Supplemental Material to:

Nucleosome landscape and control of transcription in the human malaria parasite

Nadia Ponts[1,*], Elena Y. Harris[2,*], Jacques Prudhomme[1], Ivan Wick[3], Colleen Eckhardt[3], Glenn Hicks[5], Gary Hardiman[3,4], Stefano Lonardi[2], Karine G. Le Roch[1, †]

[1] *Department of Cell Biology and Neuroscience, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.*

[2] *Department of Computer Science and Engineering, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.*

[3] *BIOGEM, School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.*

[4] *Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.*

[5] *Institute of Integrative Genomics Biology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.*

**These authors contributed equally to this work*

[†] *Corresponding author. E-mail: karine.leroch@ucr.edu*

Supplemental Methods

Sample preparation for sequencing. Libraries were prepared using the Illumina® genomic DNA sample preparation kit (cat. #FC-102-1001) following the manufacturer's instructions, with a starting amount of 3 μg of fragmented DNA. Due to the (A+T)-rich content of *P. falciparum*'s genome, the manufacturer's procedure for PCR amplification of the libraries prior sequencing was customized to ensure efficient amplification of all sequences; every PCR reaction was supplemented with 1 µL of the high fidelity and high efficiency DNA polymerase TaKaRa Ex Taq™ (the activity of the TaKaRa Ex Taq™ in the reaction buffer has been previously verified, data not shown).

Sequencing and data processing. The manufacturer's instructions were strictly followed for cluster generation (Illumina® cluster generation kit #FC-103-1001) and single-read 36-cycle sequencing (Illumina® sequencing kit #FC-104-1003) on the Illumina® genome analyzer generation I (GAI, FAIRE experiment), or generation II (GAII, MAINE experiment). The difference between the GAI and the GAII essentially resides in an improved flow cell that generates an increased number of reads and improved instrumentation. One sample that was collected 24 *hpi* and processed with FAIRE was also analyzed on the GAII to confirm the reproducibility and the robustness of the sequencing technology (Supplemental Fig. S12). For both FAIRE-seq and MAINE-seq, images were processed by the Illumina® Pipeline v0.3. The sequence reads generated by the Illumina® Pipeline, with their quality scores, were retrieved and used for mapping on *P. falciparum*'s genome.

Mapping sequenced reads. The workflow chart for the mapping procedure is given in Supplemental Fig. S3. Sequence reads generated by the Pipeline were trimmed for quality score lower than 20. In addition, the first and last base of each read was systematically removed to minimize the risk of being part of the adapter sequences. Trimmed reads shorter than 18 bases were not used for mapping. The genome from *P. falciparum* strain 3D7 (downloaded from www.plasmoDB.org, version 5.5) was used as reference genome. Illumina® ELAND was used to map reads 18-32 bases long, whereas RMAP (Smith *et al.* 2008) was employed to process reads 33-34 bases long. Sequenced reads giving a perfect match at a unique location in the reference genome were immediately mapped. The remaining reads were mapped uniquely allowing up to two mismatches, but only if their chastity was at least 0.6. Filtered reads that gave perfect matches at multiple locations in the genome of *P. falciparum* were not used in this analysis. Remaining reads were aligned with the human genome. Indeed, *P. falciparum* is cultured in human blood, which makes the presence of human DNA ineluctable.

Computing read distribution profiles for nucleosome positioning. In order to remove high-frequency noise, the raw coverage distribution of uniquely mapped reads of each chromosome and each strand was smoothed using a kernel density estimation method (Parzen 1962). Specifically, given the raw coverage *C* the smoothed coverage *S* was computed as follows:

$$S(i) = \frac{1}{h\sqrt{2\pi}} \sum_{j=i-5h}^{i+5h} C(j) e^{-0.5\left(\frac{j-i}{h}\right)^2}$$

The parameter *h* was set to be 30, so that the sliding window was 300bp. A similar smoothing procedure was previously described (Valouev *et al.* 2008). To obtain the overall smoothed coverage distribution (or coverage profile) the smoothed coverage for the positive and negative strands was summed up at each position. For each time point the total number of reads that mapped to the genome and the percentage of nucleotides actually covered by at least one read was retrieved. Smoothed distributions of reads were normalized per million of mapped reads and with the percentage of the genome actually covered by reads.

Determining the position of nucleosomes. The position of nucleosomes was determined from the coverage profiles obtained in the MAINE-seq procedure, as follows. First, *regions of coverage* were defined as the continuous spans of the genome where coverage was at least one. Then each such region of coverage was processed individually: positive and negative strand was scanned for local maxima using a sliding window. The size of the sliding window was empirically determined to be 15 bases. For each local maximum found on the positive strand (so called positive peak, *Peak*$^+$), the negative strand was scanned for a corresponding negative maximum (so called negative peak, *Peak*$^-$) within a distance range determined empirically according to the sizes of the libraries and the sizes of the sequenced reads. The chosen distance range was [-15, 180]. The half distance between the two corresponding peaks was defined as the *shift*. The center of a nucleosome was placed in the middle between the two corresponding positive and negative peaks, and the left and right boundaries were set at distance 146/2 = 73 bp around the center. In some cases, a local maximum was found only on one strand, with no matching peak on the other. Visual inspection of the dataset revealed that most of the time the missing matching peak was actually present as a shoulder of a neighbouring peak (and not as a local maximum). In these cases a nucleosome was placed according to a single detected peak, using the shift calculated for the nearest nucleosome. The center of a nucleosome was placed at the position *P*$^+$ if only positive peak was detected and at the position *P*$^-$ if only negative peak was detected, where $P^+ = pos_{Peak^+} + shift$ and $P^- = pos_{Peak^-} - shift$. Each nucleosome was assigned a *confidence score* equal to the sum of the smoothed normalized coverage in a region of 147 nucleotides centered at the assigned nucleosome location.

