Chromatin structure can introduce systematic biases in genome-wide analyses of Plasmodium falciparum

Sebastian Baumgarten, Jessica Bryant

RESEARCH ARTICLE

REVISED Chromatin structure can introduce systematic biases in genome-wide analyses of *Plasmodium falciparum* [version 2; peer review: 2 approved]

Sebastian Baumgarten[1], Jessica Bryant[2-4]

[1]Plasmodium RNA Biology Group, Pasteur Institute, Paris, Paris, 75015, France
[2]Biology of Host-Parasite Interactions Unit, Pasteur Institute, Paris, Paris, 75015, France
[3]INSERM U1201, Paris, France
[4]CNRS ERL9195, Paris, 75015, France

v2 **First published:** 10 Jun 2022, **2**:75
https://doi.org/10.12688/openreseurope.14836.1

Latest published: 15 Sep 2022, **2**:75
https://doi.org/10.12688/openreseurope.14836.2

Abstract

Background: The maintenance, regulation, and dynamics of heterochromatin in the human malaria parasite, *Plasmodium falciparum,* has drawn increasing attention due to its regulatory role in mutually exclusive virulence gene expression and the silencing of key developmental regulators. The advent of genome-wide analyses such as chromatin-immunoprecipitation followed by sequencing (ChIP-seq) has been instrumental in understanding chromatin composition; however, even in model organisms, ChIP-seq experiments are susceptible to intrinsic experimental biases arising from underlying chromatin structure.

Methods: We performed a control ChIP-seq experiment, re-analyzed previously published ChIP-seq datasets and compared different analysis approaches to characterize biases of genome-wide analyses in *P. falciparum*.

Results: We found that heterochromatic regions in input control samples used for ChIP-seq normalization are systematically underrepresented in regard to sequencing coverage across the *P. falciparum* genome. This underrepresentation, in combination with a non-specific or inefficient immunoprecipitation, can lead to the identification of false enrichment and peaks across these regions. We observed that such biases can also be seen at background levels in specific and efficient ChIP-seq experiments. We further report on how different read mapping approaches can also skew sequencing coverage within highly similar subtelomeric regions and virulence gene families. To ameliorate these issues, we discuss orthogonal methods that can be used to characterize *bona fide* chromatin-associated proteins.

Conclusions: Our results highlight the impact of chromatin structure on genome-wide analyses in the parasite and the need for caution

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 15 Sep 2022		✓ view
version 1 10 Jun 2022	✓ view	? view

1. **Joana Santos**, Université Paris-Saclay, Paris, France

2. **Elena Gómez-Díaz**, Consejo Superior de Investigaciones Científicas, Granada, Spain
   Diana C. López-Farfán, Consejo Superior de Investigaciones Científicas, Granada, Spain

Any reports and responses or comments on the article can be found at the end of the article.

when characterizing chromatin-associated proteins and features.

Keywords
Chromatin structure, ChIP-seq, Plasmodium falciparum, read alignment

This article is included in the Excellent Science gateway.

This article is included in the Genomes collection.

Corresponding authors: Sebastian Baumgarten (sebastian.baumgarten@pasteur.fr), Jessica Bryant (jessica.bryant@pasteur.fr)

Author roles: Baumgarten S: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Writing – Original Draft Preparation; **Bryant J**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 947819, to SB under the project “PlasmoEpiRNA”). JMB is supported by ANR grant PRC “ApiMORCing” and ANR grant JCJC “PlasmoVarOrg”. SB and JMB are supported by baseline funding provided by the Institut Pasteur.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Baumgarten S and Bryant J. **Chromatin structure can introduce systematic biases in genome-wide analyses of *Plasmodium falciparum* [version 2; peer review: 2 approved]** Open Research Europe 2022, **2**:75 https://doi.org/10.12688/openreseurope.14836.2

First published: 10 Jun 2022, **2**:75 https://doi.org/10.12688/openreseurope.14836.1

REVISED **Amendments from Version 1**

We have updated version 1 of our manuscript to address the reviewer comments. In particular, we added a paragraph to the discussion why intrinsic genome biases might affect input and control ChIP samples differentially and how the incorporation of additional mock IP controls can strengthen ChIP-seq analysis. We further discuss additional mapping and filtering options of different short read mappers, the effect of ChIP biases in different life-cycle stages of the parasite and how intrinsic genome biases might affect other genome-wide analysis.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the method of choice to identify where a chromatin feature associates with the genome on a global scale. However, the secondary structure and inherent sequence of the genome can be challenging for ChIP-seq analysis. Variation in DNA accessibility, similarity of multiple sequences, and DNA sequence diversity can lead to inaccurate analyses. In this study, we describe how these factors influence the analysis of genome-wide ChIP-seq data generated by Next Generation sequencing in the human malaria parasite *Plasmodium falciparum*. In particular, we observed that DNA regions associated with compact chromatin are underrepresented in samples used for ChIP normalization, leading to the identification of false enrichments across these regions. Furthermore, we show how the choice of options during the mapping of sequencing reads to the parasite genome and subsequent filtering steps can differentially affect regions with varying levels of similarity and nucleotide diversity. Together, these data highlight the sensitivity of genome-wide analyses to intrinsic chromatin features in the human malaria parasite and how orthogonal methods can be used to characterize chromatin-associated features.

Introduction

Epigenetic regulation of transcription has become a major focus in the study of *Plasmodium falciparum*, the eukaryotic parasite that causes the most severe form of human malaria. An ever-increasing number of studies have attempted to elucidate how this parasite, which has relatively few specific transcription factors compared to other eukaryotes, maintains sophisticated programs of gene regulation throughout its complex life cycle. Accordingly, chromatin-immunoprecipitation followed by sequencing (ChIP-seq) has become widely used to identify the genome-wide enrichment of putative regulatory proteins. In *P. falciparum*, ChIP-seq has been instrumental in characterizing the dynamics of histone post-translational modifications, transcription factors, and other chromatin-associated proteins involved in transcriptional activation and silencing (reviewed in [1,2]). Because of its importance to parasite pathogenicity and transmission, the formation, maintenance, and dynamics of heterochromatinization uniquely of subtelomeric regions and individual central chromosomal clusters has received special attention. Heterochromatin silences genes such as the virulence gene families encoding variant surface antigens (*e.g. var, rifin* and *stevor*) and *ap2-g,* a transcription factor that is essential for the differentiation of the parasite into the human-to-mosquito transmission stage.

As the field of epigenetics progresses, experimental standards for genome-wide studies of chromatin have been set forth by the ENCODE project, including using appropriate controls, replicates, normalization, and read depth; however, there are *P. falciparum*-specific issues that have come to our attention in recent years that mostly concern heterochromatic or highly homologous regions of the genome. In a recent report where we characterized the *var* gene interactome with enChIP, we used an epitope-tagged enzymatically inactive Cas9 (‘dead’, dCas9) that was co-expressed with a non-specific guide RNA as negative control[3]. Although the dCas9 protein had no intended target within the parasite genome, ChIP-seq analysis of this supposedly non-specific dCas9 showed a specific enrichment in heterochromatic regions across the parasite genome (Figure 1a, top). As this result was unexpected, we sought to discover an explanation in order to optimize our protocol. Performing and re-analyzing additional control experiments and using previously published ChIP-seq experiments of *bona fide* chromatin-associated proteins, we found that heterochromatic regions across the parasite genome are systematically under-represented in input samples used for ChIP-seq normalization. In combination with non-specific or inefficient immunoprecipitations, this bias can lead to the identification of enrichment in heterochromatic regions. Here, we highlight specific pitfalls and intrinsic confounding factors associated with genome-wide analyses and how orthogonal methods can be used in the characterization of chromatin-associated proteins in the human malaria parasite.

