Chappell et al. BMC Genomics (2020) 21:395
https://doi.org/10.1186/s12864-020-06787-5

BMC Genomics

RESEARCH ARTICLE Open Access

Refining the transcriptome of the human malaria parasite *Plasmodium falciparum* using amplification-free RNA-seq

Lia Chappell[1], Philipp Ross[2,3], Lindsey Orchard[2], Timothy J. Russell[2], Thomas D. Otto[1,4], Matthew Berriman[1], Julian C. Rayner[1,5] and Manuel Llinás[2,6*]

Abstract

Background: *Plasmodium* parasites undergo several major developmental transitions during their complex lifecycle, which are enabled by precisely ordered gene expression programs. Transcriptomes from the 48-h blood stages of the major human malaria parasite *Plasmodium falciparum* have been described using cDNA microarrays and RNA-seq, but these assays have not always performed well within non-coding regions, where the AT-content is often 90–95%.

Results: We developed a directional, amplification-free RNA-seq protocol (DAFT-seq) to reduce bias against AT-rich cDNA, which we have applied to three strains of *P. falciparum* (3D7, HB3 and IT). While strain-specific differences were detected, overall there is strong conservation between the transcriptional profiles. For the 3D7 reference strain, transcription was detected from 89% of the genome, with over 78% of the genome transcribed into mRNAs. We also find that transcription from bidirectional promoters frequently results in non-coding, antisense transcripts. These datasets allowed us to refine the 5′ and 3′ untranslated regions (UTRs), which can be variable, long (> 1000 nt), and often overlap those of adjacent transcripts.

Conclusions: The approaches applied in this study allow a refined description of the transcriptional landscape of *P. falciparum* and demonstrate that very little of the densely packed *P. falciparum* genome is inactive or redundant. By capturing the 5′ and 3′ ends of mRNAs, we reveal both constant and dynamic use of transcriptional start sites across the intraerythrocytic developmental cycle that will be useful in guiding the definition of regulatory regions for use in future experimental gene expression studies.

Background

There are six species of *Plasmodium* that are known to cause malaria in humans, but most of the estimated 405, 000 annual deaths are caused by *Plasmodium falciparum* [1]. Although *Plasmodium* spp. have a complex life cycle that involves both invertebrate and vertebrate hosts, it is the asexual development of the parasite in the blood that is responsible for all clinical symptoms of malaria. Blood stage development begins when a newly released, extracellular parasite (a merozoite) invades an erythrocyte, establishing the ring stage of infection, which progresses to the trophozoite stage, during which the infected erythrocyte is extensively modified to enable parasite proliferation [2]. The parasite then divides to form a connected group of daughter cells, termed the schizont, which eventually lyses the host erythrocyte, releasing the newly formed merozoites to invade new erythrocytes. Collectively, these steps are known as the intraerythrocytic developmental cycle (IDC), and take 48 h to complete in *P. falciparum.*

* Correspondence: manuel@psu.edu
[2]Department of Biochemistry & Molecular Biology and Huck Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA
[6]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA
Full list of author information is available at the end of the article

The *P. falciparum* genome is 23.3 Mb in size and encodes over 5400 genes [3]. Most parasite genes are transcriptionally regulated during the IDC, often expressed across multiple time points but with a single peak of maximum abundance per gene [4–6]. Another study has compared the IDC transcriptome profiles of three laboratory strains (3D7, HB3 and Dd2; with origins in West Africa, Latin America and Asia, respectively), demonstrating that gene expression was remarkably conserved between strains from across the globe, despite the strains being isolated at different times from disparate geographical locations [7].

These initial analyses of the *P. falciparum* IDC were based on cDNA microarray technology. The first application of RNA-seq to the *P. falciparum* IDC led to alterations in the gene models for over 10% of the ~ 5400 genes, including the identification of 121 new coding sequences [8]. This study also confirmed 75% of predicted splice sites and conservatively detected 84 cases of alternative splicing. However, the limitations of available RNA-seq technology at that time meant that the extremely AT-rich UTRs could not be detected on a genome-wide scale; this was probably caused by a combination of difficulties generating AT-rich cDNA and PCR bias against the AT-rich sequences.

The extreme AT-content of the *P. falciparum* genome remains challenging even for current sequencing and alignment technologies; within coding sequences the AT-content is ~ 75%, but in non-coding regions the AT-content rises to ~ 90–95%. Successive RNA-seq studies [9–16] each used protocols that were not fully optimised for generation of AT-rich cDNA, preventing the full extent of transcription outside of the protein-coding regions of the genome from being captured. One source of bias is random priming of extremely AT-rich RNA fragments; primer binding to these AT-rich fragments is less stable, resulting in fewer AT-rich fragments being converted to cDNA during reverse transcription [17]. The use of PCR amplification is another source of bias in RNA-seq, the effect of which is particularly pronounced with AT-rich sequences [18]. In whole genome sequencing, the AT-bias can be dramatically reduced with "PCR-free" Illumina sequencing adaptors and omission of PCR amplification steps (Kozarewa et al. [19]). Several of the previous *P. falciparum* RNA-seq studies described thousands of non-coding RNA molecules (ncRNAs) originating from the most AT-rich regions of the genome (López-Barragán et al. [9]; Vignali et al. [10]; Sorber et al. [11]; Hoeijmakers et al. [12]; Siegel et al. [13]; Broadbent et al. [14]; Toenhake et al. [16]), but gaps in sequence coverage in these regions limited assembly of complete transcripts. Many of these predicted ncRNAs have subsequently been discarded during reannotation [20].

To explore the *P. falciparum* transcriptome with minimal bias from the extreme AT-content, we have developed a directional, amplification-free RNA-seq protocol (DAFT-seq) that produces more accurate measures of gene expression. Analysis of the resulting DAFT-seq data revealed extensive transcription between coding regions, particularly of long and often overlapping UTRs. We then applied DAFT-seq to the IDCs of three strains of *P. falciparum*: 3D7 (the genome reference strain, and presumed to be of West African origin) [21], HB3 (a drug sensitive isolate from Honduras) [22] and IT (widely used for studies of antigenic variation) [23].

We identified relatively few differences in transcript levels and transcription start sites (TSSs) between these strains. To specifically capture the 5′ ends of mRNAs from the 3D7 strain, we developed a modified amplification-free RNA-seq protocol (5UTR-seq), and confirmed multiple features by sequencing long cDNA molecules from the same parasite RNA using the Pacific Biosciences (PacBio) platform. The PacBio platform has been used to sequence *Plasmodium* genomic DNA [24–29], but not yet cDNA. Collectively, these new approaches provide a new view of the *P. falciparum* transcriptome at a greater level of resolution. We provide precise definitions of the boundaries of coding transcripts and comprehensively define 5′ and 3′ UTRs, and TSS positions, on a genome-wide scale. In particular, transcription from bidirectional promoters and overlapping transcripts are common features. These data will be informative for both experimental genetic studies and for further dissecting the mechanisms of transcriptional regulation in *Plasmodium spp.*

Results

Directional, amplification-free RNA-seq (DAFT-seq) reveals extensive transcription from the 3D7 genome

We developed an optimised directional, amplification-free RNA-seq protocol (DAFT-seq; Figure S1, supplementary materials) that uses adaptors that eliminate the need for PCR (Kozarewa et al. [19]), even for low input quantities of total RNA (≥500 ng). A further critical modification was to synthesise full-length cDNA molecules, which were then fragmented to make libraries; this gave more even coverage in AT-rich regions (Figure S2). To map transcripts throughout the IDC, DAFT-seq libraries were generated from seven RNA samples taken from tightly synchronized *P. falciparum* 3D7 parasites at 8-h intervals from 0 to 48 h. For most DAFT-seq libraries we obtained around 10 million reads that mapped to the parasite genome (Table S1). Mapping of these libraries showed that the majority of each chromosome sequence is transcribed, as shown for chromosome 1 in Fig. 1a. A striking feature of the data is the extent to which the transcripts extend beyond existing annotated protein-coding exons, defining much larger 5′ and 3′

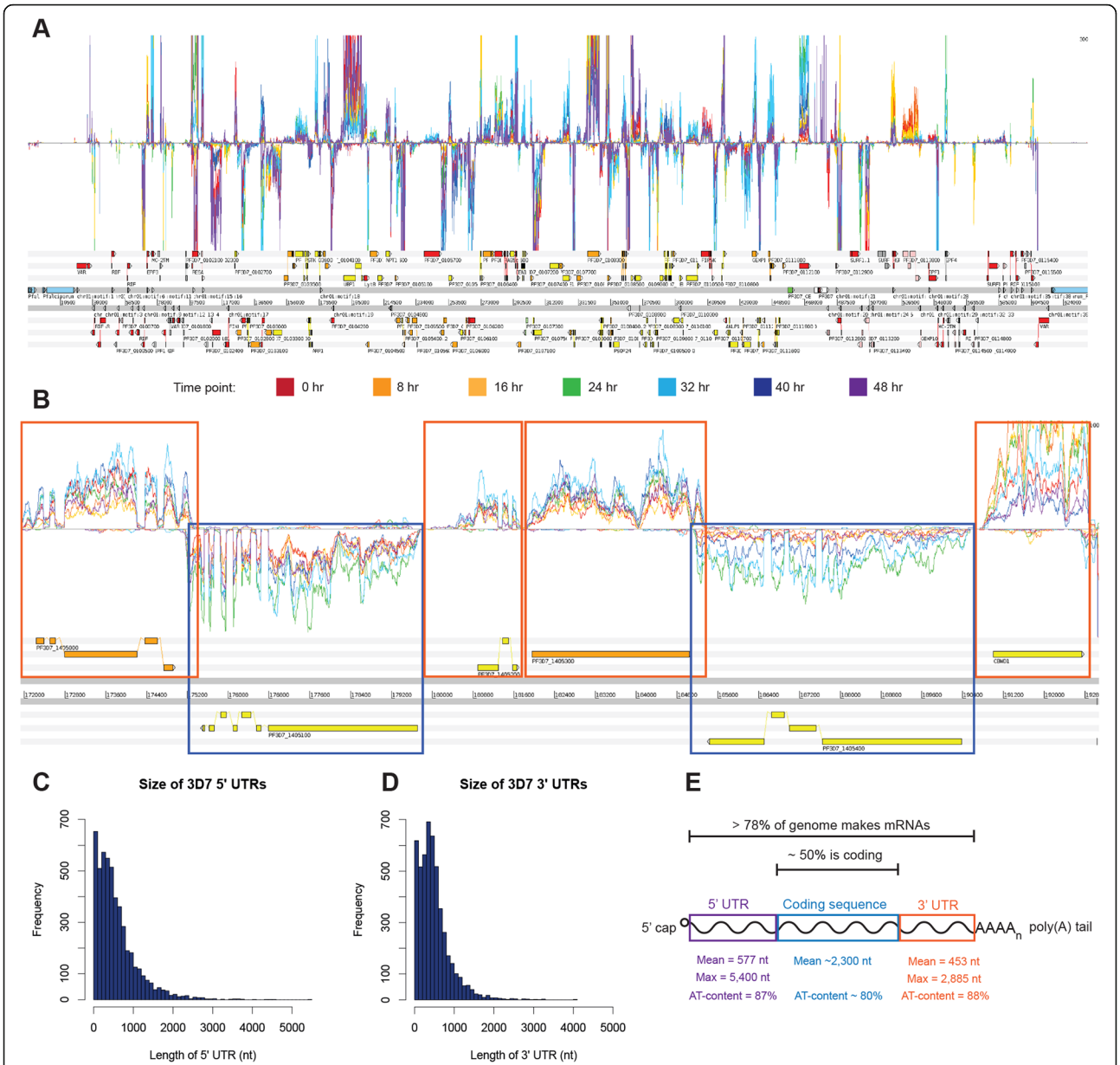

Fig. 1 Most of the *P. falciparum* genome is transcribed. **a** Overview of DAFT-seq data for the 3D7 time course for all of chromosome 1. Top panel: DAFT-seq coverage for plus strand. Coloured traces represent normalised coverage for each of the seven time points analysed. Middle panel: DAFT-seq coverage for minus strand. Lower panel: annotated gene models for Pf3D7v3 from GeneDB. Legend: colours of coverage traces from each of the seven time points. **b** Continuous coverage of DAFT-seq data allows transcript boundaries to be redefined. Orange boxes define boundaries of transcripts on the plus strand and blue boxes define boundaries of transcripts on the minus strand. Colours of coverage traces from the seven time points are the same as those shown above. **c** Size of 3D7 5′ UTRs based on continuous coverage of DAFT-seq data. See supplementary information for details of the computational method. **d** Size of 3D7 3′ UTRs based on continuous coverage of DAFT-seq data. See supplementary information for details of the computational method. **e** Summary statistics to describe the extent of the genome that is transcribed. At least 78% of the genome can be transcribed into mRNA

untranslated regions (UTRs) than previously realised (Fig. 1b,c).

