Plasmodium Falciparum Regulates Transcriptional Output by Modulating The Activity of RNA Polymerase II Rather Than Its Recruitment

Ragini Rai Khandelia
Nanyang Technological University
Aarthi Mohan
Nanyang Technological University
Lei Zhu
Nanyang Technological University
Jie Zheng
Nanyang Technological University
Mark Stephen Featherstone (msfeatherstone@gmail.com)
Nanyang Technological University https://orcid.org/0000-0003-1576-046X
Zbynek Bozdech
Nanyang Technological University

Research article

Keywords: Plasmodium falciparum, malaria, RNA polymerase II, transcription, gene regulation, C-terminal domain, serine phosphorylation, CTD

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74869/v1

Abstract

Background The asexual phase of the unicellular eukaryotic parasite *Plasmodium falciparum* takes place in the human red blood cell (RBC). During the 48 h between RBC infection and the release of new progeny – the intraerythrocytic developmental cycle (IDC) – transcripts levels for most of the parasite's ~5,772 genes are set by the integration of transcriptional and post-transcriptional processes. Transcription of eukaryotic protein coding genes is carried out by RNA polymerase II (RNAPII) whose functional states are reflected by the phosphorylation status of serine residues at positions 2 and 5 within the so-called heptad repeats within the C-terminal domain (CTD) of the catalytic subunit, RPB1; unphosphorylated or Ser5-phosphorylated RNAPII correspond to uninitiated and initiating states, while phosphorylation at Ser2 is associated with the processive elongating form of the polymerase. Transcriptional control over the IDC could impinge on RNAPII recruitment, initiation, and conversion to processive elongation. To distinguish the contribution of these potential regulatory steps to transcriptional control during infection, we used ChIP-seq and RNA-seq to determine the genome-wide distribution of three RNAPII phosphoisoforms as well as total RNAPII and correlated occupancy with mRNA levels across the three stages of the IDC.

ResultsWe find that most genes are occupied by RNAPII along the length of the coding region at all stages of the IDC, regardless of transcript output, with RNAPII accumulation of all forms strongly biased toward the coding region. Total RNAPII levels at a given gene are relatively constant across the IDC. By contrast, occupancy by the elongating form of the polymerase shows a strong positive correlation with mRNA abundance.

ConclusionsThese observations reveal that transcriptional activation during the IDC is not regulated primarily at the level of RNAPII recruitment, but rather through the conversion of RNAPII to its processive elongating form. Our results have implications for the expected roles of enhancers and transcription factors in the parasite and support a major role for the control of transcriptional pausing.

Background

In eukaryotes, the processes of gene expression include, but are not limited to, transcriptional initiation and elongation, RNA processing (capping, polyadenylation, splicing), RNA export, RNA-sequestration and stability, translation, post-translation modification, protein sequestration and degradation.

In most eukaryotes, the largest subunit of RNAPII, RPB1, harbors a C-terminal domain (CTD) composed of multiple heptad repeats of the consensus sequence (YSPTSPS). Importantly, the phosphorylation status of the serine residues at positions two (Ser2) and five (Ser5) correlates with the transcriptional status of the polymerase [1, 2]. RNAPII bearing an unphosphorylated CTD has a high affinity for the Mediator complex which retains the polymerase at the promoter in the pre-initiation complex (PIC). Phosphorylation of heptad repeats at position 5 (Ser5-P) decreases the affinity of RNAPII for the Mediator thereby leading to promoter escape. In metazoans, RNAPII then undergoes promoter-proximal pausing under the influence of elongation inhibitors. Release from pausing and entry into highly processive

elongation is concurrent with further phosphorylation of the heptad repeat at position 2 (Ser2/5-P). These phosphorylation states do not directly modify intrinsic RNAPII activity, but rather serve to couple stages in the transcriptional process to RNA processing and chromatin modification [1, 2]. For example, Ser5-P is required for recruitment of the capping enzyme complex needed to cap and protect the 5' end of the nascent mRNA chain; Ser5-P in yeast and Ser2-P in mammals promote interaction with the spliceosome; and Ser2-P in yeast facilitates interaction with the 3' end processing machinery [2]. Regardless, the different forms of the RNAPII CTD provide markers for enzymatically distinct states of the polymerase. The elongating Ser2-P modification is abundantly localized to the gene body. The Ser5-P phosphoisoform is broadly distributed over the gene body in yeast, but in metazoans displays a sharp concentration in the promoter-proximal region due to the action of elongation inhibitors resulting in RNAPII pausing. Complicating these profiles, the Ser2/5-P phosphorylation status can change as the polymerase traverses the gene body due to, for example, association with factors involved in transcription termination and 3' end processing [1, 2].

The *P falciparum* genome is one of the most A-T rich described, with A-T content as high as 90 to 95% in intergenic regions [3, 4]. The parasite is haploid throughout the infection of human tissues, providing excellent opportunities for genetic analysis of gene function. Gene organization is relatively dense, and expression patterns suggest that each gene comes under independent control, with little evidence for coordinated expression of multiple genes by a smaller number of regulatory elements. To the extent that they have been analyzed, potential transcriptional regulatory elements are located in the 5' flank within intergenic regions that range in size from a median of 680 bp for the smallest classes to greater than 1,900 bp for the largest [5]. While some studies demonstrate long 5' untranslated regions (UTRs) in *P falciparum* [5], the recent publication of genome-wide mapping of transcriptional start sites (TSS) across the IDC suggests that the majority of genes make use of broadly scattered TSS, the majority of which are close to the translational start codon, and with many genes harboring TSS well within the coding region [6].

P falciparum carries ~5,772 genes, greater than 80% of which are expressed during the IDC [7, 8]. In the course of the 48 h IDC, the parasite transitions through ring (0-16 h), trophozoite (17-32 h) and schizont stages (33-48 h; see Fig 1) culminating in the release of up to 32 merozoites equipped for the next round of infection. It follows that the differential control of gene expression is a pre-requisite to successful completion of the life cycle, and over 2,700 genes are indeed differentially expressed over the IDC as determined by microarray-based assessment of transcript levels [9-11]. Such analyses reveal that fully half of such transcripts peak just once with stage-specific timing in what has been termed "just in time" or "transcript to go" expression, though studies on translational regulation reveal a more nuanced situation [8, 10, 12-17].

The 12 subunits of RNAPII are well conserved in *P falciparum*. Relative to model eukaryotes, the largest subunit, RPB1, bears a reduced number of heptad repeats in the CTD whose consensus is highly related to those in other species (YSPTSPK) [18, 19]. We have successfully raised and used highly specific monoclonal antibodies to detect the unphosphorylated heptad repeat, the monophosphorylated Ser5-P

form and the diphosphorylated (Ser2/5-P) form of RNAPII in *P falciparum*, demonstrating that the heptad motifs are modified in the parasite as they are in other organisms. While there has been discussion as to the possible divergence of function of CTD phosphorylation in the parasite ([18-20] and reviewed in Rai 2014 [21], the simplest understanding is that the different forms are associated with the same enzymatic states of RNAPII seen in other model systems. Thus, the unphosphorylated, Ser5-P and Ser2/5-P heptad repeats would mark the pre-initiating, initiating, and elongating forms of the polymerase, respectively.

In our previous study, we used these phosphoisoform-specific monoclonal antibodies to map the differential association of RNAPII to the entire parasite genome at six 8-hour time intervals across the IDC via chromatin immunoprecipitation (ChIP) followed by the probing of pan-genomic microarrays (ChIP-chip) [21]. We made the striking observation that most genes differentially expressed across the IDC are starkly divided in the timing of RNAPII occupancy [21]. While there is a large group of genes bound by RNAPII early in the IDC, a complementary set of genes is bound by RNAPII in the late stage. This is the first and only report to reveal that the expression of the *P falciparum* genome is mechanistically divided between early and late stages. However, this study provides almost no details on the distribution of RNAPII across the length of the gene body and in intergenic regions; nor does it distinguish between mechanisms controlling RNAPII activity such as recruitment vs pausing and elongation.

In the present study, we used chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) to investigate genome-wide binding of RNAPII phosphoisoforms during the three developmental stages of the intraerythrocytic developmental cycle (IDC). In addition to the three antibodies described above, we included a commercially available antibody that recognizes a universally conserved epitope toward the N-terminus of RPB1. This antibody has the advantage of recognizing all RNAPII molecules regardless of the phosphorylation state of the CTD, thereby providing a measure of total RNAPII occupancy [22].

We find that most genes are occupied by RNAPII along the length of the coding region at all stages of the IDC, regardless of detectable transcript output. Importantly, the distribution of total RNAPII does not change substantially upon changes in transcript output determined in parallel by RNA-seq. Rather, it is changes in the level of the Ser2/5-P phosphoisoform that strongly correlate with mRNA expression levels. We observe accumulations of the CTD (pre-initiating) and Ser5-P (initiating) forms of RNAPII in the presumptive promoter regions of a subset of genes in schizonts which become highly expressed in rings, consistent with a role for promoter-proximal pausing of RNAPII in the transcriptional regulation of some genes. Last, phosphoisoform distribution at the gene 3’ end suggests that RNAPII slows and/or pauses to accommodate 3’ end processing and transcriptional termination.

Overall our data suggest that *P falciparum* does not generally control mRNA output at the level of RNAPII recruitment, but rather through the regulation of RNAPII activity as reflected in the phosphorylation status of the CTD. Our results show that transcriptional control in the parasite shares mechanisms common to other eukaryotes, and have implications for enhancer function, the role of transcription factors, and the importance of transcriptional pausing during the IDC.

