

1 Single-cell transcriptomics of malaria parasites

2 Adam J. Reid^{1,3}, Arthur M. Talman^{1,3}, Hayley M. Bennett^{1,3}, Ana R. Gomes¹, Mandy J. Sanders¹,
3 Christopher J. R. Illingworth², Oliver Billker¹, Matthew Berriman¹, Mara K. N. Lawniczak¹

4

5 ¹Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

6 ²Department of Genetics, University of Cambridge, Cambridge, UK

7 ³These authors contributed equally to this work.

8 Correspondence should be addressed to AJR (ar11@sanger.ac.uk), AMT (at10@sanger.ac.uk) or

9 MKNL (mara@sanger.ac.uk)

10

11 Abstract

12 Single-cell RNA-sequencing is revolutionising our understanding of seemingly homogeneous cell
13 populations, but has not yet been applied to single cell organisms. Here, we established a method to
14 successfully investigate transcriptional variation across individual malaria parasites. We discover an
15 unexpected, discontinuous program of transcription during asexual growth previously masked by
16 bulk analyses, and uncover novel variation among sexual stage parasites in their expression of gene
17 families important in host-parasite interactions.

18

19 Main text

20 Single-cell RNA-sequencing (scRNA-seq) has revealed astounding cell-to-cell heterogeneity,
21 uncovered previously unknown cell types, and enhanced our understanding of developmental
22 pathways in mammals ¹. However, unicellular organisms have not yet been successfully explored
23 using this method. Variation in the behaviour and developmental decision making of individual
24 unicellular organisms cannot be resolved when averaging across populations of individuals. In
25 *Plasmodium* parasites, transcriptomic analyses of isogenic strains grown under homogenous
26 conditions revealed extensive transcriptional variation particularly among genes associated with
27 host-parasite interactions ². Furthermore, disease-relevant phenotypes such as commitment to
28 sexual development ^{3,4}, parasite sequestration ⁵ and nutrient acquisition ⁶ are thought to be driven
29 by transcriptional variation between individual parasites.

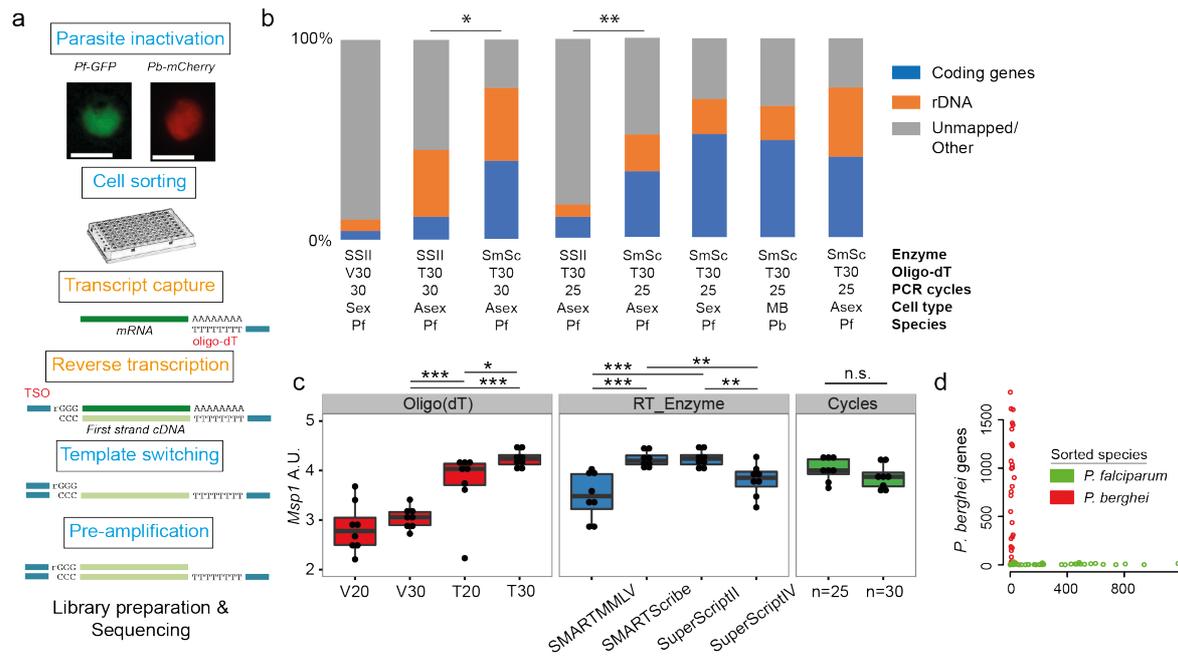
30

31 Compared to mammalian cells, *Plasmodium* parasites present several challenges to existing single
32 cell technologies, including much lower RNA content and a highly AT-biased base composition ⁷.
33 Initially we trialled the well-established Smart-seq2 method ⁸ on sorted, *Plasmodium falciparum*-
34 infected single red blood cells (Fig. 1a), but we increased the number of amplification cycles from 12

35 to 30 to account for their low RNA content. However, on average only 10% of reads mapped to
36 genes in the parasite genome and more than half of these mapped to rRNA genes (Fig. 1b, left most
37 bar). To improve yield, we used qPCR to evaluate the impact of modifying the anchor and length of
38 the oligo(dT) primer, the reverse transcription enzyme, and the number of amplification cycles (Fig.
39 1c). A longer, unanchored polyT primer (T30) significantly improved yield (Fig. 1c), and was used in
40 all further experiments. The reverse transcriptase enzymes SuperScript II and SMARTScribe gave the
41 greatest yields by qPCR (Fig. 1c). Amplification by 25 and 30 cycles appeared equivalent by qPCR (Fig.
42 1c). Therefore we sequenced libraries generated from a small number of individual cells using the
43 best two enzymes and either 25 or 30 cycles of PCR (Fig. 1b). The SMARTScribe enzyme with 25 or 30
44 cycles provided the best results, significantly increasing the number of genes detected and
45 dramatically reducing the rRNA contamination (Fig. 1b; Supplementary Table 1). Given equivalent
46 results for 25 or 30 cycles we opted to use fewer cycles for subsequent experiments. Transcript
47 coverage was not affected by GC content or length; long genes were just as well detected as short
48 genes, although we detected a slight 5' bias in transcript coverage (Supplementary Figure 1). To
49 validate our sorting protocol, a mixed population of *P. falciparum* and *P. berghei* schizonts were
50 individually sorted into wells (n=72 cells). Reads from each parasite matched the expected species
51 based on fluorescence suggesting that there was little cross cell contamination due to lysis (Fig. 1d;
52 Supplementary Figure 2).

53
54 To begin to explore variation within parasite life cycle stages we generated 144 high quality single
55 cell transcriptomes of mixed asexual and sexual (gametocyte) blood stage parasites of the rodent
56 malaria model *P. berghei*. We identified expression from, on average, 1981 genes per cell,
57 representing 33% of the total genes in the genome, on par with mammalian single cell experiments⁹
58 (Supplementary Table 1; Supplementary Figure 3). Using a combination of Principal Components
59 Analysis (PCA), k-means clustering¹⁰ and comparison to bulk transcriptome datasets^{11,12}, we
60 classified each cell as male, female, or asexual (Fig. 2a). The accuracy of our classification was
61 strongly supported by established stage-specific markers (Fig. 2b). In addition to capturing sufficient
62 transcriptional complexity to classify cells, we were also able to identify novel marker transcripts for
63 each parasite stage (Supplementary Data 1).

64
65
66
67
68



69

70

71 **Figure 1. Establishment of a robust protocol for single cell transcriptomic analysis of Plasmodium parasites.**

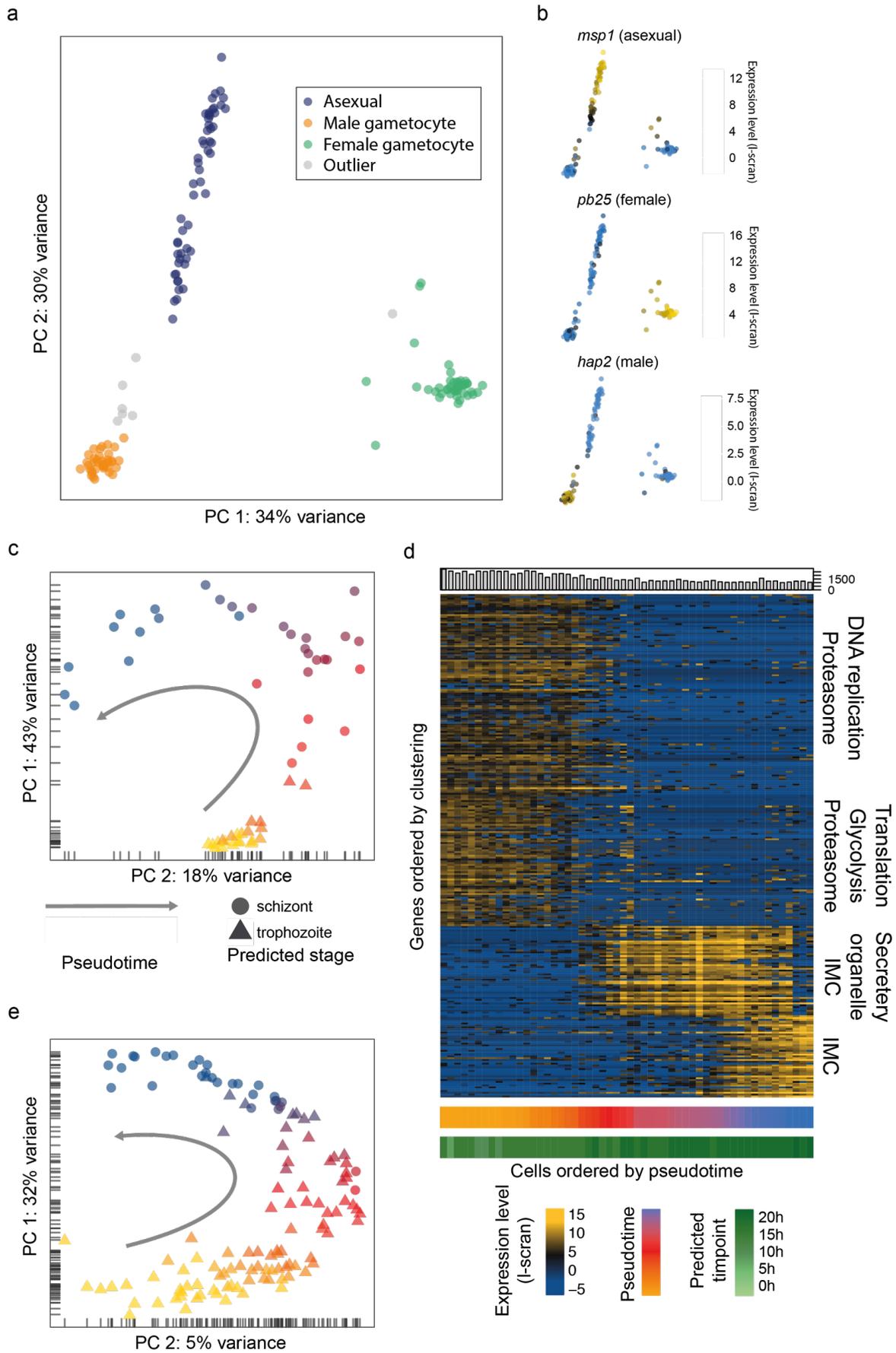
72 (a) Overview of the single cell RNAseq protocol. Steps in the original Smart-seq2 protocol⁸ that resulted in
 73 significant gains are highlighted in orange. (b) Relative numbers of reads mapping to coding RNA and rDNA for
 74 each trial, averaged over all cells in that trial. The final three bars represent the main *P. falciparum*
 75 gametocyte, *P. berghei* mixed blood and *P. falciparum* asexual datasets, respectively. Asterisks indicate
 76 selected significant differences between proportions of reads mapping to coding genes, calculated using Mann-
 77 Whitney U. (c) The protocol was evaluated using qPCR of the *msp-1* transcript (PF3D7_09303000) on sorted
 78 pools of 10 asexual parasites (n=8) (Significance from Mann Whitney test). The following reagents were tested:
 79 Oligo(dT)s containing a terminal anchoring base (A,G,C; V) or not (T) and of varying lengths (20 Ts vs. 30 Ts); 4
 80 reverse transcriptase enzymes; 25 or 30 cycles of preamplification. (d) Individually sorted *P. falciparum* and *P.*
 81 *berghei* cells from a mixed pool revealed no doublets and little contamination (see Supplementary Figure 2).

