Identifying transcript 5′ capped ends in *Plasmodium falciparum*

Philip J. Shaw[1,*], Jittima Piriyapongsa[2,*], Pavita Kaewprommal[2], Chayaphat Wongsombat[1], Chadapohn Chaosrikul[2], Krirkwit Teeravajanadet[1], Manon Boonbangyang[2], Chairat Uthaipibull[1], Sumalee Kamchonwongpaisan[1] and Sissades Tongsima[2]

[1] National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand

[2] National Biobank of Thailand (NBT), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand

[*] These authors contributed equally to this work.

ABSTRACT

Background. The genome of the human malaria parasite *Plasmodium falciparum* is poorly annotated, in particular, the 5′ capped ends of its mRNA transcripts. New approaches are needed to fully catalog *P. falciparum* transcripts for understanding gene function and regulation in this organism.

Methods. We developed a transcriptomic method based on next-generation sequencing of complementary DNA (cDNA) enriched for full-length fragments using eIF4E, a 5′ cap-binding protein, and an unenriched control. DNA sequencing adapter was added after enrichment of full-length cDNA using two different ligation protocols. From the mapped sequence reads, enrichment scores were calculated for all transcribed nucleotides and used to calculate *P*-values of 5′ capped nucleotide enrichment. Sensitivity and accuracy were increased by combining *P*-values from replicate experiments. Data were obtained for *P. falciparum* ring, trophozoite and schizont stages of intra-erythrocytic development.

Results. 5′ capped nucleotide signals were mapped to 17,961 non-overlapping *P. falciparum* genomic intervals. Analysis of the dominant 5′ capped nucleotide in these genomic intervals revealed the presence of two groups with distinctive epigenetic features and sequence patterns. A total of 4,512 transcripts were annotated as 5′ capped based on the correspondence of 5′ end with 5′ capped nucleotide annotated from full-length cDNA data.

Discussion. The presence of two groups of 5′ capped nucleotides suggests that alternative mechanisms may exist for producing 5′ capped transcript ends in *P. falciparum*. The 5′ capped transcripts that are antisense, outside of, or partially overlapping coding regions may be important regulators of gene function in *P. falciparum*.

Subjects Bioinformatics, Genomics, Microbiology, Molecular Biology, Parasitology

Keywords 5′ capped nucleotide, Full-length cDNA, *Plasmodium falciparum*, Transcriptomics, Malaria, eIF4E

Submitted 29 January 2021
Accepted 26 July 2021
Published 25 August 2021

Corresponding authors
Philip J. Shaw, philip@biotec.or.th
Jittima Piriyapongsa, jittima.pir@nstda.or.th

Academic editor
Gunjan Arora

Additional Information and Declarations can be found on page 28

DOI **10.7717/peerj.11983**

How to cite this article Shaw PJ, Piriyapongsa J, Kaewprommal P, Wongsombat C, Chaosrikul C, Teeravajanadet K, Boonbangyang M, Uthaipibull C, Kamchonwongpaisan S, Tongsima S. 2021. Identifying transcript 5′ capped ends in *Plasmodium falciparum*. *PeerJ* **9**:e11983
http://doi.org/10.7717/peerj.11983

INTRODUCTION

Malaria remains the most widespread parasitic disease of humans, with over 200 million cases in 2018 (*World Health Organization, 2019*). The majority of severe malaria cases and deaths are attributed to *Plasmodium falciparum*, which is an obligate apicomplexan protist parasite of *Anopheles* mosquito and human hosts. Motile sporozoite stage parasites in the mosquito salivary gland enter the human host during blood feeding by the mosquito, which then migrate to the liver. Following a successful liver cell invasion, parasites multiply, develop and are then released into the bloodstream as merozoites capable of infecting mature erythrocytes. Infected erythrocytes are remodeled by the parasite as it develops through ring, trophozoite and schizont stages of the intraerythrocytic development cycle (IDC) before daughter parasites egress to invade new erythrocytes.

To facilitate discoveries of parasite molecular biology, such as finding new antimalarial drug and vaccine targets, the *P. falciparum* genome was sequenced in 2002 (*Gardner et al., 2002*). The finished reference genome comprises 5,280 protein-coding genes (*Böhme et al., 2019*). A total of 3,975 or more of the parasite's genes are expressed in the IDC (*Bozdech et al., 2003*; *Otto et al., 2010*). However, the annotation of these genes is incomplete, in particular the associated RNA transcripts. Previous attempts to clone full-length complementary DNA (cDNA) corresponding to transcripts encountered problems with sequencing highly A/T rich flanking untranslated regions, which hampered efforts in mapping the transcript ends (*Lu et al., 2007*). Accurate mapping of transcript 5′ ends is challenging because reverse transcription synthesis of cDNA is initiated from the 3′ end of the RNA. Premature termination of cDNA synthesis can occur because of RNA degradation, inhibitory RNA secondary structure and limited processivity of the reverse-transcriptase enzyme (*Das et al., 2001*). Hence, the transcript 5′ end inferred from cDNA sequence can be an experimental artifact of incomplete cDNA. The 5′ ends of gene transcripts in *P. falciparum* are modified by the addition of a canonical eukaryotic 5′ cap m^7G nucleotide (*Ho & Shuman, 2001*), which can be exploited for enriching 5′- capped mRNA (*Shaw et al., 2007*; *Shaw et al., 2016*).

One of the most widely used transcriptomic methods for enriching 5′-end fragments of eukaryotic mRNA is oligo capping, in which enzymatic treatments are used to append a synthetic oligonucleotide preferentially to mRNA 5′ ends that originally had a 5′ cap prior to reverse-transcription. The 5′ appended sequence serves as a primer for the synthesis of second strand cDNA so that the resulting cDNA is enriched for full-length fragments (*Maruyama & Sugano, 1994*). Full-length *P. falciparum* cDNA fragments were enriched by the oligo-capping method, cloned and sequenced in the Full-Malaria project (*Watanabe et al., 2002*). However, the limited coverage of *P. falciparum* genes prompted further study of transcript 5′ ends in *P. falciparum* using more comprehensive next-generation sequencing (NGS) approaches. Because NGS is limited to short read lengths, full-length cDNA fragments are generated by reverse-transcription proximal to the transcript 3′ end, typically by random priming. Throughout the rest of the manuscript, the term ''full-length cDNA'' refers to a cDNA fragment with sequence extending to the original transcript 5′ end, but not necessarily extending to the transcript 3′ end. Using an oligo-capping

method adapted for NGS referred to as 5′ cap sequencing, full-length cDNAs covering the majority of *P. falciparum* genes were detected (*Adjalley et al., 2016*). However, despite the extensive sampling and depth of coverage employed, transcript 5′ ends were not detected for some genes known to be expressed in the IDC (*Adjalley et al., 2016*). One major reason for the missing 5′ end data could be 5′ end bias, in which some full-length cDNAs are inefficiently enriched. The oligo-capping enrichment method used in *Adjalley et al. (2016)* employs RNA ligase 1 enzyme, which is markedly less efficient for some RNA substrates (*Fuchs et al., 2015*).

Alternative strategies for enriching full-length cDNA are available that can complement oligo-capping and provide more comprehensive information of transcript 5′ ends. Switching mechanism at the 5′ end of the RNA transcript (SMART) is another widely used method for enriching full-length cDNA (*Zhu et al., 2001*). SMART is based on the propensity of reverse transcriptase enzymes to preferentially add extra cytosine nucleotides to the 3′ end of full-length first-strand cDNA. A template-switching oligonucleotide with complementary guanosines allows the reverse transcriptase to extend the cDNA using the oligonucleotide as a template. Second-strand cDNA synthesis is primed from the extended cDNA and the resulting double-stranded cDNA can be converted to a DNA library for NGS. Transcriptome-wide surveys of full-length *P. falciparum* cDNA enriched by SMART have been reported (*Kensche et al., 2016*; *Chappell et al., 2020*). Although a much greater proportion of the genome could be annotated as transcribed compared with previous transcriptomic studies, transcript 5′ ends were reported for 2,278 (*Kensche et al., 2016*) and 3,194 (*Chappell et al., 2020*) genes, suggesting that characterization of transcript 5′ ends in *P. falciparum* is not comprehensive by this approach. SMART enrichment employed in these studies studies is 5′ end-biased towards substrates with a terminal G (*Wulf et al., 2019*), which is reflected by the high frequency of G observed among 5′ end nucleotides (*Chappell et al., 2020*).

Cap-Trapper is an alternative to oligo-capping and SMART for enriching full-length cDNA in eukaryotes (*Carninci et al., 1996*). In this strategy, the diol group of the RNA 5′ cap is oxidized and coupled to biotin in the mRNA/cDNA products of reverse transcription. Incomplete cDNAs are depleted by ribonuclease digestion and the biotin-modified 5′ cap is used as a purification handle to enrich for full-length cDNA. In the Cap Analysis of Gene Expression (CAGE) method (*Murata et al., 2014*), NGS libraries are made from full-length cDNA fragments enriched by Cap-Trapper. To construct NGS library, adapters are ligated to the enriched first-strand cDNA. The 5′ adapter sequence used in CAGE is biased to enrich full-length cDNA with a ''cap signature'' guanosine derived from reverse-transcription of the 5′ cap m^7G nucleotide (*Ohtake et al., 2004*). CAGE is technically demanding because the presence of the cap signature on full-length cDNA is strongly dependent on the reverse-transcription conditions (*Wulf et al., 2019*), and the efficiency of Cap-Trapper enrichment is strongly dependent on RNA purity and reaction conditions (*Weiss & Curran, 2015*). Although CAGE is the most accurate of current methods for mapping transcript 5′ ends in eukaryotes (*Adiconis et al., 2018*), it has not been widely used in non-metazoan organisms, including *P. falciparum*.

In this study, we developed a transcriptomic method for mapping of transcript 5′ capped ends. In our method, incomplete cDNAs are depleted enzymatically as in CAGE. Instead of chemical modification of 5′ capped ends, full-length cDNA is enriched using the eIF4E 5′ cap-binding protein, an enrichment strategy described as CAPture (*Edery et al., 1995*). We employ two different methods for adding NGS sequencing adapter to mitigate 5′ end bias. To increase signals of transcript 5′ ends, corresponding control transcriptomic libraries are prepared from cDNA samples before full-length enrichment. The degree of 5′ cap enrichment for each transcribed nucleotide is determined from the data using statistical models. To increase the power of detection, enrichment *P*-values are combined from replicate experiments (*Promworn et al., 2017*). Data were obtained with the new method from the ring, trophozoite and schizont stages of the *P. falciparum* IDC and used to annotate transcript 5′ ends. Published transcriptomic data from other studies were used together with the new data to annotate full-length 5′ capped transcripts.

MATERIALS & METHODS

Ethics statement

Human erythrocytes and serum were obtained from donors after providing informed written consent, following a protocol approved by the Ethics Committee, National Science and Technology Development Agency, Pathum Thani, Thailand, document no. 0021/2560.

Parasite culture and mRNA enrichment

Plasmodium falciparum strain K1 (NCBI Taxonomy ID: 5839) was cultured *in vitro* in human O+ erythrocytes and medium containing pooled human serum and RPMI 1640 as described previously (*Shaw et al., 2007*). Cultured parasites were synchronized to ring stage by Percoll gradient enrichment of mature stages followed by sorbitol treatment (*Lambros & Vanderberg, 1979*). Parasites were harvested immediately (ring stage), 12 h (trophozoite stage) and 24 h (schizont stage) after sorbitol treatment. Parasites were liberated from the host cell by treatment with 0.1% (w/v) saponin. Parasite total RNA was obtained using Trizol reagent according to the manufacturer's instructions (Invitrogen). Purified total RNA was stored in 75% ethanol at −80 °C before use. On the day of library preparation, total RNA was dried and then resuspended in nuclease-free water. Total RNA concentration was estimated by Nanodrop ND1000 measurement assuming A260 = 1.0 is equivalent to 40 μg/mL RNA. Up to 100 μg of total RNA was used for mRNA enrichment using a Dynabeads oligo dT25 mRNA kit (Thermo Scientific) or mRNA magnetic beads kit (New England Biolabs) followed by treatment with 1 unit of XRN-1 nuclease (New England Biolabs) for 30 min at 37 °C. XRN-1 enriched mRNA was purified by acidic phenol:chloroform extraction and ethanol precipitation. The mRNA was redissolved in 20 μL of nuclease-free water and genomic DNA was removed using a Turbo DNA-free kit following the manufacturer's instructions (Ambion).

Synthesis of cDNA for HiSeq libraries

In vitro synthesized spike-in RNA SIRV-Set 3 (Iso Mix E0/ERCC, Lexogen) was 5′ capped with Vaccinia capping system (New England Biolabs) following the manufacturer's

instructions. A 100–200 ng sample of parasite mRNA from ring or trophozoite stage (estimated by Nanodrop measurement) was mixed with 1.5 ng of 5′ capped SIRV-Set 3 RNA. Approximately the same amounts of parasite mRNA from schizont stages were used for reverse-transcription without the addition of spike-in RNA. The mRNA samples were incubated at 94 °C for 3 min in 1x first-strand cDNA synthesis buffer supplied with SuperscriptIV reverse transcriptase enzyme (Invitrogen) to fragment the RNA. Fragmented mRNA was chilled on ice for 5 min before reverse-transcription with 50 pmol of 5′ tagged random primer (RTNGS1 (Table S1) for adapter ligation method1, RTCIRC (Table S1) for adapter ligation method2; see below) with SuperscriptIV as described by the manufacturer in a final volume of 20 μL. After the addition of SuperscriptIV enzyme, reactions were incubated at 25 °C for 5 min, followed by heating to 37 °C for 30 min, 50 °C for 30 min, and then 70 °C for 15 min. First-strand cDNA was diluted to 100 μL with 10 mM Tris 1 mM EDTA and digested with 1 μL of RNaseA/T1 mix (Thermo Scientific) for 10 min at 37 °C to remove incompletely reverse-transcribed RNA. The mRNA/cDNA hybrid was purified using an equal volume of Agencourt® AMPure® XP beads (Beckman Coulter) and recovered from beads in 100 μL nuclease-free water. A 20 μL sample of purified mRNA/cDNA was kept for processing as the unenriched control sample.

