*Nucleic Acids Research, 2018* **1**
*doi: 10.1093/nar/gky643*

# Characterization of the accessible genome in the human malaria parasite *Plasmodium falciparum*

**José Luis Ruiz[1], Juan J. Tena[2], Cristina Bancells[3], Alfred Cortés[3,4,†], José Luis Gómez-Skarmeta[2,†] and Elena Gómez-Díaz[1,5,*]**

[1]Estación Biológica de Doñana (EBD), Consejo Superior de Investigaciones Científicas, Seville 41092, Spain, [2]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas-Universidad Pablo de Olavide-Junta de Andalucía, Seville 41013, Spain, [3]ISGlobal, Hospital Clínic - Universitat de Barcelona, Barcelona, Catalonia 08036, Spain, [4]ICREA, Barcelona, Catalonia 08010, Spain and [5]Instituto de Parasitología y Biomedicina 'López-Neyra' (IPBLN), Consejo Superior de Investigaciones Científicas, Granada 18016, Spain

Received February 24, 2018; Revised July 04, 2018; Editorial Decision July 05, 2018; Accepted July 10, 2018

## ABSTRACT

**Human malaria is a devastating disease and a major cause of poverty in resource-limited countries. To develop and adapt within hosts *Plasmodium falciparum* undergoes drastic switches in gene expression. To identify regulatory regions in the parasite genome, we performed genome-wide profiling of chromatin accessibility in two culture-adapted isogenic subclones at four developmental stages during the intraerythrocytic cycle by using the Assay for Transposase-Accessible Chromatin by sequencing (ATAC-seq). Tn5 transposase hypersensitivity sites (THSSs) localize preferentially at transcriptional start sites (TSSs). Chromatin accessibility by ATAC-seq is predictive of active transcription and of the levels of histone marks H3K9ac and H3K4me3. Our assay allows the identification of novel regulatory regions including TSS and enhancer-like elements. We show that the dynamics in the accessible chromatin profile matches temporal transcription during development. Motif analysis of stage-specific ATAC-seq sites predicts the *in vivo* binding sites and function of multiple ApiAP2 transcription factors. At last, the alternative expression states of some clonally variant genes (CVGs), including *eba*, *phist*, *var* and *clag* genes, associate with a differential ATAC-seq signal at their promoters. Altogether, this study identifies genome-wide regulatory regions likely to play an essential function in the developmental transitions and in CVG expression in *P. falciparum*.**

## INTRODUCTION

The chromatin structure defines the scenario where the interactions between transcription factors (TFs) and their cognate regulatory regions take place, and this is dynamically shaped in a cell and stage-specific manner. To successfully interact with promoters, enhancers, insulators and non-coding RNAs (ncRNAs), TFs must induce chromatin remodeling of nucleosomal structures, which results in different levels of chromatin accessibility (1,2). Based on this principle, open chromatin profiling can be used to identify regulatory elements associated with specific transcriptional and epigenetic states (3). The identification of these regulatory sequences in their native chromatin environment is important for understanding how gene expression is coordinated throughout development and in response to the environment.

Malaria parasites show a complex life cycle in the human and the mosquito hosts and face dynamically changing environments during development. *Plasmodium falciparum* is the most virulent and prevalent human malaria parasite species in Africa (4). In order to successfully develop within the host, this parasite switches transcriptional programs between stages (5–7) and has evolved mechanisms to adjust its phenotype through transcriptional heterogeneity within populations. The ability of the parasite to survive in the face of changes in host conditions is tightly linked to the variant gene expression of a number of genes involved in processes such as antigenic variation, red blood cell invasion, solute transport and sexual differentiation (8,9). These genes show clonally variant gene (CVG) expression, such that individual parasites having identical genomes and under the same environment can maintain a variant gene in a different transcriptional state, and this state can be transmitted to the next generation by epigenetic mechanisms. Multiple gene families showing clonally variant expression have been characterized (6,10), many of which encode anti-

*To whom correspondence should be addressed. Tel: +34 954 466 700; Fax: +34 954 621 125; Email: elena.gomez@csic.es
†The authors declare that, in their opinion, the fourth and fifth authors contributed equally to the manuscript.

gens expressed at the surface of infected erythrocytes, such as the multicopy *var*, *rifin*, *stevor*, *surfin* and *Pfmc-2TM* CVG families (10,11). Among them, the *var* genes encode the Erythrocyte Membrane Protein 1 (PfEMP1), which is a critical virulence factor for malaria. Each individual parasite has ∼60 different *var* genes, only one of which is expressed at a time, i.e. mutually exclusive expression (12–14). Another clonally variant family is *clag*, in which two of the genes (*clag3.1* and *clag3.2*) are also expressed in a mutually exclusive manner (15). They encode for membrane proteins involved in solute transport and are linked to drug resistance (16–20).

Despite the implications for malaria pathogenesis, the role of chromatin-mediated processes in the regulation of gene expression remains poorly understood. This lack of knowledge is in part due to unique particularities of *P. falciparum* in relation to genomic architecture. For example, one key feature is its extreme AT richness, which is the highest compared to all genomes sequenced to date: >80% on average and often reaching 90–95% in intergenic regions (21). Recent models suggest that highly AT-rich sequences may inhibit the formation of higher-order structures maintaining chromatin accessible and the genome in a transcriptionally permissive state (22,23). However, during development, the parasite undergoes drastic switches in the pattern of gene expression (5,24,25). Active chromatin regions are marked primarily by H3K4me3 and H3K9ac (26,27), whereas the non-active fraction of the genome corresponds to blocks of heterochromatin that are marked by histone 3 tri-methylation at lysine 9 (H3K9me3) and the heterochromatin protein HP1 (28). Importantly, CVGs locate primarily in facultative heterochromatin regions occupying central and subtelomeric clusters. The question then arises: how these genes are switched on/off in such a heterochromatic environment?

A growing hypothesis in the field is that promoter-specific temporal regulation of transcription in *P. falciparum* occurs through site-specific recognition of DNA sequences by TFs that are uniquely produced in each stage of development. Several *cis*-regulatory elements involved in the temporal regulation of transcription have been predicted bioinformatically, but in general they are still poorly understood (29–32). Furthermore, despite some work having been done on *trans*-acting regulatory DNA sequences (33), the characterization and function of enhancer elements remain a poorly explored issue in the field. In *P. falciparum* there is only one large family of specific TFs characterized to date, the ApiAP2 family. Current evidence points to an essential role of these DNA regulatory proteins in driving *P. falciparum* developmental progression (34–44). Accordingly, the genes encoding these TFs are expressed in a stage-specific manner (45,46). The cognate motifs of several ApiAP2 TFs have been determined experimentally (47), but only a few of these have been tested *in vivo* by chromatin immunoprecipitation-sequencing (ChIP-seq) (41,42), mainly because of the difficulties in obtaining antibodies against *Plasmodium* proteins (47). Profiling of accessible chromatin can be used to reveal gene regulatory networks and to identify novel trans-acting factors and *cis*-regulatory elements. MNase-seq has been employed to characterize nucleosome-depleted regions at different time points during the intraerythrocytic developmental cycle (IDC) (48–50). This technique is useful for identifying genomic regions occupied by nucleosomes or accessible to nuclease cleavage. However, this data alone provides limited information regarding the identity and function of regulatory elements (51). FAIRE-seq is another method for mapping chromatin accessibility that has been used in *Plasmodium* but has low resolution and limited accuracy in identifying DNA–protein binding events. In consequence, many of the TF binding sites in *P. falciparum* have only been predicted bioinformatically (29–32). In this context, the Assay for Transposase-Accessible Chromatin by sequencing (ATAC-seq) has the advantage of allowing simultaneous mapping of nucleosomes and DNA-binding proteins protected sites, being much faster and more sensitive and requiring orders of magnitude less starting material than other assays (3).

In this work, we use ATAC-seq to characterize the chromatin accessibility landscape during the intraerythrocytic cycle in two transcriptionally variant subclones. Our results show that differences in chromatin accessibility recapitulate transcriptional changes during development and are predictive of clonal differences in epigenetic and expression profiles. The analysis of differentially accessible regions allowed us to identify the landscape of binding events, regulatory DNA sequences and putative TFs that are likely responsible for these changes. Altogether, our work shows the dynamics of the *P. falciparum* regulatory genome and its association with developmentally regulated and CVG expression.

## MATERIALS AND METHODS

### Parasite cultures

10G and 1.2B correspond to *P. falciparum* subclones derived from the same 3D7-A stock, which have been characterized in previous studies (6,15,52,53). The experiments were performed at different times but on the same laboratory and using the same parasite stocks. Parasites were thawed from a cryopreserved stock and cultured for two or three generations in B+ erythrocytes (3% hematocrit) under standard conditions with media containing Albumax II and no human serum. Cultures were tightly synchronized to a defined 5 h age window by purification of parasites at the schizont stage using Percoll gradients (63% Percoll), followed by sorbitol lysis 5 h later to eliminate erythrocytes infected with late asexual stages (trophozoites and schizonts). Parasites were collected at the following times post invasion: 10–15 hpi (T10, early ring stage), 20–25 hpi (T20, late ring stage), 30–35 hpi (T30, trophozoite stage) and 40–45 hpi (T40, schizont stage). Two independent biological replicates were performed. gDNA was isolated from parasites at the schizont stage using the DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's instructions.

### ATAC-seq

The ATAC-seq protocol was performed using 10 million cells for early ring, late ring and throphozoite cell stages, and 1 million cells for schizonts, which were obtained after saponin red blood cell lysis. Parasite cells were resuspended in lysis buffer in order to permeabilize membranes. The nuclei pellet was immediately resuspended in the transposition

reaction mix (25 μl of 2× TD Buffer, 1.25 μl of Tn5 Transposase and 23.75 μl of nuclease free water), and incubated for 30 min at 37°C. As a control, 50 ng of gDNA were used for the standard ATAC-seq procedure as described in the Nextera library preparation protocol by Illumina (5 min at 55°C). All transposed samples were purified using the Qiagen MiniElute Kit. Following purification, ATAC-seq library amplification was carried out with 2× KAPA HiFi mix and 1.25 μM of Nextera primers (3). The use of this enzyme eliminates mapping biases due to the high AT content of the *P. falciparum* genome (54,55). The optimal cycle number was determined using quantitative polymerase chain reaction (qPCR) with PCR conditions as originally described in (3). ATAC-seq libraries were sequenced at BGI (China) using an Illumina HiSeq2000 sequencer to obtain 20–40M of 2 × 50 bp paired-end sequencing reads per sample (Supplementary Table S1).

### Library QC

Quality control on Illumina reads was performed using FastQC (v.0.11.5, http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). The alignment and accuracy statistics were computed using QualiMap (v.2.2.1) (56) (Supplementary Table S1). We used ataqv (v.0.9.4, https://github.com/ParkerLab/ataqv) and Picard Tools (v.2.2.2, http://broadinstitute.github.io/picard/) to plot the distribution of paired-end sequencing fragment sizes. Replicability analysis was performed using deepTools2 (v.2.5.0) (57) that reports the Spearman rho correlation coefficient between each pair of '.bam' alignment files. In addition, the plotPCA function included in the DESeq2 R package (58) was used to calculate the percentage of total variance explained by differences between subclones, replicates and stages.

### External data

ChIP-seq raw data for various hPTMs: H2A.Z, H3K4me3, H3K4me1 and H3K9ac was obtained from Gene Expression Omnibus (GEO) database: accession numbers GSE23787 (26) and GSE63369 (59). Reads were aligned to the *P. falciparum* reference genome Pf3D7-28 (PlasmoDB) using Bowtie2 (v.2.3.1) (60) with default parameters except for -no- unal -no-mixed -X 2000, and trimmed 10 bases from the 3' end (-3 10) of each read. Colorspace reads corresponding to samples in GSE63369 (59) were aligned using Bowtie (v.1.2.1) (61). Aligned files were sorted and deduplicated using Samtools (v.1.4, http://samtools.sourceforge.net/). ChIP-seq peaks for AP2-I and BDP1 proteins were obtained from previous studies (41,62). Microarray expression data of the same subclones and stages was from Rovira-Graells *et al*. (6). RNA-seq data for the four time points analyzed in our study was downloaded from GEO accession number GSE66185 (48).

### Data analysis

All sequenced ATAC-seq paired reads were trimmed 10 bases from each read 3' end (-3 10) and aligned to the *P. falciparum* reference genome Pf3D7-28 (PlasmoDB) with Bowtie2 (v.2.3.1) using default parameters except for –no-unal –no-mixed -X 2000. Alignment files were filtered applying a quality threshold of 10 in the MAPQ score, sorted and deduplicated using Samtools (v.1.4). To adjust for fragment size and to make sure Tn5 cutting sites were mapped, aligned reads were shifted +4 bp for + strands and −5 bp for − strands. The extremely high AT bias of the *P. falciparum* genome might reduce mapping quality and introduce coverage biases. Furthermore, the presence multiple copy gene families, such as *var* and *clag3*, can lead to multiple sites mapping and the removal of non-unique reads. To check the impact of this mapping biases, we ran Bowtie2 with no quality threshold and found a high similarity in the percentage of ATAC-seq coverage at *var* and *clag3* families, with <10% difference between the two quality filtered alignments.

