. Moloo SK, Kabata JM, Gitire NM (2000) Study on the mechanical transmission by tsetse fly Glossina morsitans centralis of Trypanosoma vivax, T. congolense or T. *brucei brucei* to goats. *Acta Trop* 74(1):105–108.
6. Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M (2009) Widespread variation in transcript abundance within and across developmental stages of *Trypanosoma brucei*. *BMC Genomics* 10:482.
7. 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.
8. Pays E, Salmon D, Morrison LJ, Marcello L, Barry JD (2007) Antigenic variation in *Trypanosoma brucei*. *Trypanosomes: After the Genome*, eds Barry JD, McCulloch R, Mottram J, Acosta-Serrano A (Horizon Bioscience, Norfolk, UK), pp 339–372.
9. Berriman M, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309:416–422.
10. Dickin SK, Gibson WC (1989) Hybridisation with a repetitive DNA probe reveals the presence of small chromosomes in *Trypanosoma vivax*. *Mol Biochem Parasitol* 33(2): 135–142.
11. Cully DF, Gibbs CP, Cross GA (1986) Identification of proteins encoded by variant surface glycoprotein expression site-associated genes in *Trypanosoma brucei*. *Mol Biochem Parasitol* 21(2):189–197.
12. Hertz-Fowler C, et al. (2008) Telomeric expression sites are highly conserved in *Trypanosoma brucei*. *PLoS ONE* 3:e3527.
13. Carrington M, et al. (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.
14. Blum ML, et al. (1993) A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature* 362:603–609.
15. Salmon D, et al. (1997) Characterization of the ligand-binding site of the transferrin receptor in *Trypanosoma brucei* demonstrates a structural relationship with the N-terminal domain of the variant surface glycoprotein. *EMBO J* 16:7272–7278.
16. Alsford S, Wickstead B, Ersfeld K, Gull K (2001) Diversity and dynamics of the minichromosomal karyotype in *Trypanosoma brucei*. *Mol Biochem Parasitol* 113(1): 79–88.
17. Jackson AP, et al. (2010) The genome sequence of *Trypanosoma brucei* gambiense, causative agent of chronic human African trypanosomiasis. *PLoS Negl Trop Dis* 4:e658.
18. Strickler JE, et al. (1987) *Trypanosoma congolense*: Structure and molecular organization of the surface glycoproteins of two early bloodstream variants. *Biochemistry* 26:796–805.
19. 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.
20. Eshita Y, Urakawa T, Hirumi H, Fish WR, Majiwa PA (1992) Metacyclic form-specific variable surface glycoprotein-encoding genes of *Trypanosoma* (*Nannomonas*) *congolense*. *Gene* 113(2):139–148.
21. Helm JR, et al. (2009) Analysis of expressed sequence tags from the four main developmental stages of *Trypanosoma congolense*. *Mol Biochem Parasitol* 168(1): 3442.
22. 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):1–16.
23. Schell D, 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* J10:1061–1066.
24. 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(1-2):135–145.
25. Carrington M, Boothroyd JC (1996) Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol Biochem Parasitol* 81:119–126.
26. Hamilton PB, Adams ER, Malele II, Gibson WC (2008) A novel, high-throughput technique for species identification reveals a new species of tsetse-transmitted trypanosome related to the *Trypanosoma brucei* subgenus, *Trypanozoon*. *Infect Genet Evol* 8(1):26–33.
27. Hirumi H, Hirumi K (1991) In vitro cultivation of *Trypanosoma congolense* bloodstream forms in the absence of feeder cell layers. *Parasitology* 102(2):225–236.
28. Ndao M, Magnus E, Büscher P, Geerts S (2004) *Trypanosoma vivax*: A simplified protocol for in vivo growth, isolation and cryopreservation. *Parasite* 11(1):103–106.
29. Cutler DJ, et al. (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res* 11:1913–1925.
30. Rutherford K, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16:944–945.
31. Van der Ploeg LH, Cornelissen AWCA, Barry JD, Borst P (1984) Chromosomes of kinetoplastida. *EMBO J* 3:3109–3115.
32. Carver T, et al. (2008) Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 24:2672–2676.
33. Broadhead R, et al. (2006) Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature* 440(7081):224–227.
34. Koumandou VL, Natesan SK, Sergeenko T, Field MC (2008) The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. *BMC Genomics* 9:298.
35. Izquierdo L, et al. (2009) Identification of a glycosylphosphatidylinositol anchormodifying β1-3 N-acetylglucosaminyl transferase in *Trypanosoma brucei*. *Mol Microbiol* 71:478–491.
36. Wirtz E, Leal S, Ochatt C, Cross GA (1999) A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol Biochem Parasitol* 99(1):89–101.
37. Kelly S, et al. (2007) 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(1):103–109.
38. De Lange T, et al. (1984) Comparison of the genes coding for the common 5′ terminal sequence of messenger RNAs in three trypanosome species. *Nucleic Acids Res* 12:4431–4443.
39. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.
40. Poisson G, Chauve C, Chen X, Bergeron A (2007) FragAnchor: A large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. *Genomics Proteomics Bioinformatics* 5(2):121–130.
41. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305:567–580.
42. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
43. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
44. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
45. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
46. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25:1307–1320.
47. Felsenstein J (1985) Confidence-limits on phylogenies-An approach using the bootstrap. *Evolution* 39:783–791.
48. 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.
49. Freeman TC, et al. (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol* 3:2032–2042.
50. Weiller GF (2008) Detecting genetic recombination. *Methods Mol Biol* 452:471–483.
51. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681.