Results

In order to confirm the initial observation of background heterochromatin enrichment in the dCas9 control strain, we performed ChIP-seq on wild-type *P. falciparum* (3D7) ring stage parasites (12 hours post infection) using an α-HA antibody. In addition, using an identical pipeline (see Methods), we re-analyzed a control experiment that used an α-GFP antibody[4] as well as an α-HP1 ChIP-seq as positive control of a heterochromatin-associated protein[5]. For all three control experiments (*i.e.* control dCas9, α-HA and α-GFP), we observed similar enrichments across the subtelomeric and central heterochromatin clusters resembling the profile of HP1 (Figure 1a). Peak-calling analysis on the α-HA experiment using macs2 identified 1,332 significant (q-value ≤ 0.05) peaks that were significantly overrepresented within heterochromatic regions (505 peaks, χ^2-test $p < 2.2\times10^{-16}$, Figure 1a). We next asked whether such an intrinsic bias could also be observed within a ChIP-seq experiment of a specific DNA-binding protein. We re-analyzed our data of a ChIP-seq experiment that used a dCas9-HA protein with a guide RNA specific to the upstream region of 17 *var* genes[3]. Enrichment of the dCas9 protein was highly specific and robust at the targeted binding sites (Figure 1b, top and 3); however, when we visually decreased the enrichment range (y-axis), we could also observe a background enrichment across heterochromatic regions

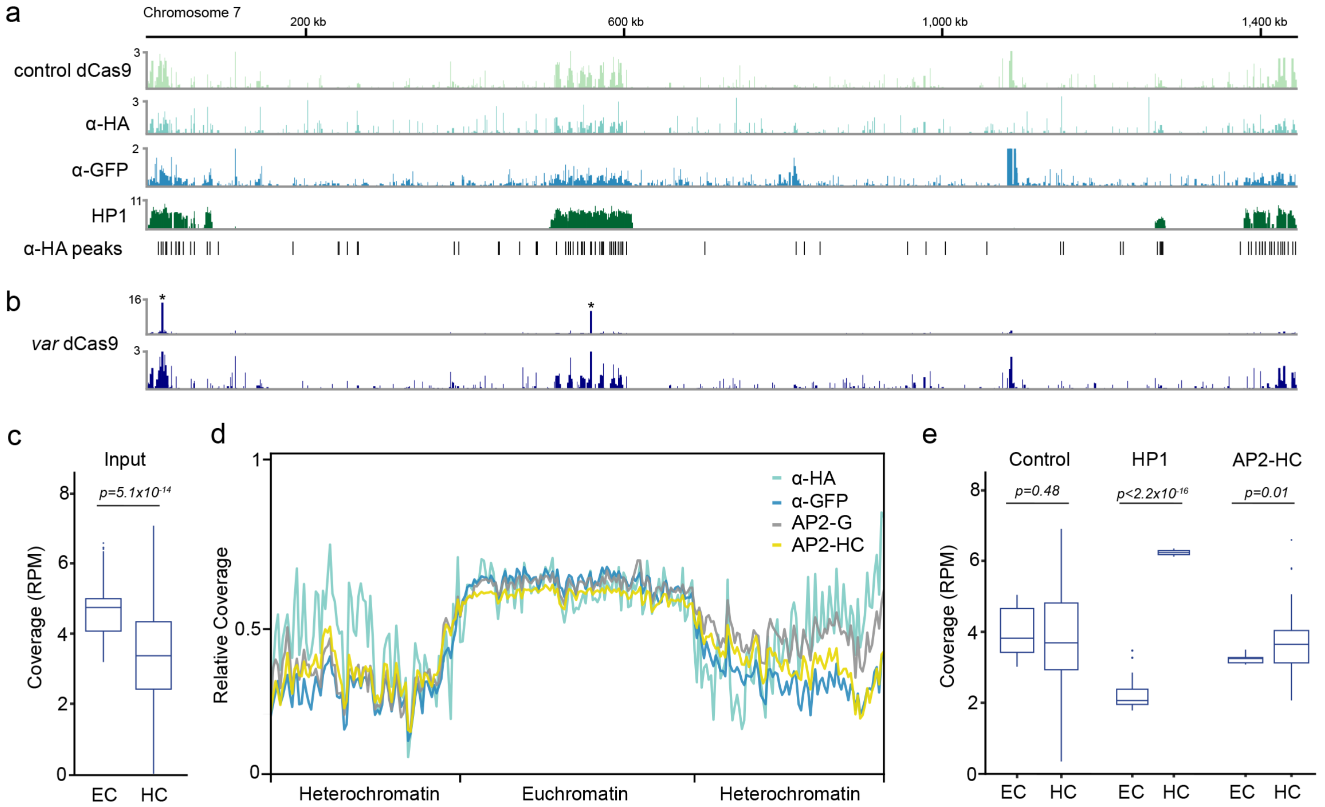

Figure 1. Enrichment of heterochromatic regions in control chromatin immunoprecipitation sequencing (ChIP-seq) experiments. a) Fold-enrichment (ChIP/input) of three control ChIP-seq experiments (control dCas9, α-HA, α-GFP) and the heterochromatin marker HP1 on chromosome 7. Significant peaks identified in the α-HA ChIP-seq experiment are indicated on the bottom. **b**) Fold-enrichment (ChIP/input) of a dCas9 ChIP-seq experiment on chromosome 7 scaled to the two peaks of intended dCas9 target sites (top, indicated with asterisks) and with decreased enrichment range (y-axis, bottom). **c**) Genome coverage (reads per million, RPM) in euchromatic (EC) and heterochromatic (HC) regions from six different ChIP-seq input libraries[3–6]. **d**) Metagene plot of the 14 nuclear chromosomes of *P. falciparum* of ChIP-seq input libraries used for normalization. **e**) Genome coverage (reads per million, RPM) in euchromatic (EC) and heterochromatic (HC) regions for three control ChIP-seq immunoprecipitations libraries (control dCas9, α-GFP, α-HA, 'control', left), HP1 (middle) and the heterochromatin-associated transcription factor AP2-HC (right).

similar to the enrichment observed for the control ChIP-seq experiments (Figure 1b, bottom).

Because immunoprecipitated chromatin is normalized to corresponding input chromatin, the observed heterochromatic enrichment in control/non-specific ChIP-seq experiments could arise either from 1) an overrepresentation of heterochromatinized regions in the immunoprecipitation or 2) an underrepresentation of heterochromatinized regions in the input sample. When we compared input samples from six different ChIP-seq experiments[3–6], we found that heterochromatic regions had significantly lower sequencing read coverage than euchromatic regions ($p = 5.1\times10^{-14}$, Figure 1c). This was particularly true for heterochromatic subtelomeric regions, where input samples from ChIP-seq experiments performed by four different laboratories all showed significantly lower coverage than the euchromatic, central chromosomal regions (Figure 1d)[3–6]. This underrepresentation of heterochromatic regions was not significant in the three control/non-specific immunoprecipitation samples ($p = 0.48$, Figure 1e). In contrast, immunoprecipitation samples for HP1 and a recently characterized heterochromatin-associated ApiAP2 transcription factor, AP2-HC, showed significantly higher read coverage in heterochromatic than in euchromatic regions ($p < 2.2\times10^{-16}$ and $p = 0.01$, respectively, Figure 1e).

