UNIVERSITY OF CALIFORNIA
RIVERSIDE

Exploring Chromatin Organization and Regulation in Human Malaria Parasites

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Cell, Molecular and Developmental Biology

by

Gayani Dinusha Batugedara

December 2018

Dissertation Committee:
Dr. Karine G Le Roch, Chairperson
Dr. Morris Maduro
Dr. Frances Sladek

The Dissertation of Gayani Dinusha Batugedara is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I want to express my most sincere gratitude to my advisor Dr. Karine Le Roch. Karine, thank you for your constant guidance and encouragement during the course of this work. The successful completion of this thesis would not have been possible without your innovative ideas and advice.

I would also like to thank my dissertation committee, Dr. Morris Maduro and Dr. Frances Sladek. Thank you for your continued support and guidance throughout my thesis work.

A most heartfelt thank you to my lab family members, past and present: Jacques Prudhomme, Evelien Bunnik, Maggie Lu, Hailey Choi, Desiree Williams, Steven Abel, Anthony Cort, Tina Wang, Raphael Reyes, Chris Conner, and Mike Lee. Thank you for always being there for me, through the good days and bad. I am lucky to have made such great friends.

Finally, I would like to thank my family and friends. Your love and support has meant the world to me. This journey would not have been possible without you all.

The text of this dissertation (or thesis), in part or in full, is a reprint of the material as it appears in "The Role of Chromatin Structure in Gene Regulation of the Human Malaria Parasite" Trends in Parasitology (2017). Gayani Batugedara and Xueqing Maggie Lu

drafted the manuscript. Karine Le Roch supervised and finalized the manuscript for publication, which forms the basis for this dissertation.

The text of this dissertation (or thesis), in part or in full, is a reprint of the material as it appears in "Unraveling the 3D genome of human malaria parasites" Seminars in Cell and Developmental Biology (2018). Gayani Batugedara drafted the manuscript. Karine Le Roch supervised and finalized the manuscript for publication, which forms the basis for this dissertation.

The text of this dissertation (or thesis), in part or in full, is a reprint of the material as it appears in "Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages" in Nature Communications (2018). Evelien M. Bunnik and Gayani Batugedara carried out the Hi-C and ChIP-seq experiments. Evelien M. Bunnik, Kate B. Cook, Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Karine G. Le Roch wrote the manuscript. William Stafford Noble and Karine G. Le Roch conceived the project, which forms the basis for this publication.

DEDICATION

For Thaththi, Ammi and Nanga.

ABSTRACT OF THE DISSERTATION

Exploring Chromatin Organization and Regulation in Human Malaria Parasites

by

Gayani Dinusha Batugedara

Doctor of Philosophy, Graduate Program in Cell, Molecular and Developmental Biology
University of California, Riverside, December 2018
Dr. Karine Le Roch, Chairperson

The human malaria parasite, one of the deadliest infectious agents in the world, still contributes significantly to the global burden of disease. In 2017, an estimated 214 million cases of infection and over 400,000 malaria-related deaths were reported, a majority of which are caused by the most lethal human malaria parasite, *Plasmodium falciparum*. Given the absence of an FDA-approved vaccine and parasite resistance to all current antimalarial drugs there is a desperate need for new therapeutic approaches.

Plasmodium falciparum has a complex life cycle that requires coordinated gene expression regulation to allow host cell invasion, transmission and immune evasion. However, this cascade of transcripts is unlikely to be regulated by the limited number of identified parasite-specific transcription factors. Increasing evidence now suggests a

major role for epigenetic mechanisms in gene expression in the parasite. Therefore, in this dissertation work, we further explore genome architecture, epigenome, proteome and transcriptome including long-non-coding RNAs (lncRNAs) to better understand the relationship between chromatin structure, genome organization and transcriptional regulation in malaria parasites.

In the first chapter, we explore genome organization in human *Plasmodium* parasite stages including the transmission stages from human to mosquito (gametocytes) and from mosquito to human (sporozoites). Our work demonstrates that genome organization is an important regulator for several parasite-specific gene families involved in pathogenesis and immune evasion, erythrocyte and liver cell invasion, sexual differentiation, and master regulators of gene expression. In the second chapter, we investigated genome organization in five malaria parasites and two related apicomplexan parasites with the goal to identify common features of genome organization and possible connections between genome architecture and pathogenicity. We show that in all malaria parasites, genome organization is dominated by the clustering of *Plasmodium*-specific gene families in 3D space. Our data highlight the importance of spatial genome organization in gene regulation and control of virulence in malaria parasites.

In the subsequent chapters, we aim to identify molecular components, specifically proteins and lncRNAs, that maintain and regulate chromatin structure in the malaria parasite. To investigate parasite proteins and protein complexes maintaining and

regulating nuclear architecture, we undertook comparative genomics analysis using twelve distinct eukaryotic genomes. We identified conserved and apicomplexan parasite-specific chromatin-associated domains (CADs) and proteins (CAPs). We validated two of our candidate proteins including a novel plant-related protein that is functionally analogous to animal nuclear lamina proteins and might have a role in heterochromatin organization. Finally, we also explore the role of lncRNAs in *P. falciparum*. In eukaryotes, lncRNAs have been shown to be pivotal regulators of genome structure and gene expression. To investigate the regulatory roles of lncRNAs in *P. falciparum,* we first explored the intergenic distribution of lncRNA using deep sequencing in nuclear and cytoplasmic subcellular locations. We then validate the subcellular localization and stage-specific expression of several putative lncRNAs at single cell resolution using fluorescence in situ hybridization (FISH) technology. Additionally, we explore the genome-wide occupancy of several candidate nuclear lncRNAs using Chromatin Isolation by RNA Purification followed by deep sequencing (ChIRP-seq) technology. Data analysis revealed that lncRNA occupancy sites within the parasite genome are focal and sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis, erythrocyte remodeling, and regulation of sexual differentiation. Discovery of these proteins and lncRNAs are the starting point for further exploration of mechanisms regulating chromatin structure and genome architecture in these deadly parasites.

Collectively, our data highlight the importance of spatial genome organization as a mechanism of transcriptional regulation in malaria parasites, and our work directly addresses one of the central outstanding questions in *Plasmodium* biology, namely, how a parasite with approximately 6,000 genes manages to control gene expression in a coordinated fashion using a limited number of transcription factors.

TABLE OF CONTENTS

Introduction **1**

References 21

Chapter 1 - Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages **32**

Preface 34

Abstract 35

Introduction 35

Results 39

Discussion 65

Materials and Methods 70

References 79

Supplemental Material 85

Chapter 2 - Comparative 3D Genome Organization in Apicomplexan Parasites **113**

Preface 115

Abstract 116

Introduction 117

Results 121

Discussion 141

Materials and Methods 148

References 153

Supplemental Material 158

Chapter 3 - Comparative Analysis of Chromatin-Associated Proteins in Apicomplexans and Characterization of a Plant-Like Nuclear Lamina Protein in *Plasmodium* 172

Preface 173

Abstract 174

Introduction 174

Results 180

Discussion 202

Materials and Methods 208

References 222

Supplemental Material 232

Chapter 4 - The Role of LncRNAs in Malaria Parasites: Deciphering the Non-Coding Code of Pathogenicity and Sexual Differentiation 234

Preface 235

Abstract 236

Introduction 237

Results 240

Discussion 254

Materials and Methods 258

References 268

Supplemental Material 274

Conclusions **276**

References 283

LIST OF FIGURES

Figure I.1: A graphical representation of the life cycle of *Plasmodium falciparum*. 4

Figure 1.1: Genome organization in *Plasmodium* parasites. 40

Figure 1.2: Changes in interaction of *pfap2* genes and invasion genes with the repressive center. 46

Figure 1.3: Silencing of genes encoding exported proteins in gametocytes through expansion of heterochromatin. 52

Figure 1.4: Formation of superdomains on chr14 in gametocytes. 56

Figure 1.5: Changes in genome organization in salivary gland sporozoites. 58

Figure 1.6: 3D genome structure correlates with gene expression. 64

Supplemental figure 1.1: Microscopy images of the transmission stages analyzed in this study. 85

Supplemental Figure 1.2: Quality measures of Hi-C libraries. 86

Supplemental Figure 1.3: Differences in genome organization between *Plasmodium* species and life cycle stages. 87

Supplemental Figure 1.4: Similarities in genome organization between biological replicates as compared to differences in genome organization between different parasite stages. 89

Supplemental Figure 1.5: Introduced translocations in the *P. vivax* genome. 91

Supplemental Figure 1.6: Misassembly metric for *P. falciparum* and *P. vivax*. 93

Supplemental Figure 1.7: Detection of simulated translocations using the misassembly metric. 95

Supplemental Figure 1.8: DNA-FISH experiments in ring stage parasites. 96

Supplemental Figure 1.9: Interaction of ApiAP2 TF genes and invasion genes with the repressive center. 97

Supplemental Figure 1.10: Restriction site resolution virtual 4C of chromosome 12 in trophozoites. 99

Supplemental Figure 1.11: Quality measures of H3K9me3 ChIP-seq libraries. 100

Supplemental Figure 1.12: H3K9me3 immunofluorescence analysis in IDC and gametocyte stages. 101

Supplementary Figure 1.13: MboI restriction site resolution contact count heatmaps of the region surrounding the location of the domain boundary in chr14. 102

Supplemental Figure 1.14: Confirmation of integrity of chromosome 14 around the domain boundary. 103

Supplemental Figure 1.15: Investigating the role of *pfap2-o3* on chromosome 14. 104

Supplemental Figure 1.16: Expression and sequence analysis of PF3D7_1430100 (PTPA). 105

Supplemental Figure 1.17: Characteristics of genome organization in *P. vivax* salivary gland sporozoites. 106

Supplemental Figure 1.18: Tagging and depletion of PfHP1 result in a loss of *var* gene interactions. 107

Figure 2.1: Overview of samples and protocol. 122

Figure 2.2: Hi-C data and 3D genome modeling. 126

Figure 2.3: Correction of the *T. gondii* genome assembly by Hi-C data. 129

Figure 2.4: Formation of domain-like structures and chromosome loops by *var* and *SICAvar* genes. 135

Figure 2.5: Correlation between genome organization and gene expression. 139

Supplementary Figure 2.1: Correlation between samples of the same organism. 158

Supplementary Figure 2.2: Hi-C data and 3D genome modeling. 159

Supplementary Figure 2.3: Robustness of 3D models to random initializations of PASTIS. 160

Supplementary Figure 2.4: Representative *P. vivax* 3D models from different initializations. 161

Supplementary Figure 2.5: Misassembly metrics for all samples used in this study. 162

Supplementary Figure 2.6: Detection of inversion and duplication events in the *T. gondii* genome. 163

Supplementary Figure 2.7: Differences in centromere interaction patterns. 164

Supplementary Figure 2.8: Formation of domain-like structures and chromosome loops by *SICAvar* genes. 165

Supplementary Figure 2.9: Comparison between Hi-C data obtained from nonsynchronous and synchronous *B. microti* cultures. 166

Supplementary Figure 2.10: Effect of deleting hallmarks of genome organization on the 3D model. 169

Supplementary Figure 2.11: Contact count fold enrichment as a function of genomic distance for all organisms included in this study. 170

Figure 3.1: Relative abundance of chromatin-binding domains in apicomplexan parasites compared to other eukaryotes. 185

Figure 3.2: Overview of chromatin-associated proteins in *Plasmodium falciparum*. 190

Figure 3.3: Chromatin enrichment for proteomics (ChEP). 193

Figure 3.4: Experimental validation of candidate CAPs. 198

Figure 3.5: ChIP-seq analysis showing genome-wide distribution of SMC3 in trophozoites. 201

Supplemental figure 3.1: Comparison of computationally identified candidate CAPs with the ChEP enriched CAPs. 232

Supplemental figure 3.2: Fold enrichment of putative proteins interacting with SMC3 in the parasite. 232

Figure 4.1: Nuclear and cytoplasmic lncRNA identification. 242

Figure 4.2: Candidate lncRNA categorization. 242

Figure 4.3: Gene expression patterns of lncRNAs. 246

Figure 4.4: RNA-FISH experiments show localization of several candidate lncRNAs. 248

Figure 4.5: Chromatin Isolation by RNA Purification (ChIRP). 250

Figure 4.6: ChIRP-seq reveals candidate lncRNA binding sites. 252

Supplemental figure 4.1: RT-PCR validation of selected lncRNAs. 274

Supplemental figure 4.2: Correlation analysis. 275

LIST OF TABLES

Table I.1: Anti-malarial drugs and corresponding *Plasmodium falciparum* drug resistant strain development 2

Table 1.1: Loci involved in long-range intrachromosomal interactions in *P. falciparum* sporozoites. 63

Supplemental Table 1.1: Number of contact counts after mapping and filtering out interactions between loci that are less than 1 kb apart. 108

Supplemental Table 1.2: Interchromosomal contact probability (ICP) and percentage of long- range contacts (PLRC) values for each Hi-C library generated in this study. 109

Supplemental Table 1.5: The sum of Hi-C contacts for significant interactions (5% FDR) between 10 kb bins containing *pfap2* genes and 10 kb bins containing virulence genes. 110

Supplemental Table 1.7: Loci involved in long-range interactions in *P. vivax* sporozoites. 111

Supplemental Table 1.8: Sequences of primers used for the generation of FISH probes. 112

Supplemental Table 1.9: Sequences of primers used to validate that chr14 is physically intact around the domain boundary. 112

Table 2.1: Description of organisms and source material included in this study. 125

Table 2.2: Colocalization of centromeres, telomeres and virulence genes. 133

Supplementary Table 2.1: Numbers of sequence reads generated in Hi-C experiments and valid interaction pairs retained for downstream analyses. 171

INTRODUCTION

Malaria

Malaria remains one of the deadliest infectious diseases worldwide. In 2017, an estimated 200 million cases of infection and over 400,000 malaria-related deaths were reported [1]. Most malaria infections occur in sub-Saharan Africa; however developing countries in South East Asia and South America are also affected. Children under the age of five and pregnant women are most susceptible to the disease and in 2015, children under the age of five accounted for approximately 70% of all malaria-related deaths.

The period between initial infection and appearance of symptoms ranges between 9-40 days, depending on the malarial species. The early symptoms of malaria include fever, chills, headache, sweats, fatigue, nausea and vomiting. Other symptoms, such as dry cough, muscle pain and enlarged spleen could also transpire. Severe cases of malaria could lead to seizures and loss of consciousness as a result of anemia, renal failure and impairment of the respiratory and nervous systems [2]. The intensity of the disease and recovery rate vary depending on the health of the affected individual, as well as the malarial species causing the infection.

Currently, malaria infections are largely treated with the anti-malarial drugs chloroquine, mefloquine, doxycycline, atovaquone, proguanil hydrochloride and artemisinins. However, the highly adaptable nature of the parasite has led to the development of drug

resistant strains. Resistance stains have now been reported for all effective anti-malarial drugs (Table I.1). The use of an artemisinin-based combination therapy (ACT), which are variations of artemesinin derivatives and combined with other anti-malarials such as mefloquine and sulfadoxin, has aided the treatment of drug-resistant infections. However, the rapid development of drug-resistant parasite strains remains of the biggest challenges facing malaria eradication.

Table I.1: Anti-malarial drugs and corresponding *Plasmodium falciparum* drug resistant strain development (Sources: WHO, 2003, Medicines for Malaria Venture, 2015 and O'Brien et al., 2011)

Past popular Anti-malarial Drug	Year introduced to market	Resistance Strain reported	Years in-between
Quinine	1632	1910	278
Chloroquine	1945	1957	12
Sulfadoxin-Pyrimethamin	1967	1967	0
mefloquine	1977	1982	5
Artemisinin	2001	2009	8

Vaccine development against malaria has also proved difficult due to the plasticity of the parasite genome for genes responsible for host-parasite infections. So far, the most advanced vaccine candidate against malaria is called RTS,S/AS01 (Mosquirix, Phase IV) and was developed by the PATH Malaria Vaccine Initiative (MV1). Phase III clinal trials of the vaccine reported the reduction of clinical malaria by 51% after the administration

of three doses [3]. Separate studies have reported the efficacy of RTS,S against malaria to be closer to 34-36% [4]. However, this vaccine has now been approved for a pilot study to vaccinate infants and children in sun-Saharan Africa [ref].

The malaria parasite

Malaria is caused by apicomplexan parasites from the genus *Plasmodium.* Among the five species of *Plasmodium* parasites that can infect humans, *P. falciparum*, *P. vivax, P. ovale, P. malariae* and *P. knowlesi*, *P. falciparum* is the most prevalent (responsible for 99% of the cases in Africa) and most deadly. *P. vivax*, while less virulent than *P. falciparum*, can remain in the body in a dormant state and can cause relapses weeks to months after the initial infection.

All *Plasmodium* species have similar, complex life cycles that involve two hosts: the *Anopheles* mosquito and the human host (Figure I.1). The human infection starts as a female mosquito takes a blood meal and injects sporozoites into the host bloodstream. The sporozoites travel into the liver, invade hepatocytes and replicate extensively to release thousands of merozoites into the blood stream. The merozoites invade red blood cells (RBCs) where they begin a 48-hour replication cycle and develop asexually [5, 6]. During the intra-erythrocytic developmental cycle (IDC) the parasite progresses through three distinct stages termed ring, trophozoite and schizont and multiply into 16-32 daughter cells by a processed known as schizogony. These daughter parasites burst out of the host cell and invade new healthy erythrocytes.

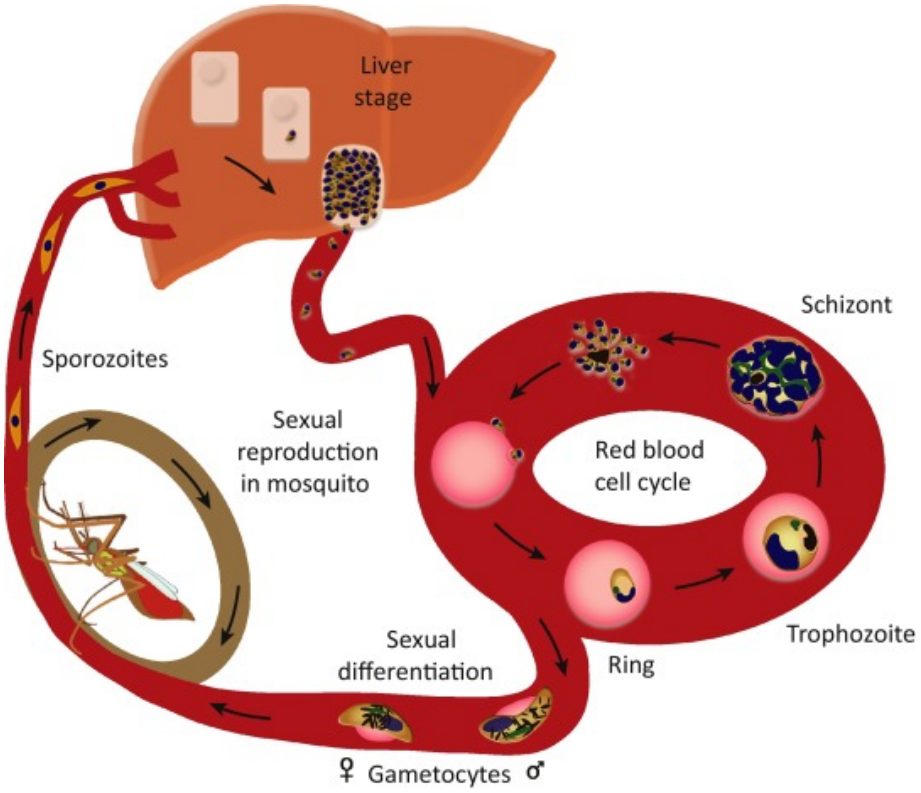


Figure I.1: A graphical representation of the life cycle of *Plasmodium falciparum*.

During the IDC, environmental stress, such as limited nutrient levels, can trigger sexual development of parasites into male and female gametocytes. The mature gametocytes can be ingested by a feeding mosquito, undergo sexual replication in the mosquito midgut and develop further into salivary gland sporozoites that can be transmitted to a new human host.

Understanding this tightly regulated multi-stage life cycle of the parasite still remains an important goal in malaria research, especially since the development of therapeutics that

could halt parasite differentiation or transmission can lead to disruption of the cycle of infection. Developmental stage transitions are regulated by coordinated changes in gene expression; however, the nature and contribution of mechanisms regulating gene expression [7-12], including the role of chromatin structure and three-dimensional (3D) genome organization in the parasite are only now starting to emerge.

The Plasmodium falciparum genome

The human malaria parasite *P. falciparum* has a relatively compact genome of twenty three million base pairs that is organized into fourteen chromosomes (per haploid genome) [13]. The *P. falciparum* genome is the most AT-rich eukaryotic genome sequenced to date, with an overall AT-composition of ~80%, rising to 90-95% in introns and intergenic regions. The distinct developmental stages of the *P. falciparum* life cycle (Figure 1) are characterized by coordinated changes in gene expression [7-12]. In eukaryotes, gene expression is partly controlled by transcription factors that bind to cell- or tissue-specific promoters to regulate transcription [14]. However, a surprisingly low number of specific transcription factors have been identified in the parasite genome [15, 16] and in particular, only a few stage-specific transcription factors have been validated [17-23]. Therefore, the coordinated cascade of transcripts observed throughout the parasite life cycle is unlikely to be regulated only by this limited collection of specific transcription factors, and suggests that additional components and mechanisms, such as post-transcriptional [24-28], translational and post-translational regulation [27, 29, 30] as

well as change of chromatin structure, may control the expression of the predicted 6,372 genes in the malaria parasite.

Transcriptional machinery in Plasmodium

Since the publication of the *P. falciparum* genome in 2002 [13], researchers have attempted to explore the transcriptional machinery of the parasite in detail. The basal transcriptional machinery, RNA polymerase II and all its subunits have been identified in the parasite [16, 31]. Additionally, a total of 23 TFII components have been found. Although four TATA-binding protein (TBP) associated factors (TAFs) have been discovered in *P. falciparum*, the parasite seems to lack the classical TFIID subunits with a histone fold domain. The histone fold domain allows the TAFs to assemble into heterodimers [16]. In yeast, the TFIID complex contains the TATA-binding protein (TBP) and TBP-associated proteins (TAFs) and more than half of these proteins contain a histone fold motif [32]. The fact that the TAFs in *P. falciparum* do not contain this motif suggests that the parasite TFIID complex is divergent from other eukaryotes. As so many TAFs are missing in *P. falciparum* compared to other eukaryotes, alternative mechanisms may be more important for transcriptional regulation in the parasite.

In eukaryotes, specific transcription factors (TFs) recruit and activate the transcription pre-initiation complex. Remarkably, in *P. falciparum*, only about 30 specific transcription factors have been identified [15, 16]. Twenty-seven of these TFs belong to the apicomplexan-specific family of transcription factors and contain a modified form of the

AP2 domain found in plant TFs (ApiAP2) [15]. These TFs are present throughout the parasite life cycle and are believed to control the transition between specific developmental stages. Some examples include AP2-G for the development of gametocytes [19, 20], AP2-Sp for sporozoite development [22], AP2-L for the development of liver-stage parasites [18], and AP2-O for the development of ookinetes in the mosquito [23]. Another member of the ApiAP2 family, PfSIP2, is shown to bind to heterochromatic regions of the genome and act as a transcriptional repressor [33]. Despite continued experimentation, it remains to be determined how ApiAP2 transcription factors recruit RNA polymerase II to sites of transcription. Similar to the lack of TAFs mentioned above, the small number of specific TFs identified in *P. falciparum* highlights the role of alternative mechanisms regulating gene expression.

It is now evident that most of the *P. falciparum* genome is maintained in a decondensed chromatin environment called euchromatin, while only a small subset, including subtelomeric regions and a few internal loci, are contained within highly condensed heterochromatin cluster(s) [34, 35]. The heterochromatin cluster(s) of the parasite genome are marked by H3K9me3 modifications and heterochromatin protein 1 (PfHP1), and harbor gene families encoding clonally variant antigens (*var*, *rifin*, *stevor* and *pfmc-2tm*), invasion gene families (*eba* and *clag*) and a few other loci such as the gametocyte-specific transcription factor *pfap2-g* during the IDC [36-40]. The presence of these repressive marks on parasite stage-specific gene families suggests that mechanisms

regulating transcription of these genes may be more conserved with higher eukaryotes, than the rest of the genes in the *Plasmodium* genome.