Results

Transcript levels are the integrated outcome of multiple distinct processes including transcription, splicing, nuclear export and stability, each of which is influenced by a variety of regulatory mechanisms. Our aim is to assess the relative contribution of RNAPII recruitment and activity to mRNA output during RBC infection. To accomplish this goal, we evaluated the occupancy of RNAPII and three CTD phosphoisoforms across the *P. falciparum* genome at multiple developmental stages of the IDC. In a previous study, we used ChIP and DNA microarray technology (ChIP on chip) to show that there are two temporal domains of RNAPII occupancy, with about half of protein-coding genes (~1800) associating with RNAPII early in the IDC and remaining genes (~1700) associating late [21]. To assess RNAPII occupancy across the entire transcription unit, we now use the same custom-generated RNAPII antibodies to perform ChIP at three time points across the IDC followed by next-generation sequencing (ChIP-seq). These mouse monoclonal antibodies are specific to the unphosphorylated CTD (anti-CTD) associated with the pre-initiating polymerase; the heptad repeat mono-phosphorylated at serine 5 (anti-Ser5-P) associated with the initiating polymerase; and the heptad repeat double phosphorylated at serine positions 2 and 5 (anti-Ser2/5-P) associated with the processive, elongating form of the polymerase [21]. In addition, we employed the commercially available N20 antibody against a universally conserved epitope at the N-terminus of RNAPII which is not known to undergo post-translational modification; N20 should therefore recognize all forms of the polymerase regardless of the phosphorylation state of the CTD [22, 23].

These antibodies were used in ChIP-seq to characterize in detail the genome-wide occupancy of RNAPII isoforms at three stages of the IDC: the ring stage at 12-16 hours post invasion (hpi), the trophozoite stage (24-28 hpi) and schizont stage (36-40 hpi). Occupancy by RNAPII as revealed by the four antibodies was visualized in two ways: as enrichment relative to input along the length of the transcription unit (determined using the bamCompare algorithm and shown as profile plots and heatmaps), and as discrete peaks of RNAPII occupancy called by the MACS2 algorithm [24, 25]. To enable correlation of RNAPII occupancy with transcript abundance, we also performed RNA-seq on the identical parasite culture.

The distribution of RNAPII across the transcription unit is stage-specific

The genome-wide read densities for RNA-seq, ChIP-seq, and input chromatin were plotted and visualized in a genome browser format [26]. Such a representation allows ready visualization of read density and is used here to illustrate some key points regarding the results elaborated later in this report. The genome browser view of three contiguous genes in a ~27 kb region of chromosome 14 is shown in figure 1. The far leftward gene, PF3D7_1410200, encodes CTP synthetase and is known to be expressed early in the IDC, consistent with the need to boost CTP pools for the biosynthetic needs of later stages. The far rightward gene, PF3D7_1410400, encodes rhoptry associated protein 1, a so-called “invasion” gene required for the infectivity of newly released merozoites. Not surprisingly, transcripts for this gene are known to accumulate late in the IDC. The middle gene, PF3D7_1410300, encodes a putative WD-repeat-containing protein, and has been observed to maintain low and relatively consistent transcript levels

across the IDC. The expected stage-associated transcript levels for each of these genes are borne out by our own RNA-seq results as plotted immediately below the positions of the gene bodies (**Fig 1**).

For simplicity, only the ChIP-seq results and input controls for two informative antibodies have been shown – those corresponding to the N20 antibody which recognizes all forms of RNAPII (total RNAPII) and the antibody recognizing the doubly phosphorylated heptad repeat (RNAPII-Ser2/5-P) associated with the actively elongating form of the polymerase. First, RNAPII, regardless of the isoform detected, is concentrated within the gene body, i.e. between the start and stop codons of the open reading frame. Second, we find that the distribution of total RNAPII (N20, upper set of green tracks) over the three genes is highly similar, regardless of the amount of transcript produced for a given gene or the stage at which it is assayed. By contrast, the far leftward gene (PF3D7_1410200; CTP synthetase) is associated with increased transcript levels at the ring and trophozoite stages and increased reads corresponding to the elongating form of RNAPII (RNAPII-Ser2/5-P; upper set of purple tracks). Conversely, the far rightward gene (PF3D7_1410400; RAP1) is associated with increased transcripts at the schizont stage and increased reads for RNAPII-Ser2/5-P at that same point in the IDC. (The plot for RNAPII-Ser2/5-P is truncated along the y-axis whose maximum was set to allow clear visualization of the weaker signals present over the other two genes.) Last, the middle gene (PF3D7_1410300; putative WD-repeat protein) yields almost no transcripts at any stage by RNA-seq, and displays no stage-specific change in the level of ChIP-seq reads obtained with either antibody. Thus, for the far leftward and rightward genes, transcript abundance is stage-specifically correlated with variation in the level of RNAPII-Ser2/5-P signal, but not that of total RNAPII. Despite significant occupancy by the elongating form of RNAPII, little to no transcript is detected for the middle gene, PF3D7_1410300, suggestive of post-transcriptional control.

Initial observations such as those presented in figure 1 would be consistent with a role for RNAPII activity (Ser2/5-P), as opposed to RNAPII recruitment (N20), in regulating the levels of transcript output. We therefore undertook an assessment of the correlation between the levels of transcript output and occupancy by each of the RNAPII isoforms (CTD, Ser5-P, Ser2/5-P) along with total RNAPII (N20).

ChIP-seq data was used to plot the occupancy of the four RNAPII isoforms across the transcription units of protein-coding nuclear genes at all three stages of the IDC. For each RNAPII isoform at each stage of the IDC, RNAPII enrichment was determined at each of 3,000 bins spanning the 5' flanking region, coding region and 3' flanking region. Genes showing RNAPII enrichment of at least 2-fold (before $\log_2$ transformation) in at least one bin defined gene sets specific for an RNAPII isoform and stage of the IDC. The averaged plots of RNAPII occupancy for each gene set were then determined (**Fig 2A**). Further details regarding filters and gene selection are provided in the figure legends, Materials and Methods and Additional File 1.

The distribution of total RNAPII as revealed by the N20 antibody provides the level of RNAPII recruitment at a given locus against which changes in RNAPII phosphorylation state can be compared. Total RNAPII shows two distinct profiles (**Fig 2A**). The first profile, exhibited in rings and trophozoites, is characterized by elevated levels of RNAPII over the gene body (GB) spanning the start codon (ATG) to the stop codon

(STOP). The second profile, found in schizonts, is distinct and shows a relatively uniform distribution along the entire length of the transcription unit from 1 kb upstream of the start codon to 1kb downstream of the stop codon, with a moderate decrease in overall abundance from 5’ to 3’. The observation of these distinctly different profiles for total RNAPII is consistent with the conclusions of our earlier study suggesting that different transcriptional mechanisms govern gene expression at early vs late times of the IDC [21].

Deviations from the total RNAPII (N20) pattern by the three RNAPII isoforms under study reveal the distribution of CTD phosphorylation (or lack thereof) and thereby the activity of RNAPII along the transcription unit. In rings, the unphosphorylated (CTD) and monophosphorylated (Ser5-P) isoforms display rather flat profiles, indicating that these isoforms are more abundant in the 5’ and 3’ flanking regions than in the gene body. This would be reasonable if a significant fraction of heptad repeats were doubly phosphorylated while the polymerase is traversing the gene body, consistent with the association of the Ser2/5-P modification with the processive elongating form of the polymerase. Supporting this interpretation, the distribution of the Ser2/5-P modification is exaggerated by comparison to that of total RNAPII, with highly elevated levels over the gene body and depleted levels in the 5’ flank and, to a lesser extent, the 3’ flank.

The distribution of the CTD and Ser5-P isoforms in schizonts is distinct from those observed in rings and trophozoites, as was observed for total RNAPII (N20). Levels of the unphosphorylated (CTD) and monophosphorylated (Ser5-P) forms are noticeably depleted in the gene body relative to the 5’ and 3’ flanks. By contrast, the Ser2/5-P isoform is unique at this stage in its persistent accumulation over the gene body relative to the flanks. The simplest interpretation of this data is that the depressed level of N20 in the gene body is reflected in lower levels of RNAPII phosphoisoforms along the gene body as well, and that transcriptional output is driven by RNAPII activity rather than recruitment.

In conclusion, a global view of expressed protein-coding genes reveals exaggerated levels of the Ser2/5-P modification over the gene body relative to total RNAPII (N20) in rings and schizonts, though not in trophozoites. Fluctuations in the levels of Ser2/5-P modification without concomitant changes in the level of total RNAPII would be consistent with a model in which transcriptional activity is modulated via the control of RNAPII enzymatic activity rather than the levels of RNAPII recruitment.

Transcript output at a given gene correlates with the abundance of the elongating form of RNAPII but not total polymerase

To refine the association between RNAPII distribution and transcript output, we defined a set of 4,990 nuclear, non-antigenic protein-coding genes and subdivided them into three classes based on their relative transcript levels. For each stage of the IDC, genes were ranked in descending order of transcript abundance and partitioned into three classes: High, Medium, and Low, with the number of genes in each class being 1,663 or 1,664. We then plotted the enrichment of individual RNAPII isoforms and total RNAPII for each of the three gene classes in each of the three developmental stages under study (**Fig. 2B**).