ATAC-seq data analysis was conducted following the prototype ATAC-seq ENCODE Uniform Processing Pipeline prototype (https://www.encodeproject.org/atac-seq/). First, nucleosome-free fragments were extracted by applying a size threshold of 130 bp. Peak calling on nucleosome-free reads was performed by MACS2 (v.2.1.1) (63) using the module 'callpeak' with the following parameters: -c -g 2.41e7 –keep-dup all -q 0.01 –nomodel –shift -35 –extsize 75 and -B. Transposed genomic DNA (gDNA) was used as a control for the peak calling. For this analysis, bam files were downsampled (Samtools v.1.4) to equal the number of reads between treatment and input (gDNA control) for each ATAC experimental sample (Supplementary Table S1). We applied a replicability filter to the output of the MACS2 peak calling analysis using the 'intersect' tool from the BEDTools software suite (v.2.25.0) (64). In addition to the independent peak calling for each biological replicate, we performed peak calling on the pooled sample and the pseudoreplicates, and kept only consensus ATAC-seq peaks if present in the three datasets (replicate, pooled and pseudoreplicate). Peaks located in mitochondrion and apicoplast chromosomes were removed. The resulting set of ATAC-seq peaks, that we considered high confidence Tn5 hypersensitive site (THSS) regions, are included in Supplementary Table S2. For visualization purposes, tracks of input (gDNA) corrected ATAC-seq signal per stage were built with MACS2 'bdgcmp' module (-m ppois) on each pair of fragment pileup and control lambda bedGraph files from the peak calling analysis.

Annotation of THSS regions to genomic features was performed using the HOMER software (v.3.12) (65). We used the gene ensemble available from PlasmoDB (Pf3D7-28) and considered the translation start site ATG as the reference point and −1 Kb upstream as the putative promoter region. This assumption is based on a recent study that mapped transcription start sites (TSSs) in *P. falciparum* at high resolution and revealed that 81% of the TSSs are positioned less than 1 Kb upstream of the start codon, with more than half of these located at distances less than 0.5 Kb (66). For the ATAC-seq enrichment analysis we used BEDTools to obtain the number of reads overlapping with: known TSSs (48,66), promoters (−1 Kb of the translation start codon, ATG), introns, exons, etc. The resulting read counts were normalized, ATAC to noise signal was corrected (ATAC/gDNA ratio) and log2-transformed using R

(v.3.4.1) (http://www.r-228 project.org/). We added a pseudocount (0.1) to avoid infinite values when control/input correcting (dividing by 0 in ratios) or when computing log2 values.

In order to assess the global relationship between the ATAC-seq signal at promoters and the transcription levels of the corresponding gene, we categorized genes into high, medium or low groups, based on their mRNA levels using previously published RNA-seq data (48). For this purpose, we applied a threshold value determined by dividing the mRNA values (RPKM normalized values) in three quantile groups according to their means (*cut2* function in the 'Hmisc' R package) (67). Average profile plots representing ATAC-seq and histone modifications enrichments (RPKM normalized and input-corrected) centered on TSSs or ATAC peaks coordinates, were built using ngs.plot (v.2.61) (68). Violin plots, boxplots and line diagrams were produced using 'ggplot2' R package (69). For comparative and visualization purposes, ATAC-seq, RNA-seq and histone ChIP-seq enrichment data at various genomic features are shown on a log2 scale. Correlation tests were performed between mRNA levels, ATAC-seq and H3K9ac/H3K4me3 ChIP-seq enrichment data using a Spearman rank correlation test in R.

Unless if otherwise specified, interval operations like intersect, shuffle, merge, flank or slop were performed using BEDTools and statistical tests and plots were performed in R using Bioconductor (http://www.bioconductor.org).

### Characterization of novel ATAC-seq regulatory regions

We classified novel THSS elements, ATAC-seq peaks located in intergenic regions, into TSS or enhancer-like regions using published H3K4me1, H3K4me3 and H3K9ac ChIP-seq data (59). Since H3K4me1 data exists only for trophozoite and schizonts, we restricted our analysis to these two stages. We used ChromHMM (70) to compute chromatin state predictions on the entire *P. falciparum* genome based on relative enrichment levels of these three histone modification marks. For the binarization of the genome, default values were used for all parameters except for -b 75. We chose a four states model as a prior assuming the following chromatin states: unspecific (high levels of enrichment for all hPTMs), depleted (low levels of enrichment for all hPTMs), TSS (H3K9ac/H3K4me3 enrichment and H3K4me1 depleted) and enhancer (H3K4me1 enrichment and H3K9ac/H3K4me3 depleted). According to the ChromHMM segmentation, most of the genome displays low enrichment signal (depl). In order to obtain a high-confidence set of regulatory regions, we first filtered out THSSs located in depleted and unspecific enrichment signal genome segments. We then ran again ChromHMM model on the filtered dataset and discarded those regions not classified as TSS-like or enhancer-like in the second prediction.

### Soft clustering analysis

In order to analyze chromatin accessibility dynamics among different stages, we conducted a soft clustering analysis using the 'Mfuzz' R package (71) on THSS-annotated promoters. Using a standard *m* fuzzy c-means parameter of 1.7, a total of 30 pre-clusters were created that summarize the variability of temporal profiles in our data. Promoters showing a pertinence membership >51% and a unique peak of chromatin accessibility at a given stage were grouped into four Mfuzz clusters with promoters showing maximum ATAC-seq enrichment levels at early rings, late rings, trophozoite and schizont parasite stages, respectively (Supplementary Figure S4). For visualization purposes, the promoters in each cluster were ordered by ATAC-seq enrichment levels at corresponding stage, and we then used this clustering order to show the association between the dynamics in chromatin accessibility, gene expression and histone modifications enrichment. For this analysis we used the microarray normalized counts obtained in a previous study of the same subclones and stages of development (6). ChIP-seq enrichment data for H3K4me3 and H3K9ac was obtained from (26).

### Differential ATAC-seq enrichment analysis

Of the 5085 total peaks, 1053 were unique to 10G and 1797 to 1.2B parasites (Supplementary Figure S7A). However, many of the ATAC-seq peaks unique to one of the two subclones are likely attributable to some variability in the performance of the MACS2 peak calling algorithm (e.g., some low enriched peaks may be similar between the two subclones but pass the peak calling criteria in only one) (Supplementary Figure S7B). To account for this issue, differential ATAC-seq accessibility analysis between 1.2B and 10G subclones was conducted using the 'DESeq2' R package (58). This analysis takes all the ATAC-peaks called in 10G and 1.2B, and tests for normalized read count differences at the peak region. In analyzing clonal differences for each developmental stage separately, we detected that a fraction of the clonal variation was due to slight differences in age between subclones. For this reason, we repeated the analysis treating the stage of development as a co-variable and we used this set of clonally variant peaks for all subsequent analyses. Heatmaps showing ATAC-seq enrichment at promoters and microarray gene expression data for the clonally variant genes were built using the 'iheatmapr' R package (72) and the *heatmap.2* function in the 'gplots' R package (73). In order to visualize quantitative differences in gene expression patterns between subclones we used microarray data from (6) and considered the log2 ratio relative to a reference pool. Due to very low transcription levels for most gene copies in multicopy gene families like *var*, the expression of a given gene in the reference pool could be higher than in the sample. To control for this, we obtained semi-quantitative estimates of *var* gene expression patterns considering microarray values for the sample channel (F635) instead of the log2 ratio relative to a reference pool when building the heatmaps. Furthermore, the telomeric PF3D7_1100100 *var* gene was excluded from the analyses because there is a deletion in the sequence in the 10G subclone. For visualization purposes, microarray expression was log2-transformed. Promoter windows for *var* genes were defined as 2 Kb upstream of the ATG sites if possible, or as the intergenic region not overlapping with exons or promoters of nearest adjacent genes. In cases where over-

lapping adjacent genes were of opposing directions, half the intergenic distance was assigned to each gene.

### Motif discovery

We carried out *de novo* motif analysis using HOMER software (v.3.12) (65) on the stage-specific THSS regions obtained by the soft clustering analysis. For this analysis, ATAC peaks specific of each time point (T10, T20, T30 and T40) were merged and the highest scoring summit was selected as summit for the merged peak. We then used the 'slop' tool from BEDTools software suite to obtain equal size 200 bp peaks from the summit coordinate. This analysis allowed us to identify DNA motif sequences enriched in the set of active regulatory sites during the intraerythrocytic development. Next, in order to investigate whether predicted binding sites for ApiAP2 TFs are among the motifs identified by HOMER, we performed similarity domain analysis with the TomTom tool (http://meme.nbcr.net/meme/doc/tomtom.html) using the previously published database of validated ApiAP2 motifs (47). Only motifs enriched in more than 5% of the targets sequences and below a threshold *P*-value of 10e-10 were considered, and results corresponding to low complexity motifs and offsets or degenerate versions of highly enriched motifs were discarded. For the TomTom analyses, those hits that aligned to the core and not to the edge of the known ApiAP2 motif, and met a minimum *P*-value of 10e-3, were selected. We used the annotatePeaks.pl module in HOMER to find motif occurrences in each stage-specific ATAC peaks sets. A Cytoscape network analysis (http://www.cytoscape.org/) was conducted to investigate links between significantly enriched DNA motif sequences at stage-specific regulatory active sites and their target genes.

### Real-time quantitative PCR

Microarray expression data of 10G and 1.2B parasite lines from a previous study (6) used for comparison with the ATAC-Seq data, was validated for a subset of CVGs (Supplementary Figure S9). Total RNA from in vitro cultures of the same parasite stocks was obtained by TRIZOL extraction followed by reverse transcription using the Tetro cDNA Synthesis Kit (Bioline), according to the manufacturers' instructions. Amplification was performed using Roche LighCycler 480. Reactions (12.5 μl) were performed in duplicate and consisted of 0.5 μl of 1/10 cDNA dilution, 0.2 μM of forward and reverse primers and 5 μl of KAPA SYBR FAST qPCR Master Mix (KapaBiosystems). Cycling conditions were 10 min initial denaturation at 95°C, followed by 40 cycles of 15 s denaturation at 95°C, 30 s annealing at 57°C and 30 s extension at 60°C. Melting curve analysis was performed to guarantee specificity of the template. The amount of cDNA was normalized using the housekeeping gene S-adenosylmethionine synthetase (PF3D7_0922200). The $2^{-\Delta\,\Delta CT}$ method (74) was used to estimate relative mRNA abundance.

## RESULTS

### Chromatin accessibility by ATAC-seq is predictive of active transcription and epigenetic states in *P. falciparum*

We generated ATAC-seq libraries using tightly synchronized parasites at four different developmental stages during the IDC: early rings (10–15 h, T10), late rings (20–25 h, T20), trophozoites (30–35 h, T30) and schizonts (40–45 h, T40) (Supplementary Table S1). For each stage, we obtained two independent ATAC-seq replicates and performed experiments in parallel with two different parasite lines, 1.2B and 10G (Figure 1A). These lines correspond to two subclones of the same 3D7 genetic background that differ in the expression of several CVGs (6,15). The similarity between replicates is high compared to the similarity between subclones and between stages (Supplementary Figures S1A and B). The similar distribution of fragment sizes suggests that chromatin is accessible to Tn5 transposase to the same degree in all samples independently of the stage of development (Figure 1A). The identified THSS regions in all samples are listed in Supplementary Table S2. The number of peaks and the intensity of the ATAC-seq signal is higher in the trophozoite stage (10G: 2229, 1.2B: 2158), followed by schizonts (10G: 1062, 1.2B: 1080), late rings (10G: 950, 1.2B: 2223) and early rings (10G: 613, 1.2B: 1042) (Supplementary Figure S1C and Table S2). The number of peaks in 1.2B ring stage parasites is higher compared to 10G, but the ATAC-seq enrichment levels are similar. Such differences are possibly due to slightly lower peak calling performance in the 10G subclone (Supplementary Figures S1C and 7B).

Distance analysis of ATAC-seq data shows that THSS regions are mostly located in a window of 1 Kb upstream from the translation start codon (ATG) and are scarce at distances greater than 2 Kb of the nearest gene (Figure 1B and Supplementary Figure S1D). This is in agreement with previous studies in *P. falciparum* showing that TSSs preferentially localize 0.5–1 Kb upstream from the ATG (48). The ATAC-seq signal is significantly enriched at previously annotated TSSs (48,66) and matches the typical profiles described in higher eukaryotes (3,75,76) (Figure 1C and Supplementary Figure S1E). Based on these findings, we annotate THSSs as promoters if the peak is located less than 1 Kb upstream from the corresponding ATG (Figure 1D and Supplementary Figure S1F); and use this annotation for all subsequent analysis (Supplementary Table S2).

