a.p.jackson@liv.ac.uk *

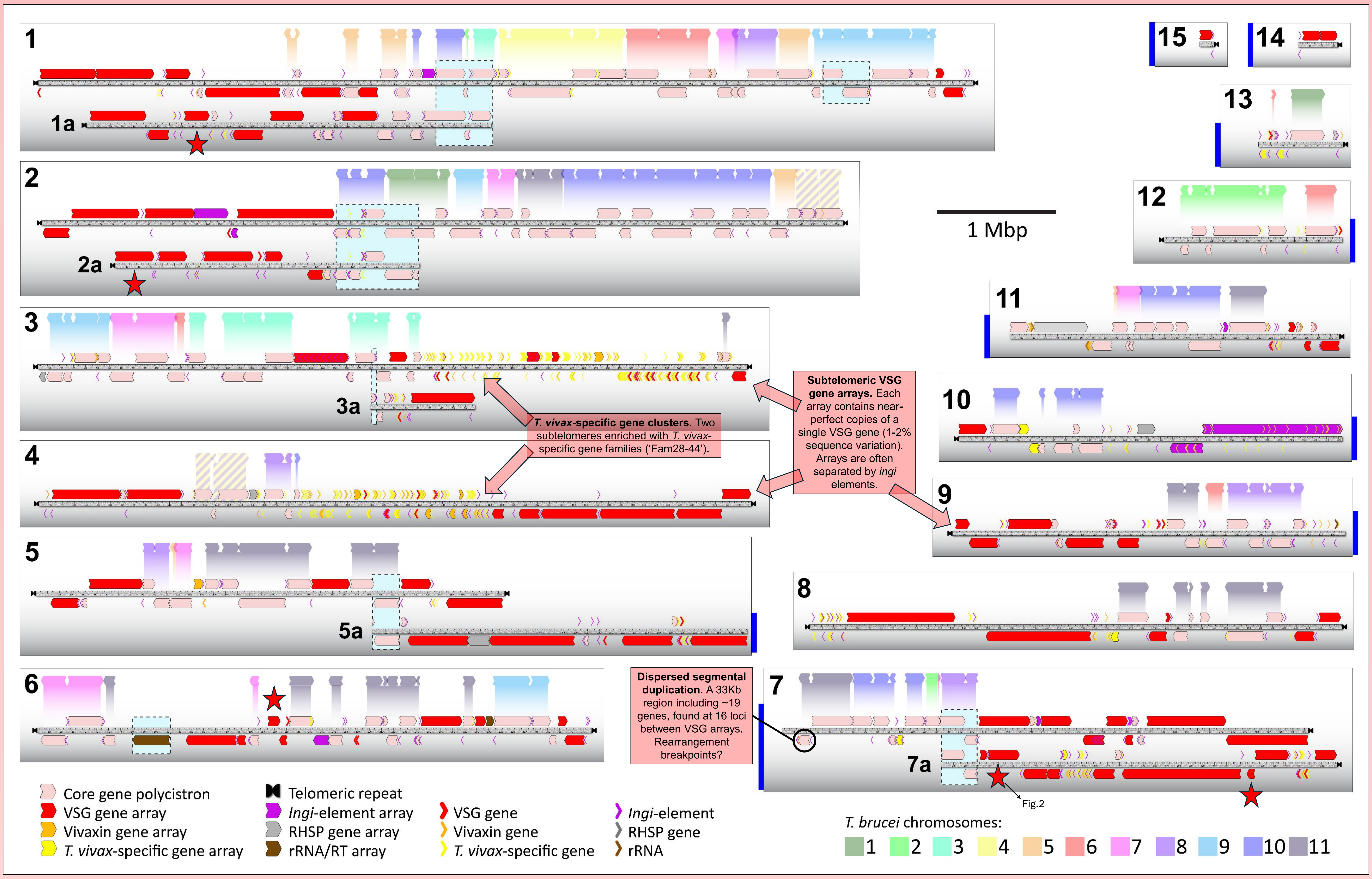
Species-specific features of the *Trypanosoma vivax* genome

#

Andrew P. Jackson*, Rachel Young#, Mark Whitehead*, Pieter Steketee#, Alistair Darby*, Liam Morrison#

Abstract. *Trypanosoma vivax* causes animal trypanosomiasis across Africa and South America. An accurate genome sequence is essential for functional genetics and the design of novel diagnostics, drugs and vaccines. Our aim is to improve the existing *T. vivax* Y486 reference genome assembly, which was scaffolded against the *T. brucei* genome assembly, to understand the species-specific features. We present a new core genome assembly in which many chromosomes are complete and articulated with sub-telomeres. As with *T. brucei*, *T. vivax* Variant Surface Glycoprotein (VSG) genes are arranged in sub-telomeric arrays, but that these are near-perfect repeats of often intact genes, (unlike *T. brucei*), and we find no evidence for a VSG expression site at telomeres or elsewhere. *T. vivax* Y486 VSG transcripts/peptides expressed in vivo did not map to telomeric loci and most VSG arrays were not expressed, consistent with allelic exclusion. However, within active VSG arrays, multiple tandem genes are expressed simultaneously. This new genome assembly shows how the *T. brucei* genome sequence is a poor guide for *T. vivax* and instead provides an accurate platform for future *T. vivax* research.