The overall distribution of all forms of RNAPII is highly similar to that described above, but now reveals a striking association with gene expression levels at the ring stage. Those genes with the highest transcript levels exhibit up to 2-fold enrichment of RNAPII occupancy in the ORF compared to genes with medium and low levels of expression. Strikingly, the positive association between Ser2/5-P and transcription is exclusive to the gene body, while occupancy in the 5' and 3' flanking regions follows a reverse trend. We also note the gradually increasing Ser2/5-P signal toward the 3' end of the ORF, possibly associated with the terminal stages of elongation and the recruitment of 3' end processing factors [2] (**Fig 2B**).

In schizonts, there is only a modest increase in Ser2/5-P signal associated with more highly expressed genes; however, the Ser2/5-P signal in both rings and schizonts remains elevated relative to that of total RNAPII (N20). Once again, these observations suggest that *P falciparum* does not achieve higher transcript output by enhanced recruitment of RNAPII to the promoter, but rather via control of RNAPII activity. They are also consistent with distinct transcriptional mechanisms operating at early vs late times of the IDC.

The above results reveal a positive correlation between Ser2/5-P occupancy and transcript abundance. To more rigorously assess the association between RNAPII occupancy and transcript output, we used the MACS2 algorithm [24, 25] to identify localized peaks of RNAPII enrichment. MACS2 derives statistically significant peaks of ChIP-seq enrichment occurring over defined windows measured in base pairs, and therefore provides a complementary method for the identification of genes displaying enriched RNAPII occupancy.

We used MACS2 to classify genes according to their association with one or more peaks of RNAPII binding. Transcript abundance for each of these classes was then compared for each of the three stages of the IDC (**Fig 3A**). Strikingly, enhanced transcript abundance is consistently associated with the presence of peaks of Ser2/5-P occupancy, but not unphosphorylated (CTD) or Ser5-P isoforms. (The statistical significance between the expression levels of different groups are provided in **Table S1**.) These results strongly support a role for increased transcription, necessarily mediated by the Ser2/5-P isoform, in achieving elevated mRNA levels at all stages of the IDC.

While these results reveal that higher levels of Ser2/5-P are associated with increased mRNA output, they do not implicate increased recruitment of RNAPII in achieving this outcome. Indeed, the findings of figures 1 and 2 suggest that overall recruitment of RNAPII, as revealed by the N20 antibody, does not vary markedly despite large changes in transcriptional output. We therefore used peak calling to compare directly the correlation between mRNA levels and total RNAPII vs. Ser2/5-P abundance across the IDC (**Fig 3B**).

For each stage, we defined four gene sets: those genes associated only with peaks of the Ser2/5-P isoform; those associated only with peaks of total RNAPII (N20); those associated with both; and those associated with neither. For each of these four classes, transcript abundance was compared and tested for statistical significance (**Fig 3B**; **Table S2)**. The results clearly show that increased transcript output is strongly correlated with enrichment of the elongating Ser2/5-P isoform but not with enrichment of total

RNAPII (N20). We conclude that *P falciparum* does not achieve higher mRNA output by increased recruitment of RNAPII, but rather through the conversion of RNAPII to its active elongating form. RNAPII distribution was independently confirmed in rings and schizonts by ChIP-qPCR using the same immunoprecipitated chromatin samples used for ChIP-seq (**Fig S1**).

A recent study reported on genome-wide measurements on transcription and mRNA decay rates across the IDC in the 3D7 parasite strain [27]. We compared our current data identifying genes having stage-specific peaks of Ser2/5-P occupancy in the T996 parasite strain with the data set from this published study (**Table S3**). Interestingly, genes having ring-stage-specific peaks of the Ser2/5-P isoform show highly statistically significant overlap with those genes that are maximally transcribed in early and late rings. Reciprocally, genes associated with schizont-stage peaks of Ser2/5-P display highly statistically significant overlap with those genes that are maximally transcribed in this same stage. By contrast, genes having peaks of total RNAPII (N20, but not Ser2/5-P) or no peaks of either form do not display a stage-specific positive correlation with transcription rate. The overall congruity of these results strongly supports the association of the Ser2/5-P isoform with elevated transcriptional output.

The accumulation of the elongating form RNAPII in the gene body and 3' flanking region correlates with transcript output

The results above demonstrate a positive correlation between the Ser2/5-P signal and transcript output. The Ser2/5-P isoform is routinely detected at lower levels in the flanking regions and is highest toward the 3' end of the gene body. To test for an association between the location of Ser2/5-P signal along the transcription unit and the level of transcript output, we correlated the position of peaks of Ser2/5 abundance called by the MACS2 algorithm with our RNA-seq data at each stage of the IDC (**Fig 4**). Genes first filtered by peak calling as for Fig 3 were then sorted based on Ser2/5-P peak position within the transcription unit followed by plotting of mRNA expression levels (RNA-seq). A large fraction of genes does not display RNAPII peaks as defined by the MACS2 algorithm but are shown at the far right of each panel for comparison. In addition, those genes having peaks exclusively in the 5' plus 3' flanks, or in all three regions at once, are poorly represented and likely to suffer from the effects of small sample size but are included for completeness.

At all stages, genes having Ser2/5-P peaks in the GB in conjunction with the absence of peaks in the 5' flank show the highest transcript output. In schizonts, concurrent peaks in the GB and 3' flank are also correlated with high mRNA levels. By contrast, the presence of peaks of Ser2/5-P signal in the 5' flank is consistently associated with lower transcript output. These results are highly statistically significant **(Table S4)**. We conclude that peaks of Ser2/5-P signal in the gene body and 3' flank region are correlated with elevated mRNA levels.

ChIP-seq reads obtained with the antibody to the Ser2/5-P isoform of RNAPII are most abundant toward the 3' end of the coding region, and peaks of the Ser2/5-P signal in the gene body and 3' flank are associated with higher transcript output. By contrast, ChIP-seq reads from the N20 isoform are only modestly elevated at the 3' end, revealing that the increased reads attributable to the Ser2/5-P isoform are

not simply due to increased RNAPII abundance. To investigate the relevance of this observation, we assessed the correlation between the strength of the Ser2/5-P signal at the gene 3' end vs. transcript output.

We defined a 3' region extending from 1 kb upstream to 1 kb downstream of the translational stop codon (STOP) and used K-means clustering to identify four patterns in the accumulation of the Ser2/5-P isoform at each of the three stages of the IDC (**Fig 5A**). Cluster 1 genes show the broadest and generally highest levels of Ser2/5-P across the 3' region spanning the stop codon. Further, for both rings (481 genes) and schizonts (712 genes), genes in cluster 1 showed 3' enrichment of Ser2/5-P at two distinct locations – one just before the STOP codon and another within 1 kb downstream. This second site of Ser2/5-P accumulation seems to overlap with the average distribution of the polyadenylation site (PAS) in *P. falciparum,* since most PAS fall within 1 kb downstream of the stop codon yielding a 3' UTR with an average length of 523 nt [7]. Increased ChIP-seq signal at the 3' end of Cluster 1 genes was confirmed independently by ChIP-qPCR using the same immunoprecipitated chromatin as used for ChIP-seq (**Fig S2**).

To assess the correlation between the distribution of the Ser2/5-P isoform and transcript output, we plotted the expression level of the mRNA derived from genes in each of the 4 clusters. At all three stages and especially in rings, genes in Cluster 1 show transcript levels that are significantly higher than for the other clusters (**Fig 5B, Table S5**). In rings, Cluster 2 genes are also expressed at significantly higher levels than genes in Clusters 3 and 4. These results extend and support the positive correlation between Ser2/5-P occupancy and transcriptional output revealed in the analyses described above.

Cluster 1 genes at each stage are defined by the broad accumulation of Ser2/5-P across the gene 3' end. We used expression heat maps to address how the expression of Cluster 1 genes defined for a given stage varied across the IDC. The results show that for Cluster 1 genes showing maximal Ser2/5-P signal in rings, a large majority shows maximum expression in that same stage (**Fig S3**). Likewise, a majority of Cluster 1 genes having maximal ChIP-seq reads in schizonts have their most abundant expression at this same stage. By contrast, trophozoite-stage Cluster 1 genes are maximally expressed either in rings or schizonts, with a smaller fraction showing highest expression in trophozoites. Functional annotation of Cluster 1 genes shows that most of them are associated with stage-specific pathways or processes.

While the Ser2/5-P isoform accumulates preferentially over the gene 3' end in schizonts, a different pattern is observed for two other isoforms. In schizonts, the CTD and Ser5-P isoforms of RNAPII are more abundant at the 5' ends of genes relative to total RNAPII (Fig 2). To assess whether this profile is associated with transcript output, we used k-means clustering to define four distinct distributions of CTD and Ser5-P signals in a region spanning 1 kb up- and downstream of the ATG in schizonts (**Fig S4A**). Cluster 1 genes had the highest accumulation of each isoform in the 1 kb region upstream of the ATG that would be expected to include the presumptive promoter for most genes [6]. The genes of other clusters had lower levels of RNAPII occupancy and in more restricted regions located either upstream

(Clusters 3 and 4) or downstream (Cluster 2) of the ATG. There is significant overlap between the four gene groups showing CTD (orange circle) and Ser5-P (blue circle) occupancy around the ATG (**Fig. S4B**).

Clusters 1, 3 and 4 show higher CTD and Ser5-P signal in the presumptive promoter region upstream of the ATG in schizonts. We assessed the mRNA levels for these three clusters in rings and schizonts. Interestingly, transcript output for the three gene sets is significantly higher than for genes of cluster 2 and is most evident in rings (**Fig S4C, Table S6**), suggesting that CTD and Ser5-P signal in schizonts anticipates higher expression in rings of the subsequent cycle.

