

5 Variant antigen diversity in *Trypanosoma vivax* is not driven by recombination

Sara Silva Pereira¹, Kayo J. G. de Almeida Castilho Neto², Craig W. Duffy¹, Peter Richards¹,
Harry Noyes³, Moses Ogugo⁴, Marcos Rogério André², Zakaria Bengaly⁵, Steve Kemp⁴, Marta
10 M. G. Teixeira⁶, Rosangela Z. Machado², Andrew P. Jackson^{1*}

¹Department of Infection Biology, Institute of Infection and Global Health, University of
Liverpool, 146 Brownlow Hill, Liverpool, L3 5RF, United Kingdom.

²Department of Veterinary Pathology, Faculty of Agrarian and Veterinary Sciences, São
15 Paulo State University (UNESP), Jaboticabal, SP, Brazil.

³Institute of Integrative Biology, University of Liverpool, Biosciences Building, Crown Street,
Liverpool, L69 7ZB, United Kingdom.

⁴Livestock Genetic Programme, International Livestock Research Institute, 30709 Naivasha
Road, Nairobi, Kenya.

20 ⁵International Research Centre for Livestock Development in the Sub-humid Zone (CIRDES),
Nº. 559, rue 5-31 angle, Avenue du Gouverneur Louveau, Bobo-Dioulasso, Burkina Faso.

⁶Department of Parasitology, Institute of Biomedical Sciences, University of Sao Paulo,
Avenue Professor Lineu Prestes, 1374 Cidade Universitaria, 05508-000, Sao Paulo, SP, Brazil.

*Corresponding author. Email: a.p.jackson@liv.ac.uk

African trypanosomes are vector-borne haemoparasites that cause African trypanosomiasis in humans and animals. Parasite survival in the bloodstream depends on immune evasion, achieved by antigenic variation of the Variant Surface Glycoprotein (VSG) coating the trypanosome cell surface. Recombination, or rather directed gene conversion, is fundamental in *Trypanosoma* 30 *brucei*, as both a mechanism of VSG gene switching and of generating antigenic diversity during infections. *Trypanosoma vivax* is a related, livestock pathogen also displaying antigenic variation, but whose VSG lack key structures necessary for gene conversion in *T. brucei*. Thus, this study tests a long-standing prediction that *T. vivax* has a more restricted antigenic repertoire. Here we show that global VSG repertoire is broadly conserved across diverse *T. vivax* clinical strains. We use 35 sequence mapping, coalescent approaches and experimental infections to show that recombination plays little, if any, role in diversifying *T. vivax* VSG sequences. These results explain interspecific differences in disease, such as propensity for self-cure, and indicate that either *T. vivax* has an alternate mechanism for immune evasion or else a distinct transmission strategy that reduces its reliance on long-term persistence. The lack of recombination driving antigenic diversity 40 in *T. vivax* has immediate consequences for both the current mechanistic model of antigenic variation in African trypanosomes and species differences in virulence and transmission strategy, requiring us to reconsider the wider epidemiology of animal African trypanosomiasis.

African trypanosomes (*Trypanosoma* spp.) are unicellular hemoparasites and the cause of African Trypanosomiasis in animals and humans¹. These parasites are transmitted by tsetse flies (*Glossina* spp.), and their proliferation in blood and other tissues leads to anaemia, immune and neurological dysfunction, which is typically fatal if untreated. The profound, negative impact of this disease on livestock productivity across sub-Saharan Africa is measured in billions of dollars annually².

Trypanosoma vivax is a livestock parasite found throughout sub-Saharan Africa and South America³⁻⁵. Although superficially like the more familiar *T. brucei*, (the species responsible for Human African trypanosomiasis), and *T. congolense* (another livestock parasite), *T. vivax* is distinct in morphology and motility⁶, cellular ultrastructure^{7,8} and genetic repertoire^{9,10}. Most conspicuously, it has a truncated life cycle in tsetse flies, lacking a procyclic stage in the insect midgut, and can be transmitted non-cyclically by other genera of hematophagous flies⁶.

Although distinct from *T. brucei*, *T. vivax* shares a defining phenotype with other African trypanosomes. Trypanosome cell surfaces are coated with a Variant Surface Glycoprotein (VSG) that undergoes antigenic variation¹¹. Trypanosome genomes encode hundreds of alternative VSG, but each cell expresses just a single variant. Periodically, new variants emerge that have dynamically switched to an alternative expressed VSG¹¹. Each VSG is strongly immunogenic but confers no heterologous protection. Thus, as antibodies clear the dominant VSG clones of the parasite intra-population, serologically distinct clones replace them, rendering cognate antibodies redundant and facilitating a persistent infection¹².

Previously, we showed that *T. vivax* VSG are distinct from those in *T. brucei* or *T. congolense*. *T. vivax* VSG genes display much greater sequence divergence, and include sub-families absent in other species (named Fam23-26 inclusive¹³). In *T. brucei*, gene conversion is crucial to switching VSG genes and generating novel antigens^{14,15}. However, sequence repeats known to facilitate gene conversion

in *T. brucei* were absent from the *T. vivax* reference genome, suggesting that the *T. brucei*-based paradigm of antigenic variation might not apply there¹⁰.

75 Experiments from the pre-genomic era revealed certain enigmatic features that corroborate the distinctiveness of antigenic variation in *T. vivax* and which remain unexplained. Animals infected with *T. vivax* self-cure more often and faster compared with other species, which was attributed to antigenic exhaustion^{16,17}. Clones expressing certain VSG re-emerged late in infection after the host had developed immunity^{3,17}. Quite unlike *T. brucei* or *T. congolense*, recovered animals displayed
80 immunity to strains from very distant locations, indicating that *T. vivax* serodemes could span countries, or even the whole continent^{18,19}. Such features prompted the prediction that antigen repertoire in *T. vivax* would be smaller than in other trypanosomes³.

Here, we address these long-standing issues by characterising antigenic diversity in clinical *T. vivax*
85 isolates. We apply the data to examine VSG recombination in parasite populations and to profile VSG expression during experimental infections in a goat model. The Variant Antigen Profile (VAP) we establish for *T. vivax* shows that VSG sequence patterns in *T. vivax* are incompatible with the current, *T. brucei*-based model for antigenic variation in trypanosomes.

90 Results

Genomes of 28 *T. vivax* clinical strains isolated from seven countries were sequenced on the Illumina MiSeq platform. Genome assemblies ranged in coverage from 32.8% to 80.4%, in sequence depth from 3.5x to 78.5x, and in contiguity (N50) from 238 to 2852 (Supplementary Table 1). Using
95 sequence homology with known VSG sequences in the *T. vivax* Y486 and *T. brucei* TREU927 reference genomes, between 40 and 436 VSG genes were recovered from assembled genome

contigs; the mean average (175) is approximately one fifth of the reference genome repertoire (N=865)¹⁰.

100 ***T. vivax* variant antigen profiles reflect genealogy**

We devised a VAP for *T. vivax* VSG gene repertoire to examine antigenic diversity across strains. The four VSG-like gene sub-families (Fam23-26)¹³ in the *T. vivax* Y486 reference sequence (hereafter called 'Y486') occurred in all genomes, in similar proportions (Supplementary Fig. 1), making them unsuitable for discriminating between strains. Therefore, we produced clusters of orthologs (COGs) for all VSG-like sequences from Y486 and 28 clinical strains (N=6235), defining a COG as a group of VSG-like sequences with ≥90% sequence identity. This produced 2038 COGs, each comprising a single gene plus near-identical paralogs from multiple strains. Most COGs (78%) were cosmopolitan (i.e. present in multiple locations, see Methods), while 441 were strain-specific (Supplementary Table 2).

VAPs based on presence or absence of VSG COGs were compared to strain genealogy and geography to examine spatio-temporal variation in VSG repertoire. Fig. 1 shows that VAP-based strain relationships matched those inferred from whole genome single nucleotide polymorphisms (SNPs), and therefore, that VAP reflects both population history and location. There is a remarkable correspondence between VAPs of Ugandan strains with those from Brazil, suggesting that these Brazilian *T. vivax* were introduced into Brazil from East Africa. The correspondence of VAPs and SNPs is particularly clear when we compare the Ugandan/Brazilian profile with those in Nigeria. While clearly divergent in their VSG repertoire, there remain 769 COGs (37%) that are shared between these locations; ('TVILV-21' possesses various COGs widespread in West Africa). Thus, *T. vivax* VSG repertoires diverge in concert with the wider genome and provide a faithful record of population history, in contrast to *T. congolense*, where the opposite effect was observed²⁰.

Global *T. vivax* VSG repertoire comprises 174 phylotypes

125

The VSG gene complements in our strain genome sequences are incomplete. So, while comparing partial strain genomes in combination provides a coherent analysis of global VSG variation, the spatial distribution of COGs, and the number of truly location-specific COGs, will increase with greater sampling. This is clear when we consider that 248 COGs (12.2%) comprise a single Y486-specific sequence, which is the only strain with a complete VSG complement. Presently, a COG-based VAP will include too many false negative ‘absences’ to reliably profile individual strains.

130

A VAP that allows comparison of any two strains must be based on universal markers that also vary in the population. COGs are not universal and sub-families do not vary; so, we reasoned that a taxon of intermediate inclusivity would satisfy both criteria. Therefore, we devised another VAP based on phylotypes, each consisting of multiple, related COGs with $\geq 70\%$ sequence identity (see Methods). 174 VSG phylotypes accommodated every VSG-like sequence we observed. Fig. 2 shows the size and distribution of these across strains and emphasizes the widespread distribution of most phylotypes, 86% (149/174) of which are cosmopolitan.

135

140

Exceptions to this trend, as structurally distinct VSG sub-families restricted to specific populations, may be epidemiologically important. Among Nigerian samples, the location with the largest sample (N=11) and so the most robust presence/absence calls, five phylotypes are unique (P94, P118, P126, P170, P173). These are not recent derivations in Nigerian *T. vivax* because they are defined by a threshold sequence identity and so, are of approximately equally age. Moreover, their positions in Fig. 2 indicate no significant difference in the node connectivity of Nigeria-specific and cosmopolitan phylotypes overall. As these phylotypes comprised only one or two COGs, we extended the analysis to COGs generally.