It is known that chromatin features such as histone modifications exert a significant impact on the compaction of chromatin (77,78). In *Plasmodium*, as in other eukaryotes, active histone modifications such as H3K9ac and H3K4me3 have been shown to be enriched in intergenic regions and colocalize with H2A.Z (26). The association of these histone modifications with transcription varies through the IDC: H3K9ac has been shown to strongly correlate with transcription all over the asexual cycle, whereas H3K4me3 is enriched in schizonts regardless of the temporal expression dynamics. Thus, we expected a genome-wide stage-dependent relationship between these active chromatin features and chromatin accessibility by ATAC-seq. Using publicly available data (26), we observed that H3K9ac, H3K4me3 and H2A.Z, are enriched around the

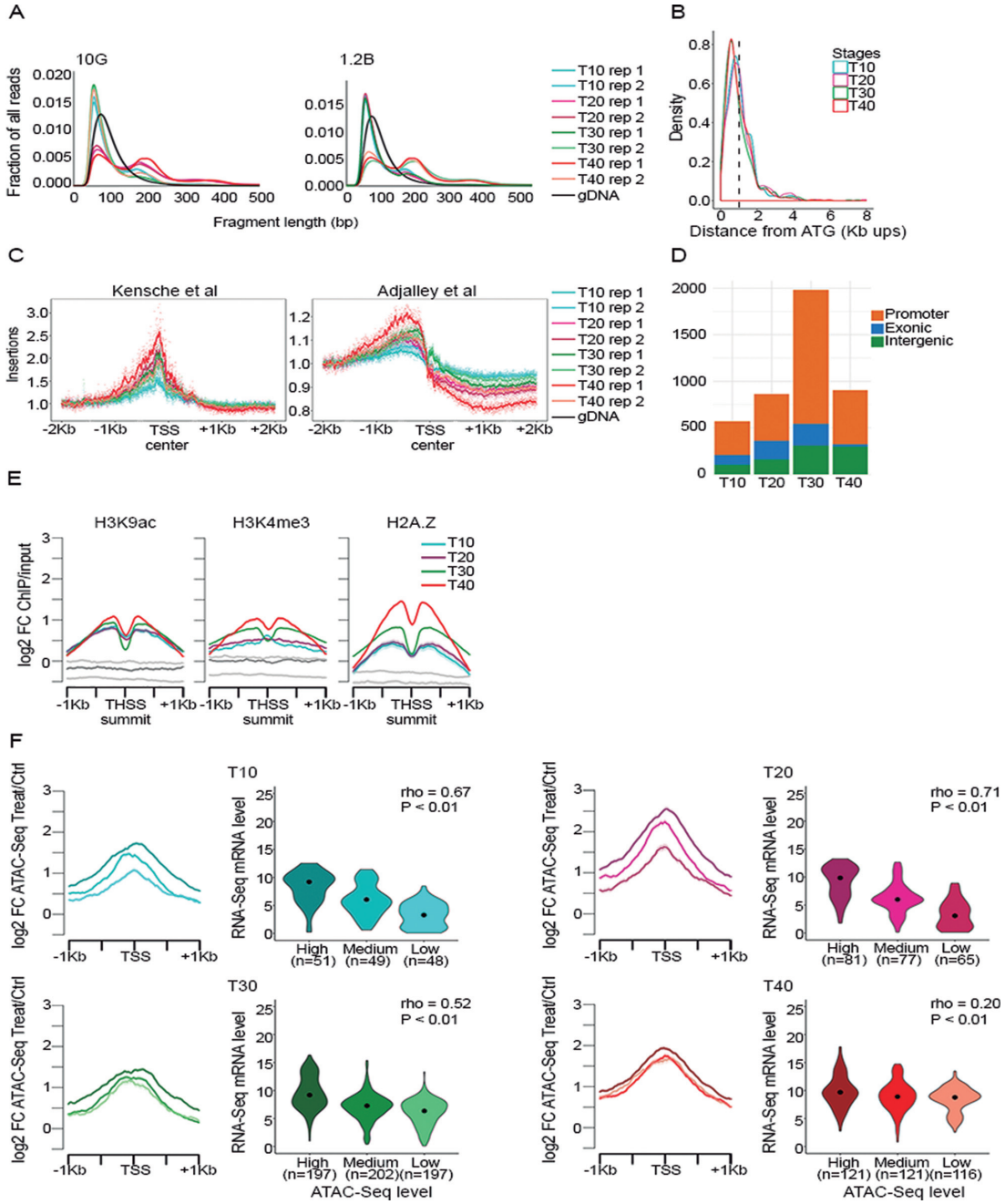


**Figure 1.** ATAC-seq in *Plasmodium falciparum*. Relationship between chromatin accessibility, histone modifications and gene expression. (**A**) ATAC-seq fragment size distribution at various developmental stages during the intra-erythrocytic developmental cycle for the 10G and 1.2B subclones. To determine the location of THSSs, we selected reads in the range of 50–130 bp corresponding to sub-nucleosomal ATAC-seq signal. (**B**) Density plot showing the position of THSSs per stage of development with respect to the ATG start codon. The dashed line indicates the putative promoter region located 1 Kb upstream. (**C**) Input corrected ATAC-seq enrichment at known TSSs (Left: 2271 TSSs by Kensche *et al.* (48), Right: 39667 TSS blocks by Adjalley *et al.* (66)) in a region comprising ±2 Kb. The average profile is shown in colored-lines to account for dispersion. ATAC-seq enrichment corresponds to different stages of the 10G subclone. (**D**) Annotation of THSSs to genomic features: Promoters, Intergenic regions and Exons. THSSs located up to 1 Kb upstream of the ATG are listed as Promoter region. (**E**) Profile plots showing the density of normalized (RPKM) and input-corrected unique reads for various histone modifications at THSSs identified in the 10G subclone. ChIP-seq data were obtained from (26). The region plotted comprises ±1 Kb around the summit. Profiles in gray represent signal at random coordinates. (**F**) ATAC-seq signal at TSSs correlates quantitatively with gene expression. Profile plots (left) show changes in levels of ATAC-seq subnucleosomal reads during the IDC in the 10G subclone. Genes are divided into three groups and ranked by their RNA-seq mRNA levels based on published data (48). The graphs represent normalized (RPKM) and input-corrected ATAC-seq reads mapped with respect to the TSSs (48). For each graph, the darkest color represents the highest mRNA levels and the lightest color represents the genes with lowest expression level. Violin plots (right) show the distribution of RNA-seq mRNA levels (48) for genes grouped by their level of ATAC-seq nucleosome-free signal at promoters (1 Kb upstream from the ATG start codon). Promoters are grouped in three classes: high, medium and low chromatin accessibility levels. Plot width accounts for the density of certain repeated values in the range. Median values of each group are marked with a black dot. The mRNA values are shown in log2 scale. The Spearman rank correlation coefficient (rho) and the corresponding *P*-value are shown for the relationship between the ATAC-seq signal at the promoter and the RNA-seq mRNA levels of the THSS-annotated gene.

THSSs. Furthermore, the association between these epigenomic features is stage-specific, being more marked in schizonts (Figure 1E and Supplementary Figure S2A).

Following the observation that the majority of THSS regions overlap with active histone marks at TSSs, we examined the quantitative relationship between chromatin accessibility and transcription genome-wide. To this end, using a public RNA-seq dataset (48), we computed the ATAC-seq signal level at known TSSs (48) in groups of genes with low, medium, and high mRNA abundance. We observe that the level of enrichment in ATAC-seq signal at the promoter is positively associated with mRNA levels of the annotated gene (Figure 1F and Supplementary Figure S2B). The same pattern is observed when categorizing promoter regions by the strength of the ATAC-seq signal into low, medium and high groups (same grouping procedure as above for the RNA-seq data above) and representing the mRNA levels of THSS-annotated genes in each ATAC-seq enrichment group (Figure 1F and Supplementary Figure S2B).

### Chromatin accessibility footprints identify novel regulatory elements in *P. falciparum*

Chromatin accessibility represents a novel and useful tool to identify novel regulatory sequences including TSSs and enhancers (79). Then, different histone post-translational modifications (hPTMs) can be used to distinguish between these elements. For example, in higher eukaryotes, histone marks H3K4me3/H3K9ac are characteristic of TSS/promoter regions, whereas H3K4me1 is a hallmark of enhancer regions, which are typically depleted in H3K4me3/H3K9ac (80–82). Based on this assumption, we investigated whether the THSS regions correspond to TSS-like or to enhancer-like regions. To this end, we first used a chromatin-state segmentation (ChromHMM) approach (70) to partition the *P. falciparum* genome based on the relative enrichment of several reported histone marks (59). Since available data for H3K4me1 exists only for trophozoite and schizonts (59), we restricted our analysis to 2273 merged THSSs from these two stages in both subclones (Supplementary Table S2). ChromHMM classified the genome into four distinct chromatin states (Figure 2A): unspecific (high levels of enrichment for all hPTMs), TSS-like (enriched in H3K4me3/H3K9ac), enhancer-like (enriched in H3K4me1 and depleted in H3K4me3/H3K9ac) and depleted (low levels of enrichment for all hPTMs). In overlapping these genomics segments with the THSS regions identified by our ATAC-seq approach (Figure 2B), we find 558 THSSs that have signatures typical of TSSs and 50 that have a chromatin state typical of enhancers (Figures 2B–D and Supplementary Table S3).

The majority of the ATAC-seq peaks with TSS-like signatures (523/558) are located <2 Kb from a downstream transcript (Figure 2E), whereas there are 35 TSS-like regions that locate more than 2 Kb upstream (Supplementary Table S3). As expected many of the TSS-like elements (269 out of the 558) coincide with a TSS previously described by others (48,66). As for the rest, we find 250 THSSs located in the promoter of genes with a known TSS, but that does not coincide with the ATAC-seq peak. We also find 12 regions that are located in the promoter of genes with no previous annotated TSS, and 27 putative novel TSS-like elements located in non-promoter regions (>2 Kb upstream of the ATG). For the 289 putative novel TSSs, the positive and significant correlation between the ATAC-seq signal enrichment and the mRNA levels of the associated gene in trophozoites suggest that these THSSs are likely to be previously unidentified TSSs (Figure 2F and Supplementary Figure S3A).

Contrary to the TSS-like elements, 24 out of the 50 enhancer-like regulatory elements are located at distances >2 Kb upstream of the ATG (Figure 2E and Supplementary Table S3). There are 12 out of the 50 enhancer-like regions that coincide with known TSSs (48,66). Amongst the remaining enhancer-like elements, we find 35 that are located upstream of genes with one or several known TSSs. Most of the known TSSs (161/172) that are associated with one of the novel enhancer-like element are not accessible. The more distal enhancer-like regions are commonly found at the ends of chromosomes, occupying subtelomeric repeats such as the TAREs and sometimes very far from the nearest gene. This is likely because TAREs are highly transcribed (and thus presumably contain promoters and enhancers) (42,83). Interestingly, we find that 22 subtelomeric *var* genes have enhancer-like THSS regions. At last, the study by Ubhe *et al*. (33) proposed a set of enhancer-like sequences based on the localized enrichment of H3K4me1 versus H3K9ac and H3K4me3 in intergenic regions. However, we do not observe a noticeable ATAC-seq enrichment around those sites, and none of the intergenic regions classified as enhancer-like coincide with our enhancer-like dataset (Supplementary Figure S3B). Contrary, 12% (57) of the enhancers described by Ubhe *et al.* intersect with 72 THSSs classified as depleted (44), unspecific (6) and TSS-like (22) by the ChromHMM analysis. We consider that our analysis combining chromatin accessibility profiling and histone modification enrichment is a much more precise and powerful approach in identifying new trans and cis-regulatory sequences in *P. falciparum*.

### Temporal variation in chromatin accessibility is predictive of stage-specific transcription

*Plasmodium falciparum* undergoes drastic switches in transcription linked to life cycle progression. Our results indicate that chromatin accessibility is dynamic and varies between stages of development (Figure 3A and Supplementary Figure S1C). To further investigate the relationship between temporal changes in chromatin accessibility and transcription through the cell cycle, we applied a soft clustering approach using Mfuzz that group promoters based on their temporal accessibility through the cell cycle. In this analysis we observe clusters of dynamic promoters accessible at different stages during development (Supplementary Figure S4). We then determined the temporal dynamics in mRNA levels of the corresponding promoter-annotated gene (6) and the H3K9ac/H3K4me3 enrichment at these open regions (26) (Figures 3B–F and Supplementary Figures S5A–E). We find a strong and significant positive correlation between changes in chromatin accessibility during development and the corresponding changes in gene expression (Figures 3B–D; Supplementary Figures S5A–C and Table S4). The Gene Ontology terms enriched at each