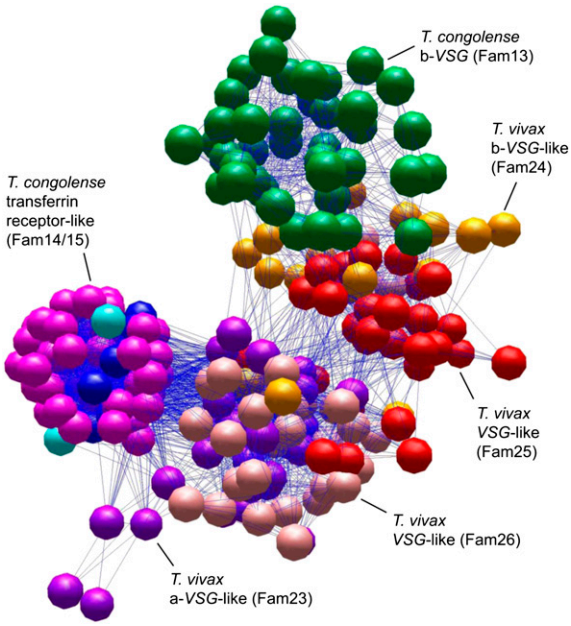
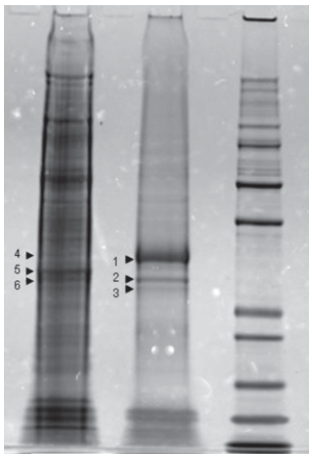


Fig. S1. A clustering network of *T. vivax VSG*-like protein sequences. The network was created with BioLayout Express v3.0 using pairwise FASTA scores for *T. vivax VSG*-like protein sequences: Fam23 (a-*VSG*–like; purple), Fam24 (b-*VSG*–like; orange), Fam25 (red), and Fam26 (pink). The illustration shows that Fam23 clusters closest to a-*VSG*–like sequences (i.e., *TFR*-like proteins) from *T. congolense* (pink) and *T. brucei* (blue), and Fam24 clusters closest to b-*VSG* sequences from *T. congolense* (green). Forming an arc of FASTA connections between a-*VSG* and b-*VSG* are Fam25 and Fam26. Thus, these two *T. vivax*-specific sequence families appear intermediate.

