

# A single-cell atlas of *Plasmodium falciparum* transmission through the mosquito

Eliana Real <sup>1,6</sup>, Virginia M. Howick <sup>2,3,4,6</sup>, Farah A. Dahalan<sup>1,6</sup>, Kathrin Witmer<sup>1,2</sup>, Juliana Cudini<sup>2</sup>, Clare Andradi-Brown<sup>1,5</sup>, Joshua Blight<sup>1</sup>, Mira S. Davidson <sup>1</sup>, Sunil Kumar Dogga<sup>2</sup>, Adam J. Reid<sup>2</sup>, Jake Baum <sup>1</sup>✉ & Mara K. N. Lawniczak <sup>2</sup>✉

Malaria parasites have a complex life cycle featuring diverse developmental strategies, each uniquely adapted to navigate specific host environments. Here we use single-cell transcriptomics to illuminate gene usage across the transmission cycle of the most virulent agent of human malaria - *Plasmodium falciparum*. We reveal developmental trajectories associated with the colonization of the mosquito midgut and salivary glands and elucidate the transcriptional signatures of each transmissible stage. Additionally, we identify both conserved and non-conserved gene usage between human and rodent parasites, which point to both essential mechanisms in malaria transmission and species-specific adaptations potentially linked to host tropism. Together, the data presented here, which are made freely available via an interactive website, provide a fine-grained atlas that enables intensive investigation of the *P. falciparum* transcriptional journey. As well as providing insights into gene function across the transmission cycle, the atlas opens the door for identification of drug and vaccine targets to stop malaria transmission and thereby prevent disease.

<sup>1</sup>Department of Life Sciences, Imperial College London, London, UK. <sup>2</sup>Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton, UK. <sup>3</sup>Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. <sup>4</sup>Wellcome Centre for Integrative Parasitology, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. <sup>5</sup>Department of Infectious Disease, Imperial College London, London, UK. <sup>6</sup>These authors contributed equally: Eliana Real, Virginia M. Howick, and Farah A. Dahalan. ✉email: [jake.baum@imperial.ac.uk](mailto:jake.baum@imperial.ac.uk); [mara@sanger.ac.uk](mailto:mara@sanger.ac.uk)

Malaria parasite transmission to and from the mosquito vector is accompanied by a precipitous drop in parasite population numbers<sup>1,2</sup>. The bottleneck that transmission affords has, therefore, long been considered an attractive target for antimalarial or vaccine interventions. In a highly orchestrated series of developmental transitions taking place inside the mosquito, parasites taken up during a blood meal complete sexual development, fertilization, and recombination; differentiate into invasive ookinetes that colonize the midgut; develop into oocysts where population expansion occurs; and emerge as invasive sporozoites that fill the mosquito salivary glands and can be transmitted to a new host during a mosquito bite<sup>3</sup>. After a relatively brief stint in the skin, some injected sporozoites will reach the liver, where they undergo a phase of profound population expansion, followed by a release of merozoites that enter the bloodstream<sup>4</sup>. This heralds the beginning of the cyclical asexual cycle, where merozoites invade erythrocytes and develop intra-erythrocytically, further expanding parasite numbers, with some parasites also developing into sexual, transmissible forms that can perpetuate the transmission cycle when taken up in the blood meal of a mosquito.

Single-cell RNA sequencing (scRNA-seq) has transformed our ability to resolve cell-level heterogeneity in complex cell populations. Recent scRNA-seq views across the full life cycle of *Plasmodium* parasites, the etiological agents of malaria, have captured fine-scale developmental transitions driving progression through the life cycle at high resolution<sup>5–9</sup>. The Malaria Cell Atlas<sup>6,8</sup> was established as a data resource and website to provide an accessible route for scientists to explore patterns of gene expression in individual parasites across multiple developmental stages and parasite species. However, its application to *P. falciparum*, the species responsible for the vast majority of human death and disease has been thus far limited to the blood stages. Here we add a chapter to the atlas, completing a scRNA-seq survey of the transmission stages of *P. falciparum* from the sexual forms transmitted to the *Anopheles* mosquito to the invasive sporozoite delivered to the human host. Understanding these key developmental transitions in parasite biology is essential in order to devise innovative strategies for tackling malaria, a task rendered more urgent by the ongoing spread of antimalarial drug resistance and lack of a vaccine that affords long-term protection.

## Results

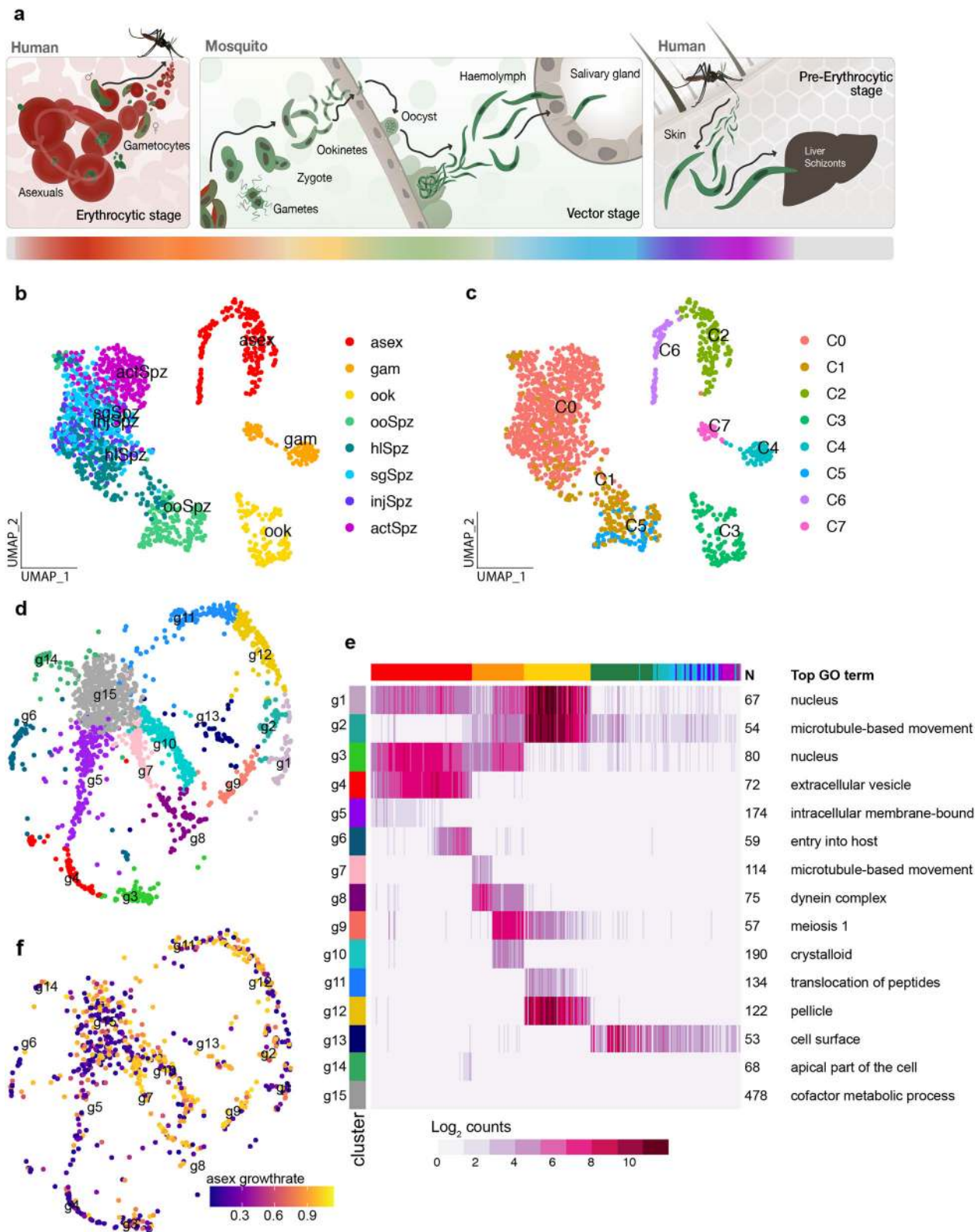
**Single-cell transcriptomics informs patterns of gene usage across the life cycle.** To capture human malaria transmission at single-cell resolution, we assembled a comprehensive scRNA-seq data set across key stages of the *P. falciparum* life cycle (Fig. 1a). Distinct morphological forms of the parasite were purified based either on the expression of known stage-specific markers (circumsporozoite protein (CSP) and ookinete surface protein P25 (Pfs25) for sporozoites and ookinetes, respectively) or the RNA and DNA content (stage V gametocytes) (Supplementary Figs. 1–3). Poor quality cells were excluded per stage based on the distribution of the number of reads and genes detected per cell (Supplementary Fig. 4; Supplementary Table 1; Supplementary Data 1–3). As observed previously<sup>8</sup>, the number of genes detected was dependent on parasite stage, ranging from 119 in sporozoites to 1497 in gametocytes. This is consistent with smaller cells having less total mRNA<sup>10</sup>, and likely reflects biological differences in the mRNA content of each cell type<sup>10</sup>. After quality control, our data set comprised 1467 single-cell transcriptomes, of 1808 originally sequenced, spanning the broad developmental transitions that underpin host-to-vector and vector-to-host transmission. We next integrated this transmission data set with our published collection of 161 single-cell transcriptomes of mixed

*P. falciparum* asexual stages<sup>6</sup>, which enabled us to reconstitute the *P. falciparum* life cycle almost entirely, with the exception of exoerythrocytic schizogony (liver stages). In order to visualize *P. falciparum* single cells across the life cycle, we performed dimensionality reduction using UMAP and clustered cell transcriptomes using Louvain clustering<sup>11–13</sup>. As expected, cells from the same stage grouped together in two-dimensional space (Fig. 1b) and formed distinct cell clusters (Fig. 1c; Supplementary Data 2). We were also able to observe clear cellular trajectories that are likely to reflect within-stage developmental paths (Fig. 1b, c; Supplementary Data 2). For example, the distribution of asexual parasites in a two-dimensional space clearly shows developmental progression along the intraerythrocytic developmental cycle (IDC) (Supplementary Fig. 5)<sup>6</sup>. Likewise, the trajectory displayed by ookinete cells is consistent with a range of developmentally intermediate states along the known ookinete maturation path (Supplementary Fig. 6)<sup>14</sup>.

To globally understand patterns of gene co-expression across the life cycle, we built a k-nearest neighbors (kNN) graph using the top highly variable genes (HVGs) in the data set. Each gene was assigned to one of 15 co-expression modules using graph spectral clustering (denoted with g, Fig. 1d; Supplementary Data 1), allowing us to predict gene function using a “guilt-by-association” approach. The majority of gene clusters showed stage-specific expression patterns and were enriched for GO terms expected for each stage (Fig. 1e; Supplementary Fig. 7; Supplementary Data 4). To further validate the predictive value of this approach, we overlaid our gene graph with data from two genome-wide essentiality screens in *Plasmodium* and found that clusters enriched for dispensable genes (normal-growth rate in ref. <sup>15</sup> and mutable in ref. <sup>16</sup>) were highly expressed in transmission stages and those enriched for essential genes were highly expressed in asexual blood stages (Fig. 1e, f; Supplementary Fig. 8; Supplementary Table 2). Taken together, this links gene usage across the life cycle to stage-specific functions allowing us to predict functionality based on co-expression patterns.