Genome assembly. PacBio Hifi data scaffolded with HiC data. 23 scaffolds include 7 complete chromosomes. Dashed blue boxes show where contiguity has been inferred manually from sequence overlap. Another 53 orphan contigs of <2kb each (not shown) contain additional VSG but no core coding sequences. Haploid genome length: 69,864,245bp.
GenBank bioproject: PRJEB108666



Subtelomeric VSG gene arrays. Each array contains near-perfect copies of a single VSG gene (1-2% sequence variation). Arrays are often separated by *ingi* elements.

***T. vivax*-specific gene clusters.** Two subtelomeres enriched with *T. vivax*-specific gene families ('Fam28-44').

Dispersed segmental duplication. A 33Kb region including ~19 genes, found at 16 loci between VSG arrays. Rearrangement breakpoints?

Figure 1. A *Trypanosoma vivax* Y486 hybrid Hifi-HiC genome assembly. Chromosomal scaffolds are represented by scaled grey bars. Polycistronic transcription units (pink) and other features are shown by coloured arrows above and below, indicating their DNA strand. Scaffolds 1-6 & 8 are complete chromosomes, often with alternative subtelomeres (labelled 'a'). Scaffolds 7,9-15 are partial chromosomes; vertical blue bars show where gaps are unclosed. Conserved synteny with *T. brucei* chromosomes (colour-coded) is indicated above each scaffold.

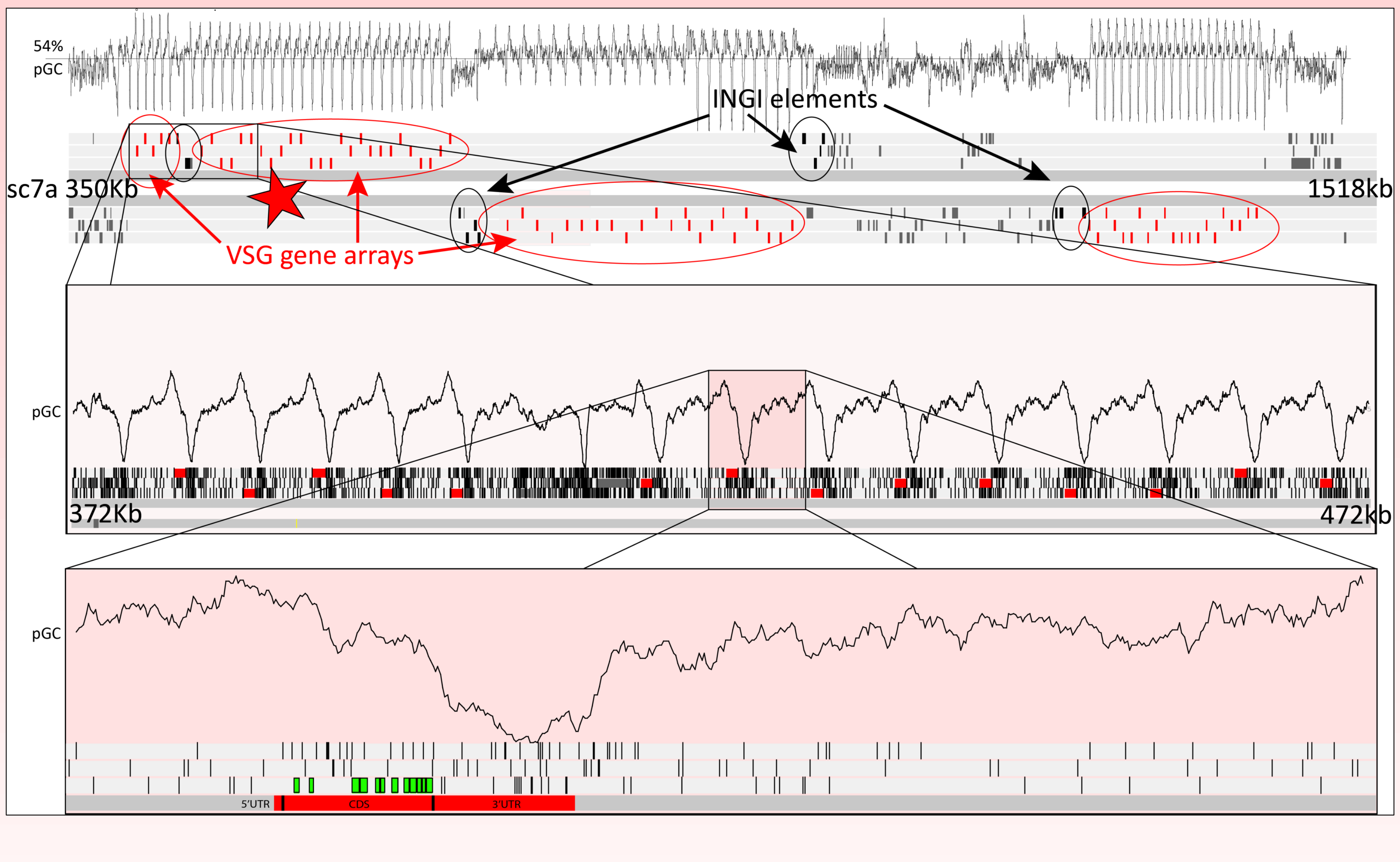


Figure 2 (right). VSG expression from subtelomeric gene arrays. Red stars in Fig.1 show the genomic locations of VSG expressed by *T. vivax* Y486, detected by proteomic analysis of a single mouse bloodstream infection. The locus shown depicts an array on sc.7a that encodes the most abundant VSG (green boxes = observed peptides). Transcriptomic analysis indicates that all array copies are transcribed. Unique peptides were detected from at least three copies.

Financial support: University of Liverpool