Finding the centromeres. Chromosomes were scanned to detect putative centromeres. Scanning parameters were empirically determined using the known centromeric regions (www.plasmoDB.org v.5.5). For both FAIRE-seq and MAINE-seq datasets (time points combined together), the average number of (A+T) and the average number of (C+G) per base were calculated for all known centromeric region. Then, chromosomes were scanned for 2 kb-long regions with (A+T) and (C+G) content within the range [average – 5*SD, average +5*SD] for both FAIRE-seq and MAINE-seq. For each time point, consecutive windows were merged into a single contig corresponding to a possible centromeric region.

Supplemental Figures and Legends

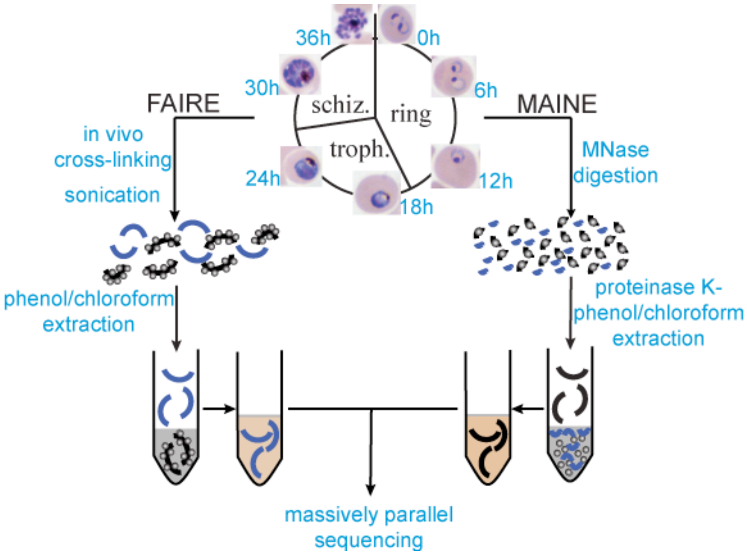
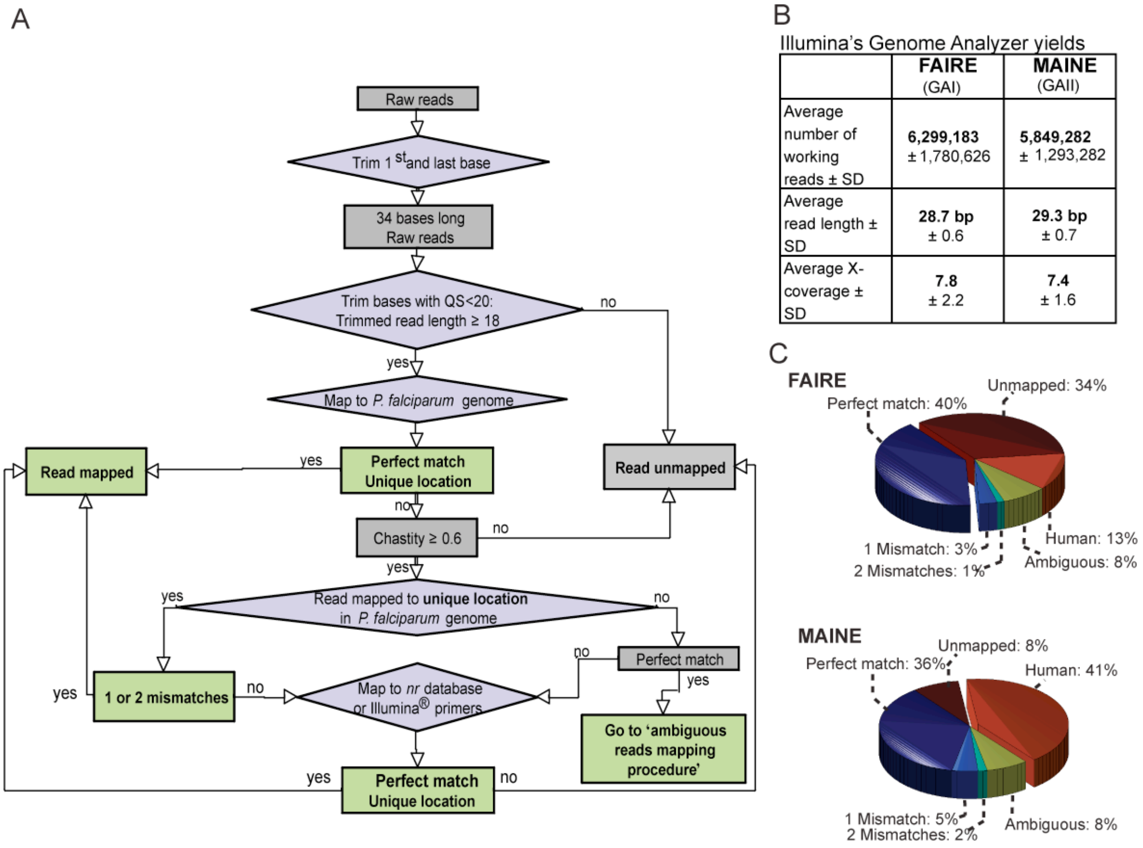