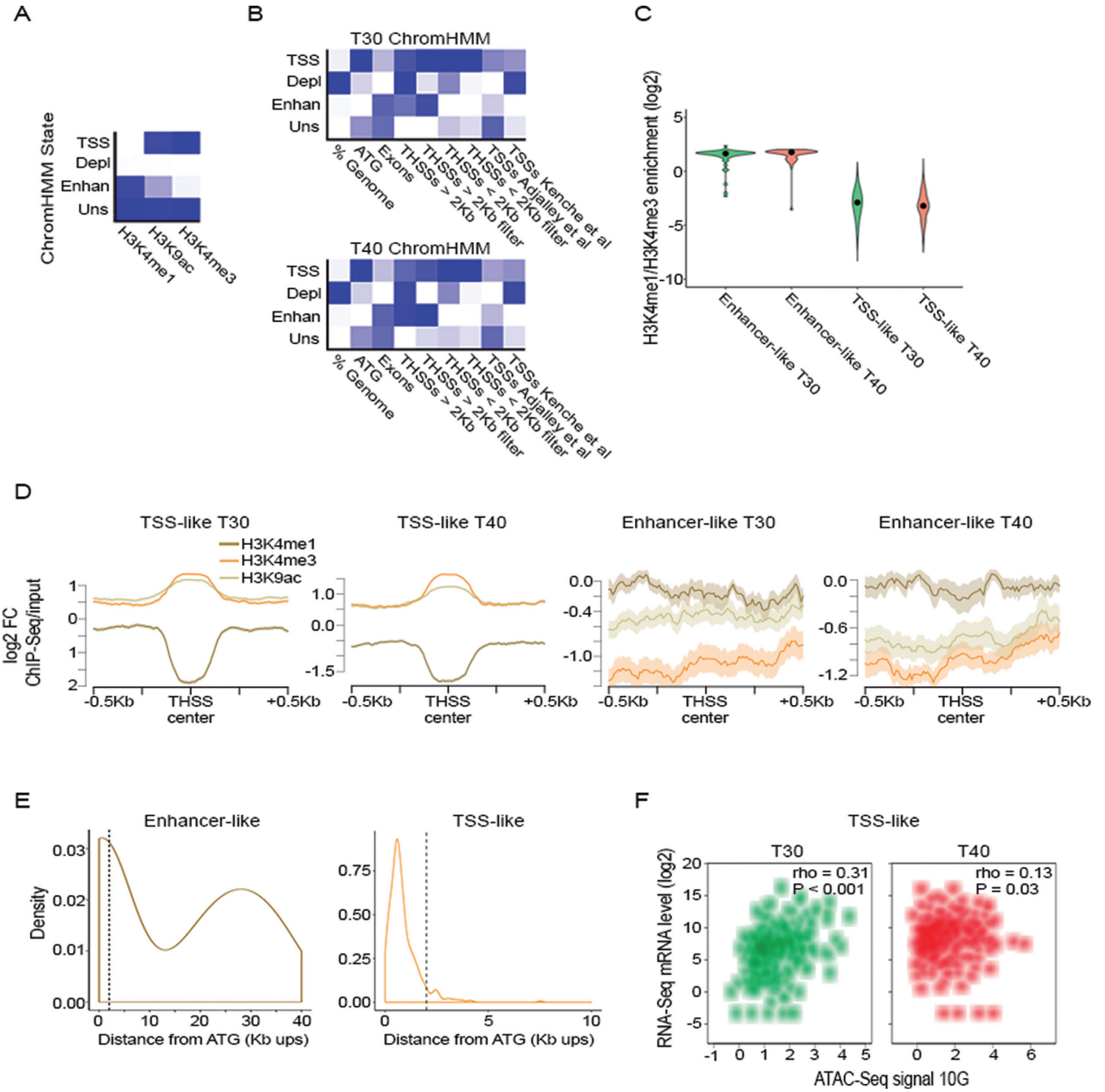


**Figure 2.** Characterization of novel regulatory regions in *Plasmodium falciparum*. (**A**) Heatmap of emission parameters from the ChromHMM analysis for a four chromatin-state model based on histone modification enrichment patterns: TSS (H3K4me3/H3K9ac enrichment), Depl (low levels of all hPTMs), Enhan (H3K4me1 enrichment) and Uns (high levels of all hPTMs). ChIP-seq data was obtained from (59). We restrict our analysis to trophozoite and schizonts. Darker blue indicates higher enrichment of a particular histone modification. (**B**) Heatmaps show the overlap of various genomic features, including THSSs and known TSSs (48,66), with the predicted chromatin states obtained for trophozoite and schizont stages. Darker blue indicates higher likelihood of pertinence to a particular chromatin state. (**C**) Violin plots showing the ratio of H3K4me1 to H3K4me3 enrichment (59) for THSS regions classified as enhancer (left) or TSS-like (right) in trophozoite and schizont stages. The ratio is log2-scaled. Median values of each group are marked with a black dot. Positive values indicate a higher enrichment of H3K4me1 relative to H3K4me3, and negative values the opposite. (**D**) Average profile plots showing H3K4me3, H3K9ac and H3K4me1 enrichment (59) at novel THSS regions in trophozoite and schizont stages that have been classified as TSS (left) and enhancer-like elements (right). The graphs represent normalized (RPKM) and input-corrected ChIP-seq read unique counts mapped 0.5 Kb with respect to the center of the THSS region. Confidence intervals are shown in light-colored shades. (**E**) Density histogram showing the distribution of distances to the ATG start codon for THSSs characterized as enhancer (left) or TSS-like (right). (**F**) Scatter plots showing the relationship between gene expression and chromatin accessibility at THSSs categorized as TSS-like by ChromHMM that do not coincide with previously annotated TSSs (48,66). ATAC-seq enrichment at the novel TSSs is compared with RNA-seq mRNA levels (48) of the corresponding downstream gene. Data corresponds to trophozoite (left) and schizont (right) stages for the 10G subclone. ATAC-seq enrichment and mRNA levels are in log2 scale. The Spearman rank correlation coefficient (rho) and the corresponding *P*-value are shown.

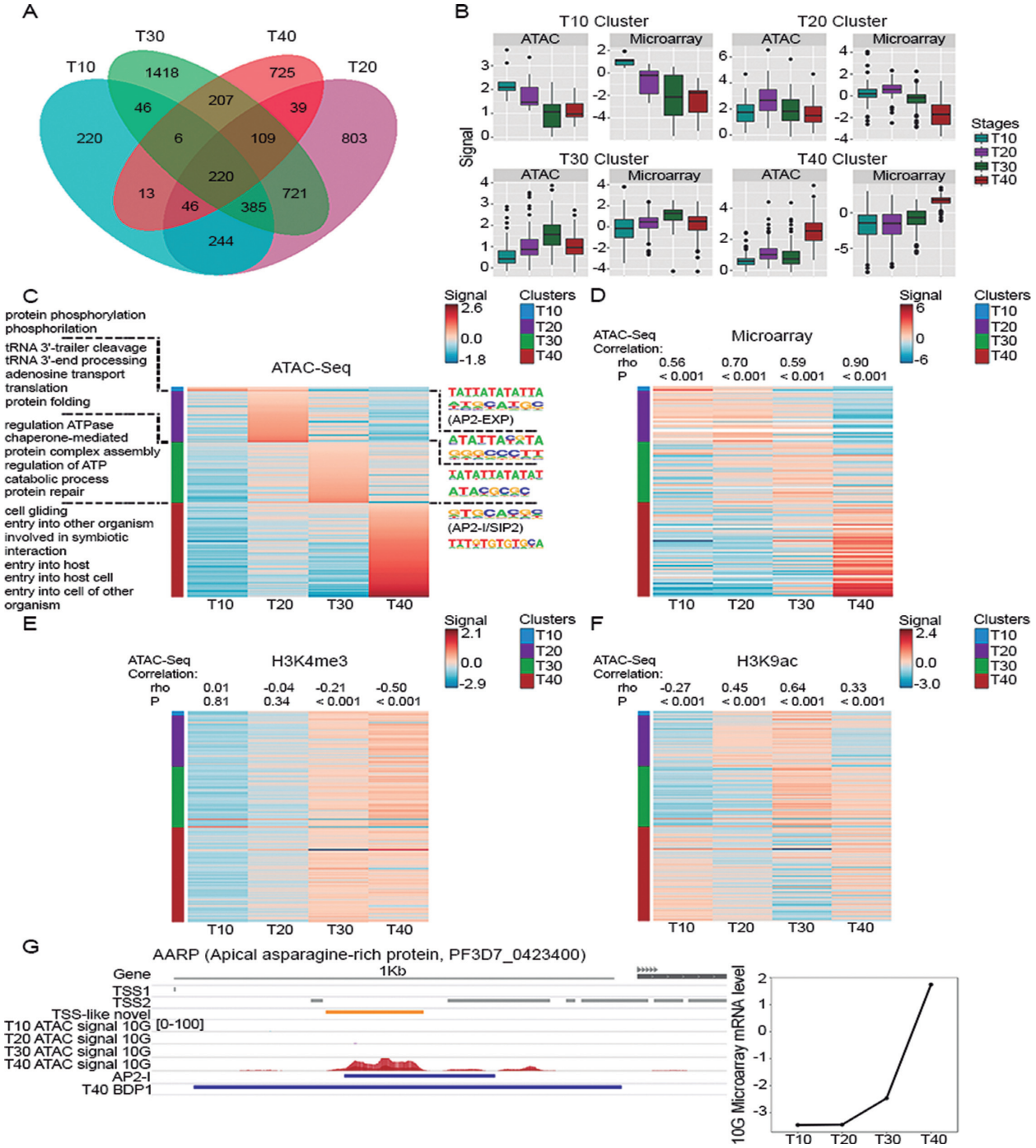


**Figure 3.** Temporal dynamics of chromatin accessibility during *Plasmodium falciparum* intraerythrocytic development. (**A**) Venn diagram showing the overlap between THSS regions identified in each developmental stage for each subclone independently. (**B**) Boxplots showing the distribution of input-corrected ATAC-seq enrichment (left) and microarray mRNA levels (right), during the IDC for clusters obtained using Mfuzz in 10G. Microarray data was obtained from (6). (**C**) Temporal changes in chromatin accessibility during the IDC in 10G. The heatmap shows ATAC-seq enrichment at accessible promoters organized by stage-specific groups. ATAC-seq enrichment is normalized (RPKM) and input-corrected. Top significant Gene Ontology terms are listed. Top stage-specific motifs enriched in each group identified by HOMER are indicated. (**D**) Heatmap based on microarray mRNA levels (6) for the set of THSS-annotated genes ordered by stage-specific cluster and ATAC-seq enrichment as in (C). Data in the heatmap is log2-scaled and mean centered. The Spearman rank correlation coefficient (rho) and the corresponding *P*-value are shown above columns for the association between the ATAC-seq enrichment at the promoter and mRNA levels (6) of the nearest downstream genes at each stage. (**E** and **F**) Heatmaps based on levels of H3K4me3 (E) and H3K9ac (F), using ChIP-seq data obtained from (26), at promoters for the set of THSS-annotated genes ordered by stage-specific cluster and ATAC-seq enrichment as in (C). Enrichment data in the heatmaps is log2-scaled and mean centered. The Spearman rank correlation coefficient (rho) and the corresponding *P*-value are shown above columns for the association between ATAC-seq levels and enrichment of H3K4me3 (E) and H3K9ac (F) at promoters at each stage. (**G**) Changes in chromatin accessibility and gene expression in the *aarp* gene (PF3D7_0423400). Tracks show normalized and input-corrected ATAC-seq signal. The location of TSSs (48,66) and the binding sites for AP2-I and BDP1 (41,62) are shown. All tracks are shown at equal scale. The plot at the right shows temporal changes in microarray mRNA levels (log2 scale) (6) during the IDC for this gene.

stage-specific cluster agree with the gene function prediction expected based on previous transcriptomic and proteomic analyses (25,84) (Figure 3C; Supplementary Figure S5B and Table S4). On the other hand, we also detect an association between accessibility and the level of enrichment of H3K9ac throughout the cell cycle. The coefficient is lower but significant for the correlation between ATAC-seq and H3K4me3 enrichment in all stages except at T40 (Figures 3E and F; Supplementary Figures S5D and E). There are multiple examples of this nature, such as the gene encoding the apical asparagine-rich protein (PF3D7_0423400) with an ATAC region only in schizonts that correlates with the expression of the gene that peaks at this developmental stage (Figure 3G). We also report promoters that contain multiple accessible regulatory sequences opening at different times during development (Supplementary Table S4). For example, the gene encoding the early transcribed membrane protein 5 (ETRAMP5) displays an invariant peak less than 1Kb upstream, and two peaks: one distal >4 Kb upstream and one proximal at the 5' start of the gene, that correlate with the transcriptional status of the gene (Supplementary Figure S5F). All this considered, we show that by combining ATAC-seq and expression data during the IDC cycle, we can identify functionally related genes that show dynamic and coordinated chromatin accessibility and gene expression at different developmental stages.

### Stage-specific motif inference at THSS regions identifies key developmental regulators

The dynamics in chromatin accessibility at TSSs and enhancers has been shown to capture DNA–protein binding events that lead to promoter activation *in vivo* (2,85). To characterize the regulatory networks controlling developmental transitions during the IDC, we first performed *de novo* motif discovery and similarity analysis on the set of developmentally dynamic promoters identified above (Mfuzz clusters, Supplementary Table S4). Amongst the top enriched motifs, we find potential binding sites for several ApiAP2 TFs including AP2-SP/AP2-EXP (10 and 40 h) and AP2-I (30 and 40 h), among others (48,66). However, we also report *de novo* predicted motifs (Figure 3C; Supplementary Figure S5B and Tables S4 and 5). In general, there is a correlation between the developmental abundance of the predicted TF and temporal enrichment of the corresponding binding site (Figure 3C; Supplementary Figure S6A and Table S5) (47). Next, we computed the occurrences of each stage-specific motif in the set of ATAC-seq peaks that are associated to each Mfuzz cluster (Supplementary Table S5), and used Cytoscape to construct gene regulatory networks that connect TF predicted binding sites with the target genes (Figure 4A and Supplementary Figure S6B). This analysis reveals co-regulated genes and interactions between predicted TFs. For example, looking at the regulatory network of late blood stage parasites (T40), we identify a number of invasion genes like *rap*, *ama*, *rama* and members of the *msp*, *clag*, *rhop* and *Rh* gene families that contain one or multiple T40-specific DNA motifs in the promoter (Figures 4A–C). We also observe that the AP2-I cognate motif GTGCACGC (named 40.1 in Supplementary Table S5) plays a central role in this gene network displaying the highest number of connections, i.e. 178 genes contain the motif (Figures 4A and B). These observations suggest that different TFs might act coordinately to control their gene expression.