P. falciparum histone and nucleosome landscape

The nucleosome landscape of *P. falciparum* is, in some ways, similar to other eukaryotes. First, nucleosome depleted regions are observed in the promoters of genes [41-43], likely to allow for binding of the transcriptional machinery. Second, intergenic regions show lower levels of nucleosome occupancy as compared to coding regions [41, 43, 44]. While some studies suggest that sequencing biases introduced by the high AT-content in the intergenic regions contribute to the differences in nucleosome occupancy [45], alternative methodologies that enrich for nucleosome depleted regions such as FAIRE-Seq [44] or ATAC-Seq [46, 47] validate this initial coverage. Additionally, lower nucleosome occupancy in intergenic regions is also observed in other eukaryotes [48-51], including *Tetrahymena thermophila*, an organism with an AT-rich genome [52]. Finally, several studies have also demonstrated that promoter regions of highly transcribed genes display more open chromatin structure compared to repressed gene regions [41, 42, 45].

While the parasite shares some nucleosome landscape features with other eukaryotes, *P. falciparum* displays several unique characteristics. First, the parasite genome lacks a strongly positioned +1 nucleosome that marks the TSS and instead, harbors strongly positioned nucleosomes at the start and end of coding regions [41, 42]. Next, histone variants H2A.Z and H2B.Z exclusively occupy intergenic regions of the parasite genome

[53, 54], instead of being restricted to active promoters similar to other eukaryotes. These histone variants are also known to bind weakly but more efficiently to AT-rich DNA sequences, and are likely an adaptation by the parasite to allow for nucleosome assembly in AT-rich regions.

Another unusual feature of the nucleosome landscape in *P. falciparum*, which remains an area of debate, is the variability of nucleosome levels during the parasite life cycle [41-44]. At the transcriptionally active trophozoite stage, nucleosome levels drop in a genome-wide fashion, and as the asexual cycle progresses towards the schizont stage, nucleosomes are reassembled and DNA is condensed in preparation for egress and invasion of new RBCs. Other studies suggest a transcription-independent nucleosome positioning that is driven by sequence specificity [45]. The reason for these conflicting observations is likely linked to data normalization and cell cycle progression. When nucleosome occupancy datasets are normalized by number of parasite nuclei at the trophozoite stage, these datasets display nucleosome depletion in a genome-wide fashion. Additionally, change in nucleosome profile throughout the parasite life cycle has been confirmed by alternative methodologies including western blots [55], mass spectrometry [41, 56, 57], MNase-Seq, FAIRE-Seq [44], and ChIP-seq [41]. It is therefore likely that genome-wide nucleosome eviction drives the massive transcriptional event at the trophozoite stage. At the schizont stage, nucleosomes are reassembled and restored to levels detected before transcriptional activation in preparation for re-invasion. At the early and late stages of the parasite erythrocytic cycle, transcriptional regulation is more

likely driven via classical mechanisms such as stage-specific transcription factors and histone PTMs. Recent work provided a mechanism for coordinated regulation of invasion genes by a transcription factor bromodomain protein (PfBDP1) [58]. PfBDP1 enrichment was observed at the transcription start sites of invasion genes at the late schizont stage and was shown to regulate invasion gene expression by binding to acetylated histone H3. Furthermore, conditional PfBDP1 knockdown caused dramatic defects in parasite invasion and growth, confirming the essentiality of this chromatin-associated factor for the coordinated expression of invasion genes in the parasite and indicating the importance of histone modifications in parasite development.

A recent mass spectrometry experiment has indeed identified a total of 232 different histone PTMs during the *P. falciparum* IDC stages, including methylations, acetylations, phosphorylations, ubiquitylations and sumoylations [57]. Many of these PTMs are novel in both *Plasmodium* and other organisms and their exact functions remain to be determined. Interestingly, the majority of the parasite genome remains in a constitutively active state marked by activating histone marks (H3K9ac and H3K4me3), while a small subset including subtelomeric regions and a few internal loci, remain in a condensed heterochromatin environment marked by silencing histone marks (H3K9me3 and H3K36me3) and heterochromatin protein 1 (PfHP1) [34, 35]. These heterochromatin environments harbor gene families encoding clonally variant antigens (*var*, *rifin*, *stevor* and *pfmc2tm*), invasion genes (*eba* and *clag*) and several other loci including *pfap2-g*, a gametocyte-specific transcription factor [36-40]. In mammalian genomes, H3K9ac and

H3K4me3 localization is restricted to active promoters [59-63], while in *P. falciparum* these modifications extend to promoters, 5' coding regions of highly transcribed genes [64, 65], intergenic regions and 'silenced' promoters [40, 64, 66].

Organization of the parasite 3D genome

An important question in *P. falciparum* genome biology is how the structural features of 3D chromatin organization are established, maintained and altered during cell cycle progression and development of the parasite. Similar to complex metazoans, the 3D genome structure of *P. falciparum* plays important roles in gene expression regulation. For decades, numerous microscopic imaging technologies have been the preferred methods for visualizing nuclear architecture in many different organisms [67-69]. Initially, immunofluorescence and FISH experiments were performed to observe global chromatin organization within the parasite nucleus [70, 71]. Earlier FISH experiments revealed that repressed gene regions, subtelomeric and a few internal loci harboring antigenic variation genes, localized to a few clusters around the parasite nucleus [35, 39]. These observations were further confirmed using Hi-C experiments that capture intra- and interchromosomal interactions in a genome-wide manner [75].

Hi-C experiments demonstrated that within the *P. falciparum* nucleus, chromosomes are arranged into folded structures, which are anchored at the centromere with both chromosome arms folding over parallel to each other (Figure 3B) [34]. Much like the 3D chromosome organization in the similarly sized budding and fission yeast, centromeres

and telomeres cluster in opposite regions of the nucleus [72, 73]. Compared to the yeast genome, the parasite nucleus displays an additional level of complexity, mostly as a result of genes involved in antigenic variation (*var* genes) located internally on five out of the fourteen chromosomes [34]. These internal *var* genes colocalize with the subtelomeric regions at the periphery of the nucleus by forming additional loops in the chromosomes. The clonally variant *var* gene expression and clustering of *var* genes in a condensed heterochromatin region is much like the epigenetic signature of the olfactory receptor genes in mice, where all but one are located in a heterochromatin environment enriched in H3K9me3 and H4K20me3 histone marks [74].

Regulation of virulence genes

Disease pathogenesis in malaria is caused by the ability of the parasite to escape the host immune response by expressing variants of antigens on the surface of the infected RBC. The *var* gene family, encoding erythrocyte membrane protein 1 (PfEMP1), remains the best characterized multigene family in the parasite to date [75]. There are approximately 60 *var* genes present in the parasite haploid genome but only one *var* gene is expressed at any given time [76]. This process of allelic exclusion contributes to constant antigenic variation and enables the parasite to evade attacks by the host immune system. Extensive research has been conducted to explore the mechanism of *var* gene regulation *in vitro* and increasing evidence indicates that these genes are regulated at the epigenetic and chromatin structure level. Silent *var* genes are marked by H3K9me3 and PfHP1 and are localized to repressed regions of the genome at the periphery of the nucleus [34, 35, 37-

40, 77, 78]. Several studies highlight proteins that are critical for *var* gene activation and silencing, and disruption of these proteins results in the loss of monoallelic *var* gene expression [39, 79-84]. Together, these results emphasize the relationship between precise nuclear organization and regulation of antigenic variation in the parasite.

As a whole, it appears that the nuclear architecture in *P. falciparum* shares features with both unicellular and multicellular organisms. While unicellular throughout most of its life cycle, the parasite needs to survive in a variety of cell types in both human and mosquito hosts. Most importantly, *P. falciparum* has to evade the host immune system. This strategy helps control expression of variable but distinct proteins at the different life-cycle stages of the parasite. For this reason, the parasite has likely developed a more complex nuclear architecture to regulate cell cycle progression, differentiation and survival compared to other unicellular organisms with a small genome size.

A combination of Hi-C experiments and advanced microscopy methodologies have revealed that parasite nuclear architecture undergoes distinct changes during its life cycle progression. Throughout the asexual, sexual and transmission stages of the parasite, the nucleus and chromatin are drastically remodeled, most likely to allow for the changes in transcriptional activity that takes place during these stages (Figure 3B). First, after invading erythrocytes, *Plasmodium* ring-stage parasites settle and grow slowly inside the RBCs. After 18 to 24 hours, the nucleus expands in size [85], reaching its maximum size and volume at the trophozoite stage, which can easily be observed using microscopy

images of Giemsa stained parasites [34]. Second, the number of nuclear pores increases from 3-7 clustered pores at the ring stage to 12-58 pores evenly distributed around the nucleus at the trophozoite stage [85]. The increased number of nuclear pores suggests the need to facilitate the transcriptionally active trophozoite stage. Finally, along with the increased nuclear volume, the chromatin structure opens up [34] together with nucleosome eviction [41, 44], chromosome intermingling [34] and increased transcriptional activity. As the parasite reaches schizogony, the contents of the nucleus along with the nuclear pores are distributed between the daughter nuclei, nucleosomes are reassembled, chromosome territories are re-established and the chromatin structure recompacts.

Proteins involved in P. falciparum nuclear organization

The structure of the eukaryotic nucleus is organized by a network of proteins, known as nucleoskeleton that anchor the contents of the nucleus to the nuclear envelope and help mediate the movement of chromosomes. It is now apparent that the nucleoskeleton confers shape and functionality to the eukaryotic nucleus [86]. Some of these proteins directly interact with chromatin by binding to DNA, histones and chromatin-remodeling complexes. Not all the proteins forming the metazoan nucleoskeleton are present in all unicellular eukaryotes including *P. falciparum*. Specifically, *P. falciparum* seems to lack lamin proteins that form the nuclear lamina on the inside of the nuclear envelope [87]. In addition, the parasite lacks CTCF proteins that insulate the boundaries of TADs and tether chromatin to the nuclear lamina [88]. However, proteins such as actin, myosin and

kinesin, motor components of the nucleoskeleon, are present in the parasite and are most likely critical for maintaining chromatin structure. They have also recently gained attention as possible drug targets as a small molecule inhibitor against *Kinesin-5* in *P. falciparum* and *P. vivax* showed no cross-reactivity in human cell lines [89].

Similar to other eukaryotes, the parasite centromeres are marked by PfCENH3, a special isoform of histone H3 [90, 91]. PfCENH3 has also been shown to interact with PfCENP-C to form the functional centromeric complex that facilitates kinetochore assembly [90-92]. Furthermore mitotic spindle integrity was lost upon disruption of the dimerization domain of PfCENP-C, suggesting that this protein is essential for chromosome segregation and cell cycle progression much like in other eukaryotes [92]. Several SMCs (structural maintenance of chromosome) critical in the assembly of cohesin and condensin complexes, which are essential for chromosome assembly and segregation in all eukaryotic organisms [93, 94], have also been identified in the *P. falciparum* genome. However, the characterization of these proteins in the parasite has yet to be completed.

Proteins involved in the maintenance of the heterochromatin cluster(s) in the parasite nucleus, have been subjected to extensive research. The most critical structural element of the parasite repressive center identified thus far has been heterochromatin protein-1 (PfHP1), which is structurally and functionally homologous to HP1 in other eukaryotes. PfHP1 specifically binds to the repressive histone mark H3K9me3. Knockdown of this protein disrupts the monoallelic expression of genes involved in antigenic variation, the

var genes, known to cluster together in heterochromatin environments [95]. SPE2-interacting protein (PfSIP2), a member of the ApiAP2 transcription factor family in *P. falciparum*, has also been shown to be involved in organization of subtelomeric heterochromatin [33]. PfSIP2 interacts with SPE2 DNA motifs present on telomeric and subtelomeric regions. DNA/RNA binding proteins PfAlba1-4 (Acetylation lowers binding affinity) family of proteins have also been implicated in heterochromatin maintenance. PfAlba 1,2 and 4 bind to TARE6 (telomere-associated repetitive element 6) [96] and PfAlba3 has been shown to bind to telomeric and subtelomeric regions as well as *var* gene promoters [97]. PfSIR2A, a histone deacetylase (HDAC) also present in the repressive center, is thought to increase DNA-binding affinity by deacetylating lysine residues in its N-terminus via interaction with PfAlba3 [97].

It is now increasingly evident that *P. falciparum* nuclear organization and chromosome dynamics are quite complex, and it is likely that many other proteins are regulating these processes. Identification and characterization of these molecular components will likely contribute to a better understanding of the molecular machinery regulating chromatin structure, gene regulation and parasite development.

Long non-coding RNAs

Emerging evidence confirms that non-protein coding transcripts, long non-coding RNAs (lncRNAs), also play a role in transcriptional regulation and 3D genome activity by affecting chromatin-remodeling events such as chromatin looping and nucleosome

positioning [98]. A well-studied lncRNA, Xist, mediates X-chromosome inactivation during zygotic development in placental mammals [99]. The expression of the Xist lncRNA on the X chromosome recruits histone-modifying enzymes that place repressive histone marks, such as H3K9 and H3K27 methylation at the Xist locus, leading to gene silencing and heterochromatin formation.

In *Plasmodium*, a variety of long non-coding RNAs (lncRNAs) transcribed from telomere-associated repetitive elements (TAREs) have been identified [100-102]. These TARE-lncRNAs reach their highest expression levels at the schizont stage. While the exact role of these lncRNAs has yet to be determined, it is likely that these transcripts are part of a network regulating and maintaining the heterochromatin environment within the parasite nucleus. Interestingly, the lncRNA TARE regions harbor ApiAP2 transcription factor PfSIP2 binding sites. Since PfSIP2 has been implicated in heterochromatin formation around subtelomeric regions, these TARE-lncRNAs may play a role in regulating *var* gene expression. The TARE-lncRNAs in *P. falciparum* are functionally similar to the eukaryotic family of non-coding RNA called telomeric repeat-containing RNA (TERRA) important for telomere maintenance and heterochromatin assembly [103], which further validates the importance of lncRNA-TAREs for regulating the repressive centers within the parasite nucleus.

The monoallelic *var* gene expression can also be regulated via transcription of lncRNAs. Two lncRNAs are transcribed from a bidirectional promoter within the intron of *var*

genes [104]. These transcripts are incorporated into chromatin after being capped but not polyadenylated. It is likely that the sense lncRNA functions to silence *var* gene expression, while the antisense lncRNA is associated with the single active *var* gene [104]. Thus, the presence of these lncRNAs adds an additional layer of complexity to epigenetic mechanisms regulating *var* gene expression.

Conclusions

An increasing amount of evidence emphasizes the importance of the epigenetic landscape and nuclear architecture in regulating gene expression in *P. falciparum* and higher eukaryotes such as human and mouse. Here we discuss the local and global genome architecture of *P. falciparum* including similarities with other eukaryotic organisms and differences that contribute to the unique biology of the parasite. As outlined above, the *Plasmodium* 3D nuclear architecture points toward a binary structure where a majority of the genome is maintained in a transcriptionally permissive euchromatin state and a small subset of genes are harbored within a transcriptionally repressed heterochromatin state. The asexual cycle of the parasite reflects large changes in genome organization that are characterized by nucleosome landscape and global histone levels. The overall genome organization in gametocytes is similar to IDC stages with a few exceptions including the localization of gametocyte-specific transcription factor, *pfap2-g*, erythrocytic remodeling and invasion genes. These results suggest that transcriptional regulation in the parasite is controlled at different layers and these layers shape the overall organization of the nucleus. However, our understanding of the parasite nucleome is far from complete. It is

important to understand the extent to which nuclear reorganization controls gene expression in *P. falciparum*. In particular, molecular components, such as proteins and lncRNAs, that are likely involved in regulating genome architecture in the parasite could serve as potential drug targets that can disrupt parasite development with high specificity and low toxicity to the host.

In the upcoming chapters of this thesis, I investigate the 3D nuclear organization of five *Plasmodium* species, two related apicomplexan parasites and the transmission stages of *P. falciparum* and *P. vivax* in order to identify possible connections between genome architecture and pathogenicity. Collectively, the data show that genome organization was dominated by the clustering of *Plasmodium*-specific gene families in 3D space. In particular, the two most pathogenic human malaria parasites shared unique features in the organization of gene families involved in antigenic variation and immune escape. Given the importance of nuclear organization and chromatin structure in parasite development, I then explore the molecular components that maintain and regulate parasite nuclear architecture. In particular, I developed complementary experimental approaches to identify and characterize proteins and lncRNAs that regulate the 3D genome architecture of the malaria agent *P. falciparum*. While significant progress has been made in elucidating the role of chromatin in parasite development and transcriptional control [38, 44, 80, 81, 84, 105-108], the molecular factors and the mechanisms underlying changes in chromatin structure and nuclear organization are not well understood. The collective explorations presented here will address this knowledge gap by uncovering key proteins

and lncRNAs involved in regulating the interaction of chromatin elements with genes involved in virulence and sexual differentiation. Identification and functional characterization of the specific factors that control these processes will facilitate development of drugs that can target chromatin structure to block parasite survival and disease transmission.

References

1. WHO: **The World Malaria Report.** http://www.who.int/malaria/publications/worldmalaria-report-2017/en/**.** 2017.

2. Sadanand S: **Malaria: an evaluation of the current state of research on pathogenesis and antimalarial drugs.** *Yale J Biol Med* 2010, **83:**185-191.

3. Rts SCTP: **Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial.** *Lancet* 2015, **386:**31-45.

4. Alonso PL, Sacarlal J, Aponte JJ, Leach A, Macete E, Milman J, Mandomando I, Spiessens B, Guinovart C, Espasa M, et al: **Efficacy of the RTS,S/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial.** *Lancet* 2004, **364:**1411-1420.

5. Rosenberg R, Wirtz RA, Schneider I, Burge R: **An estimation of the number of malaria sporozoites ejected by a feeding mosquito.** *Trans R Soc Trop Med Hyg* 1990, **84:**209-212.

6. Yuda M, Ishino T: **Liver invasion by malarial parasites--how do malarial parasites break through the host barrier?** *Cell Microbiol* 2004, **6:**1119-1125.

7. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1:**E5.

8. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, Prudhomme J, Florens L, Le Roch KG: **Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum.** *Genome Biol* 2013, **14:**R128.

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

10. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, Cui K, Barillas-Mury C, Zhao K, Su XZ: **Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum.** *BMC Genomics* 2011, **12:**587.

11. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Berriman M, et al: **New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq.** *Mol Microbiol* 2010, **76:**12-24.

12. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, Ribas de Pouplana L, Preiser PR, Bozdech Z, Cortes A: **Transcriptional variation in the malaria parasite Plasmodium falciparum.** *Genome Res* 2012, **22:**925-938.

13. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419:**498-511.

14. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.

15. Balaji S, Babu MM, Iyer LM, Aravind L: **Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains.** *Nucleic Acids Res* 2005, **33:**3994-4006.

16. Coulson RM, Hall N, Ouzounis CA: **Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum.** *Genome Res* 2004, **14:**1548-1554.

17. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M: **Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite.** *PLoS Pathog* 2010, **6:**e1001165.

18. Iwanaga S, Kaneko I, Kato T, Yuda M: **Identification of an AP2-family protein that is critical for malaria liver stage development.** *PLoS One* 2012, **7:**e47557.

19. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, Williams AE, Drought LG, Kwiatkowski DP, Baker DA, et al: **A transcriptional switch underlies commitment to sexual development in malaria parasites.** *Nature* 2014, **507:**248-252.

20. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, Religa AA, Bushell E, Graham AL, Cameron R, et al: **A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium.** *Nature* 2014, **507:**253-257.

21. Young JA, Johnson JR, Benner C, Yan SF, Chen K, Le Roch KG, Zhou Y, Winzeler EA: **In silico discovery of transcription regulatory elements in Plasmodium falciparum.** *BMC Genomics* 2008, **9:**70.

22. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I: **Transcription factor AP2-Sp and its target genes in malarial sporozoites.** *Mol Microbiol* 2010, **75:**854-863.

23. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, Waters AP, Kato T, Kaneko I: **Identification of a transcription factor in the mosquito-invasive stage of malaria parasites.** *Mol Microbiol* 2009, **71:**1402-1414.

24. Balu B, Maher SP, Pance A, Chauhan C, Naumov AV, Andrews RM, Ellis PD, Khan SM, Lin JW, Janse CJ, et al: **CCR4-associated factor 1 coordinates the expression of Plasmodium falciparum egress and invasion proteins.** *Eukaryot Cell* 2011, **10:**1257-1263.

25. Bunnik EM, Batugedara G, Saraf A, Prudhomme J, Florens L, Le Roch KG: **The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum.** *Genome Biol* 2016, **17:**147.

26. Eshar S, Altenhofen L, Rabner A, Ross P, Fastman Y, Mandel-Gutfreund Y, Karni R, Llinas M, Dzikowski R: **PfSR1 controls alternative splicing and steady-state RNA levels in Plasmodium falciparum through preferential recognition of specific RNA motifs.** *Mol Microbiol* 2015, **96:**1283-1297.

27. Kirchner S, Power BJ, Waters AP: **Recent advances in malaria genomics and epigenomics.** *Genome Med* 2016, **8:**92.

28. Vembar SS, Macpherson CR, Sismeiro O, Coppee JY, Scherf A: **The PfAlba1 RNA-binding protein is an important regulator of translational timing in Plasmodium falciparum blood stages.** *Genome Biol* 2015, **16:**212.

29. Caro F, Ahyong V, Betegon M, DeRisi JL: **Genome-wide regulatory dynamics of translation in the Plasmodium falciparum asexual blood stages.** *Elife* 2014, **3**.

30. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z: **Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites.** *Genome Biol* 2008, **9:**R177.

31. Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S: **Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium**

falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics* 2005, **6:**100.

32. Gangloff YG, Romier C, Thuault S, Werten S, Davidson I: **The histone fold is a key structural motif of transcription factor TFIID.** *Trends Biochem Sci* 2001, **26:**250-257.

33. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BT, Moes S, Bozdech Z, Jenoe P, Stunnenberg HG, Voss TS: **A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology.** *PLoS Pathog* 2010, **6:**e1000784.

34. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: **Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression.** *Genome Res* 2014, **24:**974-988.

35. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A: **Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum.** *Nature* 2000, **407:**1018-1022.

36. Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ, Hartl DL, Deitsch KW: **Epigenetic memory at malaria virulence genes.** *Proc Natl Acad Sci U S A* 2007, **104:**899-902.

37. Crowley VM, Rovira-Graells N, Ribas de Pouplana L, Cortes A: **Heterochromatin formation in bistable chromatin domains controls the epigenetic repression of clonally variant Plasmodium falciparum genes linked to erythrocyte invasion.** *Mol Microbiol* 2011, **80:**391-406.

38. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, Rojas-Meza AP, Mancio-Silva L, Leal-Silvestre RJ, Gontijo AM, Shorte S, Scherf A: **Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites.** *Cell* 2005, **121:**25-36.

39. Lopez-Rubio JJ, Mancio-Silva L, Scherf A: **Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites.** *Cell Host Microbe* 2009, **5:**179-190.

40. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AM, Cohen AM, Janssen-Megens EM, van de Vegte-Bolmer M, Selzer RR, Iniguez AL, et al: **Dynamic histone H3 epigenome marking during the**

intraerythrocytic cycle of Plasmodium falciparum. *Proc Natl Acad Sci U S A* 2009, **106:**9655-9660.

41. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, Le Roch KG: **DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum.** *BMC Genomics* 2014, **15:**347.

42. Ponts N, Harris EY, Lonardi S, Le Roch KG: **Nucleosome occupancy at transcription start sites in the human malaria parasite: a hard-wired evolution of virulence?** *Infect Genet Evol* 2011, **11:**716-724.

43. Westenberger SJ, Cui L, Dharia N, Winzeler E, Cui L: **Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes.** *BMC Genomics* 2009, **10:**610.

44. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG: **Nucleosome landscape and control of transcription in the human malaria parasite.** *Genome Res* 2010, **20:**228-238.

45. Kensche PR, Hoeijmakers WA, Toenhake CG, Bras M, Chappell L, Berriman M, Bartfai R: **The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences.** *Nucleic Acids Res* 2016, **44:**2110-2124.

46. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ: **Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions.** *Genome Res* 2015, **25:**1757-1770.

47. Toenhake CG, Fraschka SA, Vijayabaskar MS, Westhead DR, van Heeringen SJ, Bartfai R: **Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development.** *Cell Host Microbe* 2018, **23:**557-569 e559.

48. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD: **Evidence for nucleosome depletion at active regulatory regions genome-wide.** *Nat Genet* 2004, **36:**900-905.

49. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, et al: **Nucleosome organization in the Drosophila genome.** *Nature* 2008, **453:**358-362.

50. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, et al: **Genome-wide map of nucleosome acetylation and methylation in yeast.** *Cell* 2005, **122:**517-527.

51. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18:**1051-1063.

52. Beh LY, Muller MM, Muir TW, Kaplan N, Landweber LF: **DNA-guided establishment of nucleosome patterns within coding regions of a eukaryotic genome.** *Genome Res* 2015, **25:**1727-1738.

53. Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Francoijs KJ, Treeck M, Gilberger TW, Stunnenberg HG, Bartfai R: **H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the Plasmodium falciparum genome.** *Mol Microbiol* 2013, **87:**1061-1073.

54. Petter M, Selvarajah SA, Lee CC, Chin WH, Gupta AP, Bozdech Z, Brown GV, Duffy MF: **H2A.Z and H2B.Z double-variant nucleosomes define intergenic regions and dynamically occupy var gene promoters in the malaria parasite Plasmodium falciparum.** *Mol Microbiol* 2013, **87:**1167-1182.

55. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, et al: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.** *Genome Res* 2004, **14:**2308-2318.

56. Oehring SC, Woodcroft BJ, Moes S, Wetzel J, Dietz O, Pulfer A, Dekiwadia C, Maeser P, Flueck C, Witmer K, et al: **Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum.** *Genome Biol* 2012, **13:**R108.

57. Saraf A, Cervantes S, Bunnik EM, Ponts N, Sardiu ME, Chung DW, Prudhomme J, Varberg JM, Wen Z, Washburn MP, et al: **Dynamic and Combinatorial Landscape of Histone Modifications during the Intraerythrocytic Developmental Cycle of the Malaria Parasite.** *J Proteome Res* 2016, **15:**2787-2801.

58. Josling GA, Petter M, Oehring SC, Gupta AP, Dietz O, Wilson DW, Schubert T, Langst G, Gilson PR, Crabb BS, et al: **A Plasmodium Falciparum Bromodomain Protein Regulates Invasion Gene Expression.** *Cell Host Microbe* 2015, **17:**741-751.

59. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-837.

60. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR, et al: **Genomic maps and comparative analysis of histone modifications in human and mouse.** *Cell* 2005, **120:**169-181.

61. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: **A high-resolution map of active promoters in the human genome.** *Nature* 2005, **436:**876-880.

62. Nishida H, Suzuki T, Kondo S, Miura H, Fujimura Y, Hayashizaki Y: **Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell.** *Chromosome Res* 2006, **14:**203-211.

63. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40:**897-903.

64. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger TW, Francoijs KJ, Stunnenberg HG: **H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3.** *PLoS Pathog* 2010, **6:**e1001223.

65. Cui L, Miao J, Furuya T, Li X, Su XZ, Cui L: **PfGCN5-mediated histone H3 acetylation plays a key role in gene expression in Plasmodium falciparum.** *Eukaryot Cell* 2007, **6:**1219-1227.

66. Trelle MB, Salcedo-Amaya AM, Cohen AM, Stunnenberg HG, Jensen ON: **Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite Plasmodium falciparum.** *J Proteome Res* 2009, **8:**3439-3450.

67. Cremer T, Kreth G, Koester H, Fink RH, Heintzmann R, Cremer M, Solovei I, Zink D, Cremer C: **Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture.** *Crit Rev Eukaryot Gene Expr* 2000, **10:**179-212.

68. Misteli T: **Beyond the sequence: cellular organization of genome function.** *Cell* 2007, **128:**787-800.

69. Takizawa T, Meaburn KJ, Misteli T: **The meaning of gene positioning.** *Cell* 2008, **135:**9-13.

70. Horrocks P, Wong E, Russell K, Emes RD: **Control of gene expression in Plasmodium falciparum - ten years on.** *Mol Biochem Parasitol* 2009, **164:**9-25.

71. Segal E, Widom J: **Poly(dA:dT) tracts: major determinants of nucleosome organization.** *Curr Opin Struct Biol* 2009, **19:**65-71.

72. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465:**363-367.

73. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K: **Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation.** *Nucleic Acids Res* 2010, **38:**8164-8177.

74. Magklara A, Yen A, Colquitt BM, Clowney EJ, Allen W, Markenscoff-Papadimitriou E, Evans ZA, Kheradpour P, Mountoufaris G, Carey C, et al: **An epigenetic signature for monoallelic olfactory receptor expression.** *Cell* 2011, **145:**555-570.

75. Miller LH, Good MF, Milon G: **Malaria pathogenesis.** *Science* 1994, **264:**1878-1883.

76. Scherf A, Lopez-Rubio JJ, Riviere L: **Antigenic variation in Plasmodium falciparum.** *Annu Rev Microbiol* 2008, **62:**445-470.

77. Flueck C, Bartfai R, Volz J, Niederwieser I, Salcedo-Amaya AM, Alako BT, Ehlgen F, Ralph SA, Cowman AF, Bozdech Z, et al: **Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors.** *PLoS Pathog* 2009, **5:**e1000569.

78. Perez-Toledo K, Rojas-Meza AP, Mancio-Silva L, Hernandez-Cuevas NA, Delgadillo DM, Vargas M, Martinez-Calvillo S, Scherf A, Hernandez-Rivas R: **Plasmodium falciparum heterochromatin protein 1 binds to tri-methylated histone 3 lysine 9 and is linked to mutually exclusive expression of var genes.** *Nucleic Acids Res* 2009, **37:**2596-2606.

79. Coleman BI, Skillman KM, Jiang RH, Childs LM, Altenhofen LM, Ganter M, Leung Y, Goldowitz I, Kafsack BF, Marti M, et al: **A Plasmodium falciparum**

histone deacetylase regulates antigenic variation and gametocyte conversion. *Cell Host Microbe* 2014, **16:**177-186.

80. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, Thompson JK, Freitas-Junior LH, Scherf A, Crabb BS, Cowman AF: **Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum.** *Cell* 2005, **121:**13-24.

81. Tonkin CJ, Carret CK, Duraisingh MT, Voss TS, Ralph SA, Hommel M, Duffy MF, Silva LM, Scherf A, Ivens A, et al: **Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum.** *PLoS Biol* 2009, **7:**e84.

82. Jiang L, Mu J, Zhang Q, Ni T, Srinivasan P, Rayavara K, Yang W, Turner L, Lavstsen T, Theander TG, et al: **PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum.** *Nature* 2013, **499:**223-227.

83. Ukaegbu UE, Kishore SP, Kwiatkowski DL, Pandarinath C, Dahan-Pasternak N, Dzikowski R, Deitsch KW: **Recruitment of PfSET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in P. falciparum.** *PLoS Pathog* 2014, **10:**e1003854.

84. Volz JC, Bartfai R, Petter M, Langer C, Josling GA, Tsuboi T, Schwach F, Baum J, Rayner JC, Stunnenberg HG, et al: **PfSET10, a Plasmodium falciparum methyltransferase, maintains the active var gene in a poised state during parasite division.** *Cell Host Microbe* 2012, **11:**7-18.

85. Weiner A, Dahan-Pasternak N, Shimoni E, Shinder V, von Huth P, Elbaum M, Dzikowski R: **3D nuclear architecture reveals coupled cell cycle dynamics of chromatin and nuclear pores in the malaria parasite Plasmodium falciparum.** *Cell Microbiol* 2011, **13:**967-977.

86. Simon DN, Wilson KL: **The nucleoskeleton as a genome-associated dynamic 'network of networks'.** *Nat Rev Mol Cell Biol* 2011, **12:**695-708.

87. Batsios P, Peter T, Baumann O, Stick R, Meyer I, Graf R: **A lamin in lower eukaryotes?** *Nucleus* 2012, **3:**237-243.

88. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T: **The chromatin insulator CTCF and the emergence of metazoan diversity.** *Proc Natl Acad Sci U S A* 2012, **109:**17507-17512.

89. Liu L, Richard J, Kim S, Wojcik EJ: **Small molecule screen for candidate antimalarials targeting Plasmodium Kinesin-5.** *J Biol Chem* 2014, **289:**16601-16614.

90. Hoeijmakers WA, Flueck C, Francoijs KJ, Smits AH, Wetzel J, Volz JC, Cowman AF, Voss T, Stunnenberg HG, Bartfai R: **Plasmodium falciparum centromeres display a unique epigenetic makeup and cluster prior to and during schizogony.** *Cell Microbiol* 2012, **14:**1391-1401.

91. Verma G, Surolia N: **Plasmodium falciparum CENH3 is able to functionally complement Cse4p and its, C-terminus is essential for centromere function.** *Mol Biochem Parasitol* 2013, **192:**21-29.

92. Verma G, Surolia N: **The dimerization domain of PfCENP-C is required for its functions as a centromere protein in human malaria parasite Plasmodium falciparum.** *Malar J* 2014, **13:**475.

93. Freeman L, Aragon-Alcaide L, Strunnikov A: **The condensin complex governs chromosome condensation and mitotic transmission of rDNA.** *J Cell Biol* 2000, **149:**811-824.

94. Strunnikov AV, Jessberger R: **Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions.** *Eur J Biochem* 1999, **263:**6-13.

95. Brancucci NM, Bertschi NL, Zhu L, Niederwieser I, Chin WH, Wampfler R, Freymond C, Rottmann M, Felger I, Bozdech Z, Voss TS: **Heterochromatin protein 1 secures survival and transmission of malaria parasites.** *Cell Host Microbe* 2014, **16:**165-176.

96. Chene A, Vembar SS, Riviere L, Lopez-Rubio JJ, Claes A, Siegel TN, Sakamoto H, Scheidig-Benatar C, Hernandez-Rivas R, Scherf A: **PfAlbas constitute a new eukaryotic DNA/RNA-binding protein family in malaria parasites.** *Nucleic Acids Res* 2012, **40:**3066-3077.

97. Goyal M, Alam A, Iqbal MS, Dey S, Bindu S, Pal C, Banerjee A, Chakrabarti S, Bandyopadhyay U: **Identification and molecular characterization of an Alba-family protein from human malaria parasite Plasmodium falciparum.** *Nucleic Acids Res* 2012, **40:**1174-1190.

98. Bohmdorfer G, Wierzbicki AT: **Control of Chromatin Structure by Long Noncoding RNA.** *Trends Cell Biol* 2015, **25:**623-632.

99. Maclary E, Hinten M, Harris C, Kalantry S: **Long nonoding RNAs in the X-inactivation center.** *Chromosome Res* 2013, **21:**601-614.

100. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, Ribacke U, Volkman S, Duraisingh M, Wirth D, Sabeti PC, Rinn JL: **A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs.** *Genome Biol* 2011, **12:**R56.

101. Raabe CA, Sanchez CP, Randau G, Robeck T, Skryabin BV, Chinni SV, Kube M, Reinhardt R, Ng GH, Manickam R, et al: **A global view of the nonprotein-coding transcriptome in Plasmodium falciparum.** *Nucleic Acids Res* 2010, **38:**608-617.

102. Sierra-Miranda M, Delgadillo DM, Mancio-Silva L, Vargas M, Villegas-Sepulveda N, Martinez-Calvillo S, Scherf A, Hernandez-Rivas R: **Two long non-coding RNAs generated from subtelomeric regions accumulate in a novel perinuclear compartment in Plasmodium falciparum.** *Mol Biochem Parasitol* 2012, **185:**36-47.

103. Luke B, Lingner J: **TERRA: telomeric repeat-containing RNA.** *EMBO J* 2009, **28:**2503-2510.

104. Amit-Avraham I, Pozner G, Eshar S, Fastman Y, Kolevzon N, Yavin E, Dzikowski R: **Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 2015, **112:**E982-991.

105. Dzikowski R, Deitsch KW: **Active transcription is required for maintenance of epigenetic memory in the malaria parasite Plasmodium falciparum.** *J Mol Biol* 2008, **382:**288-297.

106. Dzikowski R, Li F, Amulic B, Eisberg A, Frank M, Patel S, Wellems TE, Deitsch KW: **Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites.** *EMBO Rep* 2007, **8:**959-965.

107. Frank M, Dzikowski R, Amulic B, Deitsch K: **Variable switching rates of malaria virulence genes are associated with chromosomal position.** *Mol Microbiol* 2007, **64:**1486-1498.

108. Merrick CJ, Dzikowski R, Imamura H, Chuang J, Deitsch K, Duraisingh MT: **The effect of Plasmodium falciparum Sir2a histone deacetylase on clonal and longitudinal variation in expression of the var family of virulence genes.** *Int J Parasitol* 2010, **40:**35-43.

CHAPTER 1: Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages

Evelien M. Bunnik[1,2$], Kate B. Cook[3$], Nelle Varoquaux[4,5,6,7,8], **Gayani Batugedara**[2], Jacques Prudhomme[2], Anthony Cort[2], Lirong Shi[9], Chiara Andolina[10,11], Leila S. Ross[12], Declan Brady[13], David A. Fidock[12,14], Francois Nosten[10,11], Rita Tewari[13], Photini Sinnis[9], Ferhat Ay[15], Jean-Philippe Vert[6,7,8,16], William Stafford Noble[3,17]*, Karine G. Le Roch[2]*

[1] Department of Microbiology, Immunology & Molecular Genetics, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

[2] Department of Cell Biology and Neuroscience, University of California Riverside, Riverside, CA 92521, USA

[3] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[4] Department of Statistics, University of California, Berkeley, CA 94720, USA

[5] Berkeley Institute for Data Science, Berkeley, CA 94720, USA

[6] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Fontainebleau, F-77300, France

[7] Institut Curie, Paris, F-75248, France

[8] U900, INSERM, Paris, F-75248, France

[9] Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[10] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine Research building, University of Oxford Old Road campus, Oxford, OX3 7FZ, UK

[11] Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Tak 63110, Thailand

[12] Department of Microbiology and Immunology, Columbia University Medical Center, New York, NY 10032, USA

[13] School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham, NG7 2UH, UK

[14] Division of Infectious Diseases, Department of Medicine, Columbia University, New York, NY 10032, USA

[15] La Jolla Institute for Allergy & Immunology, La Jolla, CA 92037, USA

[16] Département de mathématiques et applications, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France

[17] Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA

[$] These authors contributed equally to this study

A version of this chapter was published in *Nature Communications*, 2018.

Preface

An increasing amount of evidence now suggests that gene regulation in malaria parasites is largely regulated by epigenetic mechanisms and nuclear reorganization. *Plasmodium falciparum* and *Plasmodium vivax* are the two most prevalent malaria parasites and are both significant contributors to the global burden of disease. Understanding the biology of the transmission stages of these parasites will be extremely important for vaccine development and for the discovery of novel antimalarials that will help to break the continuous cycle of infections. Here, to further our understanding of genome organization in *Plasmodium* we performed chromosome conformation capture coupled with next generation sequencing (Hi-C) in the parasite transmission stages from human to mosquito (gametocytes) and from mosquito to human (sporozoites). As these stages are in general more challenging to study than the intraerythrocytic stages, our knowledge of their biology is limited. Our results describe many previously undocumented changes in genome organization during the parasite life cycle in connection with activation and silencing of parasite-specific gene families. Though I am not the primary contributor to the body of work presented in this chapter, the major discoveries from this work highlight the link between genome organization and gene regulation in malaria parasites and open up new avenues for targeted approaches to block parasite replication and transmission. I specifically performed ChIP-seq and immunofluorescence experiments, which contributed to figures 1.3, supplemental figures 1.11, 1.12 and 1.14.

Abstract

The development of malaria parasites throughout their various life cycle stages is coordinated by changes in gene expression. We previously showed that the three-dimensional organization of the *Plasmodium falciparum* genome is strongly associated with gene expression during its replication cycle inside red blood cells. Here, we analyze genome organization in the *P. falciparum* and *P. vivax* transmission stages. Major changes occur in the localization and interactions of genes involved in pathogenesis and immune evasion, host cell invasion, sexual differentiation, and master regulation of gene expression. Furthermore, we observe reorganization of subtelomeric heterochromatin around genes involved in host cell remodeling. Depletion of heterochromatin protein 1 (PfHP1) resulted in loss of interactions between virulence genes, confirming that PfHP1 is essential for maintenance of the repressive center. Our results suggest that the three-dimensional genome structure of human malaria parasites is strongly connected with transcriptional activity of specific gene families throughout the life cycle.

Introduction

With an estimated 438,000 deaths per year, malaria is still one of the most deadly infectious diseases, mostly targeting young children in sub-Saharan Africa1. The disease is caused by one of five parasites of the *Plasmodium* genus, of which *P. falciparum* is the most common and deadliest. *P. vivax* is also responsible for significant disease, mostly in Southeast Asia [1].

Plasmodium parasites have complex life cycles that involve a human host and a mosquito vector. Infection in humans starts when an infected female *Anopheles* mosquito takes a blood meal and transmits parasites that are present in the form of sporozoites in her salivary glands. These sporozoites are inoculated into the skin, travel to the liver and establish an infection in hepatocytes. Over a period of several days, the parasite replicates and eventually releases thousands of merozoites into the bloodstream. Alternatively, *P. vivax* can survive in the liver for weeks or years in dormant forms called hypnozoites, which can be reactivated and cause malaria relapses. Merozoites that emerge from the liver start a 48-h replication cycle in red blood cells. During this asexual intraerythrocytic development cycle (IDC), the parasite progresses through three main developmental stages: ring, trophozoite, and schizont, to produce 8–24 daughter parasites, which burst from the cell and invade new erythrocytes. During the IDC, the parasite can commit to differentiation into male and female gametocytes, which can be taken up by another mosquito. Inside the mosquito, the parasite undergoes sexual reproduction and further develops through several stages into the salivary gland sporozoites that can be transmitted to a new human host.

Understanding how the transitions between the various life cycle stages of the *Plasmodium* parasite are regulated remains an important goal in malaria research. Stage transitions are regulated by coordinated changes in gene expression, but it is still largely unknown how these changes in transcriptional profiles are controlled at the transcriptional level. Only a single family of ApiAP2 transcription factors (TFs) with

27 members has been identified, while approximately two-thirds of the TFs expected based on the size of the *Plasmodium* genome seem to be missing [2]. Several of these ApiAP2 TFs are involved in stage transitions, such as PfAP2-G, which is thought to be the main driver for gametocyte differentiation [3-5]. Our understanding of how these TFs are controlled and how various TFs may act together to form transcriptional networks is still very limited.

In recent years, considerable insight has been gained into the role of epigenetics, chromatin structure, and genome organization in gene regulation, mostly during the IDC of *P. falciparum*. Studies show that the parasite genome is largely in an active, euchromatic state [6-8], while members of several parasite-specific gene families involved in virulence (*var*, *rifin*, *stevor*, and *pfmc-2tm*), erythrocyte remodeling (*phist*, *hyp*, *fikk*, and others) and solute transport (*clag3)* are organized into heterochromatin [7, 9-11]. In particular, the family of *var* genes has received much attention, since these genes are key to pathogenesis and immune escape. Out of a total of 60 *var* genes, only a single variant is expressed within an individual parasite, while the other 59 genes are tightly repressed by a combination of isolation into a perinuclear compartment, repressive histone marks, and repressive long non-coding RNAs [7, 12-15].

Previously, we assessed genome organization at the ring, trophozoite, and schizont stages of the IDC in *P. falciparum* using Hi-C experiments (chromosome conformation

capture coupled with next-generation sequencing) [16] and compared our findings to an earlier Hi-C study in ring-stage *P. falciparum* [17]. We observed a strong association between genome architecture and gene expression, suggesting that the three-dimensional organization of the genome is very important for gene regulation. Here, we analyze the genome organization in the transmission stages of the *P. falciparum* life cycle (gametocytes and sporozoites), as well as for *P. vivax* sporozoites, and present a comparative analysis of genome organization throughout the different life cycle stages. Finally, to meaningfully compare changes in genome organization throughout the *Plasmodium* life cycle, we developed a novel statistical test to detect loci that differ significantly in their intrachromosomal contact count numbers between stages. While large-scale features of chromosome organization are preserved in the sexually differentiated stages, we observe several stage-specific specific changes, including reorganization of genes encoding rDNA, invasion proteins, and transcription factors. During gametocytogenesis, heterochromatic regions at the ends of chromosomes expand to include genes involved in host cell remodeling and a broad superdomain is created on chromosome 14. A prominent feature of genome organization in the sporozoite stage is the establishment of long-range DNA interactions for genes involved in sporozoite migration and hepatocyte invasion. Our results provide important novel insights into the connection between genome organization, heterochromatin, and stage-specific gene expression.

Results

Capturing genome conformation of Plasmodium transmission stages

To complement our previous study describing the genome architecture of *P. falciparum* during the intraerythrocytic developmental cycle (IDC) [16], we performed Hi-C experiments on three additional stages of the *P. falciparum* life cycle: early gametocytes (stage II/III), late gametocytes (stage IV/V), and salivary gland sporozoites (Figure 1.1A, Supplementary Table 1.1, and Supplementary Figure 1.1) using the tethered conformation capture methodology. In addition, to evaluate similarities and differences in genome organization between the highly pathogenic *P. falciparum* and the less virulent *P. vivax*, we generated Hi-C data for *P. vivax* salivary gland sporozoites. For each stage, we obtained high-quality data, evidenced by a log-linear relationship between contact probability and genomic distance (Supplementary Figure 1.2A), as well as interchromosomal contact probability (ICP) and percentage of long-range contacts (PLRC) values in agreement with previous studies (Supplementary Table 1.2). Biological replicates of both *P. falciparum* and *P. vivax* sporozoite stage parasites showed a high degree of similarity (Supplementary Figure 1.2B-C), demonstrating the robustness of our methodology. The data from these replicates were combined to obtain higher resolution for subsequent analyses.

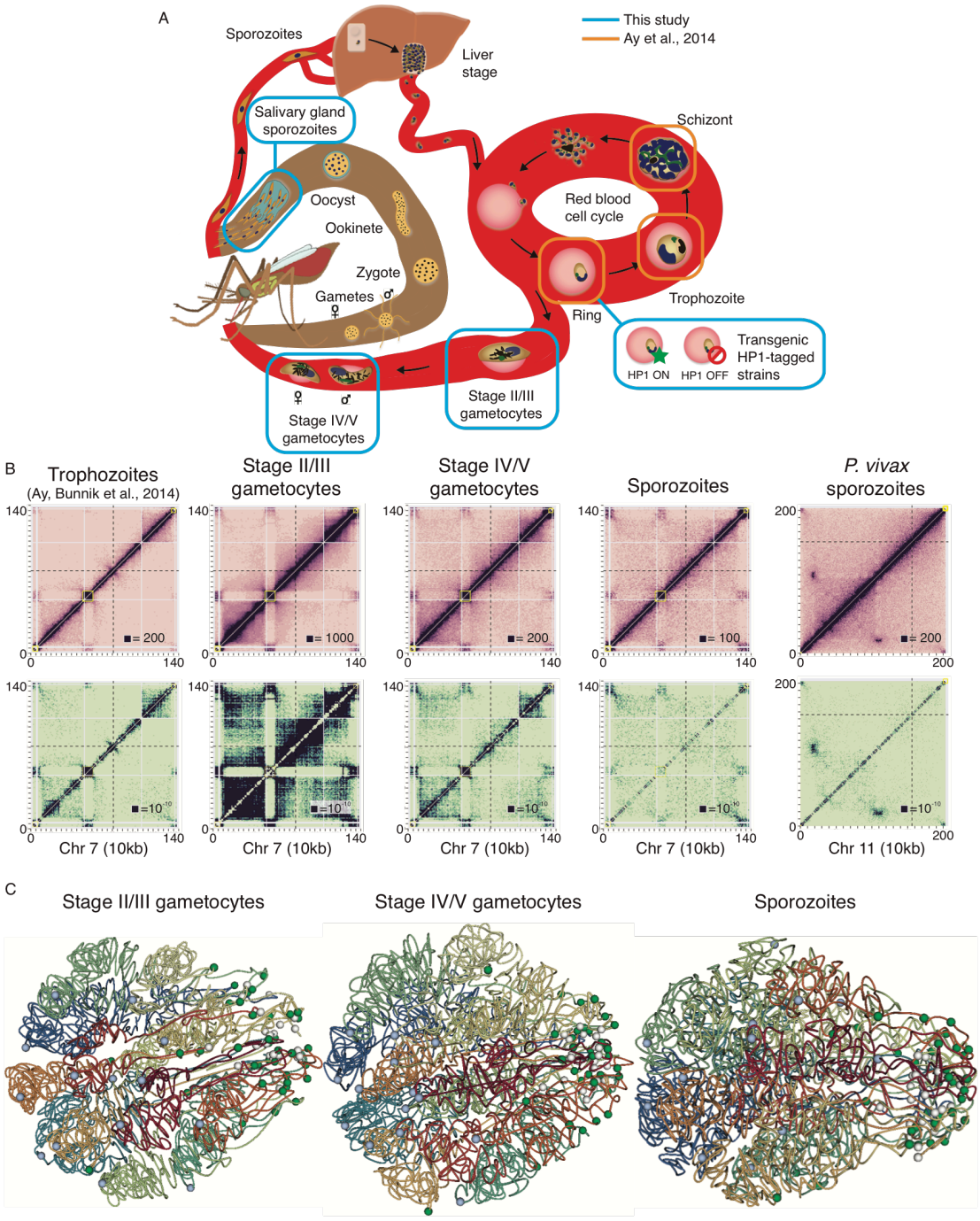

A
This study
Ay et al., 2014
Sporozoites
Liver stage
Salivary gland sporozoites
Schizont
Oocyst
Ookinete
Red blood cell cycle
Zygote
Gametes ♂ ♀
Trophozoite
Ring
Transgenic HP1-tagged strains
HP1 ON
HP1 OFF
Stage II/III gametocytes
♀ ♂
Stage IV/V gametocytes
B
Trophozoites (Ay, Bunnik et al., 2014)
Stage II/III gametocytes
Stage IV/V gametocytes
Sporozoites
P. vivax sporozoites
Chr 7 (10kb)
Chr 7 (10kb)
Chr 7 (10kb)
Chr 7 (10kb)
Chr 11 (10kb)
C
Stage II/III gametocytes
Stage IV/V gametocytes
Sporozoites

Figure 1.1: Genome organization in *Plasmodium* parasites. (A) Schematic overview of the parasite life cycle, with the samples generated in this study highlighted in blue and stages available from a previous study{Ay, 2014 #319} shown in orange. (B) ICE-normalized contact count matrices (top row) and fit-hi-c p-value matrices (bottom row) at 10 kb resolution of chromosome 7 for *P. falciparum* stages and chromosome 11 for *P. vivax* sporozoites. The boxed value indicates the maximum contact count (top row) or minimum P-value (bottom row). In all other figures comparing different stages, the contact counts were subsampled to the same total. Virulence clusters are indicated by yellow boxes, the centromere location by a dashed black line, and unmappable regions by grey in these and all other heatmaps. (C) Models of the consensus three-dimensional organization of the *P. falciparum* genome in stage II/III gametocytes, stage IV/V gametocytes and salivary gland sporozoites, with light blue spheres indicating centromeres, white spheres indicating telomeres and green spheres indicating the location of virulence gene clusters.

Next, the observed intrachromosomal and interchromosomal contacts were aggregated into contact count matrices at 10-kb resolution and were normalized using the ICE method to correct for experimental and technical biases [18] (Figure 1.1B). In addition, we identified significant contacts using fit-hi-c, which controls for the propensity of adjacent loci to have more contacts and calculates a *p*-value reflecting the probability that the number of contacts in that bin is larger than expected by chance [19]. We inferred a consensus 3D genome structure for each of the transmission stages

and the three IDC stages using Pastis [20] (Figure 1.1C). The stability of these consensus structures was assessed by generating 5000 possible structures from varying initial starting points. A principal component analysis showed strong clustering of structures from the same stage, and clear separation between structures of different stages, except for early and late gametocytes (Supplementary Figure 1.2D). These results indicate that the genome organization of early and late gametocytes is similar, while there are distinct differences between all other stages that are captured using a single representative structure for each stage.

Universal and stage-specific features of genome organization

From the contact count matrices and the consensus 3D structures, it became apparent that large-scale features of genome organization at the gametocyte stage were comparable to those of the IDC stages, including colocalization of centromeres and clustering of telomeres and virulence genes (all p-values < 0.00001, Witten–Noble colocalization test [21]). However, we observed significant intrachromosomal rearrangements, including increased interactions among virulence genes and exported proteins, repression of invasion genes, change in organization of ribosomal DNA genes, as well as the formation of large domains on chromosome 14 in close proximity to a female gametocyte-specific *pfap2* transcription factor locus. These changes will be addressed in more detail in the next sections. At the sporozoite stage of *P. falciparum*, the clustering of telomeres was conserved, but the colocalization of the centromeres was completely lost (p-value $= 0.49$, Witten–Noble colocalization test), Figure 1.1C,

rightmost panel, and Supplementary Figure 1.3A). In contrast, in *P. vivax* sporozoites, the centromeres colocalized significantly (p-value < 0.0001, Witten-–Noble colocalization test), but these interactions were only observed at the centromere itself and did not involve any of the surrounding regions (Supplementary Figure 1.3B). While our result will need to be validated by an independent approach, our Hi-C experiment is so far the only successful technique that has been able to monitor a reduction in centromere clustering in the sporozoite stage. To better visualize the large-scale differences between the various stages of the *P. falciparum* life cycle, we generated an animation of the changes in genome organization during the stage transitions, which highlights that chromosomes undergo dramatic rearrangement in sporozoites as compared to the blood stages (Supplemental Movie 1.1).

The groups of genes described in the previous sections were examined independently and were selected based on our prior knowledge of the function of these genes, rather than as the result of a systematic screen for changes in contact counts. To systematically identify changes in genome conformation between the various stages, we designed a statistical test that analyzes differences in the number of intrachromosomal contacts for each 10-kb bin in the normalized contact count matrices between pairs of stages. As expected, interactions at the sporozoite stages were most different from those in other stages of the life cycle. Several chromosomes showed large rearrangements towards the chromosome ends, for example the right arm of chromosome 4 (Supplementary Figure 1.4), involving gene families coding for

exported proteins involved in virulence and erythrocyte remodeling. The fold-change heatmaps (accessible at http://noble.gs.washington.edu/proj/plasmo3d_sexualstages/), the contact count heatmaps, and the confidence score heatmaps showed many additional differences in chromosome conformation during stage transitions, as detailed below. The interaction patterns were similar in two distinct field isolates for sporozoites and in two laboratory strains for gametocytes, suggesting that genetic variations did not influence our results. Furthermore, the Hi-C methodology has recently been used to detect translocations in genomes and to correct genome assemblies based on Illumina and/or PacBio sequencing in many organisms, including *Plasmodium knowlesi*, *Arabidopsis thaliana*, and *Aedes aegypti* [22-25]. It is therefore unlikely that the changes that we observed between life cycle stages are artifacts caused by genomic recombination during in vitro parasite culture, since none of the interchromosomal heatmaps (Supplementary Figure 1.3A) showed any evidence of such recombination events. To further validate our results, we have introduced interchromosomal and intrachromosomal translocations in the *P. vivax* genome to visualize the aberrant patterns that such recombination events would produce (Supplementary Figure 1.5). In addition, we scanned our samples using a recently published metric developed to detect genome assembly errors in Hi-C data [24] and did not detect any signs of misassembly or translocations (Supplementary Figure 1.6). Additional simulations showed that translocations of a single 10-kb bin are not easily detectable, but larger regions are detectable when there is a separation of a few bins between them (Supplementary Figure 1.7).

pfap2-g leaves the repressive center during gametocytogenesis

Previous work has shown that knockdown of heterochromatin protein 1 (PfHP1) during the IDC results in activation of the gametocyte-specific transcription factor locus *pfap2-g* and an increased formation of gametocytes [26]. PfHP1 interacts with the repressive histone mark H3K9me3 on silenced *var* genes that are colocalized in a perinuclear heterochromatic compartment. Using DNA-FISH, we observed that the *pfap2-g* locus was located in close proximity to a subtelomeric *var* gene on chr8 in >90% of the cells observed (Figure 1.2A and Supplementary Figure 1.8A), suggesting that *pfap2-g* is associated with the repressive cluster during the IDC. In agreement with this observation, the trophozoite and schizont stage fit-hi-c *p*-value heatmaps showed a significant interaction between *pfap2-g* and the nearest internal virulence gene cluster (Figure 1.2B and Supplementary Figure 1.9A). The virulence cluster and *pfap2-g* locus each straddle two ten kilobase bins, and in trophozoites, each of the four possible pairwise interactions were significant (fit-hi-c q-value < 0.05). We performed virtual 4C at MboI restriction site resolution to demonstrate that this interaction was specific for *pfap2-g* and did not involve the nearby *pfap2* PF3D7_1222400 (Supplementary Figure 1.10), although these two *ap2 TF* genes are located in neighboring 10 kb bins and therefore fall outside our limits to detect potential translocations. In addition, *pfap2-g* significantly interacted with virulence clusters on chromosomes 6 and 8 in trophozoites (fit-hi-c q-value < 0.05). In stage II/III gametocytes, no significant interactions between *pfap2-g* and virulence clusters were observed (q-values = 1.0; Figure 1.2B; Supplementary Table 2.3), indicating

that *pfap2-g* dissociates from the repressive cluster in the transition from the IDC to early gametocytes. Interactions between *pfap2-g* and virulence clusters were partially regained in the late gametocyte stage (Supplementary Table 1.3). Unfortunately, DNA-FISH experiments were unsuccessful for the gametocyte stage. An improved DNA-FISH methodology for the gametocyte stages will need to be developed to further confirm these results by an independent approach.

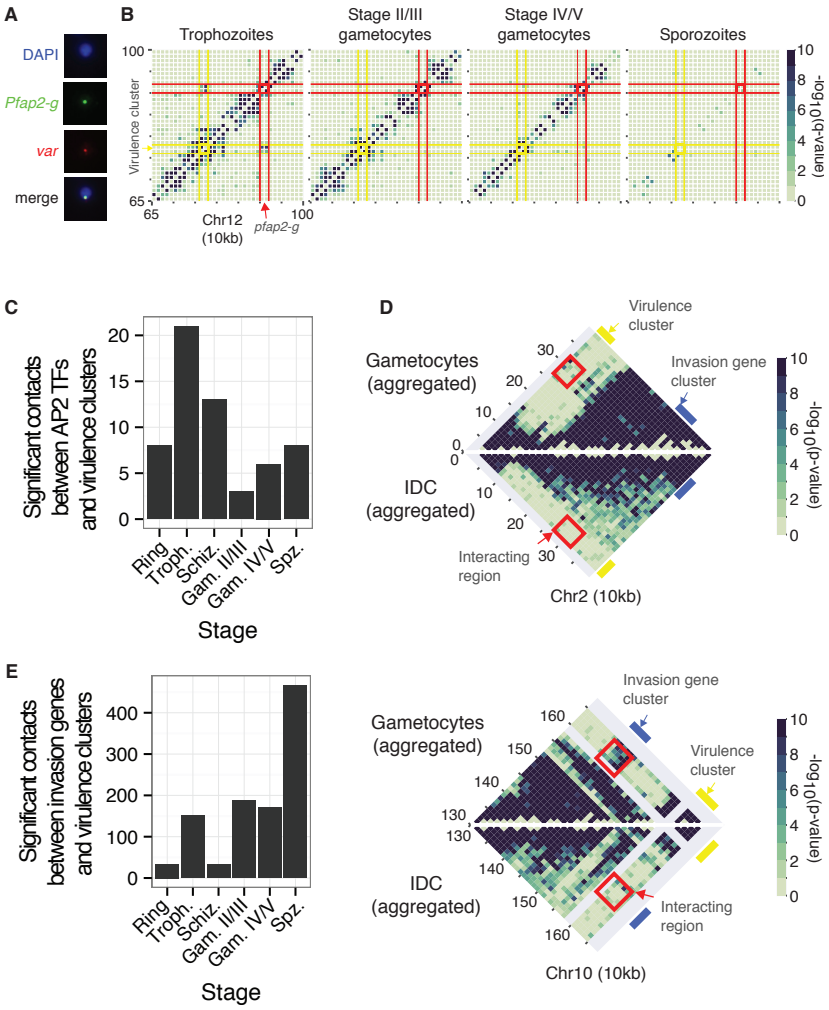


Figure 1.2: Changes in interaction of *pfap2* genes and invasion genes with the repressive center. (A) Colocalization of *pfap2-g* and *var* gene PF3D7_0800300 by DNA-FISH. Additional images are presented in Supplemental Fig. 2.6. (B) Dissociation of the gametocyte-specific transcription factor locus *pfap2-g* (red) from the nearby internal virulence gene cluster (yellow) in stage II/III gametocytes. (C) Overall reduced number of intrachromosomal and interchromosomal interactions between *pfap2* TF genes and virulence genes in gametocytes and sporozoites as compared to the IDC stages. (D) Invasion gene clusters (blue bar) on chromosomes 2 (top) and 10 (bottom) interact with subtelomeric virulence genes (yellow bar) in gametocytes, but not during the IDC. In each plot, the top triangle shows the aggregated data of both gametocyte stages, while the bottom triangle shows the aggregated data from the three IDC stages. Bins that depict interactions between virulence genes and invasion genes are highlighted by a red box. E) Increased number of intrachromosomal and interchromosomal interactions between invasion genes and virulence genes in gametocytes and sporozoites as compared to the IDC stages.

Significant contacts with virulence gene clusters at any parasite stage were observed for a total of 9 *pfap2* loci (Supplementary Table 1.3). The dissociation of *pfap2* genes from the repressive center in other stages than the IDC seemed to be a general trend: fewer significant contacts (q-value < 0.05) between *pfap2* genes and virulence gene clusters were observed in gametocytes and sporozoites as compared to the IDC stages (p-value = 0.0027, one sided sign test; Figure 1.2C and Supplementary Figure 1.9). We

used a resampling procedure to compare this result to randomly-selected bins; the *p*-value was 0.021. Note that *pfap2* PF3D7_0420300 is located in close genomic proximity to a virulence cluster in chromosome 4. To ensure that the significant decrease in *pfap2* gene loci interacting with virulence clusters is not due to this gene, we repeated the *p*-value calculation without PF3D7_0420300 and confirmed that this remained significant (*p*-value = 1.192e−07, one-sided sign test). Such changes in interactions with virulence gene clusters between stages were not observed for an unrelated gene family (histone genes, data not shown). These results are indicative of an important connection between genome organization and the activity of *pfap2* genes that drive parasite life cycle progression.

Relocation of invasion genes during gametocytogenesis

Distinct changes were observed in chromosomes 2 and 10 at loci that harbor invasion genes (Figure 1.2D). These genes encode proteins that are expressed in merozoites and mediate attachment to and entry of red blood cells and include several merozoite surface proteins, S-antigen, and glutamate-rich protein. In contrast to several other invasion genes (*pfrh4* [27], *clag*, and *eba* [28], these genes are not known to undergo clonally variant expression, and we therefore consider their association with heterochromatin in gametocytes to be a different regulatory mechanism than the epigenetic mechanisms that control expression of *pfrh4*, *clag*, and *eba* during the IDC. In gametocytes, these loci showed strong interaction with the subtelomeric regions, while these contacts were not observed in the IDC. To quantify this observation, we

assessed the number of significant contacts between invasion gene loci and virulence clusters in each life cycle stage. A larger number of significant contacts were observed between invasion genes and virulence clusters in the transmission stages than in IDC stages (p-value < 2.2e−16, sign test; Figure 1.2E and Supplementary Figure 1.9C). The lack of interactions between invasion gene GLURP on chromosome 10 (PF3D7_1035300) and *var* gene PF3D7_0800300 during the IDC was confirmed by DNA-FISH (Supplementary Figure 1.8B). As mentioned earlier, DNA-FISH experiments were unfortunately not successful for the gametocyte stage. Collectively, our data indicate that, similar to the *pfap2* gene family, the expression of invasion genes during the life cycle is correlated with association with or dissociation from repressive heterochromatin.

Expansion of subtelomeric heterochromatin in gametocytes

To further study changes in chromatin organization during gametocytogenesis, we determined the distribution of repressive histone mark H3K9me3 in late ring/early trophozoites and stage IV/V gametocytes by performing ChIP-seq on two biological replicates for each stage (Supplementary Figure 1.11A) that were combined for downstream analyses. We used two different commercially available anti-H3K9me3 antibodies to rule out that our ChIP-seq results were influenced by the antibody used. In trophozoites, H3K9me3 marking was restricted to subtelomeric regions, internal virulence gene clusters and a few additional loci (including *pfap2-g* and *dblmsp2*), as previously described for both H3K9me3 [7-8] and PfHP1 [29] (Figure 1.3A and

Supplementary Figure 1.11B-C). These same regions were occupied by H3K9me3 in gametocytes. In addition, in several chromosomes, the subtelomeric heterochromatin marking expanded to more internally located genes in gametocytes (Figure 1.3A-C and Supplementary Figure 1.11B-C). The expansion of heterochromatin at the chromosome ends was also visible in the contact count heatmaps as larger subtelomeric domains that showed strong intra-domain interactions and were depleted of interactions with the internal region of the chromosome. While not all genes in these regions were marked by H3K9me3, a total of 79 genes showed increased H3K9me3 levels in gametocytes as compared to trophozoites, 61 of which were exported proteins that may play a role in erythrocyte remodeling. These genes included members of the *phist* ($n = 15$ out of a total of 68), *hyp* ($n = 7$ out of 34) and *fikk* ($n = 7$ out of 19) families, as well as 32 other genes annotated as exported proteins or containing a PEXEL export motif (Figure 1.3D-E). Other members of gene families encoding exported proteins were marked with H3K9me3 in the IDC. However, 47 of these 61 genes have never been detected in a heterochromatic state in the IDC [7-8, 29]. At two loci on chr9 and chr14, respectively, H3K9me3 was lost in gametocytes (Figure 1.3F). The locus on chr9 is between two genes known to be involved in gametocyte differentiation (*pfgdv1* and *pfgig*) and is deleted in various gametocyte-defective *P. falciparum* strains [30]. On chromosome 14, the genes that were not marked by H3K9me3 in gametocytes encode exported proteins that have been implicated in gametocytogenesis: PF3D7_1476600, PF3D7_1477300 (*Pfg14-744*), PF3D7_1477400, and PF3D7_1477700 (*Pfg14-748*) [30]. These results imply that

H3K9me3 and chromatin structure play important roles in gene activation and silencing during gametocyte formation. To validate the 3D modeling and heterochromatin clustering, we performed immunofluorescence imaging against repressive histone mark H3K9me3. We identified a single nuclear H3K9me3 focus per nuclei in rings and schizonts (Figure 1.3G and Supplementary Figure 1.12), corresponding to the single repressive center harboring all virulence genes as predicted by our 3D models. In trophozoites, the number of foci varied, in line with nuclear expansion [16, 31], and the progression of DNA replication that takes place at the end of this stage (Figure 1.3G). Gametocytes showed either one or two H3K9me3 foci that did not seem to be associated with a male or female phenotype.

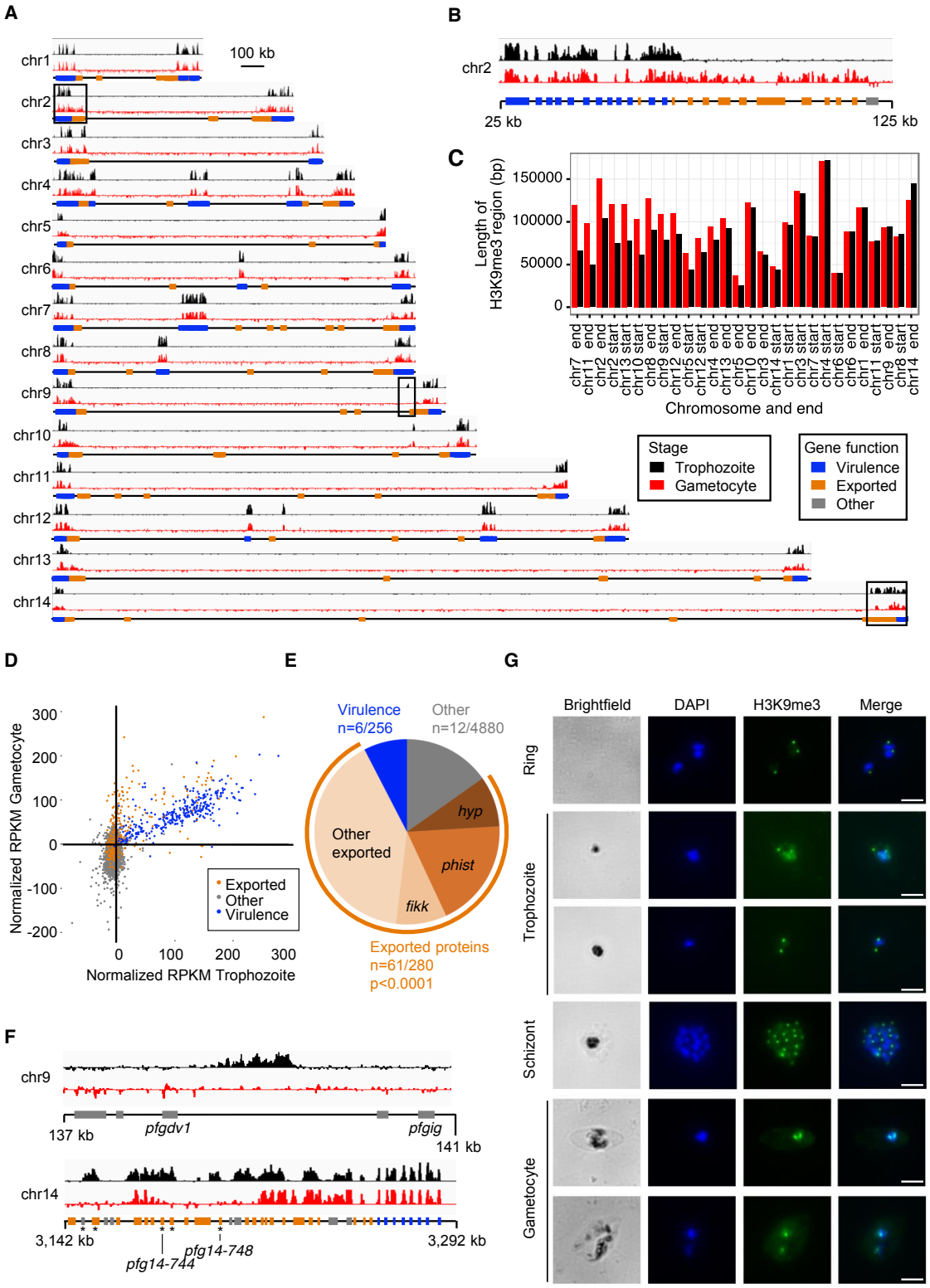

A
100 kb
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
B
chr2
25 kb
125 kb
C
Length of H3K9me3 region (bp)
150000
100000
50000
0
Chromosome and end
Stage
Trophozoite
Gametocyte
Gene function
Virulence
Exported
Other
D
Normalized RPKM Gametocyte
Normalized RPKM Trophozoite
300
200
100
0
-100
-200
Exported
Other
Virulence
E
Virulence
n=6/256
Other
n=12/4880
hyp
Other exported
phist
fikk
Exported proteins
n=61/280
p<0.0001
F
chr9
137 kb
pfgdv1
pfgig
141 kb
chr14
3,142 kb
pfg14-744
pfg14-748
3,292 kb
G
Brightfield
DAPI
H3K9me3
Merge
Ring
Trophozoite
Schizont
Gametocyte

Figure 1.3: Silencing of genes encoding exported proteins in gametocytes through expansion of heterochromatin. (A) ChIP-seq analysis of genome-wide H3K9me3 localization in trophozoites (top tracks in black) and stage IV/V gametocytes (bottom tracks in red). Results of one representative biological replicate are shown for each stage. The regions depicted in panels B and F are indicated with black boxes. (B) Expansion of H3K9me3 heterochromatin in gametocytes as compared to trophozoites, predominantly to genes encoding exported proteins. (C) Length of each subtelomeric region in which the majority of genes is marked by H3K9me3, sorted by the difference in length between these regions in trophozoites and gametocytes. (D) H3K9me3 levels per gene at the trophozoite and gametocyte stages. (E) Enrichment of genes encoding for exported proteins among genes with increased levels of H3K9me3 in gametocytes (P-value from a two-tailed Fisher's exact test). (F) Loss of H3K9me3 mark in gametocytes on chromosome 9 between gametocyte development genes *pfgdv1* and *pfgig*, as well as at gametocyte-specific genes encoding exported proteins on chromosome 14 (indicated with an asterisk). (G) Immunofluorescence analysis showing a single H3K9me3 focus in ring and schizont stages, and either one or two foci in gametocytes. Scale bar denotes 1 μm.

Formation of superdomains on a P. falciparum chromosome

Chromosome 14 showed the formation of a strong domain boundary in both early and late gametocytes, which was not observed in any of the other life cycle stages (Figure 1.4A). This separation of the chromosome into two superdomains is reminiscent of the bipartite structure of the inactivated X chromosome (Xi) in human,

rhesus macaque, and mouse [32-34]. Zooming in on the boundary region of *P. falciparum* chromosome 14 showed a sharp transition at the MboI restriction site at nucleotide position 1,187,169 (Figure 1.4B and Supplementary Figure 1.13). To demonstrate that this sharp boundary is not the result of a chromosomal translocation in the NF54 strain used for gametocyte isolations, we confirmed that the genomic region that spans the boundary region can be amplified by PCR and can be detected by Southern blot in both 3D7 and NF54 strains (Supplementary Figure 1.14). In eukaryotic genomes, genes close to the domain boundary are often associated with higher levels of transcription [35-36]. The domain boundary is located inside or near PF3D7_1430100, which encodes serine/threonine protein phosphatase 2A activator (PTPA; Supplementary Figure 1.15A). In humans, serine/threonine protein phosphatase 2A (PP2A) is one of the four major Ser/Thr phosphatases and is thought to play a complex, but mostly inhibitory role in the control of cell growth and division [37]. Gametocytes have a higher expression level of *pfptpa* than the IDC stages and express a different variant of *pfptpa* that does not contain exon 1 (Supplementary Figure 1.16A). The sequence of intron 1 is unusual and contains many motifs that are repetitive (for example, 12 repeats of motif TGTACATACACTTAT and minor variations thereof, within the 705 nt intron; Supplementary Figure 1.16B). These could be the binding sites for a lncRNA or protein involved in formation of the domain boundary.

The domain boundary is also relatively close to the *pfap2*-encoding locus PF3D7_1429200 (chr14:1,144,518–1,148,078) (Supplementary Figure 1.15A). To evaluate whether this TF could be involved in sexual differentiation, we generated a transgenic *P. berghei* strain in which the homolog of PF3D7_1429200 (PBANKA_1015500) was expressed as a GFP-tagged protein (Supplementary Figure 1.15B-C). *P. berghei* is widely used as a model for *P. falciparum*, in part because of the higher efficiency of genetic manipulations as compared to *P. falciparum*. Female gametocytes, activated female gametes, and zygotes all expressed the tagged ApiAP2 TF with nuclear localization of the protein, while the protein was completely absent in male gametocytes and gametes (Figure 1.4C), as well as in IDC stages (data not shown). These results demonstrate that this ApiAP2 TF (named PfAP2-O3 hereafter in line with a recent publication [38]) is expressed in a strict sex-specific fashion.

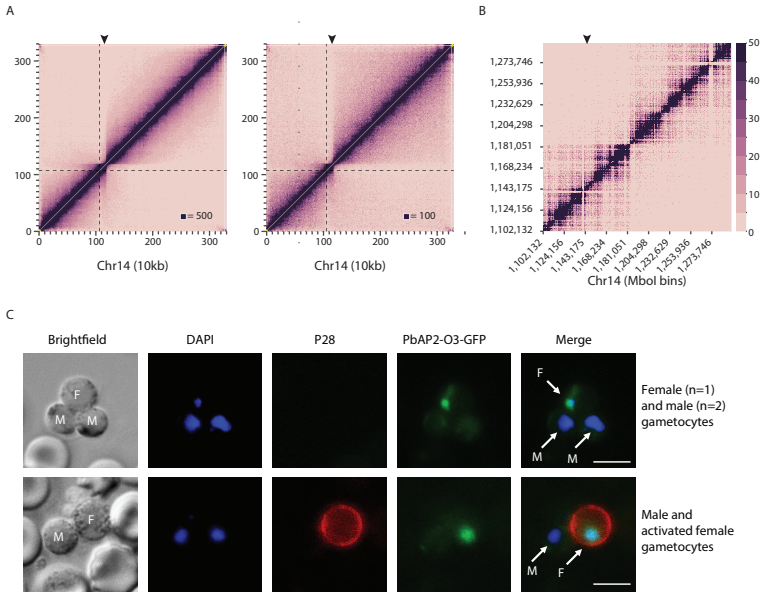


Figure 1.4: Formation of superdomains on chr14 in gametocytes. (A) ICE-normalized contact count heatmap at 10 kb resolution of early gametocyte (left) and late gametocyte (right) chromosome 14 showing the separation of the chromosome into two superdomains. The dashed line indicates the location of the centromere, and the arrowhead indicates the position of PF3D7_1429200. (B) Smaller region of chromosome 14 centered on the domain boundary that is located inside PF3D7_1430000, a conserved gene with unknown function. (C) The homolog of *pfap2* gene PF3D7_1429200 in *P. berghei* (PBANKA_1015500; *pbap2-o3*) has a nuclear localization in female gametocytes and gametes, but is not detected in male gametocytes. The top row shows male and female gametocytes. The bottom row shows a male and female gamete activated by mosquito ingestion, which triggers expression of the female-specific surface protein P28. Male (M) and female (F) parasite are indicated in the brightfield and merged images. Scale bar denotes 10 μm.

Rearrangement of chromosomes in sporozoites

Similar to the re-localization of invasion genes during the transition from the IDC to the gametocyte stage, the invasion genes also interacted more strongly with virulence genes in *P. falciparum* sporozoites than during the IDC (Figure 1.2E and Supplementary Figure 1.9C). In *P. vivax* sporozoites, a cluster of invasion genes on chromosome 10 showed depletion of interactions with other loci on the same chromosome as compared to surrounding genomic regions (Supplementary Figure 1.17B). This observation may suggest that the invasion genes also have a distinct genome organization at this stage of the *P. vivax* life cycle. However, these results may also be caused by sequence variation in invasion genes in the field isolates used in this study as compared to the reference genome, resulting in lower mapping to this region.

In addition, distinct changes were noticeable around rDNA loci. *P. falciparum* encodes four rDNA units containing single copies of the 28S, 5.8S, and 18S genes (Figure 1.5A). The units on chromosomes 5 and 7 are active during the human blood stages, whereas the units on chromosomes 1 and 13 are active in the mosquito stages. In general, sporozoites showed a large increase in the number of contacts between rDNA genes and virulence genes as compared to the IDC stages and gametocytes (Figure 1.5B and Supplementary Figure 1.9D). These changes in conformation were most visible in chromosome 7, in which the rDNA unit is located at the boundary of two large domains in the IDC stages and gametocytes, which presumably contributes

to its activation status. In sporozoites on the other hand, the separation of the chromosome into two large domains disappeared (Figure 1.5C and Supplementary Figure 1.4B). Similar changes in domain conformation can be observed around the rDNA locus on chromosome 5 (Supplementary Figure 1.9E)

.

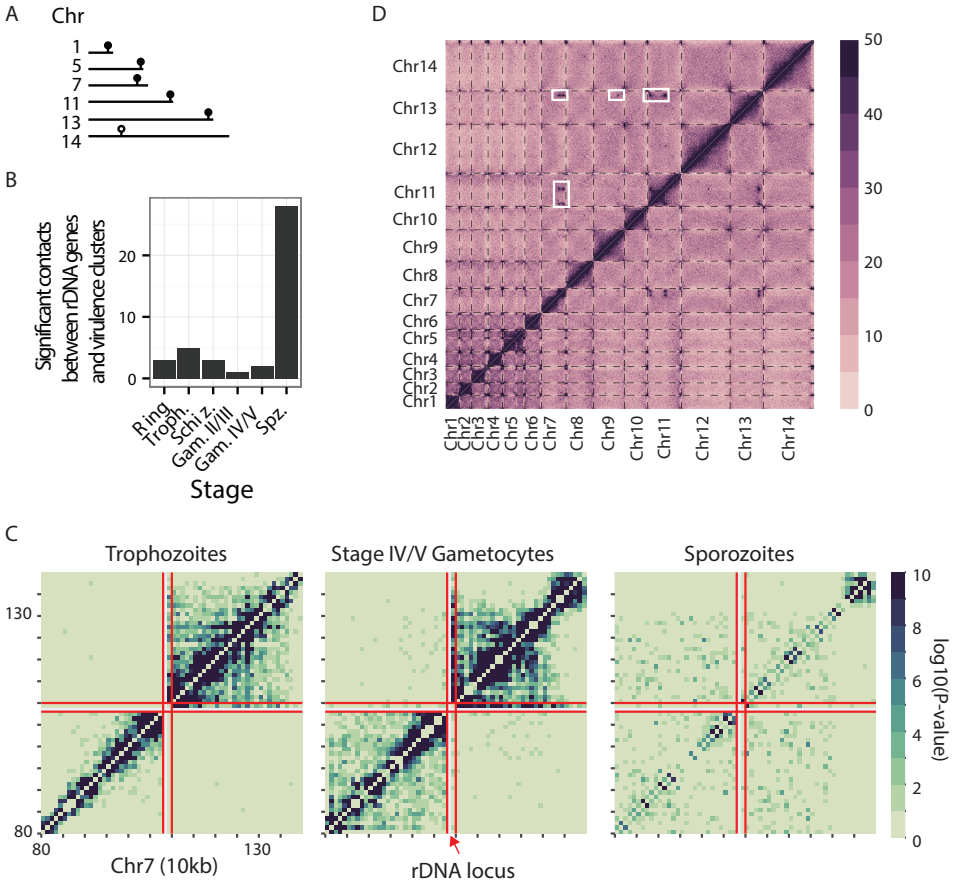


Figure 1.5: Changes in genome organization in salivary gland sporozoites. (A) Locations of rDNA genes in the *P. falciparum* genome. Units of 28S, 5.8S, and 18S genes are indicated with a filled symbol; the unit of three 5S genes is indicated with an open symbol. (B) Increased overall number of interactions between rDNA genes and virulence genes in *P. falciparum* sporozoites. (C) Loss of domain formation around the rDNA locus on chr7 in *P. falciparum* sporozoites as compared to other life cycle stages. The borders of the rDNA locus are indicated by red lines. (D) Strong interchromosomal interactions in *P. vivax* sporozoites, indicated by white rectangles. Dashed lines indicate chromosome boundaries.

Prominent features of genome organization in *P. falciparum* and *P. vivax*sporozoites were strong long-range and interchromosomal contacts that involved other genes than virulence genes and that did not seem to be present in the IDC or gametocyte stages. In *P. vivax*, strong intrachromosomal contacts were present in chromosomes 7 and 11, which also formed strong interchromosomal interactions with each other and with additional loci on chromosomes 9 and 13 (Figure 1.1B, rightmost panel, Figure 1.5B, and Supplementary Table 1.4). For *P. falciparum*, intrachromosomal interactions were observed in chromosomes 3, 4, 8, 9, 11, 13, and 14, although, interestingly, these are not homologous to the loci that participate in loops in *P. vivax*. Several of these loops involved *pfap2* loci and genes involved in sporozoite migration to the liver and in some cases in hepatocyte invasion, such as circumsporozoite protein (PfCSP), sporozoite micronemal protein essential for cell traversal (PfPLP1), thrombospondin-

related anonymous protein (PfTRAP), sporozoite protein essential for cell traversal (PfSPECT1), and gamete egress and sporozoite traversal protein (PfGEST) (Table 1.1).

A putative clonally variant gene family in P. vivax

The genome annotation of *P. vivax* is less complete than that of *P. falciparum*, and many genes have been grouped into families based solely on sequence homology. An example is the *Pv-fam-e* (also named *rad*) gene family that is closely related to the *Plasmodium* helical interspersed sub-telomeric (*phist*) gene family [39], encoding exported proteins involved in erythrocyte remodeling [40-42]. The *P. vivax* genome contains 45 *rad* genes, of which 10 and 27, respectively, are located in two separate clusters on chromosome 5. In the contact count matrix of *P. vivax* chromosome 5, the largest of the two clusters strongly interacted with the (sub-)telomeric regions and showed a depletion of interactions with all other intrachromosomal loci (Supplementary Figure 1.17C). These results suggest that *P. vivax rad* genes may be regulated by organization into facultative heterochromatin.

Table 1.1: Loci involved in long-range intrachromosomal interactions in *P. falciparum* sporozoites.

Chr	Locus (kb) [a]	Gene	Description	Pv homolog
3	115	PF3D7_0302100	Ser/Thr protein kinase 1 (PfSTPK1)	PVX_119250
3	225	PF3D7_0304600	CS protein (PfCSP)	PVX_119355
4	245	PF3D7_0404500	6-cys protein (P52)	PVX_001020
4	375	PF3D7_0407600	Conserved, unknown function	n.a.
4	425	PF3D7_0408700	Perforin -Like Protein 1 (PfPLP1)	PVX_000810
8	135	PF3D7_0801900	Conserved, unknown function	PVX_093645
8	295	PF3D7_0805200 PF3D7_0805300	Gamete release protein (PfGAMER) Conserved, unknown function	PVX_093500 PVX_093495
9	125 135	PF3D7_0902800 PF3D7_0902900 PF3D7_0903000 PF3D7_0903100 PF3D7_0903200	Serine repeat antigen 9 (PfSERA9) Conserved, unknown function Conserved, unknown function PfRER1 PfRAB7	N.a. N.a. PVX_098595 PVX_098600 PVX_098605
9	325	PF3D7_0906600	Zinc finger protein	PVX_098775
9	535 545	PF3D7_0911700 PF3D7_0911800 PF3D7_0911900 PF3D7_0912000 PF3D7_0912100	GTP-binding protein Conserved, unknown function Falstatin (PfICP) Conserved, unknown function Zinc finger protein	PVX_099025 PVX_099030 PVX_099035 PVX_099040 PVX_099045
11	225	PF3D7_1105000 PF3D7_1105100 PF3D7_1105200	Histone H4 (PfH4) Histone H2B (PfH2B) WD repeat-containing protein (PfWRAP73)	PVX_090930 PVX_090935 PVX_090940
11	335	PF3D7_1107800	ApiAP2 TF	PVX_091065
13	1,465	PF3D7_1335900	PfTRAP	PVX_082740
13	1,675	PF3D7_1342500	PfSPECT1	PVX_083025

PfHP1 is essential for virulence gene colocalization

The clustering of virulence genes seems to be a general feature of the *P. falciparum* genome that is maintained throughout its life cycle. Recently, it was shown that depletion of PfHP1 results in loss of *var* gene repression and an arrest in parasite growth [26], suggesting that this protein is essential for structural integrity of the repressive cluster. To study the effect of PfHP1 depletion on genome conformation, we performed Hi-C experiments on a transgenic *P. falciparum* strain expressing PfHP1 fused to GFP and a destabilization domain (DD), both in the presence and in the absence of Shield-1, resulting in expression or knockdown of the PfHP1 fusion protein, respectively [26]. Ring-stage parasites expressing the tagged PfHP1 protein showed a decrease in virulence gene clustering as compared to wild-type parasites (p-value = 0.026 and p-value ≤ 0.001, respectively, BH FDR-corrected Paulsen colocalization test [43]). In particular, interactions between internal and subtelomeric *var* gene clusters were lost (Supplementary Figure 1.18). This result is in agreement with an increase in internal *var* gene expression in this strain, as reported previously [26]. Virulence gene clustering was completely lost in the PfHP1-depleted strain (p-value = 0.129, BH FDR-corrected Paulsen colocalization test), in line with a generalized loss of *var* gene repression [26]. Accordingly, we observed more significantly changing virulence gene bins as compared with wild-type in the PfHP1-depleted strain (n = 71) than in the PfHP1-tagged strain (n = 18). These results confirm that PfHP1 is indeed essential for maintenance of the structure of the repressive cluster and thus for regulation of virulence gene expression.

The 3D genome structure correlates with gene expression

Finally, we explored the relationship between gene expression and 3D structure, leveraging four published expression data sets [44-47] and our 3D models of the genome architecture. As in our previous study, we applied kernel canonical correlation analysis (KCCA) [48]. KCCA is an unsupervised learning approach akin to principal component analysis that identifies a set of orthogonal gene expression components that are coherent with the 3D structure. To perform this analysis, we separated the 3D models into three distinct groups: those related to IDC (ring, trophozoite, schizont), gametocyte (early and late) and sporozoite stages. For each of these groups of time points, we extracted a gene expression component and a structure component that exhibited coherence to the expression profiles and 3D structure, such that genes whose expression is correlated with the selected gene expression component tend to be colocalized in 3D. The gene expression components for the three sets of structures were highly correlated and were dominated by the repressive center (Figure 1.6A and ref. 16). To further interpret the results of the KCCA, we extracted ranked lists of genes based on their KCCA scores: lists that rank genes based on the similarity of their gene expression profiles with the first or second gene expression components, and lists that rank genes based on the similarity of their 3D position with the first or second structure components. We then investigated whether several sets of genes were enriched in those ranked lists. *Var*, *rifin*, and exported protein genes all showed strong and significant enrichment on the first gene expression and structure components for all stages (Figure 1.6B), as expected based on their localization in or near the

repressive center. In contrast, the invasion genes were significantly enriched for the first gene expression component in gametocytes and sporozoites, and for the second gene expression and structure components in all stages (Figure 1.6B). While the average scores for this group of genes are relatively small, these results suggest that expression of these genes is also coordinated with their location within the nucleus. Gametocyte-specific genes did not show correlation to the first or second component for either gene expression or structure, which is in line with regulation of these genes by other factors, such as the gametocyte-specific transcription factor PfAP2-G, instead of localization within the nucleus.

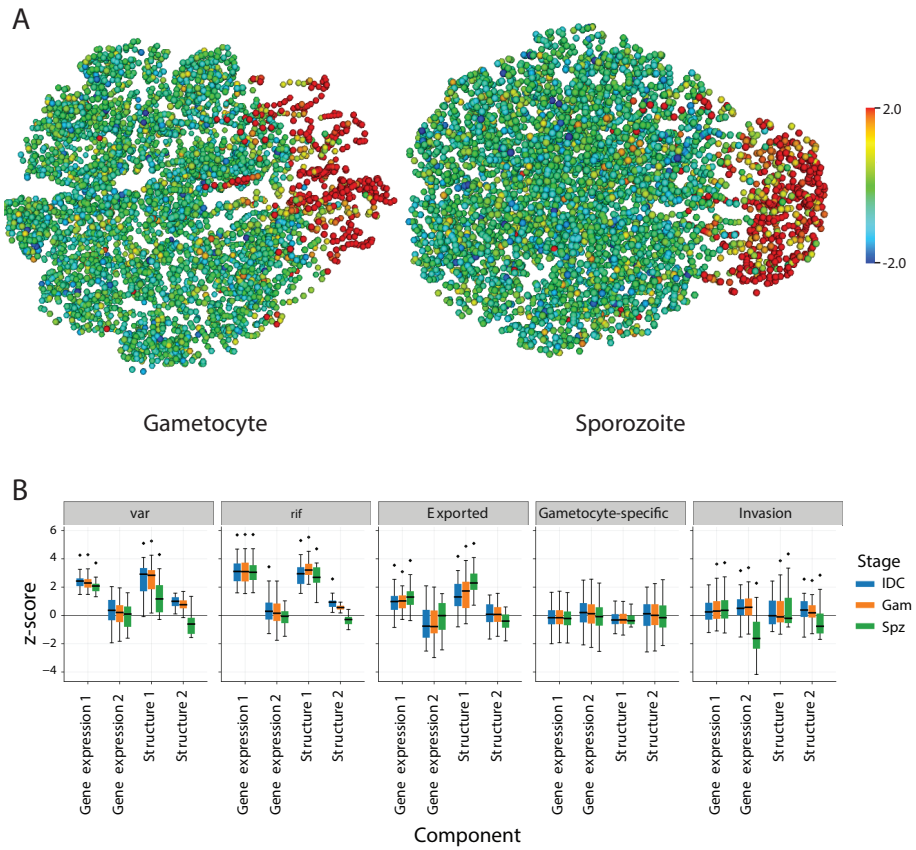


Figure 1.6: 3D genome structure correlates with gene expression. (A) Each gene is plotted by its position within the 3D structure and is colored by its standardized KCCA score (see Computational Methods) in the first gene expression component. The direction of this first gene expression component is from the telomere cluster on the right to the opposite side of the nucleus and is dominated by the repressive center (genes colored in red). (B) The average standardized KCCA score for the first and second gene expression components and the first and second structure components of specific groups of genes. Stars indicate groups of genes for which the standardized KCCA scores for both the gene expression component and the structure component were significantly different (t-test, FDR < 0.1%).

Discussion

Understanding the mechanisms involved in gene regulation during the various life cycle stages of *P. falciparum* will be important for the development of novel strategies to block parasite replication and transmission. Our previous study described the hallmarks of genome organization during the intra-erythrocytic developmental stages of *P. falciparum*, and showed that nuclear architecture correlated well with gene expression [16]. Here, we investigated chromosome conformation and chromatin structure in the stages of parasite transmission from human to mosquito (gametocytes) and from mosquito to human (sporozoites) and compared all stages to identify various subsets of genes that exhibit changes in genome organization during the complex life cycle of *Plasmodium* parasites.

Our results confirm that virulence gene families, such as *var, rifin*, and *stevor*, cluster in perinuclear heterochromatin, not only during the IDC stages, but also in gametocytes and sporozoites. Previous reports showed multiple virulence gene clusters scattered around the nucleus, either by DNA-FISH on telomere repeat sequences [12-13], IFA on H3K9me37 or IFA on proteins that bind to chromosome ends (PfSIP2-N) or subtelomeric regions (PfHP1 [26, 49]). Our population-based Hi-C data showed strong interactions between all telomeres, which could be consistent with a completely random distribution of telomeres over multiple clusters. However, our IFA results conflict with this large body of data and instead show a single H3K9me3 focus in all blood stages in which the parasite is not undergoing DNA replication, representative of a single repressive center. Similar results were previously obtained by IFA against the heterochromatin mark H3K36me31 [11] and also recently for heterochromatin mark H3K9me35 [50]. Differences in timing during the cell cycle (before or after start of DNA replication) may account for some of these differences. Indeed, we also observed multiple H3K9me3 foci during the trophozoite stage, most likely as a result of nuclear expansion and the start of DNA multiplication. We currently do not have an explanation that would reconcile our observations with those of others in the field. However, whether *var* genes cluster in one or multiple repressive centers throughout the IDC, both models are consistent with the presence of distinct heterochromatin and euchromatin regions controlling gene expression and antigenic variation in *P. falciparum*.

In gametocytes, the overall genome organization is similar to IDC stages, in line with relatively small changes in the gene expression program during gametocytogenesis [51]. We observed specific changes for *pfap2-g* [3] that dissociates from the repressive center, and for invasion genes that associate with the repressive center. In addition, subtelomeric genes encoding exported proteins are selectively silenced or activated by H3K9me3 deposition or removal, respectively. These results confirm that remodeling of the infected host cell is among the essential changes that occur during gametocytogenesis [52-53], and that an epigenetic switch provides an extra layer of transcriptional regulation for the genes involved. The presence of H3K9me3 in the relatively large intergenic region downstream of *pfgdv1* suggests that this region by itself is an important determinant in the regulation of gametocyte differentiation. A lncRNA is transcribed from this region at low levels during the IDC5 [54], and it is tempting to speculate that this transcript is upregulated in gametocytes and possibly essential for gametocyte development.

The sharp division of chromosome 14 into two superdomains is an intriguing finding that will need to be explored in more detail. The inactivated X chromosome (Xi) in mammalian cells adopts a seemingly similar structure with a ~200 kb hinge region that is organized into euchromatin, while the surrounding chromosome is in a heterochromatic state [55]. However, the domain boundary on *P. falciparum* chromosome 14 is sharp, does not have a distinct hinge region, and is not surrounded by H3K9me3 marking. In support of the hypothesis that boundary

formation is sex-specific, we discovered that the homolog of ApiAP2 transcription factor PF3D7_1429200 (PfAP2-O3) in *P. berghei* is strictly female-specific, suggesting that PfAP2-O3 controls female gametocyte differentiation. A few hundred genes are differentially expressed between male and female gametocytes [56-57] and could be under the control of this protein, either through direct transcriptional regulation or through selective stabilization of female-specific transcripts. Interestingly, disruption of *ap2-o3* in *P. berghei*resulted in differential gene expression in gametocytes and reduced gametocyte levels, although the sex ratio was not influenced3 [38]. However, *pfap2-03* is located approximately 40 kb from the domain boundary and it is not directly clear if and how expression of this gene would be influenced by the formation of these superdomains.

An alternative hypothesis is that the domain boundary is involved in regulation of the nearby gene Pf3D7_1430100 (*pfptpa*). PfPTPA has been shown to bind and activate PP2A, and to block the G2/M transition [58]. The transition from rapidly dividing asexual parasites into cell cycle arrested gametocytes is likely to require tight cell cycle regulation by a protein such as PTPA. Given the differences in *pfptpa* expression between the IDC and gametocyte stages, we speculate that the domain boundary inside or close to this gene may be important for driving the expression of this gene, which may in turn activate PP2A to block cell division in gametocytes. In ongoing efforts, we are further investigating the role of the domain boundary in gene expression and gametocyte differentiation.

From an evolutionary perspective, *P. falciparum* and *P. vivax* are highly divergent within the family of *Plasmodium* species [59-60]. Salivary gland sporozoites maintain a relatively quiescent transcriptional state, waiting for injection into the human bloodstream and invasion of a hepatocyte before ramping up transcriptional activity. Our interpretation of the numerous long-range interactions in the sporozoite stage is that the majority of the genome is transcriptionally repressed, with the exception of several active loci that colocalize in transcriptional islands, giving rise to long-range interactions. The observation that the genes involved in these long-range interactions are not each other's homologs in *P. falciparum* and *P. vivax* is suggestive of species-specific gene expression and warrants further investigation.

To fully understand transcriptional regulation, it is of great interest to unravel the causal relationship between genome organization and transcriptional activity. In multicellular organisms, evidence is accumulating that certain aspects of genome organization are independent of transcription [61]. In addition, disruptions in genome structure that bring together previously isolated promoters and enhancers can result in gene activation [62]. It will be important to determine to what extent genome organization controls transcriptional activity in *P. falciparum*. Our findings bring a new level of insight into genome dynamics during the *Plasmodium* life cycle and open up new avenues for targeted approaches towards understanding parasite gene regulation. In addition, molecules inhibiting the (re-)structuring of the genome have the potential to act as potent transmission-blocking antimalarials.

Materials and Methods

Parasite strains and cultures

The *P. falciparum* strain NF54 (obtained from the MR4 malaria repository) was cultured at 5–10% parasitemia in human O^+ erythrocytes at 5% haematocrit. The induction of gametocyte-stage parasites was performed by sorbitol synchronization and culturing in a low medium volume. Stage IV/V gametocytes at 2% parasitemia were isolated from 150 ml of culture using a percoll gradient, were cultured for one additional day, and were then isolated by magnetic purification yielding 6.25×10^8 parasites (Supplementary Fig. 2.1). To obtain sporozoites, adult female *Anopheles stephensi* mosquitoes were allowed to feed on *P. falciparum* NF54 gametocyte cultures. Sporozoites were harvested from infected mosquitoes 14–19 days later. Stage II/III gametocytes were obtained using the *P. falciparum* $NF54^{Pfs16}$ reporter gene line [63]. Gametocytes were isolated by magnetic purification, yielding 1.17×10^8 parasites with high purity (>95% gametocytes) as determined by GFP expression (Supplementary Fig. 2.1). To obtain *P. vivax* sporozoites, female *Anopheles cracens* mosquitoes were fed on blood samples drawn from *P. vivax* infected patients who had given written informed consent and who attended a Shoklo Malaria Research Unit (SMRU) clinic in Mawker Thai or Wang Pha, on the western Thailand-Myanmar border. The research protocol was approved by Oxford Tropical Research Ethics Committee and adhered to the Declaration of Helsinki. Two biological replicates containing 21,583,075 and 29,245,000 sporozoites, respectively, were used for Hi-C experiments. Fifteen days post-infection, *P.*

*vivax*sporozoites were harvested from *An. cracens* salivary glands. The construction of *P. falciparum* 3D7 transgenic strain PfHP1-GFP-DD has been described previously26. Parasites were synchronized, split into two populations at 4–12 hpi and cultured in the presence or absence of Shield-1 as described. Parasites were harvested for Hi-C at 4–12 hpi in the next cell cycle.

Hi-C procedure

Parasites were crosslinked in 1.25% formaldehyde in warm PBS for 25 min on a rocking platform in a total volume between 1 and 10 ml, depending on the number of parasites harvested. Glycine was added to a final concentration as 150 mM, followed by 15 min of incubation at 37 °C and 15 min of incubation at 4 °C, both steps on a rocking platform. The parasites were centrifuged at 660×*g* for 20 min at 4 °C, resuspended in 5 volumes of ice-cold PBS, and incubated for 10 min at 4 °C on a rocking platfrom. Parasites were centrifuged at 660×*g*for 15 min at 4 °C, washed once in ice-cold PBS, and stored as a pellet at −80 °C. For late stage gametocytes, the crosslinking protocol was slightly modified. Late gametocytes were collected in lysis buffer (25 mM Tris-HCl, pH 8.0, 10 mM NaCl, 2 mM 4-(2-aminoethyl)benzenesulfonyl fluoride HCl (AEBSF), 1% Igepal CA-360 (v/v), and EDTA-free protease inhibitor cocktail (Roche)) and incubated for 10 min at RT. After homogenization by 15 needle passages, formaldehyde was added to a final concentration of 1.25%, followed by 10 more needle passages. The protocol was then continued as for all other samples. To map the inter-chromosomal and

intrachromosomal contact counts, crosslinked parasites were subjected to the tethered conformation capture procedure, using MboI for restriction digests, as previous described [32].

DNA fluorescence in situ hydridization

DNA-FISH experiments were performed as previously described [16]. In brief, probes were prepared using Fluorescein-High Prime and Biotin-High Prime kits (Roche) according to manufacturer's instructions. Template DNA was prepared by PCR (5 min at 95 °C, 35 cycles of 30 s at 98 °C, followed by 150 s at 62 °C, and 5 min at 62 °C) using the KAPA HiFi DNA Polymerase HotStart ReadyMix. Sequences of primers used for probe generation are shown in Supplementary Table 2.5. Double sorbitol synchronized ring-stage parasites were extracted using 0.015% saponin in cold PBS, washed in cold PBS and fixed in 4% formaldehyde in PBS at RT. A monolayer of parasites was deposited on a 9 × 9 mm frame-seal slide chamber on a standard microscopy slide and air-dried. Parasites were permeabilized with 0.1% Triton X-100 in PBS. After application of the denatured probes, the slides were denatured at 80 °C for 30 min followed by hybridization at 37 °C overnight. The slides were washed, equilibrated, washed in TNT solution, stained with DAPI and mounted.

Immunofluorescence microscopy

P. falciparum IDC-stage parasites and gametocytes were fixed onto slides using 4% paraformaldehyde for 30 min at RT. Slides were washed three times using 1× PBS.

The parasites were permeabilized with 0.1% Triton-X for 30 min at RT, followed by a wash step with 1× PBS. Samples were blocked overnight at 4 °C in IFA buffer (2% BSA, 0.05% Tween-20, 100 mM glycine, 3 mM EDTA, 150 mM NaCl and 1× PBS). Cells were incubated with anti-Histone H3 antibody (ab8898 (Abcam), 1:500 or 07–442 (Millipore), 1:500) for 1 h at RT followed by anti-rabbit Alexa Fluor 488 (Life Technologies A11008; 1:500). No differences were observed in the results obtained with the two primary antibodies (Supplementary Fig. 2.12). Slides were mounted in Vectashield mounting medium with DAPI. Images were acquired using an Olympus BX40 epifluorescence microscope.

H3K9me3 ChIP-seq

Asexual parasites were crosslinked for 10 min with 1% formaldehyde in PBS at 37 °C, while gametocytes were crosslinked with 1.25% formaldehyde in lysis buffer for 25 min at RT. Chromatin was sheared using the Covaris Ultra Sonicator (S220). ChIP was performed using 2 µg of anti-Histone H3K9me3 antibody (ab8898 (Abcam) for biological replicates #1 and 07–442 (Millipore) for biological replicates #2) or no antibody as a negative control as previously described [51].

Amplification and Southern blot of region spanning the domain boundary on chr14

Genomic DNA was isolated from a mixed blood stage culture of *P. falciparum* strain 3D7, from a mixed blood stage culture of *P. falciparum* strain NF54, and from a stage II/III gametocyte culture of *P. falciparum* strain NF54 using the DNeasy Blood &

Tissue kit (Qiagen). PCR amplifications were performed using 50 ng of genomic DNA, 0.4 µM of each primer (see Supplementary Table 2.5), and HiFi HotStart ReadyMix (KAPA Biosystems) with the following program: 5 min at 95 °C, 30 cycles of 30 s at 95 °C, 30 s at 52 °C, 2 min at 62 °C and a final extension of 6 min at 62 °C. Genomic DNA (100 ng) was digested with restriction enzyme DraIII (NEB), purified using the DNA genomic DNA Clean & Concentrator kit (Zymo Research), and separated by DNA electrophoresis on a 0.8% agarose gel (100 min at 35 V). The DNA was transferred to charged nylon membrane by upward capillary transfer. The DNA probe was generated by PCR amplification using 10 ng of genomic DNA, 0.4 µM of each primer (see Supplementary Table 2.5), 0.425 mM biotinylated dCTP, and HiFi HotStart ReadyMix (KAPA Biosystems) with the following program: 3 min at 95 °C, 35 cycles of 20 s at 98 °C, 30 s at 58 °C, 30 s at 62 °C and a final extension of 5 min at 62 °C. The membrane was incubated with 25 ng/ml probe in PerfectHyb Plus hybridization buffer (Sigma) overnight at 60 °C. The membrane was blocked with 5× Casein blocking buffer (Sigma) and incubated with HRP-conjugated streptavidin (1:2000). The membrane was developed using Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare).

Validation of sex-specific ApiAP2 TF in P. berghei

Transgenic parasites endogenously expressing a GFP-tagged version of the *P. berghei*homolog of Pf3D7_1429200 (PBANKA_1015500) were generated by single homologous recombination [64]. A 1323 bp region of PBANKA_1015500 omitting the

stop codon was PCR amplified on genomic DNA using primers T2191 (5′-CCCCGGTACCGAATGCCCTAATAAATCTATTTCAATAG, KpnI site underlined) and T2192 (5′-CCCCGGGCCCATATTTTTTTGGTCGTTGGAAATTAAAC, ApaI site underlined). This was inserted upstream of the *gfp* sequence in the p277 vector using KpnI and ApaI restriction sites as underlined in the primers. The p277 vector contains the human *dhfr* cassette, conveying resistance to pyrimethamine. Before transfection, the sequence was linearized using ClaI. For the endogenously C-terminal fusion GFP-tagged parasites, a diagnostic PCR reaction was used as illustrated in Supplementary Fig. 2.15C. Primers INT T219 (5′-CAAATGATATTATCCCTTATATTGAAAG) and ol492 (5′-ACGCTGAACTTGTGGCCG) were used to determine correct fusion of the gfp sequence at the targeted locus by single homologous recombination. The presence of the full-length gene was verified using primers INT T219 and T2192. Asexual proliferation and gametocytogenesis were analysed using blood smears. Gametocyte activation and zygote formation were monitored using in vitro cultures and the surface antigen P28. Blood stages were stained in schizont medium for 60 min at 37 °C. Cells were washed once with respective medium and resuspended in PBS containing Hoechst 33342 DNA stain before being mounted for fluorescent microscopy. Six-to-eight-week-old female Tuck-Ordinary (TO) or NIHS (Harlan) outbred mice were used for all experiments. All animal work at Nottingham has passed an ethical review process and was approved by the United Kingdom Home Office. Work was carried out in accordance with the United Kingdom “Animals (Scientific Procedures) Act 1986”

and in compliance with "European Directive 86/609/EEC" for the protection of animals used for experimental purposes under UK Home Office Project Licenses (30/3248). Sodium pentabarbitol was used for terminal anesthesia and a combination of ketamine followed by antisedan was used for general anesthesia.

Computational methods

We mapped and binned reads and created and ICE-normalized contact count matrices as previously described (ref. 16). A consensus 3D genome structure was inferred for each of the transmission stages and the three IDC stages using Pastis [20]. Pastis uses a maximum likelihood approach based on modeling contacts using a negative binomial model to account for the observed overdispersion. This method generates more accurate and robust structures when compared to the optimization-based approach we used previously [16]. To identify significant contacts, we modeled the effect of genomic distance on contact count probability with a spline using fit-hi-c [19]. Because of the possibility of differing statistical power between datasets, we report both the number of significant contacts above a given threshold, as well as the percent of significant contacts meeting a criterion (e.g., between *pfap2* gene loci and virulence clusters) out of all significant contacts. Significant co-localization of gene sets was assessed using previously developed tests (refs. 21, 65)

We developed ACCOST (altered chromatin conformation statistics) to estimate the statistical significance of differences in contact counts between samples, taking as

inspiration the negative binomial-based tests used for RNA-seq data [66-68]. Briefly, we modeled each bin as a negative binomial random variable, and we estimated the relationship between the mean and variance by grouping pairs of loci that are separated by the same linear genomic distance. We adapted the model employed by DESeq to Hi-C data by using an explicit specific scaling factor corresponding to bin-specific ICE biases. In addition, we estimated variance and dispersion of the negative binomial without replicates by assuming that most bins at a given genomic distance act similarly. Contact count matrices were subsampled to account for differences in interaction counts measured at the various stages and only intrachromosomal contacts for which the sum of the contacts was above the 80% percentile were tested. The resulting fold-change values were filtered after FDR estimation to only include loci that show a two-fold or larger difference in contacts, after normalizing for the effect of genomic distance, with a false discovery rate of less than 1% (Supplementary Data 3).

Code availability

Python scripts for mapping, binning, and normalization are described in the Supplementary Information and source code is available from Bitbucket (https://bitbucket.org/noblelab/plasmo-hic-2018/). ICE normalization was done using iced (https://github.com/hiclib/iced). The source code for ACCOST is available at https://github.com/cookkate/ACCOST.

Data availability

The Hi-C and ChIP-seq sequencing data that support the findings of this study have been deposited in the NCBI Sequence Read Archive with accession numbers SRP091967 and SRP091939, respectively. Fold-change heatmaps can be accessed at http://noble.gs.washington.edu/proj/plasmo3d_sexualstages/.

References

1. WHO. The World malaria report. World Health Organization http://www.who.int/malaria/publications/world-malaria-report-2017/en/ (2017).

2. Balaji, S., Babu, M. M., Iyer, L. M. & Aravind, L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. 33, 3994–4006 (2005).

3. Kafsack, B. F. et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507, 248–252 (2014).

4. Sinha, A. et al. A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. Nature 507, 253–257 (2014).

5. Poran, A. et al. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature 551, 95–99 (2017).

6. Bartfai, R. et al. H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. PLoS Pathog. 6, e1001223 (2010).

7. Lopez-Rubio, J. J., Mancio-Silva, L. & Scherf, A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe 5, 179–190 (2009).

8. Salcedo-Amaya, A. M. et al. Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of Plasmodium falciparum. Proc. Natl Acad. Sci. USA 106, 9655–9660 (2009).
9. Chookajorn, T. et al. Epigenetic memory at malaria virulence genes. Proc. Natl Acad. Sci. USA 104, 899–902 (2007).

10. Jiang, L. et al. PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. Nature 499, 223–227 (2013).

11. Ukaegbu, U. E. et al. Recruitment of PfSET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in P. falciparum. PLoS Pathog. 10, e1003854 (2014).

12. Freitas-Junior, L. H. et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature 407, 1018–1022 (2000).

13. Ralph, S. A., Scheidig-Benatar, C. & Scherf, A. Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. Proc. Natl Acad. Sci. USA 102, 5414–5419 (2005).

14. Amit-Avraham, I. et al. Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum. Proc. Natl Acad. Sci. USA 112, E982–E991 (2015).

15. Epp, C., Li, F., Howitt, C. A., Chookajorn, T. & Deitsch, K. W. Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. RNA 15, 116–127 (2009).

16. Ay, F. et al. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 24, 974–988 (2014).

17. Lemieux, J. E. et al. Genome-wide profiling of chromosome interactions in Plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. Mol. Microbiol. 90, 519–538 (2013).

18. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003 (2012).

19. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 24, 999–1011 (2014).

20. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J. P. A statistical approach for inferring the 3D structure of the genome. Bioinformatics 30, i26–i33 (2014).

21. Witten, D. M. & Noble, W. S. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. Nucleic Acids Res. 40, 3849–3855 (2012).

22. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31, 1119–1125 (2013).

23. Jiao, W. B. et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. 27, 778–786 (2017).

24. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95 (2017).

25. Lapp, S. A. et al. PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. Parasitology 145, 1–14 (2017).

26. Brancucci, N. M. et al. Heterochromatin protein 1 secures survival and transmission of malaria parasites. Cell Host Microbe 16, 165–176 (2014).

27. Stubbs, J. et al. Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes. Science 309, 1384–1387 (2005).

28. Cortes, A. et al. Epigenetic silencing of Plasmodium falciparum genes linked to erythrocyte invasion. PLoS Pathog. 3, e107 (2007).

29. Flueck, C. et al. Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog. 5, e1000569 (2009).

30. Eksi, S. et al. Plasmodium falciparum gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development. PLoS Pathog. 8, e1002964 (2012).

31. Weiner, A. et al. 3D nuclear architecture reveals coupled cell cycle dynamics of chromatin and nuclear pores in the malaria parasite Plasmodium falciparum. Cell. Microbiol. 13, 967–977 (2011).

32. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).

33. Deng, X. et al. Bipartite structure of the inactive mouse X chromosome. Genome Biol. 16, 152 (2015).

34. Giorgetti, L. et al. Structural organization of the inactive X chromosome in the mouse. Nature 535, 575–579 (2016).

35. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).

36. Ramirez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat. Commun. 9, 189 (2018).

37. Wlodarchak, N. & Xing, Y. PP2A as a master regulator of the cell cycle. Crit. Rev. Biochem. Mol. Biol. 51, 162–184 (2016).

38. Modrzynska, K. et al. A knockout screen of ApiAP2 genes reveals networks of interacting transcriptional regulators controlling the Plasmodium life cycle. Cell. Host Microbe 21, 11–22 (2017).

39. Carlton, J. M. et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 455, 757–763 (2008).

40. Prajapati, S. K. & Singh, O. P. Remodeling of human red cells infected with Plasmodium falciparum and the impact of PHIST proteins. Blood Cells Mol. Dis. 51, 195–202 (2013).

41. Oberli, A. et al. A Plasmodium falciparum PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface. FASEB J. 28, 4420–4433 (2014).

42. Oberli, A. et al. Plasmodium falciparum Plasmodium helical interspersed subtelomeric proteins contribute to cytoadherence and anchor P. falciparum erythrocyte membrane protein 1 to the host cell cytoskeleton. Cell. Microbiol. 18, 1415–1428 (2016).

43. Paulsen, J. et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. Nucleic Acids Res. 41, 5164–5174 (2013).

44. Bunnik, E. M. et al. Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum. Genome Biol. 14, R128 (2013).

45. Le Roch, K. G. et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301, 1503–1508 (2003).

46. Lopez-Barragan, M. J. et al. Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. BMC Genom. 12, 587 (2011).

47. Otto, T. D. et al. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Mol. Microbiol. 76, 12–24 (2010).

48. Bach, F. R. & Jordan, M. I. Kernel independent component analysis. J. Mach. Learn. Res. 3, 1–48 (2003).

49. Flueck, C. et al. A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. PLoS Pathog. 6, e1000784 (2010).

50. Sindikubwabo, F. et al. Modifications at K31 on the lateral surface of histone H4 contribute to genome structure and expression in apicomplexan parasites. eLife 6, e29391 (2017).

51. Lu, X. M. et al. Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite Plasmodium falciparum. Nucleic Acids Res. 45, 7825–7840 (2017).

52. Dearnley, M. et al. Reversible host cell remodeling underpins deformability changes in malaria parasite sexual blood stages. Proc. Natl Acad. Sci. USA 113, 4800–4805 (2016).

53. Tiburcio, M. et al. Early gametocytes of the malaria parasite Plasmodium falciparum specifically remodel the adhesive properties of infected erythrocyte surface. Cell. Microbiol. 15, 647–659 (2013).

54. Broadbent, K. M. et al. Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genom. 16, 454 (2015).

55. Chadwick, B. P. DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Genome Res. 18, 1259–1269 (2008).

56. Tao, D. et al. Sex-partitioning of the Plasmodium falciparum stage V gametocyte proteome provides insight into falciparum-specific cell biology. Mol. Cell. Proteom. 13, 2705–2724 (2014).

57. Lasonder, E. et al. Integrated transcriptomic and proteomic analyses of P. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. Nucleic Acids Res. 44, 6087–6101 (2016).

58. Vandomme, A. et al. Phosphotyrosyl phosphatase activator of Plasmodium falciparum: identification of its residues involved in binding to and activation of PP2A. Int. J. Mol. Sci. 15, 2431–2453 (2014).

59. Martinsen, E. S., Perkins, S. L. & Schall, J. J. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47, 261–273 (2008).

60. Hall, N. Genomic insights into the other malaria. Nat. Genet. 44, 962–963 (2012).

61. Krijger, P. H. & de Laat, W. Can we just say: transcription second? Cell 169, 184–185 (2017).

62. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012–1025 (2015).

63. Adjalley, S. H. et al. Quantitative assessment of Plasmodium falciparum sexual development reveals potent transmission-blocking activity by methylene blue. Proc. Natl Acad. Sci. USA 108, E1214–E1223 (2011).

64. Guttery, D. S. et al. A putative homologue of CDC20/CDH1 in the malaria parasite is essential for male gamete development. PLoS Pathog. 8, e1002554 (2012).

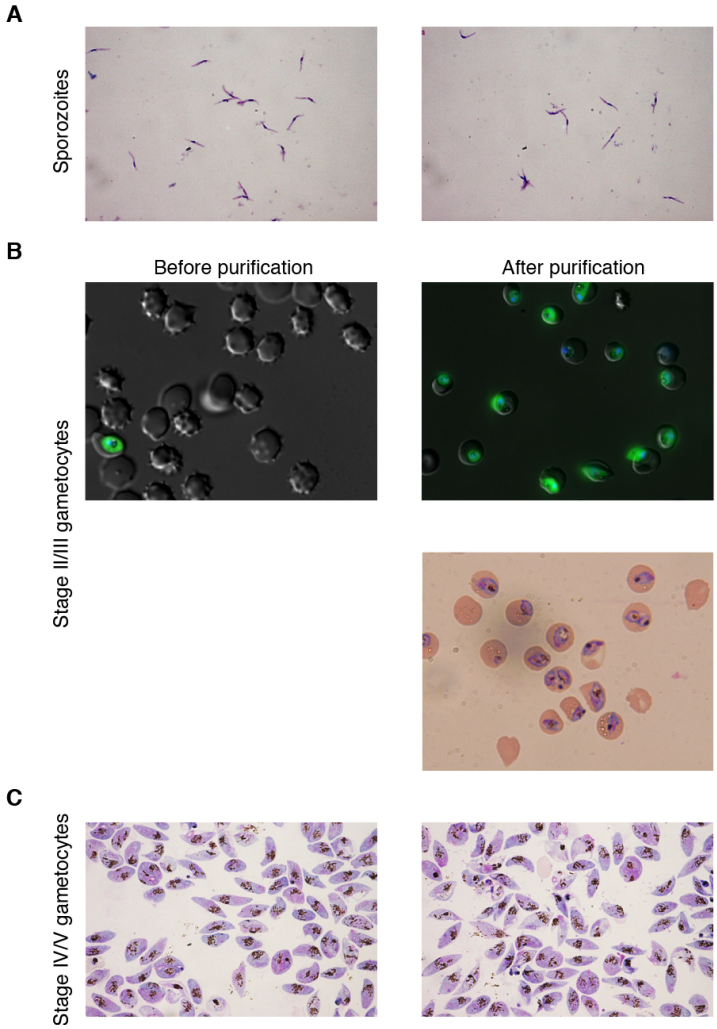
65. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. 43, 1059–1065 (2011).

66. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).