145

150 We found 130 COGS in at least 9/11 Nigeria strains and no other location. We hypothesized that, if they were relatively recent gene duplications, they would have shorter genetic distances to their closest relatives than cosmopolitan COGs. We estimated Maximum Likelihood phylogenies for each phylotype containing a Nigerian-specific COG and inferred relative divergence times using the RelTime tool in MEGA v10.0.5²¹. This showed that there was no significant difference ($p=0.35$) in the mean divergence times for Nigeria-specific COGs ($\mu=0.038\pm0.005$; $N=83$) and cosmopolitan COGs in the same phylotype ($\mu=0.041\pm0.005$; $N=212$). Therefore, Nigerian-specific COGs and phylotypes are just as ancient as lineages with cosmopolitan distributions, and do not provide evidence for population-specific gene family expansions.

160 In summary, the incompleteness of strain genomes compelled us to adopt phylotypes as a universal but variable metric to profile *T. vivax* VSG repertoire. On this basis, *T. vivax* VSG repertoire appears to be relatively conserved continent-wide. Population variation does exist, especially at COG-level, but appears to originate through differential patterns of lineage loss rather than population-specific gene family expansions, since Nigeria-specific COGs are no younger than other VSG. This degree of continent-wide conservation is quite unlike patterns seen in *T. brucei*²². Suspecting that this indicated a more fundamental difference between African trypanosome species in how antigenic diversity evolves, we examined population variation among their VSG sequences in detail.

Minimal signature of recombination in *T. vivax* VSG sequences

170 We took multiple approaches to test the hypothesis¹⁰ that *T. vivax* VSG recombine less than *T. brucei* and *T. congolense* VSG. First, we asked if VSG sequences assort. Based on the current model of antigenic switching¹¹, VSG reads from 28 clinical strains would not remain paired after mapping to Y486 because historical recombination events would have distributed them across multiple

reference loci. Fig. 3a shows that the proportion of strain read-pairs remaining paired after mapping is significantly higher in *T. vivax* (mean=92%; N=19) relative to *T. congolense* (mean=87%; $t=3.23$; $p<0.05$) and *T. brucei* (mean=76%; $t=12.8$; $p<0.001$), and is almost as high as a negative control comprising adenylate cyclase genes (mean=97%).

Reversing this approach, we examined how Y486 VSG gene sequences mapped to strain assemblies when broken into 150 bp segments. Fig. 3b shows how the outcome of segmental mapping was defined. The mean proportion of Y486 VSG that are mosaics of strain genes (i.e. 'Multi-coupled' (MC: 25%) or 'Uncoupled' (UC: 7%)) is significantly lower than in *T. congolense* (MC: 33%; $p<0.05$ UC: 31%; $p<0.001$) and *T. brucei* (MC: 39%; $p<0.001$; UC: 12%; $p<0.001$), while the number that are essentially orthologous (i.e. 'Fully-coupled' (FC: 59%)) is significantly greater (for *T. congolense*, $p<0.001$; for *T. brucei*, $p<0.001$) (Fig. 3c). Analysis of phylogenetic incompatibility in alignments of FC and MC quartets using PHI²³ corroborates the mapping patterns. Across all species, FC VSG contain little evidence for phylogenetic incompatibility and not generally more than the adenylate cyclase control (Fig. 3d). While MC VSG display phylogenetic incompatibility, *T. vivax* MC quartets displayed this less frequently ($P_{pi}=41\%$) than in *T. congolense* ($P_{pi}=65\%$) and *T. brucei* ($P_{pi}=67\%$).

While there are fewer MC VSG in *T. vivax*, this sizeable minority might still be genuine mosaics. Alternatively, other processes such as gene paralogy or substitution rate heterogeneity could account for the signature of recombination. Hence, we explicitly modelled the history of recombination within FC or MC sequence quartets using ancestral recombination graphs (ARG) and inferred the time to most recent common ancestor (TMRCA) for each quartet. Average TMRCA was significantly greater for *T. vivax* FC VSG (0.19 ± 0.17) than either *T. congolense* (0.05 ± 0.06) or *T. brucei* (0.06 ± 0.07), indicating much deeper coalescent times for *T. vivax* VSG. More importantly, the variance in TMRCA along sequence alignments is extremely small for *T. vivax* FC VSG, showing that the whole alignment shares a common ARG (Fig. 3e). Variance is greater for MC VSG, but both MC

and FC types are significantly less variable than either other species ($p < 0.001$). Both the relatively small TMRCAs and variance in TMRCAs along alignments indicates that *T. brucei* and *T. congolense* VSGs are routinely mosaics, while the coalescence of most *T. vivax* VSGs can be modelled without recombination. Interestingly, TMRCAs are significantly higher among *T. brucei* MC VSG quartets than *T. congolense* VSG ($p < 0.001$), indicating that the former may have a higher recombination rate (explored further in Supplementary Table 3).

In summary, these analyses show that retention of orthology among VSG loci across trypanosome populations varies significantly between species. Fig. 3f plots the total pairwise orthology between strains (see Methods). Around 75% of *T. vivax* VSGs are found in multiple strains as orthologs, without evidence for recombination, compared with ~40% in *T. brucei* ($p < 0.001$) and *T. congolense* ($p < 0.001$). As the VAPs indicated, *T. vivax* VSGs typically retain orthology and essentially behave like ‘normal’ genes in the population, while *T. brucei* or *T. congolense* VSGs recombine frequently, causing loss of orthology and the appearance of strain-specific mosaics throughout the population.

Strong phylogenetic effects in VSG expression *in vivo*

Broadly conserved VSG phylotypes containing little signature of historical recombination indicate that VSG mosaics do not contribute to antigenic diversity *in vivo*. We tested this by measuring VSG transcript abundance in goats experimentally infected with *T. vivax* (strain Lins²⁴) over a 40-day period. Parasitaemia and expression profiles of VSG phylotypes in four replicates are shown in Fig. 4. We observed the expected waves of parasitaemia beginning after four days and continuing approximately every three days until termination (i.e. 6-9 parasitaemic peaks). Transcriptomes were prepared for each peak and revealed 282 different VSG transcripts across all replicates (Supplementary Table 4), which belonged to 31 different phylotypes (18% of total).

Variant antigen profiling of the expressed transcripts characterised the dominant, (but more often co-dominant), VSG phylotypes across successive peaks (Fig. 4). Somewhat contrary to expectation, persistent expression of a phylotype across peaks, e.g. P24 (Supplementary Fig. 2) and P2 (Supplementary Fig. 3), or re-emergence of a phylotype after decline, e.g. P40 (Supplementary Fig. 4) and P143 (Supplementary Fig. 5), was often seen. The identity of expressed phylotypes was partly reproduced across replicates, with 12/31 phylotypes observed in all four animals, and 19 phylotypes in three animals (Supplementary Fig. 6); on 21 occasions this extended to an identical VSG sequence, (for detail, see Supplementary Fig. 2-5).

Similarly, the order of VSG expression was partly reproducible across animals. Fig. 5 displays transcript number and abundance at early, middle and late points in the experiment, mapped on to the sequence similarity network of all phylotypes. The best example of reproducibility is the dominant expression of P24 in the middle-to-late period across all animals. Other examples include a group of phylotypes (P2, P40, P142 and P143) expressed early (i.e. peak 1/2, Fig. 5a) in A2 and A3, then re-emerging later at peak 5/6 in A1-3 (Fig. 5b), and even later in A4. For detailed analysis of phylotype abundance at each time-point see Supplementary Fig. 7. Importantly, however, while phylotypes show consistency in expression through time and across replicates, individual VSG transcripts do not. Hence, while P24 was a dominant variant antigen in every replicate, the actual P24 transcript expressed was different in each case and diverged by up to 26.5%. Further examples in Supplementary Fig. 2-5 demonstrate that this was typical.

Across all peaks, groups of related transcripts of the same phylotype were commonly co-expressed at the same peak (e.g. P2 expression comprised 3.08 ± 1.97 transcripts on average, $P24 = 2.33 \pm 1.3$, $P40 = 2.67 \pm 1.12$, $P143 = 2.71 \pm 1.25$). On three occasions, the observed phylotype comprised seven distinct transcripts (P2 at peak 5 in A1, P8 at peak 8 in A4 and P135 at peak 5 in A1). Overall, only 8/31 phylotypes were only ever represented by a single transcript. This indicates that the expressed

repertoire is determined in part by sequence homology, and Supplementary Fig. 8 shows that expressed transcripts belong to significantly fewer phylotypes than simulated transcript repertoires of the same size, confirming that they are not drawn from the available repertoire by chance. For detailed examples, see Supplementary Fig. 2-5.

An obvious feature in Fig. 5 is the concentration of highly-expressed phylotypes in the bottom-left corner of the network. A complex of closely-related Fam23 phylotypes (e.g. P2, P40, P142) were expressed early in A1 and A2 (Fig. 5a-b). This was followed by Fam23 phylotypes more centrally placed (e.g. P8), and finally, Fam25 phylotypes (e.g. P24/P44) in late infection. In A3 and A4, a similar pattern occurred, except that Fam25 VSG (i.e. P44) were expressed early, followed by the Fam23 complex and then P24. This can also be seen in Supplementary Fig. 7, where phylotypes displaying reproducible profiles across replicates are often closely related (e.g. P2, P40, P142 and P143). The connectivity of nodes representing expressed phylotypes is greater than that expected by chance. The clustering coefficient of a sub-network representing all ‘expressed’ nodes across all peaks is significantly greater than randomised sub-networks of the same size ($p < 0.05$; for detail, see Supplementary Fig. 9).