In addition to the issue of heterochromatin representation, the composition of the *P. falciparum* genome also poses a challenge for ChIP-seq analyses. First, the overall GC content of the *P. falciparum* genome is approximately 19%, the lowest reported for any genome known to date[7]. Intergenic regions in particular show very low levels of nucleotide diversity and an average AT content of up to 90%. Second, subtelomeric regions and other regions containing members of multigene families share a high level of sequence similarity[7,8] and show an elevated rate of recombination events[9,10]. Both low nucleotide diversity and high sequence similarity among multiple gene loci or genomic regions complicate the read mapping step of genome-wide analyses, since the true location of a sequenced molecule originating from such a gene/region cannot be exactly inferred.

With default settings, the popular short-read aligner bowtie2 only reports one alignment per read, and the confidence of the alignment corresponding to the true origin of the sequenced molecule is given by the mapping quality (MAPQ). The higher the mapping quality, the larger the difference between the best and the second-best possible alignment of a given read. MAPQs are reported as $Q=-10*log_{10}(p)$, where p equals the probability that the reported alignment is not the true location from where the sequenced molecule originated[11]. Thus, a MAPQ value of 10 equals the probability of the reported alignment location being incorrect is 1/10, while a MAPQ value of 40 would indicate a probability of 1/10,000.

We used the input sample of the α-HA control ChIP experiment to compare the distribution of MAPQ values between alignments in euchromatic and heterochromatic regions. We found that 97% of aligned reads in euchromatic regions, but only 71% of aligned reads in heterochromatic regions had a MAPQ ≥ 31 (Figure 2a). Even more striking, the percentage of all read alignments with a MAPQ ≤ 10 was 2% for euchromatic regions and 22% for heterochromatic regions (Figure 2a). We observed a similar trend in the comparison of genic and intergenic regions, with 3% and 12% of all alignments within genic and intergenic regions, respectively, with a MAPQ ≤ 10 (Figure 2a, right). Filtering for reads with high MAPQ values (*i.e.* ≥ 30) therefore affects genic and intergenic regions disproportionally: when calculating the distribution of 100 basepair bins with a given GC content across the parasite genome, the peak of GC bins was at 18% (Figure 2b). In contrast, the alignments from the α-HA input sample described above were biased towards higher GC content, with the peak of the distribution falling within the 22% GC bin. These data suggest that experiments aiming to characterize features potentially binding within these regions (*e.g.* transcription factors) might require deeper sequencing to compensate for lower numbers of high-confidence alignments.

The discrepancy between different regions of the genome regarding the confidence of a read's true origin is exacerbated when instead of only one alignment, every possible alignment of a read is reported during read mapping and no further downstream quality filter step is included. Using the α-HA input sample, reporting of only one alignment per read (default in bowtie2) and filtering for MAPQ ≥ 30 resulted in the underrepresentation of heterochromatic regions relative to euchromatic regions as described above (p = 0.01, Figure 2c, left). Reporting all alignments for a given read (bowtie2 option -a) led to an overall increase in genome coverage for both euchromatic and heterochromatic regions. However, this increase in coverage was substantially higher for heterochromatic regions, leading to a significant overrepresentation of these regions compared to euchromatic regions (p = 0.02, Figure 2c, right; Figure 2d). On average, we found almost twice the number of alignments within heterochromatic compared to euchromatic regions when all possible alignments were reported, although heterochromatin in our analysis made up only 12% of the total genome size. Within subtelomeres, loci sharing the highest sequence similarity showed the most pronounced increase in coverage, including the conserved exon 2 of members of the *var* multigene family, and especially the telomeric repeat regions (Figure 2e). When multiple alignments of a read were reported, the average increase in coverage for genes located in euchromatic regions was only ~4 %, but ~30% for members of the *var* multigene family (Figure 2f). These data demonstrate the drawbacks of studying an organism with multigene families showing high sequence similarity and the importance of using stringent mapping parameters for analysis of Next Generation sequencing data.

Discussion

Challenges with ChIP-seq analysis surrounding heterochromatic regions and regions with similar sequences in the genome are not unique to *P. falciparum*; however, the presence of several multigene families in heterochromatic regions of the genome and low sequence diversity in intergenic regions makes *P. falciparum* a particularly difficult organism for generating robust ChIP-seq datasets. Chromatin fragmentation is a key step in any ChIP-seq protocol and can differ between samples or experiments, making it important to perform multiple replicates and to use an input sample from the same chromatin preparation used for the ChIP (see ENCODE ChIP-seq guidelines[12]). Chromatin structure affects chromatin fragmentation whether sonication or enzymatic cleavage are used, and heterochromatin tends to be more difficult to fragment by sonication than euchromatin[13]. Thus, fragments of the genome from heterochromatic regions tend to be longer than those from euchromatic regions and are not sequenced as efficiently and/or are lost when the sonicated chromatin is cleared by centrifugation. The underrepresentation of heterochromatic regions in the analyzed *P. falciparum* ChIP-seq input samples was not seen to the same extent in the immunoprecipitated DNA from non-specifically binding features or control antibodies. This leads to spurious "enrichment" of heterochromatic regions when immunoprecipitated DNA is normalized to corresponding input DNA (Figure 1). One possibility for why heterochromatin regions are less depleted in immunoprecipitated samples might be that (control) antibodies bind more often in an unspecific manner in protein-dense chromatin regions, and DNA within these regions therefore gets pulled-down more often[14]. In contrast, for true heterochromatin-associated proteins (*i.e.*, AP2-HC and HP1), we found that heterochromatic DNA was enriched compared to euchromatic DNA in the immunoprecipitated DNA sample (Figure 1e). Moreover, in the case of specific and/or effectively immunoprecipitated chromatin features, background enrichment of heterochromatic regions is negligible compared to the real ChIP signal (Figure 1b). Because heterochromatin is dynamic across the *P. falciparum* life cycle[15,16], the biases described are likely not fixed to specific chromosome coordinates, but might affect ChIP experiments differently depending on the investigated life cycle stage.

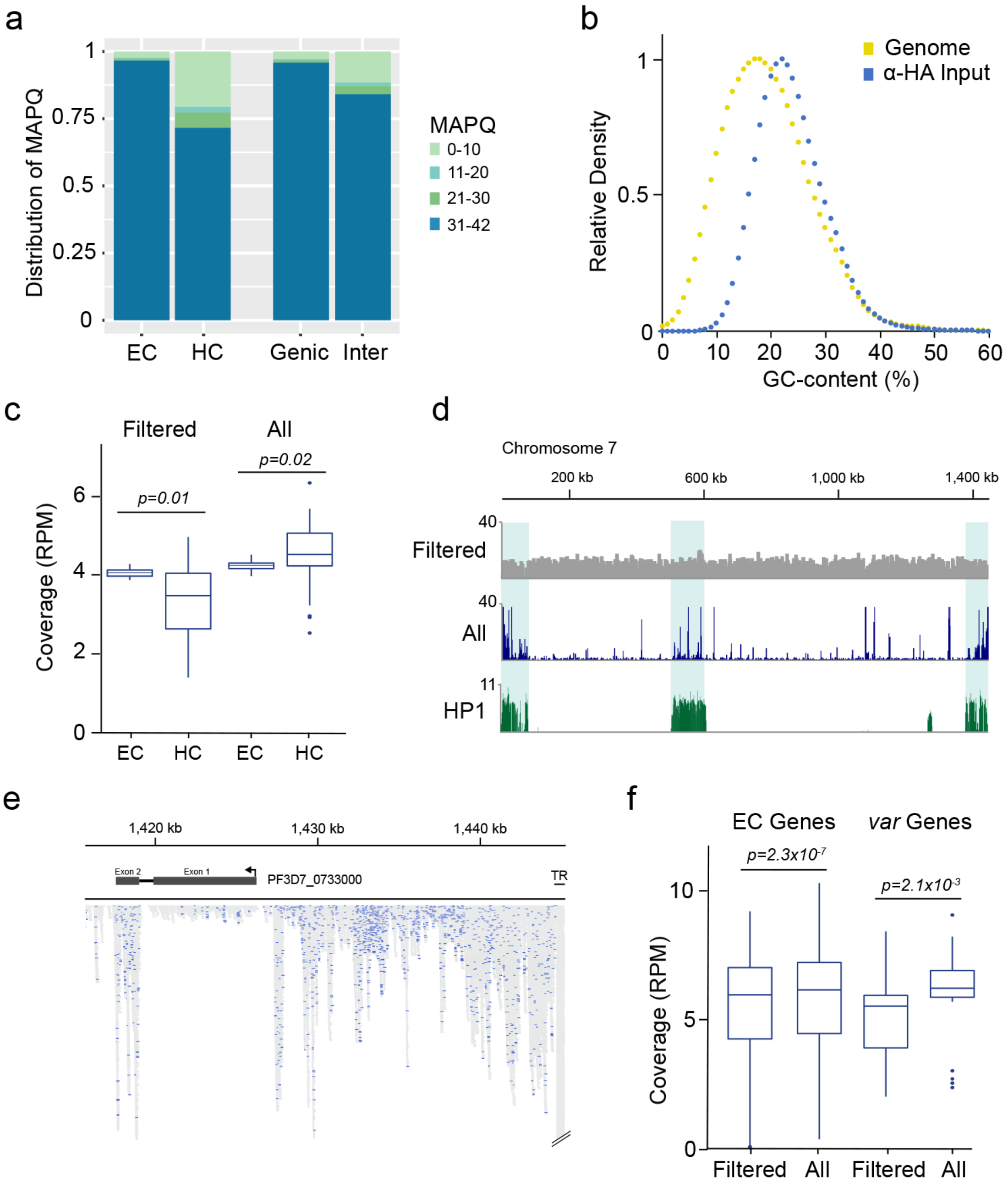

Figure 2. Read mapping biases between euchromatic and heterochromatic regions. a) Distribution of read alignment mapping quality (MAPQ) between euchromatic (EC), heterochromatic (HC), genic and intergenic (Inter) regions. **b**) Relative abundance of 100 bp windows sorted by GC content within the *P. falciparum* genome (Genome) and relative abundance of read alignments from the α-HA input library within 100 bp windows sorted by GC content (α-HA). **c**) Comparison of genome coverage (reads per million, RPM) within euchromatic (EC) and heterochromatic (HC) regions calculated from MAPQ-filtered alignments (*i.e.*, reporting one alignment per read with MAPQ ≥ 30, 'Filtered') and from all possible alignments (*i.e.*, reporting all alignments of a read mapping to multiple locations and without MAPQ-filtering, 'All'). **d**) Genome coverage of MAPQ-filtered (Filtered) and all possible (All) alignments across chromosome 7. Fold-enrichment (ChIP/input) of the heterochromatin marker HP1 is shown on the bottom. **e**) Detailed view of read alignments at the end of chromosome 7 when all possible alignments per read are reported. Blue: Alignments with a MAPQ ≥ 30. Grey: Alignments of reads mapping to multiple regions in the genome. TR: Telomere repeat. Arrow indicates direction of transcription **f**) Comparison of mapping approaches as in **c**) for genes located in euchromatic regions (EC Genes) and members of the *var* multigene family (*var* Genes). Coverage was calculated from MAPQ-filtered alignments (*i.e.*, reporting one alignment per read with MAPQ ≥ 30, 'Filtered') and from all possible alignments (*i.e.*, reporting all alignments of a read mapping to multiple locations and without MAPQ-filtering, 'All')

Another issue with analyzing ChIP-seq data in *P. falciparum* is the mapping of genomic regions with high sequence similarity and/or low nucleotide diversity. Allowing a read to map to multiple loci in the genome and reporting each alignment can result in false over-representation within multigene families (e.g. *var, rifin, stevor*) or ribosomal genes, which show relatively high levels of homology. Similar issues are prominent and have been reported previously in model systems, for example regarding the repeat targets of PIWI-interacting RNA[17]. To ensure robust and replicable data, analysis of ChIP-seq experiments should include stringent mapping and filtering steps for both input and immunoprecipitation samples and include 1) the reporting of only one alignment per sequenced read with a high alignment score, 2) the removal of PCR duplicates, and 3) filtering out alignments that might not represent the true origin of a sequenced read (*e.g.*, low MAPQ values in bowtie2). Performing Next Generation sequencing with longer and/or paired-end reads can help in assigning the origin of a sequenced read more accurately, especially with regard to heterochromatic genomic loci with high sequence similarity.

For other short read aligners, similar mapping and filtering options are available. In bwa 'mem'[18] and the RNA-seq mapping tool STAR[19], only allowing reads to map once to the genome can be set with option '-c 1' and '--outFilterMultimapNmax 1', respectively. bwa also calculates MAPQ values in a similar manner to bowtie2, whereas STAR assigns uniquely mapping reads a MAPQ value of 255. Importantly, while other genome-wide techniques (e.g. ATAC-seq) might require different and/or additional controls, the read mapping and filtering steps outlined here can also guide the analysis of such experiments.

Importantly, our peak analysis of control ChIP-seq experiments also showed that despite following stringent read-mapping and filtering procedures, one can detect seemingly specific and significant enrichments across the *P. falciparum* genome in input-normalized negative control ChIP samples. Because bioinformatic tools are limited in discriminating whether an enrichment is truly biological or originates from intrinsic biases and technical noise, purely relying on *p*-value cutoffs can lead to inaccurate conclusions. Therefore, in addition to a rigorous analysis pipeline, experimental controls can provide additional confidence in a ChIP-seq data set, especially when performing ChIP-seq for the first time with an uncharacterized chromatin feature or antibody. First, a chromatin feature of interest should be present at high enough levels at the time point in the life cycle being investigated to result in a robust immunoprecipitation. Here, a basic immunoprecipitation followed by Western blot from a crosslinked sample should be used to determine whether the feature of interest can be enriched from the nuclear or chromatin fraction. For an epitope-tagged chromatin feature, a key control ChIP-seq experiment would be to use the antibody against the epitope tag with the same amount of input chromatin, but from the wild-type parent strain that does not contain any epitope-tagged proteins. For an uncharacterized antibody, Western blots and immunoprecipitation followed by mass spectrometry could provide evidence for specificity (see ENCODE guidelines), and immunoprecipitation with immunoglobulin G (IgG) could serve as a ChIP-seq control[14]. Additional methods for validating ChIP-seq data are ChIP followed by quantitative PCR using highly specific primers (if possible).