Other gene clusters were more ubiquitously expressed, possibly reflecting shared functional strategies across the life cycle. For instance, cluster g1, which features two putative readers of post-translational histone modifications, namely bromodomain protein 1 (BDP1) and the EELM2-domain-containing PF3D7\_0519800, contains genes that tend to be expressed throughout the IDC and sexual stages, where epigenetic control of developmental processes has long been recognized<sup>17,18</sup>. Here, we find that these genes are also expressed in ookinetes, implying that epigenetic mechanisms may be more widely used across the life cycle<sup>19</sup>. Both BDP1 and EELM2 proteins have known or predicted links to apicomplexan AP2 transcription factors (ApiAP2 TFs)<sup>20–22</sup>. Significantly, two ApiAP2 TFs for which interactions with protein complexes containing BDP1 or EELM2 have been experimentally demonstrated are also expressed in this cluster (PfsAP2Tel/SP3 and PF3D7\_1107800, respectively)<sup>21,22</sup>. In addition, BDP1 has been shown to associate with the promoter regions of seven ApiAP2 TFs<sup>20</sup> whose expression we find to be correlated with gene clusters associated with asexual development and host-to-vector transmission (Supplementary Fig. 9; see Supplementary Data 5 for an analysis of enriched DNA motifs in each cluster), thus expanding the reach of this gene regulatory network. A total of 18 genes without predicted functions are co-transcribed in cluster g1 and as such could represent additional links between chromatin remodeling/targeting and AP2-dependent gene expression (Supplementary Fig. 10; Supplementary Data 1).

The majority of genes in the *P. falciparum* genome have one-to-one orthologs with *P. berghei* (80%). We, therefore, explored whether any clusters were enriched for genes that lack orthologs



between the two species. This analysis revealed that every cluster had some clade-specific genes present, possibly reflecting species-specific adaptive mechanisms present across the life cycle (Supplementary Figs. 11a, b, and 12). Cluster g15, which is the largest gene cluster by more than twofold showed low levels of expression across the life cycle and had the highest proportion of genes lacking orthologs, with a large number of *P. falciparum*

variant multigene families (e.g., *stevors*, *rifs*) associated with host interactions in blood stages (GO:0018995, Bonferroni-adjusted *p* value =  $3.58 \times 10^{-9}$ )<sup>23</sup>. Clusters associated with mosquito stages and vector-to-host transmission (g1, g2, g9, g12, g13), in contrast, had proportionally more genes with *P. berghei* orthologs. Despite being enriched for one-to-one orthologs, two of these clusters (g2, g13) showed the lowest amino-acid conservation scores across the

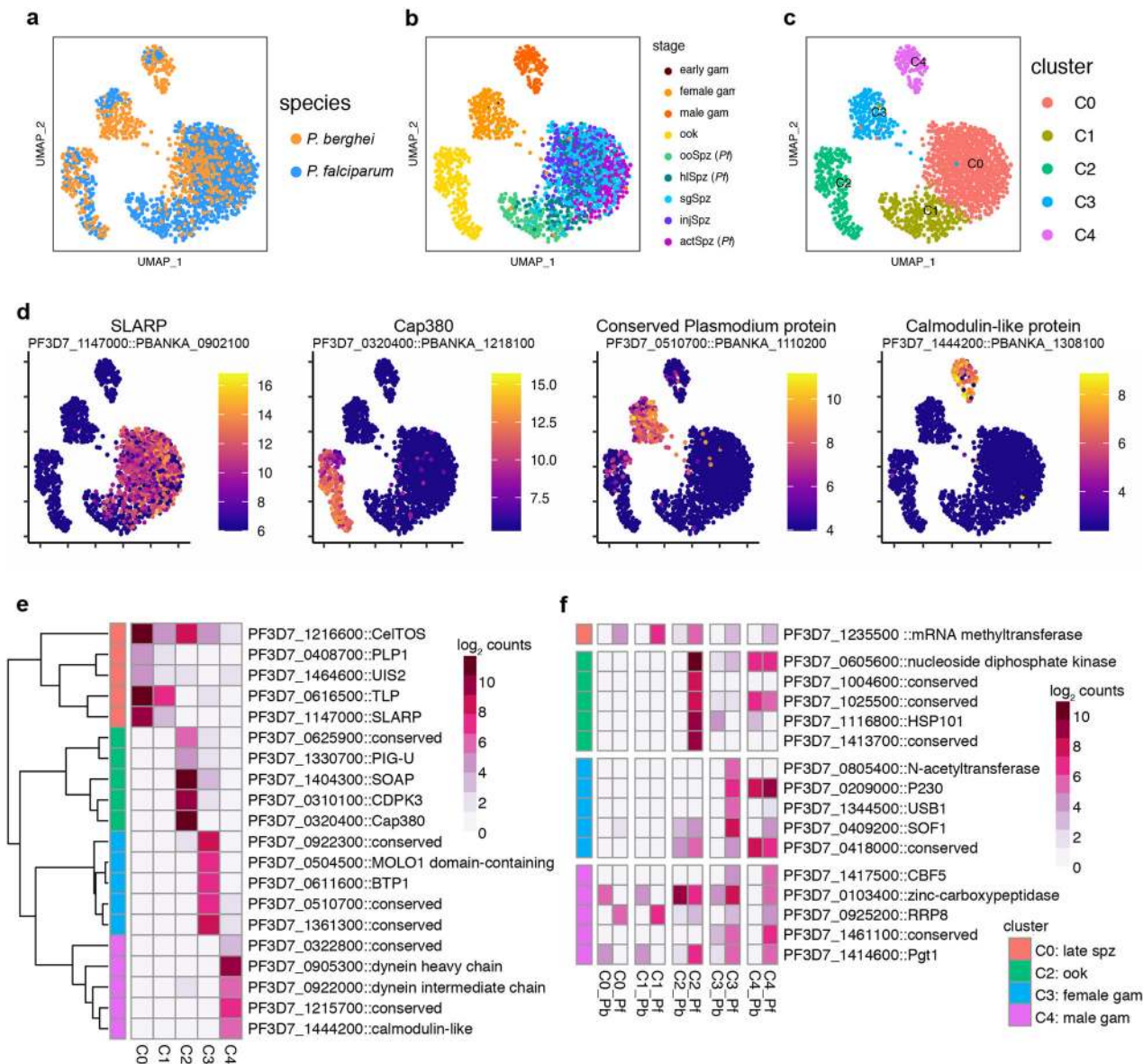
**Fig. 1 A single-cell transcriptome atlas of *P. falciparum* transmission.** **a** The life cycle of *P. falciparum* includes asexual replication and sexual development in red blood cells (erythrocytic stages). Sexual stage gametocytes are then taken up in the mosquito blood meal, where they undergo fertilization and develop into the ookinete form that crosses the mosquito midgut epithelium and begins sporogonic development along the basal lamina as an oocyst. Sporozoites are released from mature oocysts to the hemolymph where they travel to and invade the salivary glands. A subset of these sporozoites may be injected into the next human host where they migrate through the dermis on their way to the liver (pre-erythrocytic stages). The color bar below represents the different life cycle stages reflected in **(b)**. **b** UMAP of single-cell transcriptomes from across asexual and transmission stages of the parasite. Stage abbreviations: asex (asexual blood stages), gam (male and female gametocytes), ook (ookinete), ooSpz (oocyst sporozoites), hlSpz (hemolymph sporozoites), sgSpz (salivary gland sporozoites), injSpz (sporozoites released by mosquito bite), actSpz (activated sporozoites). Asexual stage transcriptomes have been previously published<sup>6</sup>. **c** Louvain clustering of single-cell transcriptomes identified eight distinct clusters. In order to get a more equal distribution of numbers of cells across clusters, clusters C0 and C1 were downsampled to 100 cells each prior to making the gene graph **(d-f)**. **d** A k-nearest neighbors (kNN) force-directed graph of 1797 highly variable genes across cell types. Each node represents a gene. Nodes are colored by their graph-based spectral clustering assignment (g1-g15). **e** A heatmap of mean expression of the genes in each cluster across 623 subsampled cells. Cells are ordered by their developmental progression along the life cycle. Cells from asexual, ookinete, and sporozoite stages were ordered according to their computed pseudotime values (Supplementary Figs. 5 and 6). Gametocytes are grouped based on cluster assignment from **(c)**, where C7 are male gametocytes and C4 are female/early gametocytes (see Supplementary Fig. 2 for more detailed analysis). The male gametocyte cells are ordered before the female gametocytes in the heatmap. Stages are represented by the colors in **(a, b)**. The number of genes (*N*) and the top-enriched GO term (Top GO term, FET one-sided) in each cluster is indicated next to the corresponding row. Bonferroni-adjusted *p* values for this analysis are shown in Supplementary Data 4. **f** The gene graph colored by the relative growth rate of orthologous knockout mutants in *P. berghei* asexual blood-stage parasites<sup>15</sup>.

gene graph (Supplementary Fig. 11c, d; Supplementary Table 2). In addition, g13, which features genes encoding surface proteins with essential roles in transmission (Supplementary Data 4), had the highest median global differentiation score<sup>24</sup> (Supplementary Fig. 11e). This suggests that protein interactions that underpin transmission might have been shaped by coevolution of parasite and host and may account, at least in part, for host specificity/tropism in *Plasmodium* infections.

**Integration across species reveals conserved and divergent patterns of gene expression.** To further explore the conservation of genes involved in *Plasmodium* transmission, we compared *P. falciparum* transmission stages with the equivalent *P. berghei* stages from Howick et al.<sup>25</sup> using scmap. The two data sets showed a good overall correlation, with *P. falciparum* cells mapping to cells of the same stage in the *P. berghei* reference data set (Supplementary Fig. 13a, b)<sup>8</sup>. As expected, cells that were not represented in the *P. berghei* transmission analysis (early sporozoite developmental stages), showed lower similarity across species than those that were common to the two data sets (Supplementary Fig. 13c). This initial comparison enabled us to assert that differences between experimental workflows, which in our case were optimized to purify non-genetically modified parasites and thus differed slightly from those used by Howick et al.<sup>12,13</sup> were not a likely source of variability between the data sets. We next used an unsupervised integration and graph-based clustering approach in Seurat to elucidate gene expression patterns across the transmission stages of the two species. Using this approach, cells from the two data sets could be matched to four clusters, with cells mixing by stage in the life cycle (Fig. 2a–c). A fifth cluster (C1) was populated by *P. falciparum* midgut and hemolymph sporozoites that were matched by very few cells in the *P. berghei* data set (Fig. 2a–c). The global structure of the integrated data set remained largely unchanged compared with the original *P. falciparum* subset, underlining the robustness of the cell identity correlations. For each cluster, we identified marker genes that are conserved between species (Fig. 2d, e). Among genes that represent the conserved transcriptome (Fig. 2d, e; Supplementary Data 1), 30–40% have no known function, but as exemplified for the ookinete subset (Supplementary Fig. 6f), we can begin to make predictions about gene function by surveying correlated gene expression patterns along highly resolved developmental trajectories. The same strategy could help to fast-track the discovery of candidates for therapeutic development, for example, by uncovering genes with expression profiles that are highly correlated with those of validated vaccine targets (e.g., Pfs25, CSP)<sup>26</sup>.