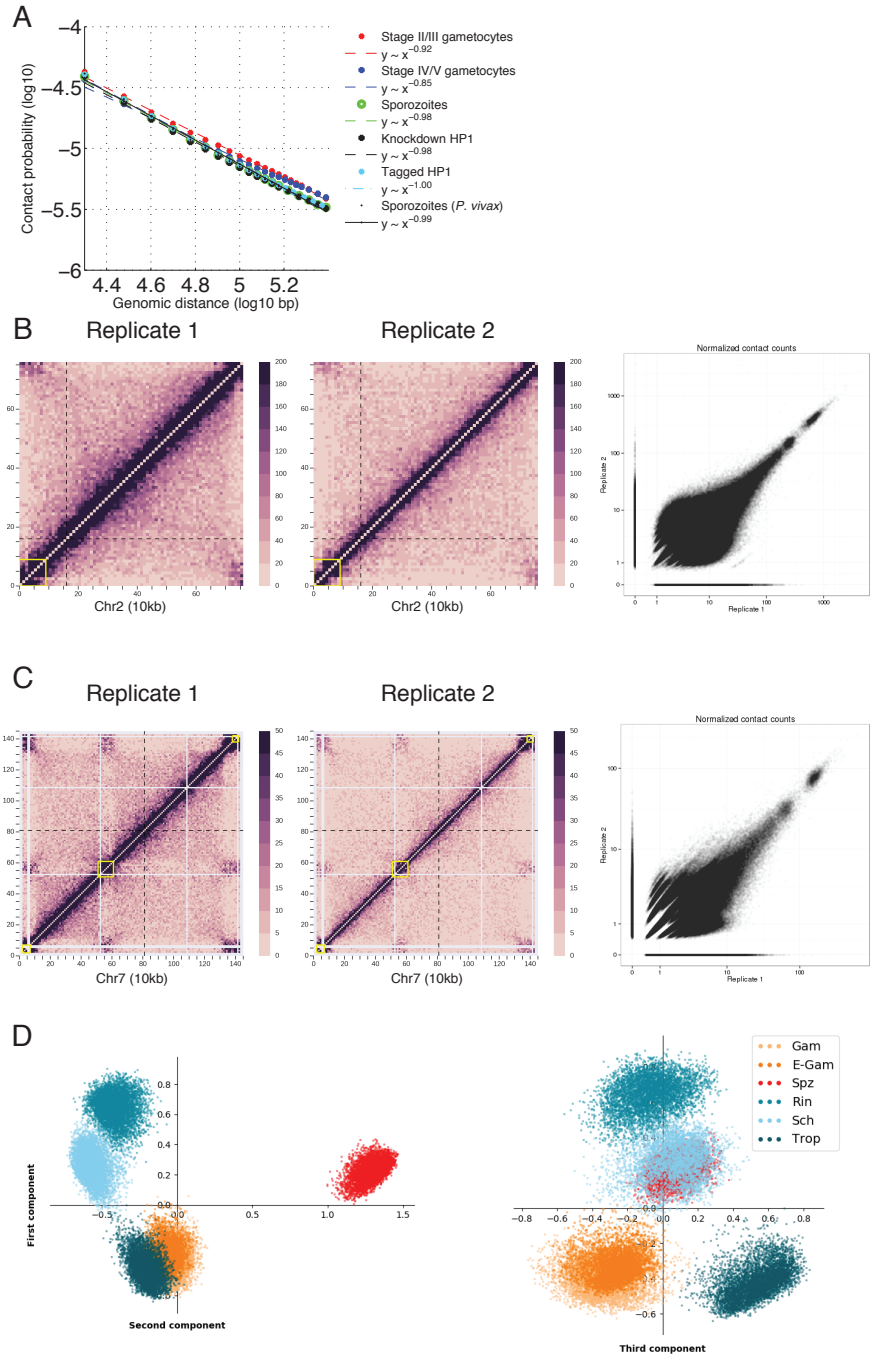
67. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

68. Drewe, P. et al. Accurate detection of differential RNA processing. Nucleic Acids Res. 41, 5189–5198 (2013).

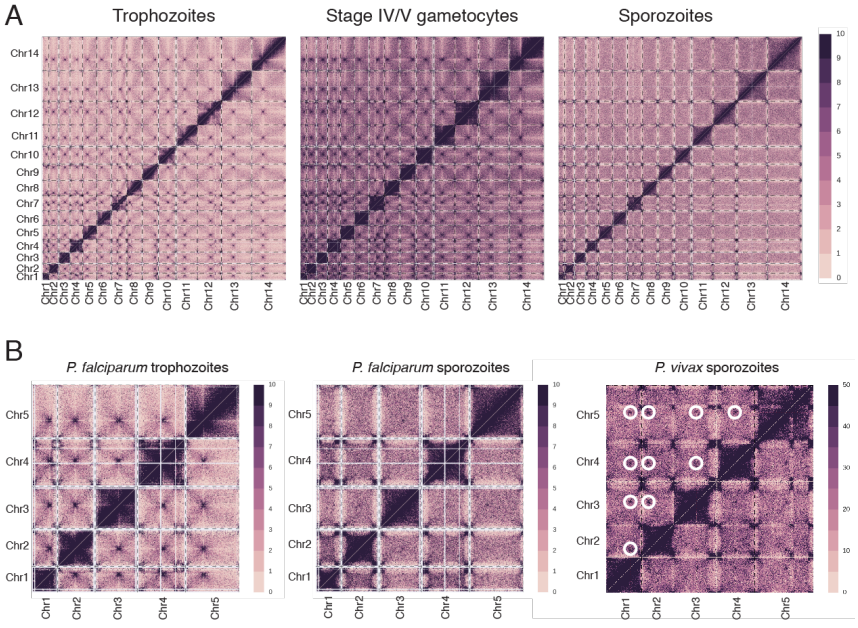
Supplemental Material



Supplemental figure 1.1: Microscopy images of the transmission stages analyzed in this study. (A) Two representative images of Giemsa-stained *P. falciparum* salivary gland sporozoites. (B) GFP-expressing stage II/III gametocytes before (left) and after (right) purification visualized by fluorescent microscopy (top) or transmitted light microscopy of Giemsa-stained parasites (bottom). (C) Two representative images of Giemsa-stained stage IV/V gametocytes after percoll gradient and magnetic purification.


A
Contact probability (log10)
Genomic distance (log10 bp)
Stage II/III gametocytes
y ~ x^-0.92
Stage IV/V gametocytes
y ~ x^-0.85
Sporozoites
y ~ x^-0.98
Knockdown HP1
y ~ x^-0.98
Tagged HP1
y ~ x^-1.00
Sporozoites (P. vivax)
y ~ x^-0.99
B
Replicate 1
Replicate 2
Chr2 (10kb)
Chr2 (10kb)
Normalized contact counts
Replicate 1
Replicate 2
C
Replicate 1
Replicate 2
Chr7 (10kb)
Chr7 (10kb)
Normalized contact counts
Replicate 1
Replicate 2
D
First component
Second component
Third component
Gam
E-Gam
Spz
Rin
Sch
Trop

Supplemental Figure 1.2: Quality measures of Hi-C libraries. (A) Log-linear relationship between contact probability and genomic distance in all Hi-C libraries generated in this study. (B) ICE-normalized contact count matrices of chromosome 2 for two biological replicates of *P. vivax* sporozoites (left and middle) and ICE-normalized contact count scatter plot for these replicates (right). (C) ICE-normalized contact count matrices of chromosome 7 for two biological replicates of *P. falciparum* sporozoites (left and middle) and ICE-normalized contact count scatter plot for these replicates (right). (D) Principal component analysis on 5,000 3D genome structures per stage generated from varying initial starting points. The first and second components are plotted on the left, while the first and third components are plotted on the right.

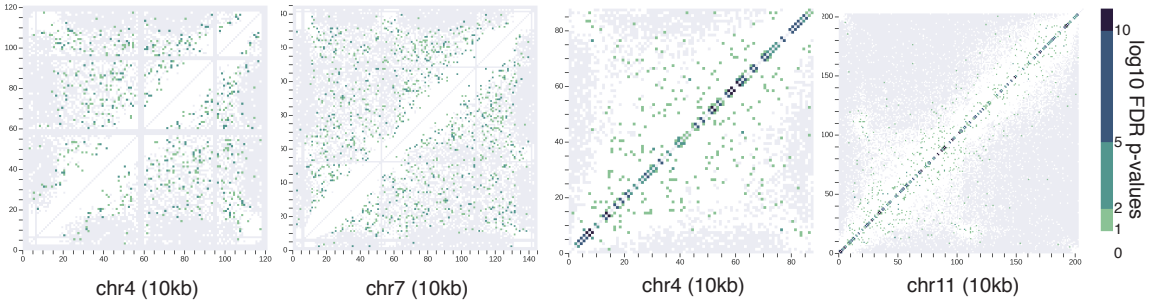


Supplemental Figure 1.3: Differences in genome organization between *Plasmodium* species and life cycle stages. (A) ICE-normalized interchromosomal contact count heatmaps at 10 kb resolution for *P. falciparum* trophozoites, stage IV/V gametocytes and sporozoites. Dashed lines indicate chromosome boundaries. (B) Interchromosomal contact count heatmaps of chromosomes 1-5 for *P. falciparum* trophozoites (left), showing strong co-localization of centromeres, *P. falciparum* sporozoites (center) with absent centromere co-localization, and *P. vivax* sporozoites (right), whose centromeres colocalize, although these interactions do not involve regions adjacent to the centromere. In all heatmaps, dashed black lines indicate chromosome boundaries, and white circles in the *P. vivax* sporozoite matrix are used to highlight inter-centromere contacts.

A

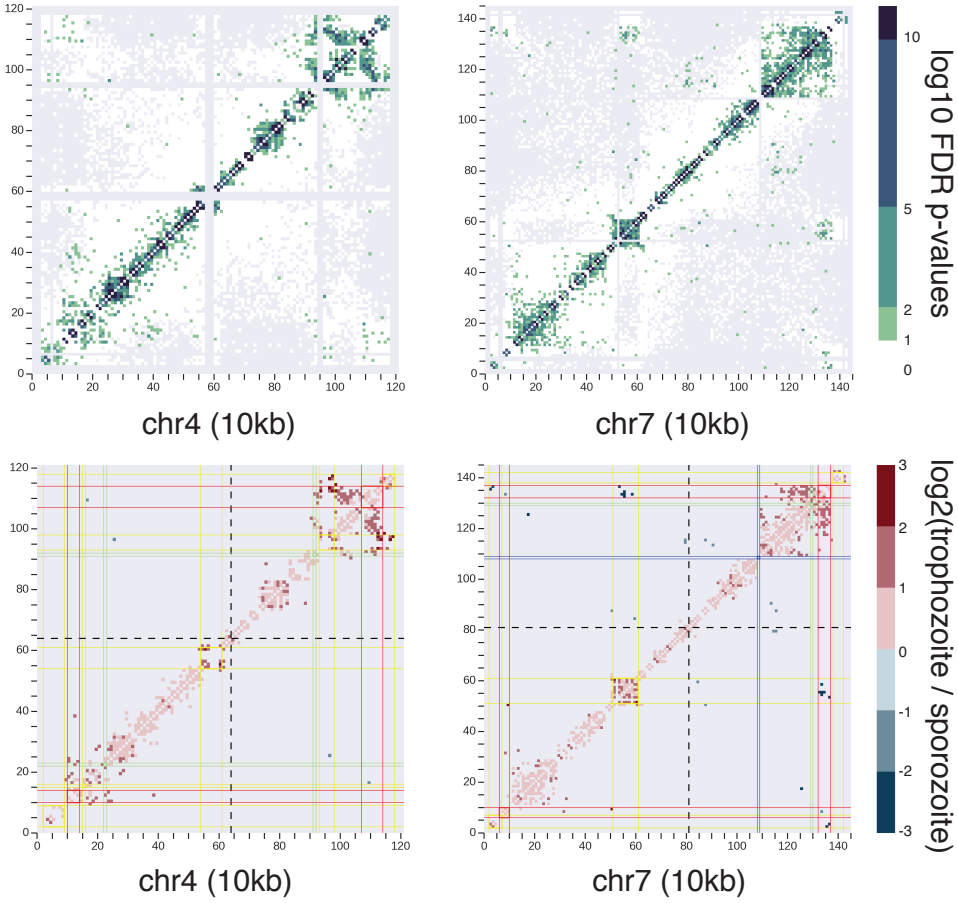
P. falciparum sporozoite replicates

P. vivax sporozoite replicates

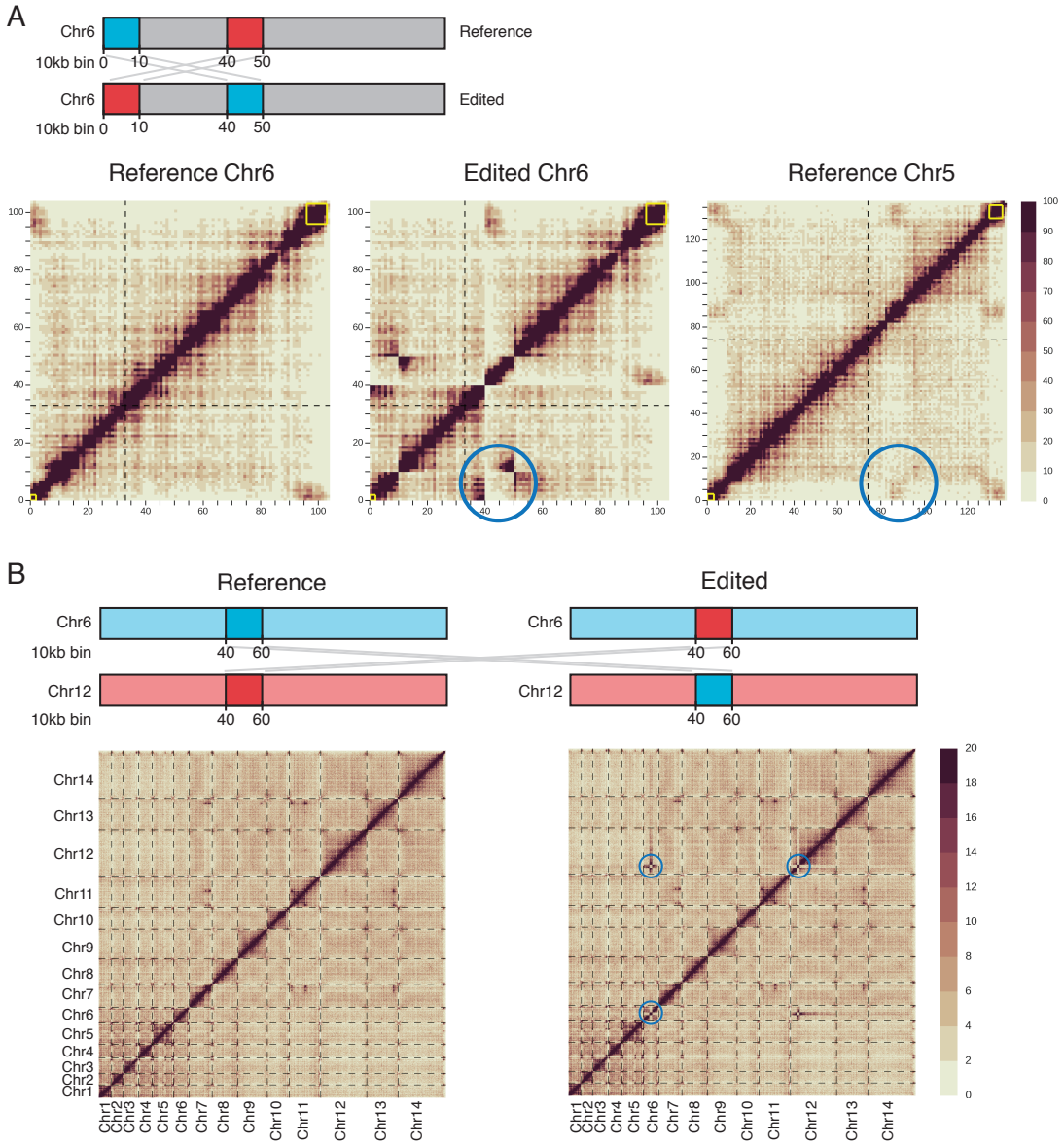


B

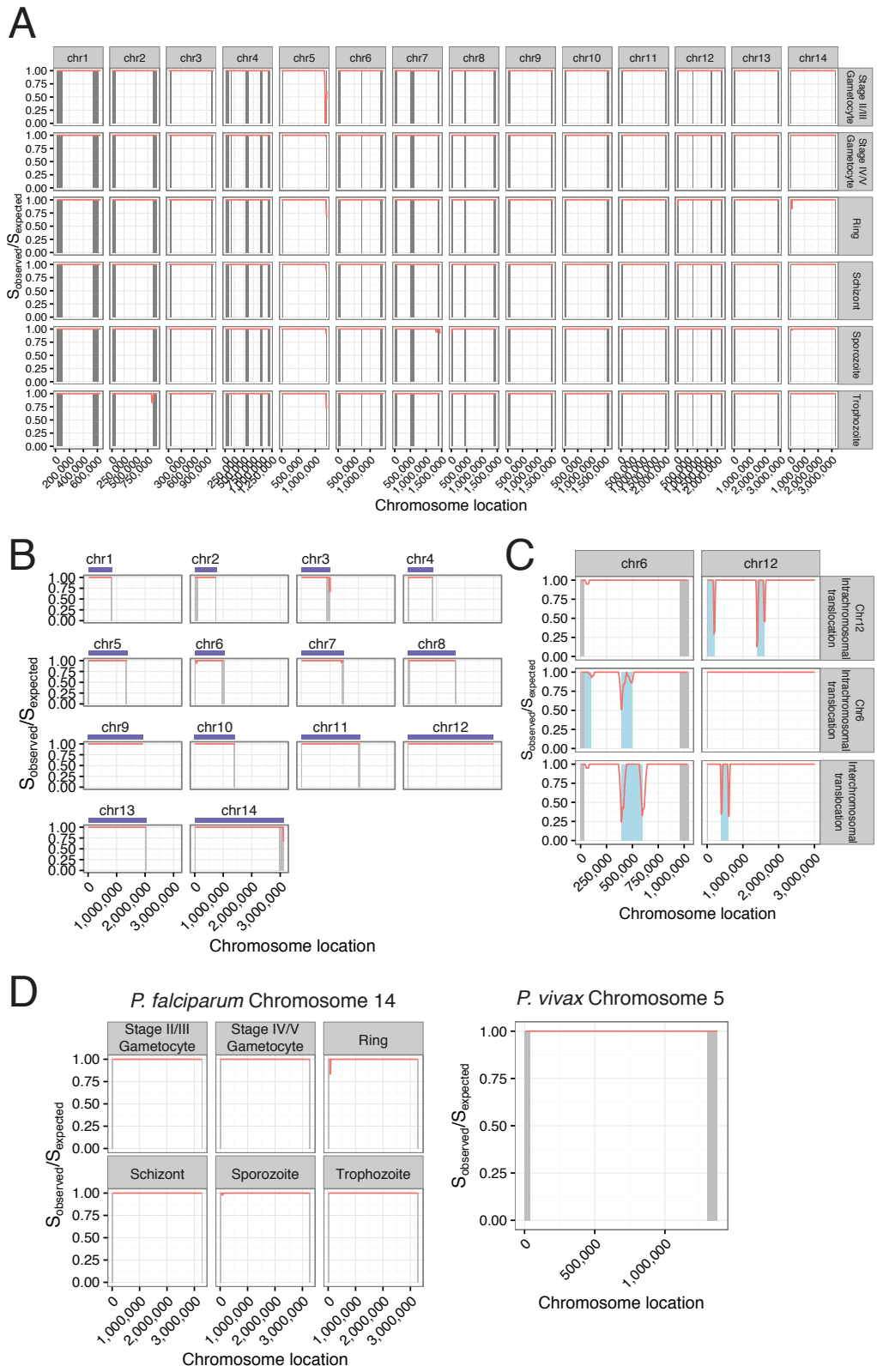
P. falciparum sporozoites vs trophozoites



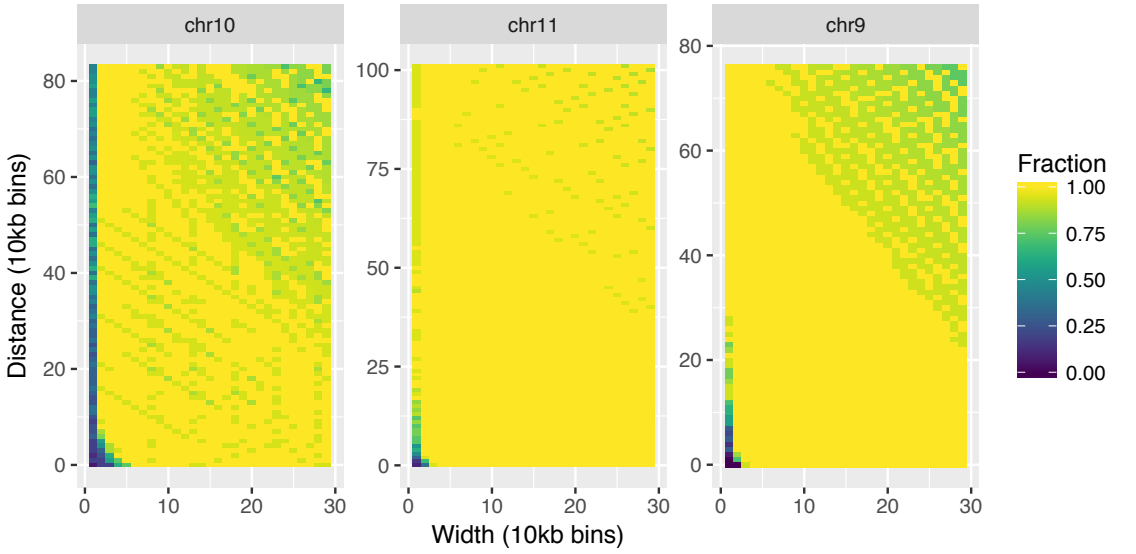
Supplemental Figure 1.4: Similarities in genome organization between biological replicates as compared to differences in genome organization between different parasite stages. (A) Heatmaps of differences in interactions between biological replicates of the sporozoite stage for *P. falciparum* (left panels for two different chromosomes) and *P. vivax* (right panels for two different chromosomes). The heatmaps are color-coded based on FDR p-values. No large-scale differences in interactions can be observed between biological replicates (compare: P. falciparum sporozoites vs trophozoites in panel B). Greyed out regions are bins that were filtered out for poor mappability, or loci for which the sum of the contact counts between the two samples was below the 80th percentile. (B) Heatmaps of all significantly changing interactions between trophozoites and sporozoites for chromosomes 4 (left) and 7 (right). The top panels show the FDR p-values. In the bottom panels, all 10 kb bins that differ significantly (at 1% FDR) in the number of interactions between the two stages are shown and are color-coded according to the direction of the change: loci with stronger interactions in trophozoites are indicated in red and in sporozoites in blue. Locations of genes of interest are bordered with color coded lines: virulence gene clusters are indicated in yellow, subtelomeric clusters of genes encoding exported proteins in red, ApiAP2 TF loci in green, and rDNA genes in blue. Centromeres are denoted by black dotted lines.


A
Chr6
10kb bin 0
10
40
50
Reference
Chr6
10kb bin 0
10
40
50
Edited
Reference Chr6
Edited Chr6
Reference Chr5
B
Reference
Edited
Chr6
10kb bin
40
60
Chr12
10kb bin
40
60
Chr6
40
60
Chr12
40
60
Chr14
Chr13
Chr12
Chr11
Chr10
Chr9
Chr8
Chr7
Chr6
Chr5
Chr4
Chr3
Chr2
Chr1
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14

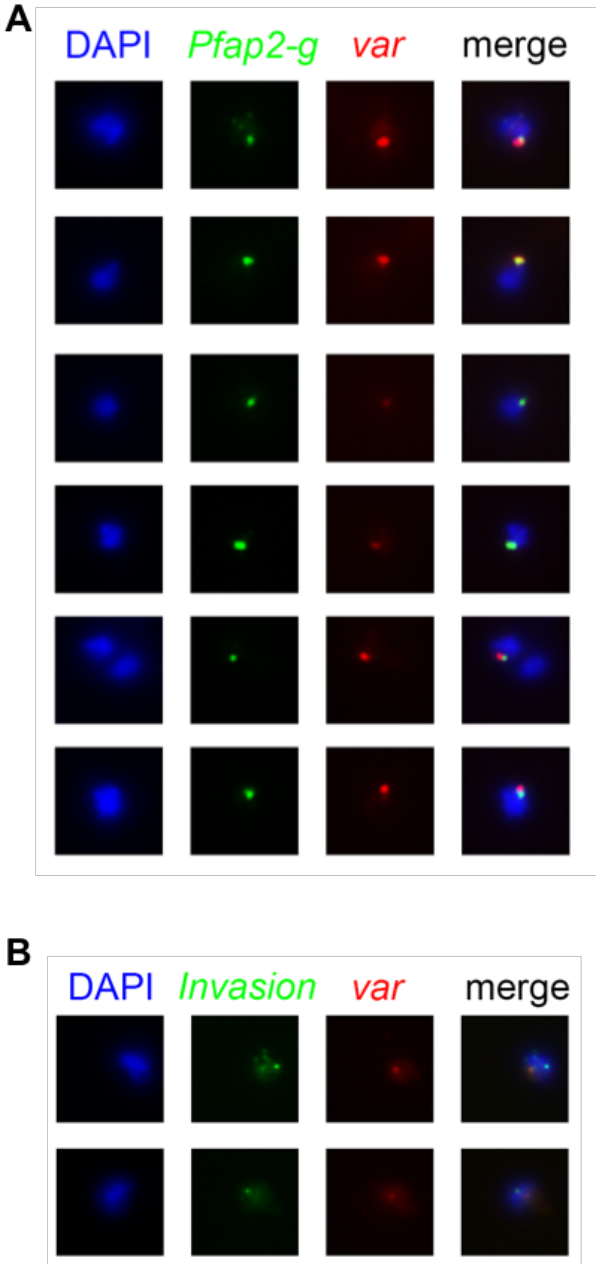
Supplemental Figure 1.5: Introduced translocations in the *P. vivax* genome. (A) A 100 kb intrachromosomal translocation in chromosome 6. A schematic representation of the introduced translocation is shown in the top. The bottom row shows three raw contact count heatmaps: observed contacts for chr6 (left), chr6 with introduced translocation (middle), and observed contacts for chr5 (right). (B) A 200 kb interchromosomal translocation between chromosome 6 and chromosome 12. A schematic representation of the introduced translocation is shown in the top. The bottom row shows the observed interchromosomal contacts (left) and the interchromosomal contact count heatmap after introduction of the translocation (right). These translocations produce aberrant signals in the contact count heatmaps (indicated with blue circles) that were not observed in any of the samples that were generated for this study. This supports our conclusion that the observed changes between life cycle stages are unlikely to be artifacts caused by genomic recombination.