A.



B.

MTAGAARILLALFAGCFCVLARNAAGGSDKA
IAGDDMGK**VCAASSALK**ATAAR**AGEEMLGAE**
ERTRHAVGNASAWQHTVQEQAANMSGTGKQA
AAWAKEETDKRIAKAADALRRVQRAALSVAR
RATRAAARI**DEMVILFTTYTSK**TSSGLACVK
AGSTR**TKPTSYGDGALTWAAGKGSTLK**GCAD
EKWTKGTTDLATDAAKLAETTKALGKLGAGT
AGKLFDGATTSSGVQHACPLLSGSGSSAITT
DYAQIYDTQYDSSTLTQMGGLWQVSAKSNAV
VLDLVEEDTDTVETSKKQPLKQLR**ADAAALW**
KSINATEPDTEEAETLEQALQQLAALPATAF
SVGTQSGERT**AEEWVQLEATLAQQK**KHAAKR
GPTRNDAGTAEQEGAQAITGGEATTVTDGRG
EETHTQATTGDKSAASSGTQRLYRAWALLAA
NTLGANARGAQHSGARGRLA

Fig. S2. Sequence of an expressed VSG in *T. vivax* Y486. (*A*) Relapse-specific protein band yields a putative VSG. Trypanosome lysates derived from days 7 and 14 of a single infection initiated with a population of unknown VSG composition were separated by SDS/PAGE and stained with SYPRO-Ruby. Bands were excised and analyzed by MS. The identities of proteins were derived from Mascot analysis, and the numbered bands were identified as follows: 1, putative VSG; 2, 3, 5, and 6, GAPDH (predicted isoforms in *T. vivax* are 35.6, 38.8, and 39.1 kDa); and 4, actin (predicted 40.9 kDa). (*B*) Tryptic peptide sequences (underlined) from band 1 in *A*, determined by MS analysis, mapped onto the protein predicted for TvY486_0027060.