In summary, the major pattern emerging from *in vivo* expression profiles is a strong phylogenetic signal on three levels. First, the identity and order of expressed phylotypes is partly reproducible, (but expression of individual transcripts is typically not). Second, phylotypes expressed at a given peak regularly comprise multiple related, but non-identical, transcripts. Finally, at the phylotype level, related phylotypes are expressed simultaneously or consecutively, manifested as clustering in Fig. 5 and Supplementary Fig. 8. Therefore, phylogeny (or sequence identity) is an important factor in explaining VSG expression profile in *T. vivax*.

No mosaics of VSG phylotypes during experimental infections

280 Expressed *VSG* in *T. brucei* include sequence mosaics, which is interpreted as evidence for recombination of *VSG* loci during infections^{15,25,26}. In *T. brucei*, *VSG* mosaics can be formed between highly divergent donors with as little as 25% identity along their entire lengths²⁶, and can implicate relatively short recombinant tracts of ~100 bp²⁷. We analysed expressed *VSG* transcript sequence mosaics by comparing 100 bp windows of each transcript to the *T. vivax* Lins genome sequence using
285 BLASTp²⁸. Typically, mosaics would be confirmed where a single transcript displayed affinities to different *VSG* genes along its length. Unfortunately, since both *VSG* transcripts and gene sequences were often fragmentary, it was common for a transcript to have multiple affinities as no single gene sequence spanned its length. Even so, without exception, the closest related sequences in every window of each transcript were other sequences in the same phylotype.

290 With sequence affinities inconclusive, we searched for reorganisation of an expressed *VSG* sequence relative to a genomic locus by mapping all read-pairs belonging to *VSG* transcripts to the *T. vivax* Lins genome. The percentage of read-pairs that mapped to unpaired genomic positions (1.06-5.63%) was greater than the percentage arising from a random selection of 100 housekeeping genes (0.01 -
295 0.05%). However, given that *T. vivax* *VSG* are arranged in tandem gene arrays of closely-related paralogs¹⁰, we reasoned that this repetitive organisation might lead to multiple mapping of reads. Indeed, the percentage of *VSG* read-pairs split after mapping is not significantly different to that of adenylate cyclases (3.43-7.53%; $p=0.892$), which do not form mosaics but are often arranged in tandem arrays²⁹.

300 Nonetheless, the few mis-mapped reads could still derive from rare mosaic transcripts. To examine these explicitly, we aligned *VSG* transcripts with the three most similar genes from the *T. vivax* Lins genome sequence using BLASTn (where three sequences >500bp in length could be obtained; N=68) and used GARD³⁰ to identify potential recombination breakpoints. The closest matches to each

transcript were again always from the same phylotype (minimum full-length sequence identity of 86%). GARD found that 54/68 alignments displayed significant topological incongruence not attributable to rate heterogeneity, indicating 1.94 ± 1.66 breakpoints on average (ranging between 0 and 7). This might suggest that mosaicism is widespread within phylotypes, however, this degree of phylogenetic incompatibility was not significantly different to adenylate cyclases (36/48 alignments with significant topological incongruence and an average of 1.87 ± 1.88 breakpoints (ranging between 0 and 8); $p=0.39$).

In summary, while most transcript alignments contained breakpoints, these only implicated very closely related sequences, and the scale of genetic admixture is comparable with other tandemly arrayed gene families. Thus, we believe that these slight topological inconsistencies are consistent with re-arrangements (real or artefactual) caused by tandem arrangement of *T. vivax* VSG. Certainly, no transcript contained evidence for mosaics of different VSG phylotypes and therefore, assortment of *T. brucei* order was sort seen.

Discussion

The current model of trypanosome antigenic variation has recombination as the driver behind novelty and persistence. Unlike *T. brucei* and *T. congolense*, we find little evidence for VSG mosaics, either historically in the population or during experimental infections. Instead, *T. vivax* VSG repertoire comprises 174 conserved phylotypes, and incomplete sorting of these lineages produces population variation. We see now that the deep ancestry of VSG lineages and lack of VSG pseudogenes in *T. vivax*¹⁰ reflect a long history without recombination.

Experiments in the twentieth century documented the progression of Variant Antigen Types (VATs) during *T. vivax* infections^{3,16,17}. VATs represent parasite clones that confer a specific, reproducible

immunity, assumed to relate to a specific *VSG*. Our results confirm the hypothesis that emerged from these experiments, that the *T. vivax* *VSG* repertoire is smaller than those of other species^{3,16}. While the number of *VSG* genes is comparable to *T. brucei* and *T. congolense*, these provide fewer unique antigens because they are often extremely similar, expressed simultaneously, and cannot recombine. This explains several features of *T. vivax* infections, including the propensity for host self-cure¹⁶ and the re-emergence of VATs late in infection¹⁷. Furthermore, 70% of phylotypes and 45% of COGs are shared between East and West Africa respectively, which could explain the widespread distribution of serodemes, that is, why immunity to VATs in East Africa provides protection against some parasite strains from Western and Southern Africa also^{19,31}.

We have defined *VSG* phylotypes as universal but variable quantities for variant antigen profiling of any *T. vivax* strain. The evolutionary conservation of many phylotypes, and their reproducible expression patterns (in contrast with individual genes), has shown that phylotypes are not merely convenient, but have biological relevance. A crucial consideration then is how phylotypes relate to VATs. If individual transcripts in a phylotype cross-react with the same antibody, then VATs are likely to be synonymous with phylotypes; which raises the question of why multiple transcripts are expressed when this confers no benefit to parasite persistence. Conversely, if all *VSG* transcripts are serologically distinct, this poses the question of why co-expression is determined by sequence homology. Either way, the relevance of *VSG* phylogeny to antigenic variation is clear. The absence of recombination means that the mechanism of *VSG* switching in *T. vivax* must be different to the *T. brucei* model. We have seen that *VSG* expression *in vivo* displays an obvious phylogenetic signal, which might be explained if co-expressed transcripts derive from the same tandem array of *VSG* paralogs, which exist throughout the *T. vivax* genome¹⁰. If so, these structures could have a central role in a distinct switching mechanism not dependent on gene conversion.

Without recombination to create mosaic *VSG* sequences, there is a fundamental limitation on antigenic diversity in *T. vivax* and therein its capacity for immune evasion. This poses profound new questions of how *T. vivax* persists long enough to transmit, (which it evidently does very successfully). Perhaps *T. vivax* has adopted a different life strategy with respect to the transmission-virulence or invasion-persistence trade-offs that govern pathogen evolution^{32,33}. One possibility is that *T. vivax* has evolved a more acute infection strategy than other species and achieves transmission over shorter periods. Some aspects support an invasion-persistence trade-off; *T. vivax* infections (where the host survives) are typically shorter than other species^{34,35}, and some haemorrhagic strains cause an extremely acute syndrome that is also hypervirulent^{36,37}.

Furthermore, where trypanosome species have been directly compared, chronic pathologies such as reduced packed cell volume^{34,35} and humoral immunosuppression³⁸ are less severe with *T. vivax*. However, there is no evidence that *T. vivax* replicates or transmits quicker, as would be expected under a trade-off. Another possibility is that the idiosyncratic life cycle and wider vector range of *T. vivax*⁶, are an adaptation to increase transmission in the absence of long-term persistence. However, in various reports, animals that survive the initial acute *T. vivax* infection are said to develop a chronic, often asymptomatic, infection during which parasites are not visible^{39–41}, but which may cause progressive neuropathy⁴². Thus, another possibility is that *T. vivax* cause long-term, chronic infections like other species, but has an alternative mechanism for persistence. Dissemination to immune-privileged sites might allow persistence at low cell densities and *T. vivax* does disseminate to the reproductive and nervous systems, but all trypanosome species have a comparable ability for disease tropism⁴³.

In conclusion, the orthology of *VSG* phylotypes across populations, and the considerable structural divergence among them, indicates that the global *T. vivax* variant antigen repertoire has remained largely unchanged over time. Crucially, we find no evidence in *T. vivax* for the vital role that recombination, or gene conversion, has in diversifying *VSG* sequences and mediating antigenic

switching in *T. brucei*. This is a major departure from the current model of antigenic variation, indicating that *T. vivax* has a distinct mechanism of immune evasion. Antigenic diversity in *T. vivax* is finite, in a way that *T. brucei* and *T. congolense* are not; this both explains the antigenic exhaustion observed during *T. vivax* infections and poses important new questions of how infections persist under such circumstances. Possibly, the lack of adaptation for persistence, so evident in *T. brucei*, reflects a fundamentally different life strategy in *T. vivax*, with profound implications for understanding virulence and transmission of this pervasive and devastating pathogen.

Methods

Ethical Considerations

This study was conducted in accordance with the guidelines of the Brazilian College of Animal Experimentation (CONCEA), following the Brazilian law for “Procedures for the Scientific Use of Animals” (11.794/ 2008 and decree 6.899/2009). Ethical approval was obtained from the Ethical Committee to the Use of Animals (CEUA) of the Veterinary and Agrarian Sciences Faculty (FCAV) of the State University of São Paulo (Jaboticabal campus) (São Paulo, Brazil) (protocol no. 001494/18, issued on 08/02/2018). The study was also approved by the Animal Welfare and Ethical Review Body (AWERB) of the University of Liverpool (AWC0103).