Figure S1 Representation of the experimental procedure

Genomic DNA samples were prepared at seven different time points across the asexual erythrocytic cycle of the human malaria parasite *P. falciparum.* After entering the host erythrocyte as a *merozoite* (time point 0 *hour post-invasion* or *hpi*) the parasite undergoes morphological changes into the *ring* stage. Then, rings mature into *trophozoites* (approximately 18 to 24 *hpi* corresponding to a transcriptionally active and DNA replication stage) followed by the *schizont* stage (approximately 36 *hpi*). Collected DNA samples were analyzed by both FAIRE and MAINE, coupled to high throughput sequencing.



Illumina's Genome Analyzer yields

	FAIRE (GAI)	**MAINE** (GAII)
Average number of working reads ± SD	**6,299,183** ± 1,780,626	**5,849,282** ± 1,293,282
Average read length ± SD	**28.7 bp** ± 0.6	**29.3 bp** ± 0.7
Average X-coverage ± SD	**7.8** ± 2.2	**7.4** ± 1.6

Figure S2 Mapping chart and sequencing metrics

(A) Representation of the stringency of the mapping procedure selected for high quality reads sequenced by the Genome Analyzer that were retained in the present study. (B) Number of working reads, their length, and the X-coverage of the genome are given as the average between all time points within an experiment, ± standard deviation (SD). For both FAIRE-seq and MAINE-seq, we obtained an average of six million working reads (sequenced reads that passed our quality criteria; see the Methods section), which represents an average coverage of 7.8-fold the *P. falciparum* genome for FAIRE-seq, and 7.4-fold for MAINE-seq. (C) For each experiment, more than 50% of the working reads were mapped to the parasite genome. Since *P. falciparum* is cultured in human blood, reads from each experiment were also mapped to the human genome. Thirteen percent of the FAIRE reads and 41% of the MAINE reads perfectly matched human DNA, leaving an average of unmapped reads of 34% for FAIRE (GAI) and 8% for MAINE (GAII). In this study, uniquely mapped reads with up to two mismatches were used.

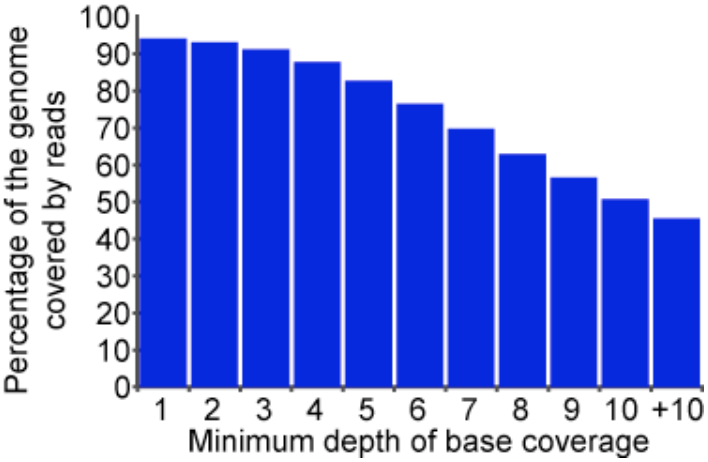


Figure S3 Percentage of the genome covered by sequencing reads

A reference sample, composed of equal proportions of FAIRE and MAINE samples for every time points, was examined for the extent of the genome coverage (number of nucleotides covered by reads) as a function of the minimum number of reads mapped to a considered position. More than 50% of the genome was mapped consistently by no less than 10 reads.

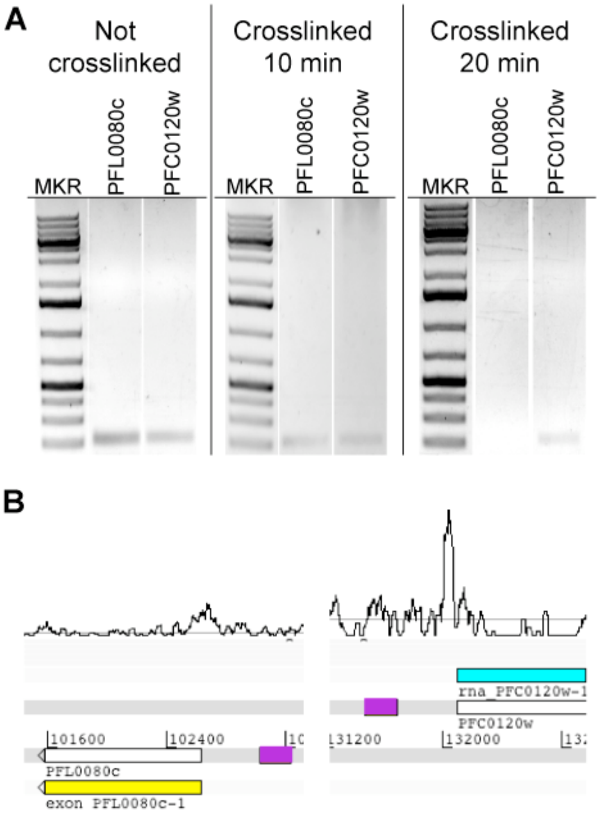


Figure S4 Validation of the FAIRE-seq experiment

(A) PCR validation - To optimize and verify the efficiency of our formaldehyde treatment, we chose two different regions expected to exhibit a different patterns of DNA enrichment with FAIRE. Regions upstream of PFC0120W (gene expressed at the schizont stage) and PFL0080c (gene expressed at the gametocyte stage) were amplified by PCR reaction. Our results show that cross-linking is efficient after 20 minutes of formaldehyde treatment. (B) Genome snapshots (Artemis software, Sanger center) of the two selected regions upstream of PFL0080c and PFC0120w. Amplified regions are represented by purple boxes (214 bp for PFL0080c and 200 bp for PFC0120w). Back curves above the gene models are FAIRE coverage profiles at 24 *hpi* for the considered region.

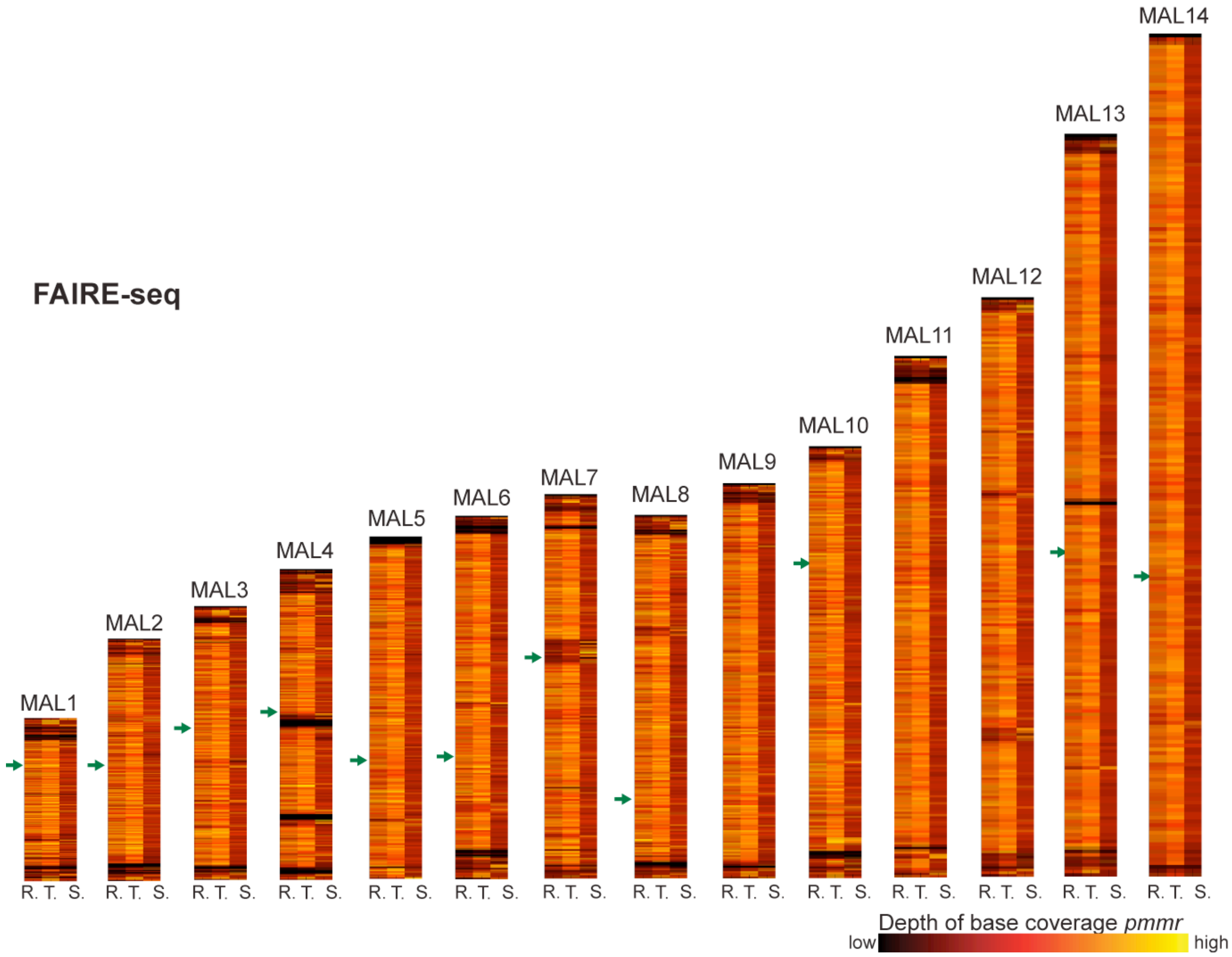


Figure S5 Chromatin structure changes are genome-wide

The general nucleosome occupancy of *P. falciparum* DNA varies across time. Each color matrix represents one chromosome of *P. falciparum* (MAL1 to MAL14). The y-axes represent the position on the considered chromosome, and the x-axes show the considered morphological stage (R. = ring, 0*hpi*; T. = trophozoite, 18*hpi*; S. = schizont, 36*hpi*). Arrows localize the position of all known centromeres (from www.plasmodb.org v5.5). Differences in color intensities, from black to yellow, reflect different depths of read coverage. At the early trophozoite stage (18*hpi*), coverage is higher in FAIRE-seq and minimum in MAINE-seq indicating an intense loosening of chromatin structure. On the contrary, the coverage is at its lowest in FAIRE-seq and highest in MAINE-seq at the late schizont stage (36*hpi*) demonstrating a maximum degree of packaging of the DNA.

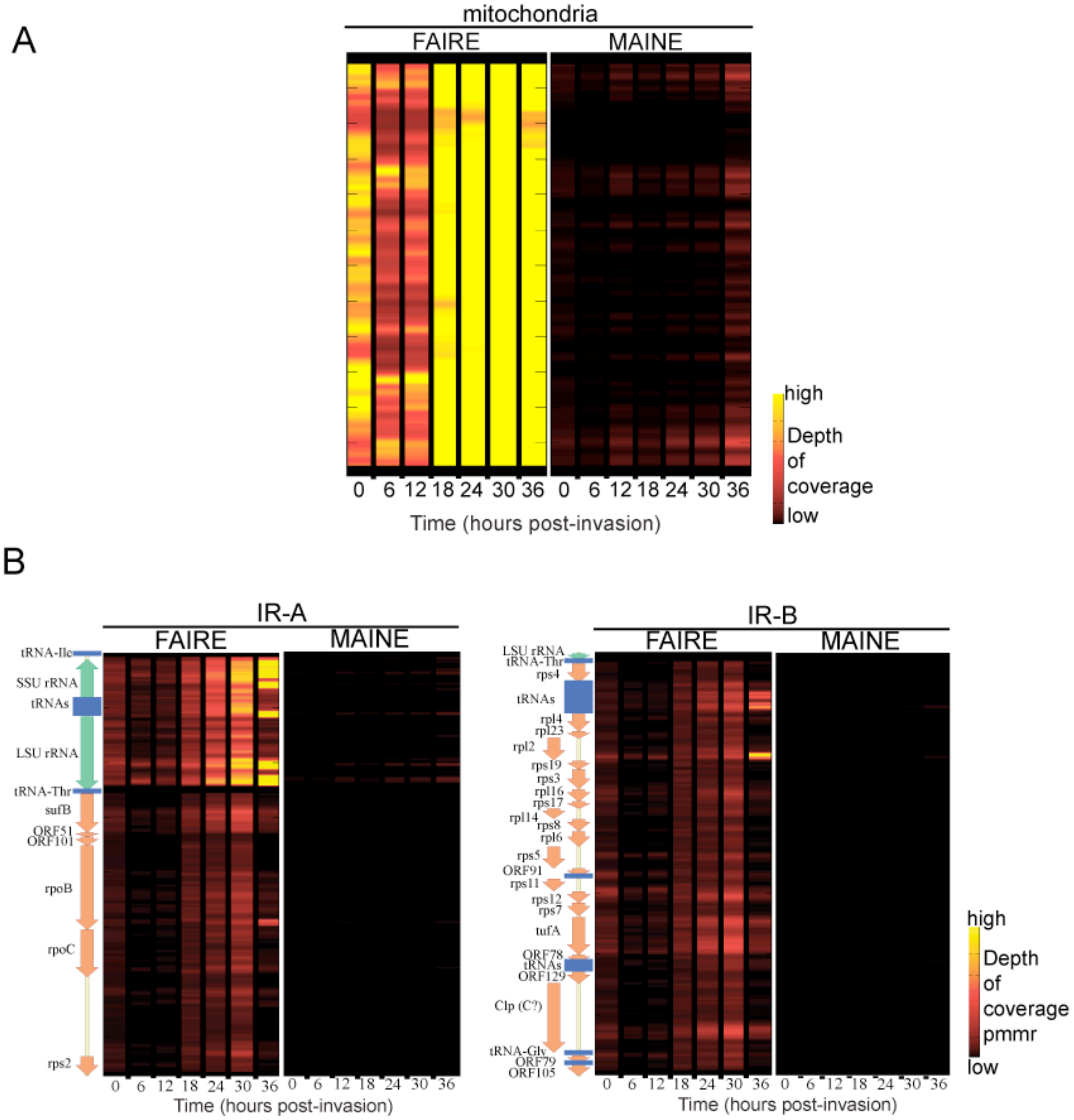


Figure S6 Chromatin status in *P. falciparum* mitochondrion and apicoplast

(A) As expected, the mitochondrion is saturated in reads. (B) *P. falciparum* gene maps of IR-A and IR-B halves of circle apicoplast DNA were retrieved (genebank accession numbers X95275 and X95276), and corresponding depths of read coverage were aligned, color coded from low to high coverage, *i.e.* black to yellow. Time is displayed on the x-axes (in hours post-invasion). Like plants chloroplasts, the apicoplast is semi-autonomous with its own 35kb circular genome and expression machinery. Chromatin remodeling in the apicoplast follows the pattern that is observed genome-wide.

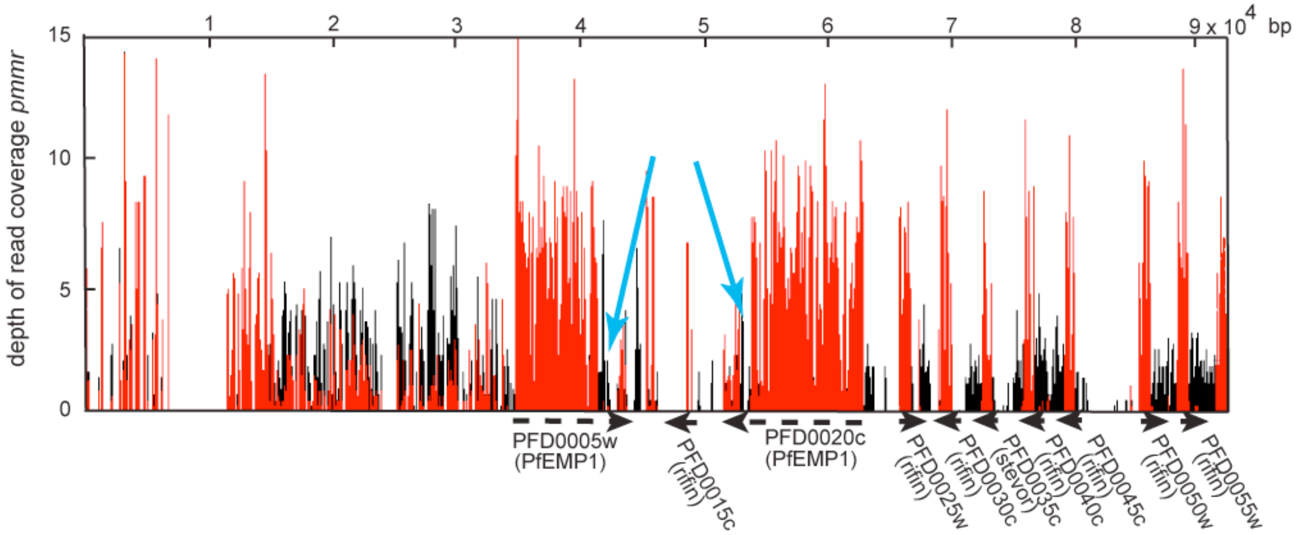


Figure S7 Chromatin status in subtelomeric regions.

Read distributions for FAIRE (black) and MAINE (red) in the 5' telomere of chromosome 4 (at 18*hpi*) are overlaid. Dashed arrows indicate genes and their orientation. Blue arrows show the location of introns covered with FAIRE-seq.

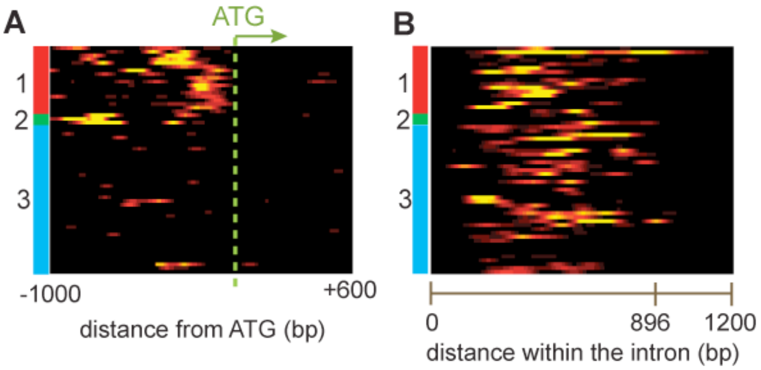


Figure S8 Representation of the chromatin status in *var* genes promoters and introns.

Chromatin status in *var* genes promoters (B) and introns (C) (dark is compacted, yellow is loosened). Each line on the y-axis represents one *var* gene. Three groups were clustered according to the presence and the position of loosened chromatin zones in the promoter (symbolized by the pink, blue and grey boxes).

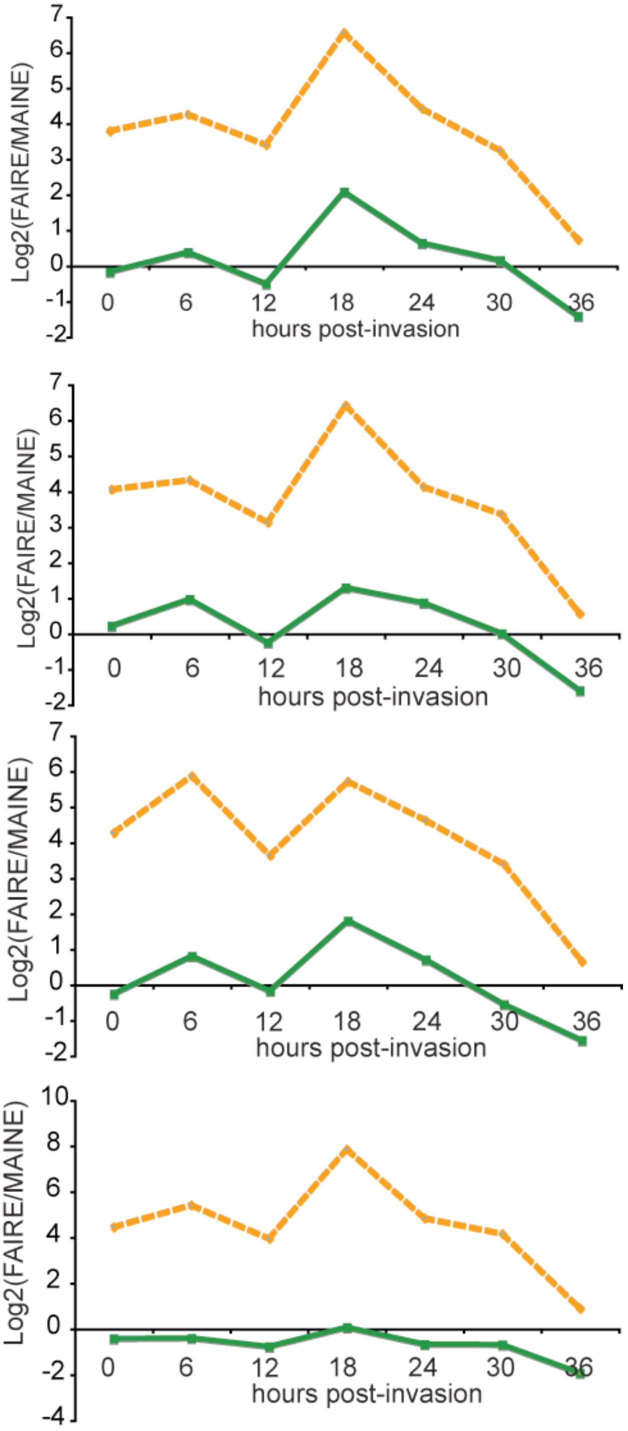


Figure S9 Intron/Promoter pairing for *var* genes

Pairing of chromatin status profiles for promoters (orange curves) and introns (green curves) from 4 sets of 60 randomly picked genes. The pairing observed in the case of *var* genes does not occur for random genes.

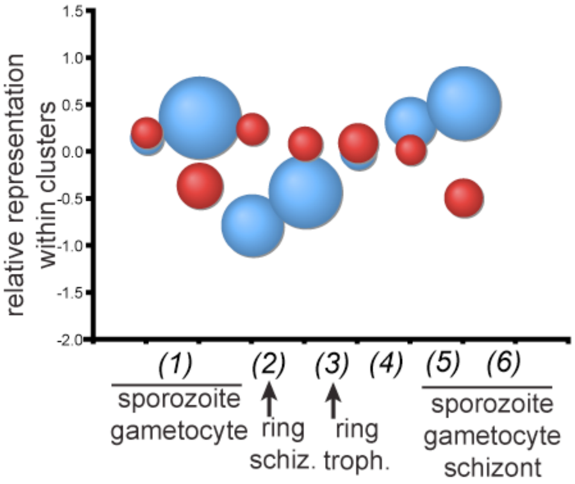


Figure S10 Enrichment of our Cluster II with the previously published functional clusters (Le Roch *et al.* 2003). Cluster II (red) is not significantly enriched in any of the functional families of genes. Cluster I (blue) is represented for comparison.

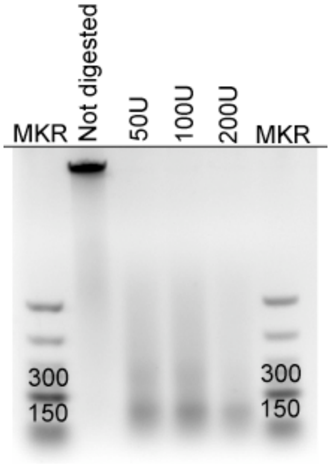


Figure S11 Optimization of the MNase digestion

We optimized the concentration of MNase to use to obtain a maximum digestion of mononucleosomes. From 50 to 150 units, di-nucleosomes are still visible at 300bp. At 200 units, a unique band of mononucleosomes at 150bp is present. Our results show that 200 units of MNase were necessary for the preparation of mononucleosomes.