We used the DAFT-seq data to define the positions of 5′ UTRs (Figure S3, Table S2) for 4982 genes in the 3D7 genome (94% of those detected as expressed at > 5 RPKM in the IDC, Table S3, method shown schematically in Figure S3). The precise boundary of each UTR likely varies slightly from transcript to transcript, but in order to annotate a single fixed position for each boundary, we estimated the true position by defining the position at which continuous RNA-seq coverage drops below a threshold of 5 reads. We used a more stringent threshold to avoid

merging adjacent UTRs on the same strand; the threshold used to define a block of continuous transcription was iteratively increased, in increments of 5 reads. This approach relies on continuous coverage along the length of a transcript, which is a feature and strength of the DAFT-seq protocol. These coverage-based UTRs generally represent the longest 5′ UTR used for the downstream parent gene, although there are some examples where an extremely AT-rich or unmappable sequence produces a break in coverage. Despite the average size of the predicted 5′ UTRs being 577 nucleotides (nt) (Fig. 1c), we identified several that were extremely long. For example, the start position of the longest 5′ UTR mapped to 5.4 kb upstream of the first protein-coding exon for polyA binding protein (*Pf3D7_1107300*). For some genes (6.5% of genes with detected 5′ UTRs), we found splice sites in their 5′ UTRs, with multiple splice sites detected in some instances.

We also predicted 3′ UTRs for 4356 genes (Table S4), again using the point at which continuous coverage of DAFT-seq data drops below a threshold of 5 reads (82% of genes detected at > 5 RPKM in the IDC, Table S3). The end of most 3′ UTRs (corresponding to polyadenylation sites) is between 500 and 1000 nt downstream of the end of the final protein-coding exons (Fig. 1d), with a mean length of 453 nt. The length of the longest measured 3′ UTR was 2885 nt downstream of the last protein-coding exon of the glycophorin binding protein (GBP) gene (*Pf3D7_1016300*).

We compared our set of longest observed 5′ UTRs to those annotated by Caro et al. (Figure S4) and those of Adjalley et al. (Figure S5). The Caro et al. 5′ UTR estimates were on average shorter than our predictions, with mean differences ranging from 7 nt (for time point 4 in the Caro et al. data set) to as much as 496 nt (for time point 5). In the Adjalley et al. study [30], TSSs were generated using thresholds of varying stringency. Using their most conservative threshold, the mean 5′ UTR length was 213 nt longer than our data set; this may be due to differences in thresholds used in analysis of slightly different data types, as their dataset also contains many TSSs per gene. We also compared our predicted coverage-based 3′ UTRs to those from a previous publication that defined 3′ UTRs by locating mapped reads with non-templated runs of adenines (polyA) [13]. Overall, the previous calls were slightly longer on average (523 nt), but covered fewer genes (3443 genes).

Based on our threshold read depth of > 5, we found that 88.5% of the 3D7 genome is transcribed during the in vitro IDC. This is higher than previously reported (78%) in a study with 30-fold more reads (~ 600 M reads) [13], emphasising the value of even coverage in DAFT-seq data, and the fact that evidence for transcription is not simply a function of overall sequencing output. In general, the regions with enhanced DAFT-seq sequence coverage relative to previous datasets were those with the highest AT-content (non-coding regions). For example, down-sampling the reads from the Siegel et al. dataset to match the smaller number of DAFT-seq reads in the present study reduced the coverage to 63% of the genome, highlighting that a greater proportion of the transcribed genome is accessible with the DAFT-seq protocol.

We identified continuous blocks of transcription (adjacent transcribed bases in the genome, each with > 5 mapped reads) in the DAFT-seq data that overlapped protein-coding genes on the same strand. The blocks of continuous transcription that overlap the boundaries of protein-coding genes cover ~ 78% of the genome (including introns), and 19% of the genome is transcribed as 5′ and 3′ UTRs of protein-coding transcripts, (summarised in Fig. 1e). We also find > 4% of the genome transcribed from both strands, which includes mRNAs on opposite, overlapping stran.

Properties of transcription start sites (TSSs) in the 3D7 strain

Although the continuous nature of DAFT-seq coverage enabled TSSs to be inferred genome-wide from RNA-seq data alone, we employed two additional strategies to better define the *P. falciparum* UTRs. First, we developed a modified RNA-seq protocol to capture the extreme 5′ ends of capped mRNA transcripts (5UTR-seq) and therefore maximise the signal for defining this region. Like DAFT-seq, 5UTR-seq used PCR-free adaptors to confer the same advantages of accessibility in AT-rich regions of the genome (Figure S6, S7). Template-switching oligos (TSOs) similar to those described in the Smart-seq2 protocol (Picelli et al. [31]; Picelli et al. [32]) were used to “tag” the 5′ end of mRNA sequences. We also used PacBio long read sequencing to determine the sequence of long, unfragmented cDNA molecules.

Precise TSSs were located for 3194 genes (Table S5) (67% of those expressed at > 5 RPKM in the IDC) using data from all time points. The 5′ UTRs determined based by 5UTR-seq had similar mean length (577 nt) to the 5′ UTRs predicted from DAFT-seq coverage (574 nt; Figure S8A). Individual TSSs predicted by the two methods were generally consistent, but there were exceptions. Where the UTR predicted from coverage appeared larger than the 5UTR-seq based prediction, we interpreted this to mean that the coverage-based UTR is the longest possible UTR, while the TSS data represented the most frequently used 5′ UTR. In contrast, a TSS identified by 5UTR-seq that is longer than a coverage-based one was hypothesised to indicate that an extremely AT-rich region had caused a short break in sequence coverage upstream of the gene. Indeed, we

observed an enrichment of long homopolymer tracts upstream of 5′ UTRs where coverage breaks occurred (Figure S9). Thus we attempted to "repair" gaps in the coverage-based UTRs with the 5UTR-seq predicted UTRs, allowing us to generate a combined "longest observed" UTR set (4499 5′ UTRs, Figure S8B,C, Table S6). These observations were also supported by PacBio reads containing TSOs, where a similar distribution of UTR sizes was observed (Figure S8D, Table S7). The longest 5′ UTR detected directly by PacBio was ~ 2600 nt (FT2, *Pf3D7_116500*, Figure S10A). We can directly detect the TSS linked with each transcript isoform using this data (Figure S10B,C).

The positions of the set of TSSs corresponding to the most complete set of 5′ UTRs (Table S6) correlated well with the genome-wide occupancy of previously characterized activating histone features H2A.z, H3K9ac, and H3K4Me3 [33], as well as the positions of the activating chromatin reader BDP1 [34]. The positions of the repressing marks HP1 [35], H3K9Me3, and H3K36Me2/3 [36] are negatively correlated with those of TSSs (Figure S11A). This result was robust to a comparison with a parallel analysis of random genomic locations (Figure S11B). A previous study found that *P. falciparum* TSSs were bordered by a small nucleosome-depleted region [15]. We also compared our TSS data to the positional nucleosome occupancy data from an additional study of *P. falciparum* chromatin [37]. This analysis revealed a depletion of nucleosomes around the TSSs at both 18 and 36 h post infection, and was also robust to a comparison with a parallel analysis of random genomic locations (Figure S12).

A detailed comparison of these multiple data sets for the same genes enables a more nuanced understanding of TSSs, including how they can vary both at a given time point or between time points; Fig. 2a shows an example of multiple TSSs for the gene encoding glyceraldehyde phosphate dehydrogenase (GAPDH, *Pf3D7_1462800*). We found that 90% of TSSs fell outside both annotated exons and introns, 9% fall within annotated exons, and 1% fall within annotated introns (Fig. 2b). This approach relied on prior knowledge of the position of start codons, and was limited to a maximum window of 2000 nt upstream of the start codon. Finally, to determine whether TSSs are constant or dynamic, we identified 422 genes (Table S8) with sufficient coverage depth (> 5 reads per TSS per time point) to independently call TSS-based 5′ UTRs at each of the seven time points from the 3D7 IDC. We found that 55% (232 genes) of these genes showed the same major TSS throughout the IDC, such as the gene encoding the 40S ribosomal protein S3 (*Pf3D7_1465900*). However, a number of genes showed distinct temporal changes of TSS usage throughout the IDC, including GAPDH (*Pf3D7_1462800*; Fig. 2a), where the distribution of TSS peaks shifts closer to the coding sequence (CDS) at the time point associated with peak mRNA levels. The converse trend is seen for the knob-associated histidine-rich protein gene (KAHRP, *Pf3D7_0202000*, Figure S13), revealing dynamic TSS use throughout the IDC.

Sequence features and genomic location of *P. falciparum* TSSs

The information in the 5UTR-seq dataset enabled us to quantify variation in position and time of TSSs associated with the same gene, but our initial analysis was limited by requiring prior knowledge of gene annotation. To address this, the CAGEr package [38] was used to cluster 5UTR-seq reads independently of gene annotation in two stages. First "tag clusters" (TCs) were formed from reads that mapped within 20 nt of each other, for each of the time points (Table S9). Depending on the time point, 89–94% of these TCs mapped outside annotated coding and intronic regions. Unlike the comparison between UTR positions for annotated genes, the genome-wide locations of the 5UTR-seq TSSs differ significantly with those in another recent study that tags the 5′ ends of mRNAs using a different approach [30]. Here, the authors reported that 49% of all "TSSs blocks" were downstream of the annotated start codons.

Next, nearby TCs (within 100 nt of each other, independent of time point) were grouped into "promoter clusters" (PCs), for each gene. At different time points, we found as many as 37–45% of genes had multiple annotated TCs. While most genes appeared to use a single TSS per time point, our data suggest that some use as many as 14 (Table S10 and Figure S14).

We calculated the nucleotide frequency in the regions surrounding the TSS-based 5′ UTRs. We found a global trend for enrichment of thymine residues upstream of TSSs, followed by an enrichment of adenine residues downstream of TSSs; this can be seen both locally (+/− 20 nt) and at greater distances (+/− 1000 nt; Fig. 2c). In addition, we find that transcription preferentially starts with a pyrimidine-purine dinucleotide, the most preferred being TG. These features are similar to those of highly expressed TSSs described for yeast, mouse, and human [39, 40], suggesting that this is a general feature of promoters that is conserved across a broad range of eukaryotes. We also found that deviations from the average base composition were localised to the site of the TSS itself and not beyond (Fig. 2c). This signature was also seen for TSSs predicted within introns and exons, albeit with a much weaker signal-to-noise ratio (Figure S15).