As a validation for our strategy to investigate TF binding events *in vivo*, we report the overlap between the ATAC-seq peaks and the binding sites of the TFs: AP2Tel, SIP2, AP2-I and other proteins such as BDP1, which have been determined experimentally (35,41,42,62) (Supplementary Table S6). In the case of AP2-I, we find that 86% of the binding sites intersect with ATAC-seq peaks (Supplementary Table S6). Importantly, the enrichment of ATAC-seq peaks at the binding site of AP2-I occurs only in late stages, when the activation of invasion genes takes place (Figure 4B and Supplementary Figure S6C). Together with AP2-I, the Bromodomain BDP1 has been described as a transcriptional activator that binds to chromatin at invasion gene promoters (62). We find 402 ATAC-seq peaks that intersect with 273 out of 799 mapped binding sites (Supplementary Table S6). An example of a target gene co-regulated by BDP1 and AP2-I is the merozoite surface protein 2 (MSP2) gene, that in our study displays a THSS region in the promoter with the GTGCA motif (Figure 4C).

All this considered, the temporal analysis of ATAC-seq data allows us to infer the regulatory DNA sequences, the TFs and the target genes that constitute the regulatory networks controlling gene expression during blood stage developmental transitions. Our results support a model in which the promoter region can be bound by one or multiple stage-specific TFs that would operate coordinately to activate different functional sets of genes.

### Differential chromatin accessibility associated with clonally variant gene expression

The 10G and 1.2B subclones of 3D7 have previously been shown to differ in the expression of several CVGs (6). We used DESeq to test for clonal variation in ATAC-seq enrichment using a 2-fold threshold for the difference between subclones. Using this criterion, we identified 140 statistical significant differentially-accessible regions between the 1.2B and 10G lines, of which 38 ATAC-seq peaks (15 specific for 1.2B and 23 for 10G) are located in gene promoters (−2 Kb from the ATG, Supplementary Table S7). Genes with differential ATAC-seq peaks include well-characterized CVGs such as *clag2*, *clag3, eba140* and members of the *var* and *phista* families (Figures 5A–C; Supplementary Figures S7C–E and Tables S7 and 8) (6,8).

Taking the list of the top-30 most differentially expressed genes between 1.2B and 10G (6), we observe a clear and significant association between differential gene expression and differential accessibility, such that an increase in mRNA levels in one subclone is generally associated with an increased accessibility in the promoter (Figure 5B). For example, the promoters of the well-characterized clonally variant invasion genes *eba140* and *clag2* genes, or a clonally variant *phista* (15,86), display various open regulatory regions only in the subclone in which the gene is highly expressed (Figure 5C; Supplementary Figure S7E and Table S7). However, we also find examples of top CVGs in which chromatin accessibility remains invariant despite the gene

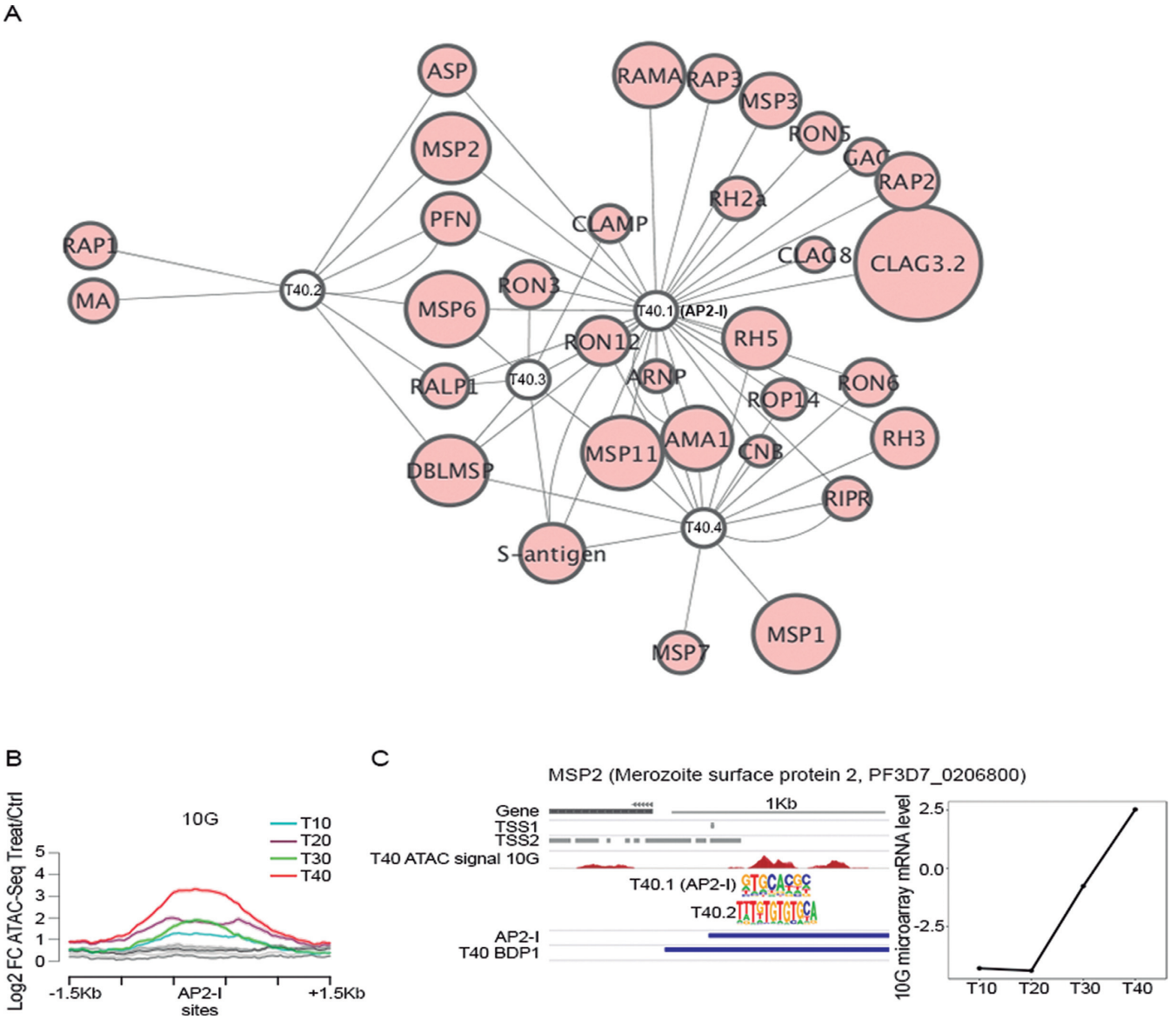


**Figure 4.** ApiAP2 TFs linked to temporal chromatin accessibility profile transitions during *Plasmodium falciparum* intraerythrocytic development. (**A**) Cytoscape network showing regulatory interactions between predicted motifs identified by HOMER in the set of schizont-specific active regulatory sites located in the promoters of invasion genes. The size of the circles is proportional to microarray mRNA levels (6) of the target gene at that particular stage of development. White color nodes represent the motifs (T40.1, T40.2, T40.3 and T40.4) named as in Supplementary Table S5. Gene symbols are indicated. In the figure, T40.1 corresponds to the AP2-I cognate motif that is present in regulatory active sites of *clag3*, *msp*, *rap* and *rama* gene families among others. (**B**) Average profile plot showing density of ATAC-seq reads at AP2-I binding sites previously characterized (34). The graphs represent normalized (RPKM) and input-corrected ATAC-seq reads mapped ±1.5 Kb of the peak region. Lines plotted correspond to different developmental stages of the 10G subclone. Profiles in gray represent signal at random coordinates. (**C**) Example of a developmentally regulated gene encoding the MSP2 protein (PF3D7_0206800), showing a stage variant THSS region that coincides with the AP2-I binding site. Tracks show normalized and input-corrected ATAC-seq signal in the 1 Kb upstream region. The location of known TSSs (48,66) and the binding sites for AP2-I and BDP1 (41,62) are included. All tracks are shown at equal scale. The plot at the right shows microarray mRNA levels (log2 scale) (6) during the IDC.

being variantly expressed. This is a set of 14 genes from multicopy gene families: nine *rif*, four *var* and one *pfmc-2TM* genes and five genes from other families. Except for the *var* genes (see below) the majority of these cases are genes that are not accessible in either clone (Supplementary Table S8), suggesting that for these genes differential expression is attributable to activation of the gene in only a small fraction of the parasites in one of the subclones.

### Dynamics of regulatory regions in mutually exclusively expressed *clag3* and *var* genes

The *clag3* genes are an example of CVGs that are expressed in a mutually exclusive manner, so only one copy is expressed at a time in individual parasites, either *clag3.1* or *clag3.2* (15,16,86–88). These genes are also developmentally regulated, thus when active, they are maximally expressed at the schizont stage. Our results show that the variant expression of *clag3* genes is associated with differential

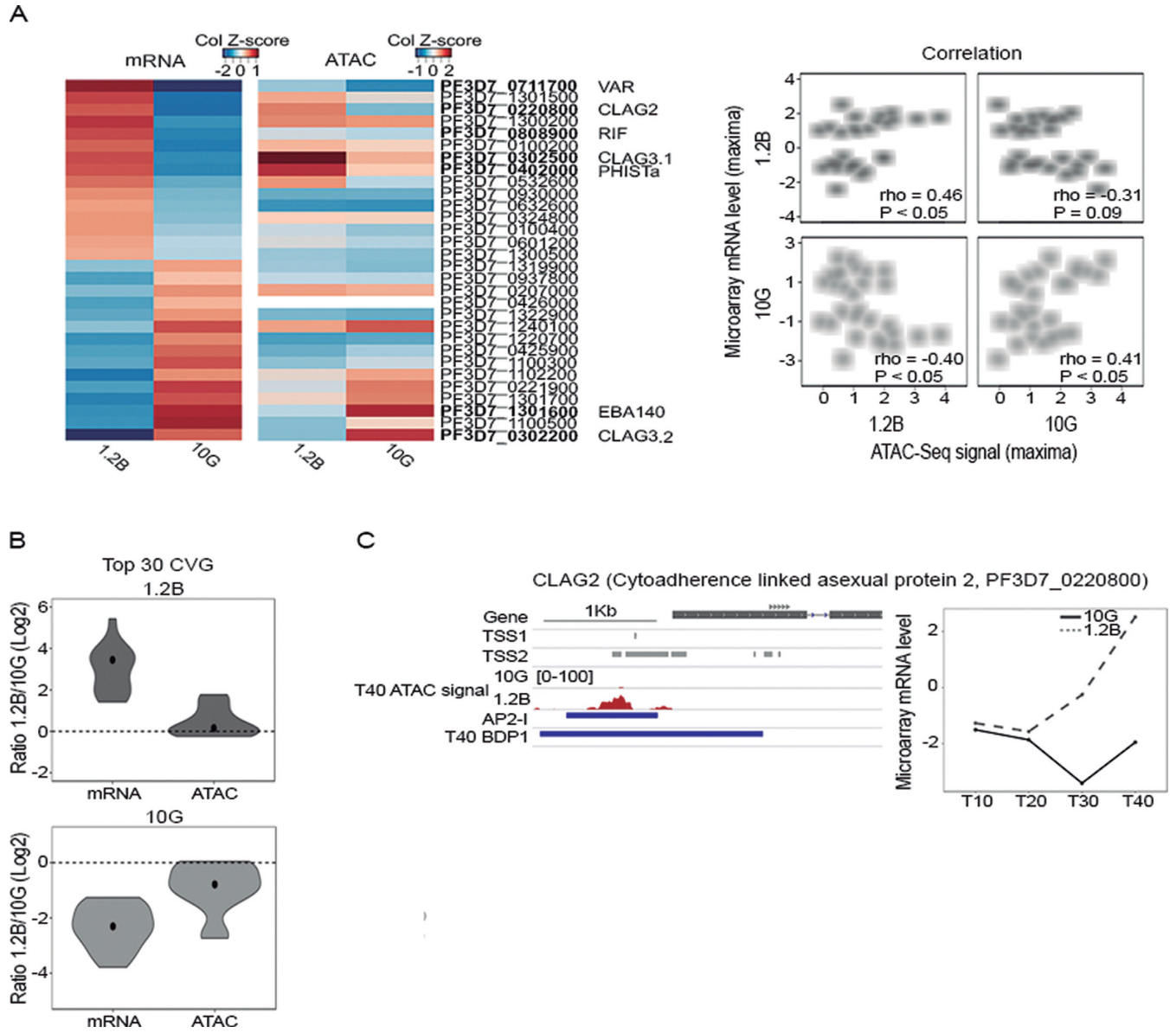


**Figure 5.** Clonal differences in chromatin accessibility and gene expression. (**A**) ATAC-seq enrichment at the top-30 most variable CVGs previously characterized (6). Graph shows normalized microarray mRNA levels in 1.2B and 10G subclones (6) and the corresponding normalized (RPKM) and input-corrected ATAC-seq reads in the promoter regions (1 Kb upstream of the ATG). Genes are ordered by mRNA levels. Values are the mRNA levels and ATAC-seq enrichment maxima through the IDC. Case examples shown in Figures 5–6 and Supplementary Figures S7 and 8 are marked in bold. The scatter plots show ATAC-seq enrichment at the promoter and mRNA levels of the peak annotated gene in each subclone. The Spearman rank correlation coefficient (rho) and the corresponding *P*-value are shown. (**B**) Violin plots showing the 1.2B/10G ratio for mRNA levels (6) and ATAC-seq enrichment at promoters for the top-30 most variable clonally variant genes previously characterized (6). The ratio is log2-scaled. Median values of each group are marked with a black dot. Positive values indicate higher levels for the 1.2B subclone and negative values indicate higher levels for the 10G subclone. (**C**) Example of changes in chromatin accessibility in the top clonally variant expressed CLAG2 (Pf3D7_0220800) encoding gene (6). Tracks show normalized/input-corrected ATAC-seq signal in the 1 Kb upstream region. The location of TSSs (48,66) and the binding sites for AP2-I and BDP1 (41,62) are included. All tracks are shown at equal scale. The plot at the right shows mRNA levels (log2 scale) (6) for each subclone during the IDC.