82

83 Whereas mature sexual parasite cells are terminally differentiated, transcriptional variation along
 84 the 24 hour asexual cycle of *P. berghei* is thought to be continuous^{11,13}, including during the
 85 transition from the growth phase (trophozoite) to the budding phase (schizont).

86 We predicted the maturation state of each individual asexual cell, using a rank-correlation approach
 87 to compare them to two published bulk RNA-seq datasets^{11,12}. We also carried out a pseudotime
 88 analysis¹⁴ using variable genes¹⁵ to order the asexual cell subset. We observed a strong
 89 concordance between the temporal and pseudotemporal predictions (Fig. 2c and 2d). Interestingly,
 90 we observed very abrupt transcriptional changes, distinct from the smooth transitions observed in
 91 bulk time course experiments (Fig. 2d, Supplementary Figure 4). To further study the dynamics of
 92 transcriptional shifts over asexual development, we analysed the transcriptomes of 155 *P.*
 93 *falciparum* late asexual stages. These cells could also be ordered in pseudotime in agreement with
 94 their predicted stage (Fig. 2e; Supplementary Figure 5) and displayed the same dramatic
 95 transcriptional profile shift in the late asexual cycle (Supplementary Figure 5). The smooth

96 transitions observed in bulk data may be attributable to averaging across slightly different asexual
97 life cycle points, which does not happen when examining single cells in pseudotime. Taking
98 advantage of our high resolution transcriptomes, we conducted a co-expression correlation analysis
99 of the ApiAp2 family of transcription factors (TFs) ¹⁶ and observed marked correlated and
100 anticorrelated clusters in a TF network (Supplementary Figure 6). These patterns were strongly
101 associated with peak expression along pseudotime and several correlations were conserved in both
102 species (Supplementary Figure 6). This suggests a possible conserved regulatory framework
103 underlying this rapid and discrete transcriptional shift in late schizogony. More comprehensive
104 sampling of the asexual cycle will be necessary to fully evaluate the pace of transitions and the role
105 of synergistic or antagonistic interactions between TFs in establishing the discrete patterns of gene
106 regulation we observe during the asexual cycle of *Plasmodium* parasites.



107

108

109 **Figure 2. Single cell RNA-seq allows dissection of parasite populations**

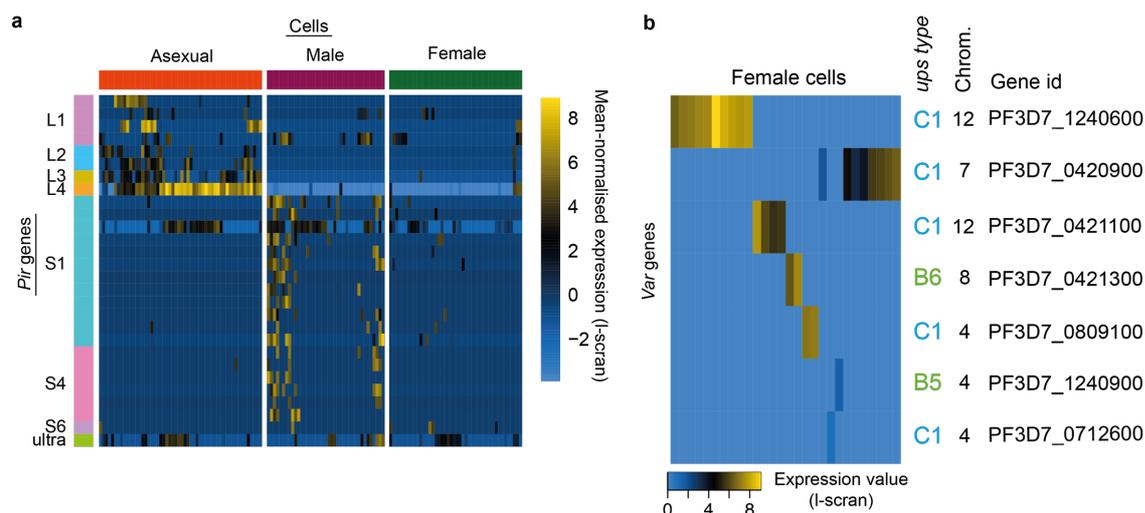
110 **(a)** A combination of Principal Components Analysis (PCA), k-means clustering and comparison to bulk RNA-seq
111 datasets was used to classify 144 high quality *P. berghei* single cells, and revealed three distinct
112 subpopulations. Outliers may represent erythrocytes infected with both sexual and asexual stages or early
113 stages in gametocyte development. **(b)** Three well-established markers of the male, female and asexual
114 lineages²¹⁻²³ are concordant with our classification. **(c)** Pseudotime ordering (using¹⁴) of the asexual cells in (a)
115 was in close agreement with bulk RNA-seq datasets (predicted timepoint from¹¹, predicted stage = consensus;
116 see Methods). **(d)** Differentially expressed genes (identified using¹⁵) were clustered along pseudotime
117 revealing groups of genes with abrupt expression profile changes during late asexual cycle. Functional
118 enrichment in the clusters was in agreement with the expected shift from the growing trophozoite to the
119 budding schizont (IMC = Inner Membrane Complex). **(e)** Pseudotime ordering (using¹⁴) of the 125 *P. falciparum*
120 late asexual cells was in close agreement with bulk RNA-seq datasets (predicted timepoint from²⁴, predicted
121 stage = consensus; see Methods).
122

123 Transcriptional variation within large gene families is known to be important for immune evasion
124 and establishment of chronic infection by *Plasmodium* in the mammalian host^{17,18}. However,
125 relatively little is known about the role of these families, or host-parasite interactions more
126 generally, in the sexual stages which are transmitted to the mosquito. We examined transcriptional
127 variation within the four *P. berghei* life stages and found 115 variable genes in females, 73 in males,
128 27 in trophozoites and 4 in schizonts (Supplementary Data 2). A full functional description of all
129 variable genes is described in Supplementary Data 3. In trophozoites, we observed variation in
130 expression of the multigene family *pir*, in particular the L-type *pir* genes¹⁸. Variation in expression of
131 these genes has recently been associated with establishment of chronic infection¹⁸. In male
132 gametocytes, but not females, we observed variability of S-type *pir* genes¹², a group for which no
133 function has yet been described (Fig. 3a; Supplementary Data 4). Their products however, have been
134 identified in male gametes¹⁹, the cells that arise from male gametocytes once ingested by a
135 mosquito. Our data suggest that sex-specific variation in expression of these genes may play a role in
136 malaria parasite transmission.

137

138

139



140

141 **Figure 3. Multigene families show variable expression in sexual stages of both *P. berghei* and *P. falciparum***
 142 **(a)** *Pir* gene expression was highly variable across male gametocytes. In addition, more *pir* genes were
 143 expressed in males than females. These are distinct subfamilies of *pir* genes from those variably expressed in
 144 asexual stages. **(b)** 28 female *P. falciparum* gametocytes expressed *var* gene transcripts from the sense strand.
 145 We only ever detected a single mRNA *var* in each cell, suggesting the maintenance of mutually exclusive *var*
 146 expression in gametocytes. The promoters of expressed *var* genes were only *upsC* and *upsB* subtypes from
 147 internal rather than subtelomeric regions of the genome.
 148

149 To further explore transcriptional variation in sexual stage biology, we carried out scRNA-seq on *P.*
 150 *falciparum* late sexual stage individuals, resulting in 191 high quality single cell transcriptomes
 151 (Supplementary Table 1, Supplementary Figure 3). Due to the small number of predicted males
 152 detected (n=5; Supplementary Figure 7), we only examined expression variation in females,
 153 detecting 448 variable genes (Supplementary Data 2). The repertoire of multigene families in *P.*
 154 *falciparum* is highly diverged from *P. berghei*²⁰ and clear *pir* gene orthologues do not exist. However,
 155 we found an enrichment for variation in the well-studied *var* gene family (14 of 60 genes,
 156 hypergeometric test, $p = 0.0006$). Expression of noncoding *var* transcripts is common and is involved
 157 in maintaining mutually exclusive expression of a single *var* gene¹⁷. Therefore, to assess the
 158 presence of true *var* mRNAs, we identified reads that confirmed intronic splicing. We discovered that
 159 although there were reads mapping to multiple *vars* within each cell, only a single *var* gene ever had
 160 reads that spanned the spliced intron. This suggests that mutually exclusive expression of *var* genes
 161 occurs in sexual stages, as it does for asexual parasites¹⁷. The *var* genes variably expressed across
 162 females were always from internal *var* gene clusters, consisting primarily of *upsC* type *vars* (Fig. 3b)
 163 but whose role is largely unexplored. In asexual stages, *var* gene products (PfEMP1) allow the
 164 parasites to adhere to the host vasculature and avoid clearance in the spleen and are important in
 165 evading the adaptive immune system¹⁷. Variable and mutually exclusive transcription of *upsC*-rich
 166 internal *var* genes in female gametocytes suggest that this variable gene family, like the *pir* genes
 167 above, may also play an important role in parasite transmission.

168

169 In this study, we have demonstrated a highly successful approach to single cell RNA-sequencing in
170 the study of a unicellular eukaryotic organism. We have used scRNA-seq to successfully resolve
171 individual parasites within a mixed population, to classify them, and to identify previously hidden
172 transcriptional variation. Our data suggest that the transcriptional cascade during development in
173 red blood cells is not continuous as previously described, but instead it occurs in discrete stages. This
174 has important implications for our understanding of gene regulation in the most pathogenic stage of
175 the parasite. We also show that variable expression of genes involved in host-parasite interactions is
176 a feature of multiple life stages of *Plasmodium* parasites. We hope that our optimisation framework
177 will assist in extending scRNA-seq to a much wider range of diverse eukaryotic cell types.

178

179

180

181

182

183

184 **Methods**

185 **Isolation of *P. berghei* parasites**

186 The constitutively mCherry-expressing *P. berghei* ANKA line, clone RMgm-928²⁵, was propagated in
187 a female 6- to 8-week-old Theiler's original outbred mouse supplied by Envigo UK. Parasites were
188 purified from an overnight (20h) 50 mL culture of 1 mL of infected blood using a 55% Histodenz
189 cushion (SIGMA), following an established schizont purification protocol detailed elsewhere²⁶.
190 Purified late stages (asexual and sexual) were pelleted at 450g for 3 minutes and incubated with 500
191 μ L of RNALater (ThermoFisher) for 5 minutes, and further diluted into 3 mL of 1x PBS prior to cell
192 sorting. All animal research was conducted under licenses from the UK Home Office and used
193 protocols approved by the ethics committee of the Wellcome Trust Sanger Institute.

194

195 ***In vitro P. falciparum* culture**

196 3D7-HTGFP, a GFP-expressing *P. falciparum* strain²⁷, was maintained in O-negative red blood cells
197 obtained from the NHSBT, using RPMI 1640 culture medium (GIBCO) supplemented with 25 mM
198 HEPES (SIGMA), 10mM D-Glucose (SIGMA), 50 mg/L hypoxanthine (SIGMA), 10% human serum
199 (obtained locally in accordance with ethically approved protocols), and gassed using a mix containing
200 5% O₂, 5% CO₂ and 90% N₂. Parasites were highly synchronised using two consecutive cycles of
201 Percoll-Sorbitol treatment²⁸. Late asexual parasites (trophozoites and schizonts) were purified on a

202 cushion of 63% Percoll (GE Healthcare). Stage V gametocytes were obtained using standard
203 gametocyte culturing²⁹ and purified magnetically with an LS column³⁰ (Miltenyi Biotec). Following
204 purification of each stage, all *P. falciparum* parasites were pelleted at 800g for 5 minutes, incubated
205 with 500 µL of RNALater (ThermoFisher) for 5 minutes, and further diluted into 3 mL of 1x PBS prior
206 to cell sorting. Parasitaemia was determined by Giemsa-stained thin blood smear.

207

208 **Cell sorting**

209 4 µl of lysis buffer (0.8 % of RNase-free Triton-X (Fisher) in nuclease-free water (Ambion)), UV-
210 treated for 30 min with a Stratalinker UV Crosslinker 2400 at 200, 000 µJ/cm², 2.5 mM dNTPs (Life
211 Technologies), 2.5 µM of oligo(dT) (IDT; see Supplementary Table 1 for detail) and 2U of
212 SuperRNAsin (Life Technologies) were dispensed into each well of the recipient RNase-free 96-well
213 plate (Abgene) immediately prior to the sort and kept on ice. In the first experiment only 2 µl of lysis
214 buffer were used but the observed cell-capture efficiency was very poor so the volume was
215 increased. Cell sorting was conducted on an Influx cell sorter (BD Biosciences) with a 70 µm nozzle.
216 Parasites were sorted by gating on single cell events and on GFP (*P. falciparum*) or mCherry (*P.*
217 *berghei*) fluorescence. A non-sorted negative control well and a positive 100-cell control well were
218 included in every plate alongside single cells. Sorted plates were spun at 200 G for 10 seconds and
219 immediately placed on dry ice.

220

221 **First and second strand cDNA synthesis and pre-amplification**

222 Cells in plates were incubated at 72 °C for 3 minutes. A reverse transcription master mix was added
223 to the samples containing 1 µM of LNA-oligonucleotide (5'-
224 AGCAGTGGTATCAACGCAGAGTACATrGrG+G-3'; Exiqon), 6 µM MgCl₂, 1M Betaine (Affymetrix), 1X
225 reverse transcription buffer, 50 µM DTT, 0.5 U of SuperRNAsin, and 0.5 µl of reverse transcriptase
226 (Supplementary Table 1). The total volume of the reaction was 10µl. The plate was incubated using
227 the following programme 1 X 42°C/90'; 10 X (42°C/2', 50°C/2'), 1 X 70°C/15'. Samples were then
228 supplemented with 1X KAPA Hotstart HiFi Readymix and 2.5 µM of the ISO SMART primer⁸ and
229 incubated using the following cycling programme 1 X 98°C/ 3'; 25 or 30 X (98°C/20'', 67°C/15'',
230 72°C/6'); 1x 72°C/ 5' (Supplementary Table 1). Samples were then purified with 1X Agencourt
231 Ampure beads (Beckman Coulter) in a Zephyr G3 SPE Workstation (Perkin Elmer) according to the
232 manufacturer's recommendation. Amplified cDNA was eluted in 10 µl nuclease-free water. Details of
233 different permutations of the protocol tested during the optimisation process are given in
234 Supplementary Table 1.

235

236 **Quality control of cDNA samples**

237 The quality of a subset of amplified cDNA samples was monitored with the high-sensitivity DNA chip
238 on an Agilent 2100 Bioanalyser. Samples were verified by qPCR using LightCycler 480 SYBR Green I
239 Master and MSP-1 primers at a concentration of 0.4 μ M (Forward: 5'-
240 TCCCAATCAGGAGAAACAGAAG-3'; Reverse: 5'-GATGGTTGTGTTGGTGGTAATG-3'), on a Roche
241 Lightcycler 480 II. Reactions were incubated according to the following cycling programme: 1x
242 95°C/10'; 45X (98°C/20'', 58°C/10'', 68°C/30''). Transcripts were quantified with the absolute
243 quantification method using a standard dilution.

244

245 **Library preparation and sequencing**

246 Libraries were prepared using the Nextera XT kit (Illumina) according to manufacturer
247 recommendations. 96 or 384 different index combinations were used to allow multiplexing during
248 sequencing. After indexing, libraries were pooled for clean-up at a 4:5 ratio of Agencourt Ampure
249 beads (Beckman Coulter). Quality of the libraries was monitored with the high sensitivity DNA chip
250 on an Agilent 2100 Bioanalyser. Empty-well controls and single cells were pooled separately from
251 100 cell controls and loaded proportionally to their expected cell content for sequencing on an
252 Illumina MiSeq or HiSeq 2500.

253

254 **Sequencing of single-cell libraries**

255 The original Smart-seq2 protocol with the Superscript II enzyme and the original oligo(dT) with an
256 anchoring base was run with 30 PCR cycles of preamplification on 10 samples. The samples included
257 a single no-cell control, five single *P. falciparum* gametocytes, two 10-cell controls and two 100-cell
258 controls. These were multiplexed, along with three samples each of individual human lung
259 carcinoma cells (A549) and sequenced on a single MiSeq run with 150bp paired end reads.

260

261 To test the effect of different reverse transcriptase enzymes and different numbers of PCR cycles, we
262 sequenced *P. falciparum* schizont libraries prepared using 1) the SmartScribe enzyme (Clontech) for
263 a) six single cells, one 100-cell control and two no-cell controls with 25 cycles of PCR, and b) six single
264 cells and one 100-cell control with 30 cycles of PCR; and 2) the SuperScript II enzyme (Thermofisher)
265 for a) six single cells, one 100-cell control and two no-cell controls with 25 cycles of PCR, and b) six
266 single cells and one 100-cell control with 30 cycles of PCR. These were multiplexed on a single MiSeq
267 run with 150bp paired end reads.

268

269 To determine whether single-cell samples might be contaminated with either additional cells or RNA
270 from lysed cells, individual mCherry *P. berghei* (RMgm-928²⁵) and GFP *P. falciparum*²⁷ schizonts
271 were mixed in a 1:1 ratio, inactivated with RNALater fixation and then sorted. A multiplex library was
272 prepared comprising 32 single *P. berghei* schizonts, two 100-cell *P. berghei* schizont controls, one no-
273 cell control, 40 single *P. falciparum* schizonts and two 100-cell *P. falciparum* schizont controls. These
274 libraries were sequenced as a multiplex pool on a single MiSeq run with 150bp paired end reads.

275

276 The *P. berghei* mixed blood stage samples comprised 182 single-cells of *P. berghei*, plus four no-cell
277 controls and six 100-cell controls. These were multiplexed with another 192 samples not analysed in
278 this work and sequenced on a single HiSeq 2500 lane using HiSeq v4 with 75bp paired end reads. The
279 *P. falciparum* gametocyte samples were sequenced as three multiplexed pools of 84, using the same
280 chemistry. Three technical duplicate samples were excluded from analysis. The *P. falciparum* asexual
281 samples were sequenced as two pools of 96, each on one Illumina HiSeq 2500 lane using HiSeq v4
282 chemistry with 75bp paired-end reads. Each batch of 96 samples contained three 100-cell controls.
283 The second batch (lane 7) contained six samples of stage I gametocytes and six samples of stage II
284 gametocytes, each with a single 100-cell control. These were not included in the analysis, leaving
285 176 single cell samples.

286

287 **Mapping reads and calculating read counts**

288 All sequencing experiments were processed in the following way. CRAM files of reads were acquired
289 from the WTSI core pipeline, converted to BAM using *samtools-1.2 view -b*, sorted using *samtools*
290 *sort -n*, converted to fastq using *samtools-1.2 bam2fq* and then deinterleaved³¹. Nextera adaptor
291 sequences were trimmed using *trim_galore -q 20 -a CTGTCTCTTATACACATCT --paired --stringency 3 -*
292 *-length 50 -e 0.1* (v0.4.1). HISAT2 (v2.0.0-beta)³² indexes were produced for the *P. falciparum* v3
293 (<http://www.genedb.org/Homepage/Pfalciparum>) or *P. berghei* v3³³ genome sequences,
294 downloaded from GeneDB³⁴, using default parameters (October 2016). Trimmed, paired reads were
295 mapped to either genome sequence using *hisat2 --max-intronlen 5000 -p 12*. For the dual sort
296 experiment we mapped against a combined reference, allowing us to exclude reads that map to
297 both genomes. SAM files were converted to BAM using *samtools-1.2 view -b* and sorted with
298 *samtools-1.2 sort*. GFF files were downloaded from GeneDB (October 2016) and converted to GTF
299 files using an in-house script. All feature types (mRNA, rRNA, tRNA, snRNA, snoRNA,
300 pseudogenic_transcript and ncRNA) were conserved, with their individual “coding” regions labelled
301 as CDS in every case for convenience. Where multiple transcripts were annotated for an individual
302 gene, only the primary transcript was considered. Reads were summed against genes using HTSeq:

303 *htseq-count -f bam -r pos -s no -t CDS* (v0.6.0;³⁵). Multimapping reads were excluded by default (-a
304 10). For downstream analysis (excluding examination of rRNA counts) transcripts not included in the
305 GeneDB cDNA sequence files were excluded. The read counts for *P. berghei* mixed blood stages, *P.*
306 *falciparum* gametocytes and *P. falciparum* asexual stages are presented in Supplementary Data 6.

307

308 **Classifying reads for quality control**

309 To determine the useful yield of different RNA amplification protocols (summarized in
310 Supplementary Table 1), we classified resulting reads into those mapping to rRNA genes, other
311 genes, unmapped or ambiguous (falling into more than one category). We concentrated here on
312 rRNA because we had observed that this was a particular problem. To do this we began with HISAT2
313 BAM files produced as described above. Total read pairs were all the unique read pair ids. Ribosomal
314 RNA reads were counted using *bedtools intersect* (v2.17.0;³⁶) to find the overlap of unique read pair
315 ids with rRNA features. Other coding reads were counted in the same way, but looking for overlap
316 with all other features. Unmapped reads were identified using *samtools view -f 0x8* (v1.2) and
317 extracting unique read pair ids. Where a read pair occurred in more than one of these lists, it was
318 counted as ambiguous.

319

320 We compared the library complexity of different iterations of our protocol in order to determine
321 whether more reads resulted in more complexity, or simply more reads from the same genes,
322 perhaps due to large numbers of PCR cycles. Different sequencing runs had very different library
323 sizes and so we downsampled the data. To maximise the number of cells included, while also
324 allowing a reasonable number of reads per cell, we chose to downsample to 50000 reads per cell. To
325 do this, 50000 counts from HTSeq were randomly sampled for each cell. Counts associated with
326 protein coding genes were enumerated and genes were called as detected if there were at least 10
327 reads mapping to them.

328

329 **Assessing bias in single-cell sequencing libraries**

330 Different library preparation and sequencing protocols exhibit different biases in representation of
331 GC/AT-rich sequences and 5' or 3' transcript ends. In order to assess such biases we took an
332 approach of using the mapped RNA-seq data to identify fragments of genes which were expressed
333 and examined the coverage of genes by these fragments. The reason for doing this, rather than
334 looking at coverage depth was that we had noticed that genes often did not have full coverage,
335 particularly when very long or expressed at a low level. This suggests that, although we would expect
336 Smartseq-2 to amplify full length transcripts, in some cases only partial transcripts survived the full

337 protocol. We used Stringtie (v1.2; default options;³⁷) to call expressed fragments from our HISAT2
338 BAM files. We then looked for Stringtie transcript features overlapping each mRNA feature in our
339 reference annotation. Where multiple Stringtie transcripts overlapped each other, these were
340 merged. We then determined, for each gene, the exonic sequence covered by the merged Stringtie
341 transcripts. The length, GC content and relative start and end of these regions was calculated.
342 Observed GC content was compared against the GC content for the whole coding region. Each
343 relative position along a coding sequence (0-100) covered by a fragment was incremented for each
344 fragment covering it. The coverage of each relative position for each gene was then normalised
345 between 0 and 1 based on the highest coverage across that coding sequence. To examine the effect
346 of gene length, we compared the length distribution of all 4943 *P. berghei* genes used in our initial
347 analysis to the 4579 which passed our filtering criteria (having at least 10 reads in at least 5 cells).

348

349 **Analysis of contamination with a dual sort of *P. falciparum* and *P. berghei* schizonts**

350 Reads for the dual sort samples were mapped as above, but to a combined reference of both
351 parasites, enabling reads that map equally well to both genomes to be discarded as their origin could
352 not be determined. Read counts were converted to FPKMs and transcripts with an FPKM ≥ 10 were
353 counted as expressed. We used these data to show that no well contained more than one cell, i.e.
354 wells with good data (a large number of expressed genes) never had similar numbers of genes from
355 both species. Furthermore no good wells contained a large number of genes from the incorrect
356 species. To explore whether contaminating genes were similar in different wells, we compared *P.*
357 *falciparum* genes identified in wells with a *P. berghei* cell sorted into them and vice versa between
358 wells. Similarity was calculated as the number of common contaminating genes with an FPKM ≥ 10 ,
359 divided by the average number of contaminating genes between the two wells. Each cell contained
360 relatively few contaminating genes. This was higher for *P. falciparum* contamination of *P. berghei*
361 (Supplementary Figure 2b) than *P. berghei* contamination of *P. falciparum* (Supplementary Figure
362 2c), suggesting *P. falciparum* cells contribute more to extracellular RNA in the medium. Different
363 cells shared relatively few contaminating transcripts, but the more commonly occurring
364 contaminants were also more highly expressed in their cells of origin (Supplementary Figures 2d,e).
365 The contamination in cells was generally very low and reflected the amount of RNA present in the
366 cells of origin.