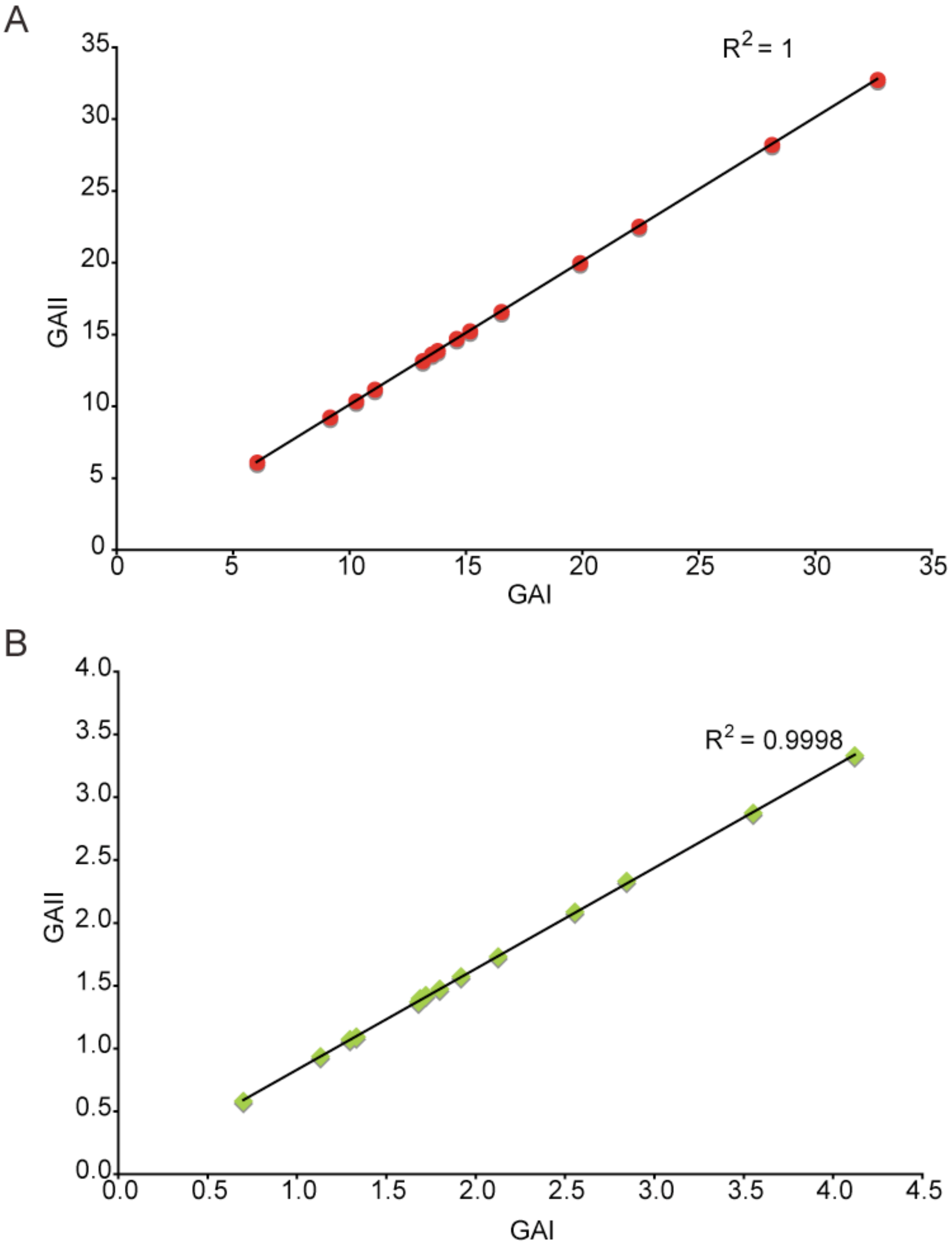


Figure S12 Assessment of the reproducibility and the robustness of the sequencing technology

To assess the reproducibility and the robustness of the sequencing technology, one sample that was collected 24 *hpi* and processed with FAIRE (GAI) was also analyzed on the GAII. (A) Correlation between the numbers of nucleotides covered. (B) Correlation between the summed and normalized depths of read coverage. The percentages of genome covered and the sequencing depths per million reads obtained with the GAI and the GAII were similar (correlation coefficients $R^2 > 0.999$) for all considered time points. Comparative analyses of the datasets produced by the two instruments are therefore possible.

Supplemental Table Legends

Supplemental Table S1. Lists of genome-wide nucleosome boundaries and confidence scores

Supplemental Table S2. Centromeres *loci* identified *de novo*

Supplemental Table S3. List of the genes that were clustered using k-means

Supplemental References

Le Roch KG Zhou Y Blair PL Grainger M Moch JK Haynes JD De La Vega P Holder AA Batalov S Carucci DJ *et al.* 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301:** 1503-1508.

Parzen E. 1962. On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33:** 1065-1076.

Smith AD, Xuan Z, Zhang MQ. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9:** 128.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829-834.