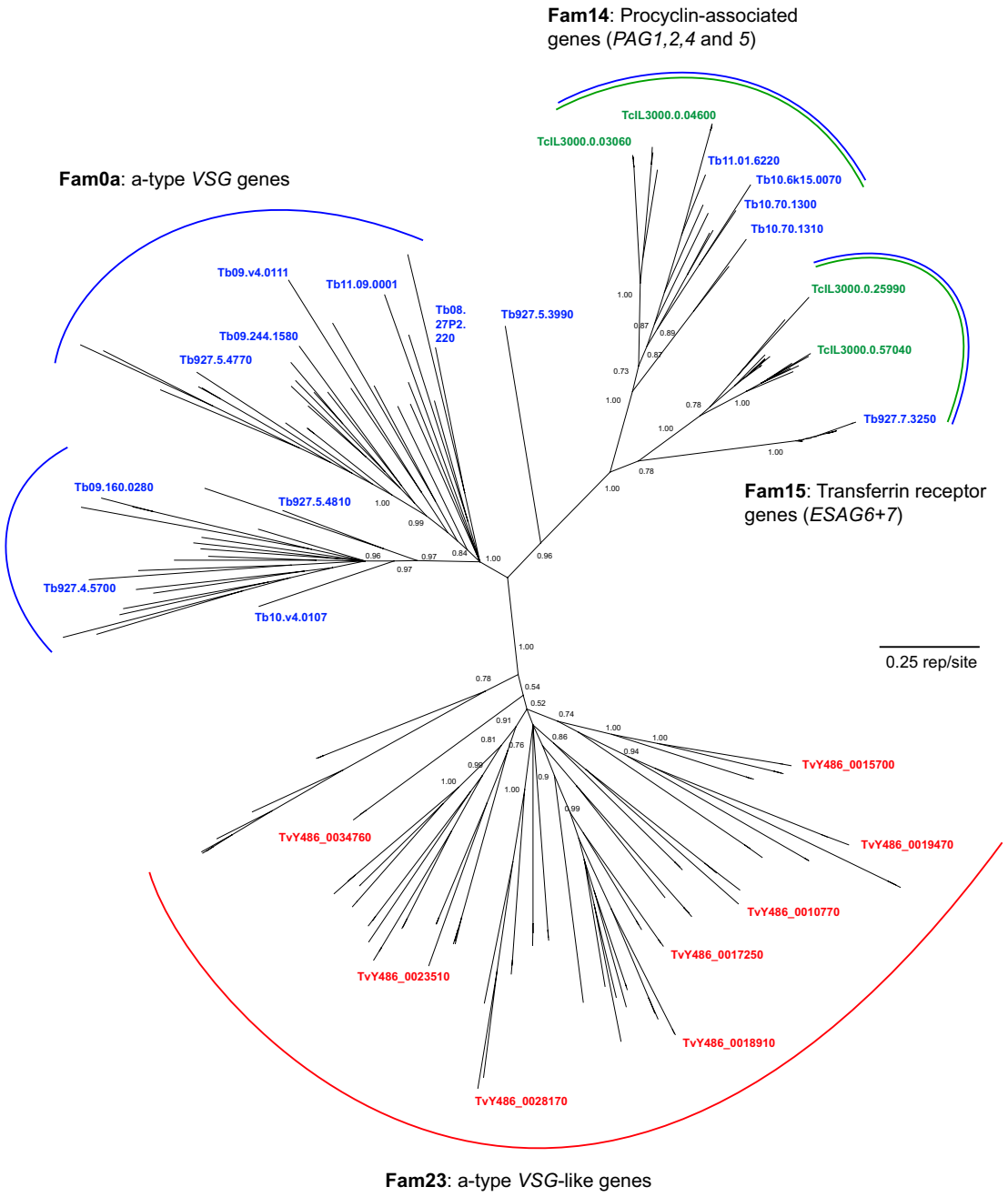


Fig. S3. a-*VSG* phylogeny. A Bayesian consensus phylogram was estimated from a multiple-protein sequence alignment of 421 characters, using MrBayes with a WAG model and corrections for rate heterogeneity. Four Markov chain Monte Carlo (MCMC) chains were run for 5,000,000 generations to encourage maximal convergence of model parameters. The tree is unrooted. Selected nodes are supported by posterior probability values and nonparametric bootstraps generated from a maximum likelihood analysis using an LG model with rate heterogeneity executed with RAxML. a-*VSG* from *T. brucei*, Fam23 from *T. vivax*, and *TFR*-like genes from both *T. brucei* and *T. congolense* form three robust clades. Example gene names use their GeneDB/TriTrypDB locus identifiers. Fam0a splits into two distinct clades corresponding to sequences with (upper) and without (lower) the GRIDE motif generally characteristic of a-*VSG* families.