367

368 **Filtering and normalisation of single-cell read count data**

369 The three main datasets (Pb mixed, Pf asex, Pf sex) were processed using Scater v1.0.4³⁸. Firstly we
370 removed genes with no counts in any cell. We then removed remaining control cells, cells with a

371 total of less than or equal to 25000 read counts and/or less than 1000 genes with at least one read.
372 Subsequently we removed genes which did not have at least 10 reads in 5 cells. For the *P. berghei*
373 dataset this resulted in 144/183 cells and 4579 unique genes detected across all cells. For the *P.*
374 *falciparum* gametocyte dataset there were 191/238 cells and 4454 unique genes after filtering and
375 for the *P. falciparum* asexual dataset 161/180 cells and 4387 unique genes. The counts were then
376 normalised using *scran*³⁹ (v1.0.3). Normalisation is required due to technical variation between
377 samples due to, for example, variable sequencing depth and capture efficiency. Single cell RNA-seq
378 read count data contain many zeroes compared to bulk RNA-seq data. These are caused by drop out
379 of low expressed genes or variation between cells and reduce the accuracy of normalisation
380 methods designed for bulk RNA-seq data. *Scran* uses a pooling approach to reduce these zeroes.
381 Furthermore, it allows an initial clustering of the data and normalisation within these clusters (e.g.
382 cell types), prior to a final normalisation step across the whole dataset. This is particularly useful for
383 our *P. berghei* data, where the asexual, male and female gametocyte cells differ greatly in their
384 expression patterns. The initial clustering step was performed with the *scran* function *quickCluster*
385 (min.size = 30). This resulted in three clusters representing the asexual, male and female gametocyte
386 populations. The *computeSumFactors* function was run using these clusters, with sizes = 20 and
387 positive = TRUE. All downstream analyses were performed with the *scran* normalised data except
388 where stated. For *P. falciparum* gametocytes, the *computeSumFactors* function was run with sizes =
389 15. For *P. falciparum* asexual stages, we set min.size = 20 for *quickCluster* and the
390 *computeSumFactors* function was run with sizes = 10.

391
392 For some applications it is necessary to normalise the data by transcript length. For instance, when
393 comparing ranked gene expression values to reference data for determining life cycle stage of a cell.
394 We therefore normalised the *scran* values by taking the exponent (2^x), multiplying by 1000 and
395 dividing by the cDNA length, determined from the GeneDB cDNA FASTA file (coding sequence only,
396 no UTRs). This is similar to the FPKM calculation, except the library size normalisation is already
397 accomplished. We refer to these values as *l-scran*, for length-normalised *scran* values.

398

399 **Determining parasite life cycle stages using bulk reference data and clustering**

400 We used several bulk RNA-seq data sets to assign a life cycle stage to each cell. For *P. berghei*
401 asexual stages we used microarray data¹¹ that captures the 24-hour asexual development cycle at
402 two-hour resolution. In their experiment Cy5 was used to label each time point while Cy3 was used
403 to label a pool of all samples. The “F635 Median - B635” values are the difference in Cy5 intensity
404 between the median foreground and the median background. This intensity value is related to the

405 actual expression level and these are the values we used. Their data were generated using the *P.*
406 *berghei* v2 genome assembly, so we remapped their probe sequences against v3 using HISAT2
407 (default parameters). We then used `htseq-count -a 200 -f sam -r name -s no` to identify the genes to
408 which the probes mapped (`cut -f1,21 probes_berghei_htseq.sam | grep PBANKA | grep -v`
409 `ambiguous > probes_berghei.map`). We then used the GPR files provided from ArrayExpress ⁴⁰
410 (accession GSE80015) and the probe map to produce a table of percentile ranks for each gene in
411 each condition.

412
413 Single cell gene expression values were converted to length normalised scran values (l-scran), as
414 described above, in order to produce more accurate rank expression levels for our scRNA-seq data.
415 We compared each single-cell expression profile against each reference data set. To reduce noise,
416 genes that do not vary greatly between conditions in the reference data were removed. For the *P.*
417 *berghei* 24h intraerythrocytic developmental cycle reference data ¹¹, genes were only included if
418 their expression profile had a mean rank of greater than 30 and less than 70 and standard deviation
419 in rank across samples of greater than 3. Genes from the query dataset with l-scran < 3 were also
420 removed. A minimum of 100 remaining genes common to both the reference and query profiles
421 were required to calculate a correlation between them. The Spearman rank correlation was used in
422 order make the microarray and RNA-seq datasets more comparable. The best correlation of a single-
423 cell expression profile with a reference expression profile was taken as the consensus stage
424 prediction for that single-cell. As new data (e.g. single-cell analysis of timepoints across the full,
425 synchronised erythrocytic development cycle) become available, benchmarking staging algorithms
426 will become feasible. Bulk RNA-seq data to classify *P. berghei* males and females directly was not
427 available. Therefore, we used bulk RNA-seq data ¹² that includes mixed-sex gametocyte samples,
428 after converting the profiles to v3 using previous id annotation from PlasmoDB ⁴¹.

429
430 To determine distinct groups of single-cells based on their expression patterns we used the
431 clustering tool SC3 ¹⁰. We used the combined Euclidean, Pearson and Spearman distance, plus the
432 combined PCA and spectral transformation. For the *P. berghei* dataset the optimal *k* was 3 (average
433 silhouette width = 0.99), with 4 being nearly as good (average silhouette width = 0.97). We found
434 that the additional cluster split the asexual parasites into trophozoites and schizonts, while both *k*
435 values retained the male and female gametocytes as separate clusters. However, there was still
436 extensive variation within these clusters so we further investigated this by excluding asexual cells
437 and clustering again. With this reduced dataset we were able to get a new, robust clustering with
438 *k*=3 (width = 0.99). Here, outliers from both the putative male and female clusters clustered

439 together, exclusive of the core of male and female clusters. Markers suggested that six of these
440 outlier cells possessed both male genes and asexual genes, while a single cell possessed both female
441 genes and asexual genes. It is possible that these cells are early gametocytes, committed schizonts
442 or cells doubly infected with both asexual and sexual parasites. These were excluded from further
443 analysis. The *markers* function of SC3 (AUROC threshold 0.85, p-value threshold 0.01) was used on
444 the initial clustering, with $k = 3$, to identify novel markers for asexuals, males and females.

445

446 Data from Lasonder and colleagues⁴² was used to classify *P. falciparum* gametocyte cells by sex. Raw
447 count data was downloaded from the Gene Expression Omnibus⁴³ (accession GSE75795) and
448 converted to FPKM. Data from Young and colleagues⁴⁴, was used to classify *P. falciparum* cells along
449 the gametocyte development time course (days 1, 2, 3, 6, 8, 12). For this dataset profiles of ranks
450 were downloaded from PlasmoDB. The Lasonder data⁴² highlighted five male cells, with the rest
451 called as females. The Young data⁴⁴ suggested that all the cells were at a consistent stage of
452 development (eight days), although resolution is lacking at the most relevant timepoints, between
453 eight and twelve days. The classification of each cell is listed in Supplementary Data 5.

454

455 Bulk RNA-seq data from Otto *et al.*²⁴ and Lopez-Barragan *et al.*⁴⁵ were used to classify 161 *P.*
456 *falciparum* asexual stage cells. RNA-seq reads from Otto *et al.*²⁴ for the 36bp Illumina libraries only,
457 were downloaded from the European Nucleotide Archive (accession ERX001048). They were
458 mapped to the *P. falciparum* 3D7 genome sequence using HISAT2 v2.0.0-beta³² and reads were
459 counted using HTSeq v0.6.0³⁵. Read counts were then converted into FPKM for subsequent analysis.
460 RNA-seq reads from Lopez-Barragan *et al.*⁴⁵ were downloaded from the European Nucleotide
461 Archive (accession SRX105940) and mapped to *P. falciparum* 3D7 transcript sequences using
462 Bowtie2 v2.2.9 (-a -X 800;⁴⁶) and eXpress v1.5.1⁴⁷. The resulting read counts were converted to
463 FPKM. The Lopez-Barragan prediction⁴⁵ was used as the consensus prediction, the prediction
464 included 6 stage II gametocytes which were removed from further pseudotime analysis (n=155).

465

466 **Assessment of gene expression variation during asexual maturation**

467 Within the 54 *P. berghei* cells identified as asexual, 277 genes were found to be variable using
468 M3Drop¹⁵ (raw count input, False Discovery Rate ≤ 0.01). L-scran expression values for this subset
469 of genes (expressionFamily=tobit()) was used to order the cells in pseudotime using Monocle 2¹⁴;
470 specifically the reduceDimension() and orderCells (num_path=2) functions were used to derive the
471 ordering of the cells. Monocle 2 identified a single cell state and the cells were ordered in a single
472 trajectory (Fig. 2c). The Monocle 2 package was further used to cluster genes in pseudotime (k-

473 means) with the clusterGenes() function. We looked for enrichment of Gene Ontology terms within
474 the four clusters identified, using topGO⁴⁸ (summarized in Fig. 2d).

475

476 For the *P. falciparum* 155 cell dataset, 361 genes were found to be variable genes with M3Drop (raw
477 count input, False Discovery Rate ≤ 0.01)¹⁵, Monocle 2 identified 2 branches defining three possible
478 trajectories, although 2 of those appeared minor (Supplementary Figure 5). Cells ordered in these
479 minor trajectories did not seem to correlate with known biological markers, such as sexual
480 commitment markers (*ap2-g* and *gvd-1*; Supplementary Figure 5) and were removed for the rest of
481 the analysis. The pseudotime analysis was repeated on the main trajectory cells (125 cells)
482 (Supplementary Figure 5). The Monocle 2 package was further used to cluster genes in pseudotime
483 (k-means) with the clusterGenes() function.

484

485 **Direct comparison of single-cells in pseudotime with bulk RNA-seq data**

486 To determine whether the same set of genes displayed different patterns across development in
487 bulk and single-cell RNA-seq experiments we made direct comparisons between these two
488 approaches. After ordering the *P. berghei* asexual cells by pseudotime, genes were ordered by their
489 peak of expression based on linear (i.e. not logged), length-normalised scran expression values. To
490 do this, expression value data, ordered by pseudotime, were normalised, then Fourier transformed,
491 sorting transcripts according to the phase of the most prominent frequency. Signal-to-noise (S/N)
492 ratios were calculated for each transformed signal and normalised with respect to the maximum
493 achievable value for the dataset. Transforms with a normalised S/N of less than 0.1 were excluded
494 from the results as lacking evidence of periodicity. The Hoo *et al.* dataset¹¹ were treated in the
495 same way, but initially ordered by time point of collection rather than pseudotime and using
496 intensity values as described above. This resulted in 1141 ordered genes for our single cell data and
497 2612 genes for the Hoo data¹¹. There were 651 shared genes, which were used to compare the two
498 datasets.

499

500 The *P. falciparum* asexual cells were ordered differently. We used the Otto *P. falciparum* asexual
501 development cycle time course data²⁴ as a reference. These data were processed as described
502 above. We used the Fourier transform approach described above, with a normalised S/N ratio of 0.5
503 to identify 4517 genes from the Otto *et al.* dataset²⁴. We then identified 336 genes common to this
504 list and the list of 361 differentially expressed genes identified across the 155 single cells. This
505 approach was taken because the window of time captured by our single-cells was too narrow to
506 identify cycling genes using the Fourier approach (all the normalised S/N ratios were very low e.g.

507 <0.05). We then generated heatmaps for the two datasets, with the genes ordered by their peak
508 time in the Otto et al. dataset²⁴ in both cases.