Sample preparation

A panel of 25 *T. vivax*-infected blood stabilates (150 µl), representing isolates from Burkina Faso (N=5), Ivory Coast (N=3), Nigeria (N=11), Gambia (N=1), Uganda (N=4), Togo (N=1), were selected from Azizi Biorepository (<http://azizi.ilri.org/repository/>) at the International Livestock Research Institute (ILRI), and the Centre International de Recherche-Développement sur l’Elevage en zone Subhumide (CIRDES) (Supplementary Table 4). In addition, genomic DNA of three Brazilian isolates

previously described^{24,44,45} was obtained from Instituto de Ciências Biomédicas (ICB) at the University of São Paulo. For samples from ILRI and CIRDES: Red blood cells were lysed with ACK lysing buffer (Gibco, UK) and discarded by centrifugation. Cells were washed twice in 1ml MACS buffer by centrifugation (10 min, 2500 rpm). The pellet was resuspended in 100 µl lysis buffer (aqueous solution of 1 M Tris-HCl pH8.0, 0.1 mM NaCl, 10 µM EDTA, 5% SDS, 0.14 µM Proteinase K). Samples were incubated at room temperature for 1 h and DNA was extracted with magnetic Sera-Mag Speedbeads (GE Healthcare Life Sciences, UK) according to the manufacturer's protocol. For samples from ICB: DNA obtained from ICB was extracted following an ammonium acetate protocol previously described³⁸ (TvBrMi) or a traditional phenol-chloroform extraction protocol (TvBrRp).

Genome sequencing and assembly

Illumina paired-end sequencing libraries were prepared from genomic DNA using the NEBNext® Ultra™ DNA Library Prep Kit according to the manufacturer's protocol (New England Biolabs, UK) and sequenced by standard procedures on the Illumina MiSeq platform, as 150 bp (ILRI) or 250 bp (ICB and CIRDES) paired ends. For each sample, the data yield from sequencing after quality filtering was between 1.69×10^6 and 1.32×10^7 read pairs. Samples were assembled *de-novo* using Velvet 1.2.10³⁹ with a kmer of 65 (ILRI and CIRDES) or 99 (ICB). These produced assemblies with n50 between 238 and 2852 bp (median=353; mean=985). Allele frequencies were inspected to ensure samples were from single infections only (Accession number: PRJNA486085).

VSG-like sequence recovery and systematics

VSG-like nucleotide sequences were retrieved from the assembled contigs files by sequence similarity search with tBLASTx²⁸. We used a database of *T. vivax* Y486 VSG as query and a significance threshold of $p > 0.001$, contig length ≥ 100 amino acids, and sequence identity $\geq 40\%$. Additionally, we queried a database of *T. brucei* a-VSG and b-VSG sequences, using the same p -value and length thresholds, to accommodate VSG genes that might be absent from *T. vivax* Y486, i.e. the possibility

that the reference is not representative of all strains. In the event, the reference proved to be

435 representative.

VSG-like sequences were translated and clustered using OrthoFinder⁴⁶ under the default settings.

Orthofinder clustered orthologous sequences from the reference and 28 strains. In practice, these clusters of orthologs ('COGs') also included near-identical in-paralogs. Sequences in each cluster

440 were aligned using Clustalx⁴⁷ and all alignments were edited to remove overhangs and short (<100 bp) sequences. Edited alignments were refined to produce COGs with >90% average sequence identity by combining COGs that were very similar or, more frequently, subdividing Orthofinder clusters that contained several orthologous groups until the average sequence divergence was <0.05. In complex cases of large Orthofinder clusters, neighbour-joining phylogenies were estimated

445 to aid sub-division. Sequences that could not be placed with any other such that sequence divergence was <0.05 were categorized as 'unclustered', (assumed to be strain specific VSG).

With the membership of COGs determined, we reverted to the original, unedited sequences to identify the longest representative as a 'type sequence' of that COG. These were combined with the

450 original, unclustered sequences and compared with Fam23-26 VSG reference sequences using BLASTp to confirm their validity and assign a subfamily. The type sequences subdivided thus: Fam23 (967), Fam24 (539), Fam25 (345) and Fam 26 (193). Sequences found not to have a satisfactory match to Fam23-26 VSG were excluded. This process produced 760 COGs (comprising 2576 sequences) and 1278 unclustered, or 'singleton' sequences. Each type sequence and singleton was

455 compared against all others using BLASTp to establish cohorts of related COGs/singletons, which we call 'phylotypes'. A BLASTp output was used to create sequence alignments for phylotypes and to estimate neighbour-joining phylogenies for each. The membership of phylotypes was manually adjusted by removing the most divergent sequences until each met a threshold of 70% average sequence identity.

460

Note that the geographical distribution VSG COGs and phylotypes is inferred from the strains in which type sequences were detected. We define a ‘cosmopolitan’ COG or phylotype as being present in more than one location, except if these locations are Brazil and Uganda, or any combination of Ivory Coast, Togo and Burkina Faso. In both cases, we judged the *T. vivax* strains to be too close to justify these as separate populations. COGs or phylotypes found only in Brazil and Uganda are considered ‘East African’ in this study. Those found only in some combination of Ivory Coast, Togo and Burkina Faso are considered ‘West African’.

465

Variant Antigen Profiling

470

To produce VAPs for each strain, we used sequence mapping to confirm the presence or absence of individual COGs. As mapping makes use of low-coverage reads that would not otherwise be integrated into VSG sequence assemblies, this was more efficient than inspecting genome contigs for sequence homology. There was an 11% increase in the observed repertoire size (an average of 87 additional VSG) when mapping relative to BLAST. Mapping indicated that most singleton sequences were present in other strains despite the absence of assembled orthologs. Of 1279 sequences that could not be placed in a COG, only 34 (2.7%) remained location-specific after mapping. For these reasons, trimmed sequence reads were aligned to the 2038 COG type sequences, using Bowtie2⁴⁸ set to -D 20 -R 3 -N 1 -L 20. A customized Perl script was used to select entries with a match length ≥ 245 nucleotides (corresponding to a 2% error rate in a 250 bp sequencing read), mapped as proper pairs, in the correct orientation, and within the expected insert size. This list was compared to the COG database and used to produce the presence/absence binary matrix that represents the *T. vivax* VAP. VAP-based strain relationships were estimated by hierarchical clustering analysis in R, using binary distance calculation and the Ward’s minimum variance method⁴⁹, and compared to the whole-genome variation phylogeny. For phylotype-based VAPs, presence/absence and distribution data were generated by summing over all constituent VSG COGs and singletons.

475

480

485

Strain variation

To estimate strain relationships based on the whole genome, MiSeq reads were retrieved and mapped against the *T. vivax* Y486 genome using BWA mem⁵⁰, converted to BAM format, sorted and indexed with SAMtools⁵¹. Sorted BAM files were cleaned, duplicates marked and indexed with Picard (<http://broadinstitute.github.io/picard/>), and Single Nucleotide Polymorphisms (SNPs) were called and filtered with Genome Analysis Toolkit suite according to the best practice protocol for multi-sample variant calling⁵². The multi-sample VCF file obtained from GATK was converted to FASTA format using VCFtools v0.1.14⁵³ and a maximum likelihood phylogeny was estimated with PHYML⁵⁴,
495 using the GTR+Γ+I model of nucleotide substitution, following Smart Model Selection⁵⁵.

T. vivax experimental infections

Five male Saanen goats of 4 to 8 months of age, housed at the Veterinary and Agrarian Sciences Faculty (FCAV) of the State University of São Paulo (Jaboticabal campus) (São Paulo, Brazil), were infected the *T. vivax* Lins²⁴ isolate. Before inoculation, parasite stabulates cryopreserved in 8% glycerol were thawed, checked for viability under a light microscope. Each animal was inoculated intravenously with approximately 6×10^6 parasites. Animals were clinically examined daily and parasitaemia was determined by microscopy as previously described⁵⁶. Animal 2 was euthanized by anesthesia overdose on day 39 post-infection (p.i.) after showing signs of health deterioration (loss
505 of appetite, lethargy and anaemia). Xylazine chlorohydrate (0.2 mg/kg) was administered intramuscularly as pre-anesthetic medication, followed by intramuscular ketamine chlorohydrate (2 mg/kg) as anesthetic. Cardio-respiratory arrest was induced by intrathecal administration of lidocaine chlorohydrate. Remaining animals were euthanized on day 45 p.i. according to the same procedure.

510

Blood collection, RNA extraction and sequencing

At each parasitaemia peak, 4 ml of blood were collected from jugular venepuncture and centrifuged for 15 min at 13,000 x g. The buffy coat was removed into a 2.0 ml LoBind microcentrifuge tube (Eppendorf, UK), 1.5 ml of ACK Lysing buffer (Gibco, UK) added, and the mixture incubated for 15 min at room temperature to lyse leftover red blood cells. Samples were centrifuged for 15 min at 13,000 x g, washed twice in PBS, pH 8.0, snap frozen in liquid nitrogen and kept at -80 °C until RNA extraction. RNA was extracted using the RNeasy Mini Kit (Qiagen, UK) according to the manufacturer's protocol, yielding a total RNA output between 117 ng and 13 µg per sample, quantified on the NanoDrop 2000 (ThermoFisher Scientific, Brazil). Up to 1 µg of total RNA was used to prepare multiplexed cDNA libraries as described⁵⁷ using the *T. vivax* splice-leader (SL) sequence⁵⁸ as the second cDNA strand primer. For samples up to day 30 p.i., the protocol of Cuypers et al. (2017)⁵⁷ was followed exactly as described, quantified using Qubit HS dsDNA (Invitrogen, UK) and the Agilent 2100 Bioanalyzer (Agilent Technologies, UK), and sequenced at Centre of Genomic Research (Liverpool, UK) on a single lane of the HiSeq 4000 platform (Illumina Inc, USA) as 150 paired ends, producing 280M mappable reads. However, as the library insert sizes produced were longer than recommended for the HiSeq 4000 platform (Illumina Inc, USA), the protocol for samples from days 30-45 p.i. was modified. Instead of adding the indexes from the Illumina Nextera index kit, adapter-ligated, SL-selected cDNA was used as input for the NEB Ultra II FS DNA library kit (NEB, UK), which includes an initial step of DNA fragmentation. Sequencing statistics are shown in Supplementary Table 1.

Transcriptome Profiling

RNAseq reads were assembled *de-novo* using Trinity⁵⁹. Transcript abundances were estimated for each sample with kallisto⁶⁰ using Trinity pre-compiled scripts. Subsequently, transcript abundances of samples from the same animal, expressed as transcripts per million, were combined and normalized based on the weighted trimmed mean of log expression ratios (trimmed mean of M values (TMM)⁶¹). TMM normalization adjusts expression values to the library size and reduces

composition bias. TMM values were used to produce transcript expression matrices for each animal.

To recover all VSG-like sequences in the transcriptomes, a sequence similarity search was performed

540 with tBLASTx²⁸ using the *T. vivax* COG database produced above as query and a significance threshold of $E > 0.001$, contig length ≥ 150 amino acids, and sequence identity $\geq 70\%$. All retrieved VSG-like sequences were manually curated to remove spurious matches. The resulting lists of VSG transcripts were used as query in a sequence similarity search to identify VSG transcripts matching the list of COGs defined in the VAP. A threshold of $E > 0.001$, contig length greater than 50 amino

545 acids, and sequence identity $\geq 98\%$ was applied. Finally, VSG transcripts were assigned a phylotype based on sequence similarity comparison to the VSG phylotype network ($\geq 70\%$ nucleotide identity across the whole gene sequence). VSG transcript abundances were combined per phylotype, resulting in a transcript expression matrix containing the abundance of each VSG phylotype over time.

550

Recombination Analysis

Fifty previously published genomes from *T. brucei* spp.^{29,62,63} and *T. congolense*²⁰ and nineteen of the *T. vivax* genomes presented in this study were used to compare signatures of recombination across species (Supplementary Table 4). VSGs and adenylate cyclase genes were extracted from genome

555 assemblies by sequence similarity search (BLASTn²⁸) using a nucleotide identity $\geq 50\%$, length ≥ 600 nucleotides, and $E < 0.001$. VSG assortment was quantified by read mapping using Bowtie2⁴⁸. VSG read-pairs were retrieved from the genomes and mapped against reference full-length VSG to calculate the proportion of strain read-pairs remaining paired after mapping. This protocol was repeated for adenylate cyclases.

560

In the segmental mapping approach, reference VSGs were broken into 150 bp fragments and mapped against the strain VSGs to calculate the frequency of reference reads remaining paired. VSG were characterized into uncoupled, multi-coupled and fully coupled, according to the estimated

number of donors. Fully coupled *VSGs* were those with at least one donor contributing to more than
565 84% of the sequence. Multi-coupled *VSGs* were those with one or more donors contributing with
more than 1 fragment (≥ 300 bp), whereas uncoupled *VSGs* were those remaining (i.e. one or more
donors contributing with 1 fragment only (i.e. ≤ 150 bp). The reference *VSGs* that were not mapped
at least once to the strain *VSGs* were considered reference-specific variants.

570 The phylogenetic signal of MC and FC *VSGs* and adenylate cyclases was calculated using phylogenetic
incompatibility (P_{pi}) in PHI²³ and compared to the P_{pi} of for two sets of simulated data (250
replicates, 16 sequences per replicate) with and without recombination. Simulated data was
generated with NetRecodon⁶⁴, under diploid settings, a population mutation rate (θ) of 160, a
heterogeneity rate of 0.05, and an expected population size of 1000. The population recombination
575 rate (ρ) was set to 0 and 96 for the non-recombinant dataset and recombinant datasets,
respectively. Both experimental and simulated sequences were divided into sequence quartets,
aligned with Muscle⁶⁵ and iteratively parsed through PHI²³. FC, adenylate cyclase and simulated
quartets were randomly generated and parsed through PHI 100 times for statistical power. MC
quartets were compiled manually with MC *VSG* and 3 donors.

580 Total sequence orthology in each trypanosome species *VSG* repertoire was calculated as the
proportion of shared nucleotides in the total number of nucleotides of the *VSG* repertoire of a given
strain. The number of shared nucleotides was extracted from the mapping output file using
genomecov from bedtools⁶⁶.

585 Estimation of ancestral recombination graphs

Ancestral recombination graphs were reconstructed for multi-coupled and fully-coupled *VSG* quartet
alignments and adenylate cyclase control quartet alignments using the ACG software package⁶⁷. The

TMRCAs were estimated along the length of each aligned quartet at 20 bp intervals using a 100 bp wide sliding window using constant recombination rate / population size models with an MCMC length of 10,000,000, burn-in of 1,000,000 and sampling frequency of 2,500. For each individual quartet the TMRCAs along the length of the alignment were summarised by calculating the mean TMRCAs. To identify evidence of recombination, which would generate a sequence with regions of differing ancestries, the variance in TMRCAs along the alignment was calculated for each individual quartet.

References

1. Giordani, F., Morrison, L. J., Rowan, T. G., De Koning, H. P. & Barrett, M. P. The animal trypanosomiasis and their chemotherapy: A review. *Parasitology* **143**, 1862–1889 (2016).
2. Shaw, A. P. M., Cecchi, G., Wint, G. R. W., Mattioli, R. C. & Robinson, T. P. Mapping the economic benefits to livestock keepers from intervening against bovine trypanosomosis in Eastern Africa. *Prev. Vet. Med.* **113**, 197–210 (2014).
3. Gardiner, P. R. Recent Studies of the Biology of *Trypanosoma vivax*. *Adv. Parasitol.* **28**, 229–317 (1989).
4. Osório, A. L. A. R. *et al.* *Trypanosoma* (Duttonella) *vivax*: Its biology, epidemiology, pathogenesis, and introduction in the New World - A review. *Mem. Inst. Oswaldo Cruz* **103**, 1–13 (2008).
5. Morrison, L. J., Vezza, L., Rowan, T. & Hope, J. C. Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species. *Trends Parasitol.* **32**, 599–607 (2016).
6. Hoare, C. A. *The Trypanosomes of Mammals. A Zoological Monograph.* (Blackwell, 1972).

doi:10.1126/science.179.4068.60

- 615 7. Vickerman, K. & Evans, A. Studies on the ultrastructure and respiratory physiology of
Trypanosoma vivax trypomastigote stages. *Trans. R. Soc. Trop. Med. Hyg.* **68**, 45 (1974).
8. Tetley, L. & Vickerman, K. Surface ultrastructure of *Trypanosoma vivax* bloodstream forms.
Trans. R. Soc. Trop. Med. Hyg. **73**, 321 (1979).
9. Van der Ploeg, L. H., Cornelissen, a W., Barry, J. D. & Borst, P. Chromosomes of
620 kinetoplastida. *EMBO J.* **3**, 3109–3115 (1984).
10. Jackson, A. P. *et al.* Antigenic diversity is generated by distinct evolutionary mechanisms in
African trypanosome species. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3416–21 (2012).
11. Horn, D. Antigenic variation in African trypanosomes. *Mol. Biochem. Parasitol.* **195**, 123–129
(2014).
- 625 12. Mugnier, M. R., Stebbins, C. E. & Papavasiliou, F. N. Masters of Disguise: Antigenic Variation
and the VSG Coat in *Trypanosoma brucei*. *PLOS Pathog.* **12**, e1005784 (2016).
13. Jackson, A. P. *et al.* A Cell-surface Phylome for African Trypanosomes. *PLoS Negl. Trop. Dis.* **7**,
(2013).
14. Robinson, N. P., Burman, N., Melville, S. E. & Barry, J. D. Predominance of duplicative VSG
630 gene conversion in antigenic variation in African trypanosomes. *Mol. Cell. Biol.* **19**, 5839–46
(1999).
15. Hall, J. P. J., Wang, H. & Barry, J. D. Mosaic VSGs and the Scale of *Trypanosoma brucei*
Antigenic Variation. *PLoS Pathog.* **9**, e1003502 (2013).
16. Nantulya, V. M., Musoke, A. J. & Moloo, S. K. Apparent exhaustion of the variable antigen
635 repertoires of *Trypanosoma vivax* in infected cattle. *Infect. Immun.* **54**, 444–447 (1986).
17. Barry, J. D. Antigenic variation during *Trypanosoma vivax* infections of different host species.
Parasitology **92** (Pt 1), 51–65 (1986).
18. Dar, F. K., Paris, J. & Wilson, A. J. Serological studies on trypanosomiasis in east africa: IV:
Comparison of antigenic types of *Trypanosoma vivax* group organisms. *Ann. Trop. Med.*

- 640 *Parasitol.* **67**, 319–329 (1973).
19. Murray, A. K. & Clarkson, M. J. Characterization of stocks of *Trypanosoma vivax*. II. Immunological studies. *Ann. Trop. Med. Parasitol.* **76**, 283–292 (1982).
20. Silva Pereira, S. *et al.* Variant antigen repertoires in *Trypanosoma congolense* populations and experimental infections can be profiled from deep sequence data with a set of universal
645 protein motifs. *Genome Res.* **28**, 1383–1394 (2018).
21. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* msw054 (2016). doi:10.1093/molbev/msw054
22. Hutchinson, O. C. *et al.* Variant Surface Glycoprotein gene repertoires in *Trypanosoma brucei* have diverged to become strain-specific. *BMC Genomics* **8**, 234 (2007).
- 650 23. Bruen, T. C., Philippe, H. & Bryant, D. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics* **172**, 2665–2681 (2006).
24. Cadioli, F. A. *et al.* First report of *Trypanosoma vivax* outbreak in dairy cattle in São Paulo state, Brazil. *Rev. Bras. Parasitol. Vet., Jaboticabal* **21**, 118–124 (2012).
25. Mugnier, M. R., Cross, G. A. M. & Papavasiliou, F. N. The *in vivo* dynamics of antigenic
655 variation in *Trypanosoma brucei*. *Science.* **347**, 1470–1473 (2015).
26. Jayaraman, S. *et al.* Application of Long Read Sequencing To Determine Expressed Antigen Diversity in *Trypanosoma brucei* Infections . 1–29 (2018).
27. Marcello, L. & Barry, J. D. Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by
660 archive substructure. *Genome Res.* **17**, 1344–1352 (2007).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).
29. Berriman, M. *et al.* The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422 (2005).
- 665 30. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: A

- genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).
31. Dar, F. K., Paris, J. & Wilson, A. J. Serological studies on trypanosomiasis in East Africa. *Ann. Trop. Med. Parasitol.* **67**, 319–329 (1973).
32. King, A. A., Shrestha, S., Harvill, E. T. & Bjørnstad, O. N. Evolution of Acute Infections and the
670 Invasion-Persistence Trade-Off. *Am. Nat.* **173**, 446–455 (2009).
33. Alizon, S., Hurford, A., Mideo, N. & Van Baalen, M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J. Evol. Biol.* **22**, 245–259 (2009).
34. Sekoni, V. O., Saror, D. I., Njoku, C. O., Kumi-Diaka, J. & Opaluwa, G. I. Comparative haematological changes following *Trypanosoma vivax* and *T. congolense* infections in Zebu
675 bulls. *Vet. Parasitol.* **35**, 11–9 (1990).
35. Mattioli, R. C., Faye, J. A. & Büscher, P. Susceptibility of *N'Dama* cattle to experimental challenge and cross-species superchallenges with bloodstream forms of *Trypanosoma congolense* and *T. vivax*. *Vet. Parasitol.* **86**, 83–94 (1999).
36. Gardiner, P. R., Assoku, R. K. G., Whitelaw, D. D. & Murray, M. Haemorrhagic lesions resulting
680 from *Trypanosoma vivax* infection in ayrshire cattle. *Vet. Parasitol.* **31**, 187–197 (1989).
37. Williams, D. J. L., Logan-Henfrey, L. L., Authié, E., Seely, C. & Mcodimba, F. Experimental Infection with a Haemorrhage-Causing *Trypanosoma vivax* in *N'Dama* and *Boran* Cattle. *Scand. J. Immunol.* **36**, 34–36 (1992).
38. Rurangirwa, F. R., Musoke, A. J., Nantulya, V. M. & Tabel, H. Immune depression in bovine
685 trypanosomiasis: effects of acute and chronic *Trypanosoma congolense* and chronic *Trypanosoma vivax* infections on antibody response to *Brucella abortus* vaccine. *Parasite Immunol.* **5**, 267–76 (1983).
39. Maikaje, D. B., Sannusi, A., Kyewalabye, E. K. & Saror, D. I. The course of experimental *Trypanosoma vivax* infection in *Uda* sheep. *Vet. Parasitol.* **38**, 267–74 (1991).
- 690 40. Fidelis Jr, O. L. *et al.* Evaluation of clinical signs, parasitemia, hematologic and biochemical changes in cattle experimentally infected with *Trypanosoma vivax*. *Brazilian J. Vet. Parasitol.*

- 2961, 69–81 (2016).
41. Parra-Gimenez, N. & Reyna-Bello, A. Parasitological, Hematological, and Immunological Response of Experimentally Infected Sheep with Venezuelan Isolates of *Trypanosoma evansi*,
695 *Trypanosoma equiperdum*, and *Trypanosoma vivax*. *J. Parasitol. Res.* **2019**, 1–9 (2019).
42. Batista, J. S. *et al.* Infection by *Trypanosoma vivax* in goats and sheep in the Brazilian semi-arid region: From acute disease outbreak to chronic cryptic infection. *Vet. Parasitol.* **165**, 131–135 (2009).
43. Barry, J. D. African Trypanosomiasis. in *Vaccination Strategies of Tropical Diseases* (ed. Liew, 700 F. Y.) 217 (CRC Press, 1989).
44. Paiva, F. *et al.* *Trypanosoma Vivax* Em Bovinos No Pantanal Do Estado Do Mato Grosso Do Sul , Brasil: I – Acompanhamento Clínico ,. *Rev. Bras. Parasitol. Veterinária* **9**, 135–141 (2000).
45. Silva, T. M. F. *et al.* Pathogenesis of reproductive failure induced by *Trypanosoma vivax* in experimentally infected pregnant ewes. *Vet. Res.* **44**, 1–9 (2013).
- 705 46. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* (2015). doi:10.1186/s13059-015-0721-2
47. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**,
710 357–359 (2012).
49. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* (1963). doi:10.1080/01621459.1963.10500845
50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* **00**, 3 (2013).
- 715 51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome

- analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013).
doi:10.1002/0471250953.bi1110s43
- 720 53. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
54. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies:
Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
55. Lefort, V., Longueville, J.E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol. Biol.*
Evol. 4–6 (2017). doi:10.1093/molbev/msx149
- 725 56. Brener, Z. Contribuição ao estudo da terapêutica experimental da doença de Chagas.
(Universidade Federal de Minas Gerais, Belo Horizonte, 1961).
57. Cuypers, B. *et al.* Multiplexed Spliced-Leader Sequencing: A high-throughput, selective
method for RNA-seq in Trypanosomatids. *Sci. Rep.* **7**, 1–11 (2017).
58. González-Andrade, P. *et al.* Diagnosis of trypanosomatid infections: Targeting the spliced
730 leader RNA. *J. Mol. Diagnostics* **16**, 400–404 (2014).
59. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a
reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
60. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- 735 61. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression
analysis of RNA-seq data. *Genome Biol.* (2010). doi:10.1186/gb-2010-11-3-r25
62. Siström, M. *et al.* Comparative genomics reveals multiple genetic backgrounds of human
pathogenicity in the trypanosoma brucei complex. *Genome Biol. Evol.* **6**, 2811–2819 (2014).
63. Weir, W. *et al.* Population genomics reveals the origin and asexual evolution of human
740 infective trypanosomes. *Elife* **5**, e11473 (2016).
64. Arenas, M. & Posada, D. Coalescent simulation of intracodon recombination. *Genetics* **184**,
429–437 (2010).
65. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.

Nucleic Acids Res. **32**, 1792–1797 (2004).

- 745 66. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
67. O’Fallon, B. D. ACG: Rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* **14**, 40 (2013).
68. Milne, I. *et al.* TOPALi v2: A rich graphical interface for evolutionary analyses of multiple
750 alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**, 126–127 (2009).

Acknowledgements

This work was supported by grants from the Biotechnology and Biological Sciences Research Council
755 (BB/M022811/1 and BB/R021139/1), an International Veterinary Vaccinology Network (IVVN) pump-priming award, a Bill and Melinda Gates Foundation Grand Challenges Explorations award (Round 11), and the Wellcome Trust (WT206815/Z/17/Z).

Author Contributions

760 Conceived and designed the experiments: SSP, APJ. Performed the experiments: SSP, HN, MO, KN.
Analysed the data: SSP, CWD, PR, APJ. Contributed reagents/materials/analysis tools: RMA, ZB, SK, RZM, MMT, APJ. Wrote the paper: SSP, APJ. Obtained funding: SSP, MMT, RZM, APJ.

Competing Interests

765 The authors declare no competing interests.

Figure legends

Fig. 1. Variant antigen profiles of *T. vivax* clinical isolates based on presence and absence of VSG

770 **gene clusters are concordant with population history.** Genome sequence reads for 28 *T. vivax* clinical strains were mapped to 2038 VSG type sequences, representing conserved clusters of orthologs (COGs) or strains-specific sequences, to determine the distribution of each VSG. Presence (red) or absence (grey) of each VSG in each strain is indicated in the central panel. Each profile is labelled with the strain name and shaded by its geographical origin. Percentage genome coverage is shown for each strain in brackets following its label. On the left, a Maximum Likelihood phylogenetic tree estimated from a panel of 21,906 whole genome SNPs using a GTR+ Γ +I model. Branch support is provided by 100 bootstrap replicates and branches with bootstrap support >70 are shown in bold. On the right, a dendrogram relating all strains according to their observed VSG repertoire is estimated from Euclidean distances between VAPs.

780

Fig. 2. The global *T. vivax* VSG repertoire is described by 174 phylotypes. A sequence homology network in which nodes represent phylotypes. Four conserved VSG sub-families (Fam23-26¹³) are indicated by pale red back-shading. Nodes are labelled by phylotype number; node size indicates the number of COGs in each phylotype, while node colour indicates the geographical distribution of the phylotype across 28 clinical isolates. Edges represent PSI-BLAST similarity scores greater than a threshold necessary to connect all phylotypes within sub-families. Structural homology of Fam23 and Fam24 with A-type and B-type *T. brucei* VSG respectively is indicated at top left. The Fig. shows that most phylotypes are cosmopolitan in nature, being found in multiple strains and in more than two regions. A minority are strain- or location-specific phylotypes, e.g. there are 10 phylotypes specific to West Africa (i.e. Ivory Coast, Togo and Burkina Faso) and another 15 phylotypes that are

790

unique to a single location, for instance five in Nigeria (P94, P118, P126, P170, P173), three in Burkina Faso (P11, P86, P120) and two in The Gambia (P110, P124).

Fig. 3. The frequency of VSG recombination differs between African trypanosome species. a. The

795 proportion of read pairs from strain VSG remaining paired after being mapped to the reference sequence for each trypanosome genome, shaded by species. Adenylate cyclase genes (AC) were included as a negative control. **b.** The definition of fully-coupled (FC) and multi-coupled (MC) VSG sequences. Reference VSG sequences were segmented and mapped to a strain genome assembly. Where $\geq 80\%$ of pseudo-reads map to the same locus (e.g. 'Donor 1'), the gene is fully coupled. 800 Where the segments map to multiple locations (e.g. 'Donor 1-3'), the gene is multi-coupled. Example *T. brucei* VSG sequence quartets are shown after TOPALi HMM analysis⁶⁸ (see Methods). The three line graphs represent the Bayesian probabilities of three possible topologies for a quartet phylogeny. A FC VSG displays the same topology along its whole length. A MC VSG displays different phylogenetic signals along its length, dependent on the identity of the sequence donor. **c.** A 805 comparison of the proportions of FC, MC, uncoupled (UC) and unmapped (UM) VSG in each trypanosome species. The median value is shown as a black bar. Statistical significance of differences in the mean are indicated by stars (independent t-test, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). **d.** Phylogenetic incompatibility among VSG genes using Phi²³. The proportion of FC and MC VSG quartet alignments showing significant phylogenetic incompatibility (P_{pi}) in MC and FC VSGs is shown, 810 shaded by species. Observed P_{pi} values for simulated sequences generated by NetRecodon⁶⁴, either with recombination ($R=2e^{-05}$) or without ($R=0$), are indicated by dashed lines. **e.** Variation in the 'time to most recent common ancestor' (TMCRA) along MC and FC VSG quartet alignments, estimated from ancestral recombination graphs constructed by ACG⁶⁷. The median value is shown as a black bar. **f.** Total sequence orthology among VSG repertoires in each species. Orthology was calculated as

815 the proportion of *VSG* base pairs fully coupled between each strain genome sequence and the reference. Number of strain genomes is shown in brackets.