A heat map generated to view the expression profile through the IDC of the genes in Cluster 1 shows that these genes have higher expression levels in rings (**Fig S4D**). Functional enrichment analysis of genes in Cluster 1 based on the Gene Ontology (GO) terms showed that the genes in this cluster are associated with transcriptional, translational and metabolic processes (**Fig S4D)**. Taken together these data show that at the schizont stage the CTD and Ser5-P RNAPII phosphoisoforms are over-represented at the presumptive promoter regions of specific gene groups, possibly in preparation for their expression in the next round of infection.

Discussion

The *P falciparum* genome encodes a wide complement of general transcription factors but lacks representatives of many specific transcription factors present in other eukaryotes. Additionally, the potential to encode numerous homologs of RNA metabolizing enzymes and the demonstrated role of RNA stability in controlling gene expression [27, 28] have raised questions regarding the extent of transcriptional control in *P falciparum* gene expression. To address the role of direct transcriptional control, and to distinguish between recruitment and activity of RNAPII in this process, we raised highly specific mouse monoclonal antibodies against different phosphoisoforms of the CTD heptad repeat, used these antibodies in ChIP-seq of synchronous parasite cultures at three stages of the IDC, and correlated the occupancy of phosphoisoforms and total RNAPII with transcript levels determined in parallel by RNA-sEq. Our results establish that transcriptional output is strongly correlated with occupancy by the Ser2/5-P isoform of RNAPII which is the isoform associated with processive transcriptional elongation in all eukaryotes studied. Additionally, elevated CTD and Ser5-P signals in schizonts anticipate higher expression in rings. By contrast, transcriptional output does not correlate with the total level of RNAPII at the gene, demonstrating that transcriptional control is at the level of RNAPII activity and not recruitment. These results support a model in which the conversion of RNAPII to the processive elongating form plays a major role in the control of gene expression in *P falciparum*.

The distribution of the Ser5-P phosphoisoform of RNAPII in *P falciparum* can be inferred by reference to total RNAPII (N20), revealing that the Ser5-P mark is broadly though modestly elevated across hundreds of base pairs of the 5’ and 3’ flanking regions relative to the gene body. Transcriptional pausing in metazoans is reflected by promoter-proximal accumulation of the Ser5-P phosphoisoform. In ChIP-seq experiments, this is evidenced by a tight peak of read density in the promoter-proximal region or just

upstream of the first (+ 1) nucleosome, contrary to what we observe for *P falciparum*. In this sense, the distribution of the Ser5-P phosphoisoform during the IDC is more similar to budding yeast which do not display transcriptional pausing, raising the possibility that pausing is not employed by the parasite. Although an absence of pausing would be consistent with the fact that neither *Plasmodium* nor budding yeast encodes the transcription elongation inhibitor NELF, transcriptional pausing has been documented in fission yeast despite the natural absence of NELF, and in other organisms under conditions of NELF depletion [29, 30]. Additionally, the parasite does encode the elongation regulator DSIF and the P-TEFb kinase that converts DSIF from an inhibitor to an activator of elongation in other organisms. In fact, the application to *P falciparum* of GRO-seq, a modified nuclear run-on approach, has provided evidence for transcriptional pausing in the vicinity of the ATG [31]. This is consistent with observations in the parasite that a large fraction of TSS map to the ATG-adjacent region [6] and that nucleosomes, which are an impediment to traversing polymerase, are generally positioned near TSS [32]. Additional clusters of TSS are scattered up to 1 kb upstream of the ATG [6] which, in conjunction with nucleosome placement at TSS, may create an extensive domain for pausing to occur and thereby account for the wide distribution of the Ser5-P phosphoisoform found here. Further, our observation that 5’ accumulation of Ser5-P in schizonts is positively correlated with transcript output in rings leads to a model in which RNAPII is paused at certain genes at the schizont stage in anticipation of ring-stage transcription. We cannot assess to what extent unphosphorylated heptad repeats and Ser5-P repeats are present on the same RPB1 C-terminal domain, but the fact that the CTD signal mirrors what we observe for Ser5-P would be consistent with a large fraction of the CTD and Ser5-P signal as being present on the same polymerase molecules. Alternatively, both pre-initiating and initiating RNAPII may be poised in schizonts for ultimate conversion to the Ser2/5-P form in rings.

Levels of total RNAPII and the Ser2/5-P phosphoisoform rise sharply beginning a few hundred bp upstream of the ATG and reach an inflection point immediately downstream of the ATG in keeping with a concentration of TSS adjacent to the start codon. Once into the coding region, Ser2/5-P levels accumulate more gradually toward the gene 3’ end. Downstream of the stop codon, CTD modifications attract the 3’ end processing machinery followed by loss of Ser2 phosphorylation likely due to dephosphorylation as seen in other species [1, 2].

While the current study was underway, two reports used alternative methods to assess the rate of transcription at all genes across the IDC. One study employed nascent RNA capture and hybridization to custom microarrays [27] to show that active transcription takes place throughout the IDC and to define non-overlapping gene sets having maximal transcription rates at one of five timepoints spanning the blood stage of infection. Strikingly, gene sets that we defined as having stage-specific peaks of the Ser2/5-P phsophoisoform show maximal overlap with these published gene sets [27] at those same stages (Table S3). For example, there is a high degree of overlap between that gene set showing an increased number of peaks of Ser2/5-P occupancy in schizonts and the gene set defined by others as being maximally transcribed in schizonts [27]. By contrast, total RNAPII accumulation does not correlate with transcription rate. This supports our model in which transcriptional output in the parasite is controlled at all stages of the IDC by the activity of RNAPII, but not by polymerase recruitment.

A separate study used global run-on sequencing (GRO-seq) to quantitate nascent RNA at multiple timepoints across the IDC with results suggesting that transcription is massively activated in trophozoites almost exclusively and that this is accomplished through the release of paused RNAPII [31]. While these results differ in important respects from our own, they are consistent with our findings that RNAPII is broadly present across the genome throughout the IDC, that total RNAPII is concentrated in the gene body in trophozoites and that the activation of RNAPII is a major regulatory step in the control of parasite transcription. We further note that genes displaying stage-specific elevation of the Ser2/5-P isoform overlap with gene sets transcribed in a stage-specific manner as defined by this group (data not shown) [31]. We do not know why results obtained via GRO-seq otherwise differ from our own results and those of others [27] but could be related in part to the run-on methodology which provokes the *in vitro* release of paused RNA polymerases.

A synthesis of our own data with that of others leads to the following model for the control of transcription rate in *P falciparum*. Recruitment and loss of RNAPII at most genes are coordinated to maintain a relatively constant total RNAPII steady-state profile across the transcription unit. Two RNAPII profiles emerge, one characteristic of rings and trophozoites, and the other of schizonts, implying structural and mechanistic differences in early vs late gene expression. Increased transcription is accomplished by the conversion of RNAPII to its processive elongating form, either by rapid conversion of pre-initiating RNAPII (no heptad repeat phosphorylation) to its initiating (Ser-5-P) and processive (Ser2/5-P) forms, or by conversion of the paused Ser5-P phosphoisoform to the processive elongating form. It follows that stage-specific enhancers would not play a regulatory role in the differential recruitment of RNAPII to the promoter since RNAPII levels remain constant at most genes. Rather, enhancers and the specific transcription factors that act through them would be expected to control the phosphorylation status of DSIF and Ser2 of the heptad repeat as has been established for the Myc family of transcription factors in metazoans [33]. This model would not necessarily apply to those parasite genes known to exist in strong on/off states such as the *var* gene family, and we have excluded the antigenic variant gene families *var*, *rifin* and *stevor* from our analysis. It would be of interest to employ the same antibodies used here in ChIP-seq analysis of clonal parasite populations each expressing different members of these gene families to better understand the role of RNAPII in antigenic variation.

Methods

Parasite culture

Asexual blood stages of *Plasmodium falciparum* T996 strain were cultured as described previously [21]. Giemsa staining of thin blood smears was carried out to monitor the parasite's developmental stage and parasitemia [34]. The cultures were synchronized with multiple rounds of D-sorbitol treatment [35]. Highly synchronized cultures were harvested for chromatin immunoprecipitation and RNA-seq analyses.

Antibodies

Highly specific, non-cross-reacting mouse monoclonal antibodies against the unphosphorylated, Ser5 phosphorylated and Ser2/5 phosphorylated heptad repeat (YSPTSPK) were raised as previously described [21]. The N20 antibody against a highly conserved epitope in the RPB1 N-terminus was obtained commercially (Santa Cruz, sc-899).

Chromatin immunoprecipitation

ChIP was carried out using formaldehyde crosslinked chromatin [21, 36]. In brief, 37% formaldehyde was directly added to synchronized parasite cultures ($2x10^9$ rings and trophozoites and $5x10^8$ schizonts) to a final concentration of 1% and incubated at 37°c for 10 min with gentle shaking. Crosslinking was terminated by addition of 1.25 M glycine to a final concentration of 0.125 M. RBCs were lysed by adding saponin to a final concentration of 0.15%. The nuclear pellet was lysed using 1% SDS and the resulting nuclear extract was sheared in Vibra Cell sonifier at 25% amplitude (10 sec on, 50 secs off) for 15 pulses to obtain DNA fragments in the range of 200-400 bp. The chromatin DNA was diluted, pre-cleared and incubated with the immunoprecipitating antibody overnight at 4°C. RNAPII-antibody-bound complexes were then bound to salmon sperm DNA/Protein A agarose slurry for 2 h at 4°C, followed by extensive washings to reduce the background.