509

510 **Determining gene expression variability within different cell types**

511 To examine gene expression variation within life stages, we used the filtered datasets and
512 considered male, female, trophozoite and schizont cells separately. For *P. berghei*, M3Drop¹⁵ was
513 used to determine gene expression variability amongst cells with FDR \leq 0.05. We identified 115
514 variable genes in *P. berghei* females, 73 in males, 27 in trophozoites and four in schizonts. Twenty
515 variable genes were shared between males and females. One variable gene from trophozoites and
516 schizonts was shared: an L1 type *pir* gene (PBANKA_0943900). Ten were shared between females
517 and trophozoites, none between females and schizonts, four between males and trophozoites, none
518 between males and schizonts. To examine functional classes enriched amongst variable genes we
519 used topGO with the weight01 algorithm, the Fisher statistic, node size = 5 and False Discovery Rate
520 \geq 0.05⁴⁸. Gene ontology terms for *P. berghei* and *P. falciparum* genes were extracted from GeneDB
521 EMBL files. Multigene families in *P. berghei* do not have associated Gene Ontology (GO) terms and so
522 we used *ad hoc* hypergeometric tests to look at their enrichment. We found that *pir* genes were
523 enriched amongst variable genes in trophozoites (2 of 135, hypergeometric test, $p=0.04$), schizonts
524 (1 of 135, hypergeometric test $p=0.005$) and males (5 of 135, hypergeometric test, $p=0.02$), but there
525 were none in females. Those in asexual stages (trophozoites and schizonts) were from the L1
526 subfamily, but those in males were from the S1 and S4 subfamilies.

527

528 Many more variable genes were present in *P. falciparum* females and so we reduced the FDR cut off
529 to 0.01 to improve specificity. We found 448 variable genes in *P. falciparum* females. We were not
530 able to analyse variability in *P. falciparum* males because we found only five of them. Amongst
531 several enriched GO terms were *modulation by symbiont of host erythrocyte* and *cytoadherence to*
532 *microvasculature, mediated by symbiont protein*. These terms refer to 14 *var* genes found amongst
533 the variable genes.

534

535 **Identifying putative functional *var* transcripts**

536 It is known that antisense transcripts are expressed from a bidirectional promoter in the intron of
537 each *var* gene¹⁷. Our protocol does not preserve information about which strand is transcribed.
538 Therefore finding that reads map to either exon of a *var* gene does not provide evidence that it is
539 functionally expressed. In fact it may indicate that the gene is silenced. In order to identify sense
540 transcripts, we looked for reads mapping over the single intron of each *var* gene. These reads that

541 include both exons must originate from transcripts including the intron, and thus indicate that the
542 gene was transcribed on the sense strand, not from antisense transcripts beginning within the
543 intron. Initially we identified reads from the HISAT2 mappings which overlapped annotated *var*
544 genes using *bedtools intersect*. From the resulting BAM file we selected those reads which included
545 an *N* in the CIGAR string, indicating a split read. We then looked for which *var* gene each read
546 overlapped and whether it was split exactly over the intron. We called expression for a *var* gene
547 where there were at least two reads mapping over the intron.

548

549 **Code availability**

550 Perl, R and C++ code for various analyses is available on request.

551

552 **Data Availability**

553 The single-cell RNA-seq reads are available from the European Nucleotide Archive (accession
554 ERP021229) and ArrayExpress (accession E-ERAD-611). Read counts and metadata are also
555 presented in Supplementary Data 5 and 6.

556

557 **Acknowledgements**

558 The Wellcome Trust Sanger Institute is funded by the Wellcome Trust (grant WT098051). CJRI was
559 supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal
560 Society (Grant Number 101239/Z/13/Z). MKNL is supported by an MRC Career Development Award
561 (G1100339). We thank Chris Newbold and John Marioni for comments that improved the
562 manuscript. The authors would like to thank the staff of the Illumina Bespoke Sequencing and Core
563 Cytometry teams at the Wellcome Trust Sanger Institute for their contribution.

564

565 The authors declare no conflict of interest.

566

567 **Contributions**

568 AJR, AMT, HMB and MKNL conceived of and designed the study. AMT and HMB carried out the
569 sorting and library preparation protocols. AMT and ARG performed the parasitological experiments.
570 AJR, AMT and HMB designed the sequencing experiments. AJR and AMT analysed the data. CJRI
571 performed the Fourier transform analysis. MJS coordinated sequencing experiments. OB supervised
572 ARG. AJR, AMT, MB & MKNL wrote the manuscript. All authors read and critically revised the
573 manuscript.

574

575 References

- 576 1. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and
577 future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).
- 578 2. Rovira-Graells, N. *et al.* Transcriptional variation in the malaria parasite *Plasmodium falciparum*.
579 *Genome Res.* **22**, 925–938 (2012).
- 580 3. Sinha, A. *et al.* A cascade of DNA-binding proteins for sexual commitment and development in
581 *Plasmodium*. *Nature* **507**, 253–257 (2014).
- 582 4. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in
583 malaria parasites. *Nature* **507**, 248–252 (2014).
- 584 5. Tembo, D. L. *et al.* Differential PfEMP1 expression is associated with cerebral malaria pathology.
585 *PLoS Pathog.* **10**, e1004537 (2014).
- 586 6. Sharma, P. *et al.* An epigenetic antimalarial resistance mechanism involving parasite genes
587 linked to nutrient uptake. *J. Biol. Chem.* **288**, 19429–19440 (2013).
- 588 7. Kissinger, J. C. & DeBarry, J. Genome cartography: charting the apicomplexan genome. *Trends*
589 *Parasitol.* **27**, 345–354 (2011).
- 590 8. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat.*
591 *Methods* **10**, 1096–1098 (2013).
- 592 9. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-
593 cell RNA-seq. *Nature* **509**, 371–375 (2014).
- 594 10. Kiselev, V. Y. *et al.* SC3 - consensus clustering of single-cell RNA-Seq data. *bioRxiv* 036558
595 (2016). doi:10.1101/036558
- 596 11. Hoo, R. *et al.* Integrated analysis of the *Plasmodium* species transcriptome. *EBioMedicine* **7**,
597 255–266 (2016).
- 598 12. Otto, T. D. *et al.* A comprehensive evaluation of rodent malaria parasite genomes and gene
599 expression. *BMC Biol.* **12**, 86 (2014).
- 600 13. Bozdech, Z. *et al.* Expression profiling of the schizont and trophozoite stages of *Plasmodium*

- 601 falciparum with a long-oligonucleotide microarray. *Genome Biol.* **4**, R9 (2003).
- 602 14. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
603 pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
- 604 15. Andrews, T. S. & Hemberg, M. Modelling dropouts allows for unbiased identification of marker
605 genes in scRNASeq experiments. doi:10.1101/065094
- 606 16. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and
607 genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from
608 the malaria parasite. *PLoS Pathog.* **6**, e1001165 (2010).
- 609 17. Guizetti, J. & Scherf, A. Silence, activate, poise and switch! Mechanisms of antigenic variation in
610 Plasmodium falciparum. *Cell. Microbiol.* **15**, 718–726 (2013).
- 611 18. Brugat, T. *et al.* Antibody-independent mechanisms regulate the establishment of chronic
612 Plasmodium infection. *Nat Microbiol* **2**, 16276 (2017).
- 613 19. Talman, A. M. *et al.* Proteomic analysis of the Plasmodium male gamete reveals the key role for
614 glycolysis in flagellar motility. *Malar. J.* **13**, 315 (2014).
- 615 20. Reid, A. J. Large, rapidly evolving gene families are at the forefront of host-parasite interactions
616 in Apicomplexa. *Parasitology* **142 Suppl 1**, S57–70 (2015).
- 617 21. Mair, G. R. *et al.* Regulation of sexual development of Plasmodium by translational repression.
618 *Science* **313**, 667–669 (2006).
- 619 22. Liu, Y. *et al.* The conserved plant sterility gene HAP2 functions after attachment of fusogenic
620 membranes in Chlamydomonas and Plasmodium gametes. *Genes Dev.* **22**, 1051–1068 (2008).
- 621 23. Moss, D. K. *et al.* Plasmodium falciparum 19-kilodalton merozoite surface protein 1 (MSP1)-
622 specific antibodies that interfere with parasite growth in vitro can inhibit MSP1 processing,
623 merozoite invasion, and intracellular parasite development. *Infect. Immun.* **80**, 1280–1287
624 (2012).
- 625 24. Otto, T. D. *et al.* New insights into the blood-stage transcriptome of Plasmodium falciparum
626 using RNA-Seq. *Mol. Microbiol.* **76**, 12–24 (2010).

- 627 25. Khan, S. M., Kroeze, H., Franke-Fayard, B. & Janse, C. J. Standardization in generating and
628 reporting genetically modified rodent malaria parasites: the RMgmDB database. *Methods Mol.*
629 *Biol.* **923**, 139–150 (2013).
- 630 26. Gomes, A. R. *et al.* A genome-scale vector resource enables high-throughput reverse genetic
631 screening in a malaria parasite. *Cell Host Microbe* **17**, 404–413 (2015).
- 632 27. Talman, A. M., Blagborough, A. M. & Sinden, R. E. A Plasmodium falciparum strain expressing
633 GFP throughout the parasite's life-cycle. *PLoS One* **5**, e9156 (2010).
- 634 28. Kutner, S., Breuer, W. V., Ginsburg, H., Aley, S. B. & Cabantchik, Z. I. Characterization of
635 permeation pathways in the plasma membrane of human erythrocytes infected with early
636 stages of Plasmodium falciparum: association with parasite development. *J. Cell. Physiol.* **125**,
637 521–527 (1985).
- 638 29. Fivelman, Q. L. *et al.* Improved synchronous production of Plasmodium falciparum gametocytes
639 in vitro. *Mol. Biochem. Parasitol.* **154**, 119–123 (2007).
- 640 30. Ribaut, C. *et al.* Concentration and purification by magnetic separation of the erythrocytic
641 stages of all human Plasmodium species. *Malar. J.* **7**, 45 (2008).
- 642 31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079
643 (2009).
- 644 32. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
645 requirements. *Nat. Methods* **12**, 357–360 (2015).
- 646 33. Fougère, A. *et al.* Variant Exported Blood-Stage Proteins Encoded by Plasmodium Multigene
647 Families Are Expressed in Liver Stages Where They Are Exported into the Parasitophorous
648 Vacuole. *PLoS Pathog.* **12**, e1005917 (2016).
- 649 34. Logan-Klumpler, F. J. *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res.*
650 **40**, D98–108 (2012).
- 651 35. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput
652 sequencing data. *Bioinformatics* **31**, 166–169 (2015).

- 653 36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
654 *Bioinformatics* **26**, 841–842 (2010).
- 655 37. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis
656 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
- 657 38. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. scater: pre-processing, quality
658 control, normalisation and visualisation of single-cell RNA-seq data in R. *bioRxiv* 069633 (2016).
659 doi:10.1101/069633
- 660 39. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA
661 sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
- 662 40. Parkinson, H. *et al.* ArrayExpress--a public repository for microarray gene expression data at the
663 EBI. *Nucleic Acids Res.* **33**, D553–5 (2005).
- 664 41. Aurrecochea, C. *et al.* EuPathDB: the eukaryotic pathogen genomics database resource.
665 *Nucleic Acids Res.* **45**, D581–D591 (2017).
- 666 42. Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of *P. falciparum*
667 gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic*
668 *Acids Res.* (2016). doi:10.1093/nar/gkw536
- 669 43. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.*
670 **41**, D991–5 (2013).
- 671 44. Young, J. A. *et al.* The *Plasmodium falciparum* sexual development transcriptome: a microarray
672 analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **143**, 67–79
673 (2005).
- 674 45. López-Barragán, M. J. *et al.* Directional gene expression and antisense transcripts in sexual and
675 asexual stages of *Plasmodium falciparum*. *BMC Genomics* **12**, 587 (2011).
- 676 46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–
677 359 (2012).
- 678 47. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing

- 679 experiments. *Nat. Methods* **10**, 71–73 (2013).
- 680 48. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene
681 expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
- 682 49. Modrzynska, K. *et al.* A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting
683 Transcriptional Regulators Controlling the Plasmodium Life Cycle. *Cell Host Microbe* **21**, 11–22
684 (2017).
- 685 50. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality
686 control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* (2017).
687 doi:10.1093/bioinformatics/btw777

688
689
690
691

692

693 Supplementary Material

694

695 Supplementary Figure 1. Analysis of biases in *Plasmodium berghei* mixed blood stage single-cell RNA-
696 seq libraries.

697 Supplementary Figure 2. Dual sorting of two parasite species shows minimal contaminating RNA

698 Supplementary Figure 3. Distribution of gene counts.

699 Supplementary Figure 4. The same subsets of transcripts show different patterns of expression

700 around the end of the asexual cell cycle in conventional bulk RNA-seq data and pseudotime

701 reconstructions of single cell RNAseq data.

702 Supplementary Figure 5. Pseudotime reconstruction of the late asexual trajectory of *P. falciparum*.

703 Supplementary Figure 6. Analysis of the co-expression pattern of the ApiAP2 family of transcription
704 factors (TFs).

705 Supplementary Figure 7. Principal Components Analysis and classification of *P. falciparum*

706 gametocyte cells

707

708 Supplementary Table 1. Reagents permuted during optimisation of the single cell RNAseq protocol

709 and stats of each treatment condition after sequencing.

710

711 Supplementary Data 1. Marker genes identifying *P. berghei* mixed stage k-means clusters

712 Supplementary Data 2. Highly variable genes in *P. berghei* female gametocytes, male gametocytes,
713 trophozoites, schizonts and *P. falciparum* female gametocytes.

714 Supplementary Data 3. GO terms enriched amongst highly variable *P. berghei* and *P. falciparum*

715 genes

716 Supplementary Data 4. *Pir* gene expression in different cell types

717 Supplementary Data 5. Samples sequenced in this study

718 Supplementary Data 6. Gene count tables for the three large datasets included in the study

719

720

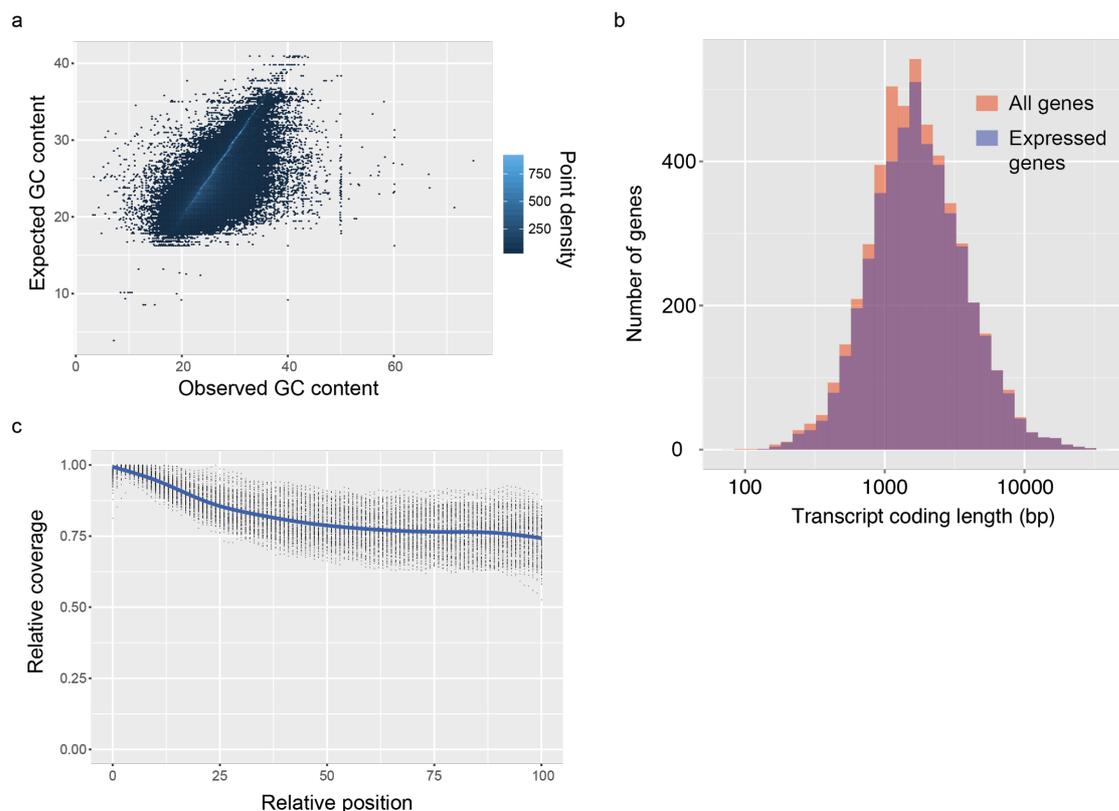
721

722

723

724 Supplementary Figures

725



726

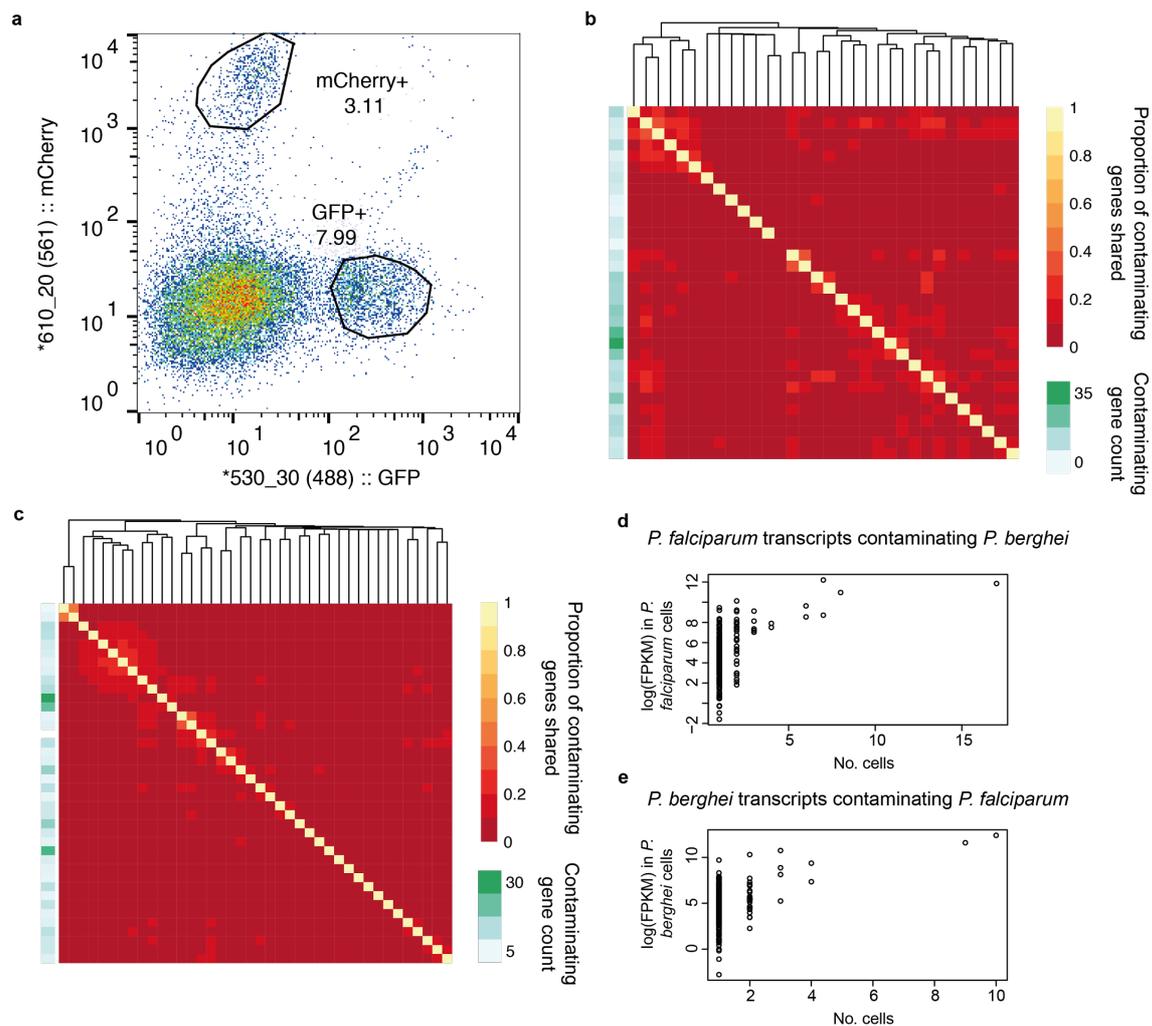
727

728 **Supplementary Figure 1. Analysis of biases in *Plasmodium berghei* mixed blood stage single-cell** 729 **RNA-seq libraries.**

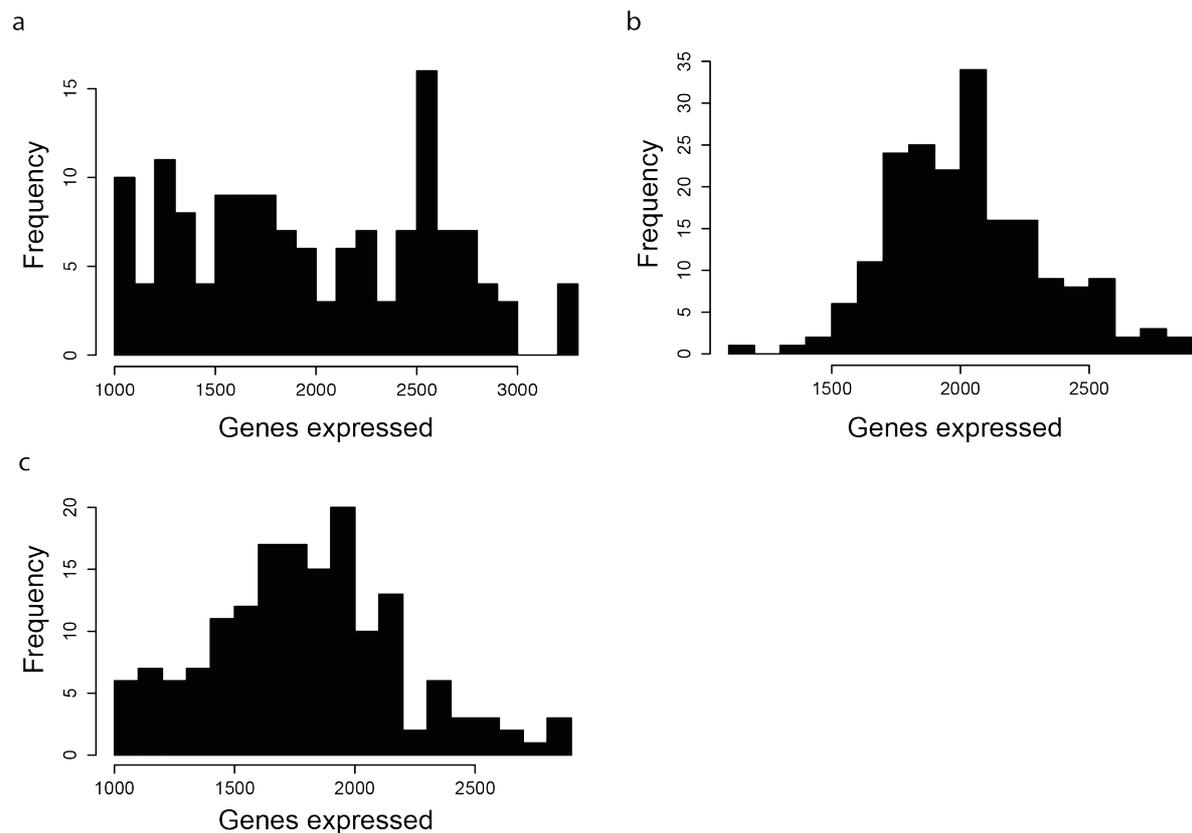
730 (a) The GC content of transcript fragments agreed well with the GC content of genes. There was no
731 apparent over- or under-representation of GC rich regions. (b) Expressed genes (those with at least
732 10 reads in at least 5 cells) were representative of average gene length, suggesting that although the
733 reverse transcriptase might not copy the whole of long transcripts, fragments of long genes are still
734 detected. (c) Sequencing library preparation often introduces end bias, where either the 5' or 3' end
735 of transcripts tend to be better covered. Our protocol introduced a small 5'-bias, which could be
736 attributable to the reverse transcription sometimes initiating within transcripts in internal polyA
737 regions, rather than in the 3' poly-A tail.