While most genes contain a primary TSS, we also identified 2157 genes with a strong distinct secondary TSS in the 5UTR-seq data set (Table S11, 68% of genes of the original 3194). This observation suggests that the

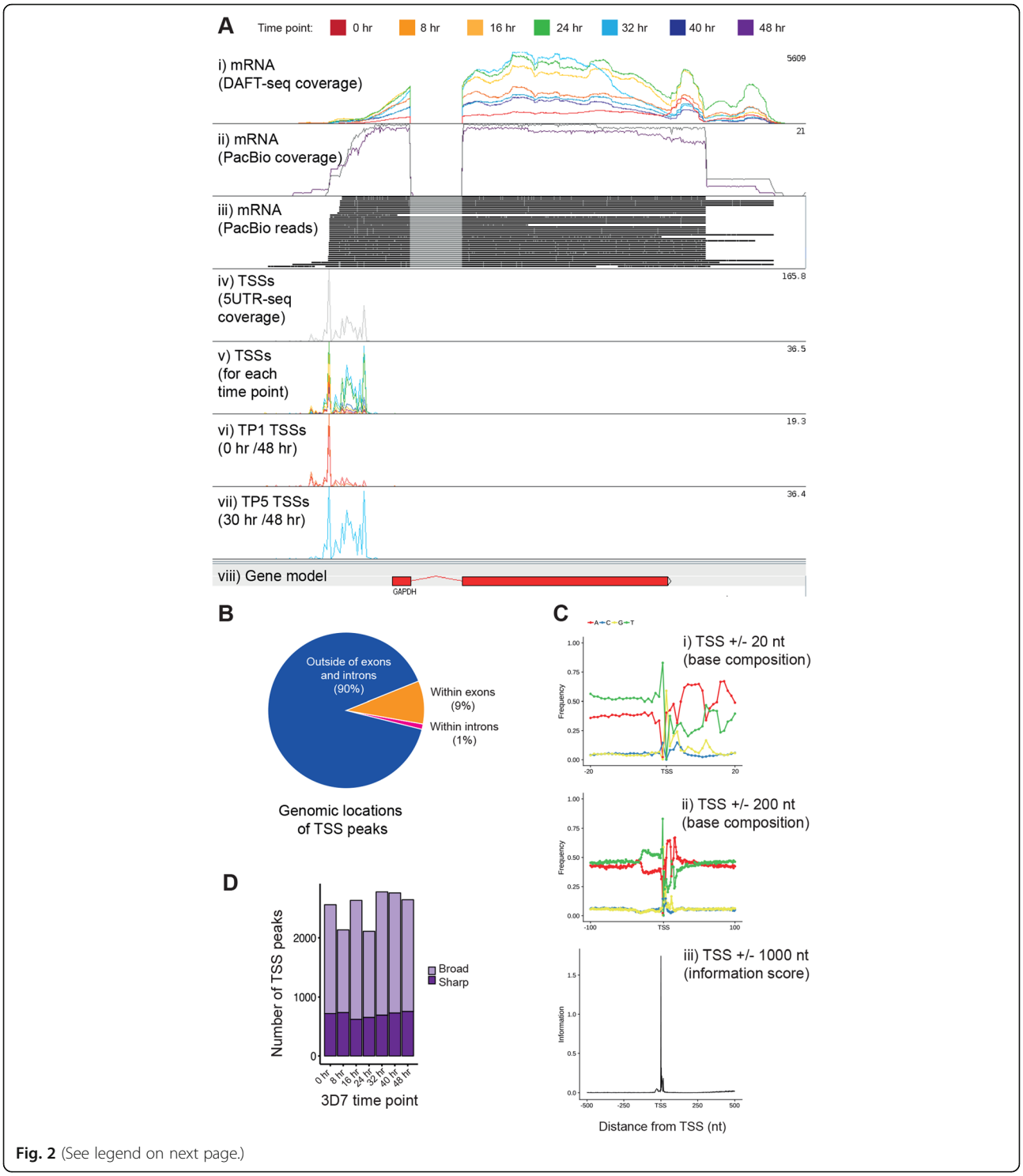

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Properties of transcription start sites (TSSs) and promoters. **a** Different library types show different properties of 5′ UTRs and TSSs for the gene encoding GAPDH (*Pf3D7_1462800*). DAFT-seq coverage (i) can be used to determine the longest possible 5′ UTR. Long read sequencing with PacBio (ii, iii) can be used to directly link a specific TSS with the rest of the transcript structure. Direct detection of TSSs with 5UTR-seq data (iv) reveals a range of different TSSs, which have different prevalences at different time points (v- vii). The first track (i) illustrates 7 DAFT-seq libraries, showing continuous coverage along the length of the gene, and variable steady state levels of mRNA throughout the time course. The next two tracks show PacBio coverage (ii) and reads (iii); these long reads can link variation in the TSS to the structure of the rest of the transcript. The fourth track (iv) shows the extreme 5′ end of mRNAs detected with all of the 5UTR-seq data. This data can be separated by time point (track v), with examination of individual time points showing that the most common TSS early in the time course (track vi) is further upstream from the coding sequence than the most common TSSs later in the time course (track vii). **b** Genomic locations of the TSS peaks identified using 5UTR-seq data. The vast majority of the TSS peaks in this data set (90%) fell outside of annotated exons and introns. A small proportion (9%) were within exons, while 1% were within introns. **c** Patterns in the base composition around TSSs were identified using the precise TSS positions inferred from the 5UTR-seq data. Windows are shown for a 20 nt distance (i) and a 100 nt distance (ii). Calculation of the information content of the base composition for a 1000 nt window shows that it peaks around the inferred TSS. **d** Number of TSS peaks in broad or sharp categories for each of the seven time points in the 3D7 time course

potential model of a single "sharp" TSS per gene is inadequate to explain the landscape of transcription initiation in *P. falciparum*. To analyse the proportion of TCs that could be categorised as "sharp" or "broad" we used TCs categorised by their interquartile width [38]. In other eukaryotes sharp promoters are often associated with initiation by RNA pol II protein complexes, whereas broad promoters are associated with activation by CpG islands in metazoans, and potentially other mechanisms linked to maintenance of an open chromatin state [39]. Analysing each time point separately, we found that many promoters (33–39%, depending on the time point) had a broad shape, establishing that *P. falciparum* transcription initiation is more variable than previously recognised (Table S9, Fig. 2d). Interestingly, the expression level from broad promoters was greater on average than that from sharp promoters, which is inconsistent with results found in other organisms and suggestive of a functional distinction between these two types of promoters (Figure S15). Despite these observations, sequence features were highly similar for both types of promoters (Figure S16). Additional studies will be required to address whether these differences are of functional relevance.

Transcription factor (TF) binding motifs relate to TSSs

Using our newly defined 5′ UTRs, we compared the relative location of all known DNA binding motifs associated with the Apicomplexan AP2 (ApiAP2) family of DNA binding proteins, which are considered the major sequence-specific transcriptional regulators present in *Plasmodium* parasites [41–43]. Previous studies predicted known ApiAP2 DNA-binding motifs within a defined distance (1–2 kb) upstream of start codons, correlating possible binding sites with the peak time of expression of the ApiAP2 proteins and their putative target genes [44–46]. We remapped all known ApiAP2 binding sequences to the 3D7 genome (Table S12) and selected motifs up to 1000 nt upstream of annotated start codons (Table S13) and the most frequently used TSSs (Table S14). Using the new TSSs significantly reduced the putative targeted binding motifs genome-wide by 33.8% and sequence search space by 38.3%. While not statistically significant, we found that motif occurrences were biased within 250 bp upstream of predicted TSSs; we speculate that this could be a feature of gene regulation within a compact genome (Figure S17). For genes with multiple predicted TSSs at different IDC timepoints, such as *kahrp* (Figure S18), independent motif searches were performed for each TSS. While overlapping and distinct motifs were observed for the long and short isoforms, future experiments will be required to functionally validate the differential role of these isoforms.

Expression of adjacent gene pairs

By improving the accuracy of UTR predictions, the extent to which the *Plasmodium* parasite uses bidirectional promoters became strikingly apparent (Fig. 3a). Bidirectional promoters have been described in multiple species, including human [47, 48] and yeast [49], where they regulate up to half of all protein-coding transcripts. In metazoans, a distance of less than 1000 bp between head-to-head genes has previously been used to define bidirectional promoters [47, 48]. In the *P. falciparum* 3D7 genome there are 1492 pairs of protein-coding genes in a "head-to-head" orientation (Table S15). Few gene pairs have overlapping 5′ UTRs (< 0 nt of sequence between TSSs), and when they do overlap, the overlap is small. The median distance between the 5′ UTRs for most pairs of head-to-head genes is 548 nt. In general, we observed positive correlations in the gene expression patterns of head-to-head gene pairs where the distance between the 5′ UTRs was less than 1000 nt (Fig. 3b). For example, the start codons for heme oxygenase (*Pf3D7_1011900*) and a putative RING zinc finger protein (*Pf3D7_1012000*) are 1499 bp apart, but their 5′ UTRs are separated by only 298 bp and their expression