Precipitated complexes were eluted from the beads and the protein-DNA cross-links were reversed by heating samples at 65°C overnight in 0.2 M NaCl. The DNA was purified using Qiagen Min Elute columns. For each antibody, three ChIP reactions were performed in parallel to obtain sufficient amount of DNA for ChIP-seq.

RNA preparation

Total RNA was isolated from three time points across the IDC (same as for ChIP), ring stage (12-16 hpi), trophozoites (24-28 hpi) and schizonts (36-40 hpi) by lysing the pelleted iRBCs in Trizol (Invitrogen). Subsequently, genomic DNA was removed by DNase I treatment and RNA was cleaned by phenol-chloroform extraction. DNA-free RNA samples were used for library preparation.

Preparation of ChIP-seq and RNA-seq libraries and sequencing

For ChIP libraries, 5 ng of ChIP or input material was used to prepare indexed Illumina sequencing libraries using the NEB Next Ultra DNA library prep kit (New England Biolabs). Adapter-ligated ChIP DNA libraries were amplified for 15 cycles. 150 bp paired-end reads were generated on the MiSeq V2 Run (150 PE).

Prior to library preparation, the quality of the RNA samples was determined on a Bioanalyzer 2,100, using a RNA 6000 Nano Chip (Agilent). Sample quantitation was carried out using Invitrogen's Ribogreen assay. Library preparation was performed according to Illumina's Stranded mRNA sample preparation protocol, following the manufacturer's recommendations.

Reverse transcription was performed using Invitrogen's Superscript III. For library enrichment, the number of amplification cycles was reduced to 12. Each library was uniquely tagged with one of Illumina's TruSeq LT RNA barcodes to allow library pooling for sequencing. The finished libraries were quantitated using Invitrogen's Picogreen assay and the average library size was determined on a Bioanalyzer 2100, using a DNA 7500 chip (Agilent). Library concentrations were then normalized to 4 nM and validated by qPCR on a ViiA-7 real-time thermocycler (Applied Biosystems), using qPCR primers recommended in Illumina's qPCR protocol, and Illumina's PhiX control library as standard. The libraries were then pooled at equimolar concentrations and sequenced in one lane on an Illumina HiSeq2500 sequencer in rapid mode at a read-length of 100 bp per paired-end.

The amplification of the ChIP DNA and cDNA libraries were performed using Kapa DNA polymerase enzyme, which has a very low A-T bias and is therefore commonly used when preparing *P. falciparum* sequencing libraries [7, 37].

Data Analysis

ChIP-seq data processing

Paired-end reads from ChIP-seq were run through FastQC (Version 0.11.4) [38] to check the quality of sequencing, followed by quality and length trimming using Trim Galore! (Version 0.4.1) [39]. Reads from ChIP and input DNA samples were mapped to PlasmoDbv9.0 using Bowtie2 (Version 2.2.6) [40]. Properly paired alignments with a MAPQ of 10 were extracted using SAMTools (Version 1.3) [41]. Read extension to fragments and input normalization by library size were done using deepTools (Version 2.3.4) [42]. deepTools was also used to generate heatmaps and profile plots. Profile plots in **Fig 2A** and **Fig 2B** required customization, and hence were generated using in-house Python scripts making use of score matrix from deepTools. Peak calling was carried out using MACS2 (Version 2.1.1) [24, 25]. Peaks were annotated to nearby gene using BEDTools (Version 2.25.0) [43]. Peaks were annotated to a gene, if the location of the peak overlaps anywhere from 1 kb up-stream to 1 kb downstream of the gene coordinates. For all downstream analysis, we have included only 5,400 nuclear, non-antigenic genes. The 5' UTR and 3' UTR were defined as 1 kb upstream of the translation start codon (ATG) and 1 kb downstream of the STOP codon, respectively.

RNA-seq

Paired-end reads from RNA-seq were run through FastQC (Version 0.11.4) to check the quality of sequencing, followed by quality control and length trimming using Trim Galore! (Version 0.4.1). RNA-seq reads were mapped to the transcriptome of the PlasmoDbv9.0 [44] using STAR-aligner (Version 2.5.1b) [45]. Properly paired alignments with a MAPQ of 10 were extracted using SAMTools (Version 1.3). Read counts in exons per gene were obtained using HTSeq-count (Version 0.6.1p1) [46], followed by DESeq2 [47] normalization and gene expression quantification as $\log_2$ values. Boxplots showing gene expression and statistical tests were carried out using in-house Python scripts. For all downstream analysis, we have included only 5,400 nuclear genes, excluding non-nuclear genes and genes encoding antigenic variants.

Comparison of RNAPII occupancy in defined genomic regions vs expression [Fig 3A, 3B and 4]

RNAPII peaks with enrichment > 1.5 in each stage and form were retained for this analysis. Peak to gene annotation was used to create classes in the following manner:

Fig. 3A: At each stage, genes were classified based on the presence of peaks from different combination of the three forms.

Fig. 3B: At each stage, gene were classified based on the presence or absence of peaks from Ser2/5-P and total RNAPII (N20).

Fig. 4: At each stage, genes were classified based on the different combination of location of Ser2/5-P peaks.

We used the two-sided Wilcoxon Rank Sum test to assess the significance of gene expression differences between the classes

Analysis of RNA-seq data [Fig 2B]

A set of 4,990 expressed genes was obtained by filtering out genes with very low expression in all three stages according to the formula:

$$Median\ (ring_{expression}, trophozoite_{expression}, schizont_{expression}) \geq log_2(2)$$

Gene expression in each stage was sorted independently in decreasing order and partitioned into three quantiles to produce the 'High' (1663 genes), 'Medium' (1663 genes) and 'Low' (1664 genes) groups. Average profile of RNAPII for these three classes of genes for each form and stage are shown with one color for each stage (Red-Ring; Blue-Trophozites; Green-Schizonts) and different opacity for each level of gene expression class (High, Medium, and Low)

The mapping statistics for all sequencing samples are given in the supplementary Table S7 and S8.

Validation of ChIP -seq by real –time PCR

qPCR was carried out with immunoprecipitated and input DNA on Applied Biosystems Step One Plus (Real-time PCR System) using Kappa SYBR green PCR master mix as per the manufacturer's instructions. ChIP enrichment was calculated by DCt method (C_t of immunoprecipitated target gene-C_t of input target gene where C_t is the threshold cycle). All PCR reactions were done in triplicate. Primers used in the study are given in Table S9.

List Of Abbreviations

bp, base pair(s)

ChIP, chromatin immunoprecipitation

CTD, C-terminal domain

GO, Gene Ontology

hpi, hours post-infection

IDC, intraerythrocytic developmental cycle

kb, kilo-base-pair(s)

RNAPII, RNA polymerase II

RBC, red blood cell

PIC, pre-initiating complex

UTR, untranslated region

TSS, transcription start site

ORF, open reading frame

GB, gene body

Declarations

Ethics approval and informed consent

This research involved the use of human blood drawn from healthy volunteers by trained healthcare professionals and was approved by the NTU Institutional Review Board and assigned the IRB approval number IRB-2013-07-020. Informed consent was obtained and documented by the signing of an approved consent form.

Consent for publication

Not applicable.

Availability of data and materials:

All the primary data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession no. GSE108716.

Competing interests:

The authors declare that they have no competing interests.

Funding:

This work was supported by a grant from the Singapore National Medical Research Council (#CBRG12nov104) to M. Featherstone and L. Wong, and AcRF Tier 2 grant from the Singapore Ministry of Education (#MOE2017-T2-2-030 (S)) to ZB.

Author's contributions:

Conceived and designed the experiments: RR and MF.

Performed the experiments: RR,

Analyzed the data: RR, AM, ZL, JZ, ZB and MF.

Contributed data/reagents/materials/analysis tools: RR, JZ, ZB and MF.

Acknowledgments

We thank Limsoon Wong and members of the Bozdech, Featherstone and Zhang labs for helpful discussions.

References

1. Bataille AR, Jeronimo C, Jacques PE, Laramee L, Fortin ME, Forest A, Bergeron M, Hanes SD, Robert F: **A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes**. *Mol Cell* 2012, **45**(2):158-170.
2. Harlen KM, Churchman LS: **The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain**. *Nat Rev Mol Cell Biol* 2017, **18**(4):263-273.
3. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al*: **Genome sequence of the human malaria parasite Plasmodium falciparum**. *Nature* 2002, **419**(6906):498-511.
4. Aravind L, Iyer LM, Wellems TE, Miller LH: **Plasmodium biology: genomic gleanings**. *Cell* 2003, **115**(7):771-785.
5. Russell K, Hasenkamp S, Emes R, Horrocks P: **Analysis of the spatial and temporal arrangement of transcripts over intergenic regions in the human malarial parasite Plasmodium falciparum**. *BMC Genomics* 2013, **14**:267.
6. Adjalley SH, Chabbert CD, Klaus B, Pelechano V, Steinmetz LM: **Landscape and Dynamics of Transcription Initiation in the Malaria Parasite Plasmodium falciparum**. *Cell reports* 2016, **14**(10):2463-2475.
7. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, Martins RM, Sismeiro O, Coppee JY, Scherf A: **Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum**. *BMC Genomics* 2014, **15**:150.