chromatin accessibility in the promoter region. The *clag3.1* paralog (PF3D7_0302500), which is expressed in the 1.2B line and silenced in 10G, shows a clear opening in the promoter that coincides with a known TSSs block (48,66,86) only in the 1.2B subclone (Figure 6A). The considerably smaller ATAC-seq peak observed in the 10G subclone likely corresponds to a subpopulation of parasites that switched from *clag3.2* to *clag3.1* expression during normal growth (88). The *clag3.1* active regulatory region has motifs similar to binding sites for the ApiAP2 PF3D7_1222400, and also overlaps with ChIP-seq peaks for AP2-I and BDP1 (Figure 6A and Supplementary Tables S4 and 5). In an analogous situation, the *clag3.2* paralog (PF3D7_0302200) is expressed only in 10G and accessibility at the promoter region is only observed in this subclone. Of the two THSS regions in the *clag3.2* promoter, the proximal peak coincides with a

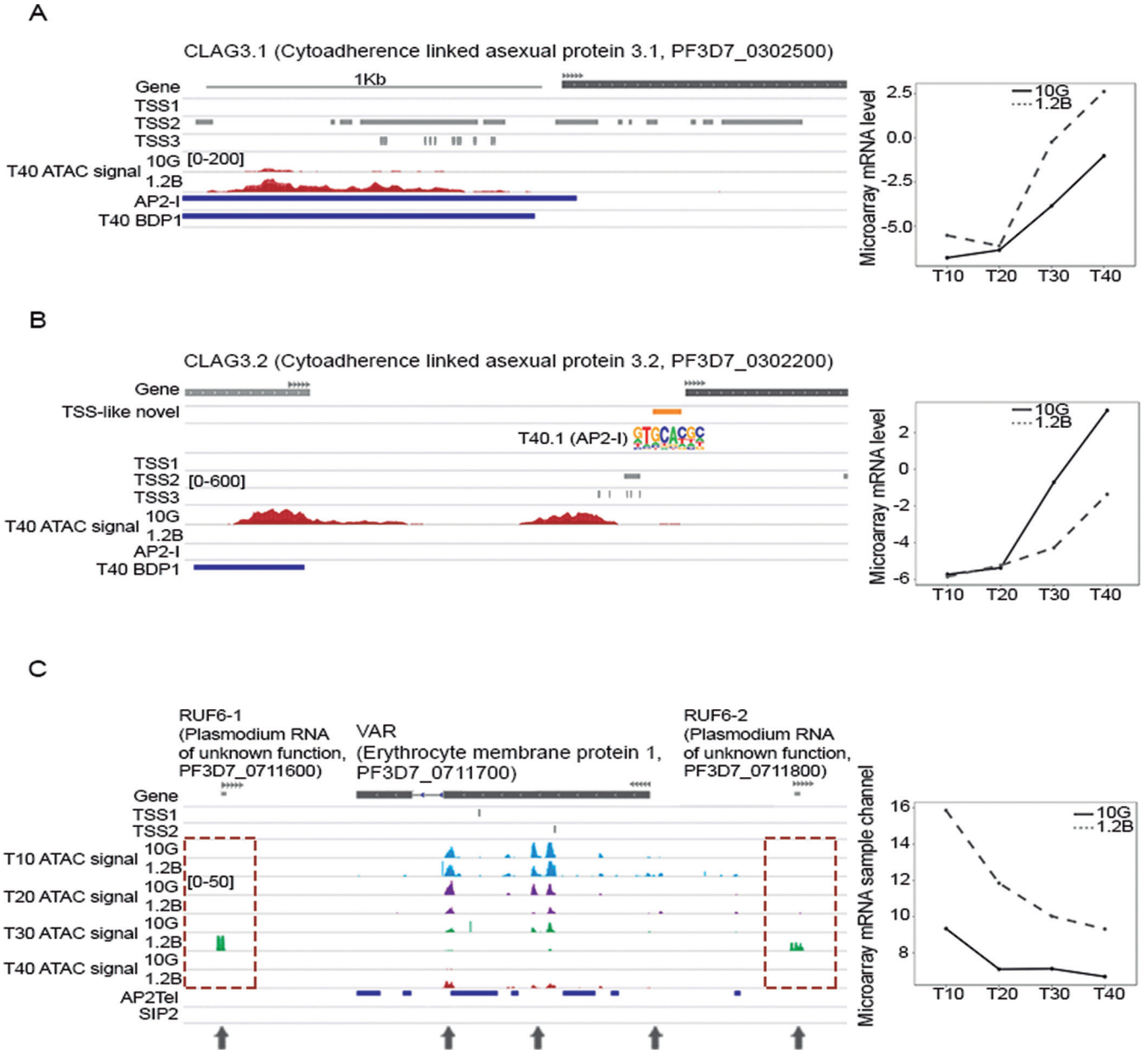


**Figure 6.** CVG expression and open and closed chromatin states at *clag3* and *var* genes. (**A** and **B**) Chromatin accessibility profiles in the region containing clonally variant *clag3.1* (PF3D7_0302500) and *clag3.2* (PF3D7_0302200) genes in chromosome 3. Tracks show normalized/input-corrected ATAC-seq signal in the upstream regions. The location of TSSs (48,66,86) and the binding sites for AP2-I and BDP1 (41,62) are shown. A DNA motif present in a stage-specific regulatory active site is included. A track with THSS regions with characteristics of novel TSS-like elements as in Figure 2 is shown. All tracks are shown at equal scale. The plots at the right represent microarray mRNA levels (log2 scale) (6) at each subclone during the IDC. (**C**) Chromatin accessibility profiles in the region of chromosome 7 containing the active *var* gene (PF3D7_0711700) that is expressed predominantly in 1.2B parasites. Two clonally variant accessible RUF6 encoding genes (PF3D7_0711600, PF3D7_0711800) flanking the 1.2B specific *var* are shown. Tracks show normalized/input-corrected ATAC-seq signal. The location of known TSSs (48,66) and the binding sites for SIP2 and AP2Tel (35,42) are included. All tracks are shown at equal scale. The plot at the right corresponds to mRNA levels data (log2 scale of microarray data sample channel, F635) (6) of the PF3D7_0711700 *var* at each subclone during the IDC.

known TSSs block. Both regions become highly accessible only in 10G schizonts (Figure 6B). The *clag3.2*-associated THSSs contain motifs that resemble the AP2-I binding site (GTGCAC), *de novo* predicted motifs (T30 stage-specific: ATACGCGC and TGCCCCTT) and also overlap with a BDP1 ChIP-seq peak (Supplementary Tables S4 and 5 and Figure 6B).

Another gene family that shows mutually exclusive expression is *var*. Previous studies using microarrays showed that the 1.2B subclone expresses only one predominant *var* gene, in contrast to the 10G subclone in which a dominantly expressed *var* was not identified (6) (Supplementary Figure S8A). The active *var* gene in 1.2B PF3D7_0711700 displays several ATAC-seq peaks in the putative promoter region, the intron, exon 1 and the downstream region (Figure 6C).

Only the two more distal peaks occur specifically in 1.2B, and both coincide with annotated members of the RNA of Unknown Function 6 (RUF6) gene family. This family includes 15 genes that encode for long ncRNAs known as GC-rich elements (89,90). Of these, only r*uf6-1* and *ruf6-2*, which are located flanking the 1.2B active *var,* become uniquely accessible in 1.2B trophozoites. All other members of the RUF6 family are non-accessible in 1.2B except for *ruf6-15,* located downstream of *var* PF3D7_1240900 and in which the upstream *ruf6-14* gene is not accessible and *ruf6-10* that flanks the 1.2B active *rif* PF3D7_0808900, one of the top-30 differentially expressed genes between 10G and 1.2B (6) (Supplementary Figure S8B).

Other *var* genes, which are silenced in the majority of parasites in both the 10G and the 1.2B subclones, also display numerous ATAC-seq peaks, some of which intersect with the binding sites of SIP2 and AP2Tel characterized in previous studies (Supplementary Figure S8C and Table S6) (91,92). The peaks that do not correlate with the active state of *var* genes may be related with the constitutive promoter activity of the intron (93,94).

Collectively, the analysis of ATAC-seq data supports a model of transcriptional regulation for *clag3* and *var* genes based on the differential use of multiple regulatory sequences. In the case of *var* genes, differential accessibility is observed at stages different from peak full-length *var* expression. Furthermore, we identify two ncRNAs (RUF6 1–2) flanking an active *var* gene that show clonally variant accessibility linked to the active state of the gene, and therefore could be involved in the mechanism of *var* mutually exclusive expression.

## DISCUSSION

Despite the high morbidity associated with *P. falciparum* infection, we still lack precise knowledge of how chromatin structure and dynamics define transcriptional regulatory networks in this human malaria parasite and which are the regulatory elements involved in differentiation and responses to the external environment. In this study, we performed genome-wide mapping of chromatin accessibility using ATAC-seq during the intraerythrocytic cycle and compared these profiles in two different laboratory subclones that differ in their epigenomic and transcriptional profiles.

Our first objective was to obtain a 'proof of concept' of the utility of ATAC-seq to map open chromatin regulatory regions in malaria parasites *in vivo*. Our results conform to the pattern reported in previous studies in other organisms by showing that most accessible regions localize up to 2 Kb upstream of the ATG and that open chromatin at these sites is predictive of active transcription (3,85,95). The integration of the ATAC-seq data with previously published ChIP-seq data for various histone marks has also allowed us to classify these regions in sites of transcription initiation (TSSs) and enhancers. Approximately half of the THSSs identified here coincide with known TSSs (48,66). In contrast, we find limited overlap with the enhancer regions identified by others (33). This is not unexpected because the method used previously to discover enhancers relies solely on a localized enrichment of H3K4me3/H3K4me1 in intergenic regions, whereas we define enhancers on the basis of both characteristic histone modifications and chromatin-accessibility patterns, as recommended by ENCODE (82). At last, our analysis also led to the high-throughput discovery of novel regulatory regions not only associated to IDC stage transitions but also linked to the clonally variant expression of virulence genes involved in host–parasite interactions.

Accessibility at regulatory regions is modulated by chromatin remodeling processes, such as histone modifications and TF binding, and it is involved in the regulation of transcriptional activity (77,78,82). The analysis of ATAC-seq data in *P. falciparum* reveals temporal alterations in the chromatin accessibility of the genome during development and a strong association between these changes and stage-specific gene expression. We also show that in general, the switch from the closed to the open state leads to gene activation. This is suggestive of a predominance of activators over repressors in *P. falciparum*. We also report a major overlap between the active regulatory sites and TF binding sites previously determined by others. For example, we show that ChIP-seq peaks of AP2-I (41), an ApiAP2 TF that has been shown to activate invasion genes in late stages, are significantly enriched in THSSs in a similar stage-specific manner (41). But how the temporality in the transcriptional regulation is achieved remains puzzling. Based on what we know from other organisms, in some cases TFs are permanently bound to their sites and poised before the target gene is actively transcribed. Another possible scenario is that binding occurs during DNA replication, and the promoters stay poised until the stage at which transcription of the gene is activated (96,97). Our results indicate that none of the two models generally apply in *P. falciparum.* Instead, our data shows very clear stage-specific chromatin accessibility profiles and changes in accessibility at stage transitions correlate with transcription changes. At last, among the predicted *cis*-regulatory sites that are responsible for inducing stage-specific gene expression, we captured the cognate motifs of several ApiAP2 TFs already known to play important roles in *P. falciparum* (47) and predict novel motif sequences. During the preparation of our manuscript, another study has been published that also conducted a similar ATAC-seq experiment in blood stage *P. falciparum* parasites, showing similar temporal accessibility dynamics and reporting the discovery of known and novel TF motifs (98). Collectively, this data supports the notion that the combined action of multiple TFs is most likely an important mechanism responsible for transcriptional regulation during *P. falciparum* blood-stage development.