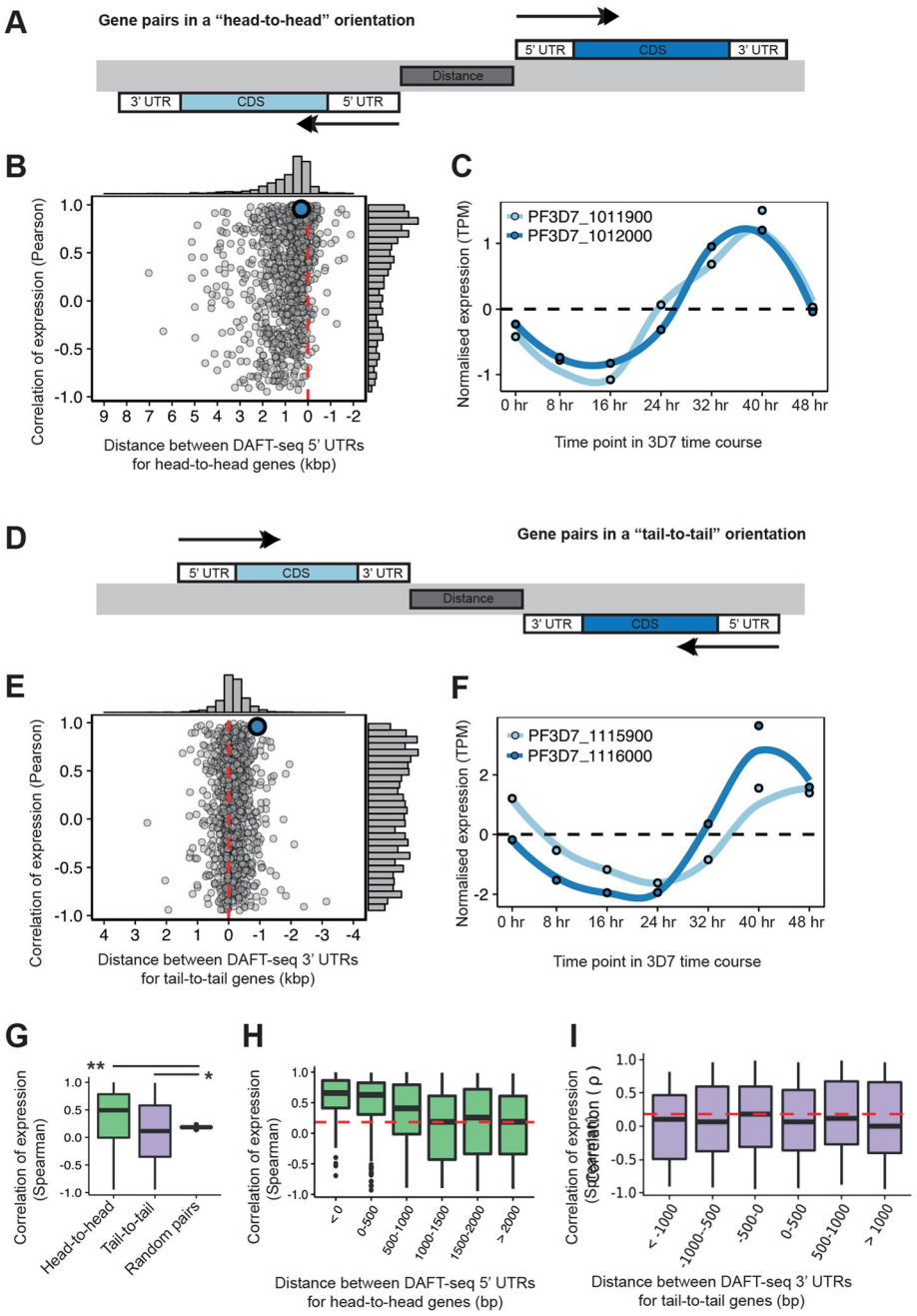

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Correlations between expression patterns of adjacent mRNA transcripts. **a** Schematic diagram of gene pairs in a "head-to-head" orientation (also known as divergent gene pairs). The black arrows represent the direction of transcription and the dark grey box between the genes represents the intervening genomic sequence that is between the longest detection version of both 5′ UTRs. **b** Correlation of gene expression (in TPM, using Pearson correlation) for 1119 pairs of head-to-head genes with annotated 5′ UTRs plotted against the distance of the intervening genomic sequence. The median intervening sequence length is 548 bp (without annotation of 5′ UTRs this distance was 1946 bp). The median correlation of expression was 0.49, with the distribution showing a positive skew. An individual region (blue) is shown in more detail in panel C. **c** Expression profiles through the 3D7 time course for a head-to-head gene pair (*Pf3D7_1011900*, heme oxygenase and *Pf3D7_1012000*, RING zinc finger protein, putative). The steady state levels of these two genes is tightly correlated, with an R value of 0.96 (measured by spearman correlation). **d** Schematic diagram of gene pairs in a "tail-to-tail" orientation (also known as convergent gene pairs). The black arrows represent the direction of transcription and the dark grey box between the genes represents the genomic sequence that is between the longest detection version of both 3′ UTRs. **e** Correlation of gene expression (in TPM, using Pearson) for 1059 pairs of tail-to-tail genes with annotated 5′ UTRs plotted against the distance of the intervening genomic sequence. The median intervening sequence length is − 124 bp, i.e. most 3′ UTR pairs overlap (without annotation of 3′ UTRs this distance is + 657 bp). The distribution of correlation values includes both strongly negative and strongly positive relationships. An individual region (blue) is shown in more detail in panel **f**. **f** Expression profiles through the 3D7 time course for a tail-to-tail gene pair (*Pf3D7_1115900*, DHHC9 and *Pf3D7_1116000*, RON4). Despite an overlap of 327 nt in the 3′ UTRs the steady state level of these genes is strongly correlated, with a spearman correlation of value 0.81. **g** Correlation of expression profiles (Spearman) of 1000 neighbouring gene pairs for head-to-head, tail-to-tail and randomly selected gene pairs. The 1000 random neighboring gene pairs were randomly sampled 1000 times from all annotated head-to-head, tail-to-tail and tandem gene pairs. Mean correlations were 0.35, 0.10, and 0.18 for head-to-head, tail-to-tail and random orientations, respectively. The Wilcoxon rank sum test was used to determine significance between groups. *P*-values of 2.2e-16 and 3.8e-3 when comparing head-to-head and tail-to-tail groups to random pairings, respectively. **h** Correlation of expression profiles (Spearman) neighbouring gene pairs for head-to-head gene pairs, binned by intervening genomic distance. **i** Correlation of expression profiles (Spearman) neighbouring gene pairs for tail-to-tail gene pairs, binned by intervening genomic distance

patterns are highly correlated (Fig. 3c). Therefore these two genes are likely to share a single bidirectional promoter. We used in silico predicted ApiAP2 TF DNA binding motifs, the genetic sequence for this newly defined promoter, and genome-wide gene expression profiles of the genes to predict that the likely ApiAP2 regulators of this promoter are encoded by the genes *Pf3D7_0730300* and *Pf3D7_1305200*. In total, we identified 401 pairs of genes that are potentially regulated by bidirectional promoters in *P. falciparum*, using an expression correlation cutoff of R > 0.5 and a distance between the 5′ UTRs of less than 1000 nt (Table S16).

The *P. falciparum* 3D7 genome also contains 1503 pairs of genes arranged in a "tail-to-tail" orientation (Table S17). We found that the 3′ UTRs overlap for many of these tail-to-tail gene pairs (Fig. 3d). In total, we identified 1036 genes (518 pairs of genes) where the 3′ UTRs overlap by at least 100 nt. Longer 3′ UTR overlaps were common; we detected 136 genes (68 pairs of genes) where the overlap was greater than 500 nt and 317 3′ UTRs which overlap the protein-coding sequence of the neighbouring gene. The largest 3′ UTR overlap was 1692 nt and is between genes encoding an acyl-CoA synthetase and a protein kinase (*Pf3D7_1238800* and *Pf3D7_1238900*, respectively). In an extreme example, 2432 nt of the protein-coding region of *Pf3D7_1244400* overlaps with the 3′ UTR of the adjacent gene (*Pf3D7_1244500).* We expected to find negative correlation for gene pairs with overlapping 3′ ends, due to the possibility of transcriptional interference, in which collision of convergently transcribing RNA polymerases leads to premature termination of transcription [50, 51]. However, in contrast to 5′ UTRs (Fig. 3c), we found little correlation between overlapping 3′ UTR distance and gene expression profiles of tail-to-tail gene pairs (Fig. 3e). We found gene pairs overlapping at the 3′ end showed strong patterns of correlation, including the example shown in Fig. 3f (*Pf3D7_1115900*, DHHC9 and *Pf3D7_1116000*, RON4). To rigorously determine if the different classes of adjacent gene pairs are coregulated, we further determined the patterns of correlations between 1000 pairs of head-to-head, tail-to-tail and randomly selected gene pairs (Fig. 3g). The mean correlations were 0.35, 0.10 and 0.18, respectively, consistent with an important role for bidirectional promoters in regulating head-to-head genes in *P. falciparum*. We further binned these correlations by distance for head-to-head (Fig. 3h) and tail-to-tail genes (Fig. 3i). Correlation of expression is inversely related to distance for head-to-head genes, which is consistent with the presence of bidirectional promoters present in the most closely adjacent 5′ UTRs. While the average correlation between tail-to-tail gene pairs was also positive, it was significantly less than randomly selected gene pairs suggesting that transcriptional interference may indeed be playing a role, albeit perhaps not a major one.

Non-coding transcripts associated with the 5′ end of mRNAs

We also identified a second form of correlated bidirectional promoters in the 3D7 genome between pairs of coding mRNAs and non-coding transcripts (ncRNAs). In several instances we could detect transcription upstream of the 5′ UTRs of mRNA transcripts on the strand opposite to that of the coding mRNA. The expression pattern of these ncRNAs are strongly correlated with the temporal expression of the adjacent coding

mRNA, suggesting that these transcript pairs are co-regulated (shown schematically in Fig. 4a). These non-coding transcripts were identified for genes where there is several kb of non-coding sequence upstream of the CDS (Fig. 4b), where there are multiple genes on the same strand (Figure S19A) and for genes in a head-to-head orientation (Figure S19B). Therefore, the most likely explanation for the observed coregulation is bidirectional transcriptional activity from the TSS that is associated with the mRNA, as has been reported in other eukaryotic species [52, 53]. To determine the genome-wide prevalence of this feature, we determined the correlation of transcript levels between all pairs of sense mRNAs and putative antisense ncRNA transcripts using a 2000 nt region upstream of the longest detected 5′ UTR (from our coverage-based 5′ UTRs). In total, we found 337 pairs of "transcriptionally linked" mRNA/ncRNA transcript pairs throughout the 3D7 genome (Table S18), which we provisionally describe as "TSS-associated RNAs" (TSSa-RNAs), similar to those previously described in mouse ES cells [53]. The functional role of these TSSa-RNAs remains to be determined.

Other studies [9, 13, 14] have described thousands of ncRNA fragments mapping to the 3D7 genome, although a recent study has called many of these into question (Böehme et al. [20]). Because the reference genome has not been annotated with these predictions in online databases, comparisons to these ncRNA are difficult. However, based on our new evidence we suggest that many, if not the majority, of these previously predicted ncRNAs can be considered "orphan fragments" of UTRs, while others represent TSS-aRNAs. However, we also find evidence of independent ncRNAs in our dataset, such as the ncRNA shown in Figure S20. After removing fragments of ncRNA which we suspect are part of annotated features, 5′ UTRs, 3′ UTRs, TSS-aRNAs, we find that 1.9% of the genome sequence has transcriptional activity consistent with currently unannotated ncRNA features (Table S19).

Novel splice sites included exitrons (exonic introns)

To identify additional functional features within our extended mRNA transcript models, we examined all spliced reads in the DAFT-seq data sets to annotate splice sites across the 3D7 transcriptome. Splice site predictions were further categorised based on overlap with features such as previously annotated introns or predicted UTRs (Fig. 5a, Table S20). We find that most introns are smaller than 200 nt (Fig. 5b). We also evaluated the prevalence of alternative splicing, and found that 6.9% of genes (365 genes) had alternative splice forms within their protein-coding regions (Table S21). Previous studies reported alternative splice forms in 1.5% of detected genes [8] and 4.5% of detected genes [11]. Figure S21 shows two alternative isoforms of the gene *Pf3D7_0316300* are captured in PacBio reads,

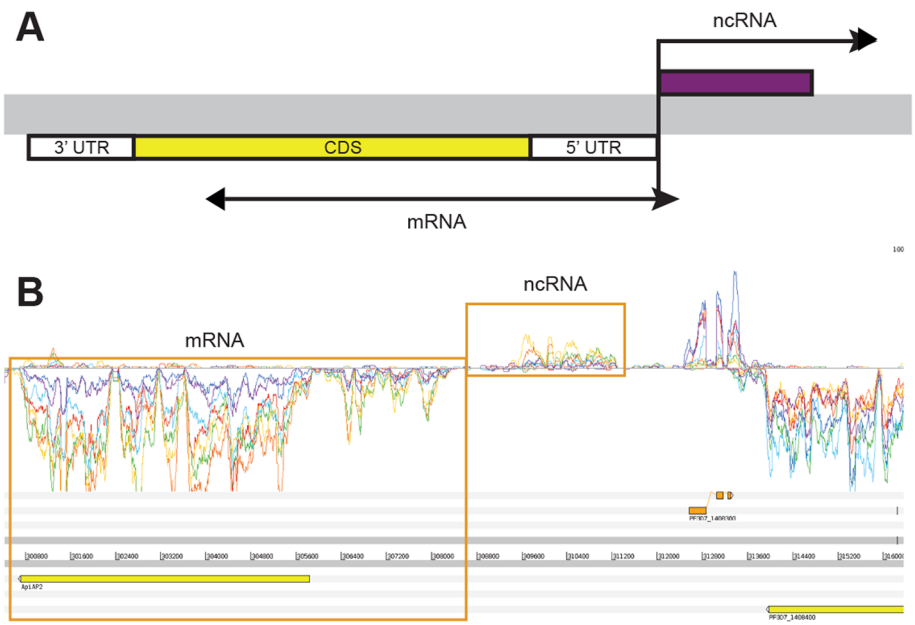

Fig. 4 Non-coding RNAs that may share promoters with nearby mRNAs. **a** Schematic showing the relative orientations of ncRNAs that may share a promoter sequence with an adjacent mRNA. Black arrows show the direction of transcription. The pair of transcripts may share a bidirectional promoter. **b** Example of a ncRNA and an mRNA (*Pf3D7_1408200*, which encodes an ApiAP2 protein) that may share a bidirectional promoter. The transcript pair is shown in the same orientation as the diagram in the top panel. The expression profiles of the two transcripts are correlated through the 3D7 time course