8. Vembar SS, Scherf A, Siegel TN: **Noncoding RNAs as emerging regulators of Plasmodium falciparum virulence gene expression**. *Current opinion in microbiology* 2014, **20**:153-161.
9. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum**. *PLoS Biol* 2003, **1**(1):E5.
10. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ *et al*: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle**. *Genome Res* 2004, **14**(11):2308-2318.
11. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL: **Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray**. *Genome Biol* 2003, **4**(2):R9.
12. Vembar SS, Macpherson CR, Sismeiro O, Coppee JY, Scherf A: **The PfAlba1 RNA-binding protein is an important regulator of translational timing in Plasmodium falciparum blood stages**. *Genome Biol* 2015, **16**(1):212.
13. Caro F, Ahyong V, Betegon M, DeRisi JL: **Genome-wide regulatory dynamics of translation in the asexual blood stages**. *Elife* 2014, **3**.
14. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, Prudhomme J, Florens L, Le Roch KG: **Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum**. *Genome Biol* 2013, **14**(11):R128.
15. Brancucci NM, Witmer K, Schmid C, Voss TS: **A var gene upstream element controls protein synthesis at the level of translation initiation in Plasmodium falciparum**. *PLoS One* 2014, **9**(6):e100183.
16. Cui L, Lindner S, Miao J: **Translational regulation during stage transitions in malaria parasites**. *Annals of the New York Academy of Sciences* 2015, **1342**:1-9.
17. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z: **Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites**. *Genome Biol* 2008, **9**(12):R177.
18. Kishore SP, Perkins SL, Templeton TJ, Deitsch KW: **An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases**. *J Mol Evol* 2009, **68**(6):706-714.
19. Yang C, Stiller JW: **Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(16):5920-5925.
20. Chapman RD, Heidemann M, Hintermair C, Eick D: **Molecular evolution of the RNA polymerase II CTD**. *Trends in genetics : TIG* 2008, **24**(6):289-296.
21. Rai R, L. Z, Chen H, Gupta AP, Sze SK, Zheng J, Ruedl C, Bozdech Z, Featherstone M: **Genome-wide analysis in Plasmodium falciparum reveals early and late phases of RNA polymerase II occupancy during the infectious cycle**. *BMC Genomics* 2014, **15**:959.
22. Karmodiya K, Pradhan SJ, Joshi B, Jangid R, Reddy PC, Galande S: **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* 2015, **8**:32.

23. Anamika K, Gyenis A, Poidevin L, Poch O, Tora L: **RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts**. *PLoS One* 2012, **7**(6):e38769.
24. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
25. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS**. *Nat Protoc* 2012, **7**(9):1728-1740.
26. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.
27. Painter HJ, Chung NC, Sebastian A, Albert I, Storey JD, Llinas M: **Genome-wide real-time in vivo transcriptional dynamics during Plasmodium falciparum blood-stage development**. *Nat Commun* 2018, **9**(1):2656.
28. Shock JL, Fischer KF, DeRisi JL: **Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle**. *Genome Biol* 2007, **8**(7):R134.
29. Booth GT, Wang IX, Cheung VG, Lis JT: **Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast**. *Genome Res* 2016, **26**(6):799-811.
30. Price DH: **Regulation of RNA polymerase II elongation by c-Myc**. *Cell* 2010, **141**(3):399-400.
31. Lu XM, Batugedara G, Lee M, Prudhomme J, Bunnik EM, Le Roch KG: **Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite Plasmodium falciparum**. *Nucleic Acids Res* 2017, **45**(13):7825-7840.
32. Kensche PR, Maria Hoeijmakers WA, Toenhake CG, Bras M, Chappell L, Berriman M, Bartfai R: **The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences**. *Nucleic Acids Research* 2015.
33. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA: **c-Myc regulates transcriptional pause release**. *Cell* 2010, **141**(3):432-445.
34. Moll K, Ljungstrom I, Perlmann H, Scherf A, Wahlgren M: **Methods in Malaria Research**. In., 6th edn; 2013.
35. Lambros C, Vanderberg JP: **Synchronization of Plasmodium falciparum erythrocytic stages in culture**. *J Parasitol* 1979, **65**(3):418-420.
36. Lopez-Rubio JJ, Siegel TN, Scherf A: **Genome-wide chromatin immunoprecipitation-sequencing in Plasmodium**. *Methods Mol Biol* 2013, **923**:321-333.
37. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP *et al*: **Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes**. *BMC Genomics* 2012, **13**:1.

38. Andrews S: **FastQC A Quality Control tool for High Throughput Sequence Data**. *http://wwwbioinformaticsbabrahamacuk/projects/fastqc/*.
39. **Trim Galore!** [http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/]
40. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Meth* 2012, **9**(4):357-359.
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
42. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis**. *Nucleic Acids Research* 2016, **44**(W1):W160-W165.
43. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
44. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS *et al*: **PlasmoDB: a functional genomic database for malaria parasites**. *Nucleic Acids Res* 2009, **37**(Database issue):D539-543.
45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
46. Anders S, Pyl PT, Huber W: **HTSeq - A Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**.
47. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.

Figures

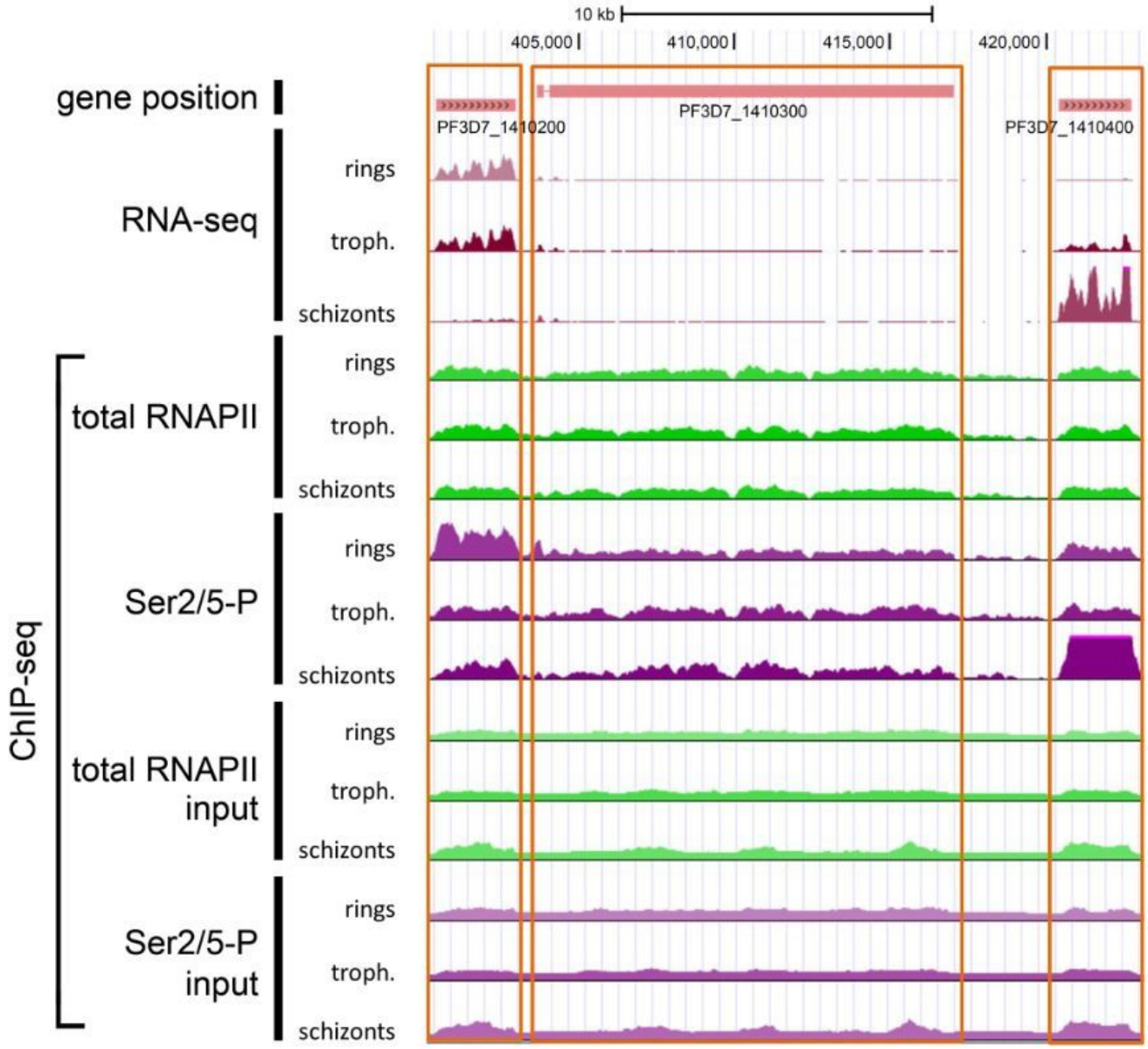

Figure 1

Distribution of transcripts and RNAPII across three representative genes. Genome browser snapshot showing distribution of transcripts and RNAPII across three representative genes on a contiguous portion of chromosome 14 of P falciparum at ring, trophozoite and schizont stages. For simplicity, only the ChIP-seq results for two informative antibodies are shown; these correspond to the N20 antibody which recognizes all forms of RNAPII (total RNAPII) and the antibody recognizing the doubly phosphorylated heptad repeat (RNAPII-Ser2/5-P) present on the CTD of the elongating form of the polymerase. The plot for ChIP-seq reads of corresponding input chromatin (before immunoprecipitation) are shown toward the bottom of the panel. The y-axis scale for all ChIP-seq results is identical, as is the case for all RNA-seq results. The extent of each of the three genes from start to stop codon is given by pink lines at the top of the figure. Exons are given by the thicker pink lines, while the single intron is given by a thin connecting pink line. The positions of the three genes and their associated data are highlighted by orange boxes. The

three genes are PF3D7_1410200 (cytidine triphosphate synthetase), PF3D7_1410300 (putative WD repeat-containing protein) and PF3D7_1410400 (rhoptry associated protein 1 (RAP1)). The plot makes use of the P. falciparum strain 3D7 genome version pf3d7v9. We made use of the publicly available genome browser at the University of California at Santa Clara [26]. Annotations were accurately redrawn to bring resolution to publication standards.

