7th May 2024

Anel Gonzalez, Rohit Satyam, Aleksandra Kurowska

3D chromatin conformation analysis upon depletion of PfAP2-P protein in *Plasmodium falciparum.*

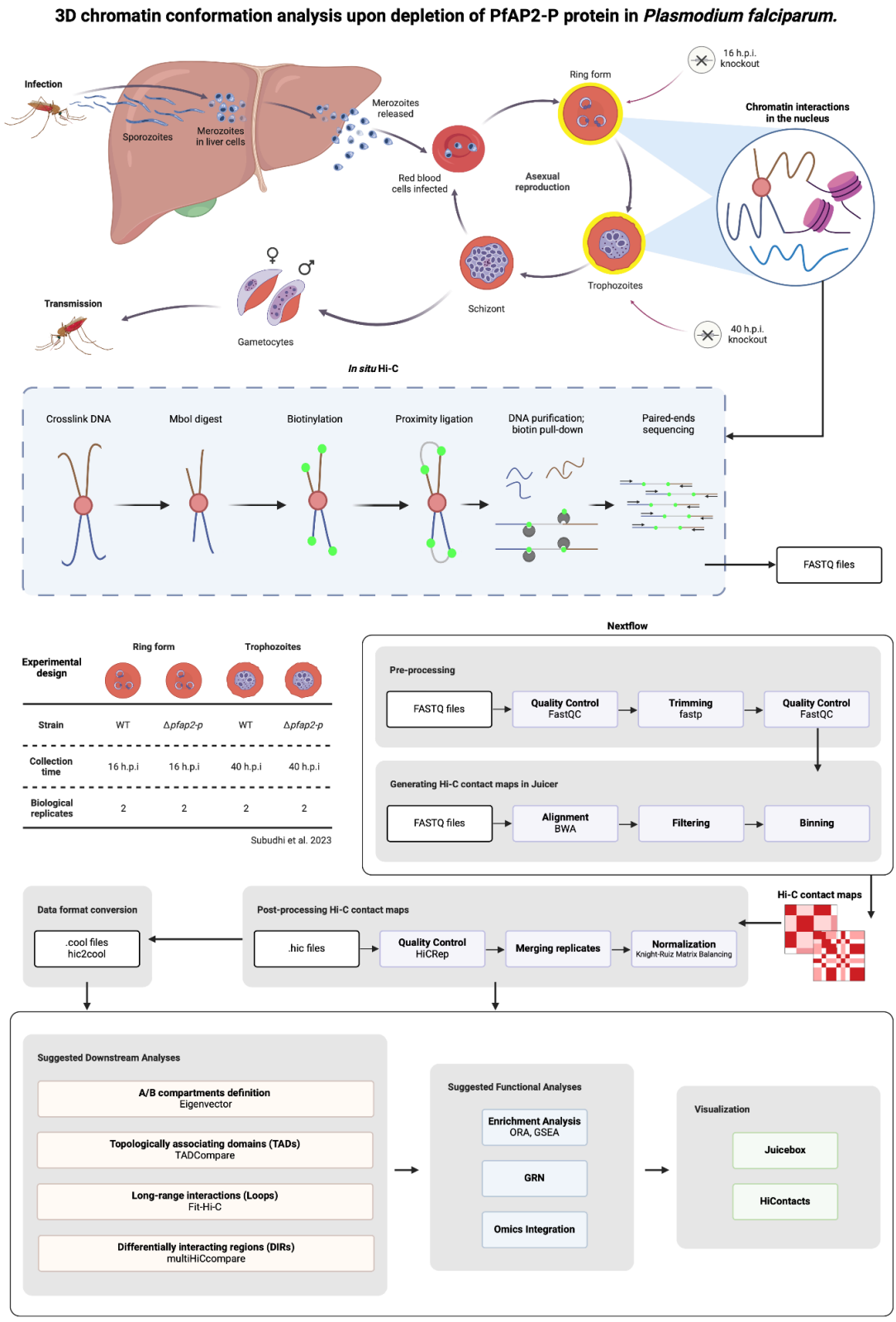
1. Introduction

The organization of chromatin in eukaryotic cell nucleus is a complex and a highly dynamic system. Understanding how the three-dimensional (3D) chromatin architecture affects gene regulation, cell cycle progression, and disease pathogenesis are crucial biological questions. To this date, there are many chromosome conformation capture (3C) technologies that are elucidating the 3D genome structure, including 4C, 5C, and high-throughput chromatin conformation capture (Hi-C). The information provided by Hi-C can be grouped into the major spatial functional units: (a) large chromatin compartments (>1 Mb): transcriptionally active compartment A, and transcriptionally inactive compartment B; (b) medium-sized structures (100–1,000 kb), known as topologically associating domains (TADs); and (c) chromatin loops (~1–10 kb) facilitating long-range interactions [1].

In this study, we focus on Hi-C application for understanding the nuclear architecture and gene regulation of the human malaria parasite, *Plasmodium falciparum*. Malaria pathogenesis involves intricate interactions among numerous proteins, yet their molecular regulation remains poorly characterized. The infection begins with a female mosquito injecting sporozoites into the host's bloodstream during a blood meal. Subsequently, sporozoites invade hepatocytes, undergo replication, and release merozoites into the bloodstream, initiating a 48-hour replication cycle within red blood cells, progressing through ring, trophozoite, and schizont stages, culminating in daughter cell multiplication via schizogony, and eventual invasion of new erythrocytes [2].

Previous studies have shown that *Plasmodium*-specific Apicomplexan AP2 transcription factor (PfAP2-P) is essential during processes of parasite development and its knockout leads to expression of genes situated within heterochromatin regions [3]. This suggests that heterochromatin may have become more accessible. Here, we specifically explore the effects of PfAP2-P deletion on chromatin structure by *in situ* Hi-C of wild type (WT) and knockout strains at two peaks of gene expression 16 and 40 hours post-invasion (h.p.i.), which represent the ring and trophozoite development stages respectively [4].

2. Graphical abstract



3. Methods

Pre-processing and 3D chromatin conformation analysis

The raw data was downloaded from SRA using fasterq-dump command-line utility. The FASTQ files were input to our in-house developed HiCFlow Nextflow workflow that can be accessed here: https://github.com/Rohit-Satyam/HiCFlow/tree/main. Briefly, the data quality was assessed using FastQC version 0.12.1 [5] and summarized using MultiQC version 1.19 [6]. Low-quality reads and Illumina adaptor sequences from the read ends were removed using fastp version 0.23.4 [7]. Trimmed FASTQ files were input to the Juicer version 1.6 pipeline for alignment and post-alignment preprocessing [8]. Paired-end Hi-C library of MboI digested genomic fragments were processed with default parameters and mapped to the *P. falciparum* genome (release-6750, https://plasmodb.org) using BWA version 0.7.17 [9]. The output .hic matrices produced were converted to .cool files using hic2cool version 0.8.3 [10]. Hi-C reads were binned to a resolution of 10 kb and assigned to the resulting n bins, thereby obtaining for each Hi-C experiment a corresponding n ´ n raw contact count matrix C. In this matrix, each row and column corresponds to a 10-kb window, and each entry indicates the number of times the two regions have been observed in a contact. We excluded 10-kb bins for which the mappability score was less than 30. Hi-C matrices were normalized using a Knight-Ruiz Matrix Balancing (KR) normalization [11], which outperforms other normalization methods [12]. 2D plots of the biological replicates were merged by mean counts scores plotted using HiContacts version 1.4.0 [13].

Differential Hi-C

Differentially interacting regions (DIRs) were analyzed with the multiHiCcompare R package version 1.20.0 [14]. The tool provides a Fast Loess joint normalization and removal of biases between multiple Hi-C datasets for comparative analysis. Then the function makes use of the edgeR [15] package for fitting a general linear model for detecting differences in Hi-C data. The hic_glm() function was run on raw datasets at 10kb resolution using QLFTest and the multiple correction test to obtain adjusted p-values.

Long-range interactions (Loops)

For the detection of significant contacts, we employed Fit-Hi-C, modeling the influence of genomic distance on contact count probability using a spline [16]. The findings from this analysis are presented as q-values assigned to individual contacts. Each q-value represents the minimum false discovery rate threshold required for considering a discovery significant.

Functional characterization

Significant Loops

Interacting regions were filtered out by FDR < 0.001 and the number of counts ranged in the top 10% (quantile 0.9). Replicates were merged by keeping only pairs present in both. Annotation was performed by overlapping PlasmoDB annotation. For the overrepresentation analysis, genes associated with either

region were submitted to topGO version 2.54.0 [17] with their associated q-value. All protein-coding genes in the genome were used as background. GO terms were retrieved from Ensembl Protist corresponding to *P. falciparum*. Relevant terms were filtered by p-value < 0.05. Regions that interacted with PfAP2-P were retrieved by selecting any pair that overlapped the bin region containing the gene mentioned above: chr11:328877-336810, corresponding to the bin on chromosome 11 (PF3D7_11_V3) with midFragment == 325000.

Differentially interacting regions

Differentially interacting regions (DIRs) were filtered by adjusted p.value < 0.05, logFC > 0.58 and subsequently, annotated by overlapping them with the PlasmoDB annotation. All the genes associated with these regions were used for overrepresentation analysis with topGO version 2.54.0 [17] using their associated adjusted p-value. All protein-coding genes were used as background. Relevant terms were filtered by p-value < 0.05. Further we explored if any var genes were present within our DIRs. We retrieved all genomic locations of var genes from the PlasmoDB annotation and overlapped them with the regions corresponding to the DIRs.

4. Data availability [18]

Hi-C data: ENA BioProject accession no. PRJNA847684. Hi-C libraries were sequenced using the Illumina NovaSeq 6000.

5. Code availability

The code for our inhouse developed HiCFlow Nextflow workflow for Hi-C pre-processing is available here: https://github.com/Rohit-Satyam/HiCFlow/tree/main. The code for downstream analysis steps is available here: https://github.com/alkurowska/BESE394E_course/tree/main/FINAL.

6. Results

As the disruption of PfAP2-P notably impacts genes linked to heterochromatin, we investigated its influence on chromatin structure. Correlation analysis of Hi-C replicates demonstrated consistent reproducibility at both time points, with intra-chromosomal interaction matrices remaining largely stable for each line (see Figure 1). Comparative examination between the WT and Δpfap2-p lines revealed a slight decrease in long-range interactions frequency and heterochromatin clusters in Δpfap2-p at 16 h.p.i. (see Figure 3). The effect is not observed at 40 h.p.i.. A genome-wide mapping of interaction alterations highlighted reduced interaction frequencies (Figure 2), particularly within internal var genes, aligned with diminished heterochromatin clustering. This observation suggests impaired chromatin compaction in PfAP2-P-deficient parasites, potentially leading to enhanced chromatin accessibility and increased var gene expression, a trend consistent with previous RNA-seq findings [18].

Figure 1. Reproducibility. Stratum-adjusted correlation coefficients.

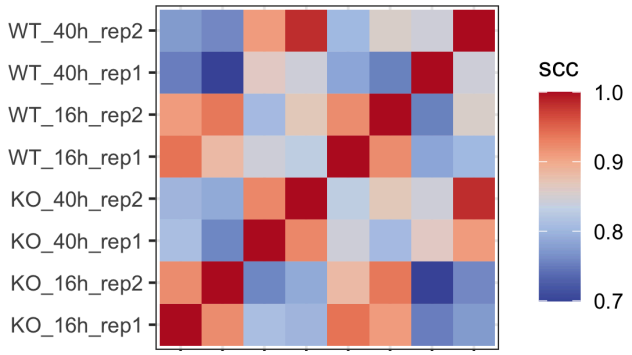


Figure 2. Significant interactions. KR-normalized contact count heatmaps at 10 kb resolution of intrachromosomal interactions for chr11:100000-500000 are given for both peak expression time points: 16 h.p.i. (top) and 40 h.p.i. (bottom). Heatmaps represent significant loops (black circles) of *Δpfap2-p* (left) and the wild type (right).