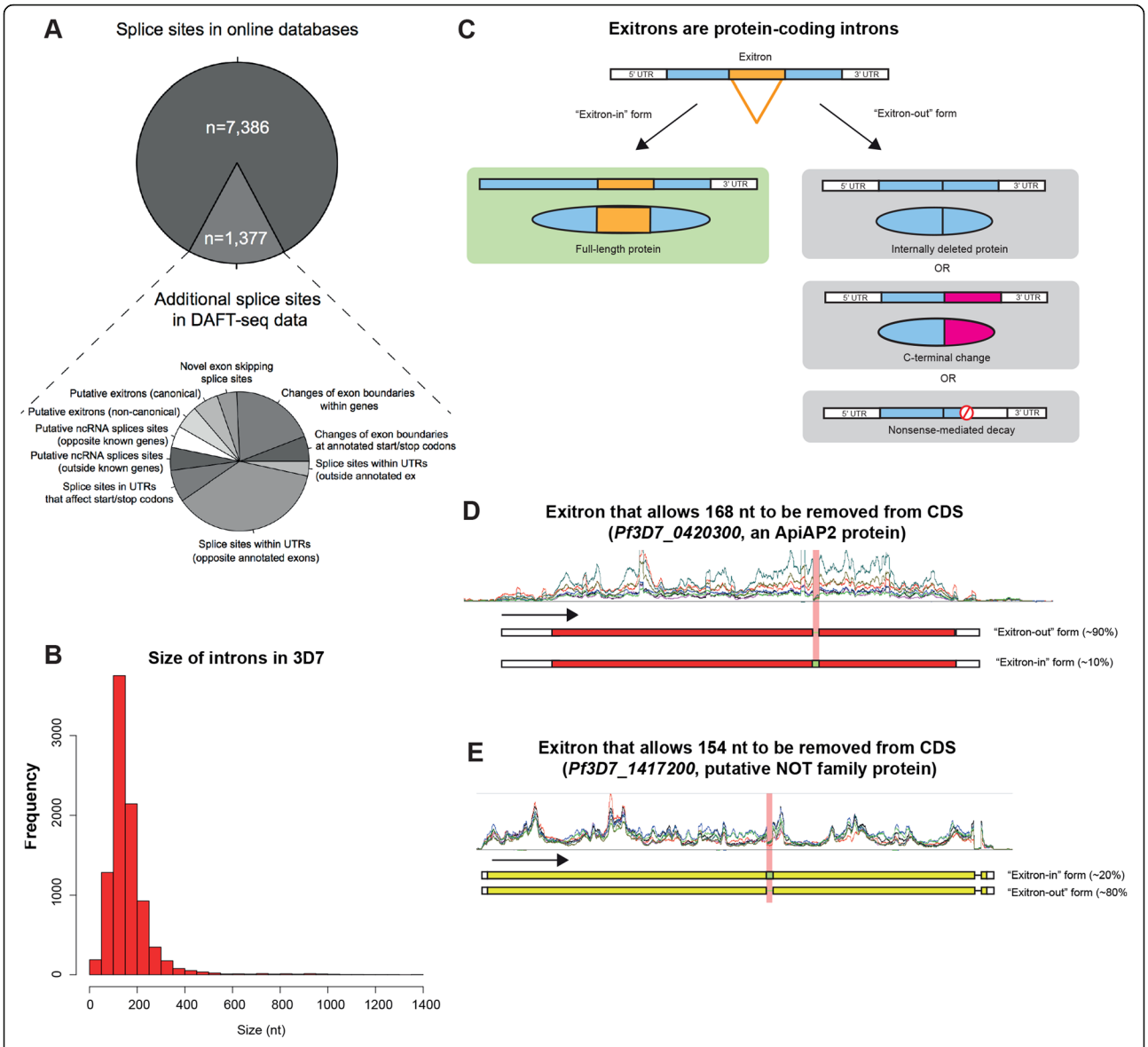

Fig. 5 Novel and updated features within *P. falciparum* genes. **a** Splice sites detected in the 3D7 time course using DAFT-seq data. A total of 8206 splice sites were supported by at least 5 reads; of these, 7386 were in online databases. There are 1377 splice sites not described in online databases (17% of the total), including 659 in UTRs. **b** Length of introns found in this study, using the DAFT-seq data for the 3D7 time course. **c** Schematic diagram of "exitrons" (protein-coding introns) modified from Marquez et al. [54]. An exitron may be retained in an mRNA, generating a fully coding mRNA sequence (left). If the excised sequence is a multiple of 3 nt the open reading frame is maintained and a full-length protein may be produced (top right). If the excised sequence is not a multiple of 3 nt this may generate a change to the C-terminus (middle left) or cause the transcript to be degraded through the nonsense-mediated decay pathway (bottom right). **d** Exitrons are present in the 3D7 transcriptome. There are examples of mRNAs where a multiple of 3 nt are excised, such as *Pf3D7_0420300* (which encodes an ApiAP2 protein). Here ~ 90% of the reads support the "exitron-out" form, with ~ 10% of the reads supporting the "exitron-in" form. **e** Exitrons are present in the 3D7 transcriptome. There are examples of mRNAs where even numbers of nt are excised, such as *Pf3D7_1417200* (which includes a putative NOT family protein). For this gene ~ 80% of the reads support the exitron-out form, with ~ 20% of the reads supporting the "exitron-in" form

where the reads capture different TSSs and different use of splice sites.

Exitrons (exonic introns) are intron-like features that can be found in protein-coding sequences whose retention or loss facilitates an additional level of regulatory control and proteome complexity (Fig. 5c). To date, exitrons have been reported only in *Arabidopsis thaliana* and human tissues [54]. Importantly, splicing of exitrons does not always retain the continuity of the protein open reading frame, resulting in protein diversification. We

detected 155 splicing events in annotated protein-coding sequences that are consistent with exitrons (Table S22). For example, for *Pf3D7_0420300*, an ApiAP2 gene, 168 nt (a multiple of 3 that retains the open-reading frame) is spliced out in ~ 90% of the detected transcripts (Fig. 5d). However, ~ 10% of the transcripts retain the intron, which has an open reading frame that is fully coding. In another gene, *Pf3D7_1417200*, the CDS is interrupted (80%) by an intron that is protein-coding and 154 nt in length (Fig. 5e). Excision of this exitron disrupts the coding sequence of the downstream protein as the reading frame is not conserved. Other exitrons we found contained non-canonical splice sites and were located in genes encoding exported proteins that may interact with the host, such as glycophorin binding protein (*Pf3D7_1016300*). The predicted exitrons were also detected in the PacBio long read sequencing data (an example is shown in Figure S22).

Comparing DAFT-seq transcriptomes for multiple *P. falciparum* strains

To define whether the extensive transcription found in 3D7 is unique to this strain, or reflects general features of *P. falciparum* biology, we measured the DAFT-seq IDC transcriptomes of the HB3 and IT strains and compared them to 3D7 by mapping the reads to the 3D7 reference genome. DAFT-seq expression data from all three strains was highly correlated to existing high-resolution hourly 3D7 microarray data, as expected [7] (Figure S23). We determined whether gene expression could be detected above a threshold of 5 TPM (tags per million mapped reads, a normalised measure of gene expression to best compare between samples) for each gene in HB3, IT and 3D7. Given that many genes within the subtelomeric regions are highly polymorphic, these regions were excluded to avoid mapping issues. This resulted in a common set of 4371 "core" genes (Fig. 6a) that were expressed by all strains (Table S23).

From the DAFT-seq data we calculated differential abundance of mRNA between the three strains (Table S24, Fig. 6b). Analysis of the functions associated with differentially expressed genes using GO terms showed that genes involved in the regulation of transcription show higher expression in 3D7 than HB3. Indeed, 15 out of 29 annotated ApiAP2 transcription factors were found to be detected at higher levels in 3D7 than in HB3 (Figure S24). In addition, we only detect transcription of AP2-G, (*Pf3D7_1222600*), the key regulator of gametocytogenesis [55, 56], in 3D7, suggesting that the HB3 and IT strains used in this study cannot produce gametocytes (Figure S25). We also considered the patterns of expression of pseudogenes in the 3 strains. There were 152 pseudogenes in our analysis, and 122 (80.2%) genes are transcribed in at least one of the three strains (TPM of at least 5 at one time point) and 62 (50.8%) were transcribed in all three strains (see also supplemental text).

An example of a gene with similar expression in all 3 strains (FIKK3, *Pf3D7_0301200*) is shown in Fig. 6c, while Fig. 6d shows differing expression profiles for ACS9 (putative acetyl-CoA synthetase, *Pf3D7_062780*), which is a member of an expanded gene family in *P. falciparum* with variable gene expression depending on the cell line. As reported previously [7], the timing of gene expression was highly correlated between *P. falciparum* strains (Fig. 6e-g), with few genes showing large differences in expression timing (Fig. 6h-j). HB3 and IT showed the strongest correlation in timing of gene expression, whereas 3D7 and IT were more correlated in amplitude (Figure S26, Table S25).

Finally, we identified UTRs and spliced transcripts for all three strains. Our UTR-detection pipeline identified fewer UTRs (~ 3800) for HB3 and IT than for 3D7 (Tables S26, S27; Figure S27), which is likely to be due to the higher quality and completeness of the underlying 3D7 reference genome sequence. Where gene expression was detectable in more than one strain, the length of the UTR calls were largely comparable (R ~ 0.8). There were slightly more splice sites observed in HB3 (9,231) than in the other two strains (8688 in 3D7, 7627 in IT; Tables S28 and S29), and we also detected more exitrons in HB3 (248) and IT (197) (Tables S30 and S31) than in 3D7 (155). It is not yet clear why there is this variation between the strains. We also observed that the relationship found for 3D7 between expression of head-to-head and tail-to-tail genes was similar across all three strains (Figure S28, Tables S32, S33, S34 and S35).