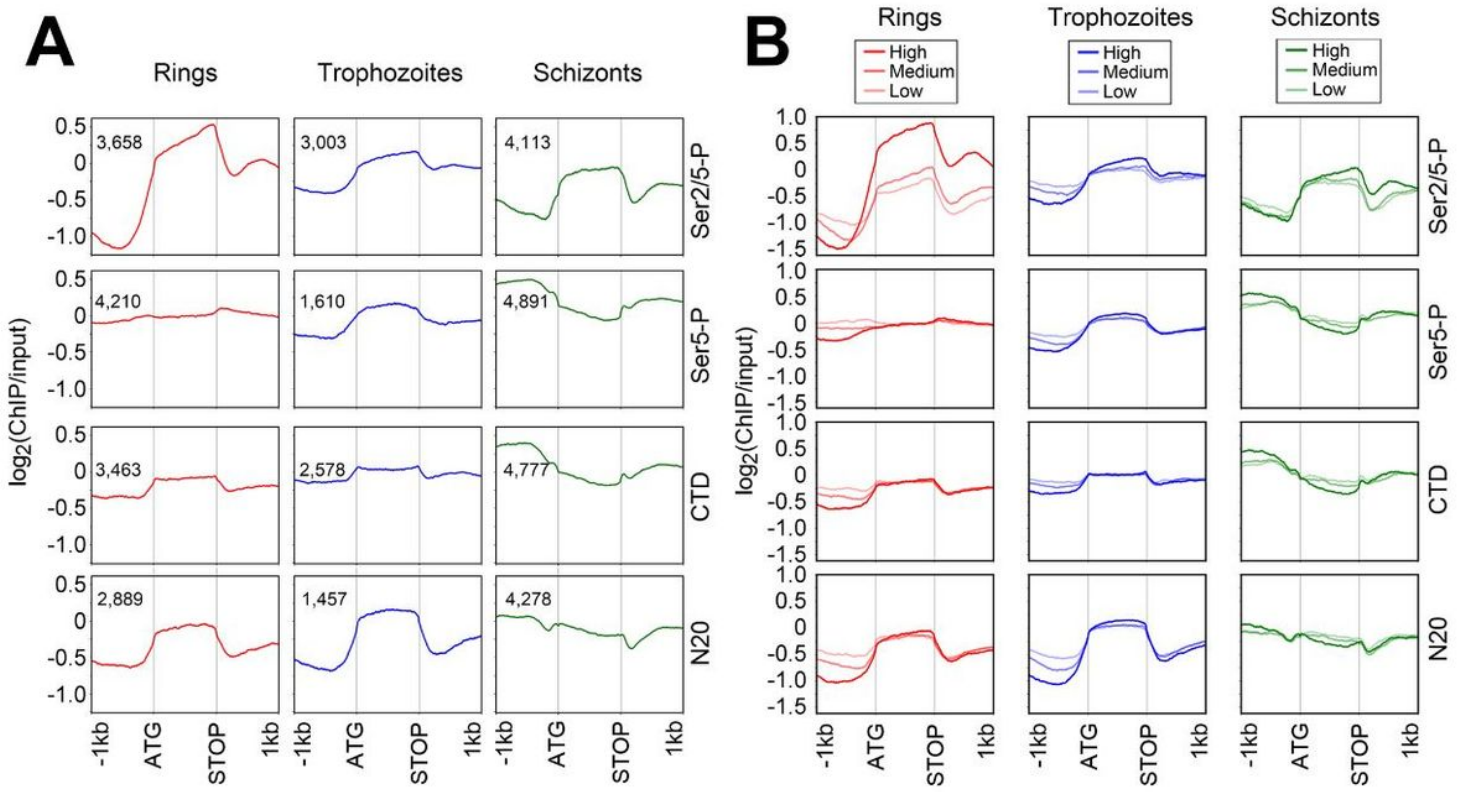

Figure 2

A. Stage-specific distribution of RNAPII along the transcription unit at each stage of the IDC. Average occupancy of total RNAPII and three phosphoisoforms across multiple transcription units at three stages of the IDC. To visualize major trends in RNAPII distribution for genes of varying length, coding regions were divided into 1,000 bins, and 5' and 3' flanking regions (always 1 kb upstream of the ATG and 1 kb downstream of the stop codon) were divided into 1,000 bins each, thereby scaling all genes uniformly to 3,000 bins. All ChIP signals were normalized to the chromatin input signal. Only those genes having at least 1 bin out of 3,000 bins with enrichment ≥ log2(2) were considered. Occupancy profiles are shown as color-coded line plots. B. Stage-specific distribution of RNAPII along the transcription unit as a function of stage-specific transcript levels. Average binding profiles of RNAPII over genes with different expression levels. Genes in each stage were partitioned into three different expression groups based on their expression levels as high, medium, and low indicated by colored lines of different opacity. The y-axis indicates the log2(ChIP/Input) ratio. The x-axis represents the coding region and 1 kb flanking regions on either side of the ATG and STOP codon.

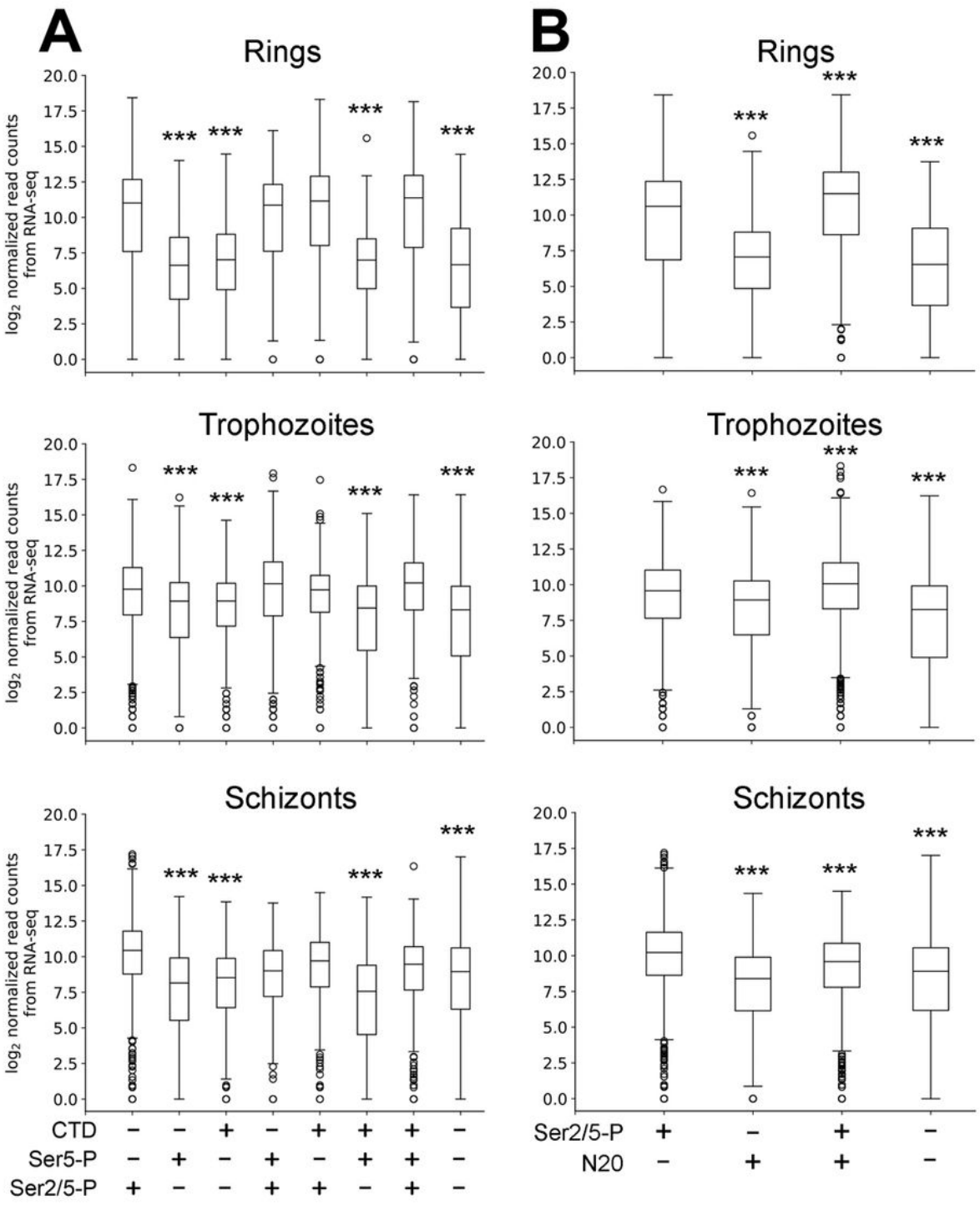

Figure 3

A. Peak calling reveals a correlation between binding by Ser2/5-P and transcript output. Box plots show the expression levels of genes with single or combinatorial binding of RNAPII (different phosphoisoforms) across the three different stages. Peaks filtered for P-value and fold change were assigned to nearby genes and the genes were partitioned among eight different groups based on their association with these peaks. The occupancy of each RNAPII phosphoisoform is shown along the x-axis