738

739



740
 741 **Supplementary Figure 2. Dual sorting of two parasite species shows minimal contaminating**
 742 **RNA** (a) Purified asexual late blood stage of GFP *P. falciparum* and mCherry *P. berghei* were mixed at
 743 a 1:1 ratio, inactivated in RNALater, and sorted individually by flow cytometry, gated on respective
 744 fluorescent channels. The proportion of shared contaminating transcripts between pairs of cells was
 745 low for *P. falciparum* transcripts in *P. berghei* cells (b) and even lower for *P. berghei* transcripts in *P.*
 746 *falciparum* cells (c). Overall there were 273 unique *P. falciparum* transcripts contaminating the *P.*
 747 *berghei* transcriptomes, although no cell had more than 37 contaminating transcripts and no pair of
 748 cells shared more than 17. There was a total of 258 *P. berghei* transcripts contaminating *P.*
 749 *falciparum* transcriptomes, although no cell had more than 32 of these and no pair of cells shared
 750 more than 10. The data suggest that *P. falciparum* schizonts cause more contamination than *P.*
 751 *berghei* schizonts. More commonly occurring, contaminating transcripts are more highly expressed
 752 in their cells of origin for both *P. falciparum* transcripts contaminating *P. berghei* cells (d) and *P.*
 753 *berghei* transcripts contaminating *P. falciparum* cells (e). This suggests that contaminants reflect the
 754 observed pool of transcripts.
 755



756

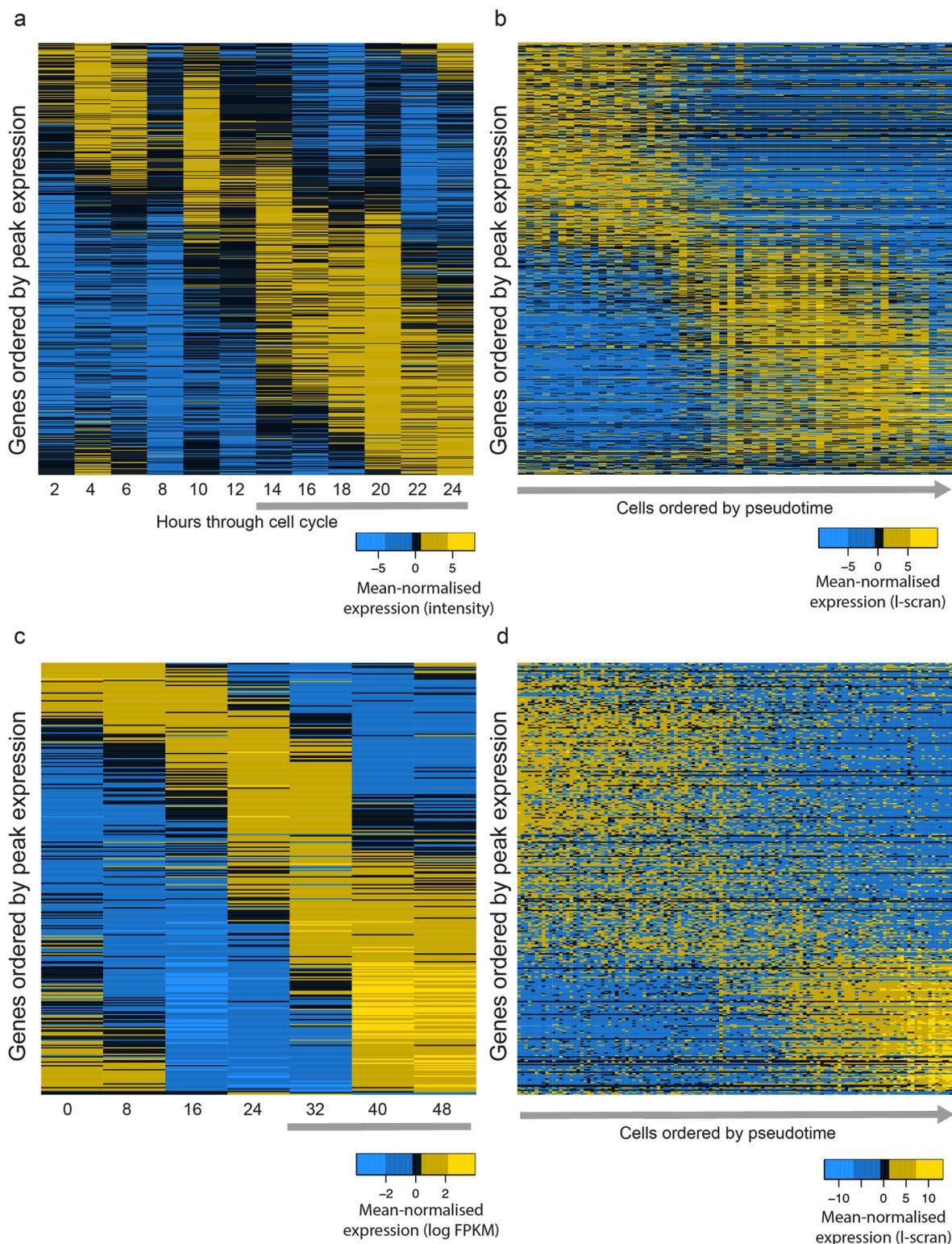
757

758 **Supplementary Figure 3. Distribution of gene counts.**

759 Histograms of expressed gene number after filtering for (a) 144 *P. berghei* mixed blood stage cells,
760 (b) 191 *P. falciparum* gametocytes and (c) 161 *P. falciparum* asexual cells. The greater distribution of
761 gene counts in *P. berghei* is due to the greater variety of cell types in this dataset. Female
762 gametocytes for instance, consistently had a greater number of genes expressed.

763

764



765
766
767
768
769
770

Supplementary Figure 4. The same subsets of transcripts show different patterns of expression around the end of the asexual cell cycle in conventional bulk RNA-seq data and pseudotime reconstructions of single cell RNAseq data.

771 A shared set of 651 genes identified as following a sigmoidal expression pattern through the
772 intraerythrocytic developmental cycle (see Methods) are shown in both bulk transcriptome data ¹¹
773 **(a)** and single cell data ordered by pseudotime **(b)** for *P. berghei*. A much more dramatic shift in gene
774 expression is observed in the single-cell transcriptome data. A similar pattern is observed between *P.*
775 *falciparum* bulk ²⁴ **(c)** and single cell **(d)** RNA-seq. In panels b and d, gene expression patterns are
776 mean-normalised I-scran values. Only late stage parasites (grey bars in bulk reference datasets) are
777 expected to be present in the single cell datasets.

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

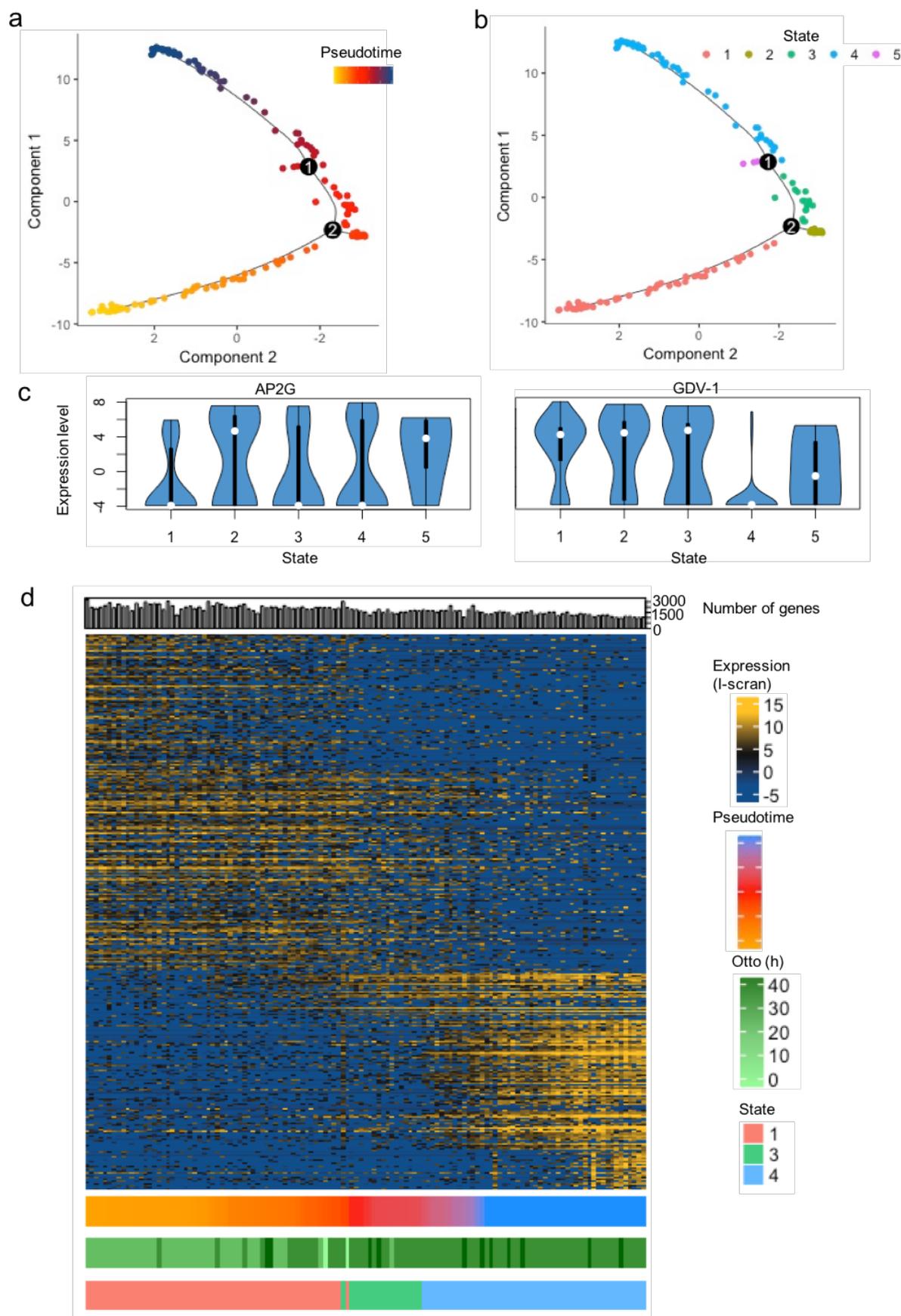
799

800

801

802

803



804

805 **Supplementary Figure 5. Pseudotime reconstruction of the late asexual trajectory of *P. falciparum*.**