Synthesis of cDNA for MiSeq libraries

Samples of approximately 1 μg of mRNA (estimated by Nanodrop measurement) were incubated at 94 °C for 5 min in 1x first-strand cDNA synthesis buffer supplied with SuperscriptIII reverse transcriptase enzyme (Invitrogen) to fragment RNA. Fragmented mRNA was chilled on ice before reverse-transcription with 50 pmol of 5′ tagged random primer (RTNGS1 for method1, RTCIRC for method2) with SuperscriptIII as described by the manufacturer in a final volume of 20 μL. Method1 reverse-transcription reactions also contained 40 μM biotin-11-dUTP (Thermo Scientific) to label first-strand cDNA with biotin. After the addition of SuperscriptIII enzyme, reactions were incubated at 25 °C for 5 min, followed by heating to 37 °C for 30 min, and then 50 °C for 30 min. First-strand cDNA was diluted to 100 μL with 10 mM Tris 1 mM EDTA and digested with 1 μL of RNaseA/T1 mix (Thermo Scientific) for 10 min at 25 °C to remove incompletely reverse-transcribed RNA. The mRNA/cDNA hybrid was purified by proteinase K treatment, phenol:chloroform extraction and desalting on a Amicon-30 column (Ambion). The volume of desalted mRNA/cDNA hybrid was adjusted to 100 μL with nuclease-free water. A 20 μL sample of desalted mRNA/cDNA was kept for processing as the unenriched control sample

Enrichment of full -length cDNA

P. falciparum eIF4E protein fused to glutathione-S transferase (GST-*Pf* eIF4E) (*Shaw et al., 2007*) was used for enriching full-length cDNA. GST-*Pf* eIF4E was expressed as recombinant protein in *Escherichia coli* and purified by SP-sepharose cation exchange chromatography as described previously (*Shaw et al., 2007*). Purified protein was buffer-exchanged using desalting columns and stored in a solution of 15% glycerol/RNase-free phosphate-buffered saline (PBS, Ambion). 5′ cap-binding activity of GST-*Pf* eIF4E protein

was determined by m^7GTP pulldown assay as described previously (*Shaw et al., 2007*), except that γ-aminophenyl-m^7GTP (C10-spacer)-agarose (Jena Bioscience) was used as the affinity support. For each full-length cDNA enrichment, approximately 100 μg of SP-sepharose purified GST-*Pf* eIF4E protein was immobilized on 50 μL of glutathione magnetic beads (Thermo Scientific) in PBS. The glutathione magnetic beads were washed four times with 0.5 mL of PBS to remove unbound protein and then resuspended in 100 μL of PBS. The remainder of the purified mRNA/cDNA sample was incubated with bead-bound GST-*Pf* eIF4E protein for 20 min at 25 °C with agitation (750 rpm Thermomixer). The beads were then washed thrice with 0.5 mL PBS and the mRNA/cDNA eluted in 100 μL of 1% sodium dodecyl sulfate/0.2 M sodium chloride. The eluted mRNA/cDNA was purified with an equal volume of Agencourt® AMPure® XP beads (Beckman Coulter) and recovered from beads in 20 μL of nuclease-free water. RNA was removed from enriched and unenriched cDNA samples by alkaline hydrolysis (15 min treatment at 65 °C with 0.2 N NaOH). Alkaline-treated cDNA was neutralized with an equal volume of 1 M HEPES pH 7.4 solution. First-strand cDNA was purified with an equal volume of Agencourt® AMPure® XP beads (Beckman Coulter) and recovered from beads in 20 μL of nuclease-free water.

DNA adapter ligation to cDNA

In some experiments, sequencing adapter was added by cDNA tailing followed by double-stranded adapter ligation (hereafter referred to as method1). A ribo-G tail was added to the purified single-stranded cDNA primed with RTNGS1 (Table S1) using terminal transferase (TdT, New England Biolabs). The tail length is limited to four Gs (*Schmidt & Mueller, 1996*). The TdT tailing reactions contained 2 mM GTP, 0.25 mM $CoCl_2$, 1 unit of TdT enzyme in 1x TdT enzyme buffer and first-strand cDNA. TdT reactions were performed in 20 μL reaction volumes for 20 min at 37 °C. Ribo-tailed cDNA was purified with an equal volume of Agencourt® AMPure® XP beads (Beckman Coulter) and recovered from beads in 20 μL of nuclease-free water.

Double-stranded DNA adapter was made by combining 4 nmol each of NGS1 and NGS1COMP oligonucleotides (Table S1) in a volume of 100 μL with 10 mM NaCl and 10 mM Tris pH 8.0. The NGS1 oligonucleotide is 5′-phosphorylated to act as a donor in ligation and blocked with a three-carbon spacer at the 3′ end to prevent concatemerization of adapters. Annealing was accomplished by heating the mixture for 3 min at 80 °C followed by slow cooling to 25 °C (−0.1 °C/min). Double-stranded DNA adapter was ligated to first-strand cDNA using T4 RNA ligase 2 (RNL2, New England Biolabs). The ligation reactions contained 40 pmol of DNA adapter, 7.5% (w/v) PEG_{6000}, 1x RNL2 enzyme buffer, 2.5 units of RNL2 enzyme and first-strand cDNA in a volume of 20 μL. The ligation reactions were performed for 99 cycles of 37 °C for 30 s, 22 °C for 30 s. After the completion of the RNL2 adapter ligation reactions, cDNA for HiSeq sequencing was diluted to 100 μL with nuclease-free water and purified with an equal volume of Agencourt® AMPure® XP beads (Beckman Coulter). Purified cDNA was recovered in 50 μL of nuclease-free water. For adapter-ligated cDNA samples sequenced on the MiSeq platform, ligation products were purified using 25 μL Streptavidin M280 magnetic beads following the manufacturer's

recommendations (Invitrogen). Synthesis of second-strand cDNA was performed on the beads using T7 DNA polymerase following the manufacturer's recommendations (New England Biolabs). The beads were resuspended in 50 μL of T7 reaction mix (0.3 mM dNTPs, 2.5 units T7 DNA polymerase, 1x T7 DNA polymerase buffer). The reaction was performed for 15 min at 37 °C. The reaction was terminated by washing the beads in 0.4 mL of washing solution (40 mM Tris pH 8.0, 10 mM $MgCl_2$). The second-strand cDNA was eluted from the beads by heating for 3 min at 99 °C in a suspension of 50 μL 1x Sodium Saline Citrate (150 mM sodium chloride, 15 mM trisodium citrate pH 7.0).

An alternative adapter ligation strategy (hereafter referred to as method2) was also employed. Purified first-strand cDNA primed with RTCIRC in which RNA had been removed by alkaline treatment was circularized with 100 U of CircLigaseII enzyme, 2.5 mM $MnCl_2$, 1 M betaine, and 1x CircLigaseII buffer in a reaction volume of 20 μL as recommended by the manufacturer (Epicentre). The reaction was incubated at 60 °C for 1 h and the reaction terminated by heating to 80 °C for 10 min. The reaction was diluted to 100 μL with nuclease-free water and cDNA purified using half the volume of Agencourt® AMPure® XP beads (Beckman Coulter). Purified cDNA was recovered in 50 μL of nuclease-free water.

Library purification and Illumina sequencing

Adapter-ligated cDNA was used as a template for PCR amplification to make sequencing library. For samples sequenced using the HiSeq platform, PCRs contained adapter-ligated cDNA template (1 μL), 2.5 pmol each of PE1 and ScriptSeq primers (Table S1), 200 μM dNTPs, 0.625 U PrimeSTAR® GXL DNA polymerase (Takara) and 1x buffer supplied with the enzyme in a reaction volume of 25 μL. The PCR program used was 98 °C for 30 s, followed by 18–25 cycles of 98 °C for 10 s, 68 °C for 60 s and a final extension of 68 °C for 5 min. For samples sequenced using the MiSeq platform, PCRs contained adapter-ligated cDNA template (1 μL), 12.5 pmol each of PE1 and ScriptSeq primers, 200 μM dNTPs, 2 mM $MgCl_2$, 0.5 units of Platinum *Taq* enzyme (Invitrogen) and 1x Platinum *Taq* buffer in a reaction volume of 50 μL. The PCR program used was 94 °C for 90 s, followed by 18–25 cycles of 94 °C for 10 s, 60 ° C for 60 s. The optimal number of PCR cycles for each sample was determined empirically by visualization of products separated by 1.5% agarose gel electrophoresis from pilot reactions conducted with varying numbers of cycles. Amplified DNA fragments 400–600 bp in size were excised from the gel and purified using a MinElute gel extraction kit (Qiagen), with the modification that agarose was solubilized in QG buffer at room temperature to prevent DNA denaturation. DNA was quantified using a Qubit™ dsDNA HS Assay kit (Invitrogen). Libraries were pooled in equimolar ratios according to the ScriptSeq index primer recommendations (Illumina).

Pooled libraries were submitted to Novogene AIT (Singapore) for Illumina standard protocol sequencing on one lane of a HiSeq2000 flow-cell (Illumina), or to the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand for MiSeq sequencing. For MiSeq sequencing, pooled libraries were processed with 1% Illumina phiX174 control and loaded onto MiSeq v3 flow cells at 10 pM. 150 bp paired-end sequencing was performed using the standard sequencing primers as recommended by the manufacturer (Illumina). For

MiSeq sequencing of libraries prepared using adapter ligation method 1, the sequencing "recipe" for the MiSeq instrument was modified to perform the first four sequencing cycles as "dark", meaning no data are recorded for the purpose of identifying sequencing clusters. This modification was necessary since the homopolymer tail added by TdT during library construction provides insufficient diversity for generating a high density of clusters. Without this modification, there will be data loss in the cluster identification step (clusters passing filter quality control) of Illumina sequencing.

Analysis of transcriptomic data from 5′ capped nucleotide enrichment experiments

The 5′ end of read1 from cDNA adapter ligation method1 data obtained using the HiSeq platform was trimmed of homopolymer sequence added by TdT using Cutadapt 1.18 (*Martin, 2011*). Read trimming was not performed for method 1 data obtained using the MiSeq platform, as signals for bases 1–4 were not recorded for this platform. Adapter sequence in the 3′ end of reads was trimmed and low-quality reads were removed using fastp with default settings (*Chen et al., 2018*). Preprocessed paired read data were aligned to the combined *P. falciparum* 3D7 v3.2 (*Böhme et al., 2019*) / SIRVomeERCCome (Lexogen) reference genome using HISAT2 (*Kim et al., 2019*), in the spliced mode guided by the *P. falciparum* / SIRVomeERCCome genome annotation with maximum intron length limited to 5,000 bp and other settings as default. Alignment summary statistics are shown in Table S2. To assess biological reproducibility, gene expression was determined for annotated *P. falciparum* genes using unenriched .bam files from each experiment as input to StringTie (*Pertea et al., 2016*). Gene expression values reported as transcripts per million were $\log_{10}$ transformed and used for sample pairwise Pearson correlations using the ggcorrplot package in R (*Kassambra, 2019*).

For analysis of 5′ cap nucleotide enrichment, .bam alignment files were used as input for the ToNER program (*Promworn et al., 2017*) with default settings, *i.e.,* depth value calculated from the read start position, pseudo-read added to depth values at all transcribed nucleotides and no filtering of positions for statistical modeling. This program calculates enrichment scores from paired feature-enriched and unenriched transcriptomic data. The enrichment scores are then transformed and fitted to a normal distribution by the Box–Cox procedure, and statistics of enrichment for each nucleotide reported. To increase statistical power, combined *P*-values from independent enrichment experiments were calculated by Fisher's method. The enrichment signals at individual nucleotides were clustered into genomic intervals representing the 5′ capped ends of major transcripts using the CAGEr package (*Haberle et al., 2015*) as follows. Clustering of nucleotide signals in CAGEr requires integer counts, which are referred to in the CAGEr manual as tags. To generate tags for clustering, the reciprocal of enrichment *P*-value was rounded to the nearest integer. To militate against program abort owing to the constraint of maximum allowed integers in R, the maximum tag count was set as 1e7. Tags were clustered with the distclus function in CAGEr with no normalization and maximum distance between merged clusters of 20 nt. The dominant_ctss position in each cluster reported by CAGEr was annotated as a 5′ capped nucleotide. CAGEr clusters overlapping known 5′ ends in

SIRV spike-in RNAs (assigned as true positives) were identified with BEDTools (*Quinlan & Hall, 2010*). The sum of tags (tpm) in each CAGEr cluster mapping to SIRVomeERCCome reference was used for Receiver Operator Characteristic (ROC) analysis with the plotROC package (*Sachs, 2017*). The optimal tpm value for filtering CAGEr clusters was identified using the OptimalCutpoints R package (*López-Ratón et al., 2014*) with default settings. For quantitative analysis of External RNA Controls Consortium (ERCC) RNA spike-ins, the CAGEr clusters reported from tag input of combined *P*-values of enrichment from four experiments were used. An RNA 5′ end was called as detected if a CAGEr cluster overlapped its known location. The empirical cumulative distribution functions of detected and undetected 5′ ends with respect to natural log of ERCC RNA concentration reported by the manufacturer (Lexogen) were determined using the base R ecdf function. The distributions of detected and undetected 5′ end groups were compared by two-sample Kolmogorov–Smirnov test in R. The natural log tpm values for CAGEr clusters overlapping detected 5′ ends were plotted against the natural log concentration of each RNA using the ggplot2 package (*Wickham, 2009*) in R.

Analysis of genomic context in the vicinity of 5′ capped nucleotides

5′ capped nucleotide signals detected as CAGEr clusters above threshold (tpm >32) in separate stages of *P. falciparum* development (including cluster signals within 100 bp of each other) were merged into 17,961 non-overlapping genomic intervals using BEDTools (*Quinlan & Hall, 2010*). Within each interval, the dominant_ctss nucleotide with highest integer count (tpm.dominant_ctss reported by CAGEr) across the sampled developmental stages was annotated as the representative 5′ capped nucleotide for genomic analysis. Genomic sequences flanking 100 bp on either side of the 17,961 5′ capped nucleotide (test set) and 17,961 random positions selected using BEDTools (*Quinlan & Hall, 2010*) (control set) were obtained from the *P. falciparum* 3D7 v3.2 reference genome (*Böhme et al., 2019*) using the seqPattern package in R (*Haberle, 2020*). Sequence analysis was performed using the kpLogo program (*Wu & Bartel, 2017*) using the control set sequences as background and other settings as default. For analysis of epigenetic features, publicly available datasets from published studies were processed to construct genome coverage files (bigwig format) as described below in sub-section analysis of external epigenetic data. Plots of average coverage for each feature were made using the genomation package (*Akalin et al., 2015*) in R. To mitigate the effect of extreme values, the top and bottom 5% of scores were clipped using the winsorize function in the genomation package.