If a region of the genome is enriched in both a test and control ChIP-seq experiment, it is still possible that the enrichment is real. To provide additional experimental support for true ChIP-seq enrichment of a chromatin feature, different orthogonal methods have been successfully used in *P. falciparum*. One option is to determine if the binding profile of the chromatin feature of interest is dynamic (*e.g.*, enrichment may change depending on the life cycle stage or growth conditions of the parasite), as in 6. Another option is to detect changes in enrichment when the chromatin feature of interest or auxiliary factor is depleted (*e.g.*, with a knockout or knockdown), as in 5. Even further support could be provided if the knockout/down affects the transcription of genes that are enriched for the chromatin feature in the ChIP-seq data, as in 3,6,20. One important caveat of this last type of analysis is that heterochromatinized multigene families expressed in a mutually exclusive manner will always appear to be differentially expressed if two different clones are used for the comparison. Thus, using an inducible knockdown/out system in the same parasite clone has to be used to obtain interpretable data for differential expression analysis of multigene families.

Using these bioinformatic and experimental techniques will allow the field of chromatin and epigenetics in *P. falciparum* to progress and reveal important processes in the gene regulation of this parasite. Maintaining high experimental and bioinformatic standards in the analysis of genome-wide features will ensure standing in the field of epigenetics and chromatin, which is so often dominated by model systems.

Methods

Chromatin immunoprecipitation

The ChIP-seq experiment for the α-HA samples was performed as described in detail in 3 using 10^9 tightly synchronized wild-type *P. falciparum* (strain 3D7) parasites harvested at 12 hours post-infection, and 75 µL protein G Dynabeads (Invitrogen 10004D) conjugated to 3 µg α-HA antibody (Abcam ab9110). fastq-dump.2 was used to download data generated previously from the NCBI Sequencing Read Archive[21] with the following accession numbers: control dCas9 input: SRR8802083[3]; control dCas9 IP: SRR8802084[3]; *var* dCas9 input: SRR8802087[3]; *var* dCas9 IP: SRR8802088[3]; α-GFP input: SRR16021005[4]; α-GFP IP: SRR16021003[4]; AP2-HC input: SRR12281322[5]; AP2-HC IP: SRR12281321[5]; HP1 IP: SRR12281320[5]; AP2-G input: SRR7903647[6].

Read mapping and filtering

Illumina sequencing adapters were trimmed from raw fastq files using trimmomatic (version 0.39)[22], removing poor-quality bases at both read ends (Phred score ≤ 20) and applying a 4 bp sliding-window trimming (option SLIDINGWINDOW:4:20). Only reads ≥ 50 nucleotides and proper read pairs (in the case of paired-end libraries) were retained. Trimmed reads were mapped to the *P. falciparum* genome[7] downloaded from plasmoDB.org (version v55)[23] with bowtie2[11] using options --end-to-end and –sensitive. With these settings, bowtie2 reports only one alignment per read that can be further filtered for 'uniqueness' by its MAPQ value. For paired-end reads, the additional options --no-mixed and --no-discordant were used. PCR duplicates were filtered from the raw alignments with samtools[24] 'fixmate' and 'markdup' (with option -r). samtools 'view' was used to filter high quality (*i.e.* more unique) sequencing alignments (MAPQ ≥ 30, option -q 30). For the MAPQ distribution (Figure 2a), read alignments were processed using the same steps without the final MAPQ filtering step. Significant peaks in the α-HA ChIP-seq experiment were identified with macs2[25] 'callpeak' using default settings and options --no-model and --extsize 150.

Genome coverage calculation

BED files for eu- and heterochromatic regions for downstream analysis were generated using the HP1 ChIP-seq data generated in 26, with heterochromatic regions being those featuring enrichment of HP1 and euchromatic regions encompassing all remaining regions of the genome. Genic regions (coding sequences, CDS) were extracted from the *P. falciparum* genome annotation file (plasmoDB GFF, version 55). Intergenic regions were computed using the *P. falciparum* genome annotation file and bedtools 'complement'[27]. Coverage plots of ChIP/input fold enrichments (Figure 1a,b) were generated using deeptool's bamCompare[28] by calculating the fold-enrichment between the IP and Input sample in 10 bp bins (option --bs 10) with options --operation 'ratio' and --normalizeUsing CPM. For the metagene plot of input samples, genome coverage was calculated using deeptool's bamCoverage in bin sizes of 1000 bp (option --bs 1000) and normalized to counts per million (option --normalizeUsing CPM). The metagene was calculated using deeptool's computeMatrix by scaling each euchromatic, central chromosome region in the 14 nuclear chromosomes to the same length and defining the flanking 80 kilobases as subtelomeric, heterochromatic regions. The graph was plotted using deeptool's plotProfile with default settings.

Average genome coverages of hetero- and euchromatic regions and genes (Figure 1c,e, Figure 2c,f) were calculated using mosdepth[24] with default settings. For between-sample comparisons, the total number of alignments were calculated using samtools flagstat and the average coverage per region was normalized to one million mapped reads (reads per million, RPM). Boxplots were generated in R using package ggplot2[29]. The GC-content analysis for the genome and the α-HA input sample was computed using 'CollectGcBiasMetrics' from the picard package.

Mapping approach comparisons

To report all possible alignments of a given read, the input sample of the α-HA ChIP-seq experiment was mapped using bowtie2 as described above and with option -a and without further MAPQ filtering. Coverage plots (Figure 2d) were generated using deeptool's bamCoverage with option --bs 10, --normalizeUsing CPM. For the comparison with the MAPQ-filtered alignments, average genome coverage in the different genomic regions and genes was calculated using mosdepth[30] and normalized to reads per million (calculated using samtools flagstat from the quality filtered alignment file). All genome coverage tracks and read alignments (Figure 1a, b and Figure 2d,e) were visualized using the Integrative Genomics Viewer (version 2.12.3)[31].

Data availability

Underlying data

NCBI BioProject: Systematic biases in genome-wide analyses of Plasmodium falciparum, https://identifiers.org/ncbiprotein:PRJNA832605

This project contains the sequencing data generated in this study.

References

1. Duraisingh MT, Skillman KM: **Epigenetic variation and regulation in malaria parasites.** *Annu Rev Microbiol.* 2018; **72**: 355–375.
PubMed Abstract | **Publisher Full Text**

2. Cortés A, Deitsch KW: **Malaria Epigenetics.** *Cold Spring Harb Perspect Med.* 2017; **7**(7): a025528.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Bryant JM, Baumgarten S, Dingli F, *et al.*: **Exploring the virulence gene interactome with CRISPR/dCas9 in the human malaria parasite.** *Mol Syst Biol.* 2020; **16**(8): e9569.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Shang X, Wang C, Fan Y, *et al.*: **Genome-wide landscape of ApiAP2 transcription factors reveals a heterochromatin-associated regulatory network during *Plasmodium falciparum* blood-stage development.** *Nucleic Acids Res.* 2022; **50**(6): 3413–3431.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Carrington E, Cooijmans RHM, Keller D, *et al.*: **The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the malaria parasite *Plasmodium falciparum*.** *iScience.* 2021; **24**(5): 102444.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Josling GA, Russell TJ, Venezia J, *et al.*: **Dissecting the role of PfAP2-G in malaria gametocytogenesis.** *Nat Commun.* 2020; **11**(1): 1503.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Gardner MJ, Hall N, Fung E, *et al.*: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature.* 2002; **419**(6906): 498–511.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. De Bruin D, Lanzer M, Ravetch JV: **The polymorphic subtelomeric regions of *Plasmodium falciparum* chromosomes contain arrays of repetitive sequence elements.** *Proc Natl Acad Sci U S A.* 1994; **91**(2): 619–623.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Freitas-Junior LH, Bottius E, Pirrit LA, *et al.*: **Frequent ectopic recombination of**

virulence factor genes in telomeric chromosome clusters of *P. falciparum.* *Nature.* 2000; **407**(6807): 1018–1022.
PubMed Abstract | Publisher Full Text

10. Deitsch KW, Del Pinal A, Wellems TE: **Intra-cluster recombination and var transcription switches in the antigenic variation of *Plasmodium falciparum.*** *Mol Biochem Parasitol.* 1999; **101**(1–2): 107–116.
PubMed Abstract | Publisher Full Text

11. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Landt SG, Marinov GK, Kundaje A, *et al.*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Res.* 2012; **22**(9): 1813–1831.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Teytelman L, Özaydin B, Zill O, *et al.*: **Impact of chromatin structures on DNA processing for genomic analyses.** *PLoS One.* 2009; **4**(8): 1–11.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Xu J, Kudron MM, Victorsen A, *et al.*: **To mock or not: a comprehensive comparison of mock IP and DNA input for ChIP-seq.** *Nucleic Acids Res.* 2021; **49**(3): e17.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Zanghì G, Vembar SS, Baumgarten S, *et al.*: **A Specific PfEMP1 Is Expressed in *P. falciparum* Sporozoites and Plays a Role in Hepatocyte Infection.** *Cell Rep.* 2018; **22**(11): 2951–2963.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Gómez-Díaz E, Yerbanga RS, Lefèvre T, *et al.*: **Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in Anopheles gambiae.** *Sci Rep.* 2017; **7**: 40655.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Marinov GK, Wang J, Handler D, *et al.*: **Pitfalls of mapping high-throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified.** *Dev Cell.* 2015; **32**(6): 765–771.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Brancucci NMB, Bertschi NL, Zhu L, *et al.*: **Heterochromatin protein 1 secures survival and transmission of malaria parasites.** *Cell Host Microbe.* 2014; **16**(2): 165–176.
PubMed Abstract | Publisher Full Text

21. Leinonen R, Sugawara H, Shumway M: **The sequence read archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D19–21.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.
PubMed Abstract | Publisher Full Text | Free Full Text

23. The Plasmodium Genome Database Collaborative: **PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium Genome Database Collaborative.** *Nucleic Acids Res.* 2001; **29**(1): 66–69.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Fraschka SA, Filarsky M, Hoo R, *et al.*: **Comparative heterochromatin profiling reveals conserved and unique epigenome signatures linked to adaptation and development of malaria parasites.** *Cell Host Microbe.* 2018; **23**(3): 407–420.e8.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Quinlan AR: **BEDTools: BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics.* 2014; **47**: 11.12.1–34..
PubMed Abstract | Publisher Full Text | Free Full Text

28. Ramírez F, Ryan DP, Grüning B, *et al.*: **deepTools2: a next generation web server for deep-sequencing data analysis.** *Nucleic Acids Res.* 2016; **44**(W1): W160–5.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York. 2016.
Publisher Full Text

30. Pedersen BS, Quinlan AR: **Mosdepth: Quick coverage calculation for genomes and exomes.** *Bioinformatics.* 2018; **34**(5): 867–868.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–26.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 23 September 2022

https://doi.org/10.21956/openreseurope.16303.r30102

✓

Elena Gómez-Díaz
Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, Granada, Spain
Diana C. López-Farfán
Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, Granada, Spain

The authors have successfully addressed my previous concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenomics, Malaria, Adaptation

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 25 July 2022

https://doi.org/10.21956/openreseurope.16025.r29565

?

Elena Gómez-Díaz
Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones

Científicas, Granada, Spain

Diana C. López-Farfán

Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, Granada, Spain

The manuscript by Sebastian Baumgarten *et al*. examines the impact of chromatin structure on ChIP-seq analysis in the malaria parasite Plasmodium falciparum. Authors evaluate how its unique genomic features impose strong bias and can lead to misleading results.

The article is clearly written, well reasoned and correct in the claims and suggestions. In particular, the main claim of this work is that heterochromatic regions appear depleted in ChIP input controls because of technical and experimental issues and this can result in a false enrichment of heterochromatic regions in the tested sample. The authors present experimental evidence of the low sequencing read coverage of heterochromatic regions compared with euchromatic regions in several ChIP input samples, and how this bias affects normalisation of telomeric and sub-telomeric regions, which are important in P. falciparum virulence and pathogenesis. In the discussion they then make recommendations to improve the Chromatin IP experiments design and analysis.

Although the claim is important, it is unclear which new insight the article and the reanalysis brings in terms of what factors are responsible for the bias, or the mitigation strategies that should be considered to eliminate that bias. Most recommendations are already covered in the ENCODE practical guidelines for the Analysis of ChIP-seq Data, for example, the use of an inespecific antibody (i.e. IgG) in addition to the input control (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/, https://pubmed.ncbi.nlm.nih.gov/24244136/).[1,2] The features that are responsible for the bias are also known in the malaria field: sequence composition (AT richness), complexity (sequence similarity of copy number genes), and chromatin structure. In fact, the original papers of the external ChIP-seq data referenced already apply most recommendations in terms of read mapping and quality control (for example see https://academic.oup.com/nar/article/50/6/3413/6548410).[3]

Perhaps it would be interesting to add further analysis on which telomeric/subtelomeric genes are more impacted or which life-cycle stages can be more affected. Also it would be great to combine ChIP-seq data with data measuring chromatin accessibility (for example ATAC-seq or MNAse-seq) to demonstrate how closed vs open regions are more or less biased in control vs test samples.

Besides, we would suggest the authors to make the recommendations in terms of design, experiments and analysis more accessible, like a guideline or protocol for non-experts.

Finally, the article would benefit from a more broad focus and perspective. For example, many of these problems raised also affect other NGS data types which are also impacted by chromatin structure and the intrinsic genomic characteristics of Pf. For example, in the case of ATAC-seq data coding regions which are GC rich compared with AT rich intergenic regions, are overrepresented due to PCR-amplification preference of GC-rich regions. In that case, it is important to include a transposed naked DNA as a control to correct the test data because of that bias in sequence composition. All mapping considerations that affect ChIP-seq data also affect ATAC-seq data for example. In that sense, why not expand recommendations to other types of NGS data like the

ATAC-seq for which there is such a well-established ENCODE manual?

Another area in which the recommendations made are very important is in CRISPR-Cas genomic experiments, very impacted by chromatin structure.

Going even further, in the light of single-cell technologies I would suggest expanding on how these problems will impact single-cell ChIP-data.

Other comments:

- Despite the fold enrichment of negative control ChIPs is lower/negligible to the specific ChIP enrichment, the peak calling analysis of control anti-HA ChIP for example, identified a lot of peaks that were significantly overrepresented within heterochromatic regions, this seems to indicate a problem with the normalization and peak calling analysis. Do authors recommend any bioinformatic analysis that can correct this false signal?

- To improve the figure legend 1, it is not fully self-explicative, you have to look in the text to fully understand it. E.g. In 1c, it is not stated that the input data is from six different ChIP-seq input libraries. In 1e, control is data from different controls ChIP together.

- The authors show that the underrepresentation of heterochromatic regions in the analyzed P. falciparum ChIP-seq input samples (Fig.1c), was not seen in the immunoprecipitated DNA from non-specific control antibodies (Fig. 1e, control), and said this leads to fake “enrichment” of heterochromatic regions when immunoprecipitated DNA is normalized to corresponding input DNA (Fig. 1a). However, it is not discussed why the genome coverage (RMP) of ChIP with control antibodies is not similar to the input chromatin. Are the euchromatin fragments lost during the ChIP process or is there a fraction heterochromatin fragments binding not specifically to the control antibodies, if the last is true, one possible solution that they did not comment on could be normalizing the ChIP experiment with the control non-specific ChIP instead of the input sample. The authors do comment on why the heterochromatin fragments could be underrepresented in the input sample (Fig 1c), e.g fragments of heterochromatic regions tend to be longer than those from euchromatic regions and are not sequenced as efficiently and/or are lost when the sonicated chromatin is cleared by centrifugation, but they don’t comment why the heterochromatin fragments are not underrepresented in the control IPs (Fig 1e. control)

References

1. Landt SG, Marinov GK, Kundaje A, Kheradpour P, et al.: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.*Genome Res*. 2012; **22** (9): 1813-31 PubMed Abstract | Publisher Full Text
2. Bailey T, Krajewski P, Ladunga I, Lefebvre C, et al.: Practical guidelines for the comprehensive analysis of ChIP-seq data.*PLoS Comput Biol*. 2013; **9** (11): e1003326 PubMed Abstract | Publisher Full Text
3. Shang X, Wang C, Fan Y, Guo G, et al.: Genome-wide landscape of ApiAP2 transcription factors reveals a heterochromatin-associated regulatory network duringPlasmodium falciparum blood-stage development. *Nucleic Acids Research*. 2022; **50** (6): 3413-3431 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenomics, Malaria, Adaptation

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 28 Aug 2022

Sebastian Baumgarten, Pasteur Institute, Paris, Paris, France

The manuscript by Sebastian Baumgarten et al. examines the impact of chromatin structure on ChIP-seq analysis in the malaria parasite Plasmodium falciparum. Authors evaluate how its unique genomic features impose strong bias and can lead to misleading results. The article is clearly written, well reasoned and correct in the claims and suggestions. In particular, the main claim of this work is that heterochromatic regions appear depleted in ChIP input controls because of technical and experimental issues and this can result in a false enrichment of heterochromatic regions in the tested sample. The authors present experimental evidence of the low sequencing read coverage of heterochromatic regions compared with euchromatic regions in several ChIP input samples, and how this bias affects normalisation of telomeric and sub-telomeric regions, which are important in P. falciparum virulence and pathogenesis. In the discussion they then make recommendations to improve the Chromatin IP experiments design and analysis.

Although the claim is important, it is unclear which new insight the article and the reanalysis brings in terms of what factors are responsible for the bias, or the mitigation strategies that should be considered to eliminate that bias. Most recommendations are already covered in the ENCODE practical guidelines for the Analysis of ChIP-seq Data, for example, the use of an inespecific antibody (i.e. IgG) in addition to the input control (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/, *https://pubmed.ncbi.nlm.nih.gov/24244136/). The features that are responsible for the bias are also known in the malaria field: sequence composition (AT richness), complexity (sequence similarity of copy number genes), and chromatin structure. In fact, the original papers of the external ChIP-seq data referenced already apply most recommendations in terms of read mapping and quality control (for example see https://academic.oup.com/nar/article/50/6/3413/6548410).*

Response: We agree that the central ENCODE ChIP guidelines are clear and identical for any kind of species investigated (as stated in the introduction). However, as there have been so few chromatin/epigenetics studies in the Plasmodium field relative to model eukaryotes, we feel the standards have not been set as stringently. What we wanted to make clear and share with the community is that even when one follows these guidelines and performs the most stringent analysis possible, one can observe seemingly “specific” heterochromatic enrichment even in negative control samples. While some groups and researchers in the field might know about the particularities of the *P. falciparum* genome, the effects such features can have on genome-wide analysis has not been described in a systematic fashion yet. Our intention with this manuscript was not to provide new insight into the specific reasons of how and why this might happen (which is covered elsewhere already[1,2]), but to make these specific biases in the *P. falciparum* genome known and clear to those who are new to chromatin biology in *P. falciparum* and/or do not have prior experience with such data but might be involved in collaborations or act as reviewers.

Perhaps it would be interesting to add further analysis on which telomeric/subtelomeric genes are more impacted or which life-cycle stages can be more affected. Also it would be great to combine ChIP-seq data with data measuring chromatin accessibility (for example ATAC-seq or MNAse-seq) to demonstrate how closed vs open regions are more or less biased in control vs test samples.

Response: In Figure 2e and 2f, we show how *var* genes that share high sequence similarity are particularly affected by different mapping approaches and have extended the text of these findings to other multigene families in the discussion. Because subtelomeric heterochromatin is dynamic between asexual and mosquito stages of the parasite, the biases in regard to sequencing coverage of input ChIP samples are likely also different between these life-cycle stages as now stated in the discussion). It is likely that closed chromatin is more refractory to sonication as reported elsewhere, which will lead to lower coverage across these regions and additional biases.[2] Notably though, it is known that intergenic regions are the most dynamic in terms of chromatin accessibility during the asexual replication of the parasite.[3] As shown in Figure 2a, these regions also feature substantially decreased mappability due to lower GC-content and sequence diversity. Therefore, intergenic regions are already biased against in terms of genome coverage (Fig. 2b) due to the intrinsic AT-richness, irrespective of chromatin accessibility. To our knowledge, there has not been a publication that includes an ATAC-seq and ChIP-seq dataset from the same parasite cell lines and time points. Thus, a direct and stringent comparison of biases in ChIP-seq samples due to open vs. closed chromatin states is not possible at this point. However, DNA sequence-based biases introduced with the Tn5 enzyme used for ATAC-seq have been analyzed and addressed in 3.

Besides, we would suggest the authors to make the recommendations in terms of design, experiments and analysis more accessible, like a guideline or protocol for non-experts.

Response: We provide fairly straightforward recommendations (often based on the ENCODE guidelines) with regard to experimental design and controls in the discussion. However, we also feel that it is very important for non-experts wanting to enter the field of epigenetics/chromatin/RNA biology in *P. falciparum* to work closely with experts and a skilled bioinformatician who have tested and established protocols which are difficult to convey with the necessary detail within the constraints of a manuscript. However, in an attempt to make the analysis description more accessible, we have added information on the performance of two other read mappers in the discussion section.

Finally, the article would benefit from a more broad focus and perspective. For example, many of these problems raised also affect other NGS data types which are also impacted by chromatin structure and the intrinsic genomic characteristics of Pf. For example, in the case of ATAC-seq data coding regions which are GC rich compared with AT rich intergenic regions, are overrepresented due to PCR-amplification preference of GC-rich regions. In that case, it is important to include a transposed naked DNA as a control to correct the test data because of that bias in sequence composition. All mapping considerations that affect ChIP-seq data also affect ATAC-seq data for example. In that sense, why not expand recommendations to other types of NGS data like the ATAC-seq for which there is such a well-established ENCODE manual?
Response: We agree that all biases that can arise during the analysis of sequencing data due to the intrinsic features of the *P. falciparum* genome (e.g. multi read-mappers, lower mapping efficiencies in AT-rich regions) likely affect different genome-wide analyses in a similar fashion, and we have added a paragraph in this regard to the updated version of the manuscript. Since we do not have the same extensive experience with ATAC-seq experiments as we do with ChIP-seq, however, we want to refrain from making recommendations for other researchers on how to perform such experiments. Moreover, we feel that the ATAC-seq publications from Toenhake et al. and Ruiz et al. have addressed these sequence-based biases adequately and make recommendations within the manuscript. [3, 4] While RNA-seq, ATAC-seq, ChIP-seq, etc. will all suffer from DNA sequence-based biases due to library preparation, NGS, and analyses, ChIP-seq involves more experimental steps that introduce potential biases such as antibody specificity, crosslinking, and sonication. This, and the fact that we have more experience with and many more datasets for ChIP-seq is why we focus on this specific experiment type in our manuscript.