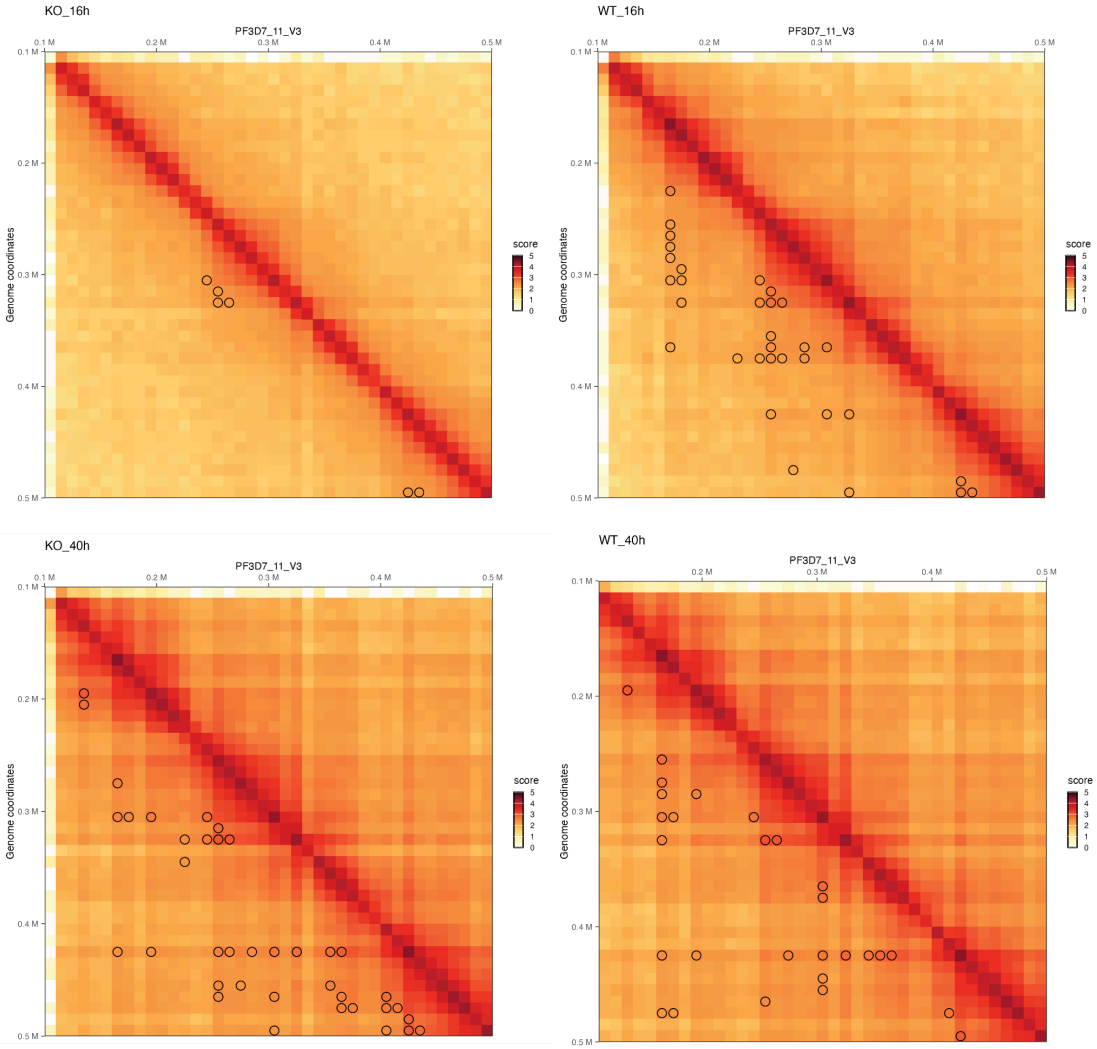
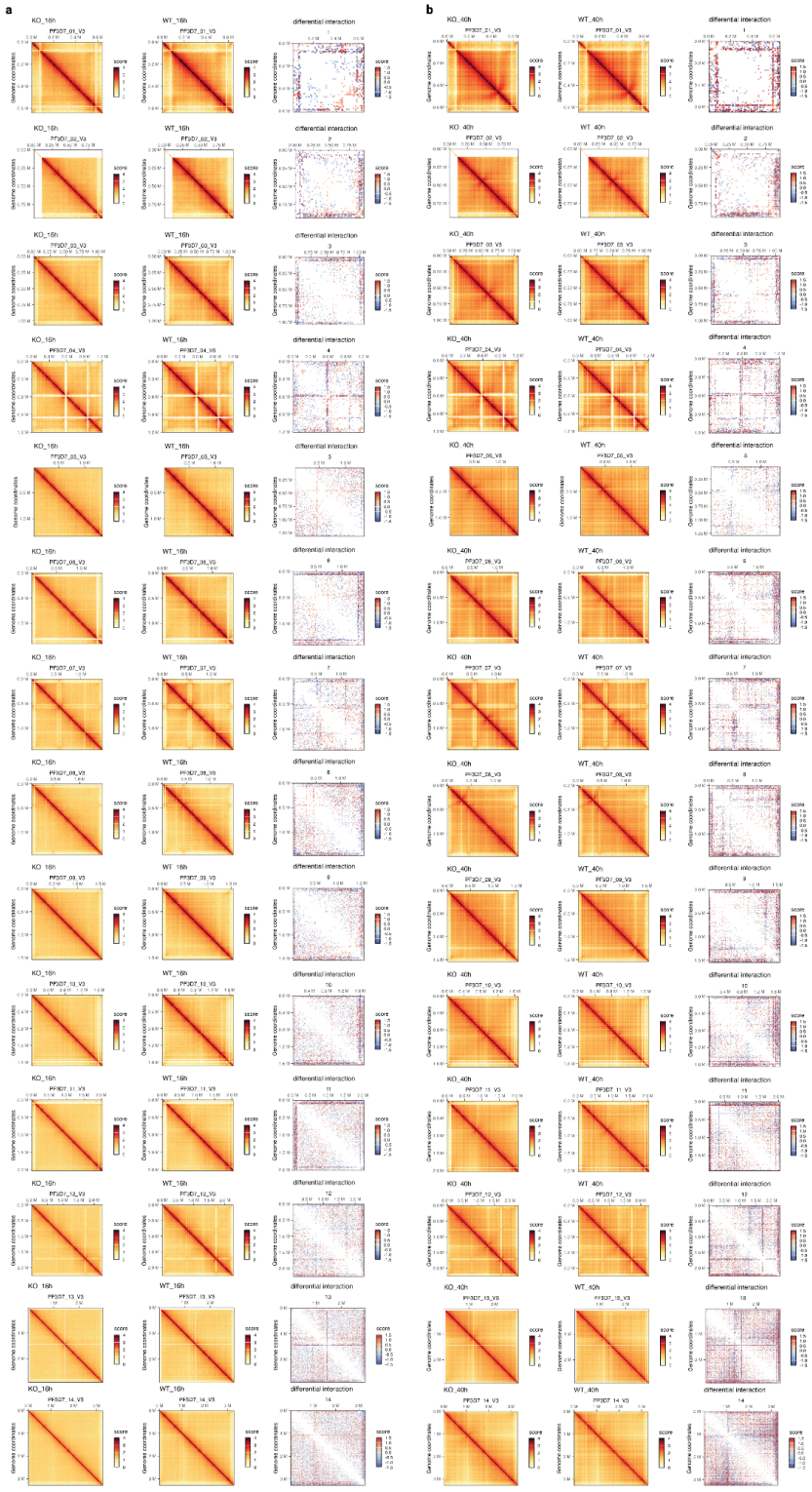


Figure 3. Depletion of PfAP2-P increases chromatin accessibility. KR-normalized contact count heatmaps at 10 kb resolution of intrachromosomal interactions for the 14 chromosomes are given for both peak expression time points: **a.** 16 h.p.i. and **b.** 40 h.p.i.. Heatmaps represent a single chromosome of *Δpfap2-p* (left), the wild type (center), and log2 fold change differential interactions (right).



7. Discussion

In summary, our study focused on utilizing Hi-C technology to investigate the nuclear architecture and gene regulation of *Plasmodium falciparum*, the human malaria parasite, with a specific emphasis on the role of the PfAP2-P gene. Our findings revealed that the deletion of PfAP2-P resulted in altered chromatin structure, characterized by reduced long-range interactions and reduced heterochromatin clustering, particularly evident at the 16-hour post-invasion time point. This observation suggests impaired chromatin compaction in PfAP2-P-deficient parasites, potentially leading to increased chromatin accessibility and enhanced expression of var genes. Our results provide valuable insights into the regulatory role of PfAP2-P in chromatin organization and gene expression in *Plasmodium falciparum*, contributing to a better understanding of malaria biology and potentially offering new avenues for therapeutic intervention.

Limitations

Plasmodium falciparum's nuclear architecture blends features of multicellular and unicellular organisms in order to adapt to its life cycle across mosquito and human hosts. The parasite genome architecture is relatively simple and does not contain well-defined compartments or TADs [2]. The typical Hi-C analysis steps to understand genomic compartmentalization do not give robust results in the context of *Plasmodium falciparum*, therefore we adapted our analysis towards the specificity of the data.

Future avenues

Integration: integration with other omics data such as RNA-seq, ChIP-seq, ATAC-seq, or DNA methylation data, to gain insights into the functional implications of chromatin interactions on gene regulation and genome organization.

Network analysis: interaction networks based on chromatin looping and other molecular data to elucidate higher-order functional relationships and regulatory networks.

8. References

[1] Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

[2] Batugedara, G., & Le Roch, K. G. (2019). Unraveling the 3D genome of human malaria parasites. *Seminars in cell & developmental biology*, *90*, 144–153. https://doi.org/10.1016/j.semcdb.2018.07.015

[3] Iwanaga, S., Kaneko, I., Kato, T., & Yuda, M. (2012). Identification of an AP2-family protein that is critical for malaria liver stage development. *PloS one*, *7*(11), e47557. https://doi.org/10.1371/journal.pone.0047557

[4] Subudhi, A. K., Green, J. L., … Pain, A. (2023). DNA-binding protein PfAP2-P regulates parasite pathogenesis during malaria parasite blood stages. *Nature microbiology*, *8*(11), 2154–2169. https://doi.org/10.1038/s41564-023-01497-6

[5] FastQC (2018). *Babraham Bioinformatics* http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[6] Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

[7] Shifu Chen. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107. https://doi.org/10.1002/imt2.107

[8] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

[9] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

[10] hic2cool https://github.com/4dn-dcic/hic2cool

[11] Knight. P., D. Ruiz. (2013) A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, Volume 33, Issue 3, Pages 1029–1047. https://doi.org/10.1093/imanum/drs019

[12] Lyu, H., Liu, E., & Wu, Z. (2020). Comparison of normalization methods for Hi-C data. *BioTechniques*, *68*(2), 56–64. https://doi.org/10.2144/btn-2019-0105

[13] Serizay, J., Matthey-Doret, C., Bignaud, A., Baudry, L., & Koszul, R. (2024). Orchestrating chromosome conformation capture analysis with Bioconductor. *Nature communications*, *15*(1), 1072. https://doi.org/10.1038/s41467-024-44761-x

[14] Stansfield, J. C., Cresswell, K. G., & Dozmorov, M. G. (2019). multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics (Oxford, England)*, *35*(17), 2916–2923. https://doi.org/10.1093/bioinformatics/btz048

[15] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

[16] Ay, F., Bailey, T. L., & Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research*, *24*(6), 999–1011. https://doi.org/10.1101/gr.160374.113

[17] Alexa A, Rahnenfuhrer J (2024). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.56.0.

[18] Subudhi, A. K., Green, J. L., … Pain, A. (2023). DNA-binding protein PfAP2-P regulates parasite pathogenesis during malaria parasite blood stages. *Nature microbiology*, *8*(11), 2154–2169. https://doi.org/10.1038/s41564-023-01497-6