In *Plasmodium*, as in other organisms, cell-to-cell heterogeneity in gene expression has been proposed to serve as a bet-hedging mechanism, allowing subpopulations of parasites to vary their phenotype and survive in a changing environment (6,99,100). Changes in chromatin structure involving nucleosome occupancy and histone marking, mainly H3K9ac/me3, have been associated with clonal variation in gene expression (12,13,15,28). However, the precise mechanism for the generation and maintenance of such variability remains poorly understood, as are the regulatory elements involved. To this end, in this study, we examined and compared the chromatin accessibility landscape

between two subclones of the same genetic background that display clonal variation in gene expression at several virulence genes with important functions in host–parasite interactions (6). This allowed us to identify genes that show differential accessibility in the promoter between subclones and in which the switch between the open/closed state is associated to a pattern of differential gene expression (6).

Similar to what we observe for stage-variant genes, a gain in accessibility in CVGs in one subclone is generally linked to the active state, but the accessibility status changes dynamically during the IDC development. These findings suggest that the alternate use of one or more *cis*-regulatory sequences in a subclone-specific manner underlies transcriptional heterogeneity (101). In agreement, we report overlap between clonally variant ATAC-seq peaks and known binding sites for putative repressors like SIP2 (35), and activators like AP2-I (41). We also predict binding sites for AP2-EXP which has been proposed to function as a repressor in *Plasmodium berghei* (102,103). It is generally accepted that heterochromatin/euchromatin transitions are an important mechanism in determining promoter accessibility in CVGs (8,28,86). But these transitions can also be driven by a transcription-based mechanism (104). Multiple binding of sequence-specific transcription factors in repetitive regions has been shown to modulate the production of regulatory ncRNAs that control heterochromatin formation in plants, yeast and mammals (105,106).

The *var* and *clag3* families comprise CVGs that show mutually exclusive expression and are differentially expressed between the two subclones studied here. In these two families, the silent transcriptional state has been linked to a loss of the activating marks H3K9ac and H3K4me3 in the promoter, and a gain of H3K9me3 and HP1 (28,86,87,107,108). Our data add new insights into the TF binding events linked to clonal transcriptional variation in these families. For *clag3* genes, we report subclone-specific opening of the promoter region in late-stage parasites that coincides with the active expression of the corresponding gene at this stage of development (41,62). Regarding *var* genes, we observed several accessible peaks in the upstream region, the exon and the intron of the single *var* gene expressed in the 1.2B subclone (6), but this pattern seems uncorrelated with the active state since it is similar in the 10G subclone (in which the gene is silenced) and resembles the profile observed at several other silent *var* genes. In contrast, we observed clonally variant accessibility of two RUF6 genes flanking the active *var* in 1.2B, such that they are accessible only when this *var* gene is active. Furthermore, other members of this gene family that are positioned adjacent to silent *var* genes in various chromosomes are not accessible by ATAC-seq. RUF6 genes encode GC-rich long ncRNAs have been proposed to be responsible for the *var* counting mechanism that drives mutually exclusive expression. In particular, it has been shown that these GC-rich elements are located at the perinuclear expression site of central and subtelomeric *var* genes, and that over-expression of RUF6 ncRNAs can activate *var* genes in *trans* and disrupts the monoallelic *var* expression pattern (89). Although with the data available this is still speculative, our data suggest the intriguing possibility that the binding of still unidentified factors to RUF6 genes may be a key event for *var* activation and play a role in controlling mutually exclusive expression. Future research should identify the factors involved and reconcile the reported activity in *trans* (89) with our observation that only the RUF6 genes flanking the active *var* gene are accessible.

In summary, open chromatin profiling by ATAC-seq represents a novel strategy to study accessibility dynamics and transcriptional regulation in malaria parasites. The application of this technique in *P. falciparum* allows us to identify the regulatory elements that are likely responsible for the temporal regulation of transcription *in vivo*. Furthermore, our ATAC-seq data contribute to close the loop for the association between TF binding, chromatin and transcriptional states, providing new insights into mechanisms of CVG regulation and mutually exclusive expression. The challenge ahead is to apply this approach to other parasite life-stages: mosquito, liver and sexual stages, and identify ways in which these regulatory landscapes can be manipulated for malaria eradication.

## DATA AVAILABILITY

ATAC-seq data are deposited in the GEO database under accession number GSE109599.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank V.G. Corcés for useful discussion and his guidance on the experimental design and data analysis. We are grateful to E. de la Calle-Mustienes and N. Rovira-Graells for laboratory technical assistance, and H.C. Liedtke for edits on the manuscript.

## FUNDING

Spanish Ministry of Economy and Competitiveness Grants [BFU2015-65000-R to E.G.-D., BFU2016-74961-P to J.L.G.-S., BFU2014-58449-JIN to J.J.T.]; Andalusian Government Grant [BIO-396 to J.L.G-S.]; European Research Council ERC Grant EvoLand [740041 to J.L.G.-S.]; Spanish Ministry of Economy and Competitiveness through the Agencia Estatal de Investigación (AEI), cofunded by the European Regional Development Fund (ERDF/FEDER), European Union (EU) [SAF2013-43601-R, SAF2016-76190-R]; Severo Ochoa Fellowship [BES-2016-076276 to J.L.R.]; Spanish Ministry of Economy and Competitiveness Ramon y Cajal Grant to E.G.-D. Unidad de Excelencia María de Maetzu [MDM-2016-0687]; Government of Catalonia, CERCA Program. Funding for open access charge: Spanish Ministry of Economy and Competitiveness Grant [BFU2015-65000-R].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.

2. Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.
3. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
4. WHO (2016) *World Malaria Report 2016*. World Health Organization, Geneva.
5. Otto,T.D., Wilinski,D., Assefa,S., Keane,T.M., Sarry,L.R., Böhme,U., Lemieux,J., Barrell,B., Pain,A. and Berriman,M. (2010) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Mol. Microbiol.*, **76**, 12–24.
6. Rovira-Graells,N., Gupta,A.P., Planet,E., Crowley,V.M., Mok,S., Ribas de Pouplana,L., Preiser,P.R., Bozdech,Z. and Cortes,A. (2012) Transcriptional variation in the malaria parasite Plasmodium falciparum. *Genome Res.*, **22**, 925–938.
7. Painter,H.J., Carrasquilla,M. and Llinas,M. (2017) Capturing in vivo RNA transcriptional dynamics from the malaria parasite Plasmodium falciparum. *Genome Res.*, **27**, 1074–1086.
8. Cortes,A. and Deitsch,K.W. (2017) Malaria epigenetics. *Cold Spring Harb. Perspect. Med.*, **7**, a025528.
9. Voss,T.S., Bozdech,Z. and Bartfai,R. (2014) Epigenetic memory takes center stage in the survival strategy of malaria parasites. *Curr. Opin. Microbiol.*, **20**, 88–95.
10. Dzikowski,R. and Deitsch,K.W. (2009) Genetics of antigenic variation in Plasmodium falciparum. *Curr. Genet.*, **55**, 103–110.
11. Scherf,A., Lopez-Rubio,J.J. and Riviere,L. (2008) Antigenic variation in Plasmodium falciparum. *Annu. Rev. Microbiol.*, **62**, 445–470.
12. Dzikowski,R., Li,F., Amulic,B., Eisberg,A., Frank,M., Patel,S., Wellems,T.E. and Deitsch,K.W. (2007) Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. *EMBO Rep.*, **8**, 959–965.
13. Chookajorn,T., Ponsuwanna,P. and Cui,L. (2008) Mutually exclusive var gene expression in the malaria parasite: multiple layers of regulation. *Trends Parasitol.*, **24**, 455–461.
14. Scherf,A., Hernandez-Rivas,R., Buffet,P., Bottius,E., Benatar,C., Pouvelle,B., Gysin,J. and Lanzer,M. (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J.*, **17**, 5418–5426.
15. Cortes,A., Carret,C., Kaneko,O., Yim Lim,B.Y., Ivens,A. and Holder,A.A. (2007) Epigenetic silencing of Plasmodium falciparum genes linked to erythrocyte invasion. *PLoS Pathog.*, **3**, e107.
16. Mira-Martinez,S., Rovira-Graells,N., Crowley,V.M., Altenhofen,L.M., Llinas,M. and Cortes,A. (2013) Epigenetic switches in clag3 genes mediate blasticidin S resistance in malaria parasites. *Cell Microbiol.*, **15**, 1913–1923.
17. Mira-Martinez,S., van Schuppen,E., Amambua-Ngwa,A., Bottieau,E., Affara,M., Van Esbroeck,M., Vlieghe,E., Guetens,P., Rovira-Graells,N., Gomez-Perez,G.P. *et al.* (2017) Expression of the plasmodium falciparum clonally variant clag3 genes in human infections. *J. Infect. Dis.*, **215**, 938–945.
18. Nguitragool,W., Bokhari,A.A., Pillai,A.D., Rayavara,K., Sharma,P., Turpin,B., Aravind,L. and Desai,S.A. (2011) Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. *Cell*, **145**, 665–677.
19. Sharma,P., Wollenberg,K., Sellers,M., Zainabadi,K., Galinsky,K., Moss,E., Nguitragool,W., Neafsey,D. and Desai,S.A. (2013) An epigenetic antimalarial resistance mechanism involving parasite genes linked to nutrient uptake. *J. Biol. Chem.*, **288**, 19429–19440.
20. Basore,K., Cheng,Y., Kushwaha,A.K., Nguyen,S.T. and Desai,S.A. (2015) How do antimalarial drugs reach their intracellular targets? *Front. Pharmacol.*, **6**, 91.
21. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.
22. Silberhorn,E., Schwartz,U., Loffler,P., Schmitz,S., Symelka,A., de Koning-Ward,T., Merkl,R. and Langst,G. (2016) Plasmodium falciparum nucleosomes exhibit reduced stability and lost sequence dependent nucleosome positioning. *PLoS Pathog.*, **12**, e1006080.
23. Lorch,Y., Maier-Davis,B. and Kornberg,R.D. (2014) Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev.*, **28**, 2492–2497.
24. Bozdech,Z., Llinas,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol.*, **1**, E5.
25. Le Roch,K.G., Johnson,J.R., Florens,L., Zhou,Y., Santrosyan,A., Grainger,M., Yan,S.F., Williamson,K.C., Holder,A.A., Carucci,D.J. *et al.* (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res.*, **14**, 2308–2318.
26. Bartfai,R., Hoeijmakers,W.A., Salcedo-Amaya,A.M., Smits,A.H., Janssen-Megens,E., Kaan,A., Treeck,M., Gilberger,T.W., Francoijs,K.J. and Stunnenberg,H.G. (2010) H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS Pathog.*, **6**, e1001223.
27. Gupta,A.P., Chin,W.H., Zhu,L., Mok,S., Luah,Y.H., Lim,E.H. and Bozdech,Z. (2013) Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum. *PLoS Pathog.*, **9**, e1003170.
28. Lopez-Rubio,J.J., Mancio-Silva,L. and Scherf,A. (2009) Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe*, **5**, 179–190.
29. Elemento,O., Slonim,N. and Tavazoie,S. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.
30. Militello,K.T., Dodge,M., Bethke,L. and Wirth,D.F. (2004) Identification of regulatory elements in the Plasmodium falciparum genome. *Mol. Biochem. Parasitol.*, **134**, 75–88.
31. Wu,J., Sieglaff,D.H., Gervin,J. and Xie,X.S. (2008) Discovering regulatory motifs in the Plasmodium genome using comparative genomics. *Bioinformatics*, **24**, 1843–1849.
32. Harris,E.Y., Ponts,N., Le Roch,K.G. and Lonardi,S. (2011) Chromatin-driven de novo discovery of DNA binding motifs in the human malaria parasite. *BMC Genomics*, **12**, 601.
33. Ubhe,S., Rawat,M., Verma,S., Anamika,K. and Karmodiya,K. (2017) Genome-wide identification of novel intergenic enhancer-like elements: implications in the regulation of transcription in Plasmodium falciparum. *BMC Genomics*, **18**, 656.
34. Yuda,M., Iwanaga,S., Shigenobu,S., Mair,G.R., Janse,C.J., Waters,A.P., Kato,T. and Kaneko,I. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol. Microbiol.*, **71**, 1402–1414.
35. Flueck,C., Bartfai,R., Niederwieser,I., Witmer,K., Alako,B.T., Moes,S., Bozdech,Z., Jenoe,P., Stunnenberg,H.G. and Voss,T.S. (2010) A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathog.*, **6**, e1000784.
36. Iwanaga,S., Kaneko,I., Kato,T. and Yuda,M. (2012) Identification of an AP2-family protein that is critical for malaria liver stage development. *PLoS One*, **7**, e47557.
37. Coleman,B.I., Skillman,K.M., Jiang,R.H.Y., Childs,L.M., Altenhofen,L.M., Ganter,M., Leung,Y., Goldowitz,I., Kafsack,B.F.C., Marti,M. *et al.* (2014) A Plasmodium falciparum histone deacetylase regulates antigenic variation and gametocyte conversion. *Cell Host Microbe*, **16**, 177–186.
38. Sinha,A., Hughes,K.R., Modrzynska,K.K., Otto,T.D., Pfander,C., Dickens,N.J., Religa,A.A., Bushell,E., Graham,A.L., Cameron,R. *et al.* (2014) A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. *Nature*, **507**, 253–257.
39. Martins,R.M., Macpherson,C.R., Claes,A., Scheidig-Benatar,C., Sakamoto,H., Yam,X.Y., Preiser,P., Goel,S., Wahlgren,M., Sismeiro,O. *et al.* (2017) An ApiAP2 member regulates expression of clonally variant genes of the human malaria parasite Plasmodium falciparum. *Sci. Rep.*, **7**, 14042.
40. Modrzynska,K., Pfander,C., Chappell,L., Yu,L., Suarez,C., Dundas,K., Gomes,A.R., Goulding,D., Rayner,J.C., Choudhary,J. *et al.* (2017) A knockout screen of ApiAP2 genes reveals networks of interacting transcriptional regulators controlling the plasmodium life cycle. *Cell Host Microbe*, **21**, 11–22.
41. Santos,J.M., Josling,G., Ross,P., Joshi,P., Orchard,L., Campbell,T., Schieler,A., Cristea,I.M. and Llinás,M. (2017) Red blood cell

invasion by the malaria parasite is coordinated by the PfAP2-I transcription factor. *Cell Host Microbe*, **21**, 731–741.
42. Sierra-Miranda,M., Vembar,S.S., Delgadillo,D.M., Avila-Lopez,P.A., Herrera-Solorio,A.M., Lozano Amado,D., Vargas,M. and Hernandez-Rivas,R. (2017) PfAP2Tel, harbouring a non-canonical DNA-binding AP2 domain, binds to Plasmodium falciparum telomeres. *Cell Microbiol.*, **19**, e12742.
43. Zhang,C., Li,Z., Cui,H., Jiang,Y., Yang,Z., Wang,X., Gao,H., Liu,C., Zhang,S. and Su,X.-Z. (2017) Systematic CRISPR-Cas9-mediated modifications of plasmodium yoelii ApiAP2 genes reveal functional insights into parasite development. *Mbio*, **8**, e01986-17.
44. Kafsack,B.F., Rovira-Graells,N., Clark,T.G., Bancells,C., Crowley,V.M., Campino,S.G., Williams,A.E., Drought,L.G., Kwiatkowski,D.P., Baker,D.A. *et al.* (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature*, **507**, 248–252.
45. Balaji,S., Babu,M.M., Iyer,L.M. and Aravind,L. (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.*, **33**, 3994–4006.
46. Painter,H.J., Campbell,T.L. and Llinas,M. (2011) The Apicomplexan AP2 family: integral factors regulating Plasmodium development. *Mol. Biochem. Parasitol.*, **176**, 1–7.
47. Campbell,T.L., De Silva,E.K., Olszewski,K.L., Elemento,O. and Llinas,M. (2010) Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.*, **6**, e1001165.
48. Kensche,P.R., Hoeijmakers,W.A., Toenhake,C.G., Bras,M., Chappell,L., Berriman,M. and Bartfai,R. (2016) The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Res.*, **44**, 2110–2124.
49. Ponts,N., Harris,E.Y., Prudhomme,J., Wick,I., Eckhardt-Ludka,C., Hicks,G.R., Hardiman,G., Lonardi,S. and Le Roch,K.G. (2010) Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res.*, **20**, 228–238.
50. Bunnik,E.M., Polishko,A., Prudhomme,J., Ponts,N., Gill,S.S., Lonardi,S. and Le Roch,K.G. (2014) DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics*, **15**, 1.
51. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.
52. Cortes,A. (2005) A chimeric Plasmodium falciparum Pfnbp2b/Pfnbp2a gene originated during asexual growth. *Int. J. Parasitol.*, **35**, 125–130.
53. Cortés,A., Benet,A., Cooke,B.M., Barnwell,J.W. and Reeder,J.C. (2004) Ability of Plasmodium falciparum to invade Southeast Asian ovalocytes varies between parasite lines. *Blood*, **104**, 2961–2966.
54. Oyola,S.O., Otto,T.D., Gu,Y., Maslen,G., Manske,M., Campino,S., Turner,D.J., MacInnis,B., Kwiatkowski,D.P. and Swerdlow,H.P. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics*, **13**, 1.
55. Lopez-Rubio,J.J., Siegel,T.N. and Scherf,A. (2013) Genome-wide chromatin immunoprecipitation-sequencing in Plasmodium. *Methods Mol. Biol.*, **923**, 321–333.
56. Okonechnikov,K., Conesa,A. and García-Alcalde,F. (2015) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.
57. Ramirez,F., Ryan,D.P., Gruning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dundar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
58. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
59. Karmodiya,K., Pradhan,S.J., Joshi,B., Jangid,R., Reddy,P.C. and 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. *Epigenet. Chromatin*, **8**, 32.
60. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
61. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
62. Josling,G.A., Petter,M., Oehring,S.C., Gupta,A.P., Dietz,O., Wilson,D.W., Schubert,T., Langst,G., Gilson,P.R., Crabb,B.S. *et al.* (2015) A plasmodium falciparum bromodomain protein regulates invasion gene expression. *Cell Host Microbe*, **17**, 741–751.
63. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
64. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
65. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
66. Adjalley,S.H., Chabbert,C.D., Klaus,B., Pelechano,V. and Steinmetz,L.M. (2016) Landscape and dynamics of transcription initiation in the malaria parasite plasmodium falciparum. *Cell Rep.*, **14**, 2463–2475.
67. Harrel,F.E. (2018) Hmisc: Harrel Miscellaneous. *R package version 4.1-1*. Frank E Harrel Jr, with contributions from Chalers Dupont and many others.
68. Shen,L., Shao,N., Liu,X. and Nestler,E. (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.
69. Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.
70. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
71. Kumar,L. and M,E.F. (2007) Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, **2**, 5–7.
72. Schep,A.N. and Kummerfeld,S.K. (2017) iheatmapr: Interactive complex heatmaps in R. *JOSS*, **2**, 359.
73. Warnes,G.R., Bolker,B., Bonebakker,L., Gentleman,R., Liaw,W.H.A., Lumley,T., Maechler,M., Magnusson,A., Moeller,S., Schwartz,M. *et al.* (2015) gplots: various R programming tools for plotting data. *R package version 2.17.0*.
74. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.*, **29**, e45.
75. Corces,M.R., Trevino,A.E., Hamilton,E.G., Greenside,P.G., Sinnott-Armstrong,N.A., Vesuna,S., Satpathy,A.T., Rubin,A.J., Montine,K.S., Wu,B. *et al.* (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, **14**, 959–962.
76. Corces,M.R., Buenrostro,J.D., Wu,B., Greenside,P.G., Chan,S.M., Koenig,J.L., Snyder,M.P., Pritchard,J.K., Kundaje,A., Greenleaf,W.J. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48** 1193–1203.
77. Zhou,V.W., Goren,A. and Bernstein,B.E. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7–18.
78. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
79. Daugherty,A.C., Yeo,R.W., Buenrostro,J.D., Greenleaf,W.J., Kundaje,A. and Brunet,A. (2017) Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res.*, **27**, 2096–2107.
80. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
81. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of

transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
82. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
83. Broadbent,K.M., Broadbent,J.C., Ribacke,U., Wirth,D., Rinn,J.L. and Sabeti,P.C. (2015) Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics*, **16**, 454.
84. Florens,L., Washburn,M.P., Raine,J.D., Anthony,R.M., Grainger,M., Haynes,J.D., Moch,J.K., Muster,N., Sacci,J.B., Tabb,D.L. *et al.* (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature*, **419**, 520–526.
85. Davie,K., Jacobs,J., Atkins,M., Potier,D., Christiaens,V., Halder,G. and Aerts,S. (2015) Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.*, **11**, e1004994.
86. Crowley,V.M., Rovira-Graells,N., Ribas de Pouplana,L. and Cortes,A. (2011) Heterochromatin formation in bistable chromatin domains controls the epigenetic repression of clonally variant Plasmodium falciparum genes linked to erythrocyte invasion. *Mol. Microbiol.*, **80**, 391–406.
87. Comeaux,C.A., Coleman,B.I., Bei,A.K., Whitehurst,N. and Duraisingh,M.T. (2011) Functional analysis of epigenetic regulation of tandem RhopH1/clag genes reveals a role in Plasmodium falciparum growth. *Mol. Microbiol.*, **80**, 378–390.
88. Rovira-Graells,N., Crowley,V.M., Bancells,C., Mira-Martinez,S., Ribas de Pouplana,L. and Cortes,A. (2015) Deciphering the principles that govern mutually exclusive expression of Plasmodium falciparum clag3 genes. *Nucleic Acids Res.*, **43**, 8243–8257.
89. Guizetti,J., Barcons-Simon,A. and Scherf,A. (2016) Trans-acting GC-rich non-coding RNA at var expression site modulates gene counting in malaria parasite. *Nucleic Acids Res.*, **44**, 9710–9718.
90. Chakrabarti,K., Pearson,M., Grate,L., Sterne-Weiler,T., Deans,J., Donohue,J.P. and Ares,M. Jr (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA*, **13**, 1923–1939.
91. Voss,T.S., Kaestli,M., Vogel,D., Bopp,S. and Beck,H.P. (2003) Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters. *Mol. Microbiol.*, **48**, 1593–1607.
92. Voss,T.S., Tonkin,C.J., Marty,A.J., Thompson,J.K., Healer,J., Crabb,B.S. and Cowman,A.F. (2007) Alterations in local chromatin environment are involved in silencing and activation of subtelomeric var genes in Plasmodium falciparum. *Mol. Microbiol.*, **66**, 139–150.
93. Epp,C., Li,F., Howitt,C.A., Chookajorn,T. and Deitsch,K.W. (2009) Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. *RNA*, **15**, 116–127.
94. Deitsch,K.W. and Dzikowski,R. (2017) Variant gene expression and antigenic variation by malaria parasites. *Annu. Rev. Microbiol.*, **71**, 625–641.
95. Natarajan,A., Yardimci,G.G., Sheffield,N.C., Crawford,G.E. and Ohler,U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
96. Gaertner,B., Johnston,J., Chen,K., Wallaschek,N., Paulson,A., Garruss,A.S., Gaudenz,K., De Kumar,B., Krumlauf,R. and Zeitlinger,J. (2012) Poised RNA polymerase II changes over developmental time and prepares genes for future expression. *Cell Rep.*, **2**, 1670–1683.
97. Zaidi,S.K., Young,D.W., Montecino,M.A., Lian,J.B., van Wijnen,A.J., Stein,J.L. and Stein,G.S. (2010) Mitotic bookmarking of genes: a novel dimension to epigenetic control. *Nat. Rev. Genet.*, **11**, 583–589.
98. Toenhake,C.G., Fraschka,S.A., Vijayabaskar,M.S., Westhead,D.R., van Heeringen,S.J. and Bartfai,R. (2018) Chromatin Accessibility-Based characterization of the gene regulatory network underlying plasmodium falciparum Blood-Stage development. *Cell Host Microbe*, **23**, 557–569.
99. Beaumont,H.J., Gallie,J., Kost,C., Ferguson,G.C. and Rainey,P.B. (2009) Experimental evolution of bet hedging. *Nature*, **462**, 90–93.
100. Seco-Hidalgo,V., Osuna,A. and De Pablos,L.M. (2015) To bet or not to bet: deciphering cell to cell variation in protozoan infections. *Trends Parasitol.*, **31**, 350–356.
101. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
102. Yuda,M., Iwanaga,S., Shigenobu,S., Kato,T. and Kaneko,I. (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites. *Mol. Microbiol.*, **75**, 854–863.
103. Martins,R.M., Macpherson,C.R., Claes,A., Scheidig-Benatar,C., Sakamoto,H., Yam,X.Y., Preiser,P., Goel,S., Wahlgren,M., Sismeiro,O. *et al.* (2017) An ApiAP2 member regulates expression of clonally variant genes of the human malaria parasite Plasmodium falciparum. *Sci. Rep.*, **7**, 14042.
104. Johnson,J.L., Georgakilas,G., Petrovic,J., Kurachi,M., Cai,S., Harly,C., Pear,W.S., Bhandoola,A., Wherry,E.J. and Vahedi,G. (2018) Lineage-determining transcription factor TCF-1 initiates the epigenetic identity of T cells. *Immunity*, **48**, 243–257.
105. Bulut-Karslioglu,A., Perrera,V., Scaranaro,M., de la Rosa-Velazquez,I.A., van de Nobelen,S., Shukeir,N., Popow,J., Gerle,B., Opravil,S., Pagani,M. *et al.* (2012) A transcription factor–based mechanism for mouse heterochromatin formation. *Nat. Struct. Mol. Biol.*, **19**, 1023.
106. Martienssen,R.A., Kloc,A., Slotkin,R.K. and Tanurdzic,M. (2008) Epigenetic inheritance and reprogramming in plants and fission yeast. *Cold Spring Harb. Symp. Quant. Biol.*, **73**, 265–271.
107. Flueck,C., Bartfai,R., Volz,J., Niederwieser,I., Salcedo-Amaya,A.M., Alako,B.T., Ehlgen,F., Ralph,S.A., Cowman,A.F., Bozdech,Z. *et al.* (2009) Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog.*, **5**, e1000569.
108. Lopez-Rubio,J.J., Gontijo,A.M., Nunes,M.C., Issar,N., Hernandez Rivas,R. and Scherf,A. (2007) 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol. Microbiol.*, **66**, 1296–1305.