Fig. 4. *VSG* phylotype expression during experimental *T. vivax* Lins infections in a goat model

(**N=4**). Parasitaemia (black line) is shown in the upper graph. Parasite RNA was isolated at peaks in
820 parasitaemia, indicated as black dots. The number of unique *VSG* transcripts (red line) observed in each transcriptome are plotted on the same axis. The lower line graph shows the combined transcript abundance for each *VSG* phylotype (shaded according to key) through the experiment (days post infection) for four replicates animals (1-4 from top to bottom). Note that phylotypes can comprise several, distinct transcripts of variable abundance. Across all peaks in all animals, a
825 phylotype was represented by a single transcript in 105/196 observations, (average=1.88±1.26 SD). However, across the 31 expressed phylotypes, only eight (P3, P13, P14, P16, P38, P141, P151 and P178) occur as single transcripts on every occasion when they were observed. Thus, while a slight majority of phylotypes are represented by only one transcript at a given peak, most phylotypes are present as multiple transcripts at some point. Phylotypes that were dominant (i.e. superabundant)
830 are labelled adjacent to the pertinent lines. A superabundant *VSG* was defined as having an expression level at least 10 times that of the next most abundant *VSG*, and this was observed at 15/28 peaks. For example, P24 is 128 times more abundant than P44 at peak 5 in A1, and P1 is 32 times more abundant than P155 at peak 7. The classical expectation of *VSG* expression is that a peak will be defined by a single superabundant *VSG* like this; often, however, several co-dominant *VSG*
835 phylotypes occurred with comparable expression levels, for example at peak 1 in A1 and A2.

Fig. 5. Expression of *VSG* phylotypes in the context of sequence similarity. Combined transcript abundance for expressed phylotypes are plotted on to the phylotype sequence similarity network at
a. early (Peak 1), **b.** middle (peaks 4-7), and **c.** late (last peak) infection stages respectively. Data from

840 four replicate animals are shown (A1-A4 from top to bottom). Nodes represent phylotypes and are labelled by phylotype number. Node size indicates the number of unique expressed transcripts, while node shade indicates the combined transcript abundance ($\log_2$ CPM). The classical expectation of VSG expression is that a dominant VSG should subside in abundance and disappear as the host acquires antibody-mediated immunity. However, phylotypes were seen to persist across peaks

845 and/or re-emerge later in the experiment; for instance, P40, P24 and P33 are present at all three time-points in A1, A2 and A3 respectively. Similarly, P2 is expressed strongly at the beginning and re-emerges at the end of infections in A1 and A2. Likewise, P44 is expressed at both the beginning and end of infection in A4. Since only three time-points are shown, it should be noted that these phylotypes were not present at all peaks, so this could represent re-emergence rather than

850 persistence. In cases where sufficient nodes were expressed, the clustering coefficient (C) for their sub-network was calculated. This observed value was compared to mean average C for 100 randomized sub-networks of the same size. The ratio of the observed and expected (by chance) clustering coefficients for expressed sub-networks is shown where a calculation was possible. This value typically exceeds one showing that expressed nodes cluster more than random selections.

855 When considered over all peaks, the clustering coefficient of expressed nodes is significantly higher than coefficients of randomised sub-networks of the same size (see Supplementary Fig. 9 for further details).