Another area in which the recommendations made are very important is in CRISPR-Cas genomic experiments, very impacted by chromatin structure.
Response: This is very true, but is already well described and discussed elsewhere.[5,6]

Going even further, in the light of single-cell technologies I would suggest expanding on how these problems will impact single-cell ChIP-data.
Response: Since single-cell ChIP-seq relies on enzymatic (i.e. MNase) chromatin fragmentation that is also susceptible to chromatin structure, it is likely that similar intrinsic biases exist with this method. However, we (and nobody else in the *P. falciparum* field to our knowledge) have not tested this in the lab and therefore do not have the experimental evidence to make a statement in this regard.

Despite the fold enrichment of negative control ChIPs is lower/negligible to the specific ChIP enrichment, the peak calling analysis of control anti-HA ChIP for example, identified a lot of peaks

that were significantly overrepresented within heterochromatic regions, this seems to indicate a problem with the normalization and peak calling analysis. Do authors recommend any bioinformatic analysis that can correct this false signal?
Response: The peak calling analysis was performing as expected, because it does detect differences in coverage and read clustering between the IP and input sample after integrated background normalization. We included this analysis to highlight the fact that even when following a stringent mapping and filtering approach, one cannot rely on *p*-values alone, and that the observation described in this manuscript is not an issue that can primarily and exclusively be solved *in silico.* To ensure that the ChIP signal is truly of biological origin and to mitigate the risk of analyzing background enrichments, additional controls (e.g. IgG or negative control ChIP, verification of antibody specificity, ChIP followed by Western Blot) should be combined with orthogonal methods that help in cross-verifying ChIP signals. We emphasized this point in the discussion of the revised version of the manuscript. However, as shown in Fig 1a in the case where background enrichment co-occurs with specific enrichments, more stringent fold-enrichment and *p*-value cut-offs can exclude such peaks from downstream analysis.

To improve the figure legend 1, it is not fully self-explicative, you have to look in the text to fully understand it. E.g. In 1c, it is not stated that the input data is from six different ChIP-seq input libraries. In 1e, control is data from different controls ChIP together.
Response: We have added additional explanation to this legend and also changed the color schemes in Figure 1d to discriminate data better from those shown in 1a.

The authors show that the underrepresentation of heterochromatic regions in the analyzed P. falciparum ChIP-seq input samples (Fig.1c), was not seen in the immunoprecipitated DNA from non-specific control antibodies (Fig. 1e, control), and said this leads to fake "enrichment" of heterochromatic regions when immunoprecipitated DNA is normalized to corresponding input DNA (Fig. 1a). However, it is not discussed why the genome coverage (RMP) of ChIP with control antibodies is not similar to the input chromatin. Are the euchromatin fragments lost during the ChIP process or is there a fraction heterochromatin fragments binding not specifically to the control antibodies, if the last is true, one possible solution that they did not comment on could be normalizing the ChIP experiment with the control non-specific ChIP instead of the input sample. The authors do comment on why the heterochromatin fragments could be underrepresented in the input sample (Fig 1c), e.g fragments of heterochromatic regions tend to be longer than those from euchromatic regions and are not sequenced as efficiently and/or are lost when the sonicated chromatin is cleared by centrifugation, but they don't comment why the heterochromatin fragments are not underrepresented in the control IPs (Fig 1e. control)
Response: The heterochromatin fraction of control IP samples analyzed in this manuscript is still lower than those of the respective euchromatin samples (Fig 1 e, 'control'), although to a lesser degree as in the input sample and not significantly (p = 0.48). One possibility for why heterochromatin regions are less depleted in IP samples might be that control antibodies bind more often in an unspecific manner in protein-dense chromatin regions, and DNA within these regions therefore gets pulled-down more often, as was shown in other systems.[7] We added this point to the discussion of the updated manuscript version. Indeed, normalization to a control IP is another possibility to mitigate the effects of intrinsic biases, yet also comes with the introduction of other biases such as the differential pulldown of chromatin regions.[7] Control IP samples also feature lower sequence complexity

and can therefore lead to uneven background coverage and overamplification, which can also lead to uneven background coverage.[1] We added a comment to the discussion of the updated version of the manuscript about how combining control IP and input DNA normalizations can help to additionally mitigate the risk of identifying false enrichments in ChIP-seq experiments of *Plasmodium*.

References

1. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
2. Teytelman, L. *et al.* Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* **4**, 1–11 (2009).
3. Toenhake, C. G. *et al.* Chromatin accessibility-based characterization of the gene regulatory network underlying *Plasmodium falciparum* blood-stage development. *Cell Host Microbe* **23**, 557-569.e9 (2018).
4. Ruiz, J. L. *et al.* Characterization of the accessible genome in the human malaria parasite *Plasmodium falciparum*. *Nucleic Acids Res.* **46**, 9414–9431 (2018).
5. Bryant, J. M., Baumgarten, S., Glover, L., Hutchinson, S. & Rachidi, N. CRISPR in Parasitology: Not Exactly Cut and Dried! *Trends Parasitol.* **35**, 409–422 (2019).
6. Lee, M. C. S., Lindner, S. E., Lopez-Rubio, J.-J. & Llinás, M. Cutting back malaria: CRISPR/Cas9 genome editing of *Plasmodium*. *Brief. Funct. Genomics* **18**, 281–289 (2019).
7. Xu, J. *et al.* To mock or not: a comprehensive comparison of mock IP and DNA input for ChIP-seq. *Nucleic Acids Res.* **49**, e17 (2021).

Competing Interests: No competing interests were disclosed.

Reviewer Report 22 June 2022

https://doi.org/10.21956/openreseurope.16025.r29558

✓ **Joana Santos**
Université Paris-Saclay, Paris, France

This is a well-written paper describing a rigorous study arising from an observation of the two authors when analysing ChIP-Seq data from a control experiment they conducted on the human malaria parasite *Plasmodium falciparum*. They re-analyse previously generated data plus generate new ChIP-Seq data and conclude that there is a strong bias against the mapping of heterochromatin regions in input samples from parasite chromatin. This is explained by the fact that the genome is very AT-rich, particularly in intergenic regions, and most heterochromatic regions are subtelomeric and highly similar. To counter-act this and prevent errors when analysing ChIP-Seq data, the authors suggest a series of aspects that should be taken into account when analysing data and the appropriate controls to include when doing ChIP-Seq experiments, in particular, and characterising parasite chromatin-associated proteins, in general.

I think this article will improve the quality of the ChIP-Seq data generated by the malaria community and it will be seen as the "bible" of ChIP-Seq data analysis in the field.

I have a few comments:

- In the introduction, the full name of ap2-g should be included to those unfamiliar with it and the way it is written it isn't clear that heterochromatic regions are rare in the parasite.

- This paper basically defines guidelines on how to better analyse ChIP-Seq data. To make these clear, I would add a table with them: input should be from same genomic prep, etc.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: parasitology, malaria, Toxoplasma, genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.