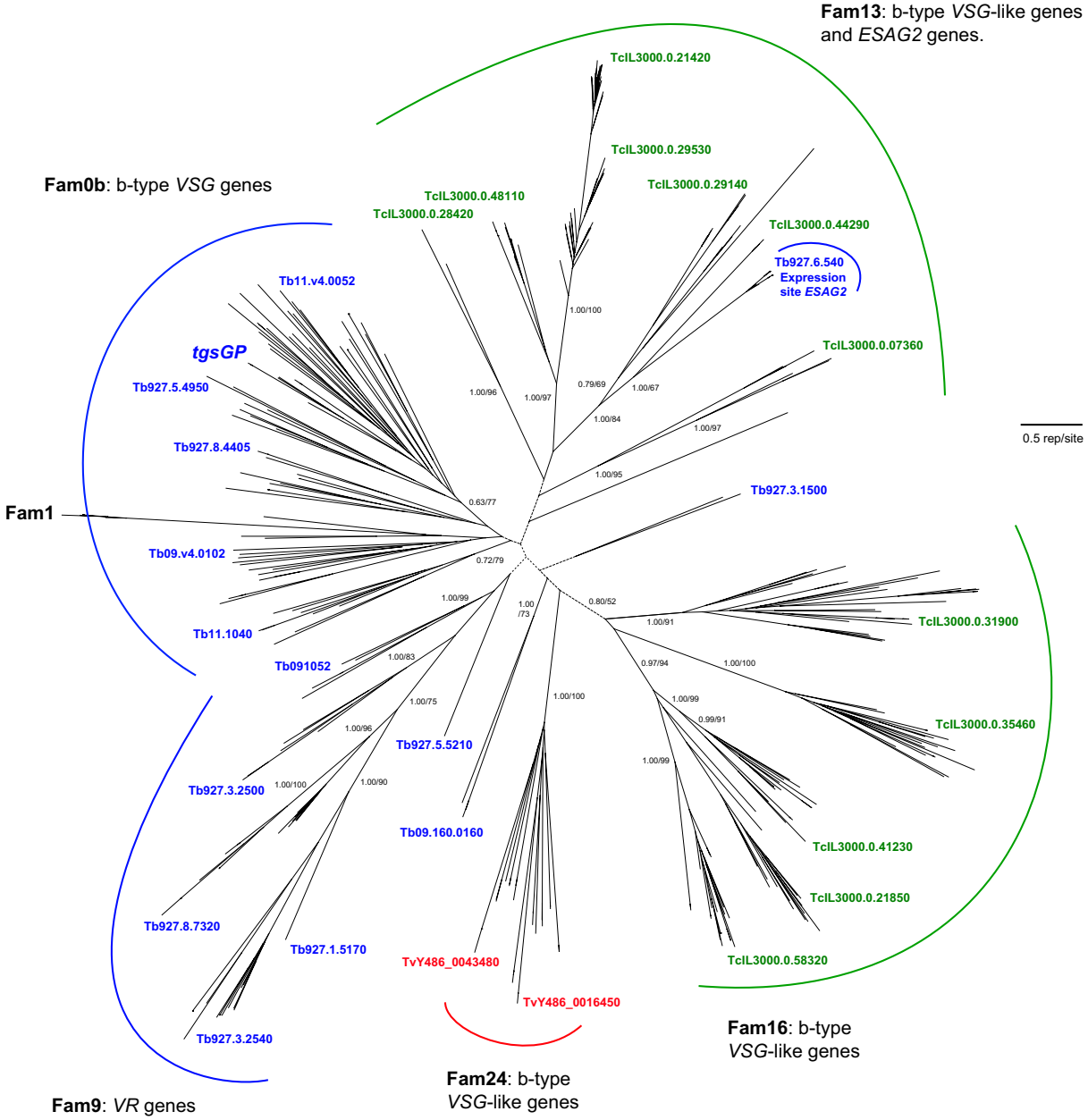


Fig. S4. b-*VSG* phylogeny. A maximum likelihood phylogram was estimated from a multiple-protein sequence alignment of 660 characters by using RAxML with an LG model and corrections for rate heterogeneity. The tree is unrooted. Basal nodes are particularly unstable and have very little or no support; these regions are indicated by dashed branches. Selected nodes are supported by nonparametric bootstraps and posterior probability values generated from a Bayesian analysis of the same dataset in MrBayes. The tree contains major clades that correspond with BLAST-defined surface phylome families. In this tree, canonical *T. brucei VSG* include the atypical *VR* Fam1 gene family. The b-*VSG*–derived *tgsGP* sequence is also shown. All *VR* genes are positioned outside of canonical *T. brucei VSG* but are not monophyletic. *ESAG2* from *T. brucei* is nested within Fam13 *T. congolense* genes.