Mismatched base analysis of -1 positions

The first base of all uniquely aligned reads (with secondary alignments removed from alignment .bam files using SAMtools) was extracted from all datasets. For reads mapping to the reference (–) strand, the complement of the first base was taken. Bases mismatched to the reference sequence (corresponding to soft-clipped bases in the alignment files) were counted for two groups. The first group comprised reference positions one base upstream of dominant 5′ capped nucleotides in non-overlapping genomic intervals (17,961 positions), and the second group all other reference positions. Contingency tables were constructed for

each dataset for the count of the tested mismatched bases and all other mismatched bases for both groups of nucleotides (Table S3). One-tailed Fisher's exact tests were performed of the alternative hypothesis that the true odds ratio is greater than 1. One-tailed tests were performed because mismatched bases under-represented in the first position of reads are not of interest. Bonferroni-corrected *P*-values less than 0.001 were considered significant. For analysis of reference sequence patterns at -1 positions with high counts of mismatched reads, positions with 10 or more aligned reads and >50% of reads with mismatched first base in any dataset were selected.

Unsupervised cluster analysis of 5′ capped nucleotides

17,961 5′ capped nucleotides in non-overlapping genomic intervals annotated from our data (see above) were clustered using epigenetic data, including occupancies of H2A.Z, H3K9 acetylation and H3K4 trimethylation (*Bártfai et al., 2010*) and H2B.Z (*Hoeijmakers et al., 2013*). For details of how chromatin mark occupancies were determined, see below in sub-section analysis of external epigenetic data. Principal Component Analysis (PCA) scores were calculated using the MacroPCA package in R (*Hubert, Rousseeuw & VandenBossche, 2019*). The data were pre-processed by transformation with natural logarithms, scaling by unit variance, means-centering and removal of rows with too many missing values. The first three principal components explained 83.3% of the variance; hence, the PCA scores from the first three principal components were used for clustering. To determine if more than one cluster existed in the data, statistical tests of unimodality were performed on PC1 scores using the multimode package in R (*Ameijeiras-Alonso, Crujeiras & Rodríguez-Casal, 2019*). Test *P*-values less than 0.05 were considered significant. Clustering was performed using the cross-entropy clustering (CEC) package in R (*Tabor & Spurek, 2014*; *Spurek et al., 2017*). The algorithm employed in this program is a hybrid of k-means and Gaussian mixed model, and can thus separate clusters with a variety of shapes. The settings used for CEC clustering were: Gaussian distribution models unconstrained, nstart =1000, initial clusters =5 and card.min =20%. To determine independently how many relevant clusters are present, the PCA scores from PC1, PC2 and PC3 were analysed using the NbClust package in R (*Charrad et al., 2014*). Cluster indices were determined using the Euclidean distance matrix and k-means method for $n = 2$ to $n = 6$ clusters. All indices were calculated except Gplus and Tau, which could not be determined owing to computational constraint, and Gamma which is only applicable for hierarchical clustering.

Analysis of external transcriptomic data

To assess the reproducibility of 5′ capped nucleotides in *P. falciparum*, data were analyzed from independent published studies using different methods for 5′ capped nucleotide enrichment. *P. falciparum* 5′ cap sequencing data reported in *Adjalley et al. (2016)* were downloaded from the NCBI GEO database series under accession number GSE68982 using sratoolkit.2.10.0-ubuntu64 (*SRA Toolkit Development Team, 2014*). The read1 fastq files were pre-processed with UMI-tools (*Smith, Heger & Sudbery, 2017*) to move the 8 bp unique molecular index to the read header. Pre-processed paired data files were filtered and trimmed of 3′ adapter sequence using fastp (*Chen et al., 2018*) with

default settings. Data from *P. falciparum* SMART-enrichment experiments reported in *Kensche et al. (2016)*, European Nucleotide Archive accession numbers ERR861692 and ERR861693, were uploaded to the public server at usegalaxy.org. Read1 were filtered and trimmed using the Barcode Splitter tool to retain reads that started with smartseq2 adapter (AAGCAGTGGTATCAACGCAGAGTACATGGG), allowing up to three mismatches or indels. The filtered read1 and read2 from 5′ cap sequencing and SMART-enrichment experiments were processed using fastp (*Chen et al., 2018*) with default settings. Processed paired 5′ cap sequencing and SMART-enrichment data from fastp were aligned to the *P. falciparum* 3D7 v3.2 reference genome (*Böhme et al., 2019*) using HISAT2 (*Kim et al., 2019*) in the spliced mode guided by the genome annotation (PlasmoDB release 44) with maximum intron length limited to 5,000 bp and other settings as default. PCR duplicates in 5′ cap sequencing data were removed using UMI-tools (*Smith, Heger & Sudbery, 2017*) with the–unmapped-reads =use option. Unmapped and read2 reads were removed from .bam files and a merged .bam file was created from all experiments in each study using SAMtools (*Li et al., 2009*).

The filtered .bam files from individual experiments were used to obtain read depth at each position in the genome with BEDTools (*Quinlan & Hall, 2010*), with -strand and -5 options. The read depth values were used as ctss input for CAGEr (*Haberle et al., 2015*) for detection of 5′ capped nucleotide signals. CAGEr was run with the distclus function, no normalization and maximum distance between merged clusters of 20 nt. The CAGEr analysis results of 5UTR-seq experiments in *P. falciparum* were obtained from Table S9 reported in a previous study (*Chappell et al., 2020*). A merged .bam file of aligned data from all 5UTR-seq experiments was provided by Dr. Lia Chappell.

The dominant_ctss nucleotide genomic locations and the corresponding tpm.dominant _ctss values reported for CAGEr clusters in each dataset were concatenated and used to construct .bed and .bedgraph files of 5′ capped nucleotides for each study. In addition to CAGEr clusters, the read depth at each position in the genome was obtained from the merged .bam file from all experiments in each study using BEDTools (*Quinlan & Hall, 2010*) with -strand and -5 options. To assess the agreement of 5′ cap nucleotide assignment in our data with other studies, agreement was scored if the depth value from the combined .bam file of all experiments in the same study was two or greater.

Direct RNA sequencing data from *P. falciparum* obtained using the Oxford Nanopore MinION platform reported in *Lee et al. (2021)* under SRA accession number SRR11094274 were uploaded to the public server at usegalaxy.org and aligned to the *P. falciparum* 3D7 v3.2 reference genome (*Böhme et al., 2019*) using Minimap2 (*Li, 2018*) with settings long-read spliced alignment, maximum intron 5 kb, and search GT-AG on the transcript strand only. The .bam alignment file was used as input to the Full-Length Alternative Isoform analysis of RNA (FLAIR) program (*Tang et al., 2020a*) with the–nvrna option using reference genome sequence and annotation (PlasmoDB release 44) for correction of isoform sequences. The corrected and inconsistent isoform sequences outputted by FLAIR were concatenated to create a single transcript annotation .gtf file. To assess whether transcripts annotated by FLAIR were 5′ capped, the transcript 5′ end locations were compared with 5′capped nucleotides annotated from 5′ cap enrichment data using BEDTools (*Quinlan & Hall,*

2010). Genomic locations in .bed file format were constructed for each 5′ cap enrichment study from the dominant_ctss nucleotide reported in all CAGEr clusters. FLAIR transcripts were considered full-length (5′ capped) if the transcript 5′ end was 20 nt or closer to a dominant_ctss nucleotide.

Analysis of external epigenetic data

Epigenetic data for analysis of 5′ capped nucleotide genomic context were obtained from published studies, including chromatin immunoprecipitation sequencing (ChIP-seq) data reported in *Bártfai et al. (2010)*, *Hoeijmakers et al. (2013)*, *Karmodiya et al. (2015)*, *Lu et al. (2017)*, *Tang et al. (2020b)*, *Bhowmick et al. (2020)* and micrococcal nuclease-digested chromatin sequencing (MNase-seq) reported in *Kensche et al. (2016)* were obtained from the NCBI database under series accession numbers GSE23867, GSE39702, GSE63369, GSE85478, PRJNA612099, GSE142803 and GSE66185, respectively. The .csfasta files in accession GSE63369 were converted to .fastq using Cutadapt 1.18 (*Martin, 2011*) and aligned to the *P. falciparum* 3D7 v3.2 reference genome (*Böhme et al., 2019*) using bowtie−1.2.3-linux-x86_64 (*Langmead et al., 2009*) with the following options: -C -v 2 –best –strata m 3. The data from the other studies were uploaded to the public server at usegalaxy.org and processed using fastp (*Chen et al., 2018*) with default settings. For MNase-seq data, read length was trimmed with fastp to a maximum of 72 bp. Processed read data were aligned to the *P. falciparum* 3D7 v3.2 reference genome (*Böhme et al., 2019*) using bowtie2 (*Langmead & Salzberg, 2012*) with default settings. The .bam files for experimental replicates or multiple runs of the same library were merged using SAMtools (*Li et al., 2009*).

In order to calculate nucleosome occupancy from MNase-seq data at each sampled timepoint of development, sonicated genomic DNA control datasets with matching insert size distributions were created. The count of fragments with the same insert size in the alignment file generated from each MNase-seq experiment was obtained using SAMtools (*Li et al., 2009*). Next, insert size distributions were determined from the count of inserts in 31 bins of varying insert size (32–52, 53–72, 73–82, 83–92, 93–102, 103–12, 113–122, 123–132, 133–142, 143–152, 153–162, 163–172, 173–182, 183–192, 193–202, 203–212, 213–222, 223–232, 233–252, 253–272, 273–292, 293–312, 313–332, 333–352, 353–372, 373–392, 393–412, 413–432, 433–452, 453–472, and 473–500 bp). Alignment .bam files of sonicated genomic DNA control with defined insert sizes were made using BamTools (*Barnett et al., 2011*). The alignment files of sonicated genomic DNA control with defined insert sizes were randomly sampled to the equivalent depth in the MNase-seq experimental dataset and merged to create a genomic DNA control with matching insert size distribution using SAMtools (*Li et al., 2009*).

Normalization factors for each sample dataset from all epigenetic experiments were calculated by dividing the reference genome size (23,332,839 bp) by the total depth of coverage in .bam files obtained using SAMtools (*Li et al., 2009*). The read depth at each position in the genome was obtained using BEDTools (*Quinlan & Hall, 2010*). Nucleotide occupancy for each epigenetic feature of interest was calculated as the ratio of normalized read depth in each test dataset to developmental stage-matched control. To prevent division

by zero, a pseudo depth value equal to 0.001 was added to every position for test and control datasets. For ChIP-seq data reported in *Lu et al. (2017)*, RNA polymerase II (RNApolII) occupancy was calculated by subtraction of timepoint-matched normalized IgG negative control read depth from normalized ChIP-seq read depth. Nucleotide occupancy values were used to create genome coverage files in .bigwig format using the bedGraphToBigWig tool (*Kent et al., 2010*).

Genomic intervals annotated as *P. falciparum* accessible chromatin from Assay for Transposase-Accessible Chromatin (ATAC-seq) experiments were obtained from the supplementary information files reported in *Toenhake et al. (2018)*, *Ruiz et al. (2018)* and *Yin et al. (2020)*. The accessible chromatin intervals reported in each study were intersected using BEDTools (*Quinlan & Hall, 2010*).

Data availability

Custom scripts written for analysis of transcriptomic data are available from the GitHub repository: https://github.com/BSI3/5CAPture-seq. Sequencing data generated in this study are available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database under series accession number GSE103036.

RESULTS

Outline of the method for identifying 5′ capped nucleotides

We present a new method for identifying 5′ capped nucleotides from mRNA transcripts. mRNA is purified from the biological sample and used to synthesize cDNA. A portion of the cDNA is reserved as unenriched for generation of a control library and the remainder is enriched for full-length cDNA (Fig. 1A). Full-length cDNA fragments are enriched using recombinant eIF4E protein, which is capable of binding to the 5′ cap nucleotide in first-strand cDNA products (*Edery et al., 1995*). We enriched full-length cDNA using recombinant *P. falciparum* eIF4E protein, which has an affinity for the 5′ cap nucleotide similar to mammalian eIF4E (*Shaw et al., 2007*). To mitigate 5′ end biases when adding adapter, we employ two different adapter ligation protocols, namely cDNA tailing followed by double-stranded adapter ligation (adapter ligation method1) and intramolecular ligation (adapter ligation method2) (Figs. 1B, 1C). 5′ end bias patterns were observed among the unaligned reads, in which C was markedly lower for the first base of method1 HiSeq libraries compared with downstream bases (Fig. S1), and G markedly lower for the first base of all method2 libraries (Figs. S1 and S2).

Accuracy and sensitivity of the method

To assess the accuracy and sensitivity of the proposed method, we performed four experiments with *P. falciparum* mRNA spiked with 5′ capped synthetic RNAs, including two replicates for each adapter ligation protocol. The reads mapped to the *in silico* SIRVomeERCCome reference sequence were used to determine enrichment signals for synthetic RNAs with known 5′ ends (Fig. 2A). The performance for detecting 5′ capped ends was assessed by Receiver Operator Characteristic (ROC) plot (Fig. 2B). As the same batch of synthetic RNA was used for the four spike-in experiments, we also tested whether