as + (presence of peaks) and – (absence of peaks). The y-axis provides the scale for log-transformed expression levels at each of the three stages of the IDC. The P-values were calculated by the two-tailed Wilcoxon Rank-Sum Test (Table S1). ***indicates a P-value of < 0.001 in comparisons to the expression levels for genes having peaks of Ser2/5-P. Peaks of RNAPII occupancy for all stages and forms were filtered for a P-value < 0.05 and fold-enrichment of > 1.5 on a raw scale (i.e. before log2 transformation). B. Gene-associated peaks of Ser2/5-P, but not those of N20, are strongly correlated with transcript output. Box plots show the expression levels of genes having peaks of occupancy by the Ser2/5-P RNAPII phosphoisoform, total RNAPII or both, across the three stages of the IDC. The analysis was performed as for Fig 3A. Statistically significant differences between expression levels are indicated by connecting lines and the corresponding p values. Refer to Table S2 for a full list of P-values determined by the two-tailed Wilcoxon Rank-Sum Test. ***indicates a P-value of < 0.001 in comparisons to the expression levels for genes having peaks of Ser2/5-P.

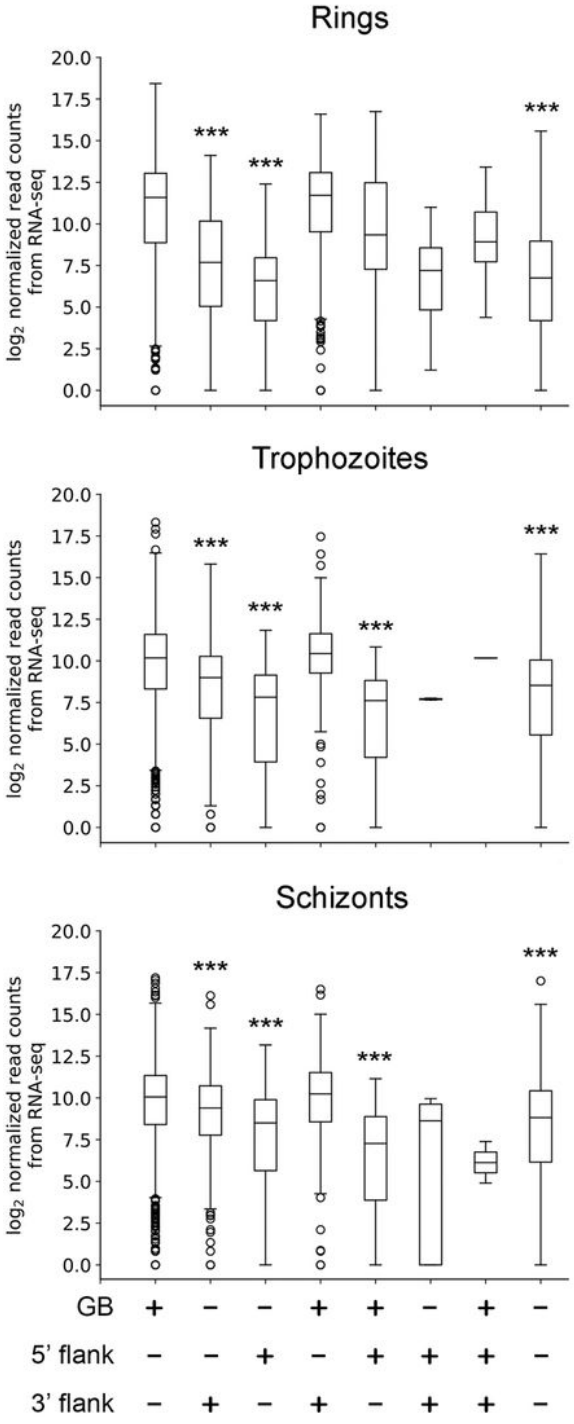

Figure 4

The position of peaks of Ser2/5-P occupancy along the transcription unit correlates with transcript output. Box plots show distribution of gene expression levels of genes having MACS2-defined peaks of Ser2/5-P signal in three regions across the transcription unit: the coding region or gene body (GB), the region extending 1 kb upstream of the ATG (5' flank), and 1 kb downstream of the stop codon (3' flank). Peaks of RNAPII Ser2/5-P isoform in the indicated genomic location are shown as + (present) and –

(absent). The y-axis provides the scale for the log2 transformed gene expression levels from RNA-seq data. Data are shown for each of the three IDC stages assayed. ***indicates significantly different expression levels at P-values < 0.001 in comparisons with the expression level of the GB gene set. All P-values calculated using the two-sided Wilcoxon Rank-Sum Test are given in Table S4.

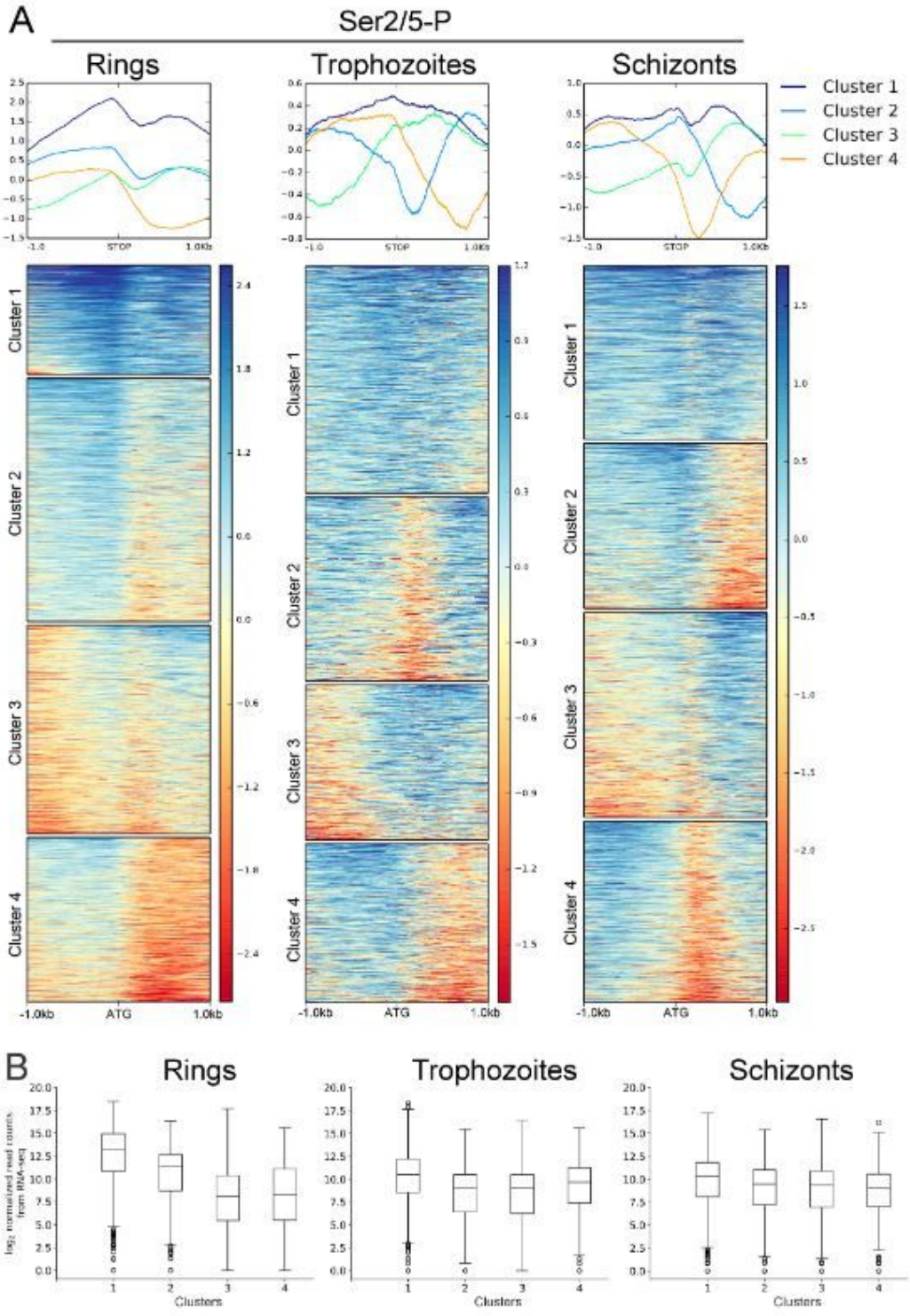

Figure 5

Stage-specific accumulation of Ser2/5-P at the 3’ end correlates with stage-specific transcript output in rings and schizonts. A. Heat map generated after K-means clustering of RNAPII Ser2/5-P enrichment in region 1 kb upstream and downstream of the STOP codon across the three stages during IDC. The color scale indicates the level of enrichment. K4-means clustering generates four clusters (1-4) across each stage. B. The box plots represent relative mRNA levels of genes in each cluster during ring, trophozoite and schizont stages. *** indicates statistically significant differences between the expression level of cluster 1 genes and each of the other clusters with P-values of < 0.0001 (two-tailed Wilcoxon Rank Sum Test). The full set of P-values is provided in Table S5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiguresLegendsTablesAddFiles.pdf