Single-cell transcriptomics reveals expression profiles of *Trypanosoma brucei* sexual stages

Virginia M. Howick[1,2,3*], Lori Peacock [4,5], Chris Kay [4], Clare Collett [4$], Wendy Gibson [4#], Mara K.N. Lawniczak [3#]

[1] Institute of Biodiversity, Animal Health, and Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK
[2] Wellcome Centre for Integrative Parasitology, University of Glasgow, Glasgow G12 8QQ, UK
[3] Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[4] School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK
[5] Bristol Veterinary School, University of Bristol, Langford BS40 5DU, UK
[$] Present address: UK Health Security Agency, Porton Down, Salisbury SP4 0JG, UK

[#] These authors contributed equally to this work
[*] Correspondence: VMH, Virginia.Howick@glasgow.ac.uk

Running head: Single-cell transcriptomics of trypanosomes

Abstract

Early diverging lineages such as trypanosomes can provide clues to the evolution of sexual reproduction in eukaryotes. In *Trypanosoma brucei*, the pathogen that causes Human African Trypanosomiasis, sexual reproduction occurs in the salivary glands of the insect host, but analysis of the molecular signatures that define these sexual forms is complicated because they mingle with more numerous, mitotically-dividing developmental stages. We used single-cell RNA-sequencing (scRNAseq) to profile 388 individual trypanosomes from midgut, proventriculus, and salivary glands of infected tsetse flies allowing us to identify tissue-specific cell types. Further investigation of salivary gland parasite transcriptomes revealed fine-scale changes in gene expression over a developmental progression from putative sexual forms through metacyclics expressing variant surface glycoprotein genes. The cluster of cells potentially containing sexual forms was characterized by high level transcription of the gamete fusion protein HAP2, together with an array of surface proteins and several genes of unknown function. We linked these expression patterns to distinct morphological forms using immunofluorescence assays and reporter gene expression to demonstrate that the kinetoplastid-conserved gene Tb927.10.12080 is exclusively expressed at high levels by meiotic intermediates and gametes. We speculate that this protein, currently of unknown function, plays a role in gamete formation and/or fusion.

Keywords: African trypanosomes; single cell RNA sequencing; transcriptomics; tsetse fly; *Glossina*

Introduction

The African tsetse-transmitted trypanosomes are single-celled parasites that cause human and animal diseases, which are a heavy burden for many countries in sub-Saharan Africa. These trypanosomes survive in both the tsetse and mammalian host by taking on distinct morphological forms that suit the diverse metabolic and immune environments they encounter [1]. When blood infected with *Trypanosoma brucei* is imbibed by the tsetse fly (genus *Glossina*), trypanosome blood stream forms (BSF) rapidly change their transcriptional profile, including switching off Variant Surface Glycoprotein (VSG) transcription and upregulating expression of other surface proteins such as procyclins [2,3] They also switch their metabolism from dependence on glucose processed via glycolysis in the glycosome to exploitation of amino acids such as proline via the mitochondrial TCA cycle [4]. Trypanosomes then multiply as procyclics in the fly midgut before migrating anteriorly, first colonising the proventriculus or cardia, the valve between the foregut and anterior midgut, and then the paired salivary glands (**Figure 1A**) [5,6]. Here trypanosomes attach and proliferate as epimastigotes characterised by BARP surface proteins [7], before final differentiation into infective metacyclics that are inoculated into a new host via the saliva.

Additionally, the salivary glands are the location of the non-obligatory sexual cycle of *T. brucei*, which involves meiosis and the production of haploid gametes [8,9]. As trypanosomes are early diverging eukaryotes, their sexual processes are of particular interest because they provide insights into the evolution of sexual reproduction and meiosis. Although the morphologies of the meiotic division stages and gametes have been described [8–10], little is known about the transcriptional dynamics that characterise the sexual stages because these cells are a minority of the heterogenous cell population in the salivary glands. Sexual stages are found during the early phase of establishment of salivary gland infection, with numbers peaking about three weeks after fly infection [8,9]. The sexual cycle appears to be a sideshow in the normal mitotic developmental program, as it occurs in clonal trypanosome lines and does not need to be triggered by external factors such as the presence of another strain.

Single-cell RNAseq opens the door to study heterogenous populations of single-celled parasites by delineating expression patterns of individual cells allowing us to understand continuous developmental processes, cell-type specific patterns of co-expression and bet-hedging strategies [11–17]. Recent studies have profiled *T. brucei* populations using single-cell droplet-based approaches from BSF culture to profile the development of stumpy forms [18] as well as *in vivo* salivary gland parasites in order to identify a potential vaccine candidate among mature metacyclics [19] and the dynamics of *VSG* expression in the developing metacyclic parasites [20]. These studies complement the previous bulk transcriptomic studies in *T. brucei* that identified the major changes in transcriptional patterns over time and metacyclic development using either whole infected salivary glands or *in vitro*-derived metacyclics from the RBP 6-inducible system [21–28]. Although these studies have been essential to our understanding of the dynamics of gene expression in *T. brucei*, we still lack an understanding of the molecular processing that characterize meiosis and sexual development in kinetoplastids.

Here we have exploited scRNAseq to investigate transcriptomes of the sexual stages of *T. brucei* that occur transiently in the heterogeneous trypanosome population in the fly salivary glands. We used a modified Smart-seq2 protocol [12,13,29] to profile *T. brucei* cells from

different tsetse tissues (midgut, proventriculus, salivary glands) at different time points during development. We tied the observed transcriptomic profiles to specific developmental stages, validated by immunofluorescence, and identified cell-type specific markers, which revealed the dynamics of surface protein expression as well as a new candidate gene that may be involved in sexual development.

Results and Discussion

Generation of high-quality transcriptomes from *in vitro* procyclic forms

To confirm that the modified Smart-seq2 protocol produces high-quality data for *T. brucei*, we initially profiled 46 single-cell transcriptomes from *in vitro* procyclic forms. We found a mean of $2.6x10^6$ mapped reads per cell and a mean detection of 1756 genes per cell **(Figure S1A, B)**, which is a greater number of genes per cell than recently published data from Hutchinson et al 2021 [20] that used a droplet-based method on the same parasite stage (1258 genes per cell). This further supports the use of Smart-seq2 to get in depth transcriptomes (high-coverage and full-length) in a low-throughput, targeted fashion compared to droplet-based methods that have fewer genes detected but are higher throughput [12,30]. Additionally, we observed high expression of genes that encode known procyclic surface antigens including GPEET and EP1-3 [2,31] **(Figure S1C).** These data support the utility of our protocol to profile single-cell expression profiles in kinetoplastids.

Transcriptomes of fly developmental stages

Having confirmed that the modified Smart-seq2 protocol would produce high-quality data from *T. brucei in vitro* procyclic forms**,** we profiled parasite transcriptomes isolated from diverse tsetse tissues at different timepoints, and from two *T. brucei* strains as outlined in **Figure 1A**. After quality control **(Figure S2)**, we obtained a total of 388 single-cell parasite transcriptomes: 78 from the midgut, 34 from the proventriculus and 276 from the salivary glands **(Figure 1A).** Parasites from the midgut and proventriculus were isolated from flies infected with *T. brucei* strain1738 dissected day 21 post infection (pi); 53 parasites from the SG were also derived from these flies. A further 62 1738 parasites from the salivary glands were collected on day 40 pi; these cells were derived from free (spill-out from tissue) or attached (enzymatically disassociated) cell populations to capture populations dominated by metacyclics or attached epimastigotes and premetacyclics, respectively **(Figure 1A)**. A further experiment dissected on day 24 pi aimed to analyse salivary gland parasites from an experimental cross between strains 1738 and J10; 86 cells were from single infections of 1738 (38) or J10 (48), and 75 were from the experimental cross. We additionally used two cell preservation methods in the collection of these data to allow for more flexibility in processing time. Although small differences were observed in the number of genes detected between preservation methods, this was confounded by timepoint, preventing us from fully understanding the impact of the preservation techniques alone (**Figure S3**).

To understand transcriptional variation at the single-cell level across tissues, strains, and time, we performed dimensionality reduction with all 388 cells using UMAP **(Figure 1B).** We observed that cells grouped by their tissue of origin, with clusters representing midgut and proventriculus trypanosomes and two groups of salivary gland parasites that we hypothesized could represent different cell-types or stages **(Figure 1B).** This idea was supported by the

distribution of time points across the two salivary gland groups with 21- and 24-day pi cells distributed throughout the two groups, while the 40-day pi cells (squares) occupied the centre and right-hand area; the left-hand area therefore represents early salivary gland developmental stages, such as sexual stages, which are frequent at 21 – 24 days pi, but are relatively scarce at 40 days pi compared to epimastigotes and metacyclics (Peacock et al, 2014).

We next used consensus clustering to partition the cells into six clusters based on the top average silhouette score in SC3 [32] **(Figure 1C)**. The midgut and proventriculus parasites each formed a single cluster (clusters C1 and C2 respectively), whereas the salivary gland cells divided into four clusters (C3-C6). We identified 238 marker genes across the six clusters allowing us to assign potential cell types (AUROC > 0.75, $p < 0.01$**, Table S1, Figure 1D)**. C1 and C5 had very few significant marker genes and C6 had none (hence not included in **Figure 1D**). This is likely due to different overall levels of transcriptional activity across the cell types as we observed fewer genes per cell for these clusters **(Table S1)**. For further characterization, we additionally identified the top 200 genes expressed in each cluster **(Table S2).** Based on both the cluster markers and top genes of each cluster, we were able to assign putative cell-types. C1 and C2 showed expression patterns consistent with midgut procyclics and proventricular forms, respectively, based on known marker genes and bulk transcriptomic data [21,33] **(Table S2).** C1 was characterized by a single marker gene, *FHc* (Tb927.3.4500), a fumarate hydratase, which catalyses conversion of fumarate to malate in the TCA cycle **(Figure 1)** [4,34]. *FHc* was also the most significant marker gene for the midgut forms when integrated with *T. brucei* single-cell data from [20], supporting the cell-type assignment across datasets **(Figure S4, Table S3)**. C1 also expressed several genes encoding surface proteins at high levels (*EP1*: Tb927.10.10260, *EP3-2:* Tb927.6.520, and *EP2:* Tb927.10.10250), as well as three Proteins Associated with Differentiation (*PAD1:* Tb927.7.5930, *PAD2:* Tb927.7.5940, *PAD7:* Tb927.7.5990) **(Table S2)**, which are implicated as sensors of environmental stimuli and trigger differentiation [35]. C2 had several marker genes associated with transport (e.g. amino acid transporter *AATP11* and purine nucleotide transporter *NT10*, **Figure 1**), both also highly expressed in C1 cells **(Table S2)**. Procyclin *EP2* was identified as a marker gene for this cluster, though both *EP1* and *EP3* (Tb927.6.480) were also highly expressed **(Table S2)**.

Cluster C3 comprised day 21/24 pi early salivary gland developmental stages (**Figure 1B, C**) including potential sexual forms. Notably, we observed high expression of the gamete fusion protein HAP2 (Tb927.10.10770), which is known to be expressed in meiotic intermediates and gametes **(Table S1)** [10]. An analogous cluster was also identified by [20], which showed high expression of *HAP2* and *HOP1* (Tb927.10.5490), a meiosis-specific protein. Integration across these two datasets showed the cells from both studies clustered together at a granular level **(Figure S4, Table S3**). Other notable marker genes for this cluster were two leucine-rich repeat (LRR) protein genes (Tb927.9.14570, Tb927.7.7180), with a third also highly expressed **(Table S2)**; as LRRs are protein recognition motifs, this could be significant in gamete interactions. Procyclin genes *EP1, EP3* and *GPEET* (Tb927.6.510) **(Table S2)** were highly expressed, together with *BARP* genes, which is the characteristic surface protein of epimastigotes [7].