In addition to shared patterns in genome activity with broadly conserved *Plasmodium* genes, our integrated analysis also revealed non-conserved patterns of gene expression. In all stages, we identified genes whose transcripts could be detected in over 70% of *P. falciparum* cells but showed no or only residual expression in *P. berghei* cells of the same stage (Fig. 2f; Supplementary Data 1). Parallel analysis of previously published asexual blood-stage transcriptomes from both species<sup>6,8</sup>, as well as different strains of *P. falciparum*<sup>27</sup>, showed that genes that were differentially expressed between species were not *P. falciparum* strain-specific in expression (Supplementary Fig. 14). This suggests that species-specific expression patterns observed in transmission stages are not driven by intra-species genotypic variation in gene expression. Together, with the genes that do not have orthologs in *P. berghei* (Supplementary Fig. 11a, b; Supplementary Data 1), these differentially expressed genes may represent species-specific adaptations that allow *P. falciparum* to interact optimally with the diverse host environments it meets as it transmits from vector-to-host and vice-versa. For example, *P. falciparum* ookinetes (C2\_Pf in Fig. 2f) displayed a striking pattern of species-specific expression with 26 genes (several of which have no known function) expressed in >70% of Pf cells and in <10% of Pb cells (Supplementary Data 1). Conversely, in sporozoites, a single one-to-one ortholog—encoding a putative mRNA methyltransferase (PfMT-A70.2)—showed *P. falciparum*-specific expression (Fig. 2f). PfMT-A70.2 was recently identified as part of a methyltransferase complex that modulates gene expression during the IDC through N<sup>6</sup>-methyladenosine methylation of mRNA transcripts<sup>28</sup>. Our observation that PfMT-A70.2 expression is highly correlated with mosquito stages in *P. falciparum* but not in *P. berghei* (Fig. 2f), hints at unexplored species-specific mechanisms to fine-tune gene expression during the parasite journey in the mosquito.

**Trajectory analysis of sporozoite development identifies a putative egress network.** Understanding how sporozoite infectivity towards the human host develops is likely to inform the identification of untested antimalarial and vaccine candidates. We, therefore, sampled these stages more densely to capture a finely resolved developmental trajectory that recapitulated the sporozoite journey through the mosquito, with its different anatomical, cellular, and physicochemical environments<sup>29</sup>. We initially re-clustered sporozoite cells and confirmed through differential abundance testing that cluster composition was not affected by experimental batch effects (Supplementary Fig. 15).

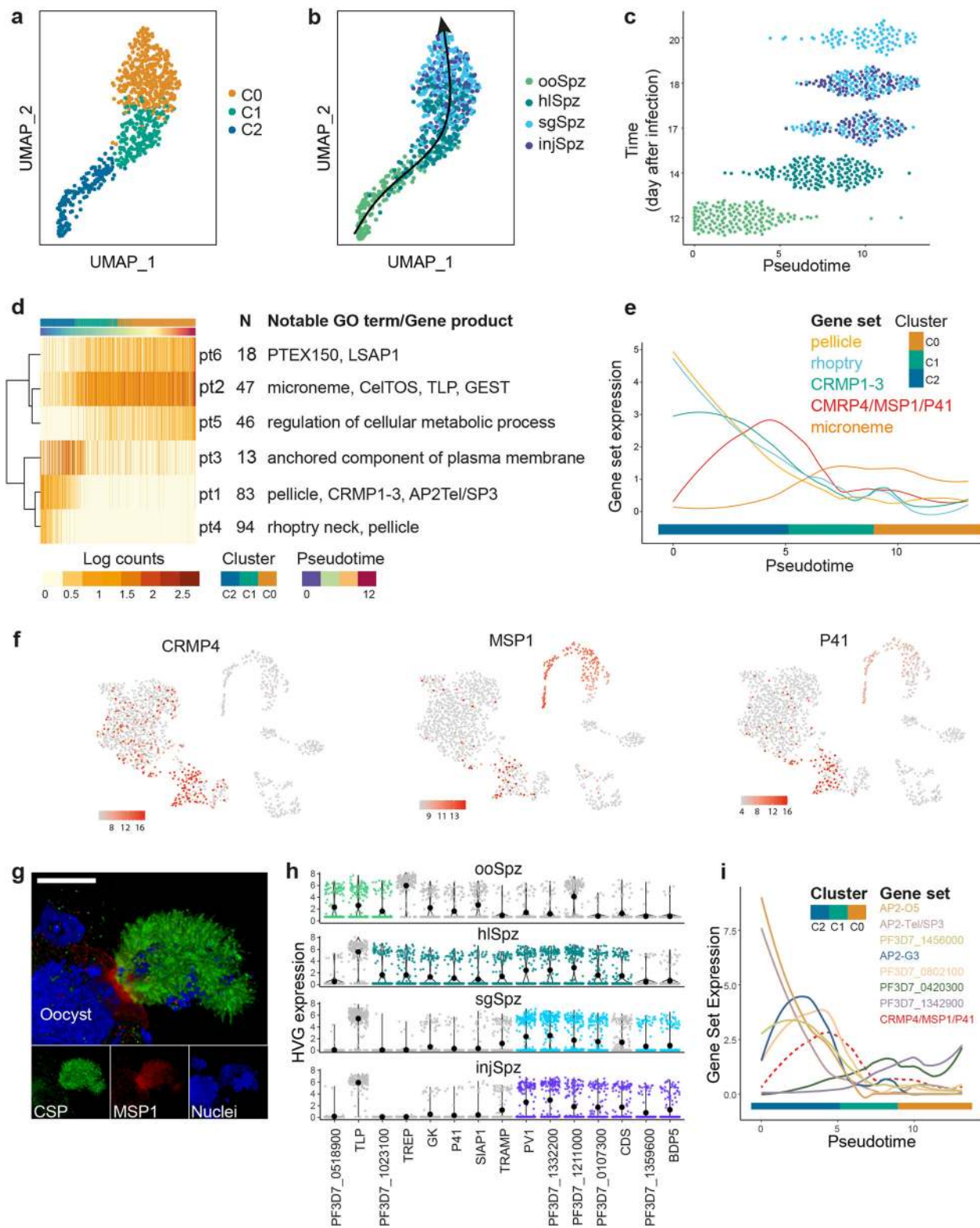


**Fig. 2 Integration of *P. falciparum* and *P. berghei* transmission data sets.** To identify conserved and divergent patterns of gene expression *P. falciparum* single-cell data was integrated with the *P. berghei* transmission stage data using one-to-one orthologs. **a–c** UMAP of integrated data colored by species (**a**), collection stage (**b**), and cluster assignment (**c**). Collection stages that were not included in the *P. berghei* data set are marked in legend with (*Pf*) in (**b**). Cluster C1 was not included in further differential expression analysis because it was composed primarily of *P. falciparum* oocyst and hemolymph sporozoites. We observed that cells clustered by cell types that were shared across the two data sets. *P. falciparum* gametocytes appeared more tightly clustered compared to *P. berghei*, within the gametocyte clusters, likely owing to differences in sample sizes for these stages (Pb female: 127, Pb male: 74, Pf female: 52, Pf male: 33). **d** Expression (log<sub>2</sub> counts) of the top-conserved marker gene, based on the lowest maximum *p* value (Wilcoxon rank-sum test) across the two species, for clusters 0 (SLARP), 2 (Cap380), 3 (conserved *Plasmodium* protein PF3D7\_0510700), and 4 (calmodulin-like protein) on the UMAP. **e** Heatmap of the top five conserved marker genes based on Bonferroni-adjusted *p* value (Wilcoxon rank-sum test) across the two species, for each cluster. The expression values represent the mean log counts for each cluster. **f** A heatmap of the top one-to-one orthologs that are more highly expressed in *P. falciparum* (*Pf*) compared with *P. berghei* (*Pb*) (based on adjusted *p* value and expression in at least 70% of *P. falciparum* cells). The expression values represent the mean log counts for each species in each cluster. Heatmaps are annotated with the *P. falciparum* gene identifier for the ortholog and manual annotation. Genes annotated as conserved are conserved proteins with unknown function in *Plasmodium*.

Next, sporozoites were oriented along a pseudotime trajectory where the relative ordering of each individual cell was defined by its transcriptional state (Fig. 3a, b; Supplementary Data 2). This enabled us to explore fine transcriptional changes associated with a continuum of development instead of discrete time points, which were observed to be heterogeneous (Fig. 3c). The pseudotime trajectory aligned with the expected direction of development, with cells broadly segregating by their anatomic site and developmental time point (Fig. 3b, c). Differential expression

(DE) analysis uncovered 301 genes with expression significantly associated with pseudotime (*q* value < 0.0001; Supplementary Data 1). Approximately 20% of these DE transcripts overlapped with genes previously identified as upregulated in oocyst sporozoites (UOS) or upregulated in infectious sporozoites (UIS) (Supplementary Fig. 16a, b; Supplementary Data 1)<sup>30</sup>, but many others have never been linked with sporozoite biology.

To reveal patterns of gene activity along the pseudotime trajectory, we grouped genes with similar expression dynamics



into co-expression modules (denoted pt1–6, Fig. 3d; Supplementary Data 1). This analysis revealed two largely mutually exclusive transcriptional programs, with the midgut-to-hemocoel transition (C2–>C1) marking an abrupt shift in gene expression. Gene modules (pt1,4) featuring known salivary gland invasion genes (Cysteine Repeat Modular Protein, CRMP1-2)<sup>31</sup>, components of the rhostry neck and inner membrane complex were the first to downregulate expression. This was followed by a transient peak

(pt3) showing high temporal correlation with sporozoite release into the hemocoel and preceding the transcriptional changes that characterized later development (Fig. 3d, e). Of note, one of the 13 genes that constitute the pt3 module, CRMP4, has been linked to sporozoite egress in *P. berghei*<sup>32</sup>. Two other genes within this module—MSP1 and the 6-Cys family member p41<sup>33,34</sup>—encode merozoite surface proteins that have never been associated with mosquito stages. Although the function of p41 has not been

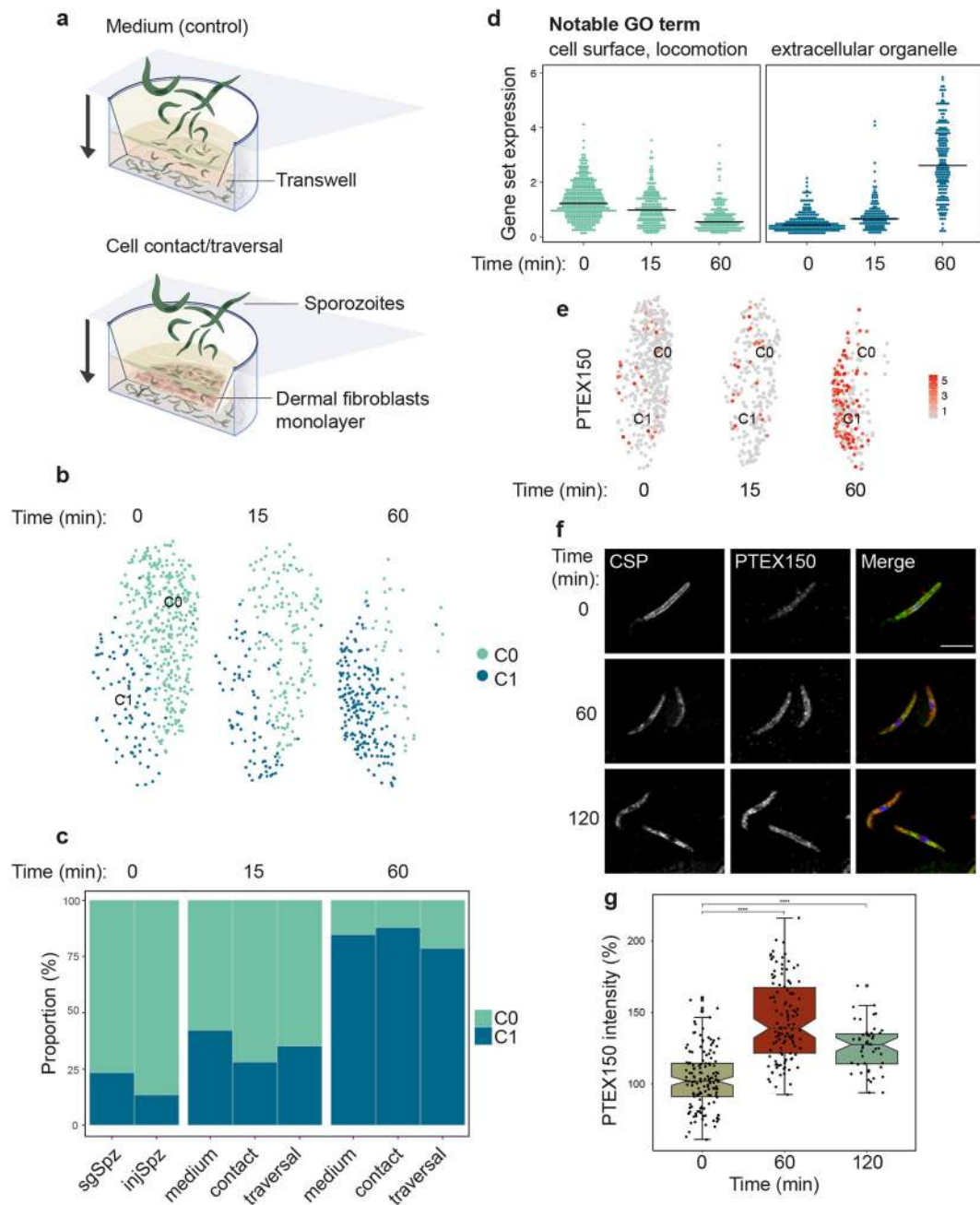
**Fig. 3 The sporozoite journey in the mosquito.** **a, b** UMAP representation of single-cell sporozoite transcriptomes at different stages of the sporozoite developmental journey colored by cluster assignment (**a**) and sorted stage (**b**). The arrow in **B** represents the direction of development. **c** Sporozoites aligned along pseudotime. Overall, the pseudotime trajectory correlates with the direction of development indicated by the time of sample collection in the *y* axis. **d** Heatmap showing the mean expression of differentially expressed (DE) genes over pseudotime, arranged into six pseudotime clusters (pt1–pt6). Each row in the heatmap represents the average expression of all genes in one cluster (the number of genes in each cluster, *N*, is indicated). Genes in each cluster show similar expression dynamics along the developmental trajectory. **e** Gene modules representative of enriched GO terms (see Supplementary Data 4 for Bonferroni-adjusted *p* values) in pseudotime clusters are shown. Transient expression of CRMP4, p41, and MSP1 (GO term anchored component of plasma membrane in **d**) correlates with the release of sporozoites into the hemocoel. **f** UMAP representations of the complete life cycle (as shown in **1b**) with the expression (log counts) of CRMP4, p41, and MSP1 highlighted. **g** 3D rendering of a 12-day oocyst immunostained with anti-CSP (green), anti-MSP1 (red), and Hoechst (blue). The image captures the egress of sporozoites (CSP positive) from the oocyst. Released sporozoites express MSP1 on their surface. Note that the lack of signal in sporozoites that remain inside the oocyst is likely a result of poor antibody penetration across the oocyst wall. Scale bar, 10  $\mu$ m. The image is representative of 6 egress events (out of 9), each releasing hundreds of sporozoites, captured from 19 infected midguts in two independent infections. **h** Expression of highly variable genes (HVG) across sporozoite developmental stages. HVGs were independently determined for each stage using M3Drop with a false discovery rate (FDR) < 0.05. Only HVGs that are also associated with the pseudotime trajectory are shown. HVGs are colored by their stage association, as per (**b**). **i** Expression of DE ApiAP2 TFs over pseudotime. AP2-Tel/SP3 and AP2-O5 expression levels steadily decline from an earlier peak in ooSpz until egress (marked by the C2–C1 transition and expression of CRMP4/MSP1/P41). Three ApiAP2 TFs are transiently expressed before egress, whereas another two are upregulated after egress.

elucidated, MSP1 facilitates merozoite egress by destabilizing the erythrocyte submembrane cytoskeleton<sup>35</sup>. Its co-regulation with CRMP4 suggests an unexpected mechanistic link between sporozoite and merozoite egress (Fig. 3f). The existence of a functional link to egress is further supported by evidence that MSP1 is expressed on the surface of sporozoites at the time of their release from the oocyst (Fig. 3g; Supplementary Movie 1), which not only validates our scRNA-seq findings but also places MSP1 within the right spatial and temporal window for such a role. Given their highly correlated expression, we hypothesize that genes in this module form the core of a sporozoite egress functional network and anticipate that future studies into the role of MSP1 may uncover parallels between events that culminate in the rupture of the oocyst wall and that of the erythrocyte membrane.

**HVGs highlight heterogeneity of sporozoite populations.** After their release into the hemocoel, sporozoites experienced a dramatic shift in gene expression. Several transcripts previously identified as UIS<sup>30</sup> begin to be expressed during this stage or, in some cases (e.g., CelTOS, TLP), even before egress (Fig. 3d; Supplementary Fig. 16a; Supplementary Data 1). Approximately half of the transcripts expressed or upregulated after egress are absent from recently published sporozoite proteomes (Supplementary Fig. 16d; Supplementary Data 1)<sup>30</sup>, which would suggest that at least some might be under translational repression, although further work will be necessary to determine case-by-case, if lack of detection in proteomics studies truly represents the absence of protein expression. A small proportion (6%) of developmentally regulated transcripts show highly variable expression between individual parasites, with salivary gland and injected sporozoites displaying significant cell-to-cell heterogeneity in the expression of genes implicated in protein export to the host cell (PV1<sup>36</sup>) and epigenetic mechanisms (BDP5<sup>37</sup>), amongst others (Fig. 3h; Supplementary Data 1). A pairwise correlation analysis did not find evidence of co-regulated expression of HVGs, suggesting that HVGs are not co-transcribed in a subset of sporozoites, and, instead, individual sporozoites activate the expression of different HVGs (Fig. 3h; Supplementary Fig. 17; Supplementary Data 1). Given the strong association of these HVGs with sporozoite development, it is tempting to speculate that the observed differences in expression between individual sporozoites could result in different readiness levels with respect to navigating the journey to the liver and/or establishing a new replication niche once they have reached their destination.

**Trajectory analysis implicates multiple ApiAP2 transcription factors in sporozoite development.** Consistent with the key role of ApiAP2 TFs in driving all major developmental switches during the parasite life cycle<sup>38</sup>, seven ApiAP2 transcripts were found to be dynamically regulated along the sporozoite developmental trajectory (Supplementary Fig. 18). Of the five ApiAP2 TFs expressed during late midgut sporogony (cluster C2), only one—PfalP2-Tel/SP3 (PF3D7\_0622900)—had been previously implicated in sporozoite development. Gene knockout studies in *P. berghei* showed that AP2-Tel/SP3 has a role in sporozoite release<sup>39</sup> and indeed, we observed that PfAP2-Tel/SP3 expression preceded the transcription of egress genes (Fig. 3i). Future studies should examine whether PfAP2Tel/SP3 has a role in regulating the expression of egress-related genes and to what extent recent findings mapping PfAP2Tel/SP3 to the telomeric ends of *P. falciparum* chromosomes<sup>21</sup> might be linked to this putative activity. In contrast to PfAP2-Tel/SP3, AP2-G3, and AP2-O5, which have known roles in gametocytogenesis and in regulating ookinete motility, respectively, have not been previously implicated in sporozoite development<sup>40</sup>. Together with AP2-G3, two CA-repeat-binding ApiAP2 transcription factors<sup>41</sup>, PF3D7\_0802100, and PF3D7\_1456000, show a peak in expression prior to egress (Fig. 3i; Supplementary Fig. 18), raising the possibility that all three might regulate the expression of genes necessary for sporozoite release or, alternatively, drive the transcriptional response post egress. Expression of the putative transcription factors PF3D7\_0420300 and PF3D7\_1342900, in contrast, did not occur until after egress (Fig. 3i; Supplementary Fig. 18). Attempts to knock out the former in *P. falciparum* have previously proved unsuccessful, suggesting PF3D7\_0420300 has an essential role in asexual development<sup>16</sup>. How ApiAP2 TFs active in multiple stages of the life cycle achieve stage-specific outcomes is not well understood but may involve combinatorial binding between members of the ApiAP2 family<sup>38</sup>. It is noteworthy that ApiAP2 TF expression patterns during sporogony are consistent with the AP2 TF cascade model, whereby ApiAP2 TFs are themselves under the control of other ApiAP2 TFs. Thus, our analysis expands the constellation of ApiAP2 TFs with putative roles in sporozoite development and suggests that a far more complex regulatory network than previously thought—involving specific combinations of ApiAP2 TFs possibly acting cooperatively—might be required to orchestrate the dynamic transcriptional patterns associated with the sporozoite journey in the mosquito.

**A shift in the host environment corresponds with gradual changes in gene expression.** As sporozoites migrate away from



the site of injection, they can spend up to 2 hours probing their environment, actively migrating in and out of dermal fibroblasts and phagocytic cells in search of a blood vessel through which they can gain access to the bloodstream<sup>42–44</sup>. Being devoid of a protective vacuole, sporozoites migrating in the dermis are particularly exposed to the onslaught of host immunity<sup>45</sup>. Indeed, the most advanced efforts at inducing immunity against pre-erythrocytic parasite stages have targeted the sporozoite in the skin<sup>26</sup>. Despite this, the skin stage of the sporozoite journey remains poorly understood and the effects of the dermal micro-environment and cell traversal on the transcriptional activity of the sporozoite have yet to be explored. To capture this elusive transcriptome, we allowed salivary gland sporozoites to migrate through transwell  $\mu$ -pores coated with a monolayer of primary skin fibroblasts, before collecting sporozoites that had either traversed or simply contacted the cell monolayer (Fig. 4a). Although our clustering and differential gene expression analysis

revealed that sporozoites adapted their transcriptional activity to cues from the host-like microenvironment, medium-only controls were sufficient to induce a gradual shift in gene expression and neither contact with dermal fibroblasts nor cell traversal was found to have an additional impact (Fig. 4b, c). The observed changes (Supplementary Data 1) were consistent, although not as wide-ranging, with the documented shifts in gene expression experienced by sporozoites exposed to host-like conditions for prolonged periods of time<sup>46</sup> (Supplementary Fig. 19; Supplementary Data 2). This suggests that host-induced changes in transcription may occur gradually in a sequential fashion. After 1 h, salivary gland sporozoites had downregulated the expression of several genes known to be involved in cell traversal (GEST, SPECT1, PLP1, TLP)<sup>44,46–48</sup>. Concomitantly, a cohort of genes associated with protein folding<sup>49</sup> and export<sup>36,50</sup>, including HSP70/90, PTEX150, and PV1, was upregulated (Fig. 4d, e; Supplementary Data 1). PTEX150, which has not been previously



**Fig. 4 Sporozoite activation in the mammalian microenvironment.** **a** Schematics of the transwell sporozoite activation assay. The skin environment was reproduced by transferring sporozoites to transwell inserts filled with fibroblast growth medium and coated with human primary fibroblasts or kept cell-free at 37°C. Sporozoites were allowed to interact with this setup for 15 or 60 min, after which they were collected from both the top and bottom compartments of the transwell. To reach the bottom compartment in the presence of the fibroblast monolayer sporozoites had to deploy their cell traversal activity, as the transwell pores were blocked by the tight cell monolayer. “Medium” refers to sporozoites exposed to medium alone, “contact” refers to sporozoites that were collected above the cell monolayer, and “traversal” refers to sporozoites that passed through the cell monolayer and were collected from the bottom compartment. Salivary gland sporozoites kept in insect medium and injected sporozoites were considered as time 0. **b** UMAP representations of sporozoite transcriptomes colored by cluster assignment before and after activation. **c** Quantification of sporozoite activation based on the cluster assignments in **(b)**. **d** Average expression of down- and upregulated genes during sporozoite activation. Each panel comprises all DE genes that present a fold change of at least 1.5 between C0 and C1 and are expressed in >25% of the cells (Supplementary Data 1). The top GO terms in each module are indicated (GO:0009986, GO:0040011, GO:0043230) and the corresponding Bonferroni-adjusted *p* values are summarized in Supplementary Data 4. **e** UMAP plots of single-cell sporozoite transcriptomes with the expression of PTEX150 before (0 min) and after activation (15 and 60 min) shown. Expression levels ( $\log_2$  counts) are color-coded as per the scale. The cell cluster assignments from **(b)** are indicated. **f** Expression of PTEX150 protein in *P. falciparum* sporozoites before and after activation. Sporozoites were fixed at the indicated time points and immunostained with anti-CSP (green), anti-PTEX150 (red), and Hoechst (blue). Scale bar, 5  $\mu\text{m}$ . **g** Quantification of PTEX150 protein levels. The boxplot shows normalized mean fluorescence intensities, with the notch representing the 95% confidence interval of the median and whiskers denoting the data range within 1.5 $\times$  of the interquartile range. A total of 140, 114, and 47 individual sporozoites were analyzed at 0, 60, and 120 min, respectively. For 0 and 60 min,  $n = 3$  independent experiments; for 120 min,  $n = 1$  with two replicates. Statistical significance was assessed using one-way ANOVA with Tukey multiple corrections. \*\*\*\*0 min vs 60 min,  $p = 3.27 \times 10^{-28}$ ; 0 min vs 120 min,  $p = 6.45 \times 10^{-11}$ .

implicated in sporozoite biology, is a core component of the *Plasmodium* translocon of exported proteins (PTEX) complex<sup>50</sup>, responsible for protein trafficking into the host cell following merozoite invasion. Upregulation of PTEX150, which we confirmed by immunofluorescence analysis (Fig. 4f, g), occurred within a timeframe when sporozoites are still en route to the liver, suggesting that protein export to the host cell might be one of the earliest events following the productive invasion of a hepatocyte. This would mirror merozoite invasion, where the PTEX complex is inserted into the newly formed parasitophorous vacuole membrane within minutes<sup>51</sup>, enabling the export of effector proteins into the erythrocyte to build a suitable niche for replication and immune evasion<sup>52</sup>. It is worth noting that this activated phenotype was also observed in a small proportion of salivary gland sporozoites before their ejection from the mosquito (Fig. 4b, c, e). Whether this is the result of spurious signaling events that disrupt sporozoite latency, or instead represents a developmentally regulated process that renders sporozoites transmission-ready awaits to be established.

## Discussion

Here, we present a comprehensive single-cell analysis of the transcriptional journey of the human malaria parasite *P. falciparum* through its transmission cycle. The transcriptomes presented here provide an in-depth characterization of gene expression across distinct phenotypic stages of development and add a chapter to the Malaria Cell Atlas, which is an interactive data resource supporting the exploration of individual *Plasmodium* parasite transcriptomes now across the life cycle from two malaria parasite species. This data set captures key aspects of parasite development within the mosquito and beyond, providing insights into how those developmental processes may be regulated, as well as elucidating potential targets for disrupting the transmission cycle to control malaria. Exploration of correlated gene expression trajectories within this data set can be the starting point to discover the function of genes that lack functional annotation or reveal unexpected potential functions for other more well-studied genes. For example, our finding of a possible role for MSP1, a vaccine candidate implicated in merozoite egress, in the release of sporozoites from the oocyst to the mosquito hemocoel, suggests that a conserved egress machinery may be deployed at multiple points in the life cycle. Another example of note is that of PTEX150, whose pattern of expression following sporozoite activation suggests an early role in liver stage

biology. Recent observations that EXP2, the pore-forming component of the PTEX complex, is necessary for sporozoite invasion of hepatocytes<sup>53</sup> together with the fact that PTEX150 acts as an adaptor between EXP2 and a third translocon component, HSP101<sup>54</sup>, raise the possibility that PTEX150 might play a role in or shortly after hepatocyte invasion as part of the PTEX translocon complex. Strategies that target this complex might thus be effective at disrupting asexual replication as well as blocking the onset of liver stage infection. More generally, the global view of gene usage across the life cycle made possible by the Malaria Cell Atlas sheds light on core functional strategies that will help inform the development of previously unappreciated multistage antiplasmodial interventions. Such multistage functional classes represent attractive targets for drug development or other interventions aimed at both the host and the mosquito vector.

Transmission of parasites between humans and mosquitoes represents the major population bottlenecks in the malaria life cycle<sup>1</sup>. After an infectious blood meal, developmental transitions within the mosquito midgut are accompanied by a steep reduction in parasite numbers (down to the tens of units). At the other end of the transmission cycle, not only is the number of sporozoites injected into the skin during a bite generally low, but an even smaller number find their way out of the site of injection and reach the liver to initiate a new phase of population expansion within a hepatocyte<sup>45</sup>. The factors that determine whether an individual parasite will survive this journey are unclear, but one intriguing possibility is that the ability to interact with and respond to the host environment in a productive manner could vary from parasite to parasite. Of key importance, we identify developmentally regulated transcripts that show highly variable expression across cells of the same stage and could theoretically account for differential phenotypes within a population. Amongst the genes with variable expression in salivary gland sporozoites is PV1, a translocon accessory protein<sup>36</sup> that is also a hallmark of activated sporozoites. It is thus tempting to speculate that sporozoites that express PV1 prior to the transmission might have a competitive advantage compared with the wider majority that does not activate PV1 expression until transmission to the host has occurred. Transcriptional heterogeneity within sporozoite populations has also been observed in a study of *P. berghei* mosquito stages published while this work was under consideration<sup>55</sup>. Although it is not clear whether population heterogeneity is driven by the same highly variable transcripts in the two species, it appears that transcriptional heterogeneity is a

conserved feature of sporozoite development. Phenotypic validation may reveal whether there is a transcriptional signature that correlates with infectiousness and predicts the likelihood of successful transmission. Further interrogation of this data set will serve as a major accelerator for the discovery of future targets to block human malaria transmission and ultimately prevent the establishment of malaria disease.

## Methods

### Parasite maintenance and isolation

*P. falciparum* maintenance and gametocyte induction. *P. falciparum* gametocytes were induced and cultured as before<sup>56</sup>. In brief, asexual blood-stage cultures were maintained in asexual culture medium (Roswell Park Memorial Institute (RPMI) 1640 with 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) (Life Technologies), 50  $\mu\text{g L}^{-1}$  hypoxanthine (Sigma), 0.3  $\text{g L}^{-1}$  L-glutamine (Sigma), and 10% A + human serum (Interstate Blood-Bank)) with a hematocrit (HC) of 5%. Gametocyte cultures were induced at 2.5% parasitemia and maintained for 15–17 days in gametocyte culture medium (RPMI 1640 with 25 mM HEPES (Life Technologies), 50  $\mu\text{g L}^{-1}$  hypoxanthine (Sigma), 2  $\text{g L}^{-1}$  sodium bicarbonate (Sigma), 0.3  $\text{g L}^{-1}$  L-glutamine, 5% A + human serum (Interstate Blood-Bank) and 0.5% AlbuMAX II (Life Technologies), 5% HC) with daily media changes.

*Standard membrane feeding assay.* *Anopheles stephensi* mosquitoes were reared under standard conditions (26–28 °C, 65–80% relative humidity, 12:12 h light/darkness photoperiod). Adults were maintained on 10% fructose, except for the 12 h before an infectious blood meal, when mosquitoes were starved. Mosquito infections were carried out using a standard membrane feeding assay (SMFA), as previously described<sup>57</sup>. In brief, 15–17-day-old gametocyte cultures with at least 2% stage V gametocytaemia were diluted in a 50% mixture of fresh RBCs and A + human serum and then used to feed starved female mosquitoes. Mosquitoes were allowed to feed for 12 min, after which they were kept without fructose for 24 h, so that only blood-fed mosquitoes would survive. From then on, mosquitoes were given fresh fructose daily until the day of dissection.

*Retrieval of gametocytes.* At day 16 post induction, gametocytes were loaded on a 70% Percoll gradient and spun at 500  $\times$  g for 10 min without brake at 38 °C. Gametocytes were washed once in gametocyte medium at 38 °C and the pellet resuspended in 1 ml of suspended animation (SA) buffer<sup>58</sup> supplemented with 2  $\mu\text{g/ml}$  Hoechst 33342 and 4  $\mu\text{g/ml}$  Pyronin Y. In the presence of Hoechst 33342, interactions between Pyronin Y and DNA are disrupted and Pyronin Y mainly stains RNA<sup>59</sup>. Gametocytes were stained at 4 °C for 20 min. After 20 min, cells were washed twice in SA and resuspended in fresh SA, passed through a strainer with single gametocytes sorted into a 96-well plate using a BD FACSAriaIII at 4 °C. PyroninY was detected using a 561 nm laser with a 586/16 nm filter, and Hoechst 33342 was detected using the 488 nm laser with the 450/50 nm filter (Supplementary Fig. 2). Gametocytes were collected from parasites deriving from either an NF54 genetic background (African) or a recent, culture-adapted field isolate, APL5G (Cambodian)<sup>60</sup>.

*Retrieval of ookinetes.* In all, 24 h after SMFA, mosquitoes were dissected, and the blood bolus removed from the midguts. The collected blood were placed into ice-cold PBS supplemented with 2 $\times$  protease inhibitors (Halt™ Protease Inhibitor Cocktail, Thermo Fisher) and 50 mM ethylenediaminetetraacetic acid. For the first pilot flow cytometry experiment (Supplementary Fig. 1A), uninfected RBCs and activated gametocytes cultures were placed in ookinete medium (RPMI-HEPES supplemented with 100  $\mu\text{M}$  xanthurenic acid, 200  $\mu\text{M}$  hypoxanthine, and 10% bovine serum albumin, pH7.4) for antibody- and DNA staining and used as controls. Each sample was split into four aliquots, one aliquot remained unstained and the others were incubated with either 1:500 anti-Pfs25-cy3 antibody<sup>56</sup>, 1:1000 DRAQ5 DNA stain (BioLegend), or both, respectively, at 4 °C in the dark for 30 min. Samples were then washed twice in ice-cold PBS and resuspended in ice-cold PBS, before being analyzed on a BD Fortessa using the 561 nm laser and filter 586/15 for cy3 and the 640 nm laser and 730/45 filter for DRAQ5, respectively. For the sorting, ookinetes collected 24 and 48 h after a feed were prepared as above and sorted into a 384-well plate. One-fifth of the plate was sorted from the P1 population and the rest from the P2 population (Supplementary Fig. 1B, C). Oocysts per midgut were counted on day 10 post-feed from 41 mosquitoes and revealed a prevalence of infection of 63%.

*Retrieval of sporozoites.* Sporozoites were collected on different days from three anatomic sites in the mosquito to capture different stages of development. sgSpz were collected on three different days, representing three independent infections. injSpz originated from two independent infections, whereas ooSpz and hlSpz derived from a single experiment. Before dissections, mosquitoes were cold anesthetized (4 °C) and killed inside a secure glove box by beheading. Oocyst and salivary gland sporozoites were obtained by dissecting mosquito midguts and salivary glands, 12 and 17–20 days post blood meal, respectively. Hemolymph sporozoites were released from mosquito carcasses on day 14 after removal of both

salivary glands and midguts. All samples were initially homogenized with a disposable pestle to release sporozoites from the mosquito tissues. After a filtration step with a 20  $\mu\text{m}$  filter (Pluriselect), sporozoites were loaded onto a 3 ml Accudenz cushion and centrifuged at 2500  $\times$  g for 20 min<sup>61</sup>. All the above procedures were carried out at 4 °C in Schneider's insect medium. After centrifugation, sporozoites were collected from a thin layer on top of the Accudenz cushion (mosquito debris separated to the bottom), washed once with cold Schneider's buffer (12,000  $\times$  g, 3 min, 4 °C), and resuspended in the same buffer supplemented with 1% fetal bovine serum (FBS). Before FACS sorting, sporozoites were immunostained for 30 min with anti-CSP (2A10 clone, MR4) diluted at 1:200, followed by a 20-minute incubation with an Alexa Fluor 488-conjugated anti-mouse antibody (1:300, Jackson ImmunoResearch) and DRAQ5 (1:1000, BioLegend). All incubations and washes were carried out at 4 °C. Single cells were sorted directly into NEB lysis buffer on a BD FACSAria III equipped with a 100- $\mu\text{m}$  nozzle.

Sporozoites released through a mosquito bite were collected by allowing 300–500 mosquitoes to feed on a warm fructose solution (80 g/L) supplemented with 10% serum for 5 min. After confirming that sporozoites had been released and that no mosquito debris were present in the sample, sporozoites were stained for 20 min with Draq5 at 4 °C and sorted directly into NEB lysis buffer.

*Cell traversal assay.* Sporozoites released from salivary glands on day 20 after a blood meal were purified as before and resuspended in fibroblast growth medium (PromoCell) supplemented with heat-inactivated bovine serum (Sigma) to a final concentration of 10%. In all, 50,000 sporozoites were added per well of a 24-well plate fitted with Transwell inserts (5  $\mu\text{m}$  pore size, Corning) coated with 100,000 human primary dermal fibroblasts (PromoCell) to form a tight monolayer. Sporozoites were allowed to migrate for 15 or 60 min and were then recovered separately from the top and bottom Transwell compartments. On average, 35% of the input was recovered from the bottom Transwell compartment. Some control wells did not contain fibroblasts, so that the effect of the medium on the sporozoites could be evaluated. Before FACS sorting, sporozoites were immunostained as described above, with all steps carried out in fibroblast growth medium with 10% FBS at 4 °C. Single cells were sorted as before.

*Immunofluorescence analysis.* Sporozoites were isolated as described above and resuspended in Dulbecco's modified Eagle's medium with or without supplementation with heat-inactivated bovine serum (Sigma), at either RT or 37 °C. At the indicated time points, 50,000 sporozoites were spun down (3000 rpm, 5 min) on a glass coverslip coated with poly-L-lysine (Sigma) and allowed to settle for 10 min before fixation with 4% paraformaldehyde (PFA, Pierce) for 20 min. Sporozoites were then permeabilized for 5 min with ice-cold methanol (Sigma) and blocked for 1 h with a solution of 5% normal donkey serum (Jackson ImmunoResearch) and 0.01% Triton X-100 (Sigma) in PBS. Primary antibodies were diluted in blocking solution and incubated overnight at 4 °C. Anti-PTEX150 was a kind gift from Paul Gilson (1:500, Burnet Institute, Australia) and anti-CSP (1:200, clone SAI131-5D5) was obtained from MR4. After three washes with blocking solution, samples were incubated with AlexaFluor-conjugated secondary antibodies (1:500, Jackson ImmunoResearch) and Hoechst 33342 (Invitrogen) for 1 h at room temperature (RT) and washed again. The coverslips were then mounted on microscope slides using Fluoromount-G (Thermo Fisher) and imaged with a Leica SP8 inverted confocal microscope equipped with a  $\times 63$  objective. For quantification of PTEX150 expression in sporozoites, images were processed and analyzed using custom macros in Fiji version 2.1.0/1.53c.

For imaging oocysts, the midguts of infected mosquitoes were dissected on day 12 post infection and fixed with 4% PFA for 1 h at RT. Samples were permeabilized with acetone for 5 min, washed, and blocked overnight at 4 °C with 5% normal donkey serum, 0.1% Triton X-100 in PBS. Primary antibodies (1:500 rabbit anti-MSP1-19, a gift from Paul Gilson; 1:200 mouse anti-CSP, MR4) were diluted in blocking solution and incubated overnight at 4 °C. After three 15-minute washes in blocking solution, samples were incubated with AlexaFluor-conjugated secondary antibodies (1:500, Jackson ImmunoResearch) and Hoechst 33342 (Invitrogen) for 1 h at RT. Following two more washes in blocking solution and a final wash in PBS, midguts were mounted on microscope slides using Fluoromount-G. Z-stack images were acquired on a Leica SP8 inverted microscope equipped with a  $\times 63$  objective and processed using Velocity software.

*scRNA-seq.* Cells were sorted in 96- or 384-well plates on FACSAria III into NEB lysis buffer (0.1 $\times$  cell lysis buffer E6428B-SG, 0.05 $\times$  RNase-inhibitor Murine E6429B-SG, and 0.85 $\times$  nuclease-free water E6433B-SG, from New England Biolab). The ookinetes were sorted into a 384-well plate, and all other samples were sorted into 96-well plates. The 384-well plate contained 2  $\mu\text{L}$  lysis buffer, whereas the 96-well plates contained 4  $\mu\text{L}$ . Reverse transcription and PCR were performed in the Sanger Institute single-cell sequencing pipeline using the NEBNext Single Cell/Low Input RNA Library Prep Kit for Illumina (E6420L) with 26 PCR cycles. cDNA was purified using an AMPure XP bead clean-up on either an Agilent Bravo or Hamilton Star. Reaction volumes were doubled in line with the starting volume of the lysis buffer in 96-well plates for cDNA generation, amplification, and purification. The samples were quantified using the Mosquito and Omega plate reader to calculate the volume of cDNA needed for library preparation. Libraries were made using the scRNA cDNA-

XP plates, bravo platforms, and NEB Ultra II FS reagents (E7805L). Libraries had an additional eight cycles of PCR. Libraries were then pooled to 384 samples prior to AMPure XP bead Clean-up on a Hamilton Star. The samples were run on an Agilent Bioanalyzer and normalized to 32.8 nM prior to sequencing. Samples were sequenced on a HiSeq 2500 with 75 bp paired-end reads.

### scRNA-seq mapping and analysis

**Single-cell transcriptome mapping and read counting.** Data acquisition and mapping followed the pipeline described in ref. <sup>6</sup>. In brief, reads were trimmed using trimalore (cutadapt v 1.18)<sup>62</sup>. Trimmed reads were mapped to the *P. falciparum* v3 genome (<https://www.sanger.ac.uk/resources/downloads/protozoa/> October 2016) with HISAT2 (v 2.1.0)<sup>63</sup> using `hisat2 --max-intronlen 5000 p 12`. Mapped reads were then summed against transcripts with featureCounts in the Subread package (v 2.0.0)<sup>64</sup> using `featureCounts -p -t CDS -g transcript_id`.

**Quality control and normalization.** Given the high variability of genes detected across the life cycle<sup>8</sup>, low-quality cells were identified based on the distribution of the number of reads and genes detected per cell within each parasite stage (gametocytes, ookinetes, sporozoites). For gametocytes, cells with fewer than 500 genes and 10,000 reads per cell were removed. For ookinetes, cells with fewer than 400 genes and 5000 reads per cell were removed. For sporozoites, cells with fewer than 40 genes per cell and 5000 reads per cell were removed (Supplementary Fig. 4, Supplementary Table 1). Transcriptomes were normalized using a deconvolution size factor with scran (v 1.16.0)<sup>65</sup> using `quickCluster()` and `calculateSumFactors()` to account for the large differences in detection between parasite stages. Visual inspection of the cell-wise relative log expression plots (Supplementary Fig. 4) shows normalization by scran smoothed profiles across stages.

**Cell clustering and projection.** To delineate parasite transcriptomes into cell types, we performed graph-based Louvain clustering in Seurat (v 3.1.5)<sup>13</sup> with different resolution depending on the analysis (overview of all cells and integration: resolution = 0.5; sporozoite and ookinete development: resolution = 0.6; sporozoite activation: resolution = 0.4). Transcriptomes were visualized using UMAP nonlinear dimensionality reduction in either Scater (v1.16.0)<sup>66</sup> with `n_neighbors=5`, `min_dist=1`, `spread=3` (global analysis, Fig. 1) or Seurat with `n_neighbors=5`, `min_dist=2`, and `spread=3` (integration, Fig. 2) or default parameters (all other analyses).

**Gene graph and clustering.** To create the kNN gene graph, cells from cluster 0 and 1 (Fig. 1c, sporozoite clusters) were downsampled to 100 cells each in order to get a more even representation across cell types. Feature selection was performed on the expression matrix of these 623 transcripts to identify the top 2000 biologically relevant genes using `selectFeature()` in `scmap` (v 1.10.0)<sup>25</sup>. Transcripts were further filtered by removing those with fewer than 200 total reads across the 623 cells, resulting in 1797 genes that were used to generate the gene graph. The counts matrix was normalized by gene by dividing the mean counts for each gene and log scaling in order to reduce the amount in which clusters were driven by total expression level. The kNN graph was made using this gene normalized expression matrix with the nearest neighbors subpackage in python's `scikit-learn` (v 0.2.22.post1) with `n_neighbors=6` and a manhattan distance metric<sup>67</sup>. Spectral clustering was then performed on the kNN graph using the `SpectralClustering` subpackage of `scikit-learn` (v 0.2.22.post1)<sup>68,69</sup>. The kNN graph was then rendered in Gephi with force atlas 2 layout algorithm in `linlog` mode with scale of 0.2 and gravity of 0.8<sup>70</sup>. To test for functional class enrichment in each gene cluster, a Gene Ontology (GO) analysis was conducted using the GO analysis tool on PlasmoDB. Pearson's correlations between ApiAP2 TFs expressed across the data set and each gene cluster were computed using the `cor` function in R.

**Integration with *P. berghei* data.** To identify conserved and divergent patterns of expression in *Plasmodium* transmission stages, we compared our *P. falciparum* data to *P. berghei* data from<sup>8,71</sup>. To do this, we first subsetted similar stages across the two data sets. (*P. falciparum*: gametocytes, ookinetes, and all sporozoite collections; *P. berghei*: gametocytes, ookinetes, and gland and injected sporozoites). We then built expression matrices for each data set based on one-to-one orthologs as described in<sup>8</sup>. Using these orthologous expression matrices, we integrated the two data sets using Seurat's integration function by first identifying anchors with the `FindIntegrationAnchors()` and then using `IntegrateData()`<sup>13</sup>. Cells were visualized and clustered based on the integrated expression values as described above. Analysis of conserved markers and DE between species was then performed on each cluster using the `FindConservedMarkers()` and `FindMarkers()` functions in Seurat<sup>13</sup>. This procedure was also performed for asexual stage parasites from<sup>6,8</sup> and compared with *P. falciparum* strain-specific expression patterns from<sup>27</sup>. Strain-specific expression within each matched cluster was identified using `FindMarkers()` in Seurat. Clusters were matched across data sets using the `scmapCell()` function in `scmap` (v 1.10.0)<sup>25</sup> with the integrated smart-seq2 data as the reference index and the multi-strain data as the query data set. Each cell was given a cluster and/or stage assignment based on the top-matched cell from the reference index. `Scmap` was also used to compare *P. berghei* and *P. falciparum* transmission stages using the *P. berghei* orthologous expression matrix as the reference index.

**Pseudotime.** Ookinetes and sporozoites were ordered along two independent developmental trajectories using Slingshot (v 1.4.0)<sup>72</sup>. In brief, Seurat clusters and UMAP cell embeddings were used as input to build a minimum spanning tree to infer the order of the clusters. Slingshot then fits principal curves to estimate pseudotime values for each cell. In both cases, Slingshot identified a single trajectory that described the progression of the cells along developmental pseudotime. To identify genes differentially expressed as a function of pseudotime, we used the `differentialGeneTest` function in Monocle 2 (v 2.14.0)<sup>73</sup>. DE genes were clustered hierarchically using the `heatmap` package with a Pearson correlation distance measure (`clustering_distance_rows = "correlation"`). Enrichment of GO terms in each pseudotime cluster was tested using the GO analysis tool on PlasmoDB. To assess whether developmentally regulated genes showed variable expression between individual cells of the same developmental stage, we identified HVGs using M3Drop (v 1.12.0), which exploits the observation that dropout-rates are correlated with expression level and identifies outliers of this null expectation as highly variable and of potential biological importance<sup>74</sup>. For sporozoites, ooSpz, hlSpz, sgSpz, and injSpz cells were analyzed separately. For the analysis of HVGs in ookinete cells, we first used a general linear model to regress out the effect of pseudotime, as previously described<sup>8</sup>. HVGs with a *q* value < 0.05 were then intersected with the pseudotime-regulated gene subset.

**Correlation analysis.** Pearson and Spearman correlations were computed using the `cor` or `correlatePairs` functions in R. Statistical significance was assessed using the `cor.test` function.

**Motif discovery.** To find short DNA motifs enriched in the gene clusters from Fig. 1d, the 1000 bp upstream of the start codon of each gene in each cluster were analyzed with DREME<sup>75</sup> (MEME suite 5.3.3). The negative data set included the equivalent 1000 bp region of all genes outside the cluster being analyzed. Motifs were considered enriched if the *e* value was lower than 0.05. Tomtom<sup>76</sup> (MEME suite 5.3.3) was used to compare the top-scoring motif of each cluster with transcription *cis*-regulatory elements described in *P. falciparum*<sup>77–80</sup>.

**Differential abundance testing.** To test for batch effects, we performed differential abundance analysis on single-cell salivary gland sporozoite transcriptomes derived from three independent experimental batches using the `miR` package (v 0.1.0)<sup>81</sup> to define cell neighborhoods on a kNN graph (*k* = 20, *d* = 10) and test for differential abundance in each neighborhood using a negative binomial general linear model framework.

**Comparison with sporozoite bulk transcriptomes.** Sporozoite single cells were assigned one of two stages (ooSpz or sgSpz) based on gene expression correlations between scRNA-seq (this study) and bulk<sup>30</sup> transcriptomes. Pearson and Spearman correlations were computed using the `corr` package in R. To determine whether a transcript was likely to be translationally repressed (TR), we intersected scRNA-seq pseudotime gene expression profiles with the published proteomes of ooSpz and sgSpz<sup>30</sup>. Sporozoite transcripts that were expressed in our data set but absent in the proteome of the same stage were labeled as TR (Supplementary Data 1). Alternatively, sporozoite transcriptomes were assigned to one of two transcriptional activation states (salivary gland or activated) by computing gene expression correlations between this and *P. vivax* bulk RNA-seq data<sup>46</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

Raw sequence data are available from the European Nucleotide Archive (accession number ERP124136). Expression matrices and supporting files are available on Zenodo at <https://doi.org/10.5281/zenodo.4719664> and data are searchable via the MCA website: [www.malaria-cellatlas.org](http://www.malaria-cellatlas.org). *P. berghei* Malaria Cell Atlas data are available at <https://doi.org/10.5281/zenodo.2843883>.

### Code availability

The code used for data analysis is available at <https://doi.org/10.5281/zenodo.4719664>.

Received: 18 March 2021; Accepted: 28 April 2021;

Published online: 27 May 2021

### References

1. Graumans, W., Jacobs, E., Bousema, T. & Sinnis, P. When is a plasmodium-infected mosquito an infectious mosquito? *Trends Parasitol.* **36**, 705–716 (2020).

2. Smith, R. C., Vega-Rodríguez, J. & Jacobs-Lorena, M. The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. *Mem. Inst. Oswaldo Cruz* **109**, 644–661 (2014).
3. Baton, L. A. & Ranford-Cartwright, L. C. Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito. *Trends Parasitol.* **21**, 573–580 (2005).
4. Prudêncio, M., Rodriguez, A. & Mota, M. M. The silent path to thousands of merozoites: the Plasmodium liver stage. *Nat. Rev. Microbiol.* **4**, 849–856 (2006).
5. Poran, A. et al. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **551**, 95–99 (2017).
6. Reid, A. J. et al. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, e33105 (2018).
7. Brancucci, N. M. B. et al. Probing Plasmodium falciparum sexual commitment at the single-cell level. *Wellcome Open Res.* **3**, 70 (2018).
8. Howick, V. M. et al. The Malaria Cell Atlas: single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **365**, eaaw2619 (2019).
9. Sa, J. M., Cannon, M. V., Caleon, R. L., Wellems, T. E. & Serre, D. Single-cell transcription analysis of Plasmodium vivax blood-stage parasites identifies stage- and species-specific profiles of expression. *PLoS Biol.* **18**, e3000711 (2020).
10. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends Genet.* **28**, 560–565 (2012).
11. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv [stat.ML]* <https://joss.theoj.org/papers/10.21105/joss.00861> (2018).
12. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
13. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019). e21.
14. Siciliano, G. et al. Critical steps of Plasmodium falciparum ookinete maturation. *Front. Microbiol.* **11**, 269 (2020).
15. Bushnell, E. et al. Functional profiling of a Plasmodium genome reveals an abundance of essential genes. *Cell* **170**, 260–272 (2017). e8.
16. Zhang, M. et al. Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis. *Science* **360**, eaap7847 (2018).
17. Ay, F. et al. Multiple dimensions of epigenetic gene regulation in the malaria parasite Plasmodium falciparum: gene regulation via histone modifications, nucleosome positioning and nuclear architecture in P. falciparum. *Bioessays* **37**, 182–194 (2015).
18. Josling, G. A. & Llinás, M. Sexual development in Plasmodium parasites: knowing when it's time to commit. *Nat. Rev. Microbiol.* **13**, 573–587 (2015).
19. Witmer, K., Fraschka, S. A., Vlachou, D., Bártfai, R. & Christophides, G. K. An epigenetic map of malaria parasite development from host to vector. *Sci. Rep.* **10**, 6354 (2020).
20. Josling, G. A. et al. A Plasmodium Falciparum bromodomain protein regulates invasion gene expression. *Cell Host Microbe* **17**, 741–751 (2015).
21. Sierra-Miranda, M. et al. PfAP2Tel, harbouring a non-canonical DNA-binding AP2 domain, binds to Plasmodium falciparum telomeres. *Cell Microbiol.* **19** (2017).
22. Hillier, C. et al. Landscape of the Plasmodium interactome reveals both conserved and species-specific functionality. *Cell Rep.* **28**, 1635–1647 (2019).
23. Guizetti, J. & Scherf, A. Silence, activate, poise and switch! Mechanisms of antigenic variation in Plasmodium falciparum. *Cell. Microbiol.* **15**, 718–726 (2013).
24. MalariaGEN et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. *Wellcome Open Res.* **6**, 42 (2021).
25. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
26. Duffy, P. E. & Patrick Gorres, J. Malaria vaccines since 2000: progress, priorities, products. *NPJ Vaccines* **5**, 48 (2020).
27. Heaton, H. et al. Souporecell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
28. Baumgarten, S. et al. Transcriptome-wide dynamics of extensive m6A mRNA methylation during Plasmodium falciparum blood-stage development. *Nat. Microbiol.* **4**, 2246–2259 (2019).
29. Smith, R. C. & Jacobs-Lorena, M. Plasmodium–Mosquito Interactions: A Tale of Roadblocks and Detours. In *Advances in Insect Physiology* (ed. Simpson, S. J.) vol. 39, Chapter 4, 119–149 (Academic Press, 2010).
30. Lindner, S. E. et al. Transcriptomics and proteomics reveal two waves of translational repression during the maturation of malaria parasite sporozoites. *Nat. Commun.* **10**, 4964 (2019).
31. Thompson, J. et al. Plasmodium cysteine repeat modular proteins 1–4: complex proteins with roles throughout the malaria parasite life cycle. *Cell. Microbiol.* **9**, 1466–1480 (2007).
32. Douradinha, B. et al. Plasmodium cysteine repeat modular proteins 3 and 4 are essential for malaria parasite transmission from the mosquito to the host. *Malar. J.* **10**, 71 (2011).
33. Gilson, P. R. et al. Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite Plasmodium falciparum. *Mol. Cell. Proteom.* **5**, 1286–1299 (2006).
34. Taechalerpaisarn, T. et al. Biochemical and functional analysis of two Plasmodium falciparum blood-stage 6-cys proteins: P12 and P41. *PLoS ONE* **7**, e41937 (2012).
35. Das, S. et al. Processing of Plasmodium falciparum merozoite surface protein msp1 activates a spectrin-binding function enabling parasite egress from RBCs. *Cell Host Microbe* **18**, 433–444 (2015).
36. Morita, M. et al. PV1, a novel Plasmodium falciparum merozoite dense granule protein, interacts with exported protein in infected erythrocytes. *Sci. Rep.* **8**, 3696 (2018).
37. Hoeijmakers, W. A. M. et al. Epigenetic reader complexes of the human malaria parasite, Plasmodium falciparum. *Nucleic Acids Res.* **47**, 11574–11588 (2019).
38. Jeninga, M. D., Quinn, J. E. & Petter, M. ApiAP2 transcription factors in apicomplexan parasites. *Pathogens* **8**, 47 (2019).
39. Modrzynska, K. et al. A knockout screen of ApiAP2 genes reveals networks of interacting transcriptional regulators controlling the Plasmodium life cycle. *Cell Host Microbe* **21**, 11–22 (2017).
40. Zhang, C. et al. Systematic CRISPR-Cas9-mediated modifications of Plasmodium yoelii ApiAP2 genes reveal functional insights into parasite development. *MBio* **8**, e01986-17 (2017).
41. Toenhake, C. G. et al. Chromatin accessibility-based characterization of the gene regulatory network underlying Plasmodium falciparum blood-stage development. *Cell Host Microbe* **23**, 557–569 (2018).
42. Yamauchi, L. M., Coppi, A., Snounou, G. & Sinnis, P. Plasmodium sporozoites trickle out of the injection site. *Cell. Microbiol.* **9**, 1215–1222 (2007).
43. Hopp, C. S. et al. Longitudinal analysis of Plasmodium sporozoite motility in the dermis reveals component of blood vessel recognition. *Elife* **4**, e07789 (2015).
44. Amino, R. et al. Host cell traversal is important for progression of the malaria parasite through the dermis to the liver. *Cell Host Microbe* **3**, 88–96 (2008).
45. Sinnis, P. & Zavala, F. The skin: where malaria infection and the host immune response begin. *Semin. Immunopathol.* **34**, 787–792 (2012).
46. Roth, A. et al. Unraveling the Plasmodium vivax sporozoite transcriptional journey from mosquito vector to human host. *Sci. Rep.* **8**, 12183 (2018).
47. Talman, A. M. & Lacroix, C. PbGEST mediates malaria transmission to both mosquito and vertebrate hostmmi\_7823 462.474. *Mol. Microbiol.* **82**, 462–474 (2011).
48. Moreira, C. K. et al. The Plasmodium TRAP/MIC2 family member, TRAP-Like Protein (TLP), is involved in tissue traversal by sporozoites. *Cell. Microbiol.* **10**, 1505–1516 (2008).
49. Daniyan, M. O., Przyborski, J. M. & Shonhai, A. Partners in mischief: functional networks of heat shock proteins of Plasmodium falciparum and their influence on parasite virulence. *Biomolecules* **9**, 295 (2019).
50. de Koning-Ward, T. F. et al. A newly discovered protein export machine in malaria parasites. *Nature* **459**, 945–949 (2009).
51. Riglar, D. T. et al. Spatial association with PTEX complexes defines regions for effector export into Plasmodium falciparum-infected erythrocytes. *Nat. Commun.* **4**, 1415 (2013).
52. Matthews, K. M., Pitman, E. L. & de Koning-Ward, T. F. Illuminating how malaria parasites export proteins into host erythrocytes. *Cell. Microbiol.* **21**, e13009 (2019).
53. Mello-Vieira, J., Enguita, F. J., de Koning-Ward, T. F., Zuzarte-Luís, V. & Mota, M. M. Plasmodium translocon component EXP2 facilitates hepatocyte invasion. *Nat. Commun.* **11**, 5654 (2020).
54. Ho, C.-M. et al. Malaria parasite translocon structure and mechanism of effector export. *Nature* **561**, 70–75 (2018).
55. Sinnis, P., Vega-Rodríguez, J. & Serre, D. Transcriptional heterogeneity and tightly regulated changes in gene expression during Plasmodium berghei sporozoite development. *Proc. Natl. Acad. Sci. USA* **118**, e2023438118 (2021).
56. Delves, M. J. et al. Routine in vitro culture of P. falciparum gametocytes to evaluate novel transmission-blocking interventions. *Nat. Protoc.* **11**, 1668–1680 (2016).
57. Ponnudurai, T. et al. Infectivity of cultured Plasmodium falciparum gametocytes to mosquitoes. *Parasitology* **98**, 165–173 (1989).
58. Nijhout, M. M. & Carter, R. Gamete development in malaria parasites: bicarbonate-dependent stimulation by pH in vitro. *Parasitology* **76**, 39–53 (1978).
59. Shapiro, H. M. Flow cytometric estimation of DNA and RNA content in intact cells stained with Hoechst 33342 and pyronin Y. *Cytometry* **2**, 143–150 (1981).

60. Witmer, K. et al. Transmission of Artemisinin-Resistant Malaria Parasites to Mosquitoes under Antimalarial Drug Pressure. *Antimicrob. Agents Chemother.* **65**, (2020)
61. Kennedy, M. et al. A rapid and scalable density gradient purification method for *Plasmodium* sporozoites. *Malar. J.* **11**, 421 (2012).
62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
63. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
64. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
65. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
66. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
67. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
68. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
69. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).
70. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. in *Third International AAI Conference on Weblogs and Social Media* (2009).
71. Howick, V. M. & Reid, A. J. *Howick/MalariaCellAtlas: Malaria cell atlas*. <https://doi.org/10.5281/zenodo.2843883>. (2019)
72. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
73. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
74. Andrews, T. S. & Hemberg, M. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/bty1044> (2018).
75. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
76. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
77. De Silva, E. K. et al. Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA* **105**, 8393–8398 (2008).
78. Young, J. A. et al. In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics* **9**, 70 (2008).
79. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6**, e1001165 (2010).
80. Yuda, M., Iwanaga, S., Shigenobu, S., Kato, T. & Kaneko, I. Transcription factor AP2-Sp and its target genes in malarial sporozoites. *Mol. Microbiol.* **75**, 854–863 (2010).
81. Dann, E., Henderson, N. C., Teichmann, S. A. & Morgan, M. D. Milo: differential abundance testing on single-cell data using k-NN graphs. *bioRxiv* <https://doi.org/10.1101/2020.11.23.393769> (2020).

## Acknowledgements

The authors wish to acknowledge the support from the core sequencing pipelines at the Wellcome Sanger Institute and the FACS and imaging facilities at Imperial College London, namely Jessica Rowley, Jane Srivastava, and Radhika Patel for single-cell sorting, and Stephen Rothery for assistance with image analysis. The Facility for Imaging by Light Microscopy (FILM) at Imperial College London is part supported by funding from the Wellcome Trust (grant 104931/Z/14/Z) and BBSRC (grant BB/L015129/1). We thank Mark Tunnicliff for the maintenance and provision of *A. stephensi* mosquitoes and Alisje Churchyard, Sabrina Yahya, and Catherine Ducker for assistance with experiments or figures. J. Baumis supported by the Wellcome Trust (Investigator Award 100993/Z/13/Z) and the Bill & Melinda Gates Foundation (OPP1200274). The Wellcome Sanger Institute is funded by the Wellcome Trust (grant 206194/Z/17/Z), which supports M.K.N.L. and A.J.R. 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). C.A.B. is supported through a studentship from the Wellcome Trust (Grant 220123/Z/20/Z).

## Author contributions

E.R., V.M.H., F.A.D., and K.W. designed and performed all experiments. E.R. and V.M.H. analyzed the data with help from K.W., J.C., C.A.B., A.J.R., and M.S.D. J. Blight designed the schematics and helped with experiments. S.K.D. and V.M.H. built the interactive web tool. J. Baum and M.K.N.L. sourced funding, provided scientific direction, and helped with experimental design and manuscript preparation. E.R. and V.M.H. wrote the manuscript with contributions from other authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-23434-z>.

**Correspondence** and requests for materials should be addressed to J.B. or M.K.N.L.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021