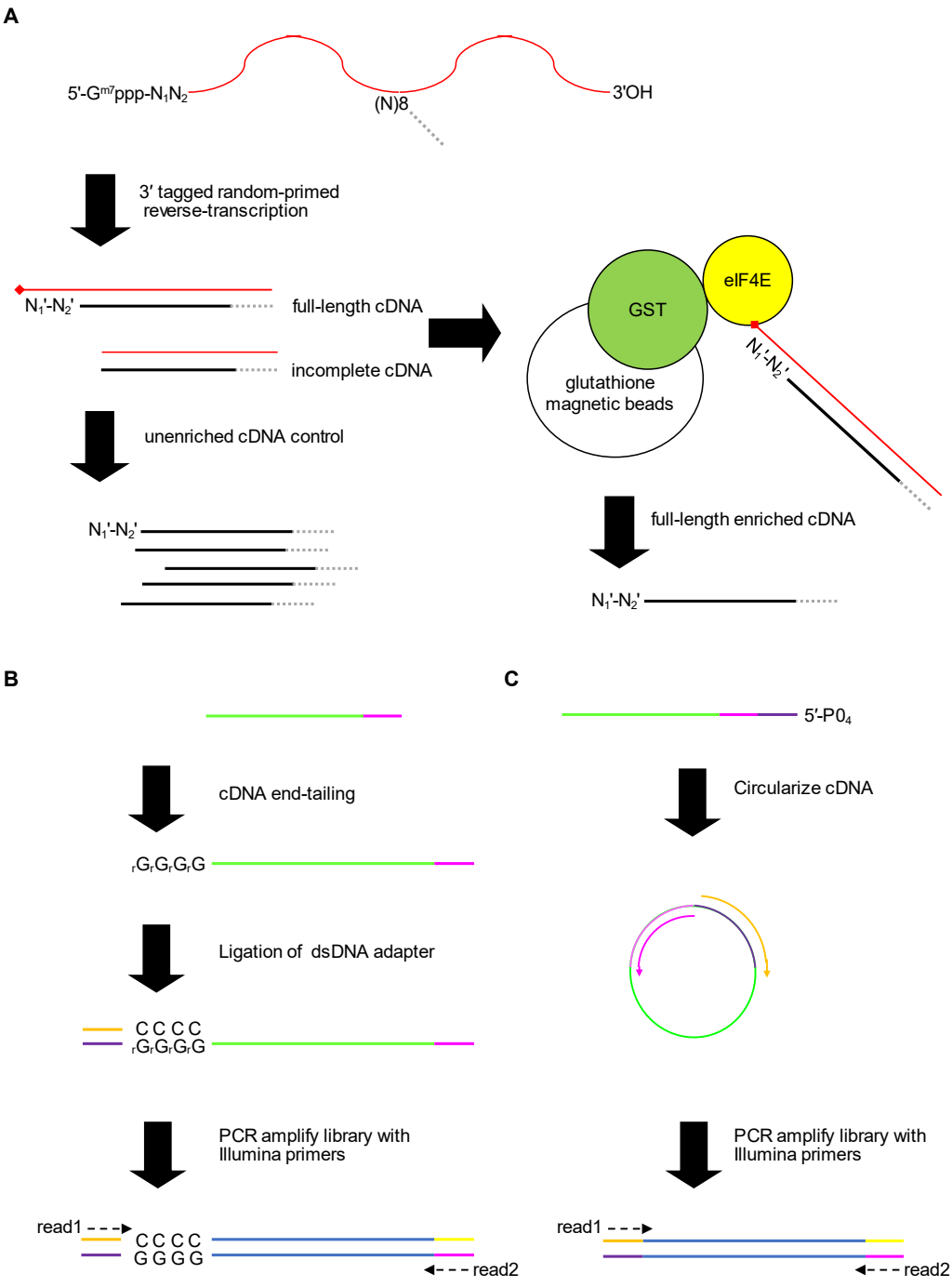

Figure 1 **Schematic of the 5′ cap enrichment transcriptomic method.** (A) 5′ capped mRNA is reverse-transcribed using a random primer with 3′ sequence tag (magenta). Full-length cDNA extending to the 5′ cap structure (lollipop) terminates in a 3′ nucleotide (N'_1) complementary to authentic 5′ mRNA end nucleotide (N_1). A portion of the cDNA sample is reserved as unenriched for generating a control library, whereas the remainder is enriched for full-length cDNA by CAPture (continued on next page…)

Full-size DOI: 10.7717/peerj.11983/fig-1

Figure 1 (…continued)
(*Edery et al., 1995*). Recombinant protein of glutathione S-transferase (GST) fused to 5′ cap binding protein (eIF4E) is immobilized on glutathione magnetic beads. Incomplete cDNA is removed by washing and full-length cDNA eluted from the beads. Unenriched control and full-length enriched cDNA are processed in parallel for 5′ adapter ligation and construction of sequencing libraries. (B) 5′ adapter ligation method1. First-strand cDNA primed using RTNGS1 (Table S1) is 3′ tailed using terminal transferase and rGTP. Up to four riboguanosines (rG) are added to the 3′ end of the cDNA (*Schmidt & Mueller, 1996*). Double-stranded DNA adapter with overhanging complementary Cs is ligated to the first-strand cDNA using RNA ligase 2 enzyme. The adapter is made by annealing top and bottom strand oligonucleotides. Adapter bottom strand oligonucleotide is 5′ phosphorylated and 3′ blocked to prevent adapter self-ligation. DNA library for Illumina next-generation sequencing (NGS) is made by PCR amplification using Illumina primers corresponding to adapter sequences. Paired-end sequencing is performed using standard read1 and read2 primers. (C) 5′ adapter ligation method2. First-strand cDNA primed using RTCIRC (Table S1) with 5′ phosphate ligation donor group (5′-PO_4) is intra-molecularly ligated using CircLigase 2 enzyme. The 3′ tag sequence in RTCIRC contains both 5′ and 3′ adapter sequences (depicted in dark purple and magenta, respectively). DNA library for NGS is made by PCR amplification and paired-end sequenced as for method1.

combining *P*-values from replicate experiments could improve performance. Individual experiments with method2 gave slightly higher performance than method1 as shown by the greater area under ROC curves. The greatest performance was achieved by combining *P*-values from all four experiments. From the ROC curve of all four experiments combined, the optimal cutpoint for classifying 5′ capped nucleotide signals was determined (Fig. 2B). We assessed the sensitivity of the method for detecting 5′ capped ends in absolute terms by comparing the distributions of the ERCC spike-in RNA concentrations for detected *versus* undetected 5′ ends, which were significantly different (Fig. 2C). Moreover, the 5′ capped nucleotide signals for ERCC spike-in RNAs showed a significant positive correlation with the known RNA concentrations indicating that the enrichment signals are quantitative and reflect RNA abundance (Fig. 2D). From the analysis of spike-in RNA enrichment data, we determined the performance of the method using different adapter-ligation protocols and showed that combining *P*-values from all available replicate experiments gives the best performance.

Annotation of 5′ capped nucleotides in *P. falciparum*

We performed 12 experiments in total for identifying 5′ capped ends in *P. falciparum* mRNA isolated from different stages of the IDC. Six experiments were conducted with high sequence coverage on the HiSeq platform (including the four with synthetic RNA spike-in) and six at lower coverage on the MiSeq platform. To assess variability among samples, gene expression profiles were compared. All pairwise combinations of samples from the same stage of the IDC (including samples prepared using different library and sequencing protocols) showed Pearson r values greater than 0.5, whereas some pairwise combinations of samples from different stages showed Pearson r less than 0.5 (Fig. 3A). Enrichment *P*-values were thus combined from four replicate experiments of each stage of development to increase the power of detection. 5′ capped nucleotide signals were annotated for ring, trophozoite and schizont stages of *P. falciparum* IDC based on the threshold of 5′ capped nucleotide signal determined from analysis of spike-in RNA (Table S4). The 5′ capped nucleotide signals detected for *P. falciparum* encompass genomic

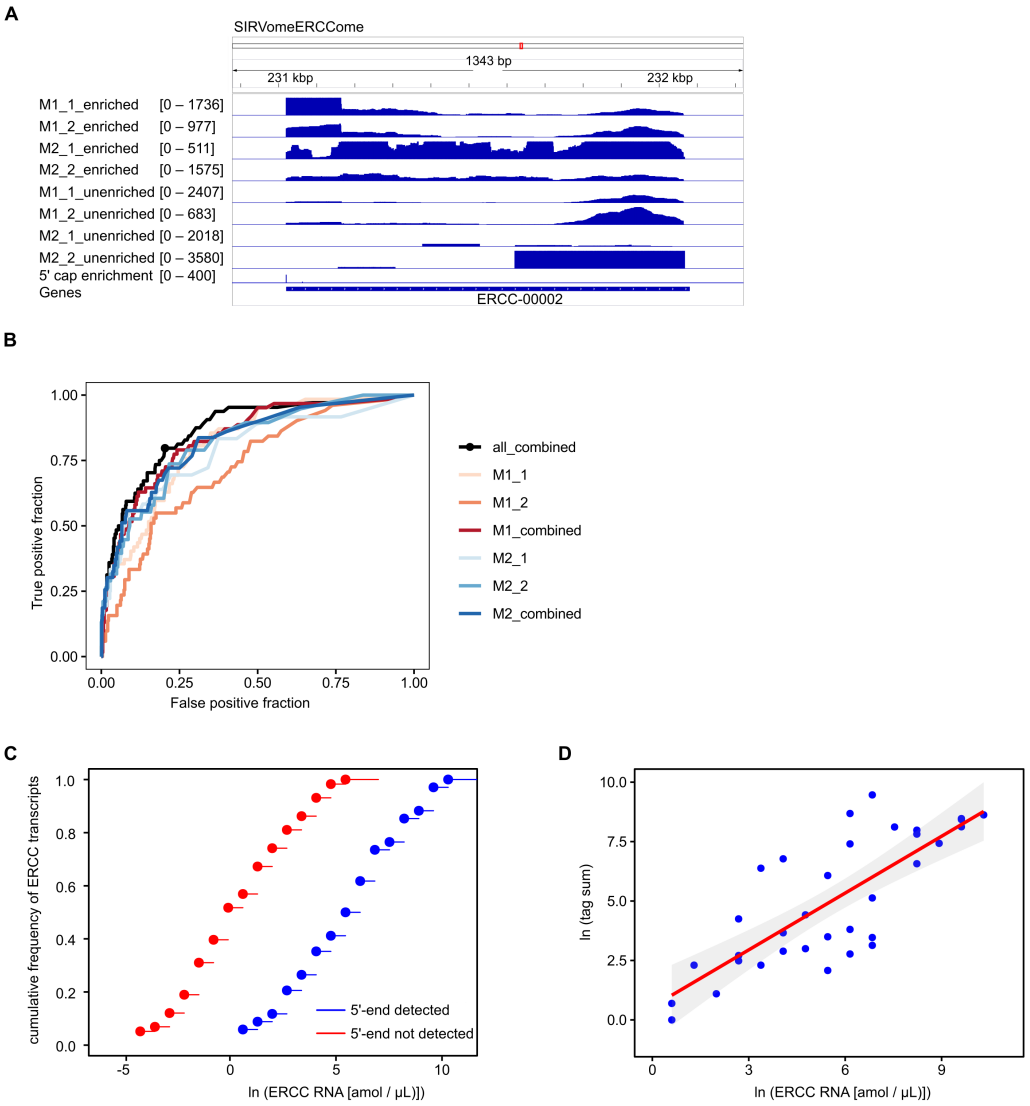

Figure 2 **Validation of 5′ cap enrichment using synthetic RNA.** *Plasmodium falciparum* mRNA was spiked with SIRV set 3 (a pool of synthetic RNAs with known 5′ capped ends). Two experiments were conducted using 5′ adapter ligation method1 (M1_1 and M1_2) and two with 5′ adapter ligation method2 (M2_1 and M2_2). To test whether the power to detect was increased by combining enrichment *P*-values from different experiments, combined *P*-values were calculated from M1_1 and M1_2 experiments (M1_combined), M2_1 and M2_2 experiments (M2_combined) and all four experiments (all_combined). Enrichment *P*-values were transformed to integer counts and clusters of enrichment signals were identified by analysis with CAGEr (*Haberle et al., 2015*). CAGEr cluster signals were used for further analyses. (A) Example of transcriptomic data for the ERCC-00002 synthetic RNA. Reads from enriched and unenriched libraries were aligned to the reference sequence (SIRvomeERCCome). Data tracks underneath the chromosome line show normalized read depth for each library, with the range of depth values in the interval shown indicated in brackets. The bottom data track shows the integer counts of 5′ capped nucleotide enrichment from four experiments (5′ cap enrichment) with the range of integer counts in the interval shown indicated in the bracket. Figure produced using IGV software (*Robinson et al., 2011*). (continued on next page…)

Full-size DOI: 10.7717/peerj.11983/fig-2

Figure 2 (…continued)
(B) Receiver Operator Characteristic (ROC) plots. The black circle on the "all_combined" plot represents the optimal cut-point of enrichment signal (CAGEr cluster tag sum) equal to 32. (C) Empirical cumulative distribution frequency plots of natural log RNA concentration of External RNA Controls Consortium (ERCC) synthetic RNAs. The ERCC RNAs are the subset of synthetic RNAs with a single transcript annotated for each *in silico* gene. The 5′ end was detected for 34 ERCC RNAs, whereas the 5′ end was not detected for 58 ERCC RNAs. The distributions of the two groups were compared by two-sample Kolmogorov–Smirnov test (P-value $= 1.142e^{-07}$). (D) Correlation of natural log (CAGEr cluster tag sums) with natural log concentration of ERCC synthetic RNAs with detected 5′ end. The red line represents linear regression of the data, $P = 3.941e^{-08}$. 95% confidence intervals are indicated by the gray bands.

regions of variable width (Table S4). In other transcriptomic studies of eukaryotic 5′ capped nucleotides, the majority of signals correspond with core promoters, or genomic regions of varying width in which multiple transcription start sites (TSS) are used at different frequencies (*Zhao et al., 2011*). The interquantile signal width reported by CAGEr can be used for the analysis of 5′ capped nucleotide distribution (*Haberle et al., 2015*). Bimodal distributions comprising sharp (<10 bp) and broad (>10 bp) peaks were observed for each sampled stage of the *P. falciparum* IDC; however, the majority of signals showed broad peak widths (Fig. 3B). Many of the 5′ capped nucleotide signals from the three stages of development overlapped the same genomic regions, which made it difficult to separate signals of alternative 5′ ends used dynamically throughout the IDC. Therefore, to simplify the genomic analysis we collapsed the signals across different stages into 17,961 non-overlapping intervals. The dominant 5′ capped nucleotide (5CN) with the strongest signal across the IDC in each genomic interval was investigated in more detail.

Sequence patterns in the vicinity of dominant 5′ capped nucleotides

Eukaryotic promoter regions are characterized by A/T richness and the presence of position-specific short motifs such as upstream TATA boxes and pyrimidine/purine at the $-1/+1$ position (where +1 denotes TSS) (*Müller & Tora, 2014*). To test whether TSS sequence patterns existed at 5CN, base compositions were investigated. The average base composition at 17,961 5CN and 100 flanking genomic bases showed elevated A composition, in particular at +1 and positions immediately downstream compared with random (Fig. 4A).