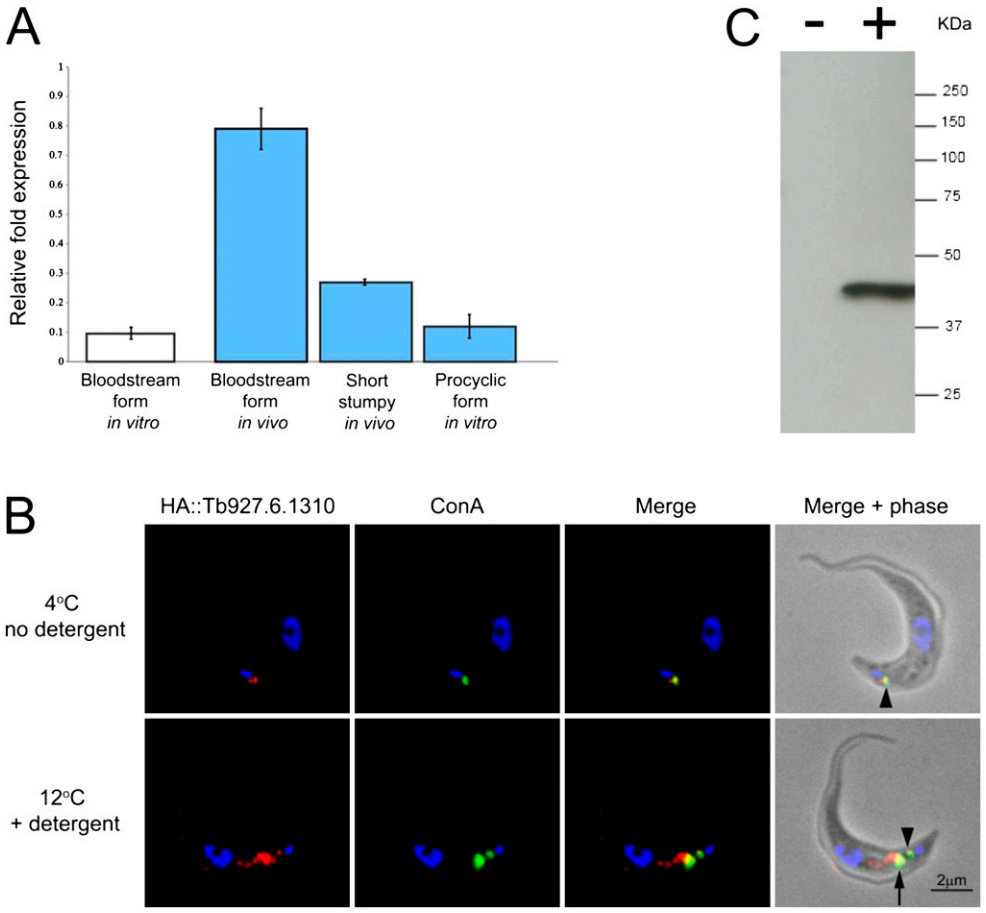


Fig. S5. Expression of a *T. brucei*-specific *VR* protein (Fam1). (*A*) qRT-PCR analysis of Fam1 RNA expression in three different life-cycle stages. RNA levels are relative to a constitutive gene (Rab11) RNA expression level set to 1.0. Error bars show the SEM for each triplicate experiment. (*B*) Immunofluorescence analysis. Tb927.6.1310 was N-terminally HA epitope-tagged and expressed in bloodstream-form cells. Cells expressing HA::Tb927.6.1310 were loaded with FITC–concanavalin A (FITC-ConA) in serum-free media and incubated at either 4 °C (*Upper*) or 12 °C (*Lower*). ConA is restricted to the flagellar pocket at the lower temperature, whereas it accumulates within Rab5A-positive early endosomes at 12 °C. Columns (from left to right): HA epitope tag (red), FITC-ConA (green), fluorescence merge, and merged images from phase and fluorescence. DAPI stain of the nucleus and kinetoplast is shown in blue. HA::Tb927.6.1310 colocalizes with ConA at the flagellar pocket (arrowhead) and early endosomes (arrow). Note that internal structures were not visualized in *Upper*. The abundance of Tb927.6.1310 mRNA is rather low, and the data here suggest that, whereas the protein is able to access the cell surface (the flagellar pocket is a specific subdomain of the plasma membrane), the Tb927.6.1310 protein is restricted from the bulk plasma membrane. (*C*) Western blot analysis of whole-cell lysates from nontransfected (–) and transfected (+) trypanosomes.