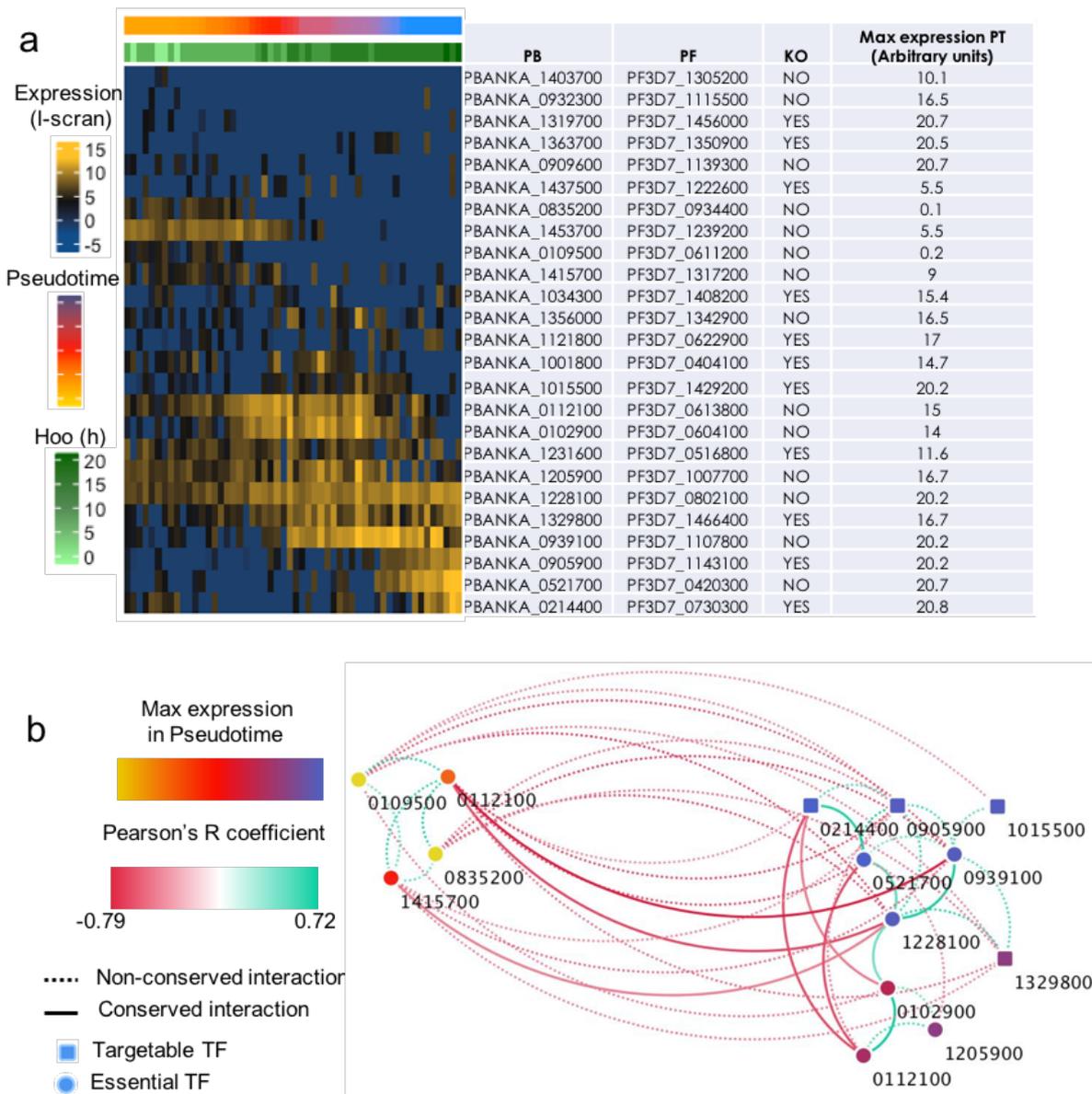
806 PCA of 155 *P. falciparum* cells colored by pseudotime (**a**) or Monocle state (**b**); identified trajectory
807 branches are displayed as circled number 1 and 2. (**c**) Expression of sexual commitment markers
808 *ap2-g* (PF3D7_1222600) and *gdv-1* (PF3D7_0935400) in cells of different states. (**d**) Differentially
809 expressed genes were plotted along pseudotime for cells in the main trajectory (States 1, 3 and 4).
810 The number of genes per cell is displayed on top of the heat map, whilst the pseudotime, the
811 maturation prediction²⁴ and the cell state are displayed on the side of the heatmap. The transition
812 between trophozoites and schizonts is associated with a hard transcriptional shift, as seen for *P.*
813 *berghei* (Fig. 2d).

814

815

816

817



818

819 **Supplementary Figure 6. Analysis of the co-expression pattern of the ApiAP2 family of**

820 **transcription factors (TFs).**

821 (a) Expression of *P. berghei* ApiAP2 genes in pseudotime. The *P. falciparum* ortholog is indicated

822 (PF), as well as its established or custom short name (Pb), it's peak pseudotime expression (Max

823 expression PT) and the ability to disrupt the TF as observed in a recent genetic screen (KO)⁴⁹. (b) A

824 network analysis was conducted using significant positive and negative interactions ($p < 0.05$ by

825 Pearson's correlation) of 25 different TFs and weighted according to their Pearson correlation

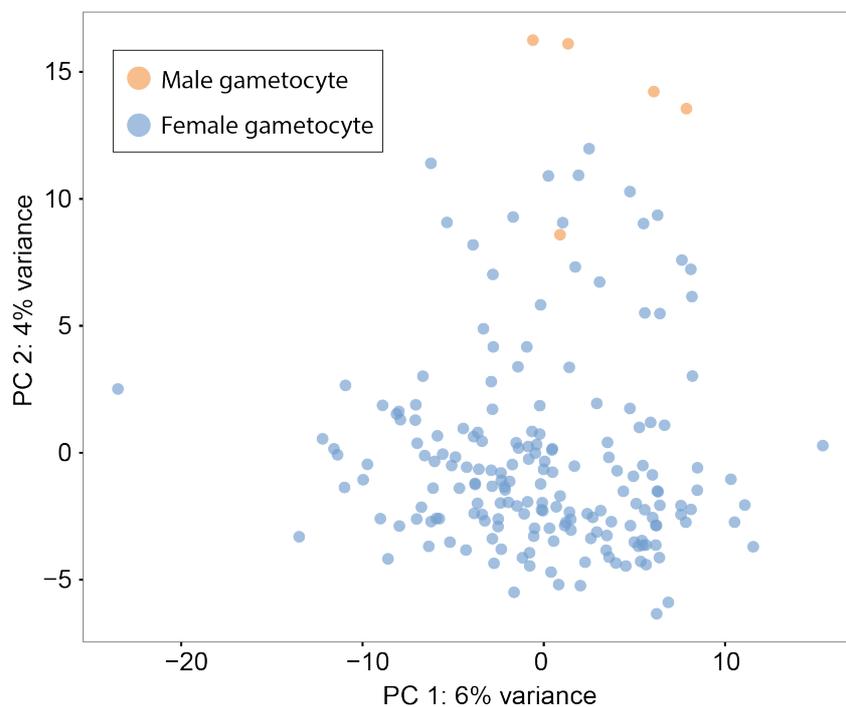
826 coefficient. A similar analysis in *P. falciparum* revealed that some of these co-expression interactions

827 are conserved within the genus (solid lines). Maximum TF expression along pseudotime appears to

828 be important to the structure of the network, strongly suggesting a coordinated regulation cascade

829 of different members of the family during the trophozoite to schizont transition.

830
831
832



833
834

835 **Supplementary Figure 7. Principal Components Analysis and classification of *P. falciparum***
836 **gametocyte cells**

837 A combination of Principal Components Analysis (PCA), k-means clustering and comparison to bulk
838 RNA-seq datasets was used to classify 191 high quality *P. falciparum* gametocytes. A consensus of
839 clustering and comparison to bulk RNA-seq allowed us to distinguish male gametocytes and female
840 gametocytes.

841

842
843

844
845
846

Supplementary Tables/Data

Conditions tested	Protocol	SSII, V30, 30 cycles	SSII, T30, 30 cycles	SmSc, T30, 30 cycles	SSII, T30, 25 cycles	SmSc, T30, 25 cycles			
	Cells	Sexual	Asexual	Asexual	Asexual	Asexual	Sexual	Mixed blood	Asexual
	Species	Pf	Pf	Pf	Pf	Pf	Pf	Pb	Pf
Lysis buffer volume	2 μ l	✓							
	4 μ l		✓	✓	✓	✓	✓	✓	✓
Oligo Dt (IDT)	Anchored 30 bp	✓							
	Non-Anchored 30 bp		✓	✓	✓	✓	✓	✓	✓
Reverse transcriptase	Superscript II (Life Technologies) 10U	✓	✓		✓				
	Smartscribe (Clontech) 5U			✓		✓	✓	✓	✓
Cycle number	25				✓	✓	✓	✓	✓
	30	✓	✓	✓					
Sequencing machine	HiSeq						✓	✓	✓
	MiSeq	✓	✓	✓	✓	✓			
Sequencing results summary	% rRNA	5.7	33.5	36.2	6.4	18.4	17.8	16.7	34.8
	% coding genes	4.4	11.3	39.3	10.5	33	51.7	49	40.5
	% other	90	55.2	24.4	83.1	48.6	30.5	34.2	24.6
	Median genes detected for 50k reads	25	84	145	174	181	502.5	NA	NA
	Total cells	5	6	6	6	6	237	182	174
	Cells passing filters	NA	NA	NA	NA	NA	191	144	161
	Median gene count	NA	NA	NA	NA	NA	2011	1922.5	1793

847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867

Supplementary Table 1. Reagents permuted during optimisation of the single cell RNAseq protocol and stats of each treatment condition after sequencing.

Different combinations of the protocol were tested by sequencing. Initial trials were performed with 2 μ l of lysis buffer, this was increased to 4 μ l to augment capture efficiency. Permutations of the protocol that were tested were a terminal anchoring base (A,G,C; **V**) or not (**T**), 2 reverse transcriptase enzymes (Smartscribe (SmSc); Superscript II (SII)) and 25 or 30 cycles of preamplification. Both sexual and asexual cells of *P. berghei* and *P. falciparum* were tested. For each sequenced dataset, we calculated the mean percentages of rRNA, mRNA and other reads across the cells. For some samples we also downsampled the data to 50,000 reads per cell to allow comparison of the number of genes detected. This was done to determine differences in the complexity of each library. For the three larger datasets produced (*P. falciparum* gametocytes, *P. berghei* mixed blood stages, and *P. falciparum* asexual stages), we provide the numbers of pre- and post-filtered cells and median number of genes in those filtered cells.

868 **Supplementary Data 1. Marker genes identifying *P. berghei* mixed stage k-means clusters**

869 Marker genes are those for which expression is reliably associated with a particular cluster. They
870 were identified for clusters in the k = 3 means clustering of *P. berghei* mixed blood stages using the
871 *markers* function of the SC3 package. Cluster 1 corresponds to asexuals, cluster 2 to males and
872 cluster 3 to females. The majority consensus is the cell type most common in that cluster. However,
873 the only alternative cells in these clusters are the small number of outliers. AUROC is the Area Under
874 the Receiver Operating Characteristics curve and is a measure of the reliability of the marker. We
875 used an AUROC cut off of 0.85 and a *p*-value cut off of 0.01.

876 See Excel file “Supplementary Data 1.xls”

877

878 **Supplementary Data 2. Highly variable genes in *P. berghei* female gametocytes, male gametocytes,
879 trophozoites, schizonts and *P. falciparum* female gametocytes.** The *p* and *q* (corrected) values given
880 for each gene are those determined using M3Drop.

881 See Excel file “Supplementary Data 2.xls”

882

883 **Supplementary Data 3. Gene Ontology terms enriched amongst highly variable *P. berghei* and *P.*
884 *falciparum* genes.**

885 Enriched GO terms were determined using topGO. The total terms are those in the whole set of
886 transcripts and multiple-hypothesis testing corrected significance values (*q* values) are shown.

887 See Excel file “Supplementary data 3.xls”

888

889 **Supplementary Data 4. *Pir* gene expression in different *P. berghei* cell types**

890 *Pir* gene expression values for *P. berghei* mixed blood stages are shown, relating to Figure 3a.

891 Expression is measured by length normalised scran, or l-scran values.

892 See Excel file “Supplementary Data 4.xls”

893

894 **Supplementary Data 5. Samples sequenced in this study**

895 A full list of the samples related to this study is presented, linking the data we present to sequence
896 data identifiers used by the European Nucleotide Archive (*sanger_sample_id*). Type: SC = single cell,
897 Hcell = 100 cells, NoCell = no cell control. Run = internal Illumina run number. Lane = Illumina lane
898 number. Tag = Internal Nextera barcode number. *is_control* = is the sample a control sample?
899 *pass_filter* = did the sample pass our filtering criteria. *consensus* = consensus annotation for that
900 cell. For *P. berghei* mixed blood stages: *sc3_k4* = cluster identifier for k means clustering (k=4),
901 *sc3_k3* = cluster identifier for k means clustering (k=3), *hoo* = best matching sample from Hoo *et al.*¹¹

902 bulk microarray dataset of cell cycle, otto = best matching sample from Otto *et al.*¹² bulk RNA-seq
903 dataset of life cycle stages. For *P. falciparum*: sc3_female_k3 = cluster identifier for k means
904 clustering of female-only samples (k=3), lasonder = best matching samples from⁴² bulk RNA-seq
905 dataset of male and female parasites, young = best matching samples from Young *et al.* bulk
906 microarray dataset⁴⁴ of gametocyte development. For *P. falciparum* asexual stages, lopez = best
907 matching sample from Lopez-Barragnan *et al.*⁴⁵, otto = best matching sample from Otto *et al.*²⁴ and
908 pseudotime_state = pseudotime state, where states 2 and 5 were filtered out.

909 See Excel file "Supplementary Data 5.xls"

910

911 **Supplementary Data 6. Gene counts for the datasets included in the study**

912 Gene counts for all cells (included those we failed) for the *P. berghei* mixed blood stages, the *P.*
913 *falciparum* gametocytes, and the *P. falciparum* asexuals are presented here. These data, along with
914 the corresponding metadata in Supplementary Data 5 can be used to replicate and/or extend our
915 analysis using, for instance, the *scater* package⁵⁰.

916 See Excel file "Supplementary Data 6.xls"