Aligned reads may possess non-reference (mismatched) 5′ terminal nucleotides that can affect the accuracy of annotating transcript 5′ ends. For example, analysis of CAGE data requires correction for the non-reference G present on the first base of the majority of aligned reads, which originates from the cap signature (*Haberle et al., 2015*). The odds ratio of mismatched base counts from read1 starts aligned one base upstream of 5CN (-1 position) was compared with all other transcribed positions. Odds ratios significantly greater than 1 were observed for mismatched A(8/12 datasets), C(2/12 datasets), G(7/12 datasets) and T(1/12 datasets) (Table S3). Although mismatched bases are more common at -1 than at other transcribed positions, it should be noted that the majority of aligned reads do not start with a mismatch (Table S3). To test whether base mismatching of aligned reads could be influenced by reference sequence, sequence patterns were investigated

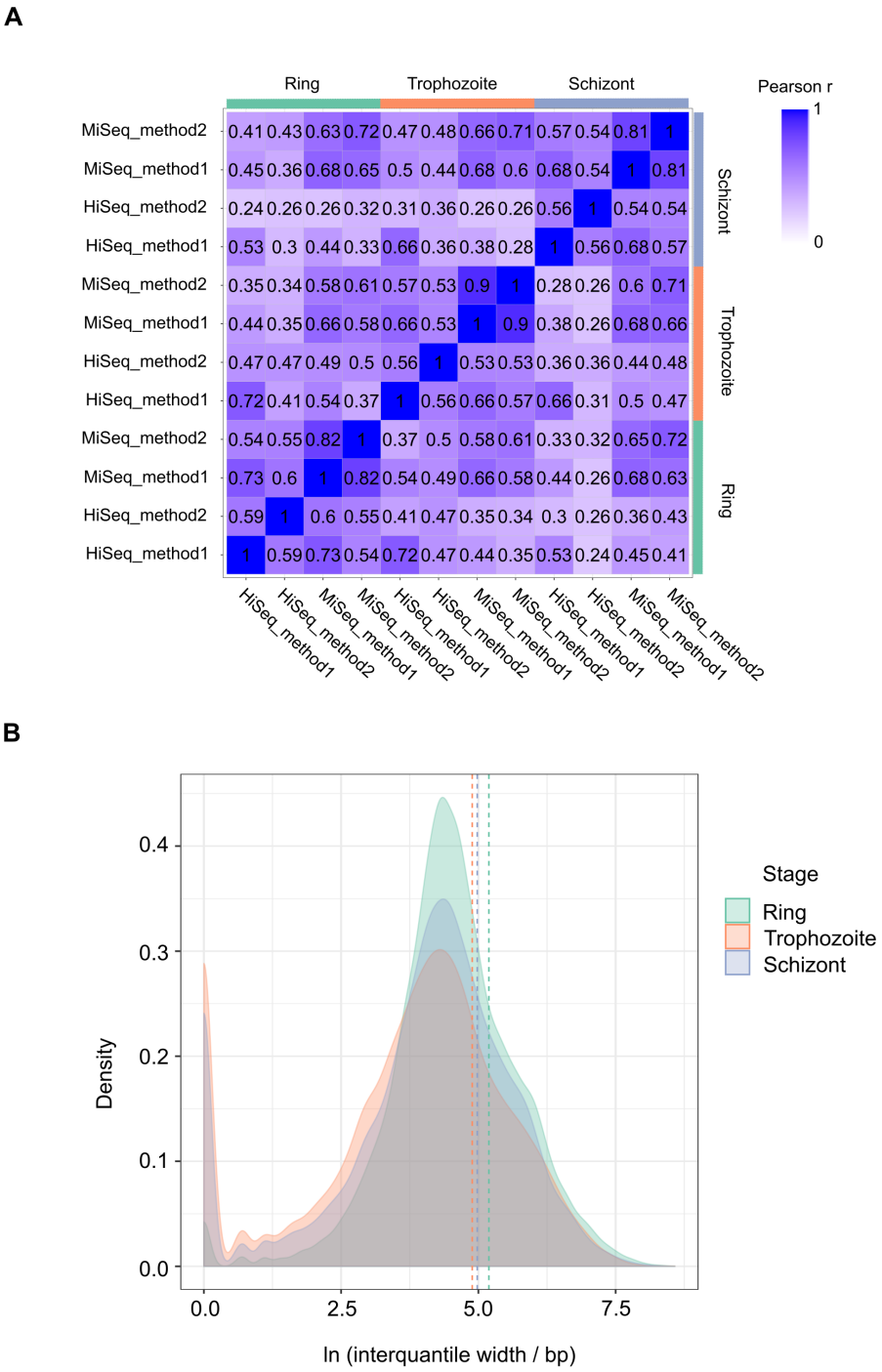

Figure 3 **Transcriptomic data and corresponding 5′ capped nucleotide signals in *Plasmodium falciparum*.** Twelve transcriptomic experiments were conducted for ring, trophozoite and schizont stages of *P. falciparum* development. Sequence data were obtained for six experiments using the Illumina HiSeq platform and six with the Illumina MiSeq platform. Sequencing libraries were generated using adapter ligation method1 and method2 for each sampled stage and sequencing platform. (continued on next page…)
Full-size DOI: 10.7717/peerj.11983/fig-3

Figure 3 (…continued)
Library pairs of unenriched control and 5′ cap enriched were generated for each experiment. (A) Pairwise correlations of experimental data. Expressions of annotated genes were determined from unenriched control library data and used to determine pairwise Pearson's correlation. Samples from the same stage of development are grouped as shown by the bars next to the correlation matrix. The sequencing platform used is indicated by the prefix (HiSeq or MiSeq) and the adapter ligation protocol by the suffix (method1 or method2). The Pearson r value for each combination is shown in the matrix and colored according to the scale shown on the right. (B) Distributions of interquantile cluster widths of 5′ capped nucleotide signals. 5′ cap enrichment was determined for all transcribed nucleotides from the paired transcriptomic data using statistical models. The combined enrichment *P*-values of four replicate experiments for each development stage were transformed to integer counts and clustered using CAGEr (*Haberle et al., 2015*). The interquantile width distribution of 5′ capped nucleotide signals passing threshold (CAGEr cluster tpm > 32) is shown by kernel density plot fitted to natural log of cluster width (bp). The dashed lines indicate the mean cluster width for each stage (green, ring 180 bp; orange, trophozoite 133 bp; blue, schizont 145 bp).

among genomic regions flanking 5CN with high frequencies (>50%) of aligned reads with mismatched first bases (Fig. 4B). Among genomic regions in the vicinity of all 5CN, we observed significant over-representation of A-rich motifs at the 5CN (+1) and downstream positions. CA was the most significant over-represented motif at the −1 position. Different over-represented motifs were identified among genomic regions with high frequencies of aligned reads with mismatches at the -1 position. T, AA, TA and CA were identified as the most significant over-represented motifs among sequences with high frequencies of non-reference A, C, G and T at -1 positions, respectively.

Genomic features in the vicinity of dominant 5′ capped nucleotides

9,334 (52%) of the 5CN are located within *P. falciparum* protein-coding exons (Table S4). Given that eukaryotic TSS are typically located outside of protein-coding regions, many of the exonic 5CN may represent transcriptional noise unrelated to TSS. We hypothesized that epigenetic information could be used to classify TSS in a manner naïve to gene annotation. Eukaryotic core promoter regions possess a distinctive chromatin architecture, characterized by high occupancy of nucleosomes with variant histone H2AZ, and histone modifications including H3K9 acetylation and H3K4 trimethylation (*Jiang & Pugh, 2009*; *Müller & Tora, 2014*). We used published *P. falciparum* epigenetic data of H2A.Z, H3K9 acetylation and H3K4 trimethylation (*Bártfai et al., 2010*) and H2B.Z (*Hoeijmakers et al., 2013*) for unsupervised clustering of 5CN. H2B.Z is an apicomplexan-specific histone variant that colocalizes with H2A.Z (*Hoeijmakers et al., 2013*; *Petter et al., 2013*). The *P*-values from unimodal tests were below the threshold of significance (Silverman critical bandwidth test $P = 0.01$; Hall and York critical bandwidth test $P < 2.2e^{-16}$; Ameijeiras-Alonso et al. excess mass test $P < 2.2e^{-16}$; Cheng and Hall excess mass test $P < 2.2e^{-16}$; Fisher and Marron Cramer-von Mises test $P < 2.2e^{-16}$; Hartigan and Hartigan dip test $P < 2.2e^{-16}$). Hence, we rejected the null hypothesis that the data are unimodal and conclude that more than one cluster exists. Two clusters of 5CN (C1, 6,450 positions; C2, 11,075 positions) were resolved by unsupervised clustering (Fig. S3; Table S4). Furthermore, the number of relevant clusters by the majority vote of clustering indices was two (Fig. S3). C1 5CN overlap genomic regions with locally high occupancies of H2AZ, H2BZ, H3K9 acetylation and H3K4 trimethylation chromatin marks across the

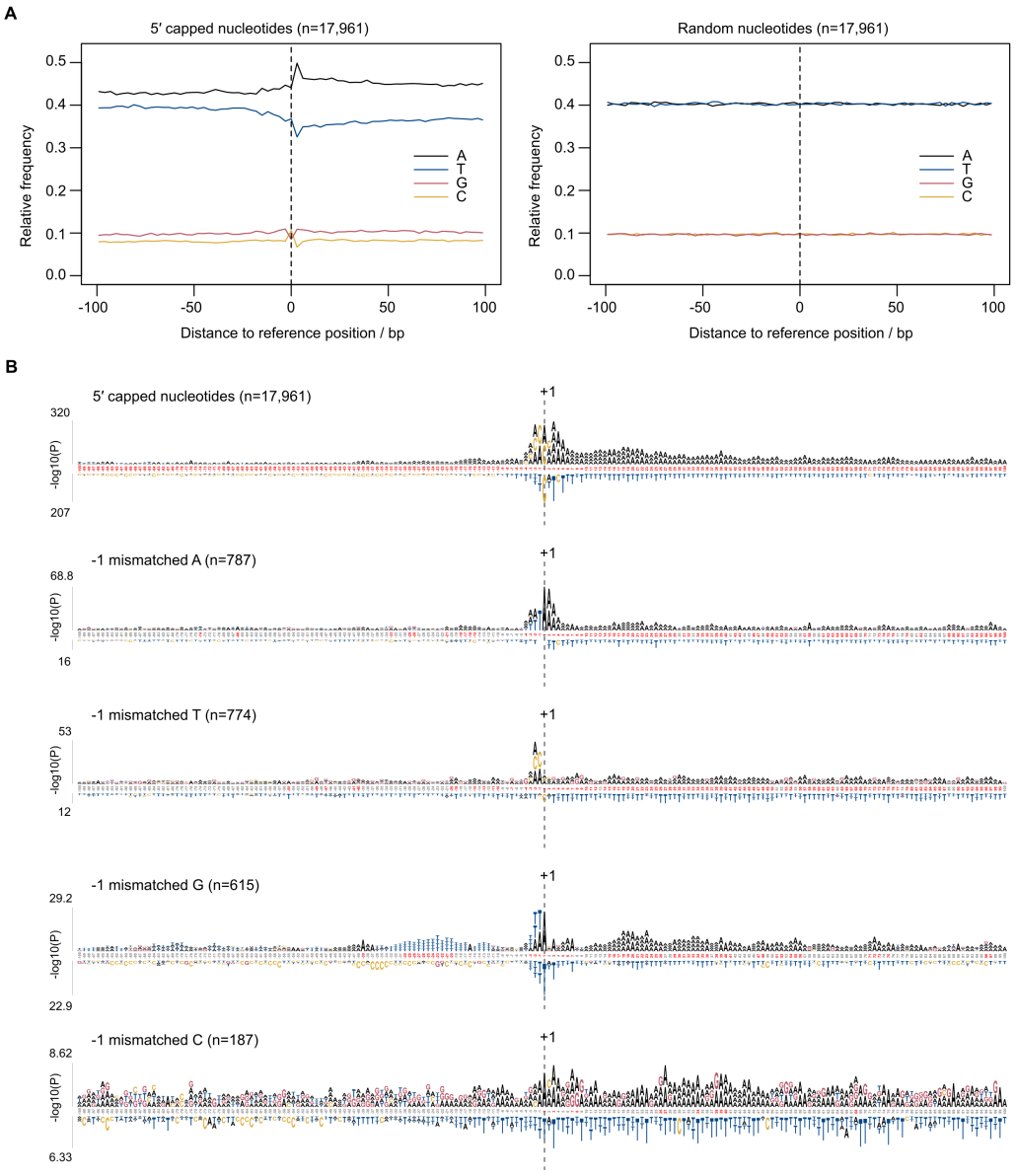

Figure 4 **Sequence patterns in the vicinity of 5′ capped nucleotides.** *Plasmodium falciparum* 3D7 genomic sequences flanking 17,961 dominant 5′ capped nucleotides (5CN) annotated from 5′ cap enriched transcriptomic data and 17,961 randomly selected positions were used for sequence analysis. (A) Average base-composition plots in the vicinity of 5CN (left) and random positions (right). Plots were generated using the seqPattern package in R. (B) Position-specific patterns of 1–4 nt k-mers in the vicinity of 5CN identified by kpLogo (*Wu & Bartel, 2017*) using random genomic sequences as background. The height of each k-mer symbol represents the $-\log_{10}P$- value from statistical testing. Over-represented k-mers are shown above the number line and under-represented below. Bases in the number line highlighted in red indicate positions with Bonferroni-corrected significant k-mers. Sequences flanking all 17,961 5CN were analysed (top) and the subset of sequences with high frequencies (>50%) of aligned reads starting with a mismatched base 1 bp upstream (−1 position) of 5′ capped nucleotides. Patterns are shown for sequences with mismatched A, T, C and G reads at the −1 position.

Full-size DOI: 10.7717/peerj.11983/fig-4

IDC, whereas C2 5CN show low occpancies of these chromatin marks. (Fig. 5A). C1 5CN overlap high occupancies of other histone acetylation and methylation marks, with the notable exception of H3K4me1, which is elevated among C2 5CN (Fig. S4). C1 and C2 5CN are distinguished by other epigenetic features, including elevated nucleosome, PfGCN5 histone acetyltransferase and RNApolII occupancies for C1 (Fig. S5). A higher proportion of 5CN are located outside of protein-coding exons for C1 than C2 (Fig. 5B), and the average distance to the nearest accessible chromatin feature is less for C1 than C2 (Fig. 5C). Analysis of genomic sequence in the vicinity of 5CN showed different patterns of motifs among C1 and C2 (Fig. 5D). T-rich motifs are over-represented upstream of C1 5CN, whereas they are under-represented in the vicinity of C2. The most significant over-represented motifs at the −1 position are TA and CA for C1 and C2, respectively. Previous studies of transcript 5′ ends in *P. falciparum* highlighted patterns of genomic features similar to the C1 group, including elevated H2AZ, H3K9 acetylation and H3K4 trimethylation chromatin marks and high local A/T contents (*Adjalley et al., 2016*; *Kensche et al., 2016*; *Chappell et al., 2020*) Correspondingly, a greater proportion of 5CN in the C1 group than the C2 group are supported by data in the other studies (Fig. 5E; Table S4).