Several *BARP* genes were the prominent marker transcripts in cluster C4 **(Figure 1D)**, identifying this cluster as salivary gland epimastigotes, consistent with previous studies

[19,20]. The majority of the later time point (day 40 pi) cells were found in clusters C5 and C6 (**Figure 1B, C**), showing that these clusters represent the later salivary gland developmental stages including mature metacyclics and/or their immediate precursors, pre- and nascent metacyclics. However, the identities of these clusters remain unclear. C5 had only three significant marker genes, two of which encode zinc finger proteins and one hypothetical protein, and C6 none (**Figure 1D, Table S2**). The zinc finger protein genes, *ZC3H11* (Tb927.5.810) and *ZC3H45* (Tb927.11.8470) were also identified as biomarkers of pre-metacyclics (Meta 1) in Vigneron et al 2020 [19], and ZC3H11 was also identified as a highly expressed gene in purified, culture-derived mature metacyclics [27]. Several other highly expressed genes of mature metacyclics ([27]) were also identified in our data for either C5 or C6 (e.g. *ZFP2*, *HSP 110*; **Table S2**). It is noteworthy that cluster C5 is predominantly J10, while cluster C6 is predominantly 1738 (**Figure 1D**), and hence these clusters may also represent strain-specific rather than stage-specific expression differences. Additionally, mature and nascent metacyclics both have VSG on the surface [36], and it is reasonable to suppose that pre-metacyclics already transcribe *VSG*s. In support of this, both previous scRNAseq studies found high levels of expression of *VSG*s in cells identified as pre-metacyclics (Meta 1, Vigneron et al, 2020; Pre-metacyclic, Hutchinson et al, 2021). Here, *VSG*s were not identified as abundant transcripts in either C5 or C6, likely because of poor match between 1738/J10 *VSG* transcripts and the Tb927 reference genome, as *VSG* repertoires are strain-specific. We next endeavoured to identify the metacyclic *VSG* repertoire of 1738 and J10, in order to confirm the cell-types of C5 and C6.

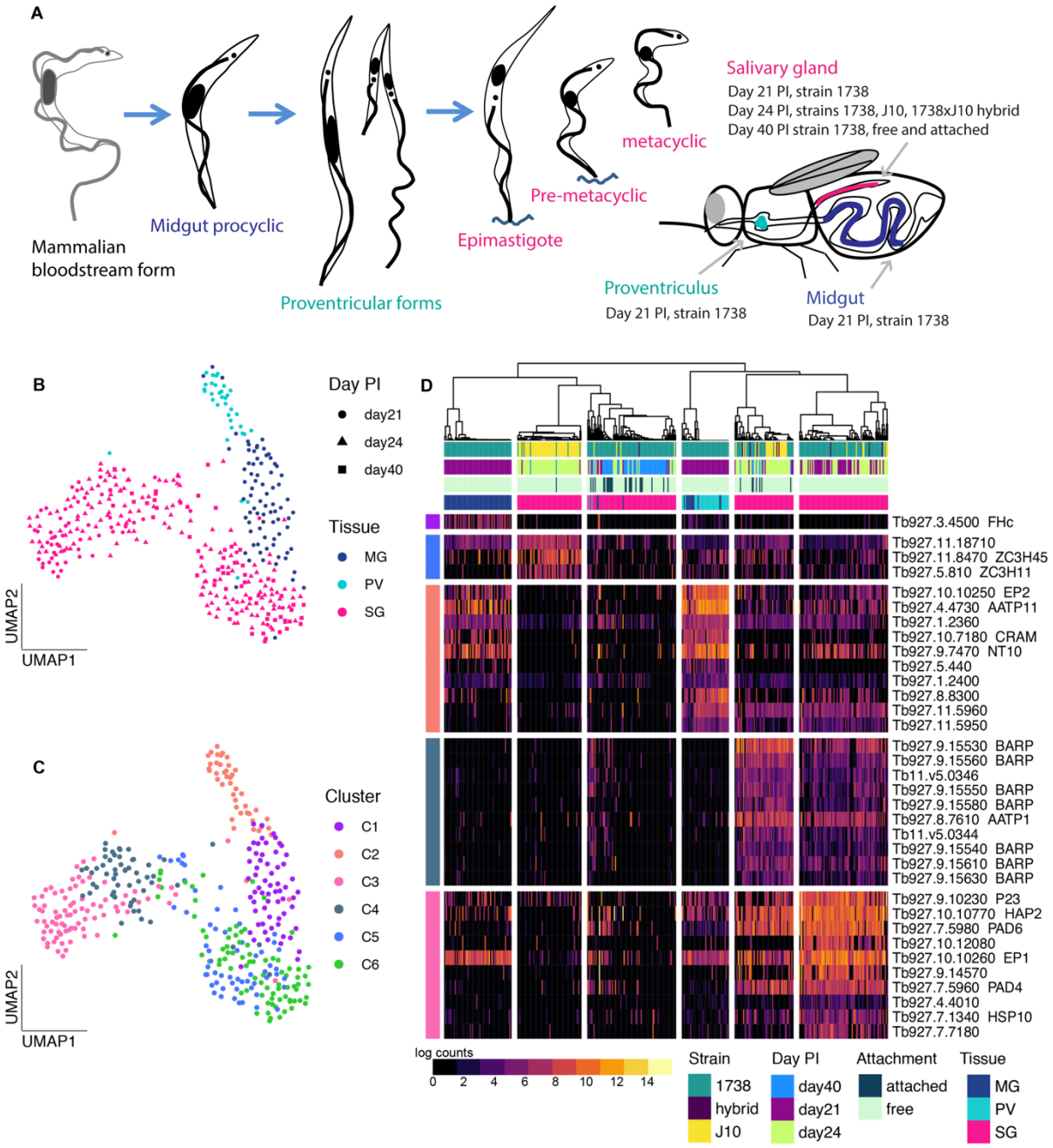

Figure 1. scRNA-seq analysis of trypanosome developmental stages in tsetse.
(A) A schematic of the trypanosome life cycle and collections of the single parasite transcriptomes from midgut (MG; blue), proventriculus (PV; turquoise) and salivary glands (SG; pink) from different time points and strains. Trypanosomes show two conformations: trypomastigote with kinetoplast (small black dot) posterior to nucleus (*e.g.* bloodstream form, procyclic, metacyclic) and with kinetoplast anterior to nucleus (*e.g.* epimastigote). **(B)** A UMAP of the 388 cells that passed QC across collections, coloured by tissue of origin. **(C)** The UMAP coloured by cluster assignment. **(D)** A heatmap of the top significant marker genes from each of the five clusters that had marker genes (AUROC >0.75 & adjusted *p*-value < 0.01).

Expression of mVSG transcripts

Mature metacyclics can be unequivocally identified by their lack of expression of the cell surface proteins GPEET, EP and BARP, replaced by expression of a single *mVSG* gene in each cell [27], but pre-metacyclics also transcribe *mVSG*s at high level in [20]. To explore

mVSG expression in salivary gland-derived cells here, we first needed to identify the *mVSG* repertoires of strains 1738 and J10, which differ from those of previously published strains. We built a *de novo* transcriptome assembly based on all reads from the 388 tsetse-derived cells and identified putative *mVSG*s by comparison of the resulting ORFs to the total *VSG* repertoire of each strain, previously identified using an HMM on full-genome Illumina sequence data [37]. These assembled *mVSG* transcripts were then mapped to 1738 or J10 contigs to place them in a genomic context. Using this method, we identified 11 *mVSG*s that were expressed in our dataset all of which contained an upstream *mVSG* promoter based on the consensus sequence [38,39]. Additionally, downstream telomeric repeats were present in six of these contigs (**Figure 2A**). Although several *ESAG*s were found on the contigs, these were up- not downstream of the promotor and therefore not part of the *mVSG* expression site. The presence of these characteristic features of *mVSG* gave us further confidence that the identified transcripts were originating from *bona fide mVSG*. Interestingly, J10 and 1738 shared one *mVSG* (DN18105), suggesting some level of conservation across strains. This is the first identification of the *mVSG* repertoire in these strains.

Individual parasite transcriptomes were then mapped to the assembly to generate counts for each putative *mVSG*. *MVSG* transcripts were expressed by most cells in cluster C5, followed in order by C4, C6 and C3, with negligible expression in C1 and C2 (**Figure 2B, C**). Overall levels of expression were highest in cluster C5 (**Figure 2B, C, D**), with 38% (16/42 cells) expressing more than one *mVSG* (**Figure 2B, C, D**). Multiple *mVSG*s were also expressed by 38% (6/16) and 21% (4/19) cells in clusters C4 and C6 respectively, and a single cell in C3 (**Figure 2C, 2D**). [20] confirmed expression of two *mVSGs* in pre-metacyclics using single molecule mRNA-FISH and put forward a model where multiple *mVSG*s are transcribed at low levels initially, with a single *mVSG* dominating expression in the mature metacyclic forms. Based on this model, C4, C5 and C6 all contain a high proportion of pre-metacyclics, as well as some mature metacyclics. The *mVSG*s expressed varied over development and between strains, with DN1222 being the dominant transcript in C5, which were primarily J10 cells, and DN16022 being the dominant transcript in C6, which were primarily 1738 cells (**Figure 2D**). Observed expressed *mVSG*s are largely consistent with the *VSG*s present in their strain (when there is sufficient read depth), but low read count and partial coverage result in ambiguous assignment due to sequence conservation between *VSG*s.

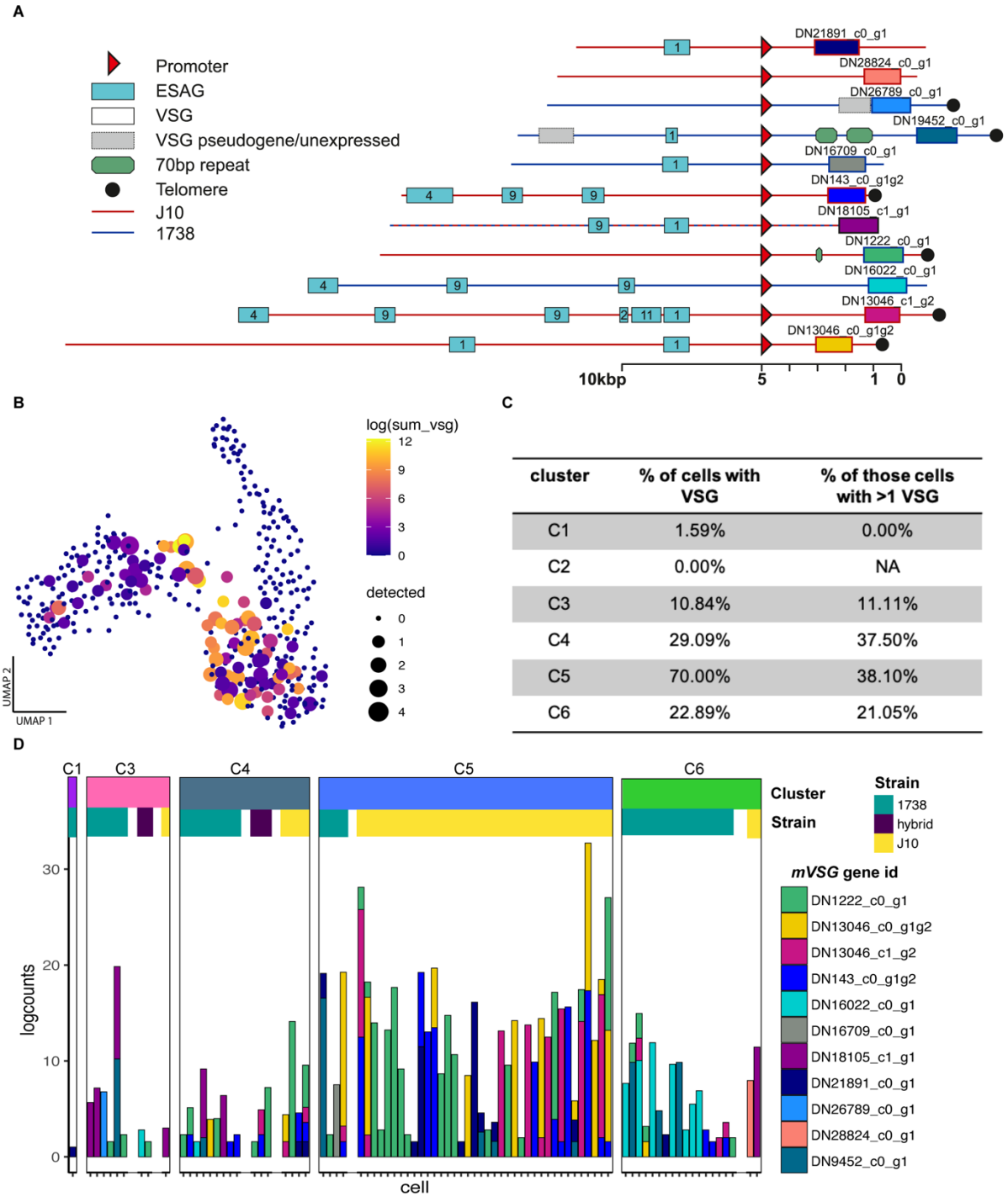

Figure 2. mVSG expression in fly-derived trypanosomes

(A) The genomic context of 11 *mVSG*s identified in strains 1738 and J10. The sequences can be found in supplementary file 1. (**B)** The transcript abundance of *mVSG* across 388 fly-derived trypanosome cells on the UMAP coloured by the logged sum of the *mVSG* counts in each cell and sized by the number of different *mVSG* detected in that cell. (**C)** The breakdown of *mVSG* expression per cluster. C5 had the highest proportion of cells expressing *mVSG* and the greatest proportion of those cells expressing multiple *mVSG*. (**D)** A barchart of all cells expressing *mVSG* (>1 read) organised by cluster and strain. Strain-specific expression of *mVSG* was seen at high levels in C5, which is primarily composed of J10.

Pseudotime analysis of salivary gland development

To understand fine-scale changes in expression patterns during development of salivary gland parasites, we focused on the 161 salivary gland cells of strain 1738 collected over three time points (day 21, 24, 40 pi). The UMAP projection of these cells showed a general correspondence with the clusters identified in **Figure 1C**, with the left-hand group of cells representing clusters C3 and C4 (day 21/24 pi, early and late epimastigotes) and the right-hand group predominantly representing cluster C6 which included most of the day 40 pi transcriptomes (**Figure 3A, B**). The small branch connecting these two groups contained many cells collected from the dissociated salivary gland tissue, which likely represent attached epimastigotes and premetacyclics. Although we cannot rule out that these parasites were trapped unattached cells, their enrichment at this bottleneck in the UMAP indicates their importance in the developmental transition to metacyclics (**Figure 3C**).

We next used Slingshot to order these cells in pseudotime, revealing a trajectory running from left to right from gametes and early epimastigotes to metacyclics (**Figure 3D**). We discovered 692 genes that were differentially expressed over this trajectory (**Figure 3E, Table S4**) and used hierarchical clustering to identify modules of co-expressed genes. Modules expressed early in development were enriched for genes involved in negative regulation of mitotic cell cycle and ATP metabolism (modules 2 and 13, **Figure 3E, Table S4**), while middle-late modules were enriched for genes involved in translation and the ribosome, perhaps necessitated by the changes in surface proteins and metabolism associated with differentiation from epimastigotes to metacyclics (**Figure 3E, Table S4**).

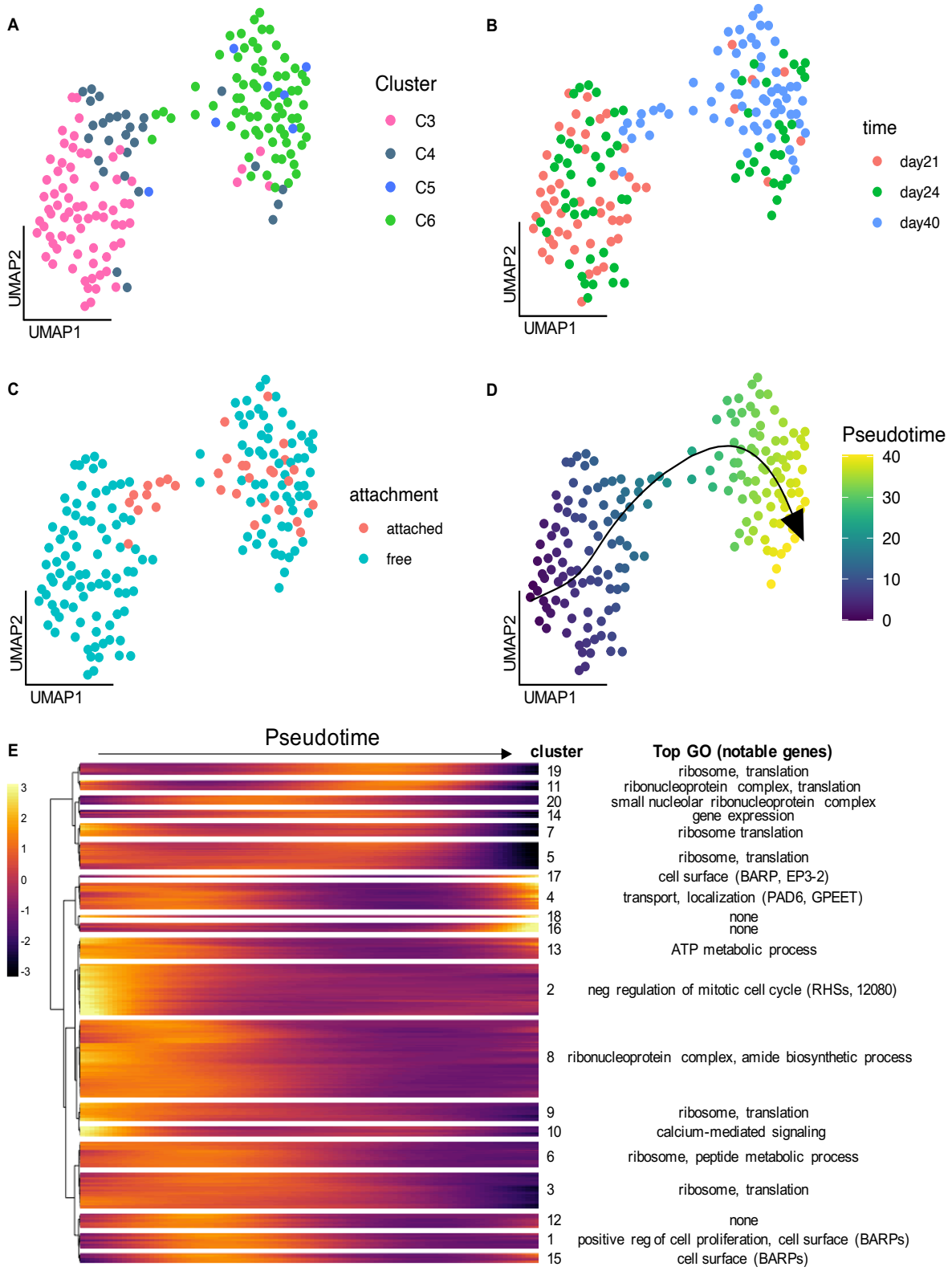

Figure 3. Pseudotime trajectory analysis of developing salivary gland parasites

1738 strain parasites collected from the salivary gland at day 21, 24 and 40 pi were used to map fine-scale changes in gene expression over development. **(A-D)** A UMAP of the 161 1738 salivary glands parasites coloured by global cluster assignment from Figure 1 **(A),** day PI **(B),** attachment treatment **(C)** and pseudotime assignment **(D). (E)** A heatmap of 20 clusters of genes differentially expressed over the pseudotime trajectory from **(D).**

Identification of hybrid progeny and strain-specific expression

To investigate potential cell-types and cell-type specific responses that could be involved in sexual reproduction at day 24 post-infection, we collected J10 and 1738 parasites from both single-strain infected and co-infected tsetse. In the co-infected treatment, we sorted cells from both strains based on fluorescence (1738 GFP+, J10 RFP+), and sorted a small number of RFP+/GFP+ potential hybrid parasites (16 sorted, 14 passed QC). To confirm strain assignment, we used Souporcell to cluster different genotypes based on SNPs in the RNA-seq reads [40]. The two genotype clusters identified were each composed of one of the strains based on fluorescent identification with FACS (1738 GFP+ = cluster 0; J10 RFP+ = 1), and the potential hybrid progeny were identified as inter-genotypic doublets (clusters 0/1 and 1/0), with alleles present from both strains (**Figure 4A)**. This confirmed that six of the RFP+/GFP+ cells were genuine hybrids, while a further three hybrid cells were identified from the RFP-/GFP+ or RFP+/GFP- groups; as the GFP and RFP genes are present on only one homologue, four hybrid genotypes with respect to fluorescent protein genes are expected [41]. Looking at a UMAP of all day 24 pi cells, we observed some separation of strain in both the early and late epimastigote clusters (C3 and C4) and clusters C5 and C6 (**Figure 4B**). However, we did not see clear clustering based on the infection treatment (co- vs single-infection), suggesting that there is no strong transcriptomic response to presence of another parasite strain (**Figure 4C**). In order to understand if the observed strain-specific clustering was a result of different cell-type composition or differential expression between strains within a cell-type, we integrated the data across strains using Seurat v3 [42]. Using this method, we were able to co-cluster the early epimastigote cells and identify 11 genes differentially expressed between the two strains (**Figure S5, Table S5**). However, the later stage cells seen in J10 had no representation of 1738, suggesting this cell-type is unique to this strain at day 24 pi, which could be observed if the strains have different developmental rates (**Figure S5**).

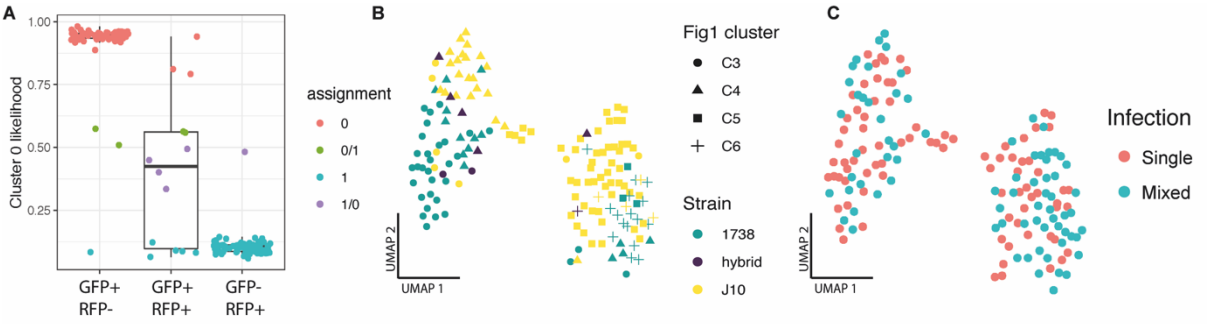

Figure 4. Classification of hybrid progeny

Souporcell was used to assign genotypes based on SNPs found between the two strains. The two genotype assignments (0,1) were each primarily composed of one of the strains based on fluorescent identification with FACS (1738 GFP+/RFP- = cluster 0; J10 GFP-/RFP+ = 1), and the potential hybrid progeny were classified as inter-genotypic doublets (clusters 0/1 and 1/0). The likelihood ratio of cluster 0 assignment is shown for each of the three sorted populations **(A).** The UMAP of day 24 mixed- and single-infection experiments coloured by strain assignment and shaped by Figure 1 cell cluster assignment **(B)** and infection treatment **(C).**

Transcript levels of procyclin and candidate novel sexual stage genes correlate with protein expression *in vivo*.

A primary aim of this study was to identify the sexual stages of *T. brucei* and our results support the hypothesis that cluster C3 (**Figure 1**) represents meiotic intermediates and gametes,

which are abundant around day 21 pi [8–10]. Looking at expression of genes encoding proteins known to be essential for sexual reproduction, we found high levels of expression of *HAP2* and also *GEX1* in cluster C3, with some signal from the meiosis-specific genes *DMC1* and *HOP1* (**Figure 5A**). Surprisingly, these cells also expressed the procyclin gene *GPEET*, which is considered to be a marker of early procyclics in the tsetse midgut, replaced by EP procyclins in late procyclics [31,43]. *GPEET*, *EP1*, *HAP2* and *GEX1* all have the highest expression in cluster C3 (**Figure 5B**). We used immunofluorescence to tie these observations to specific morphological forms and to validate the presence of GPEET on the surface of salivary gland parasites (**Figure 5C, Table S6**). We found that GPEET, together with EP and BARP were present in >90% of the meiotic dividers and gametes (identified by morphological features), and as expected absent in metacyclics (**Figure 5C**). Epimastigotes showed a similar pattern to the sexual forms but lower total proportions. Additionally, we looked at proventricular parasites and found expression of EP and GPEET but no BARPs, further confirming our gene expression data is matched at the protein level for these markers.

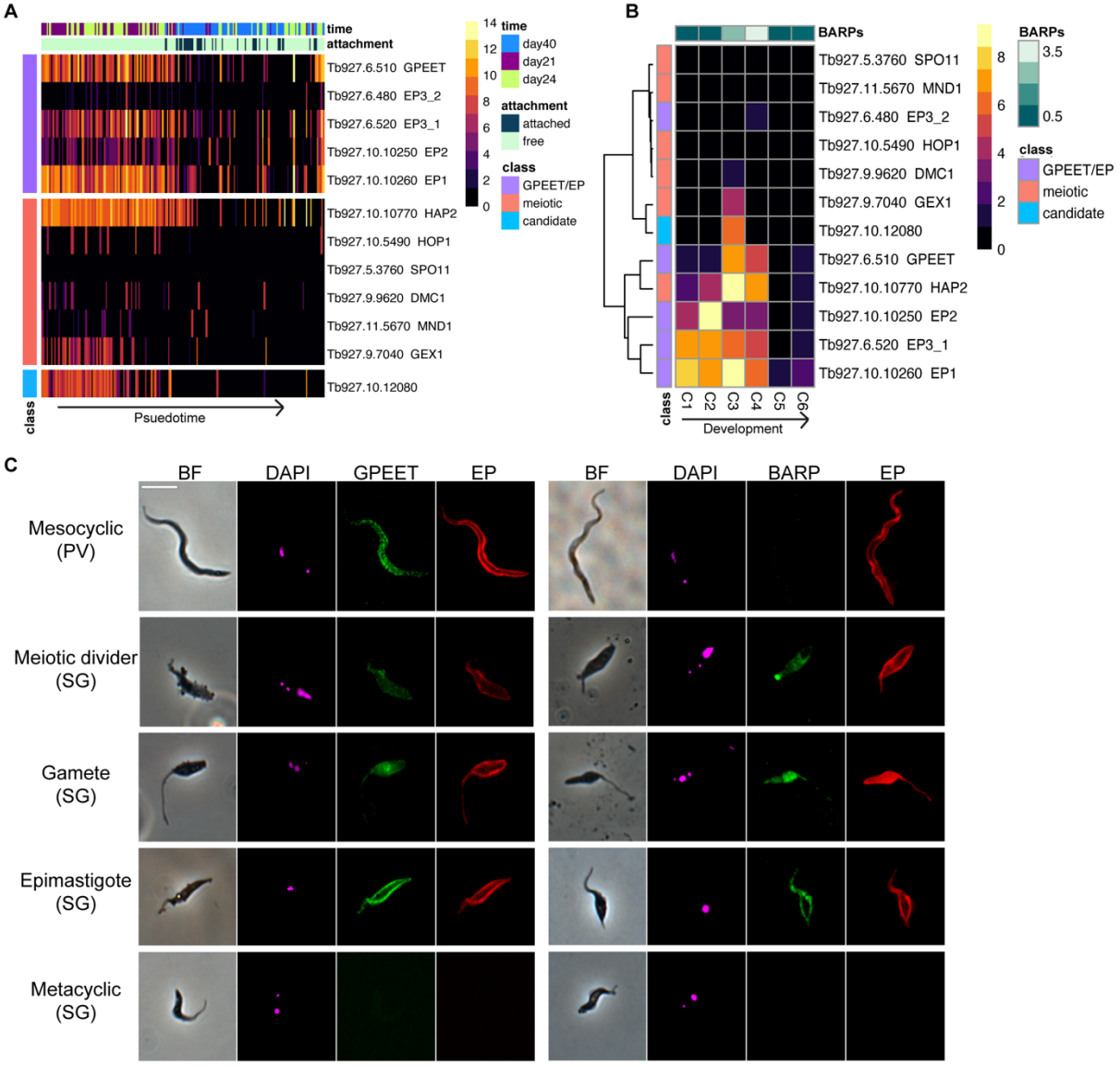

Figure 5. Expression of surface antigens and genes involved in sexual reproduction throughout development in the tsetse fly.

We observed co-expression of procyclic surface antigen genes and *HAP2* in early parasite development in the salivary glands **(A)** and this general pattern of expression was also seen in proventricular forms

(C2) as well as putative epimastigotes (C4) that also had high expression of *BARP*s **(B).** Immunofluorescence assays confirmed that these surface proteins corresponded to their transcriptional profiles and were present on the epimastigote and sexual stages **(C).**

Cluster C3 was additionally characterised by strong and unique expression of Tb927.10.12080 (**Figure 1D, Figure 5A,B, Figure 6A**), and we hypothesized that this gene may play a role in sexual development. It encodes a hypothetical protein devoid of recognisable functional domains that is well-conserved in other trypanosomes including *T. congolense, T. vivax, T. cruzi* and *T. grayi,* and the C-terminal domain in more distantly related members of the trypanosomatid family such as *Leishmania* spp. The gene falls upstream of two RNA binding proteins (Tb927.10.12090 (*RBP7a*); Tb927.10.12100 (*RBP7b*)) and the recently identified long non-coding RNA, *grumpy*, which all play a role in stumpy form differentiation [44,45]. Along with Tb927.10.12080, this genomic region could potentially act as a hotspot for differentiation and developmental processes across the parasite life cycle. Localisation data for Tb927.10.12080 from Tryptag.org is equivocal, showing punctate cytoplasmic fluorescence for the N-terminal tagged protein and a mitochondrial location in a proportion of cells for the C-terminal tagged protein. To investigate expression of this protein during development in the tsetse fly, we used the 3' UTR to regulate expression of GFP driven by the procyclin promotor (**Figure S6**). At 20-22 days post infection, there was very little detectable expression in midgut procyclics or proventricular forms, but strong expression in salivary gland trypanosomes (**Figure 6B**). Overall, of the parasites from the salivary gland scored, 39% (182/464) showed expression (**Table S7)**. These included meiotic intermediates and both 1K1N and 2K1N gametes but not metacyclics or unattached epimastigotes (**Figure 6C,D**). Cell types involved in the early stages of meiosis, such as meiotic dividers and 3N cells with one diploid and two haploid nuclei, had lower percentages of cells expressing (10% and 50% respectively) than did those involved in the later stages of meiosis and in gametes (77-80%; **Figure 6C** & **Table S7**). At 37-38 days post infection, the percentage of fluorescent trypanosomes dropped (103/682; 15%) and these cells were misshapen with no recognisable gametes or sexual intermediates. Further work is needed to understand the functional role of Tb927.10.12080 in meiosis and sexual development; however, its unique pattern of transcript and protein expression indicate it could play a vital role in processes that allow for genetic exchange in *T. brucei* and perhaps more broadly in kinetoplastids.

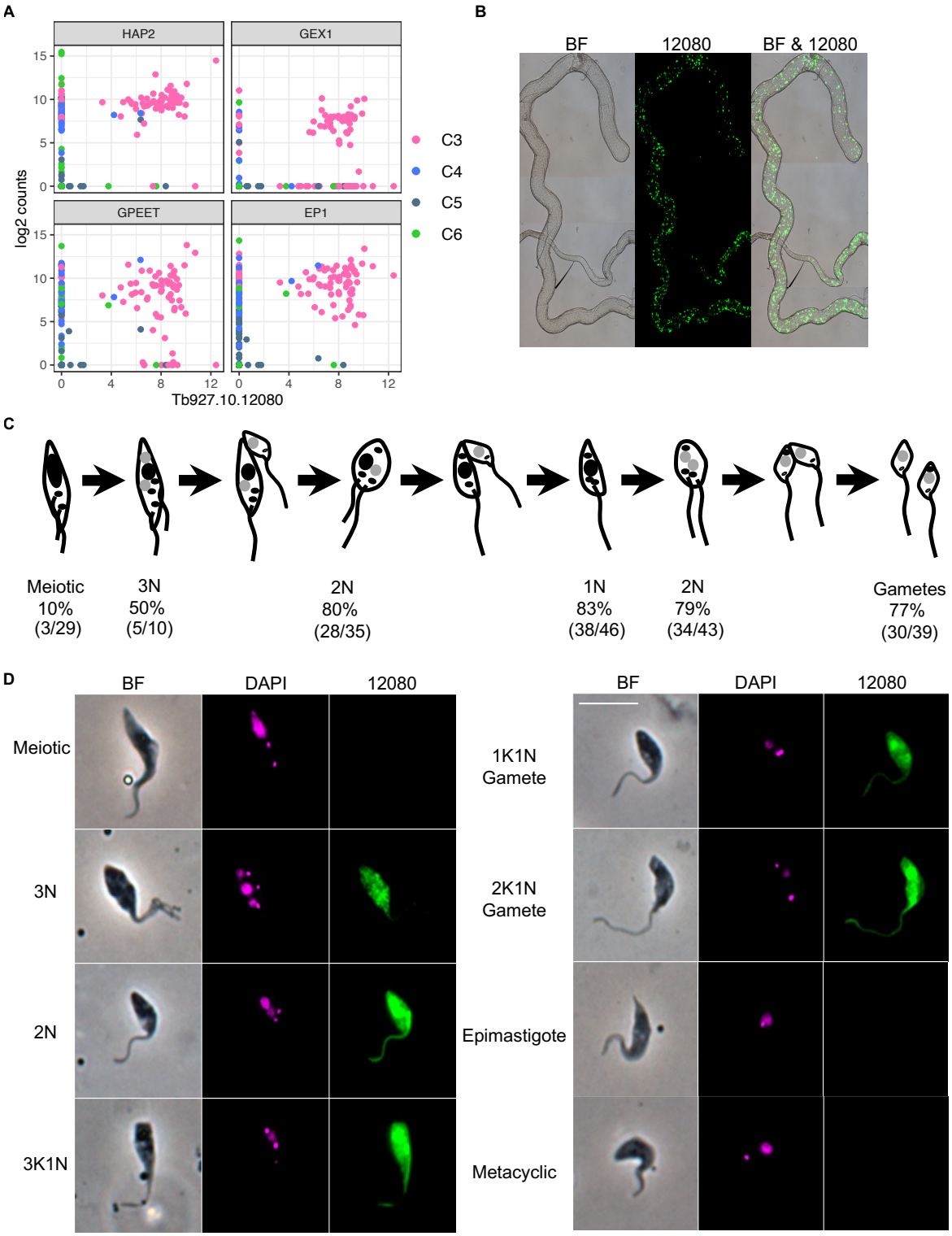

Figure 6. Expression of Tb927.10.12080 coincides with sexual forms.

(A) Co-expression of Tb927.10.12080 with genes encoding HAP2 and GEX1, proteins associated with gamete and nuclear fusion in eukaryotes, and the surface antigen genes GPEET and EP1; co-

expression is seen in a subset of cells from C3 (**Figure 1D**). **(B)** Salivary gland dissected 21 days pi that is infected with *T. brucei* 1738 expressing GFP::Tb927.10.12080-3'UTR transcribed from the procyclin promotor. **(C)** Diagram showing major cell types observed during meiosis in *T. brucei* (adapted from Peacock et al 2021); nuclei are shown in black (4C or 2C DNA contents) or grey (1C, haploid) and kinetoplasts are shown as smaller black dots. Values beneath are the numbers and percentages of cells recorded for each cell type; both 1K1N and 2K1N gametes are included in the gamete total. Full data are presented in **Table S7**. **(D)** Trypanosomes from tsetse fly salivary gland spill-out 16-21 days pi with *T. brucei* 1738 expressing GFP::Tb927.10.12080-3'UTR transcribed from the procyclin promotor. Left to right: phase contrast, DAPI, GFP::Tb927.10.12080-3'UTR. The scale bar represents 10 μm.

Conclusion

Here we applied single-cell RNA sequencing to explore the heterogeneous trypanosome populations in the tsetse fly. These data provide a resource for the parasitology community, which we have made available via an interactive website (http://cellatlas.mvls.gla.ac.uk/). Additionally, this data set allowed us to elucidate the transcriptional profiles of key life cycle stages in the salivary glands including the sexual stages. From this mixture of cell types, we were able to identify a cluster of cells that shared a particular transcriptomic profile characterized by high expression of the gene encoding the gamete fusion protein HAP2, together with several unstudied genes. One of these was a kinetoplastid-conserved gene Tb927.10.12080, which was exclusively expressed at high levels by meiotic intermediates and gametes. We speculate that this protein, currently of unknown function, plays a role in gamete formation and/or fusion.

Materials and Methods

Data collection

Trypanosome culture and tsetse infection

The following tsetse-transmissible strains of *Trypanosoma brucei brucei* were used: *T. b. brucei* J10 (MCRO/ZM/73/J10) and 1738 (MOVS/KE/70/EATRO 1738); each was genetically modified to express a fluorescent protein gene (J10 RFP, 1738 GFP). Mating between these strains has been demonstrated previously [41] and 1738 reliably produces large numbers of gametes around day 21 post-infection [9]. Procyclic form (PF) trypanosomes were grown in Cunningham's medium (CM) [46] supplemented with 15 % v/v heat-inactivated foetal calf serum, 5 μg/ml hemin and 10 μg/ml gentamycin at 27°C. Tsetse flies (*Glossina pallidipes*) were infected with PF trypanosomes, maintained and dissected as described previously [8].

Parasite isolation from tsetse tissues for scRNA-seq

Free swimming parasites were obtained from *G. pallidipes* by separately pooling tissues into CM (5 midguts in 500 μl CM; 5 proventriculi in 50 μl CM; 20 sets of salivary glands in 50 μl CM). Tissues were incubated at RT for 10 minutes prior to filtration through a 100 μm filter. Cells were washed once with 1 ml CM prior to preservation or sorting. At day 40 pi, the parasites attached to the salivary glands were isolated by disassociation of the tissue after the

10-minute incubation period. Forceps were used to transfer the tissue to an enzymatic solution consisting of 200 µl Collagenase IV (1 mg/ml) and 25 µl of Elastase (4 mg/ml). The sample was then incubated at 30°C for 40 min with shaking at 300 rpm. During the incubation, the tissue was disrupted by pipetting up and down 40 times every 15 minutes at first with a p1000 pipette set to 150 µl and then with a p200 set to 100 µl once the tissue started to break up.

Cell preservation

A subset of cells was preserved prior to cell sorting to allow for greater flexibility in the time between collections and FACS. The day 40 pi salivary glands parasites (attached and free) cells were fixed by adding 200 µl of dithio-bis(succinimidyl propionate) (DSP; Lomant's reagent) dropwise to the cell pellet as described in [47]. DSP fixed samples were incubated at room temperature for 30 minutes prior to adding 4 µl 1 M Tris-HCl pH 7.5. Samples were then stored at 4°C for up to 24 hours. Prior to sorting, DTT was added to a final concentration of 50 mM. The day 24 pi salivary gland parasites (cross vs single infection) were preserved by resuspending the cell pellet in 200 µl Hypothermosol-FRS (BioLifeSolutions) [48]. Samples were then stored at 4°C for 5 hours prior to sorting.

Cell sorting, library preparation and sequencing

All parasite cells were sorted within 24 hours of collection on an Influx cell sorter (BD Biosciences) with a 200 µm nozzle or a Sony SH800 with 100 µm chip. Parasites were sorted based on RFP and/or GFP fluorescence into nuclease-free 96 well plates containing lysis buffer as described previously [12]. Sorted plates were spun at 1000 g for 10 seconds and immediately placed on dry ice. Reverse transcription, PCR, and library preparation were performed as described in [12]. Cells were multiplexed to 384 and sequenced on a single lane of Illumina HiSeq2500 v4 with 75 bp paired-end reads.

Immunofluorescence

Salivary glands, proventriculi and midguts from infected flies (20-22 days pi) were pooled separately into CM and incubated at room temperature (RT) for 10 minutes (to allow trypanosomes to swim out of tissue) and then filtered through a 100 µm filter with PBS. Trypanosomes were pelleted by centrifugation and resuspended in 100 µl PBS. Cells were fixed overnight at 4ºC by adding 100 µl 6% paraformaldehyde, 0.1% glutaraldehyde in PBS, and then washed twice with PBS before resuspension in 50 µl PBS. Cell suspensions were pipetted onto 2 x 10 mm coverslips, allowed to settle for 20 mins in a humid chamber, and then liquid was removed and replaced by 2% BSA in PBS. After 30 mins liquid was removed and cells incubated with 2% BSA in PBS containing diluted antibody for 30 mins at RT. Rabbit anti-GPEET (1:1000) and rabbit anti-BARP (1:1000) were a kind gift from Isobel Roditi, University of Bern, Switzerland; mouse anti-EP mAB (1:100) was from Cedarlane. Cells were washed three times with PBS and incubated with 2% BSA in PBS containing anti-rabbit FITC (1:1000) and anti-mouse TRITC (1:1000) for 30 mins at RT. Cells were washed three times with PBS, briefly air dried, stained with DAPI in VECTASHIELD mounting medium (Vector Laboratories) and viewed using a DMRB microscope (Leica) equipped with a Retiga Exi camera (QImaging) and Volocity software (PerkinElmer). The whole area of the coverslip was scanned systematically from top to bottom, capturing FITC, TRITC, DAPI and phase contrast images of each trypanosome. Digital images were analysed using ImageJ (http://rsb.info.nih.gov/ij).

Tb927.10.12080 gene expression

The 3' UTR of Tb927.10.12080 was amplified from genomic DNA of 1738 using the primers 5'- GATCCTCGAGTAGTGGCGAGTGTTTACAACAGTGTC and 5'- GATCGGGCCCCTTGTGCGGATCCAAACAA, and inserted immediately downstream of a GFP gene driven by the procyclin promotor from plasmid backbone pHD449, which is designed for insertion into the tubulin locus (**Figure S6**) [49,50]. The plasmid construct was used for stable transfection of procyclic 1738 and following hygromycin selection, clones were tsetse fly transmitted as described by Peacock et al 2011, 2014. Flies were dissected 10 to 40 days pi and organs viewed by fluorescence microscopy and imaged live or fixed in 2% paraformaldehyde and stained with DAPI in VECTASHIELD mounting medium (Vector Laboratories).

scRNAseq data analysis

Mapping and generation of expression matrices

Nextera adaptor sequences were trimmed from fastq files using trim_galore (*-q 20 -a CTGTCTCTTATACACATCT --paired --stringency 3 --length 50 -e 0.1)* (v 0.4.3) [51]. Trimmed reads were mapped using HISAT2 (*hisat2 --max-intronlen 5000 -p 12*) (v 2.1.0) [52] to the *T. b. brucei* 927 genome. The GFF was converted to GTF using the UCSC genome browser tool [53]. Reads were then summed against genes using HTseq (*htseq-count -f bam -r pos -s no -t CDS*) (v 0.7.1) [54].

Assembly of VSG transcripts

Because there is a lack of conservation of *VSG*s across *T. brucei* strains, we built a *de novo* transcriptome assembly to identify the *mVSG* transcripts expressed in 1738 and J10. First, we merged the BAM files across the 388 cells and converted to FASTQ using bedtools (v. 2.29.2) [55]. Using Trinity (v. 2.1.1) [56] to assemble the transcripts from this merged file, we detected 53521 'genes' with a mean contig length of 800 bp. We then mapped each cell to this assembly using RSEM (v. 1.3.3) to generate a counts matrix and used Transdecoder (v. 5.5.0) to detect open reading frames [57]. BLASTp (v. 2.9.0) was used to match to putative *VSG*s that had been curated independently from whole genome data as described below. Transcripts with >90% identity were used for further analysis.

Genomes for the parent strains J10 and 1738 were assembled from 76 bp paired read Illumina data using SPADES under default parameters. Predicted *mVSG*s were identified to genomic loci, using BLAST against the assembled contigs with a percent identity across the entire transcript >95%, alignments of a raw score of greater than 1000 were further investigated. Additional open reading frames were identified by BLAST alignment of the curated 927 annotated CDS set. Nhmmer was used to identify putative *mVSG* promoters from the alignments [38,39].

Quality control and normalization

Quality control and visualisation was performed in Scater (v. 1.12.2) [58]. Cell quality was assessed based on the distribution of genes detected per single-cell transcriptome. Cells with fewer than 40 genes or more than 3000 genes detected were removed, as well as cells that had fewer than 1000 total reads. These QC thresholds allowed us to keep more cells in the

analysis that are likely to be less transcriptionally active such as mature metacyclics. Out of 515 parasites isolated from tsetse tissue that were sequenced, 388 passed quality control and were used for downstream analyses. Raw count data was normalized using a deconvolution size factor in Scran (v. 1.16.0) [59] to account for differences in overall level of expression between cell-types.

Cell clustering, projection, and marker genes

In order to unbiasedly group transcriptomes based on similar expression profiles, 388 cells collected from the tsetse were clustered using K-means clustering in SC3 (v. 1.12.0) [32]. Dimensionality reduction was performed in Scater (v. 1.12.2) [58] using Principal Component Analysis (PCA) with the 500 most variable genes based on log2 normalised expression values. Marker genes were identified for each cluster using SC3 (AUROC >0.75 & adjusted *p*-value < 0.01).

Identification of hybrid parasites and data integration

To select different parasite genotypes (J10, 1738, or J10x1738 cross) in the mixed infection treatment, we first FACS sorted based on GFP+ (1738), RFP+ (J10), or GFP+/RFP+ (hybrid) expression. We then used souporcell (*-k 2 -p 2*) (v2.0) [40] to confirm genotype assignment based on SNP profiles from the scRNA-seq.

To identify genes that were differentially expressed between the two strains, we used Seurat (v3.1.5) to integrated the data by identifying anchors with the FindIntegrationAnchors() and IntegrateData() functions. The data was then clustered using FindNeighbors() with the top 30 principal components and FindClusters() with a resolution of 0.5. Differential expression was performed using the FindMarkers() function (adjusted *p*-value < 0.001). The same integration methods were used to compare the data to [20] except that the top 20 principal components were used, the cluster resolution was 0.8 and the FindConservedMarkers() function identified markers found in both studies (max *p*-value < 0.001).

Pseudotime and differential expression

To assess developmental progression in 1738 salivary glands parasites from day 21-, 24-, and 40-days pi, Slingshot (v. 1.8.0) was used to infer a pseudotime developmental trajectory [60]. Genes differentially expressed over this trajectory were identified using the associationTest() function in TradeSeq (v. 1.4.0) [61].

Data availability

Raw sequence data are available on the European Nucleotide Archive under accession ERP132258. The expression matrix and associated code will be available on Github at https://github.com/vhowick/tryps_single_cell. The data are explorable via the Glasgow Cell Atlas website at http://cellatlas.mvls.gla.ac.uk/.

Acknowledgements

We thank the UK Biotechnology and Biological Sciences Research Council (BB/R016437/1 and BB/R010188/1) and Wellcome Trust (http://www.wellcome.ac.uk/) for funding. The Wellcome Sanger Institute is funded by the Wellcome Trust (Grant 206194/Z/17/Z), which supports M.K.N.L. V.M.H. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant 220185/Z/20/Z). We gratefully acknowledge the

generous supply of tsetse pupae from the International Atomic Energy Agency, Vienna. We thank the Sanger Institute flow cytometry core facility for technical support and guidance. We thank Thomas Otto and Jesse Ropp for development and deployment of the interactive website.

Author Contributions

VMH, LP, CK, and CC designed and performed all experiments. VMH, LP, CK, and WG analysed the data. WG and MKNL sourced funding, provided scientific direction, and supported experimental design. VMH, LP, CK and WG wrote the manuscript with contributions from all authors.

References

1. Matthews KR. The developmental cell biology of Trypanosoma brucei. J Cell Sci. 2005;118: 283–290. doi:10.1242/jcs.01649

2. Vassella E, Acosta-Serrano A, Studer E, Lee SH, Englund PT, Roditi I. Multiple procyclin isoforms are expressed differentially during the development of insect forms of Trypanosoma brucei. J Mol Biol. 2001;312: 597–607. doi:10.1006/jmbi.2001.5004

3. Bütikofer P, Vassella E, Mehlert A, Ferguson MAJ, Roditi I. Characterisation and cellular localisation of a GPEET procyclin precursor in Trypanosoma brucei insect forms. Mol Biochem Parasitol. 2002;119: 87–95. doi:10.1016/s0166-6851(01)00398-x

4. Bringaud F, Rivière L, Coustou V. Energy metabolism of trypanosomatids: adaptation to available carbon sources. Mol Biochem Parasitol. 2006;149: 1–9. doi:10.1016/j.molbiopara.2006.03.017

5. Van Den Abbeele J, Claes Y, van Bockstaele D, Le Ray D, Coosemans M. Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. Parasitology. 1999;118 (Pt 5): 469–478. doi:10.1017/s0031182099004217

6. Sharma R, Peacock L, Gluenz E, Gull K, Gibson W, Carrington M. Asymmetric cell division as a route to reduction in cell length and change in cell morphology in trypanosomes. Protist. 2008;159: 137–151. doi:10.1016/j.protis.2007.07.004

7. Urwyler S, Studer E, Renggli CK, Roditi I. A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of Trypanosoma brucei. Mol Microbiol. 2007;63: 218–228. doi:10.1111/j.1365-2958.2006.05492.x

8. Peacock L, Ferris V, Sharma R, Sunter J, Bailey M, Carrington M, et al. Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly. Proc Natl Acad Sci U S A. 2011;108: 3671–3676. doi:10.1073/pnas.1019423108

9. Peacock L, Bailey M, Carrington M, Gibson W. Meiosis and haploid gametes in the pathogen Trypanosoma brucei. Curr Biol. 2014;24: 181–186. doi:10.1016/j.cub.2013.11.044

10. Peacock L, Kay C, Farren C, Bailey M, Carrington M, Gibson W. Sequential production of gametes during meiosis in trypanosomes. Commun Biol. 2021;4: 555. doi:10.1038/s42003-021-02058-5

11. Poran A, Nötzel C, Aly O, Mencia-Trinchant N, Harris CT, Guzman ML, et al. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature. 2017;551: 95–99. doi:10.1038/nature24280

12. Reid AJ, Talman AM, Bennett HM, Gomes AR, Sanders MJ, Illingworth CJR, et al. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. Elife. 2018;7: e33105. doi:10.7554/eLife.33105

13. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K, et al. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. Science. 2019;365. doi:10.1126/science.aaw2619

14. Real E, Howick VM, Dahalan FA, Witmer K, Cudini J, Andradi-Brown C, et al. A single-cell atlas of Plasmodium falciparum transmission through the mosquito. Nat Commun. 2021;12: 3196. doi:10.1038/s41467-021-23434-z

15. Witmer K, Dahalan FA, Metcalf T, Talman AM, Howick VM, Lawniczak MKN. Using scRNA-seq to Identify Transcriptional Variation in the Malaria Parasite Ookinete Stage. Front Cell Infect Microbiol. 2021;11: 68. doi:10.3389/fcimb.2021.604129

16. Sà JM, Cannon MV, Caleon RL, Wellems TE, Serre D. Single-cell transcription analysis of Plasmodium vivax blood-stage parasites identifies stage- and species-specific profiles of expression. PLoS Biol. 2020;18: e3000711. doi:10.1371/journal.pbio.3000711

17. Nötzel C, Kafsack BFC. There and back again: malaria parasite single-cell transcriptomics comes full circle. Trends Parasitol. 2021. doi:10.1016/j.pt.2021.07.011

18. Briggs EM, Rojas F, McCulloch R, Matthews KR, Otto TD. Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing. Nat Commun. 2021;12: 5268. doi:10.1038/s41467-021-25607-2

19. Vigneron A, O'Neill MB, Weiss BL, Savage AF, Campbell OC, Kamhawi S, et al. Single-cell RNA sequencing of Trypanosoma brucei from tsetse salivary glands unveils metacyclogenesis and identifies potential transmission blocking antigens. Proc Natl Acad Sci U S A. 2020. doi:10.1073/pnas.1914423117

20. Hutchinson S, Foulon S, Crouzols A, Menafra R, Rotureau B, Griffiths AD, et al. The establishment of variant surface glycoprotein monoallelic expression revealed by single-cell RNA-seq of Trypanosoma brucei in the tsetse fly salivary glands. PLoS Pathog 17(9): e1009904. (2021) https://doi.org/10.1371/journal.ppat.1009904

21. Savage AF, Cerqueira GC, Regmi S, Wu Y, El Sayed NM, Aksoy S. Transcript expression analysis of putative Trypanosoma brucei GPI-anchored surface proteins during development in the tsetse and mammalian hosts. PLoS Negl Trop Dis. 2012;6: e1708. doi:10.1371/journal.pntd.0001708

22. Telleria EL, Benoit JB, Zhao X, Savage AF, Regmi S, Alves e Silva TL, et al. Insights into the trypanosome-host interactions revealed through transcriptomic analysis of parasitized tsetse fly salivary glands. PLoS Negl Trop Dis. 2014;8: e2649. doi:10.1371/journal.pntd.0002649

23. Kolev NG, Ramey-Butler K, Cross GAM, Ullu E, Tschudi C. Developmental progression to infectivity in Trypanosoma brucei triggered by an RNA-binding protein. Science. 2012;338: 1352–1353. doi:10.1126/science.1229641

24. Savage AF, Kolev NG, Franklin JB, Vigneron A, Aksoy S, Tschudi C. Transcriptome Profiling of Trypanosoma brucei Development in the Tsetse Fly Vector Glossina morsitans. PLoS One. 2016;11: e0168877. doi:10.1371/journal.pone.0168877

25. Naguleswaran A, Doiron N, Roditi I. RNA-Seq analysis validates the use of culture-derived Trypanosoma brucei and provides new markers for mammalian and insect life-cycle stages. BMC Genomics. 2018;19: 227. doi:10.1186/s12864-018-4600-6

26. Ramey-Butler K, Ullu E, Kolev NG, Tschudi C. Synchronous expression of individual metacyclic variant surface glycoprotein genes in Trypanosoma brucei. Mol Biochem Parasitol. 2015;200: 1–4. doi:10.1016/j.molbiopara.2015.04.001

27. Christiano R, Kolev NG, Shi H, Ullu E, Walther TC, Tschudi C. The proteome and transcriptome of the infectious metacyclic form of Trypanosoma brucei define quiescent cells primed for mammalian invasion. Mol Microbiol. 2017;106: 74–92. doi:10.1111/mmi.13754

28. Doleželová E, Kunzová M, Dejung M, Levin M, Panicucci B, Regnault C, et al. Cell-based and multi-omics profiling reveals dynamic metabolic repurposing of mitochondria to drive developmental progression of Trypanosoma brucei. PLoS Biol. 2020;18: e3000741. doi:10.1371/journal.pbio.3000741

29. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9: 171–181. doi:10.1038/nprot.2014.006

30. Briggs EM, Warren FSL, Matthews KR, McCulloch R, Otto TD. Application of single-cell transcriptomics to kinetoplastid research. Parasitology. 2021;148: 1223–1236. doi:10.1017/S003118202100041X

31. Bütikofer P, Ruepp S, Boschung M, Roditi I. "GPEET" procyclin is the major surface protein of procyclic culture forms of Trypanosoma brucei brucei strain 427. Biochem J. 1997;326 (Pt 2): 415–423. doi:10.1042/bj3260415

32. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14: 483–486. doi:10.1038/nmeth.4236

33. Naguleswaran A, Fernandes P, Bevkal S, Rehmann R, Nicholson P, Roditi I. Developmental changes and metabolic reprogramming during establishment of infection and progression of Trypanosoma brucei brucei through its insect host. PLoS Negl Trop Dis. 2021;15: e0009504. doi:10.1371/journal.pntd.0009504

34. Coustou V, Biran M, Besteiro S, Rivière L, Baltz T, Franconi J-M, et al. Fumarate is an essential intermediary metabolite produced by the procyclic Trypanosoma brucei. J Biol Chem. 2006;281: 26832–26846. doi:10.1074/jbc.M601377200

35. Dean S, Marchetti R, Kirk K, Matthews KR. A surface transporter family conveys the trypanosome differentiation signal. Nature. 2009;459: 213–217. doi:10.1038/nature07997

36. Vickerman K. Developmental cycles and biology of pathogenic trypanosomes. Br Med Bull. 1985;41: 105–114. doi:10.1093/oxfordjournals.bmb.a072036

37. Sistrom M, Evans B, Bjornson R, Gibson W, Balmer O, Mäser P, et al. Comparative genomics reveals multiple genetic backgrounds of human pathogenicity in the

678 Trypanosoma brucei complex. Genome Biol Evol. 2014;6: 2811–2819.
679 doi:10.1093/gbe/evu222

680 38. Ginger ML, Blundell PA, Lewis AM, Browitt A, Günzl A, Barry JD. Ex vivo and in vitro
681 identification of a consensus promoter for VSG genes expressed by metacyclic-stage
682 trypanosomes in the tsetse fly. Eukaryot Cell. 2002;1: 1000–1009.
683 doi:10.1128/EC.1.6.1000-1009.2002

684 39. Cross GAM, Kim H-S, Wickstead B. Capturing the variant surface glycoprotein repertoire
685 (the VSGnome) of Trypanosoma brucei Lister 427. Mol Biochem Parasitol. 2014;195: 59–
686 73. doi:10.1016/j.molbiopara.2014.06.004

687 40. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust
688 clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat
689 Methods. 2020. doi:10.1038/s41592-020-0820-1

690 41. Gibson W, Peacock L, Ferris V, Williams K, Bailey M. The use of yellow fluorescent
691 hybrids to indicate mating in Trypanosoma brucei. Parasit Vectors. 2008;1: 4.
692 doi:10.1186/1756-3305-1-4

693 42. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic
694 data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:
695 411–420. doi:10.1038/nbt.4096

696 43. Vassella E, Den Abbeele JV, Bütikofer P, Renggli CK, Furger A, Brun R, et al. A major
697 surface glycoprotein of trypanosoma brucei is expressed transiently during development
698 and can be regulated post-transcriptionally by glycerol or hypoxia. Genes Dev. 2000;14:
699 615–626. Available: https://www.ncbi.nlm.nih.gov/pubmed/10716949

700 44. Guegan F, Neves D, Sequeira M, Notredame C, Figueiredo LM. A long non-coding RNA
701 controls parasite differentiation in African trypanosomes. Biorxiv. 2020.
702 doi:10.1101/2020.05.03.074625

703 45. Mony BM, MacGregor P, Ivens A, Rojas F, Cowton A, Young J, et al. Genome-wide
704 dissection of the quorum sensing signalling pathway in Trypanosoma brucei. Nature.
705 2014;505: 681–685. doi:10.1038/nature12864

706 46. Cunningham I. New culture medium for maintenance of tsetse tissues and growth of
707 trypanosomatids. J Protozool. 1977;24: 325–329. doi:10.1111/j.1550-
708 7408.1977.tb00987.x

709 47. Attar M, Sharma E, Li S, Bryer C, Cubitt L, Broxholme J, et al. A practical solution for
710 preserving single cells for RNA sequencing. Sci Rep. 2018;8: 2151. doi:10.1038/s41598-
711 018-20372-7

712 48. Wang W, Penland L, Gokce O, Croote D, Quake SR. High fidelity hypothermic
713 preservation of primary tissues in organ transplant preservative for single cell
714 transcriptome analysis. BMC Genomics. 2018;19: 140. doi:10.1186/s12864-018-4512-5

715 49. Biebinger S, Wirtz LE, Lorenz P, Clayton C. Vectors for inducible expression of toxic gene
716 products in bloodstream and procyclic Trypanosoma brucei. Mol Biochem Parasitol.
717 1997;85: 99–112. Available: https://www.ncbi.nlm.nih.gov/pubmed/9108552

718 50. Bingle LE, Eastlake JL, Bailey M, Gibson WC. A novel GFP approach for the analysis of
719 genetic exchange in trypanosomes allowing the in situ detection of mating events.
720 Microbiology. 2001;147: 3231–3240. doi:10.1099/00221287-147-12-3231

721 51. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
722 EMBnet.journal. 2011;17: 10–12. Available:
723 https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

724 52. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory
725 requirements. Nat Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317

726 53. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief
727 Bioinform. 2013;14: 144–161. doi:10.1093/bib/bbs038

728 54. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput
729 sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

730 55. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
731 Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

732 56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
733 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol.
734 2011;29: 644–652. doi:10.1038/nbt.1883

735 57. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
736 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
737 generation and analysis. Nat Protoc. 2013;8: 1494–1512. doi:10.1038/nprot.2013.084

738 58. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control,
739 normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33:
740 1179–1186. doi:10.1093/bioinformatics/btw777

741 59. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA
742 sequencing data with many zero counts. Genome Biol. 2016;17: 75. doi:10.1186/s13059-
743 016-0947-7

744 60. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and
745 pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19: 477.
746 doi:10.1186/s12864-018-4772-0

747 61. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al.
748 Trajectory-based differential expression analysis for single-cell sequencing data. Nat
749 Commun. 2020;11: 1201. doi:10.1038/s41467-020-14766-3

750
751
752

Supplementary Information

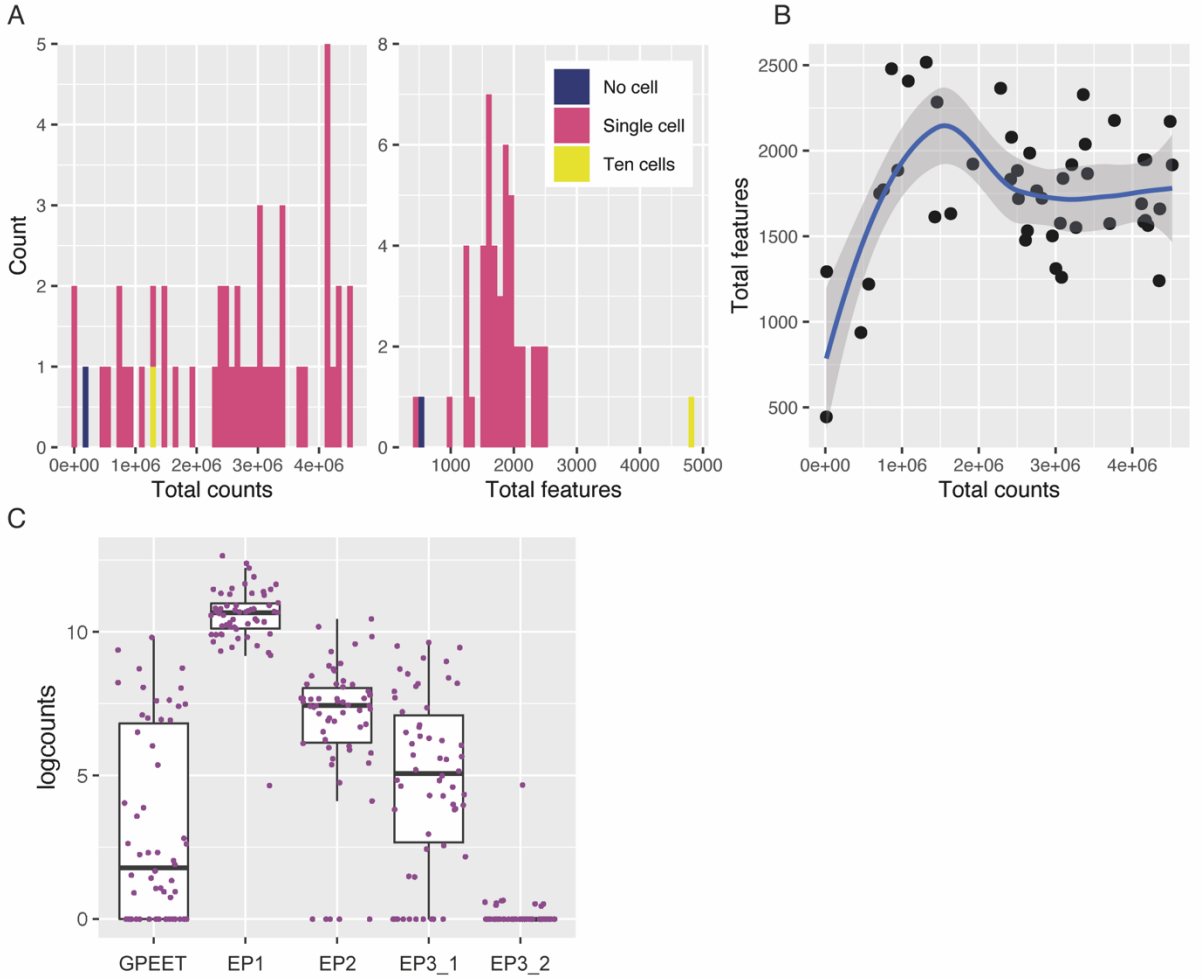

Figure S1. Quality assessment and expression of marker genes in procyclic culture singe-cell transcriptomes

Forty-eight transcriptomes were generated using Smart-seq2 from parasites in a procyclic culture including a no cell and ten cell control. **(A)** The distribution of the total counts and total features (genes) detected in these 48 transcriptomes. **(B)** The total features plotted against total counts for the 46 single-cell transcriptomes shows a plateau as features and counts increase, suggesting that sequencing was saturated for these cells. We detected a mean of 2.6×10^6 reads and 1756 features per single-cell transcriptome. **(C)** Expression of procyclic surface antigen genes *GPEET* (Tb927.6.510), *EP1* (Tb927.10.10260), *EP2* (Tb927.10.10250), *EP3_1* (Tb927.6.520), *EP3_2* (Tb927.6.480).

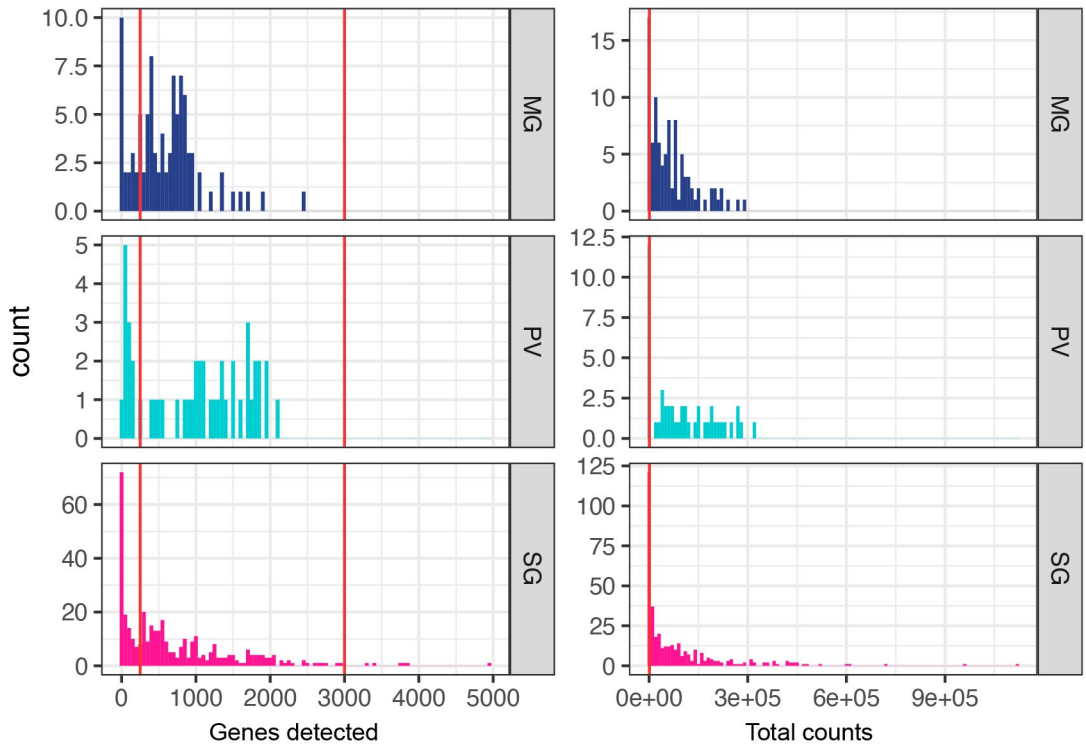

Figure S2. Quality control of insect stage parasites

The distribution of genes detected (left) and counts (right) in each cell across the three insect tissues: midgut (MG), proventriculus (PV), and salivary glands (SG). Cells with fewer than 40 or more than 3000 genes per cell were removed. Additionally, cells with fewer than 1000 reads were removed. Cut-offs are represented by the red vertical lines in each histogram. After QC we detected a mean of 889 genes per cell and $1.1x10^5$ counts per cell.

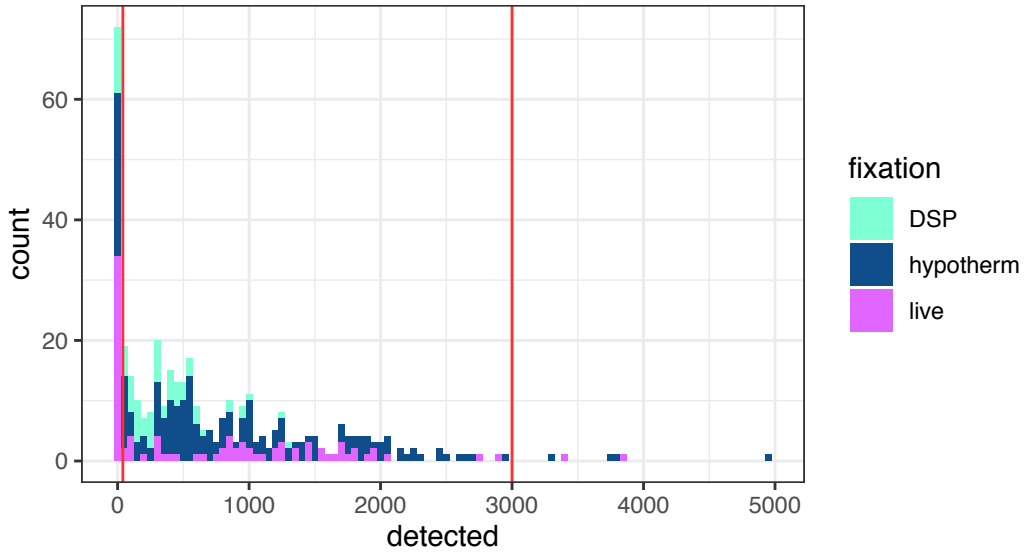

Figure S3. Assessment of preservation methods of salivary gland parasites

The distribution of features detected in salivary gland cells across the two preservation treatments (DSP and hypothermosol) compared to live parasites. Although there were slight differences in detection between the different treatments, caution must be taken in interpreting

781 these differences as the fixation methods are confounded with the different timepoints
782 collected (DSP: day 40; hypothermosol: day 24; live: day 21).

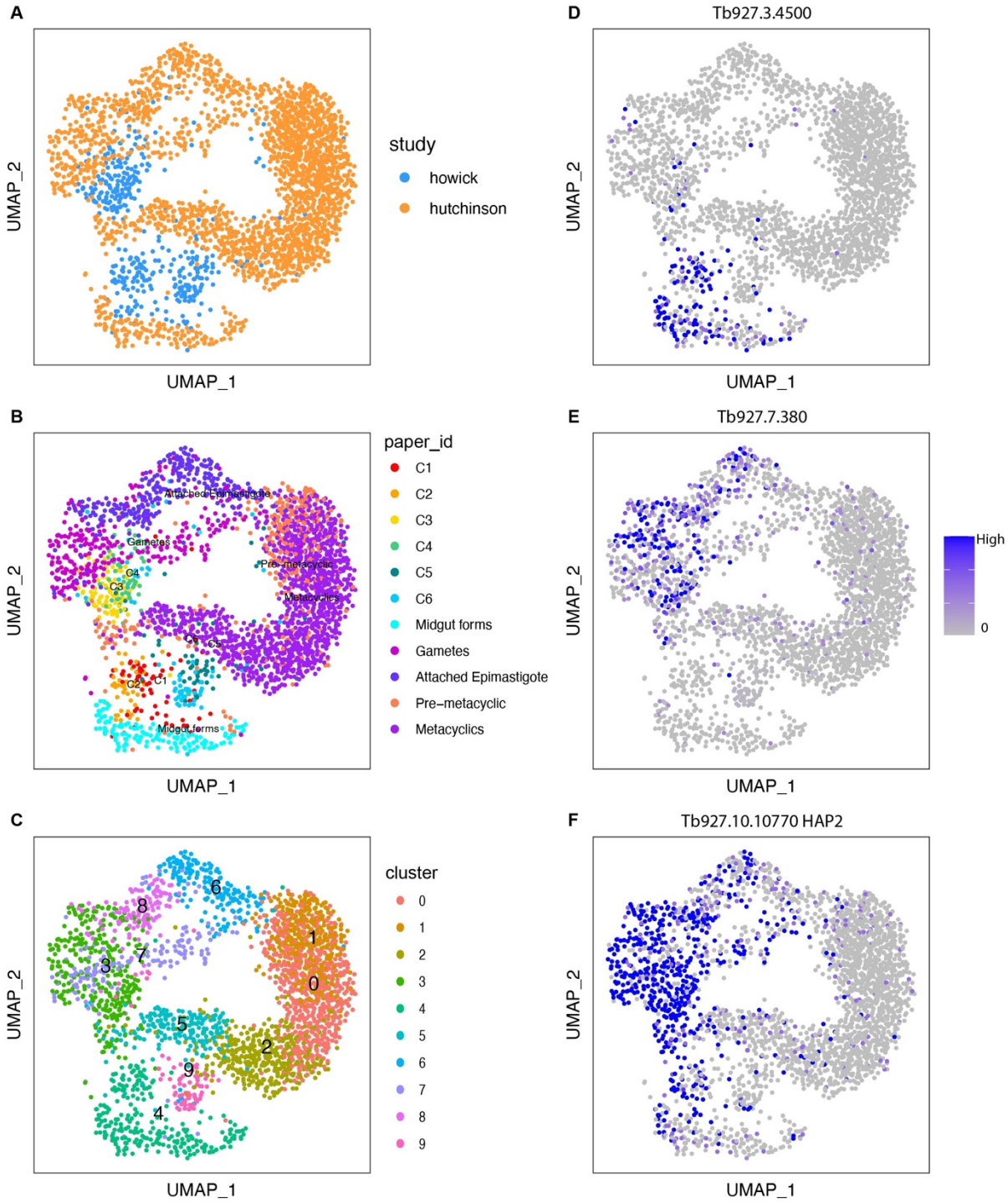

Figure S4. Integration with Hutchinson et al

All 388 tsetse gland transcriptomes were integrated with Hutchinson et al data collected from salivary glands using Seurat's data integration function. Plots show the UMAP of integrated data coloured by study **(A),** cluster identity from the different studies (paper_id) **(B),** integrated cluster assignment **(C),** or gene of interest **(D-F).** *FHc* (Tb927.3.4500) was the top marker gene (based on adjusted p-value) for the midgut and proventricular form cluster 4 **(D).** Tb927.7.380 (hypothetical protein, conserved) was the top marker gene for cluster 3 which contained gamete and epimastigote forms. *HAP2* (Tb927.10.10770) **(F)** was not a marker gene for the gamete cluster likely because of its ubiquitous expression across non-metacyclic forms. Although we were able to identify conserved marker genes across the two studies,