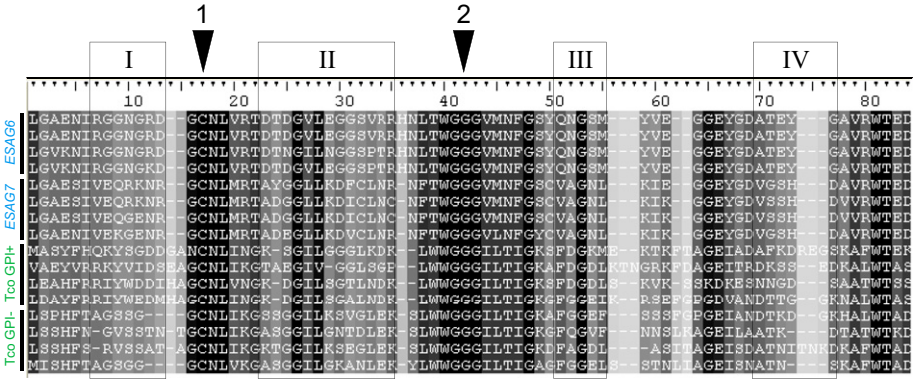
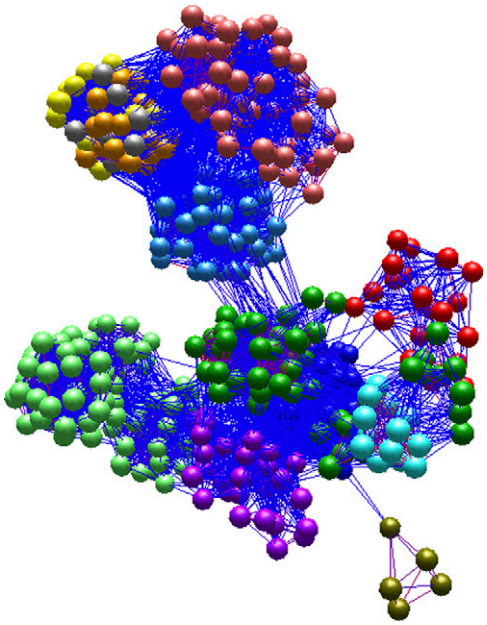


Fig. S6. An amino acid sequence alignment for the ligand-binding domain of the TFR (*ESAG6/7*) in *T. brucei* and homologous sequences from *T. congolense*. The most conserved residues are shaded darkest, and they describe a series of conserved glycine and tryptophan residues (labeled at top); the conserved motif WGGG denotes the site of a conserved peptidic turn. Four variable regions (I–IV) that are required for transferrin binding are shown (1). Glycine residues in II and III are conserved throughout. The scale refers to positions in the complete Fam15 sequence alignment, available from the cell-surface phylome website (http://www.genedb.org/Page/trypanosoma_surface_phylome).

1. Salmon D, et al. (1997) Characterization of the ligand-binding site of the transferrin receptor in *Trypanosoma brucei* demonstrates a structural relationship with the N-terminal domain of the variant surface glycoprotein. *EMBO J* 16:7272–7278.



Movie S1. A 3D rendering of the sequence-similarity network of *VSG*-like sequences presented in Fig. 1. Spheres represent individual sequences shaded according to subfamily as in Fig. 1. Lines connecting spheres represent pairwise maximum likelihood protein sequences extracted 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.

Movie S1

Table S1. Comparison of genome properties for *T. brucei* 927, *T. congolense* IL3000, and *T. vivax* Y486

Table S1

Table S2. Breaks in colinearity along core chromosomal regions

Table S2

Table S3. Exceptions to conservation of African trypanosome genome content by topic

Table S3

Table S4. All species-specific gene families derived from an OrthoMCL analysis of African trypanosomes

Table S4

Table S5. Gene families with predicted cell-surface expression comprising the cell-surface phylome

Table S5

Table S6. Percentage coverage of *T. vivax* genes by transcriptomic (RNAseq) data (cell-surface phylome families listed first)

Table S6

Table S7. Sequence properties of *VSG* genes in each trypanosome species

Table S7

Table S8. Proportion of sequence alignments showing significant PI P_{pi} using three sampling strategies

Table S8