Patterns of 5′ capped nucleotides among genes and associated transcripts

The transcriptomic surveys from data generated in this study and others (*Adjalley et al., 2016*; *Kensche et al., 2016*; *Chappell et al., 2020*) provide comprehensive information of transcript 5′ capped ends. In order to annotate transcripts, data of complete (end-to-end) transcript structures are required, which can be difficult to obtain from short-read cDNA data owing to breaks in coverage at highly AT-rich or unmappable regions of the *P. falciparum* genome (*Chappell et al., 2020*). To our knowledge, the largest available dataset of complete transcript structures was obtained by Nanopore long-read direct RNA sequencing (*Lee et al., 2021*). In direct RNA sequencing, the transcript is sequenced from the 3′ end. However, the 5′ cap modification cannot be identified from the data as a variant base call, and it is therefore not possible to determine if the transcript sequence is complete (extends to the original transcript origin) or not. Transcripts inferred from direct RNA sequencing are often truncated because of RNA degradation, electronic noise generated during the sequencing process and low signals near to the 5′ end (*Soneson et al., 2019*; *Workman et al., 2019*). We created *P. falciparum* transcript annotations by FLAIR analysis of the direct RNA sequencing data (Data S1). Full-length transcripts were identified from the correspondence of transcript 5′ end locations with 5CN from cDNA data (Table S5; Fig. 6). A total of 4,512 transcript 5′ ends (42%) corresponded with 5CN from one or more studies. The greatest number of transcripts annotated by FLAIR corresponded with 5CN from our data (2482), followed by 5CN from (*Kensche et al., 2016*; *Adjalley et al., 2016*; *Chappell et al., 2020*) with 1,720, 1,580, and 1,230 transcripts, respectively. Some transcript 5′ ends corresponded with 5CN from more than one study, but only 148 transcript 5′ ends corresponded with 5CN from all four studies (Fig. 6A). The majority of full-length transcripts overlap annotated genes/coding regions, with minorities of antisense and novel transcripts (Fig. 6B). Multiple full-length transcript isoforms with alternative 5′ ends are

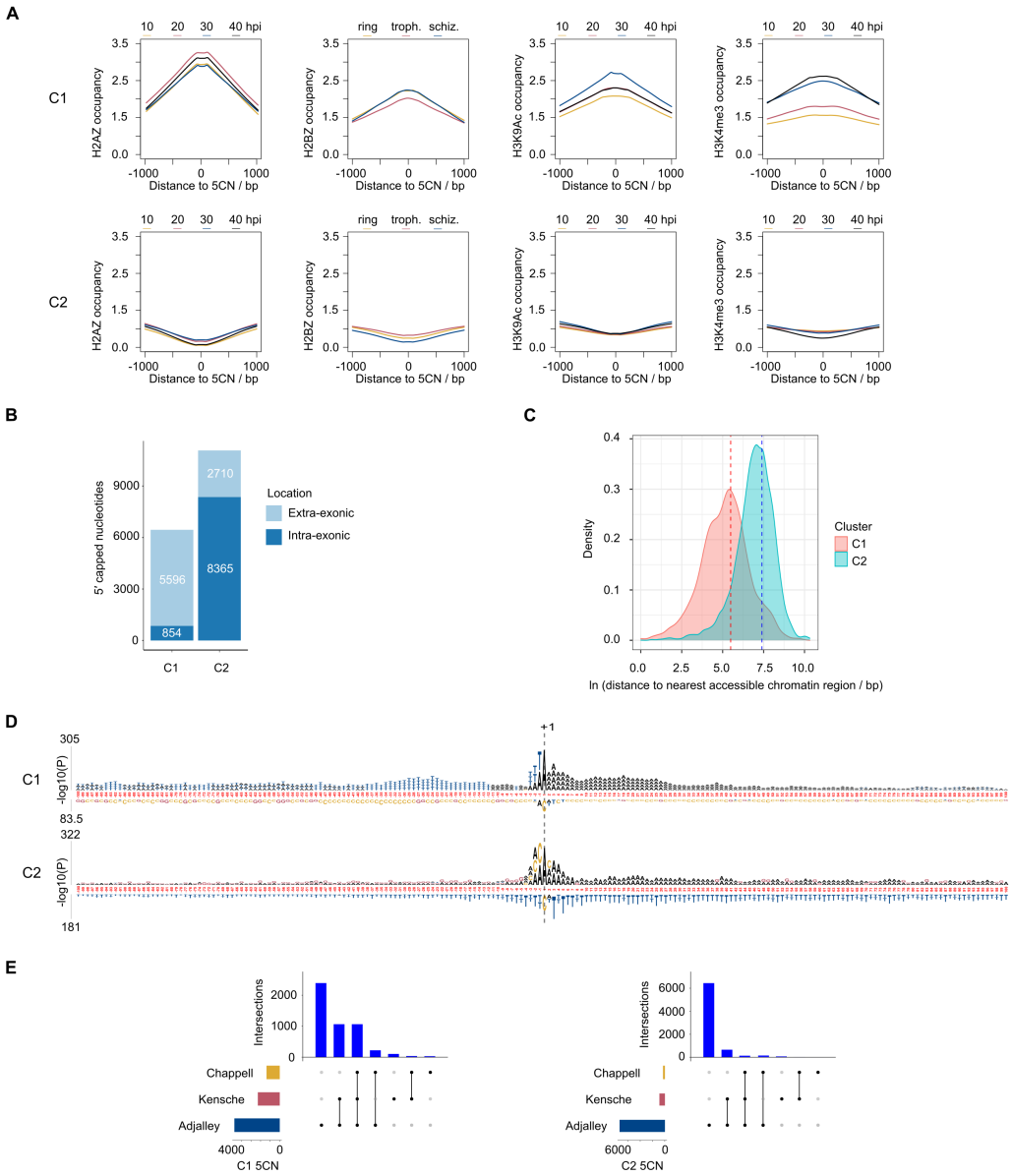

Figure 5 **Clustering of *Plasmodium falciparum* 5′ capped nucleotides by genomic features.** Genomic features were analyzed for dominant 5′ capped nucleotides (5CN) annotated from 5′ cap enriched transcriptomic data. Two clusters of 17,525 5CN (C1 and C2) were resolved by unsupervised clustering (Fig. S3) using chromatin-immunoprecipitation sequencing (ChIP-seq) data of H2AZ, H2BZ, H3K9ac and H3K4me3 occupancies (*Bártfai et al., 2010*; *Hoeijmakers et al., 2013*). (A) Epigenetic marks in the vicinity of 5CN. Score matrices were constructed from ChIP-seq data for each 5CN in the genome and 1,000 bp flanks, and plots of average scores for cluster C1 (top) and C2 (bottom) made using the genomation package (*Akalin et al., 2015*) in R. Data from different stages of the intra-erythrocytic development cycle were plotted on the same axes for each type of chromatin feature shown. Numbering on the *y*-axes refers to average normalized occupancies. (B) Compositions of cluster C1 and C2 5CN with respect to annotated protein-coding exons. (continued on next page…)

Full-size DOI: 10.7717/peerj.11983/fig-5

Figure 5 (…continued)
(C) Distributions of 5CN distance to nearest accessible chromatin region (annotated peak signal from Assay for Transposase-Accessible Chromatin sequencing data (*Toenhake et al., 2018*; *Ruiz et al., 2018*; *Yin et al., 2020*) are shown by kernel density plot fitted to natural log of distance (bp). The mean distance is shown by the dashed line for each distribution (C1, 243 bp; C2, 1,641 bp).(D) Position-specific patterns of 1–4 nt k-mers in the vicinity of 5CN identified by kpLogo (*Wu & Bartel, 2017*) using 17,961 random genomic sequences as background. The height of each k-mer symbol represents the $-\log_{10}P$- value from statistical testing. Over-represented k-mers are shown above the number line and under-represented below. Bases in the number line highlighted in red indicate positions with Bonferroni-corrected significant k-mers.(E) Intersection of 5CN with 5′ capped nucleotide enrichment data from independent studies. Intersection of 5CN with 5′ cap nucleotide enrichment data from 5′ cap sequencing (*Adjalley et al., 2016*), and SMART enrichment (*Kensche et al., 2016*; *Chappell et al., 2020*) was scored if read depth was two or greater. Intersections are shown by UpSetR plot (*Conway, Lex & Gehlenborg, 2017*), which is a matrix-based representation of set sizes and their intersections. Matrix rows represent the sets (5CN annotated from data in this study) and columns represent intersections of 5CN supported by data in other studies. Set sizes are shown by the horizontal bar plots on the left. All combinations of set intersections are shown by the matrix cell on the right, in which sets that are part of a given intersection are represented by black-filled circles. Sets that are not part of the intersection are shown as light gray circles. The sets considered in each intersection (black circles) are connected by vertical black lines to emphasize the column-based relationships. Bars above the matrix columns represent the number of 5CN in each intersection. Intersections of 5CN with other studies are shown separately for C1 (left) and C2 (right).

apparent for *P. falciparum* genes, including the PF3D7_1434200 (calmodulin) gene shown in Fig. 6C. The 5′ ends of two full-length coding transcripts map to a transcription initiation region upstream of the calmodulin gene mapped previously by RNA-ligase mediated 5′ rapid amplification of cDNA ends (*Polson & Blackman, 2005*). The untranslated regions of *P. falciparum* gene transcripts can extend for several hundred bases beyond the terminal coding exons (*Chappell et al., 2020*), such that the transcripts can overlap coding exons of adjacent genes. In the example shown in Fig. 6D, the 3′ untranslated region of a full-length coding transcript for the PF3D7_1214600 gene overlaps, and is antisense to the adjacent PF3D7_1214500 gene. Although the *P. falciparum* genome is gene-dense with median intergenic distance less than 2 kb (*Russell et al., 2013*), novel full-length transcripts not overlapping coding exons are present. In the example shown in Fig. 6E, a novel transcript initiates near to a full-length coding transcript of the PF3D7_0419600 gene on the opposite strand.

DISCUSSION

Accurate annotation of eukaryotic genes, in particular those of less well-characterized organisms such as *P. falciparum,* requires detailed transcript information. A major challenge for transcript annotation is identifying 5′ ends because truncated RNA or cDNA sequences are common experimental artifacts in transcriptomic data. Previous transcriptomic studies of *P. falciparum* employing methods for enrichment of full-length cDNA were not comprehensive as transcript 5′ ends could not be detected for some genes known to be expressed in the IDC (*Adjalley et al., 2016*; *Kensche et al., 2016*; *Chappell et al., 2020*). To provide a more complete catalog of transcript 5′ ends in *P. falciparum*, we developed a novel method to complement the available data. In our method, full-length cDNA is enriched using 5′ cap binding protein eIF4E (*Edery et al., 1995*). Instead of eIF4E

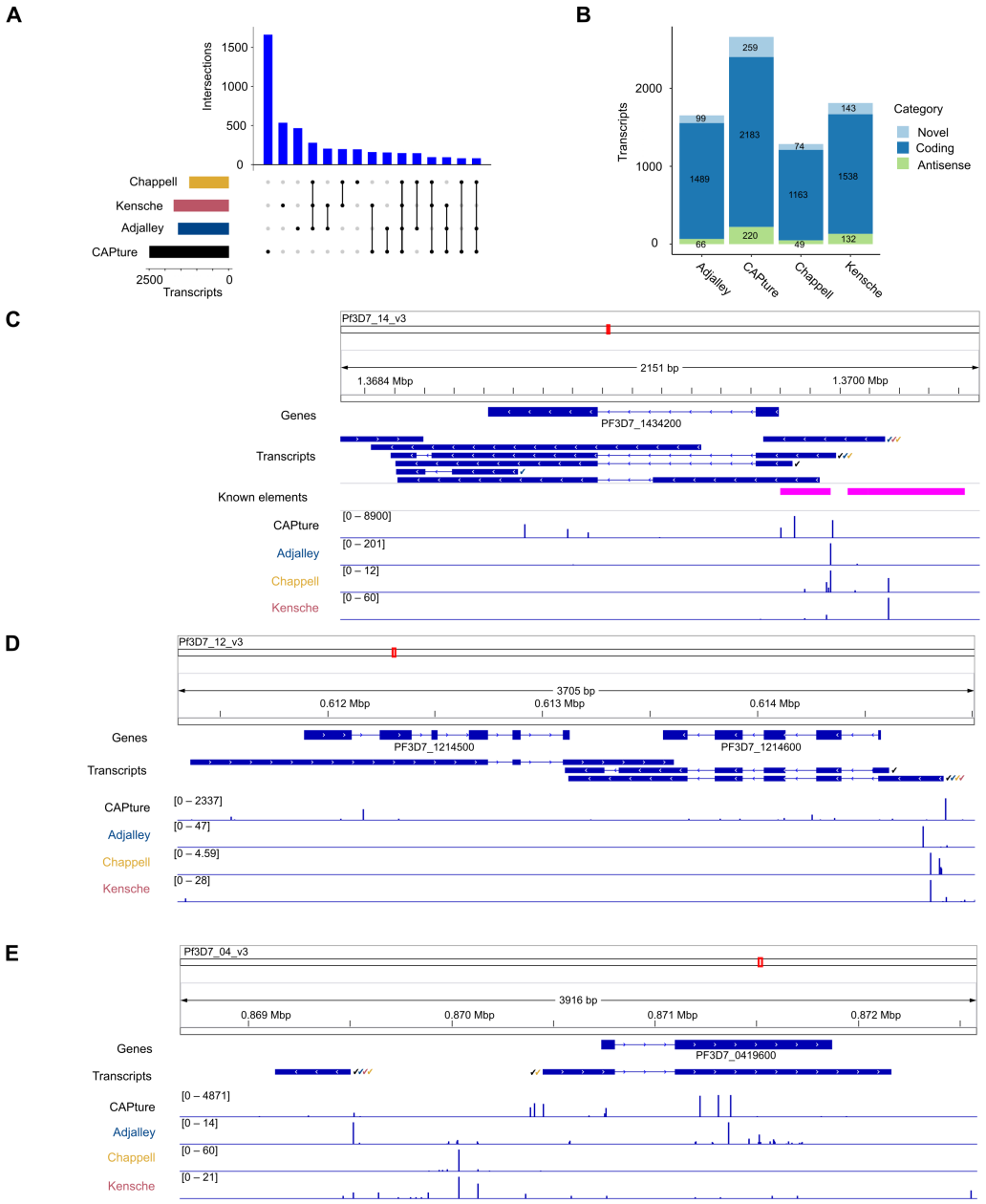

Figure 6 **Annotation of 5′ capped transcripts in *Plasmodium falciparum*.** *P. falciparum* transcripts were annotated from FLAIR analysis (*Tang et al., 2020a*) of direct RNA sequencing data (*Lee et al., 2021*). 5′ capped transcripts were identified by the overlap of 5′ end nucleotide position with 5′ capped nucleotide enrichment data. (A) Intersection of transcript 5′ ends with 5′ capped nucleotide enrichment data from independent studies. Intersections of transcript 5′ ends are shown with data from this study (labeled CAPture), 5′ cap sequencing ((*Adjalley et al., 2016*), labeled Adjalley), and SMART enrichment ((*Kensche et al., 2016*), labeled Kensche; (*Chappell et al., 2020*), labeled Chappell). Intersections are shown by UpSetR plot (*Conway, Lex & Gehlenborg, 2017*), which is a matrix-based representation of set sizes and their intersections. (continued on next page…)

Full-size DOI: 10.7717/peerj.11983/fig-6

Figure 6 (…continued)
Matrix rows represent the sets (transcripts) and columns represent intersections of transcript 5′ ends with 5′ cap nucleotide positions annotated from each study. Set sizes are shown by the horizontal bar plots on the left. All combinations of set intersections are shown by the matrix cell on the right, in which sets that are part of a given intersection are represented by black-filled circles. Sets that are not part of the intersection are shown as light gray circles. The sets considered in each intersection (black circles) are connected by vertical black lines to emphasize the column-based relationships. Bars above the matrix columns represent the number of transcripts in each intersection.(B) Categories of 5′ capped transcripts supported by 5′ capped nucleotide enrichment data from different studies. The numbers of transcripts in each category (coding, antisense, and novel) are indicated by stacked barplots for each 5′ capped nucleotide enrichment study. Parts C –E show examples of 5′ capped transcripts in genomic context. Figures were generated using IGV software (*Robinson et al., 2011*). Annotated genes, including the boundaries of coding exons (blue bars) and introns (blue lines) are shown in the Genes tracks. Transcripts annotated by FLAIR are shown in the Transcripts tracks. The signals of dominant 5′ capped nucleotides from full-length cDNA enrichment studies are indicated in the tracks labeled CAPture (this study), Adjalley (5′ cap sequencing reported by (*Adjalley et al., 2016*), Chappell (5UTR-seq reported by (*Chappell et al., 2020*) and Kensche (SMART enrichment reported by (*Kensche et al., 2016*)). The 5′ capped nucleotide signals are shown as values of tpm.dominant_ctss reported by combined CAGEr (*Haberle et al., 2015*) analysis of data across intraerythrocytic stages of *P. falciparum* development with the range values in the interval shown indicated in the bracket in each track. Full-length transcripts with 5′ ends that correspond to dominant 5′ capped nucleotides are indicated by tick marks next to the transcript, which are colored according to which full-length cDNA enrichment study is in agreement (CAPture, black; Adjalley, blue; Chappell, yellow; Kensche, red). (C) PF3D7_1434200 (calmodulin) gene region. Known elements are indicated by magenta boxes, including 5′ ends mapped by 5′ rapid amplification of cDNA ends (5′-RACE) (*Polson & Blackman, 2005*) from the most proximal to the most distal and upstream promoter/enhancer element mapped by reporter gene analysis (*Crabb & Cowman, 1996*). (D) Example of antisense transcripts for the PF3D7_1214500 gene, which are also coding transcripts for the adjacent PF3D7_1214600 gene. (E) Example of a novel transcript upstream of the PF3D7_0419600 gene.

protein, other agents with selective 5′ cap binding affinity can be used for enriching 5′ capped RNA (*Bhardwaj et al., 2019*). To separate enriched 5′ end signals from noise, the unenriched background within the body of each transcript can be modeled from local windows assumed to represent the same transcript (*Bhardwaj et al., 2019*). This data analysis method though is not likely to be effective in organisms such as *P. falciparum* with gene-dense genomes, in which multiple overlapping transcripts of varying abundance can arise from the same genomic region (*Van Lin et al., 2001*). In our method, the unenriched cDNA is sequenced in parallel as a control, which allows for statistical modeling of enrichment at the nucleotide level. Moreover, the power to detect 5′ end signals can be increased by combining enrichment *P*-values from replicate experiments as shown using spike-in RNA (Fig. 2). Although performance is increased with more replicates, the accuracy of the method is limited by the low efficiency of biochemical enrichment of full-length cDNA, as the eIF4E protein used for enrichment has micromolar affinity for the mRNA 5′ cap (*Shaw et al., 2007*).

The processes of adding sequencing adapter in oligo-capping and SMART are integral to the process of enrichment for full-length cDNA. Therefore, it is not possible to obtain a matching unenriched cDNA control for statistical modelling with these methods. We employed two different protocols for adding sequencing adapter to cDNA ends which retain transcript 5′ terminal nucleotides and can be used with unenriched and full-length enriched cDNA. However, the short homopolymer tail at the beginning of read1 from method1

libraries can be problematic for some NGS platforms. We circumvented the problem of low diversity caused by the homopolymer tail for method1 libraries by sequencing in one lane of a flow cell shared with more diverse libraries in other lanes (HiSeq platform) or using a custom dark-cycle sequencing recipe (MiSeq platform). The observed lower frequency of C for the first base of method1 HiSeq libraries compared with downstream bases (Fig. S1) is a consequence of the bioinformatic trimming of the homopolymer tail appended to the cDNA end. If the tail is not trimmed, read alignment accuracy could be affected because of mismatches to the reference. The observed lower frequency of G for the first base of method2 libraries compared with downstream bases (Figs. S1 and S2) is consistent with the reported Circligase ligation bias (*Kwok et al., 2013*) against the complementary C ligation acceptor on first-strand cDNA in method2. The different 5′ end biases for method1 and method2 could limit detection of certain 5′ capped nucleotides, although if both protocols are used they can complement each other (Fig. 2).

A bimodal distribution of transcript 5′ end signal was observed from our data, suggesting the presence of sharp and broad 5′ end signals in *P. falciparum*. Sharp and broad 5′ end signals were also described in other *P. falciparum* studies (*Chappell et al., 2020*; *Adjalley et al., 2016*), although the width distribution inferred from the data depends on the noise filtering and clustering parameters employed. Locally high A contents were observed at 5CN (Fig. 4A), in agreement with the patterns of *P. falciparum* transcript 5′ ends described in other studies (*Chappell et al., 2020*; *Adjalley et al., 2016*; *Kensche et al., 2016*). Aligned reads starting with mismatched bases are significantly more prevalent at the -1 position in our data than other transcribed regions (Fig. 4B). Because of the locally biased base content in the vicinity of 5CN, reads with mismatched T or A may arise from technical errors in these regions, *e.g.*, template slippage during reverse-transcription, PCR, or sequencing by synthesis reactions. For -1 positions with high frequencies of reads starting with mismatched G, the over-represented TA motif suggests a biological reason, most likely the cap signature of reverse-transcribed 5′ cap m^7G.

Two groups of 5CN in *P. falciparum* (C1 and C2) were identified by unsupervised clustering of our data. Although determining the mechanisms that generate the different types of 5CN is beyond the scope of the present study, the genomic contexts of 5CN provide clues as to their origins. C1 likely represents TSS since they overlap genomic regions with high local occupancies of H2AZ, H3K9Ac and H3K4me3 histone modifications (Fig. 5A, Figs. S4, S5) that are strongly associated with transcription initiation in other eukaryotes (*Müller & Tora, 2014*). Other evidence to support the conclusion that C1 represents TSS includes the following epigenetic patterns consistent with TSS in other eukaryotes: (i) elevated occupancies of other histone acetylation marks (H3K14Ac, H3K18Ac, H3K27Ac; Fig. S4) that are associated with eukaryotic transcriptional activation (*Zhao & Garcia, 2015*), (ii) patterns of elevated occupancy of H3K4me2, but reduced occupancy of H3K4me1 (Fig. S4) and reduced occupancy of H3 (Fig. S5) that are associated with TSS (*Koch et al., 2007*), (iii) elevated nucleosome and RNApolII occupancies (Fig. S5) that are associated with TSS (*Zhao et al., 2011*), and (iv) elevated occupancy of PfGCN5 (Fig. S5), the *P. falciparum* homolog of GCN5 protein (*Bhowmick et al., 2020*), which is a component of the SAGA transcription coactivator complex that acetylates histones at core

promoters and enhancers (*Cheon et al., 2020*). Furthermore, C1 nucleotides are located mostly outside of protein-coding exons (Fig. 5B) and are closer to accessible chromatin (Fig. 5C) where transcription is more likely to initiate. The TA over-represented motif at the -1 position for the C1 group (Fig. 5D) is consistent with a eukaryotic transcription initiation motif (*Müller & Tora, 2014*), and the upstream T-rich and downstream A-rich motifs are consistent with eukaryotic TSS locator motifs (*Lubliner, Keren & Segal, 2013*).

The C2 group of 5CN lacks epigenetic features of TSS (Fig. 5A). C2 nucleotides are also more prevalent in coding regions (Fig. 5B) and are further away from accessible chromatin (Fig. 5C) where transcription is less likely to initiate. C2 nucleotides are less well supported by other methods (Fig. 5E; Table S4) and many could therefore represent technical noise. However, the significant over-representation of sequence motifs including CA at the -1 position (Fig. 5D) is indicative of a biological phenomenon that requires further investigation. C2 nucleotides show elevated occupancy of H3K4me1 (Fig. S4), which is elevated distal to TSS in eukaryotes (*Koch et al., 2007*; *Zhao et al., 2011*). High occupancy of H3K4me1 is observed at the start sites of enhancer RNAs (eRNAs) in eukaryotes, which represent a class of generally low-abundance non-coding RNA; C2 nucleotides are however unlikely to represent eRNA start sites, as they generally do not overlap genomic regions associated with accessible chromatin and the H3K27Ac chromatin mark characteristic of eRNA start sites (*Lewis, Li & Franco, 2019*). Rather than TSS or 5′ ends of eRNAs, C2 could represent transcript 5′ ends generated post-transcriptionally by co-translational cleavage and cytoplasmic recapping (*Trotman & Schoenberg, 2019*). This is speculative though since it is not known if mRNA capping occurs post-transcriptionally in the *P. falciparum* cytoplasm. However, co-translational cleavage could occur as truncated transcript isoforms are evident in *P. falciparum* polysomal fractions (*Bunnik et al., 2013*).

The 5′ ends of 4,512 transcripts annotated from direct RNA sequencing corresponded to 5CN from cDNA data, suggesting that these transcripts are 5′ capped. However, the lack of corroborative cDNA data for the majority of transcripts from direct RNA sequencing suggests that they are truncated. It should be noted though that relatively few annotated transcript 5′ ends correspond to 5CN from more than one study (Fig. 6A), suggesting that verification of transcript 5′ capped ends from short-read cDNA sequencing data is challenging, perhaps because of dynamic usage of transcript initiation sites during development (*Adjalley et al., 2016*; *Chappell et al., 2020*), different 5′ end biases among full-length cDNA enrichment methods (see above), stochastic variation in TSS usage (*Xu, Park & Zhang, 2019*), and the difficulty in aligning short cDNA reads to extremely AT-rich regions in the *P. falciparum* genome (*Chappell et al., 2020*). Although the majority of *P. falciparum* transcripts overlap coding regions (Fig. 6B), many transcripts with alternative 5′ capped ends partially overlap the annotated open reading frame (Fig. 6C, Fig. 6D). These alternative transcripts could possess functions for production of N-terminal truncated protein isoforms, or regulatory non-coding RNA (*Trotman & Schoenberg, 2019*). On the other hand, transcripts with alternative 5′ capped ends could represent RNA decay intermediates (*Arribere & Gilbert, 2013*).

CONCLUSIONS

A new transcriptomic approach for identifying mRNA 5′ capped nucleotides was developed to comprehensively annotate *P. falciparum* 5′ capped transcripts expressed in intraerythrocytic stages of the life cycle. Two groups of 5CN were annotated from data generated using the new method with distinctive epigenetic and genomic sequence patterns. The correspondence of 5CN with transcript 5′ ends inferred from direct RNA sequencing revealed patterns of 5′ capped transcripts, including the widespread occurrence of transcripts with alternative 5′ ends that partially overlap gene coding regions.

ACKNOWLEDGEMENTS

We thank Drs. Lia Chappell and Vern Lee for providing *P. falciparum* 5UTR-seq and direct RNA transcriptomic data, respectively and Illumina Inc. for assistance with dark cycle recipe sequencing.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Platform Technology Management section, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand. [P1201270 and P1551103, both projects jointly to Philip J. Shaw and Jittima Piriyapongsa]; the Thailand Research Fund [RSA5880064 to Chairat Uthaipibull]; and the National Science and Technology Development Agency, (Thailand) [P1300832 to Chairat Uthaipibull, P1850116 (Research Chair Grant) and P1450883 to Sumalee Kamchonwongpaisan]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Platform Technology Management section, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand.: P1201270, P1551103.
The Thailand Research Fund: RSA5880064.
The National Science and Technology Development Agency, (Thailand): P1300832, P1850116, P1450883.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Philip J. Shaw conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Jittima Piriyapongsa conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- Pavita Kaewprommal, Chadapohn Chaosrikul, Krirkwit Teeravajanadet and Manon Boonbangyang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Chayaphat Wongsombat performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Chairat Uthaipibull, Sumalee Kamchonwongpaisan and Sissades Tongsima analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

Human erythrocytes and serum were obtained from donors after providing informed written consent, following a protocol approved by the Ethics Committee, National Science and Technology Development Agency, Pathum Thani, Thailand, document no. 0021/2560.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Sequencing data generated in this study are available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database: GSE103036.

Data Availability

The following information was supplied regarding data availability:

Custom scripts written for analysis of transcriptomic data are available at GitHub: https://github.com/BSI3/5CAPture-seq.

Sequencing data generated in this study are available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database: GSE103036.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11983#supplemental-information.

REFERENCES

Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, Shi X, Jacques J, Lancaster MA, Pan JQ, Regev A, Levin JZ. 2018. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nature Methods* **15**:505–511 DOI 10.1038/s41592-018-0014-2.

Adjalley SH, Chabbert CD, Klaus B, Pelechano V, Steinmetz LM. 2016. Landscape and dynamics of transcription initiation in the malaria parasite *Plasmodium falciparum*. *Cell Reports* **14**:2463–2475 DOI 10.1016/j.celrep.2016.02.025.

Akalin A, Franke V, Vlahovi ek K, Mason CE, Schubeler D. 2015. genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**:1127–1129 DOI 10.1093/bioinformatics/btu775.

Ameijeiras-Alonso J, Crujeiras RM, Rodríguez-Casal A. 2019. Mode testing, critical bandwidth and excess mass. *TEST* **28**:900–919 DOI 10.1007/s11749-018-0611-5.

Arribere JA, Gilbert WV. 2013. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Research* **23**:977–987 DOI 10.1101/gr.150342.112.

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**:1691–1692 DOI 10.1093/bioinformatics/btr174.

Bártfai R, Hoeijmakers WAM, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger T-W, Françoijs K-J, Stunnenberg HG. 2010. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLOS Pathogens* **6**:e1001223 DOI 10.1371/journal.ppat.1001223.

Bhardwaj V, Semplicio G, Erdogdu NU, Manke T, Akhtar A. 2019. MAPCap allows high-resolution detection and differential expression analysis of transcription start sites. *Nature Communications* **10**:3219 DOI 10.1038/s41467-019-11115-x.