794 separation remained in the UMAP for all cell-types **(A-B)** and only the non-metacyclic forms
795 co-clustered across the two studies and only at a granular level. The metacylic forms likely did
796 not cluster together because of different VSG repertoires, and the separation across other
797 cell-types may be due to time point, strain-specific expression patterns, or collection methods.
798 Conserved marker genes for clusters 3 and 4 can be found in **Table S3.**
799

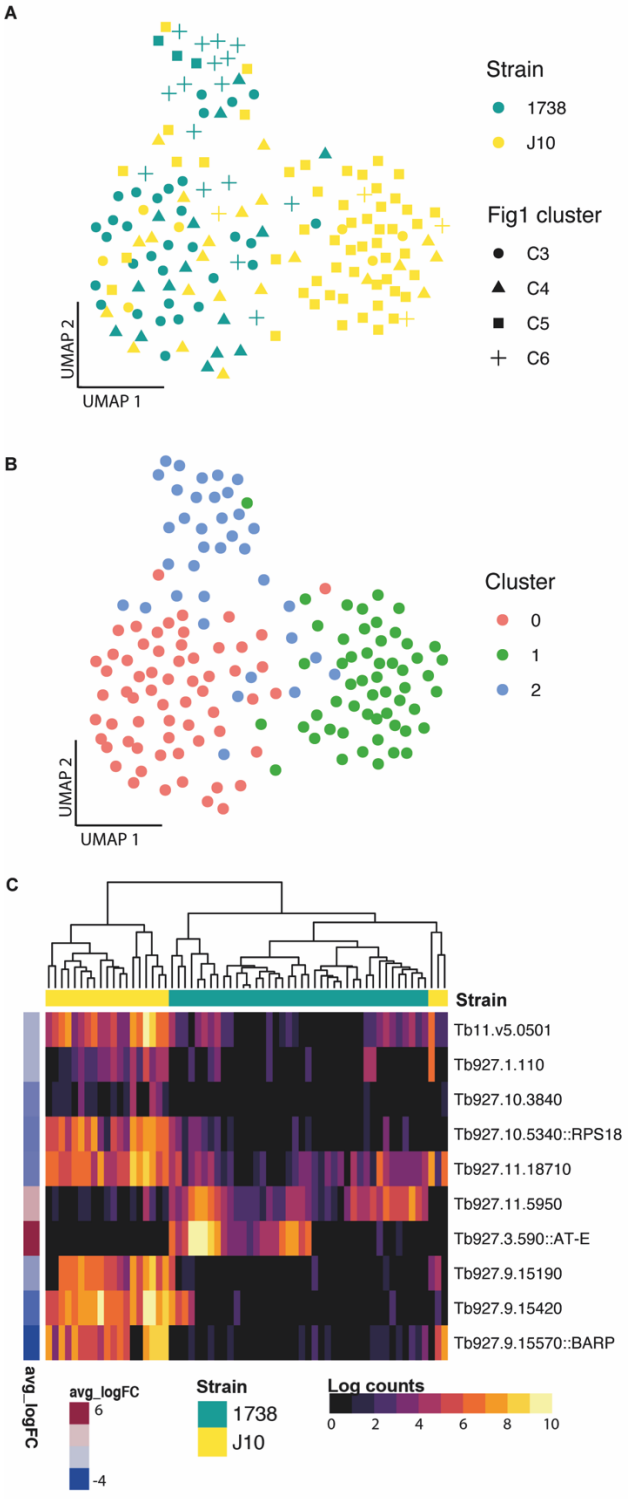

Figure S5. Strain-specific gene expression (A). A UMAP of the day 24 1738 and J10 SG parasites integrated by strain. Points are coloured by strain and shaped by Figure 1 cluster. **(B).** The integrated UMAP coloured by new cluster from the integration analysis. Cluster 0 has a representation of both strains, whereas cluster 1 and 2 are composed primarily of 1738 or J10, respectively. **(C)** Differential expression was performed between strains within cluster 0. The ten genes were differentially expressed between the two strains are displayed on a heatmap.

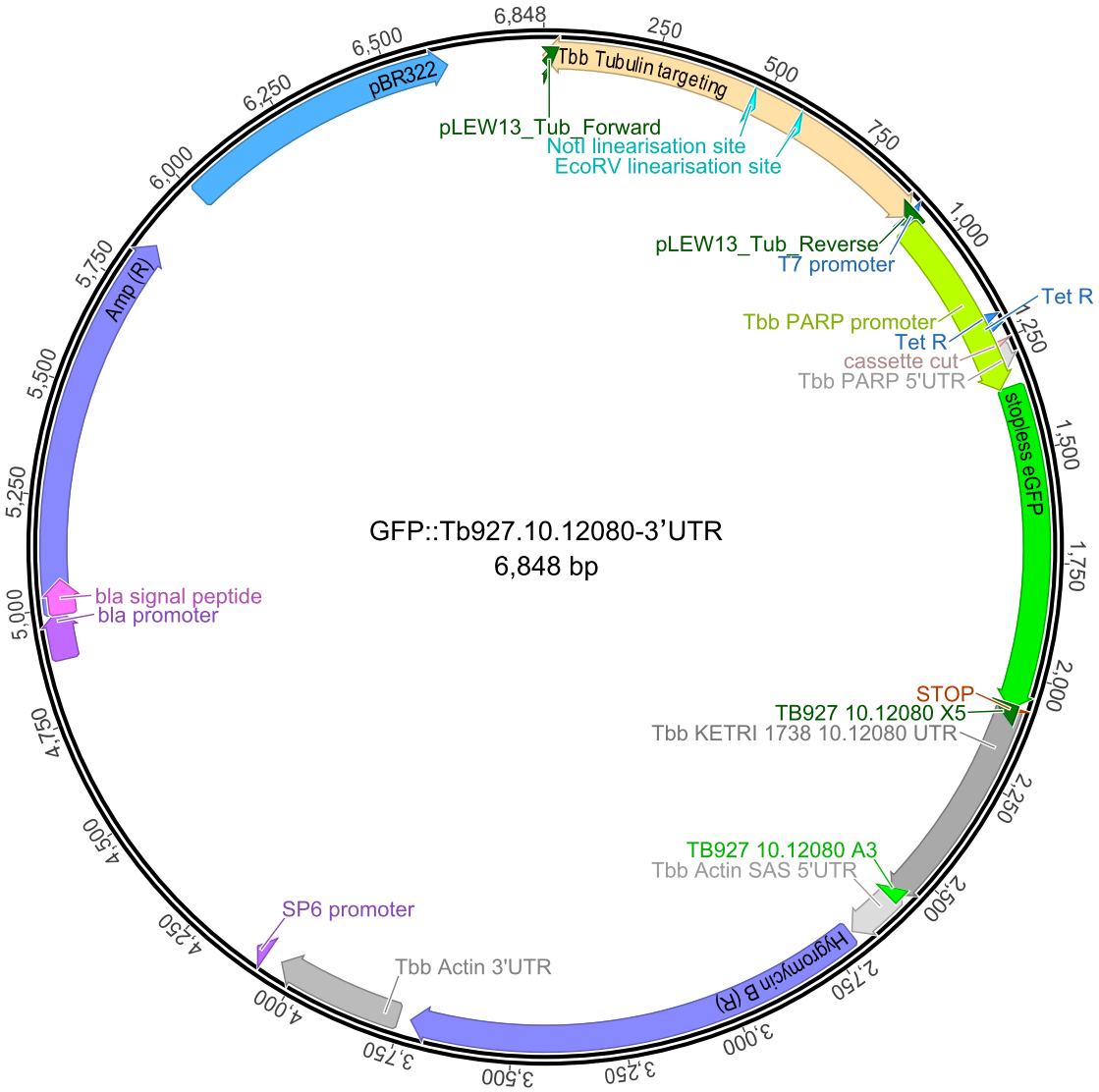

Figure S6. Plasmid map for GFP::Tb927.10.12080-3'UTR construct

Life cycle selective expression of Tb927.10.12080 was investigated through a reporter construct where the expression of GFP was controlled by ~500 bp of UTR downstream of the gene. For this study a stable transformant line was generated in 1738 using the 3' UTR from its endogenous gene and integrated into the tubulin locus.

Supplemental data tables

(A) Table S1 Marker genes from Figure 1
(B) Table S2 Top 200 genes expressed in each cluster
(C) Table S3 Integration with Hutchinson marker genes
(D) Table S4 DE over pseudotime
(E) Table S5 DE between strains
(F) Table S6 IFA data
(G) Table S7 12080 data

827

828 **Table S6.** Surface protein expression of trypanosomes from salivary glands from tsetse
829 dissected 19-21 days post infected feed, using immunofluorescence.

Cell type	GPEET- GFP				EP-RFP			Total	Notes	
	No	V.faint	Faint	Yes	No	Faint	Yes	cells		
1K1N gamete		2	5	8	0	0	15	15	all GPEET&EP	**Total gametes:**
2K1N gamete	2	3	3	11			19	19	82% GPEET&EP	**90% GPEET&EP, 10% only EP**
2K1N meiotic		1		1			2	2	all GPEET&EP	
3K1N				3			3	3		
4K1N				1			1	1		**Total intermediates:**
1K2N		2		2			4	4		**1/22 (5%) only EP;**
2K2N		1		3			4	4		**21/22 (95%) GPEET&EP**
3K2N	1	1		5			7	7		
4K2N				1			1	1		
1K3N				1			1	1		
3K3N				1			1	1		
1K1N epi	2		2	2	1		5	6	1/6(17%)none;1/6(17%)EP;4/6(66%)GPEET&EP	
1K1N short epi			1			1	0	1	GPEET&EP	
1K1N trypo	3	1	1	1			6	6	3/6 (50%) only EP; 3/6 (50%) GPEET&EP	
2K2N div trypo	1				1		0	1	no GPEET or EP	
1K1N meta	6				6		0	6	no GPEET or EP	
Total	15	11	12	40	8	1	69	78		
Cell type	**BARP- GFP**				**EP-RFP**			**Total**	**Notes**	
	No	**V.faint**	**Faint**	**Yes**	**No**	**V.faint**	**Yes**	**cells**		
1K1N gamete	4		1	10			15	15	73% BARP&EP	**Total gametes:**
2K1N gamete	4	3	2	9			18	18	78% BARP&EP.	**76% BARP&EP, 24% only EP**
2K1N meiotic	1		2	2			5	5	80% BARP&EP	
3K1N	1	2					3	3		
1K2N	2		1	3			6	6		**Total intermediates:**
2K2N	1	2	4	5			12	12		**7/36 (19%) only EP;**
3K2N	2			1			3	3		**29/36 (81%) BARP&EP**
4K2N		1					1	1		
5K2N			1	1			2	2		
2K3N				2			2	2		
3K3N		1	2	2			5	5		
4K3N	1			1			2	2		
1K1N epi		1	2	6			9	9	BARP&EP	
2K1N div epi				3			3	3	BARP&EP	
2K2N div epi	2	1		3			6	6	2/6 (33%) only EP; 4/6 (67%) BARP&EP	
1K1N trypo	2	2	1	2			7	7	2/7 (29%) only EP; 5/7 (71%) BARP&EP	
1K1N pre-meta			1	1	1	1	0	2	1/2 (50%) only BARP; 1/2 (50%) BARP&EP	
1K1N meta	8				8		0	8	no BARP or EP	
Total	28	13	17	51	9	1	99	109		

830
831
832

833 **Table S7.** Cell types recovered from tsetse salivary gland exudate 16-21 days post infection
834 with *T. brucei* 1738 expressing GFP::Tb927.10.12080-3'UTR scored for GFP fluorescence.
835

Cell Type	**Fluorescent**	**Non-Fluorescent**	**Total Cells**
Epimastigote/Dividing epimastigote	0 (0%)	86 (100%)	86
Trypomastigote/Dividing trypomastigote	0 (0%)	24 (100%)	24
Asymmetric Divider	0 (0%)	19 (100%)	19
Pre-metacyclic/Dividing pre-metacyclic	0 (0%)	29 (100%)	29
Metacyclic	0	3 (100%)	3
Trypomastigote-Epimastigote divider	0	4 (100%)	4
Meiotic Divider	3 (10%)	26 (90%)	29
Meiotic intermediate 3N	5 (50%)	5 (50%)	10
Meiotic intermediate 2N	28 (80%)	7 (20%)	35
Meiotic intermediate 3K1N or 4K1N	38 (83%)	8 (17%)	46
Meiotic intermediate - final division to two gametes	34 (79%)	9 (21%)	43
1K1N Gamete	11 (65%)	6 (35%)	17
2K1N Gamete	19 (86%)	3 (14%)	22
Unidentifiable	44 (45%)	53 (55%)	97
TOTAL	**182 (39%)**	**282 (61%)**	**464**

836
837