Figure 1

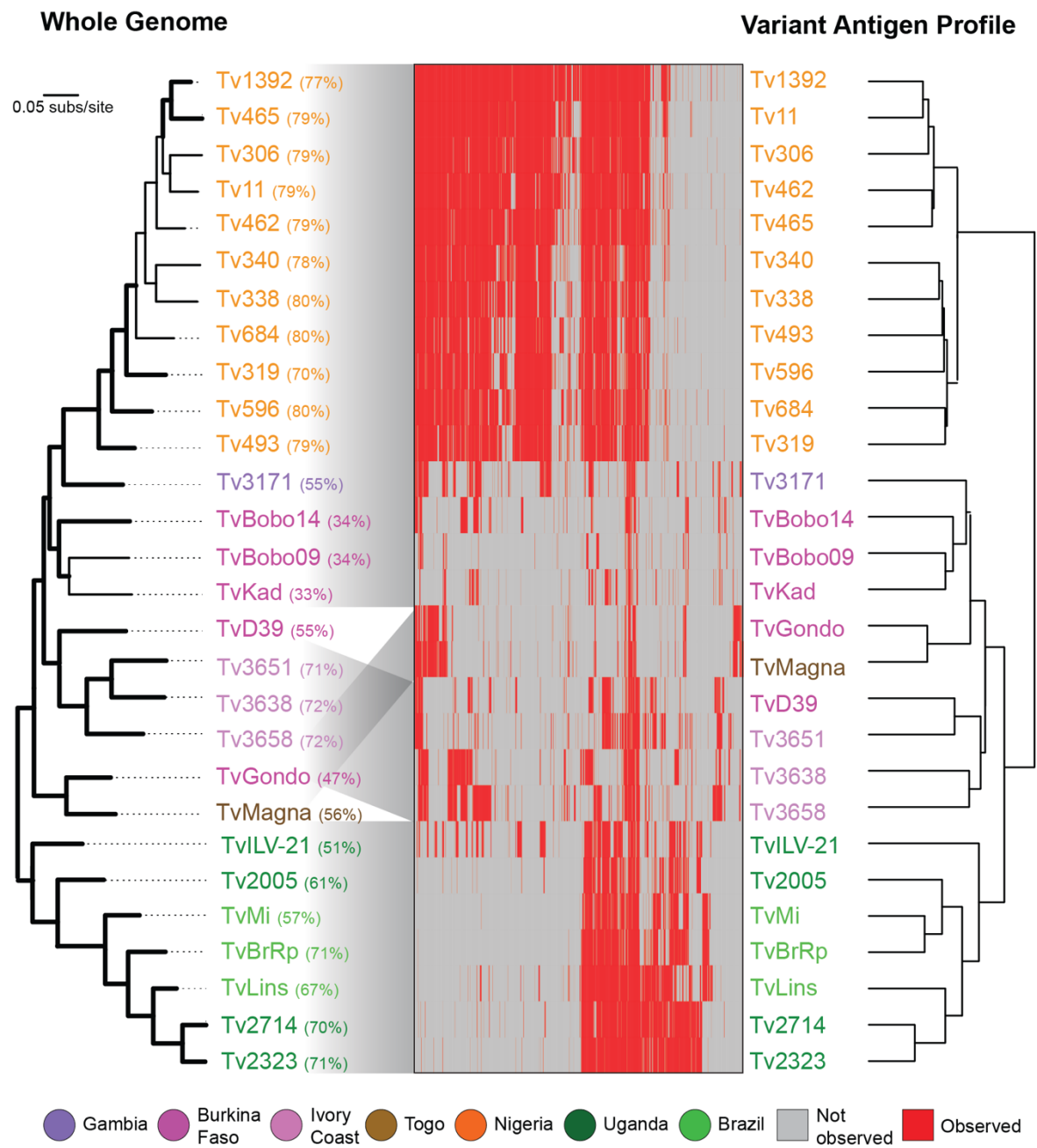


Figure 4

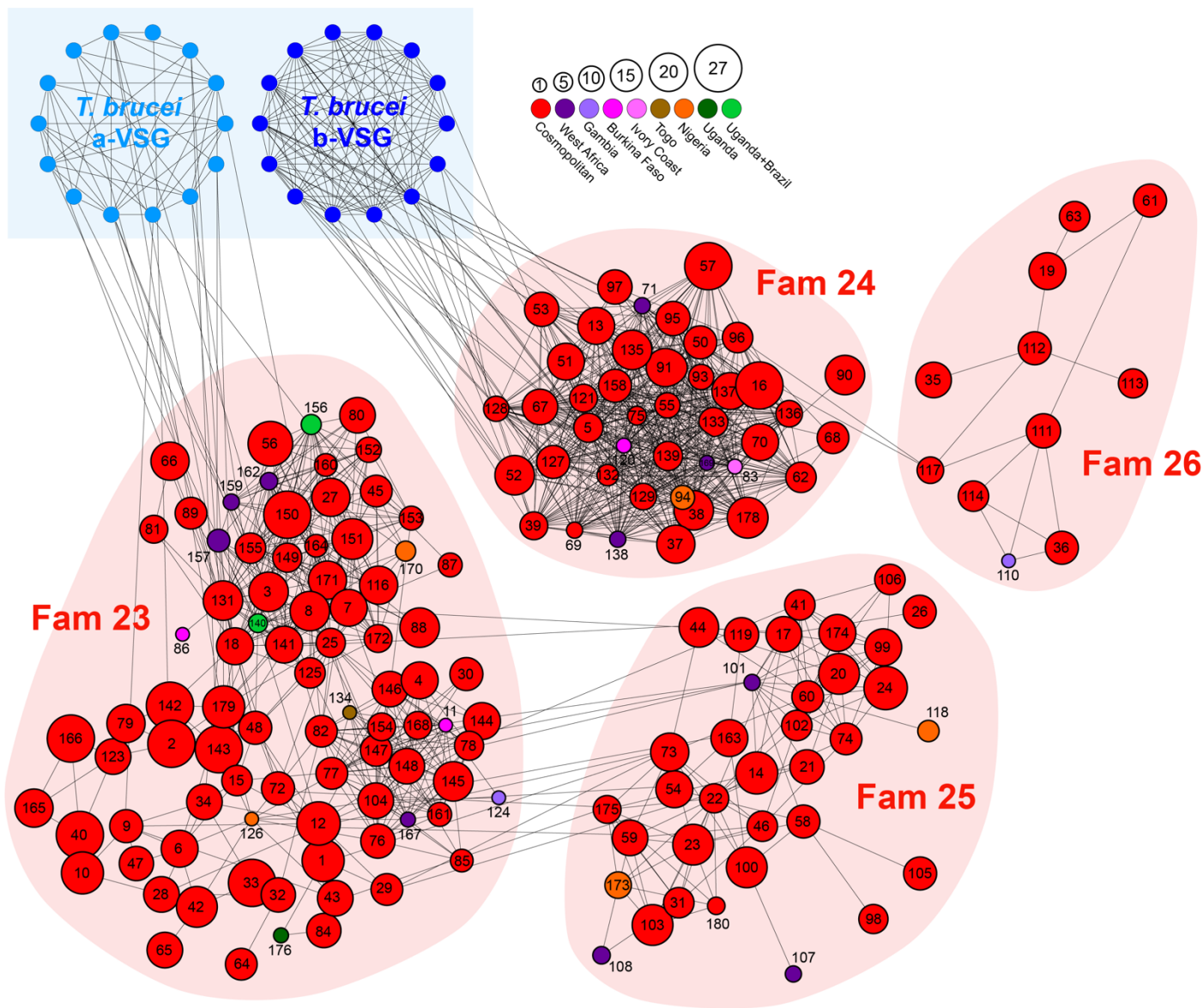


Figure 3

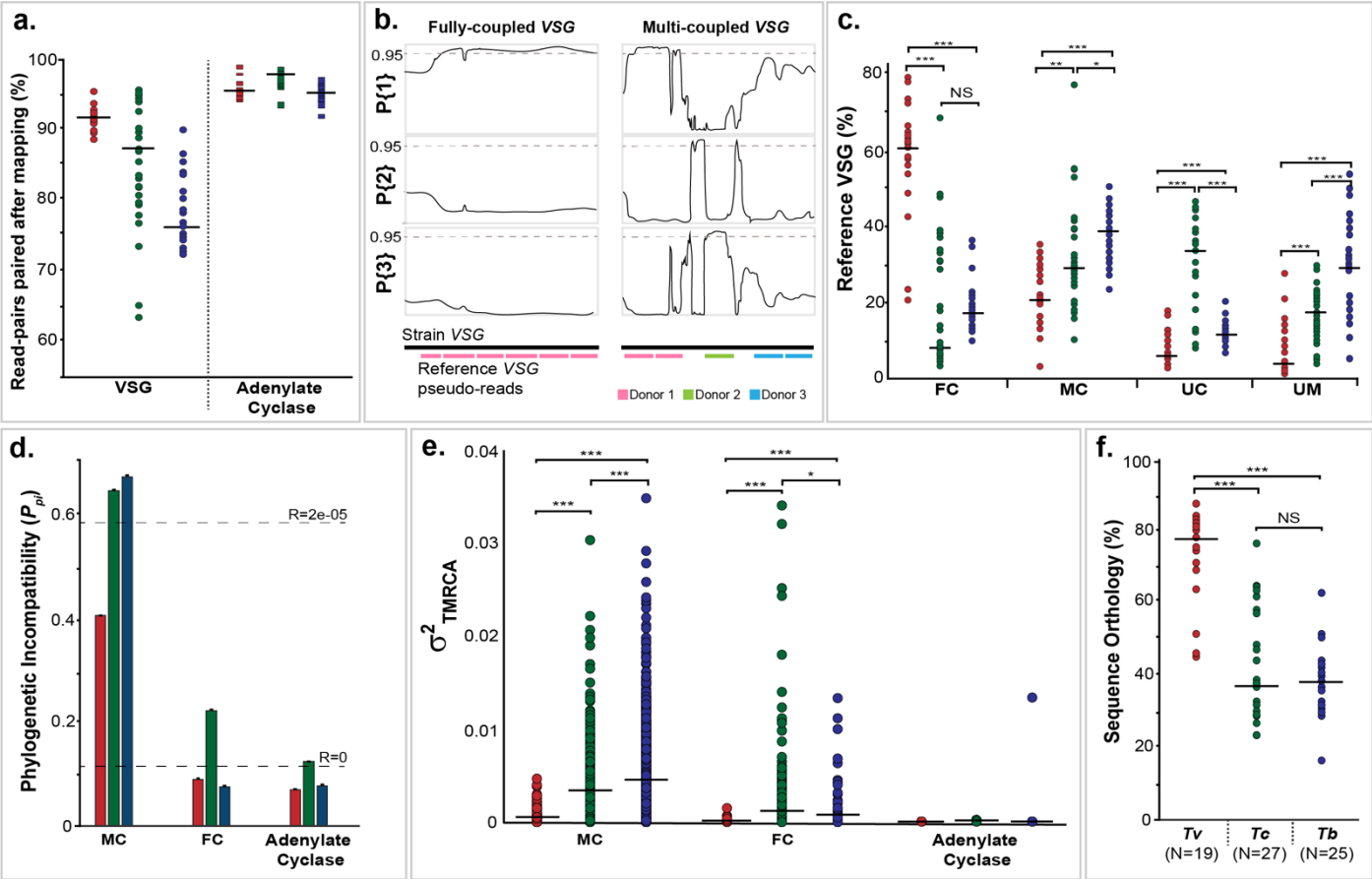
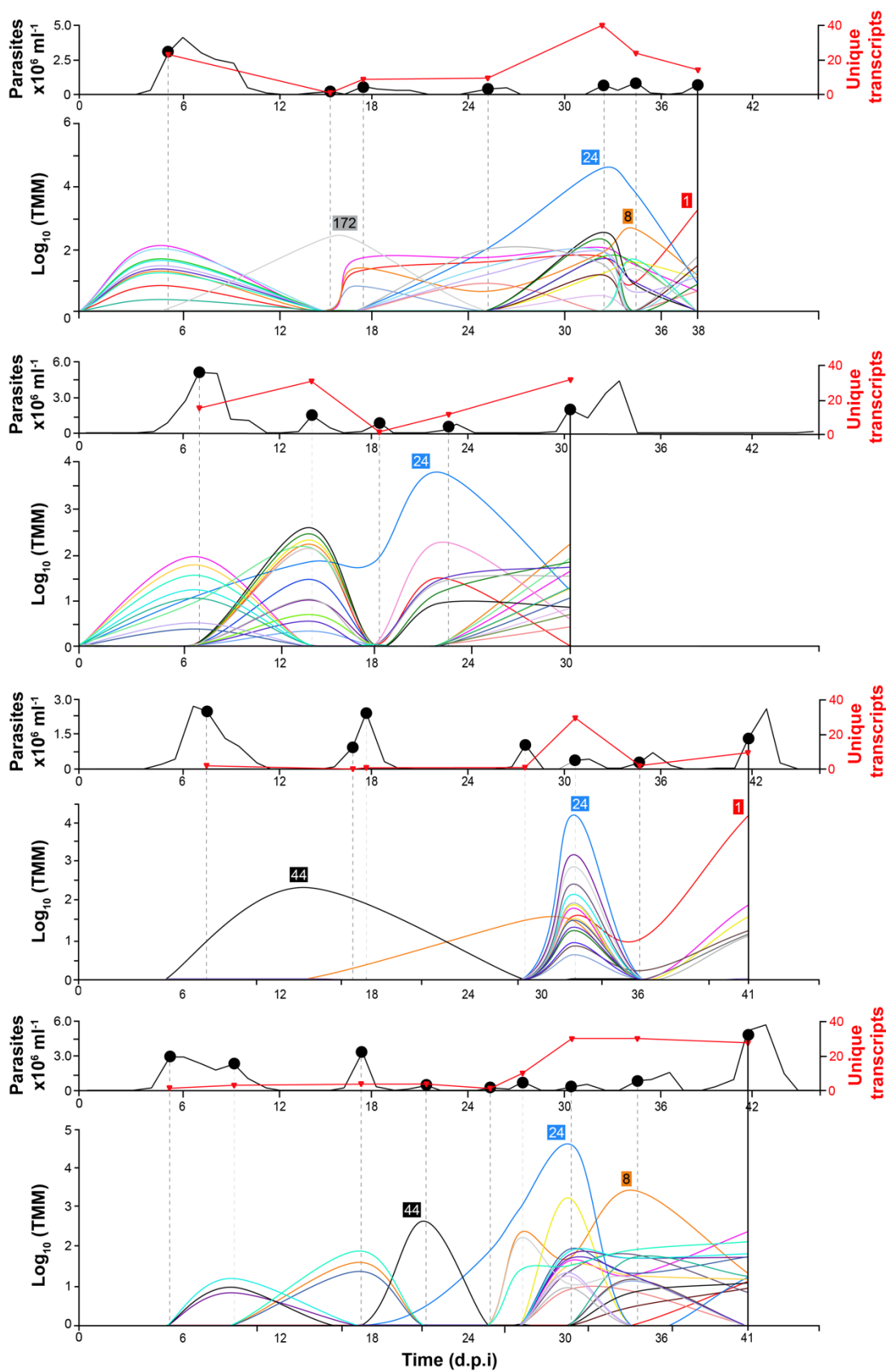


Figure 4



P1	P13	P20	P33	P44	P141	P155	P172
P2	P14	P23	P37	P87	P142	P165	P178
P3	P16	P24	P38	P123	P143	P166	P179
P8	P18	P27	P40	P135	P151	P171	

Figure 5

bioRxiv preprint doi: <https://doi.org/10.1101/733998>; this version posted August 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