Discussion

The combined use of DAFT-seq, 5UTR-seq and PacBio cDNA protocols enabled the *P. falciparum* transcriptome to be systematically re-evaluated without the GC-content biases induced by PCR amplification and random priming. These new data revealed that 89% of the 3D7 genome is transcribed throughout the IDC. The 5′ and 3′ UTRs we identified (averaging roughly 600 nt each) are large compared to the average length of a protein-coding sequence (around 2000 nt), suggesting that much of the genome is not intergenic "space", but is associated with some function or biochemical activity. A key feature of the DAFT-seq data is that the coverage is near continuous from the extreme 5′ end of a transcript through to the 3′ end, allowing linkage of features throughout a transcript. We suggest that this unambiguous linking of the 5′ and 3′ UTRs to the main gene body makes our set of predicted 5′ and 3′ UTRs highly useful to researchers wanting to identify mRNA boundaries. Although the transcriptomes

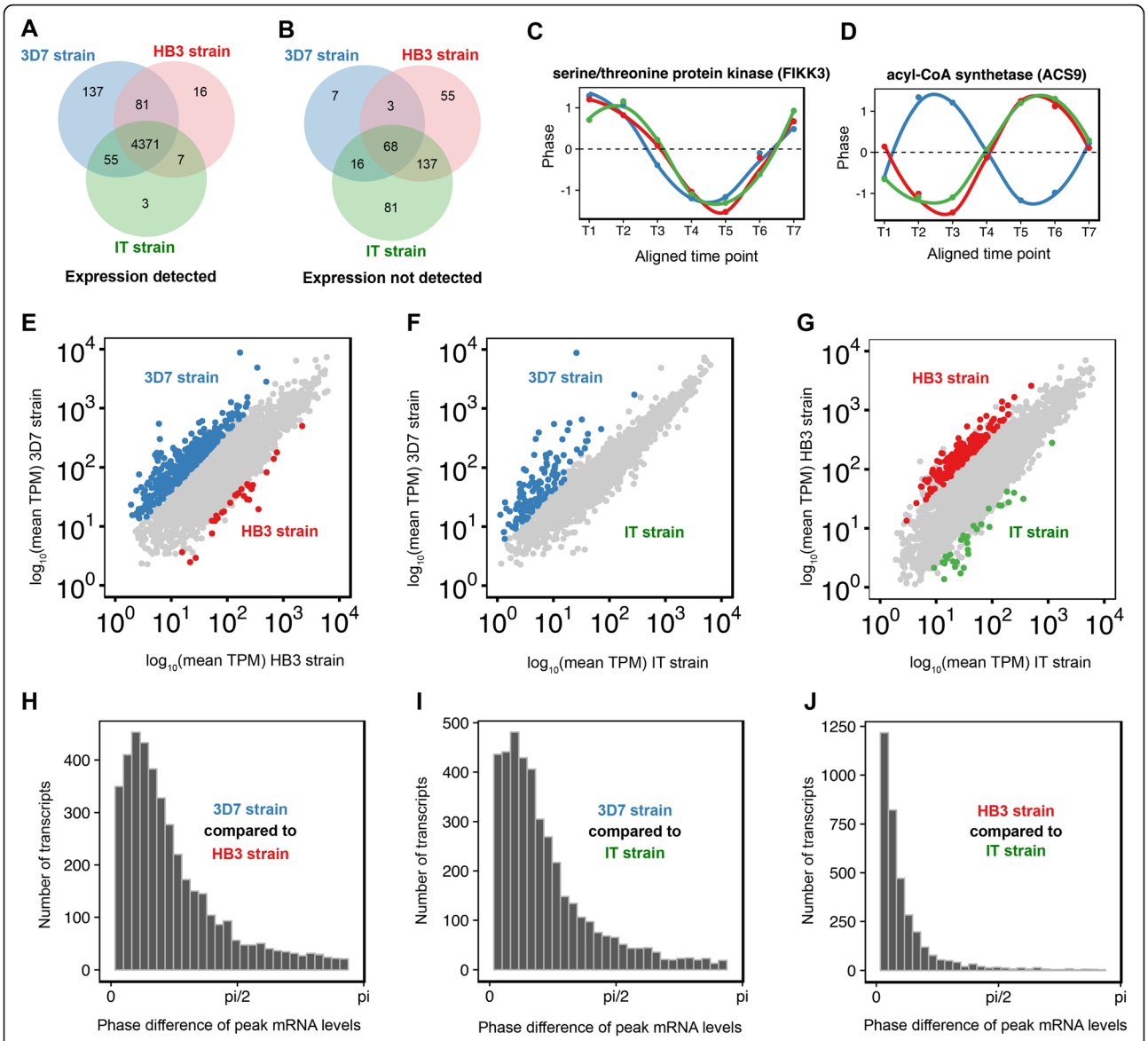

Fig. 6 Comparison of blood stage transcriptomes from the 3D7, HB3, and IT strains of *P. falciparum*. **a** Number of transcripts with expression detected in each of the three *P. falciparum* strains. Detection was based on a minimum expression of 5 TPM for at least one time point within the time course. Most transcripts (78%) were detected in all three strains. **b** Number of transcripts with no expression detected in each of the three *P. falciparum* strains. Detection was based on a minimum expression of 5 TPM for at least one time point within the time course. **c** Expression profile of the FIKK3 gene, which shows similar timing/phase of expression in each of the three strains. **d** Expression profile of the ACS9 gene, which shows a different timing/phase of expression in the 3D7 strain compared to both HB3 and IT. **e** Identification of genes that are differentially expressed between the 3D7 strain and the HB3 strain. Coloured data points highlight genes with at least a 2 fold difference in expression and FDR < = 0.05. **f** Identification of genes that are differentially expressed between the 3D7 strain and the IT strain. **g** Identification of genes that are differentially expressed between the HB3 strain and the IT strain. **h** Phase (timing) differences in genes between the 3D7 strain and the HB3 strain. The mean phase difference was 0.747 rad (5.7 h). **i** Phase differences between the 3D7 strain and the IT strain. The mean phase difference was 0.697 rad (5.3 h). **j** Phase differences between the HB3 strain and the IT strain. The mean phase difference was 0.304 rad (2.3 h)

reported in this work comprehensively represent *P. falciparum* asexual stages in vitro, it remains to be determined whether they represent all of the transcriptome of the parasites in the human host, although significant overlap is anticipated based on short-term ex vivo transcriptome studies [57–60]. Genes that were outside of the scope of our analysis include those associated with antigenic variation, such as the *var* gene family.

These newly defined transcription start sites will be especially useful to researchers wanting to generate knock-out or tagged parasites lines, and to researchers using single cell RNA-seq protocols that target either

the extreme 5′ or 3′ ends of mRNAs. In other eukaryotic systems both the sequence and structure of the 5′ and 3′ UTRs of transcripts are known to play key roles in gene regulation [61, 62]. The relative position and sequence of the UTRs in *Plasmodium* parasites is therefore likely to be constrained by functional requirements, in which particular nucleotides will be crucial for the formation of tertiary structures in the mRNA molecules recognised by RNA-binding proteins. Perhaps structural constraints imposed by tertiary structures of 3′ UTRs affect the amino acid usage in overlapping protein-coding sequences? We found relatively few differences between the 5′ and 3′ UTR sequences used by the three strains of *P. falciparum* investigated in this study, suggesting that UTR position or sequence motifs are highly conserved between different parasites.

The extensive and internally validated mapping of TSSs in these data allowed us to extend this knowledge and conclude that there is a "zone" of transcriptional initiation, within which there can be temporal variation in TSS position throughout the IDC. Combining these new data with additional genome-wide data sets such as nucleosome occupancy [15] and ATAC-seq [16, 63] may lead to further mechanistic insights into the control of transcription initiation in *P. falciparum*. Perhaps different regions of DNA are accessible or bound by different regulatory proteins during the IDC, or certain regions of the UTRs are important for the binding of post-transcriptional regulators that could stabilise or destabilise the transcript.

The genome-wide locations of 5UTR-seq TSSs we determined differ significantly with those in another recent study [30] which reported that 49% of all "TSSs blocks" were downstream of the annotated ATG start codons. In contrast, 90% of TSSs (analysed using all data and a 2000 nt upstream window) and 89–94% of the TCs (analysed per time point) we found mapped outside protein-coding sequences. We suggest that these differences may be due to methodological differences at two key points in the protocols used. Firstly, we selected polyA+ transcripts rather than those solely containing a 5′ cap. Secondly, Adjalley et al. use PCR amplification in their library production protocols; PCR amplification of AT-rich DNA usually enriches for GC-rich molecules and depletes AT-rich ones [64, 65]. This bias would be likely to enrich for TSS reads in relatively GC-rich exons (~ 80% AT) compared to the most AT-rich non-coding regions (~ 90% AT), where continuous UTR sequences were only reliably detected once we omitted the PCR amplification step from our protocol.

The strand-specific nature of the DAFT-seq protocol allows unequivocal confirmation of transcription from the non-coding strand of genes. However, the more precisely defined boundaries of transcriptional units suggests that many of the "antisense transcripts" that were observed in previous studies are either TSSa-RNAs or overlapping 3′ UTRs. There are at least two biological explanations for the presence of TSSa-RNAs. One explanation is that they are a normal byproduct of promoter activity, similar to that described previously for higher eukaryotes [52]. A second explanation is that these transcripts have some function in gene regulation, which we speculate could include binding by regulatory proteins containing an RNA recognition motif (RRM), many of which are encoded in the *P. falciparum* genome [66], but which are largely uncharacterised [67]. Additionally, our identification of independently regulated ncRNAs may help future studies of gene regulation, as seen with an ncRNA antisense to GDV1 [68] (which can be observed in our data, Figure S29), which is involved in the regulation of sexual commitment.

The putative exitrons detected suggest an additional mechanism of regulatory complexity, enabling the parasite to generate multiple protein isoforms from even single exon genes. This is of particular relevance for genes that have multiple binding partners or domains, such as transcription factors. Even for proteins with single interactors, inclusion or exclusion of key protein sequence may determine the activity or localisation of the protein.

Conclusions

Together, the complementary approaches applied in this study allow a refined description of the transcriptional landscape of *P. falciparum* and demonstrate that very little of the densely packed *P. falciparum* genome can be considered inactive or redundant. These precise definitions of 5′ and 3′ UTRs will be useful in guiding the definition of regions for amplification in experimental gene expression studies. In addition, the analysis identifies multiple regions where transcriptional units are very close or even overlap. This information will be of particular use for experimental genetics approaches aimed at deleting or altering specific genes, by highlighting regions where the insertion of gene modification constructs may have unintended functional consequences on adjacent genes. We anticipate that identifications of these mRNA transcript overlaps will motivate reanalysis of unexplained mutant phenotypes, where extended gene models can provide additional insights into the regulation of adjacent genes.

Methods

Parasite culture and RNA extraction

P. falciparum strains 3D7, HB3 and IT were cultured in O+ human erythrocytes and 10% human serum in RPMI-based media, using standard methods [69]. The 3D7 (MRA-102) and HB3 (MRA-155) parasite strains were obtained through BEI Resources (NIAID, NIH),

and the IT strain was sourced from Joe Smith (University of Washington, Seattle, USA). The IT strain was originally isolated from Brazil, but that due to cross contamination of strains is currently carrying a SE Asian genotype. Cross contamination was originally reported by Robson et al. [70] and was then verified years later by expanded SNP detection by Mu et al. [71] to be of SE Asian origin. Therefore, the IT strain we worked with is a strain of SE Asian origin. RNA extractions used the TRIzol reagent as previously described [72]. RNA was quality controlled and quantified using an Agilent Bioanalyzer 2100 Nano RNA chip.

Directional, amplification-free RNA-seq (DAFT-seq) libraries

To select for polyA+ RNA (mRNA), *P. falciparum*-derived total RNA was bound to magnetic oligo-d(T) beads and purified. Full-length mRNA was primed with oligo d(T) primers, and reverse transcribed using Superscript II (Life). Second strand synthesis used deoxynucleotides dATP, dUTP, dGTP and dCTP to encode directional information. The resulting cDNA was sheared using a Covaris AFA sonicator, and consecutive library preparation steps (dA-tailing, end repair and adapter ligation) were performed in the same well using a "with-bead" approach (reagents from NEB, equivalent to kit E6040). To avoid amplification bias (as described [19]), we used barcoded sequencing adaptors (Bioo Scientific), followed by 2 rounds of cleanup with AmpureXP beads (Beckmann Coulter) into EB buffer. To produce directional libraries, second strand cDNA was digested using USER enzyme mix (NEB). Prior to sequencing on an Illumina HiSeq2000 (100 bp paired-end), qPCR was used to quantify all libraries. The reads were mapped to version 3 of the 3D7 reference genome [20] using directional parameters in TopHat2 [73] and a maximum intron size of 5000 nt.

5UTR-seq libraries

PolyA+ RNA was isolated using oligo d(T)-coated magnetic beads. Superscript II reverse transcriptase was used to synthesise first strand cDNA using oligo-d(T) primers and in the presence of template switching oligos, which had the same sequence as those in the Smart-seq2 protocol [31, 32]. Template-switching oligos (TSOs) were used to "tag" the end of the cDNA sequences; this tag is used to prime second strand cDNA synthesis. The resulting cDNA was fragmented, made into RNA-seq libraries using an amplification-free protocol based on the DAFT-seq protocol, then were sequenced using 100 bp paired-end reads on an Illumina HiSeq2000. 5UTR-seq reads marking the TSS were identified by mapping using SMALT [74] and identified soft-clipped reads containing the TSO sequence. We removed G residues at the end of the TSO sequence that corresponded to the bases in the oligo that were involved in template-switching. We analysed reads in defined windows (up to 2500 nt) upstream of the annotated translation start site, and limited to finding the single most frequently used TSS at each time point. The threshold used for determining expression was at least 5 reads mapping to the same genomic position. Secondary TSSs were defined to be those with at least half the reads of the main TSS.

Transcription start site (TSS) clustering and analysis

5UTR-seq alignments were collapsed onto their 5′ ends to define genome-wide CAGE-defined TSSs (CTSSs) at each sampled time point. CTSSs were then analyzed using the CAGEr bioconductor package (version 1.12). CTSSs were first normalized by library size and tags-per-million (TPM) were calculated at each position. Positions across the genome where the TPM was greater than 1 were clustered based on distance (20 base-pairs apart or less) into TSS clusters (TCs). Following TSS clustering per time point, each TC was also clustered based on distance (100 base-pairs apart or less) into consensus promoter clusters (PCs). These were then filtered by those with an expression level of at least 5 TPM at one time point during the IDC. The resulting PCs were then clustered by their expression profiles using a self-organizing map and predefining the dimensions to 1 by 7. TCs and PCs were annotated using the BEDtools software suite and the BEDtools closest command to identify the nearest downstream gene for each cluster. Promoter shape was defined by the interquartile width (between 10th and 90th percentile) of each TC as in [75]. TCs with a width of < 10 bp were categorized as sharp, and the rest as broad.

Regulatory element analysis

Protein-binding microarray derived sequence-specific motifs for 24 ApiAP2 domains were obtained from [46] and trimmed to their 6 most informative, consecutive bases or core motifs (see supplementary text). The core motifs were used to search promoters extracted from version 3 of the *Plasmodium falciparum* reference genome using FIMO of the MEME suite and a threshold of 1×10^{-2} [76]. Promoters were defined as 1000 bp regions upstream of most frequently used, predicted TSSs.

Custom exploration of the extended transcriptome

RPKM values were calculated using in-house Perl scripts (Lia Chappell) that use the BEDtools suite [77]. Detection of new RNA sequences including UTRs and ncRNAs was achieved using a custom approach described further in the see supplementary materials, which also used the BEDTools suite to

manipulate blocks of continuous coverage and consider overlaps with other transcriptomic features.

Detection of splice sites (including) exitrons

A custom script (Lia Chappell) was used to identify spliced reads based on the CIGAR string in the DAFT-seq BAM files. Comparison with existing annotation allowed the detection of previously annotated splice sites. Splice sites in features such as UTRs and ncRNAs were detected by searching for spliced reads overlapping these features. Exitrons were identified by searching for spliced reads mapping within protein-coding exons. At least 5 reads were needed to support detection of each splice site.

Comparative strain analyses

Detection of expression was defined for a threshold of 5 TPM (tags per million mapped reads, a normalised measure of gene expression). Differential transcript abundance was calculated by first comparing mean Log2(TPM + 1) values across the IDC between strains using the Student's t-test for each gene. Multiple testing correction was performed using a false discovery rate, as implemented in the qvalue Bioconductor package (version 2.2.2) [78]. Differentially abundant genes were identified as having an absolute Log2 fold-change of at least 2 and a q-value less than or equal to 0.05. Correlations between genes were calculated using the Pearson correlation coefficient. Phases were calculated by implementing a non-parametric multidimensional scaling approach as demonstrated in [14] using the MASS R package available through CRAN (version 7.3–51.1).

Use of published genomics tools

The Artemis genome browser [79, 80] was used to visualise RNA-seq data. Enriched GO terms were identified using the TopGO tool [81]. A q-value cutoff of 0.05 was applied.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06787-5.

Additional file 1: Supplementary Figures. Figure S1. DAFT-seq schematic. **Figure S2.** PolyA+ vs Random hexamer priming of mRNA in the DAFT-seq protocol. **Figure S3.** Method for calling UTRs from DAFT-seq coverage data. **Figure S4.** Comparison of Chappell et al. 5′ UTRs with the Caro et al. study. **Figure S5.** Comparison of TSSs with the Adjalley et al. study. **Figure S6.** 5UTR-seq schematic. **Figure S7.** Method for calling 5′ UTRs with 5UTR-seq data. **Figure S8.** Alternative 5′ UTR sets. **Figure S9.** Homopolymer tracts. **Figure S10.** PacBio reads show the same features of the transcriptome as DAFT-seq data. **Figure S11.** Occupancy of covalent histone marks, histone variants, heterochromatin protein 1 (HP1) and the BDP1 chromatin reader at 3D7 TSSs. **Figure S12.** Nucleosome occupancy at 3D7 TSSs. **Figure S13.** Distinct temporal changes of TSS usage in the IDC (KAHRP). **Figure S14.** The TSS frequency and shape landscape for 3D7. **Figure S15.** TSS motifs in introns and exons. **Figure S16.** Sequence features of sharp and broad promoters. **Figure S17.** ApiAP2 motif occurrences relative to predicted TSSs and annotated translation start sites in the 3D7 strain. **Figure S18.** ApiAP2 motifs around the KAHRP gene. **Figure S19.** Exitrons in PacBio reads. **Figure S20.** Alternative splicing in PacBio reads. **Figure S21.** TSS associated non-coding RNAs. **Figure S22.** Long non-coding RNAs. **Figure S23.** Alignment of DAFT-seq data to microarray data. **Figure S24.** ApiAP2 expression in the 3 *P. falciparum* strains. **Figure S25.** AP2-G expression in the 3 *P. falciparum* strains. **Figure S26.** Amplitude change distributions of all genes in the 3 *P. falciparum* strains. **Figure S27.** Coverage-based UTRs for the 3 *P. falciparum* strains. **Figure S28.** Analysis of expression patterns of adjacent gene pairs for the 3 *P. falciparum* strains. **Figure S29.** A spliced non-coding RNA opposite GDV1.

Additional file 2.
Additional file 3.
Additional file 4.
Additional file 5.
Additional file 6.
Additional file 7.
Additional file 8.
Additional file 9.
Additional file 10.
Additional file 11.
Additional file 12.
Additional file 13.
Additional file 14.
Additional file 15.
Additional file 16.
Additional file 17.
Additional file 18.
Additional file 19.
Additional file 20.
Additional file 21.
Additional file 22.
Additional file 23.
Additional file 24.
Additional file 25.
Additional file 26.
Additional file 27.
Additional file 28.
Additional file 29.
Additional file 30.
Additional file 31.
Additional file 32.
Additional file 33.
Additional file 34.
Additional file 35.
Additional file 36.

Abbreviations

DAFT-seq: Directional, amplification-free RNA-seq; UTR: Untranslated region; IDC: Intraerythrocytic developmental cycle; ncRNA: Non-coding RNA; 5UTR-seq: Five-prime untranslated region tagged RNA-seq; TSO: Template-switching oligo; TSS: Transcription start site; CTSS: CAGE-detected transcription start site; TC: Tag cluster; PC: Promoter cluster; CAGE: Capped analysis of gene expression; TSSa-RNA: Transcription start site-associated RNA; TPM: Transcripts per million (mapped reads); RPKM: Reads per kilobase per million (mapped reads); ApiAP2: Apicomplexa Apetela 2 transcription factors

Acknowledgements
We would like to thank Joe Smith for providing us with the IT parasite line. We would like to acknowledge Chris Newbold for helpful discussions throughout this project, Ulrike Böhme for assistance with genome annotation, and Mandy Sanders for coordinating the sequencing. We also acknowledge the Sanger DNA Pipelines Bespoke Library team for quantifying and loading of sequencing libraries. We thank Adam Reid, Martin Hunt, Richard Bartfai and David Conway for helpful advice.

Authors' contributions
LC generated DAFT-seq, 5UTR-seq, and PacBio libraries, and designed and executed algorithms for UTR and TSS positions and for TSSa-RNA predictions, transcript quantification, isoform identification, and stranded splice-site classification. PR designed and executed algorithms for analysis of neighboring genes and bidirectional promoters, TSS clustering and annotation analyses, regulatory element analysis, and comparative strain analyses. TJR analysed the positions of TSSs compared with published genomics datasets. TO designed the algorithm for the identification of splice sites in mapped reads. LO cultured parasite lines and extracted and purified RNA. LC, PR, ML, MB and JR wrote the manuscript. JR, MB, ML, and TO designed the study. The author(s) read and approved the final manuscript.

Funding
M.L. received support from the Burroughs Wellcome Fund for Investigators in Pathogenesis of Infectious Disease (1007041.02), National Institutes of Health Grant (1DP2OD001315), and Center for Quantitative Biology Grant (P50 GM071508). L.C., T.D.O., M.B. and J.R. were supported by the Wellcome Trust through a core grant to the Wellcome Sanger Institute (206194).

Availability of data and materials
The DNA sequencing data resulting from this work are available in the European Nucleotide Archive (ENA) repository and are publicly accessible under study accession number: ERP001570. The processed data for UTRs is available online at GeneDB (ftp://ftp.sanger.ac.uk/pub/genedb/releases/latest/Pfalciparum/) and will also be made available through the *Plasmodium* genome resource PlasmoDB.org [82]. The resulting datasets supporting the conclusions of this article are included within the article and its supplementary files. The code used for custom computational analysis is available through *http://github.com/LiaChappell/DAFT-seq* and *https://llinaslab.github.io/sanger_rnaseq*. We have included key annotation files in the same directories as the relevant code.

Ethics approval and consent to participate
Not applicable. The three strains of *P. falciparum* described in this study are from established cell lines collected more than thirty years ago.

Consent for publication
Not Applicable.

Competing interests
No authors declare any competing interests.

Author details
[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK. [2]Department of Biochemistry & Molecular Biology and Huck Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA. [3]Present Address: Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL 60637, USA. [4]Present Address: Institute of Infection, Immunity and Inflammation, MVLS, University of Glasgow, Glasgow G12 8TA, UK. [5]Present Address: Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [6]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA.