Bhowmick K, Tehlan A, Sunita xx, Sudhakar R, Kaur I, Sijwali PS, Krishnamachari A, Dhar SK. 2020. Plasmodium falciparum GCN5 acetyltransferase follows a novel proteolytic processing pathway that is essential for its function. *Journal of Cell Science* **133**:jcs236489 DOI 10.1242/jcs.236489.

Böhme U, Otto TD, Sanders MJ, Newbold CI, Berriman M. 2019. Progression of the canonical reference malaria parasite genome from 2002–2019. *Wellcome Open Research* **4**:58 DOI 10.12688/wellcomeopenres.15194.2.

Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, De Risi JL. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLOS Biology* **1**:e5 DOI 10.1371/journal.pbio.0000005.

Bunnik EM, Chung D-WD, Hamilton M, Ponts N, Saraf A, Prudhomme J, Florens L, Le Roch KG. 2013. Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biology* **14**:R128 DOI 10.1186/gb-2013-14-11-r128.

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**:327–336 DOI 10.1006/geno.1996.0567.

Chappell L, Ross P, Orchard L, Russell TJ, Otto TD, Berriman M, Rayner JC, Llinás M. 2020. Refining the transcriptome of the human malaria parasite *Plasmodium falciparum* using amplification-free RNA-seq. *BMC Genomics* **21**:395 DOI 10.1186/s12864-020-06787-5.

Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* **61** DOI 10.18637/jss.v061.i06.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–i890 DOI 10.1093/bioinformatics/bty560.

Cheon Y, Kim H, Park K, Kim M, Lee D. 2020. Dynamic modules of the coactivator SAGA in eukaryotic transcription. *Experimental & Molecular Medicine* **52**:991–1003 DOI 10.1038/s12276-020-0463-4.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**:2938–2940 DOI 10.1093/bioinformatics/btx364.

Crabb BS, Cowman AF. 1996. Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the United States of America* **93**:7289–7294 DOI 10.1073/pnas.93.14.7289.

Das M, Harvey I, Chu LL, Sinha M, Pelletier J. 2001. Full-length cDNAs: more than just reaching the ends. *Physiological Genomics* **6**:57–80 DOI 10.1152/physiolgenomics.2001.6.2.57.

Edery I, Chu LL, Sonenberg N, Pelletier J. 1995. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Molecular and Cellular Biology* **15**:3363–3371 DOI 10.1128/mcb.15.6.3363.

Fuchs RT, Sun Z, Zhuang F, Robb GB. 2015. Bias in ligation-based small RNA Sequencing library construction is determined by adaptor and RNA structure. *PLOS ONE* **10**:e0126049 DOI 10.1371/journal.pone.0126049.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**:498–511 DOI 10.1038/nature01097.

Haberle V. 2020. seqPattern: visualising oligonucleotide patterns and motif occurrences across a set of sorted sequences. *Available at https://bioconductor.org/packages/seqPattern/* (accessed on 6 May 2021).

Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. 2015. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Research* **43**:e51–e51 DOI 10.1093/nar/gkv054.

Ho CK, Shuman S. 2001. A yeast-like mRNA capping apparatus in *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the United States of America* **98**:3050–3055 DOI 10.1073/pnas.061636198.

Hoeijmakers WAM, Salcedo-Amaya AM, Smits AH, Françoijs K-J, Treeck M, Gilberger T-W, Stunnenberg HG, Bártfai R. 2013. H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome: two histone variants demarcate promoters in *P. falciparum*. *Molecular Microbiology* **87**:1061–1073 DOI 10.1111/mmi.12151.

Hubert M, Rousseeuw PJ, VandenBossche W. 2019. MacroPCA: an All-in-One PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics* **61**:459–473 DOI 10.1080/00401706.2018.1562989.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* **10**:161–172 DOI 10.1038/nrg2522.

Karmodiya K, Pradhan SJ, Joshi B, Jangid R, Reddy PC, Galande S. 2015. A comprehensive epigenome map of *Plasmodium falciparum* reveals unique mechanisms

of transcriptional regulation and identifies H3K36me2 as a global mark of gene suppression. *Epigenetics & Chromatin* **8**:32 DOI 10.1186/s13072-015-0029-1.

Kassambra A. 2019. ggcorrplot: visualization of a correlation matrix using ggplot2. R package version 0.1.3. *Available at https://github.com/kassambara/ggcorrplot* (accessed on 6 May 2021).

Kensche PR, Hoeijmakers WAM, Toenhake CG, Bras M, Chappell L, Berriman M, Bártfai R. 2016. The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Research* **44**:2110–2124 DOI 10.1093/nar/gkv1214.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**:2204–2207 DOI 10.1093/bioinformatics/btq351.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**:907–915 DOI 10.1038/s41587-019-0201-4.

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research* **17**:691–707 DOI 10.1101/gr.5704207.

Kwok CK, Ding Y, Sherlock ME, Assmann SM, Bevilacqua PC. 2013. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Analytical Biochemistry* **435**:181–186 DOI 10.1016/j.ab.2013.01.008.

Lambros C, Vanderberg JP. 1979. Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *The Journal of Parasitology* **65**:418–420 DOI 10.2307/3280287.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25 DOI 10.1186/gb-2009-10-3-r25.

Lee VV, Judd LM, Jex AR, Holt KE, Tonkin CJ, Ralph SA. 2021. Direct Nanopore Sequencing of mRNA Reveals Landscape of Transcript Isoforms in Apicomplexan Parasites. *MSystems* **6**:e01081–20 DOI 10.1128/mSystems.01081-20.

Lewis MW, Li S, Franco HL. 2019. Transcriptional control by enhancers and enhancer RNAs. *Transcription* **10**:171–186 DOI 10.1080/21541264.2019.1695492.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094–3100 DOI 10.1093/bioinformatics/bty191.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079 DOI 10.1093/bioinformatics/btp352.

López-Ratón M, Rodríguez-Álvarez MX, Suárez CC, Sampedro FG. 2014. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software* **61**:1–36 DOI 10.18637/jss.v061.i08.

Lu XM, Batugedara G, Lee M, Prudhomme J, Bunnik EM, Le Roch KG. 2017. Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite *Plasmodium falciparum*. *Nucleic Acids Research* **45**:7825–7840 DOI 10.1093/nar/gkx464.

Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su X. 2007. cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics* **8**:255 DOI 10.1186/1471-2164-8-255.

Lubliner S, Keren L, Segal E. 2013. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Research* **41**:5569–5581 DOI 10.1093/nar/gkt256.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**:10–12 DOI 10.14806/ej.17.1.200.

Maruyama K, Sugano S. 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**:171–174 DOI 10.1016/0378-1119(94)90802-8.

Müller F, Tora L. 2014. Chromatin and DNA sequences in defining promoters for transcription initiation. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1839**:118–128 DOI 10.1016/j.bbagrm.2013.11.003.

Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, Itoh M. 2014. Detecting expressed genes using CAGE. In: Miyamoto-Sato E, Ohashi H, Sasaki H, Nishikawa J, Yanagawa H, eds. *Transcription factor regulatory networks*. *Methods in molecular biology*, New York: Springer New York, 67–85 DOI 10.1007/978-1-4939-0805-9_7.

Ohtake H, Ohtoko K, Ishimaru Y, Kato S. 2004. Determination of the capped site sequence of mRNA based on the detection of cap-dependent nucleotide addition using an anchor ligation method. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* **11**:305–309 DOI 10.1093/dnares/11.4.305.

Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Böhme U, Lemieux J, Barrell B, Pain A, Berriman M, Newbold C, Llinás M. 2010. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology* **76**:12–24 DOI 10.1111/j.1365-2958.2009.07026.x.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* **11**:1650–1667 DOI 10.1038/nprot.2016.095.

Petter M, Selvarajah SA, Lee CC, Chin WH, Gupta AP, Bozdech Z, Brown GV, Duffy MF. 2013. H2A.Z and H2B.Z double-variant nucleosomes define intergenic regions and dynamically occupy *var* gene promoters in the malaria parasite *Plasmodium falciparum*: dynamic occupation of *var* promoters by H2B.Z/H2A.Z. *Molecular Microbiology* **87**:1167–1182 DOI 10.1111/mmi.12154.

Polson HEJ, Blackman MJ. 2005. A role for poly(dA)poly(dT) tracts in directing activity of the *Plasmodium falciparum* calmodulin gene promoter. *Molecular and Biochemical Parasitology* **141**:179–189 DOI 10.1016/j.molbiopara.2005.02.008.

Promworn Y, Kaewprommal P, Shaw PJ, Intarapanich A, Tongsima S, Piriyapongsa J. 2017. ToNER: a tool for identifying nucleotide enrichment signals in feature-enriched RNA-seq data. *PLOS ONE* **12**:e0178483 DOI 10.1371/journal.pone.0178483.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842 DOI 10.1093/bioinformatics/btq033.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**:24–26 DOI 10.1038/nbt.1754.

Ruiz JL, Tena JJ, Bancells C, Cortés A, Gómez-Skarmeta JL, Gómez-Díaz E. 2018. Characterization of the accessible genome in the human malaria parasite *Plasmodium falciparum*. *Nucleic Acids Research* **46**:9414–9431 DOI 10.1093/nar/gky643.

Russell K, Hasenkamp S, Emes R, Horrocks P. 2013. Analysis of the spatial and temporal arrangement of transcripts over intergenic regions in the human malarial parasite *Plasmodium falciparum*. *BMC Genomics* **14**:267 DOI 10.1186/1471-2164-14-267.

Sachs MC. 2017. plotROC: a tool for plotting ROC curves. *Journal of Statistical Software* **79**:1–19 DOI 10.18637/jss.v079.c02.

Schmidt WM, Mueller MW. 1996. Controlled Ribonucleotide Tailing of cDNA ends (CRTC) by terminal deoxynucleotidyl transferase: a new approach in PCR-Mediated analysis of mRNA sequences. *Nucleic Acids Research* **24**:1789–1791 DOI 10.1093/nar/24.9.1789.

Shaw PJ, Kaewprommal P, Piriyapongsa J, Wongsombat C, Yuthavong Y, Kamchonwongpaisan S. 2016. Estimating mRNA lengths from *Plasmodium falciparum* genes by Virtual Northern RNA-seq analysis. *International Journal for Parasitology* **46**:7–12 DOI 10.1016/j.ijpara.2015.09.007.

Shaw PJ, Ponmee N, Karoonuthaisiri N, Kamchonwongpaisan S, Yuthavong Y. 2007. Characterization of human malaria parasite *Plasmodium falciparum* eIF4E homologue and mRNA 5′ cap status. *Molecular and Biochemical Parasitology* **155**:146–155 DOI 10.1016/j.molbiopara.2007.07.003.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research* **27**:491–499 DOI 10.1101/gr.209601.116.

Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nature Communications* **10**:3359 DOI 10.1038/s41467-019-11272-z.

Spurek P, Kamieniecki K, Tabor J, Misztal K, Śmieja M. 2017. R Package CEC. *Neurocomputing* **237**:410–413 DOI 10.1016/j.neucom.2016.08.118.

SRA Toolkit Development Team. 2014. SRA Toolkit. *Available at https://ncbi.github.io/sra-tools/* (accessed on 6 May 2021).

Tabor J, Spurek P. 2014. Cross-entropy clustering. *Pattern Recognition* **47**:3046–3059 DOI 10.1016/j.patcog.2014.03.006.

Tang AD, Soulette CM, Van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020a. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications* **11**:1438 DOI 10.1038/s41467-020-15171-6.

Tang J, Chisholm SA, Yeoh LM, Gilson PR, Papenfuss AT, Day KP, Petter M, Duffy MF. 2020b. Histone modifications associated with gene expression and genome accessibility are dynamically enriched at *Plasmodium falciparum* regulatory sequences. *Epigenetics & Chromatin* **13**:50 DOI 10.1186/s13072-020-00365-5.

Toenhake CG, Fraschka SA-K, Vijayabaskar MS, Westhead DR, Van Heeringen SJ, Bártfai R. 2018. Chromatin accessibility-based characterization of the gene regulatory network underlying *Plasmodium falciparum* blood-stage development. *Cell Host & Microbe* **23**:557–569 DOI 10.1016/j.chom.2018.03.007.

Trotman JB, Schoenberg DR. 2019. A recap of RNA recapping. *Wiley Interdisciplinary Reviews: RNA* **10**:e1504 DOI 10.1002/wrna.1504.

Van Lin LH, Pace T, Janse CJ, Birago C, Ramesar J, Picci L, Ponzi M, Waters AP. 2001. Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. *Nucleic Acids Research* **29**:2059–2068 DOI 10.1093/nar/29.10.2059.

Watanabe J, Sasaki M, Suzuki Y, Sugano S. 2002. Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene* **291**:105–113 DOI 10.1016/S0378-1119(02)00552-8.

Weiss B, Curran JA. 2015. CAP+ selection: a combined chemical–enzymatic strategy for efficient eukaryotic messenger RNA enrichment via the 5′ cap. *Analytical Biochemistry* **484**:72–74 DOI 10.1016/j.ab.2015.04.039.

Wickham H. 2009. *ggplot2: elegant graphics for data analysis*. New York: Springer DOI 10.1007/978-0-387-98141-3.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, Sadowski N, Holmes N, De Jesus JG, Jones KL, Soulette CM, Snutch TP, Loman N, Paten B, Loose M, Simpson JT, Olsen HE, Brooks AN, Akeson M, Timp W. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods* **16**:1297–1305 DOI 10.1038/s41592-019-0617-2.

World Health Organization. 2019. World malaria report 2019. Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO.

Wu X, Bartel DP. 2017. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Research* **45**:W534–W538 DOI 10.1093/nar/gkx323.

Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, Corrêa IR, Guan S. 2019. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *Journal of Biological Chemistry* **294**:18220–18231 DOI 10.1074/jbc.RA119.010676.

Xu C, Park J-K, Zhang J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLOS Biology* **17**:e3000197 DOI 10.1371/journal.pbio.3000197.

Yin S, Fan Y, He X, Wei G, Wen Y, Zhao Y, Shi M, Wei J, Chen H, Han J, Jiang L, Zhang Q. 2020. The cryptic unstable transcripts are associated with developmentally regulated gene expression in blood-stage *Plasmodium falciparum*. *RNA Biology* **17**:828–842 DOI 10.1080/15476286.2020.1732032.

Zhao Y, Garcia BA. 2015. Comprehensive catalog of currently documented histone modifications. *Cold Spring Harbor Perspectives in Biology* **7**:a025064 DOI 10.1101/cshperspect.a025064.

Zhao X, Valen E, Parker BJ, Sandelin A. 2011. Systematic clustering of transcription start site landscapes. *PLOS ONE* **6**:e23409 DOI 10.1371/journal.pone.0023409.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* **30**:892–897 DOI 10.2144/01304pf02.