Received: 9 December 2019 Accepted: 19 May 2020
Published online: 08 June 2020

References

1. World Health Organization. World Malaria Report 2018. Geneva: World Health Organization; 2018.
2. Cowman AF, Tonkin CJ, Tham W-H, Duraisingh MT. The molecular basis of erythrocyte invasion by malaria parasites. Cell Host Microbe. 2017;22:232–45.
3. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:498–511.
4. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 2003;1:E5.
5. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 2003;301:1503–8.
6. Painter HJ, Chung NC, Sebastian A, Albert I, Storey JD, Llinás M. Genome-wide real-time in vivo transcriptional dynamics during *Plasmodium falciparum* blood-stage development. Nat Commun. 2018;9:2656.
7. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. Nucleic Acids Res. 2006;34:1166–73.
8. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Böhme U, et al. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. Mol Microbiol. 2010;76:12–24.
9. López-Barragán MJ, Lemieux J, Quiñones M, Williamson KC, Molina-Cruz A, Cui K, et al. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. BMC Genomics. 2011;12:587.
10. Vignali M, Armour CD, Chen J, Morrison R, Castle JC, Biery MC, et al. NSR-seq transcriptional profiling enables identification of a gene signature of *Plasmodium falciparum* parasites infecting children. J Clin Invest. 2011;121:1119–29.
11. Sorber K, Dimon MT, DeRisi JL. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res. 2011;39:3820–35.
12. Hoeijmakers WAM, Salcedo-Amaya AM, Smits AH, Françoijs K-J, Treeck M, Gilberger T-W, et al. H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome. Mol Microbiol. 2013;87:1061–73.
13. Siegel TN, Hon C-C, Zhang Q, Lopez-Rubio J-J, Scheidig-Benatar C, Martins RM, et al. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC Genomics. 2014;15:150.
14. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genomics. 2015;16:454.
15. Kensche PR, Hoeijmakers WAM, Toenhake CG, Bras M, Chappell L, Berriman M, et al. The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res. 2016;44:2110–24.
16. Toenhake CG, Fraschka SA-K, Vijayabaskar MS, Westhead DR, van Heeringen SJ, Bártfai R. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying *Plasmodium falciparum* Blood-Stage Development. Cell Host Microbe. 2018;23:557–69.e9.
17. Chappell LVL. Novel approaches for transcriptome analysis in *Plasmodium* parasites. Ph.D: University of Cambridge; 2014. http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.648817.
18. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6:25533.
19. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009;6:291–5.
20. Böhme U, Otto TD, Sanders M, Newbold CI, Berriman M. Progression of the canonical reference malaria parasite genome from 2002–2019. Wellcome Open Res. 2019;4:58 https://doi.org/10.12688/wellcomeopenres.15194.2.
21. Walliker D, Quakyi I, Wellems T, McCutchan T, Szarfman A, London W, et al. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science. 1987;236:1661–6.
22. Bhasin VK, Trager W. Gametocyte-forming and non-gametocyte-forming clones of *Plasmodium falciparum*. Am J Trop Med Hyg. 1984;33:534–7.
23. Berendt AR, Simmons DL, Tansey J, Newbold CI, Marsh K. Intercellular adhesion molecule-1 is an endothelial cell adhesion receptor for *Plasmodium falciparum*. Nature. 1989;341:57–9.
24. Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, et al. Complete telomere-to-telomere de novo assembly of the *Plasmodium*

falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. DNA Res. 2016;23:339–51.
25. Dara A, Travassos MA, Adams M, Schaffer DeRoo S, Drábek EF, Agrawal S, et al. A new method for sequencing the hypervariable *Plasmodium falciparum* gene var2csa from clinical samples. Malar J. 2017;16:343.
26. Bryant JM, Baumgarten S, Lorthiois A, Scheidig-Benatar C, Claës A, Scherf A. Genome assembly of a NF54 clone using single-molecule real-time sequencing. Genome Announc. 2018;6 https://doi.org/10.1128/genomeA.01479-17.
27. Otto TD, Böhme U, Sanders M, Reid A, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed isolates reveal highly structured subtelomeres. Wellcome Open Res. 2018;3:52.
28. Benavente ED, Oresegun DR, de Sessions PF, Walker EM, Roper C, Dombrowski JG, et al. Global genetic diversity of var2csa in *Plasmodium falciparum* with implications for malaria in pregnancy and vaccine development. Sci Rep. 2018;8 https://doi.org/10.1038/s41598-018-33767-3.
29. Campino S, Marin-Menendez A, Kemp A, Cross N, Drought L, Otto TD, et al. A forward genetic screen reveals a primary role for *Plasmodium falciparum* reticulocyte binding protein homologue 2a and 2b in determining alternative erythrocyte invasion pathways. PLoS Pathog. 2018;14:e1007436 https://doi.org/10.1371/journal.ppat.1007436.
30. Adjalley SH, Chabbert CD, Klaus B, Pelechano V, Steinmetz LM. Landscape and dynamics of transcription initiation in the malaria parasite *Plasmodium falciparum*. Cell Rep. 2016;14:2463–75.
31. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
32. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using smart-seq2. Nat Protoc. 2014;9:171–81.
33. Bártfai R, Hoeijmakers WAM, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, et al. H2A.Z Demarcates Intergenic Regions of the *Plasmodium falciparum* Epigenome That Are Dynamically Marked by H3K9ac and H3K4me3. PLoS Pathog. 2010;6:e1001223 https://doi.org/10.1371/journal.ppat.1001223.
34. Josling GA, Petter M, Oehring SC, Gupta AP, Dietz O, Wilson DW, et al. A *Plasmodium falciparum* Bromodomain protein regulates invasion gene expression. Cell Host Microbe. 2015;17:741–51.
35. Fraschka SA, Filarsky M, Hoo R, Niederwieser I, Yam XY, Brancucci NMB, et al. Comparative Heterochromatin Profiling Reveals Conserved and Unique Epigenome Signatures Linked to Adaptation and Development of Malaria Parasites. Cell Host Microbe. 2018;23:407–20.e8.
36. Jiang L, Mu J, Zhang Q, Ni T, Srinivasan P, Rayavara K, et al. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. Nature. 2013;499:223–7.
37. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, et al. DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite *Plasmodium falciparum*. BMC Genomics. 2014;15:347.
38. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 2015;43:e51.
39. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38:626–35.
40. Lubliner S, Keren L, Segal E. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. Nucleic Acids Res. 2013;41:5569–81.
41. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. 2005;33:3994–4006.
42. Painter HJ, Campbell TL, Llinás M. The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. Mol Biochem Parasitol. 2011;176:1–7.
43. Jeninga M, Quinn J, Petter M. ApiAP2 transcription factors in Apicomplexan parasites. Pathogens. 2019;8:47 https://doi.org/10.3390/pathogens8020047.
44. Elemento O, Slonim N, Tavazoie S. A universal framework for regulatory element discovery across all genomes and data types. Mol Cell. 2007;28:337–50.
45. De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, Bulyk ML, et al. Specific DNA-binding by apicomplexan AP2 transcription factors. Proc Natl Acad Sci U S A. 2008;105:8393–8.
46. Campbell TL, De Silva EK, Olszewski KL, 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. 2010;6:e1001165.
47. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. Genome Res. 2004;14:62–6.
48. Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, et al. Complex loci in human and mouse genomes. PLoS Genet. 2006;2:e47.
49. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009;457:1033–7.
50. Prescott EM, Proudfoot NJ. Transcriptional collision between convergent genes in budding yeast. Proc Natl Acad Sci U S A. 2002;99:8796–801.
51. Wang L, Jiang N, Wang L, Fang O, Leach LJ, Hu X, et al. 3′ Untranslated regions mediate transcriptional interference between convergent genes both locally and ectopically in *Saccharomyces cerevisiae*. PLoS Genet. 2014;10:e1004021.
52. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.
53. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, et al. Divergent transcription from active promoters. Science. 2008;322:1849–51.
54. Marquez Y, Höpfler M, Ayatollahi Z, Barta A, Kalyna M. Genome Res. 2015;25(7):995–1007 https://doi.org/10.1101/gr.186585.114. Epub 2015 May 1. PMID: 25934563.
55. Kafsack BFC, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature. 2014;507:248–52 https://doi.org/10.1038/nature12920.
56. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, et al. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. Nature. 2014;507:253–7.
57. Almelli T, Nuel G, Bischoff E, Aubouy A, Elati M, Wang CW, et al. Differences in gene transcriptomic pattern of *Plasmodium falciparum* in children with cerebral malaria and asymptomatic carriers. PLoS One. 2014;9:e114401.
58. Lemieux JE, Gomez-Escobar N, Feller A, Carret C, Amambua-Ngwa A, Pinches R, et al. Statistical estimation of cell-cycle progression and lineage commitment in *Plasmodium falciparum* reveals a homogeneous pattern of transcription in ex vivo culture. Proc Natl Acad Sci U S A. 2009;106:7559–64.
59. Hoo R, Bruske E, Dimonte S, Zhu L, Mordmüller B, Sim BKL, et al. Transcriptome profiling reveals functional variation in *Plasmodium falciparum* parasites from controlled human malaria infection studies. EBioMedicine. 2019;48:442–52.
60. Tonkin-Hill GQ, Trianty L, Noviyanti R, Nguyen HHT, Sebayang BF, Lampah DA, et al. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen-encoding var genes. PLoS Biol. 2018;16:e2004328.
61. Leppek K, Das R, Barna M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018;19:158–74 https://doi.org/10.1038/nrm.2017.103.
62. Mayr C. Regulation by 3′-Untranslated regions. Annu Rev Genet. 2017;51:171–94 https://doi.org/10.1146/annurev-genet-120116-024704.
63. Ruiz JL, Tena JJ, Bancells C, Cortés A, Gómez-Skarmeta JL, Gómez-Díaz E. Characterization of the accessible genome in the human malaria parasite *Plasmodium falciparum*. Nucleic Acids Res. 2018;46:9414–31.
64. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, et al. Optimal enzymes for amplifying sequencing libraries. Nat Methods. 2012;9:10–1.
65. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics. 2012;13:1.
66. Reddy BPN, Shrestha S, Hart KJ, Liang X, Kemirembe K, Cui L, et al. A bioinformatic survey of RNA-binding proteins in *Plasmodium*. BMC Genomics. 2015;16:890.
67. Bunnik EM, Batugedara G, Saraf A, Prudhomme J, Florens L, Le Roch KG. The mRNA-bound proteome of the human malaria parasite *Plasmodium falciparum*. Genome Biol. 2016;17:147.

68. Filarsky M, Fraschka SA, Niederwieser I, Brancucci NMB, Carrington E, Carrió E, et al. GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. Science. 2018;359:1259–63.
69. Trager W, Jensen JB. Human malaria parasites in continuous culture. Science. 1976;193:673–5.
70. Robson KJH, Walliker D, Creasey A, McBride J, Beale G, Wilson RJM. Cross-contamination of *Plasmodium* cultures. Parasitol Today. 1992;8:38–9 https://doi.org/10.1016/0169-4758(92)90075-d.
71. Mu J, Awadalla P, Duan J, McGee KM, Joy DA, McVean GAT, et al. Recombination hotspots and population structure in *Plasmodium falciparum*. PLoS Biol. 2005;3:e335.
72. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987; 162:156–9.
73. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
74. SMALT. https://sourceforge.net/projects/smalt/. Accessed 29 Nov 2016.
75. Nepal C, Hadzhiev Y, Previti C, Haberle V, Li N, Takahashi H, et al. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res. 2013;23:1938–50.
76. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
77. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
78. Dabney A, Storey JD, Warnes GR. qvalue: Q-value estimation for false discovery rate control. R package version 1.26. 0; 2011.
79. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA. BamView: visualizing and interpretation of next-generation sequencing read alignments. Brief Bioinform. 2013;14:203–12.
80. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16:944–5.
81. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22:1600–7.
82. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37(Database issue):D539–43.

Publisher's Note