A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
Stage II/III Gametocyte
Stage IV/V Gametocyte
Ring
Schizont
Sporozoite
Trophozoite
Sobserved/Sexpected
Chromosome location
B
C
Chr12 Intrachromosomal translocation
Chr6 Intrachromosomal translocation
Interchromosomal translocation
D
P. falciparum Chromosome 14
P. vivax Chromosome 5

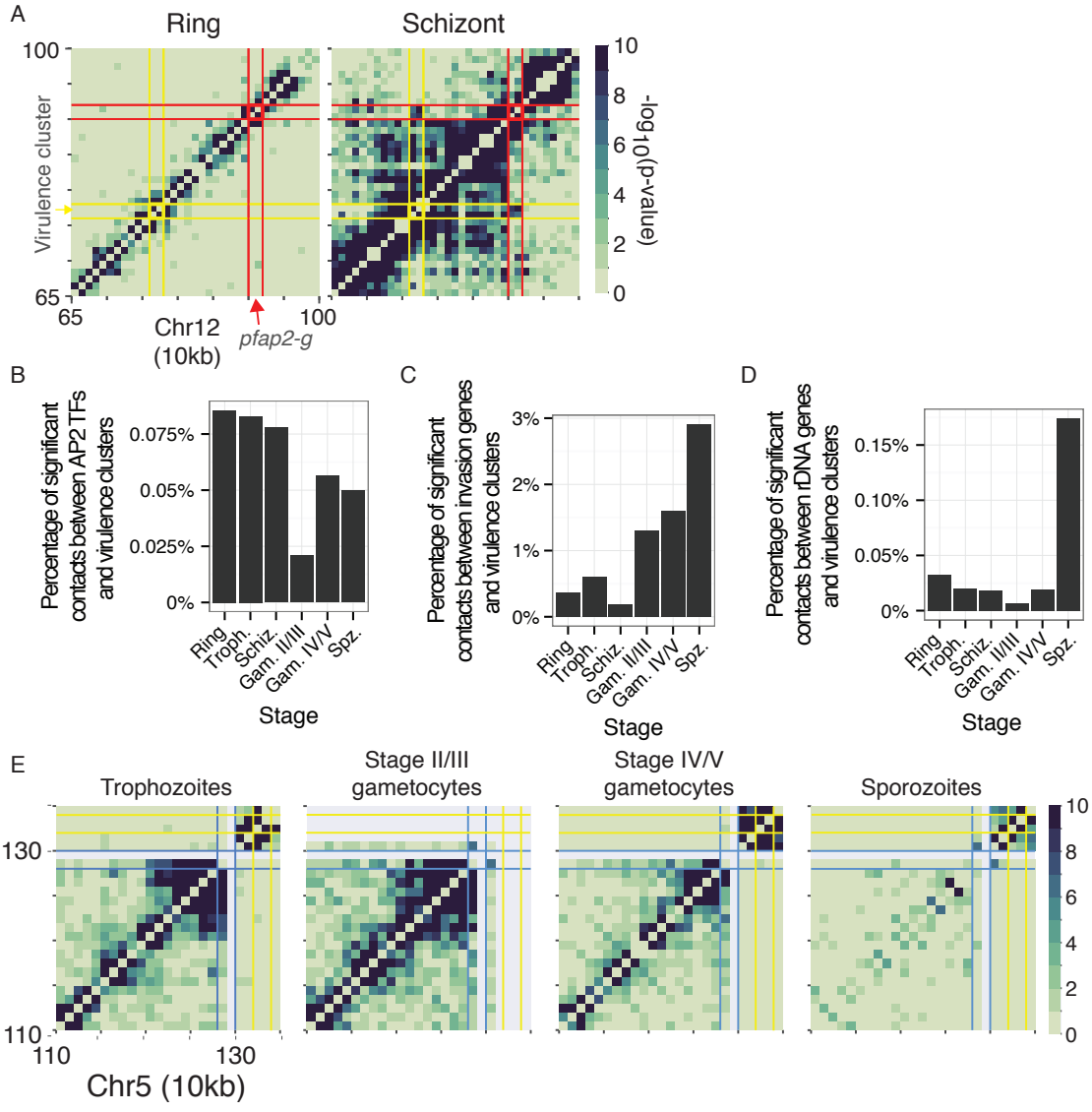
Supplemental Figure 1.6: Misassembly metric for *P. falciparum* and *P. vivax*. (A) The misassembly metric $S_{observed}/S_{expected}$ for all chromosomes of *P. falciparum* life cycle stages included in this study. Virulence clusters are shaded as grey bars. The region chr5:1,310,000-1,343,557 is likely deleted in the *P. falciparum* NF54^{Pfs16} strain used to prepare stage II/III gametocytes, explaining the drop in the misassembly metric at the right telomere of chr5 in this sample. (B) The misassembly metric $S_{observed}/S_{expected}$ for all chromosomes in *P. vivax* sporozoites. Grey bars indicate the VIR gene clusters, and the purple bar at the top indicates the extent of the chromosome. (C) Misassembly metric on *P. vivax* data with simulated translocations. Clusters of VIR genes at the telomeres are shaded in grey, and the introduced translocations are shown in blue. Top row: Intrachromosomal translocation between 0-20,000 bp and 140,000-160,000 of chromosome 12. Middle row: Intrachromosomal translocation between 0-10,000 bp and 40,000-50,000 of chromosome 6. Bottom row: Interchromosomal translocation between 40,000-60,000 bp of chromosome 6 and 40,000-60,000 of chromosome 12. (D) The same data as (A) and (B), focusing on chromosome 14 of *P. falciparum* (left) and chromosome 5 of *P. vivax* (right).



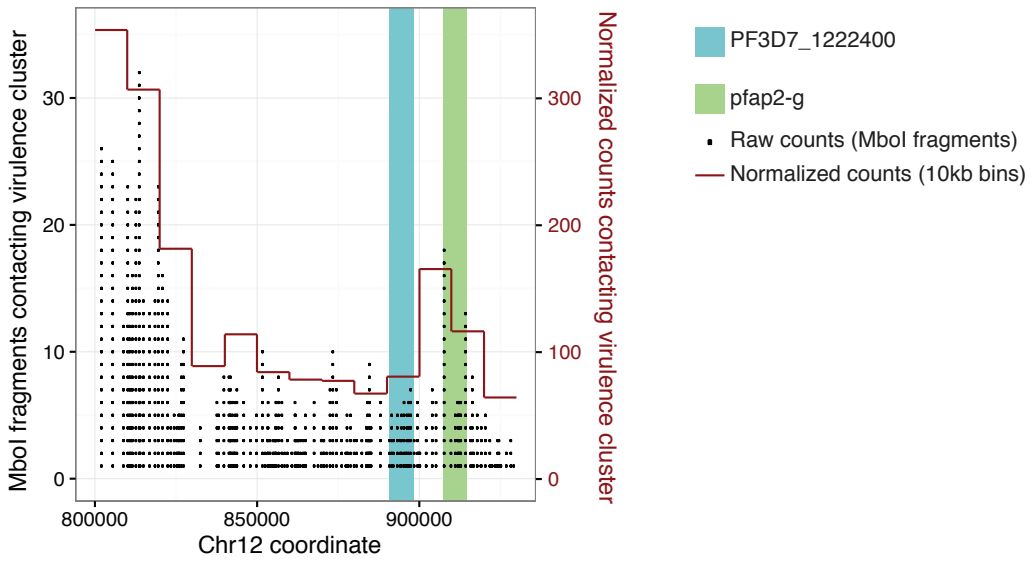
Supplemental Figure 1.7: Detection of simulated translocations using the misassembly metric. For each of three chromosomes in *Plasmodium falciparum* (chr9, chr10, and chr11), we generated translocations of widths down to a single 10kb bin, and with interchromosomal distances between the translocated regions down to zero bins (ie, right next to each other). Simulations were performed at locations across the chromosome (n = 28,571 for chr9, n = 35,647 for chr10, and n = 52,873 for chr11). Using the misassembly metric $S_{observed}/S_{expected}$, we examined trophozoite Hi-C data across those chromosomes, and tallied the frequency of cases where the lowest value of the misassembly metric corresponded to the edge of one of the translocated regions. These frequencies were plotted as colors against the widths and distances tested.



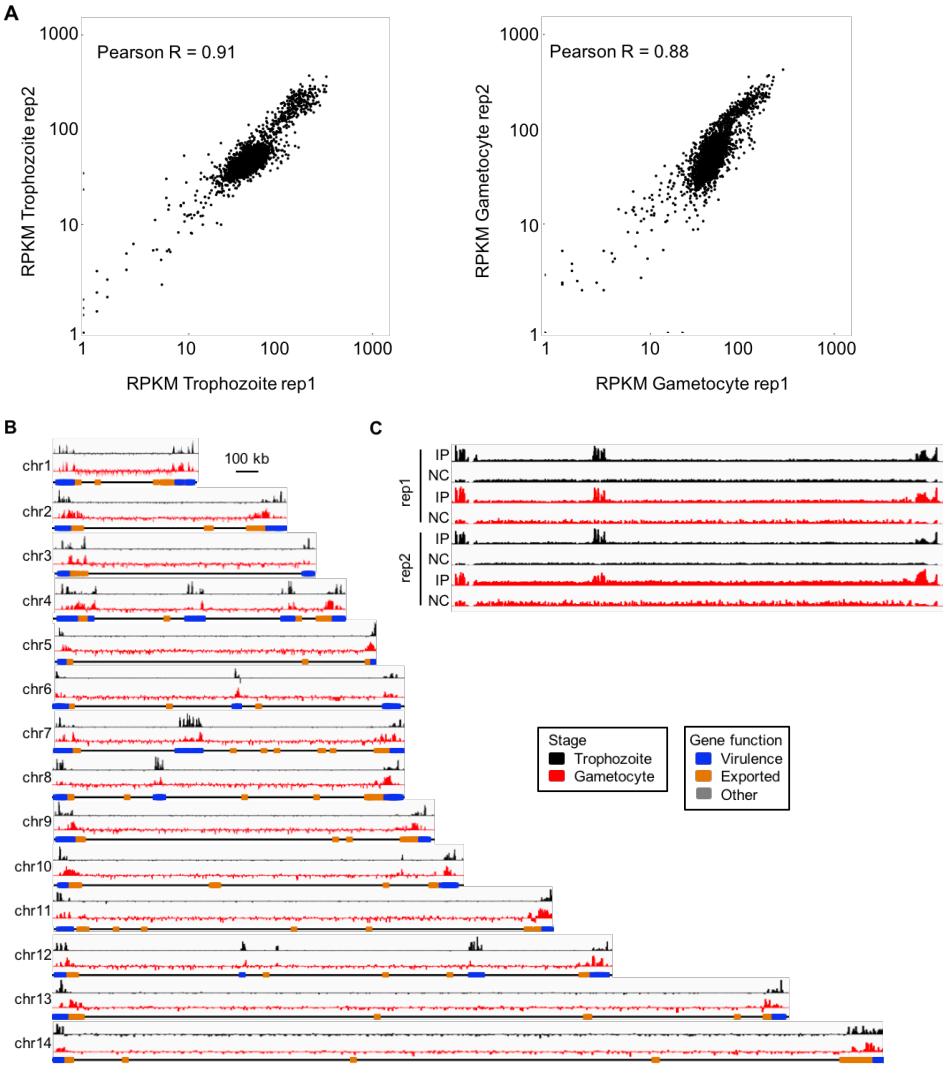
Supplemental Figure 1.8: DNA-FISH experiments in ring stage parasites. (A) Colocalization of *pfap2-g* and *var* gene PF3D7_0800300. Images are representative of visual inspection of >100 ring stage parasites. (B) Non-colocalizaton of invasion gene GLURP (chr10:1,399,195 – 1,402,896) and *var* gene PF3D7_0800300.


A
Ring
Schizont
100
65
Virulence cluster
65
100
Chr12
(10kb)
pfap2-g
10
8
6
4
2
0
-log10(p-value)
B
Percentage of significant contacts between AP2 TFs and virulence clusters
0.075%
0.05%
0.025%
0%
Ring
Troph.
Schiz.
Gam. II/III
Gam. IV/V
Spz.
Stage
C
Percentage of significant contacts between invasion genes and virulence clusters
3%
2%
1%
0%
Ring
Troph.
Schiz.
Gam. II/III
Gam. IV/V
Spz.
Stage
D
Percentage of significant contacts between rDNA genes and virulence clusters
0.15%
0.10%
0.05%
0%
Ring
Troph.
Schiz.
Gam. II/III
Gam. IV/V
Spz.
Stage
E
Trophozoites
Stage II/III gametocytes
Stage IV/V gametocytes
Sporozoites
130
110
110
130
Chr5 (10kb)
10
8
6
4
2
0

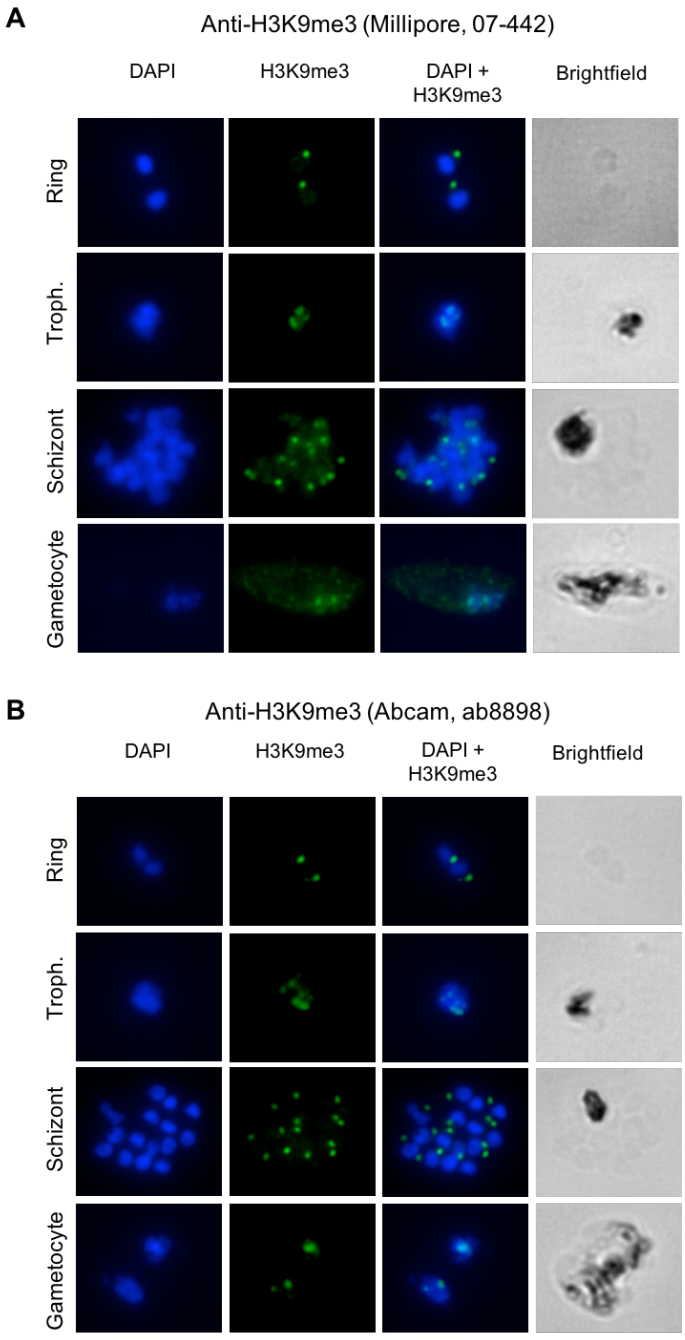
Supplemental Figure 1.9: Interaction of ApiAP2 TF genes and invasion genes with the repressive center. (A) Significant interactions between *pfap2-g* and the nearby internal *var* gene locus were also observed at the schizont stage (right panel) but not in the ring stage (left panel). A possible explanation for this observation is that the total number of significant interactions at the ring stage (n=16,705) is lower than at the trophozoite stage (n=25,457) and the schizont stage (n=160,176). The early gametocyte stage has a large number of significant interactions (n=209,345) and it is very unlikely that the absence of significant interactions between *pfap2-g* and *var* in this stage is caused by a lack of depth in the data. (B-D) Significant interactions between virulence genes and (B) *pfap2* genes, (C) invasion genes, and (D) rDNA genes are expressed as a percentage of all significant interactions within the genome. (E) Loss of domain formation around the rDNA locus on chr5 in *P. falciparum* sporozoites as compared to other life cycle stages. The borders of the rDNA locus are indicated by blue lines.



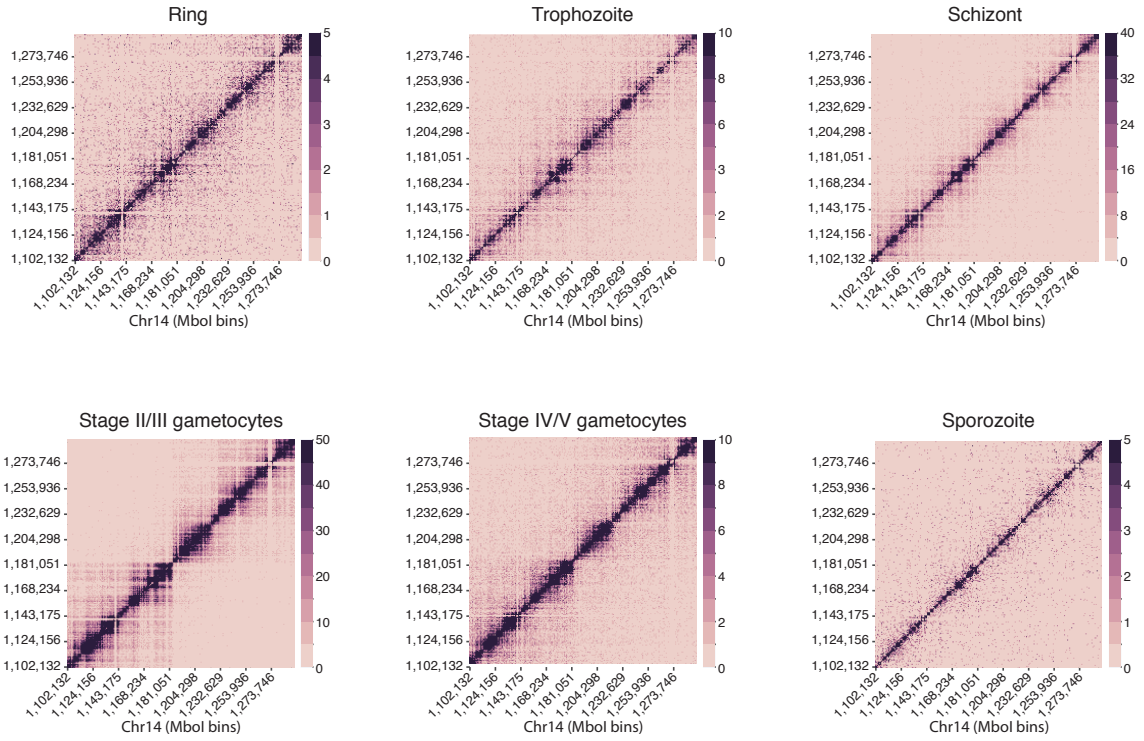
Supplemental Figure 1.10: Restriction site resolution virtual 4C of chromosome 12 in trophozoites. The chr12:764448-784830 virulence cluster was used as "bait" to collect reads mapping to individual MboI bins. Black dots represent individual reads and the left-hand axis represents the number of raw counts. The 10kb resolution ICE-normalized counts are plotted as red bars (each bar = one 10kb bin) with the right-hand axis representing the normalized contact count. The blue shaded area is the PF3D7_1222400 gene; the green area is *pfap2-g* (PF3D7_1222600).



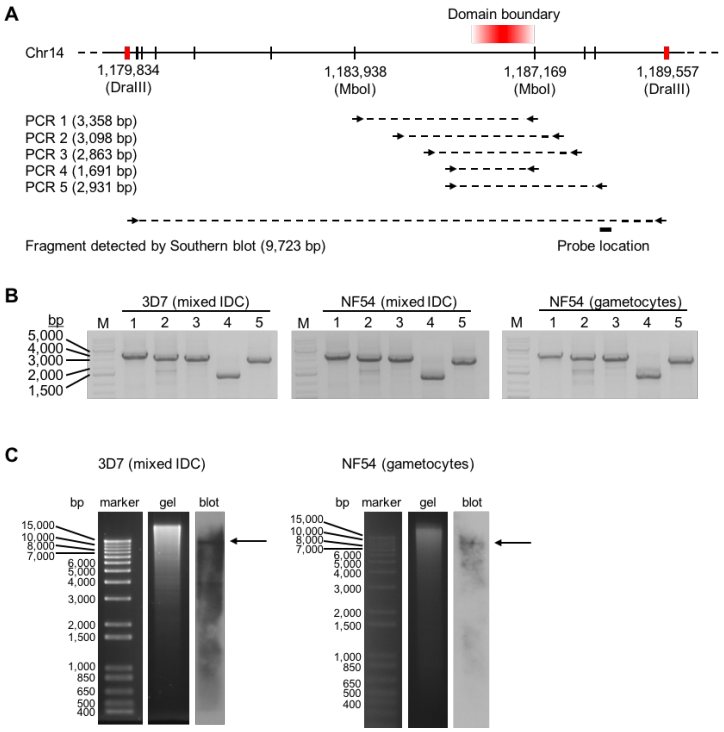
Supplemental Figure 1.11: Quality measures of H3K9me3 ChIP-seq libraries. (A) RPKM scatter plots for two biological replicates of *P. falciparum* late ring/early trophozoite stage (left) and gametocyte stage (right). (B) H3K9me3 ChIP-seq genome browser tracks of biological replicate 2 for trophozoite (top tracks in black) and stage IV/V gametocytes (bottom tracks in red). Results are similar to the tracks of biological replicate 1 presented in Figure 3A. (C) Raw read coverage of chromosome 8 for samples (IP) and negative controls (NC).



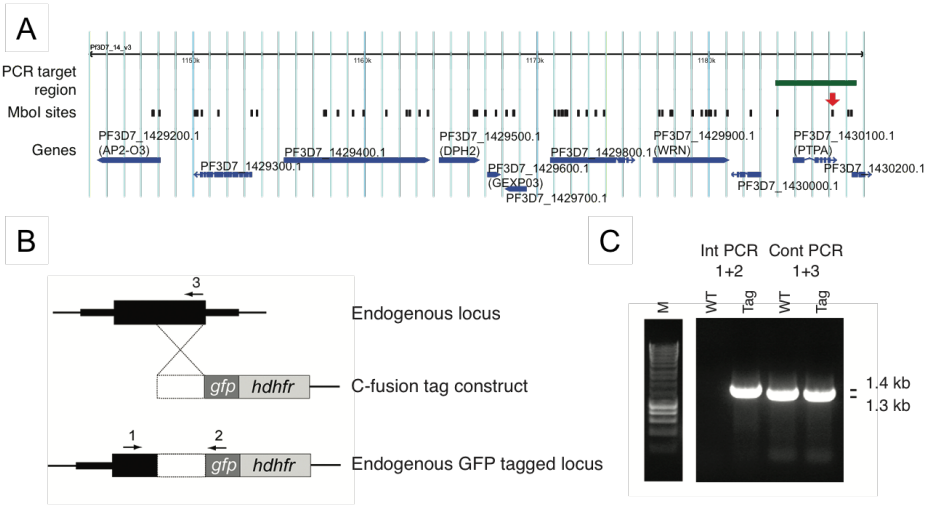
Supplemental Figure 1.12: H3K9me3 immunofluorescence analysis in IDC and gametocyte stages. The results of these experiments are highly comparable for anti-H3K9me3 antibodies Millipore 07-442 (A) and Abcam ab8898 (B).



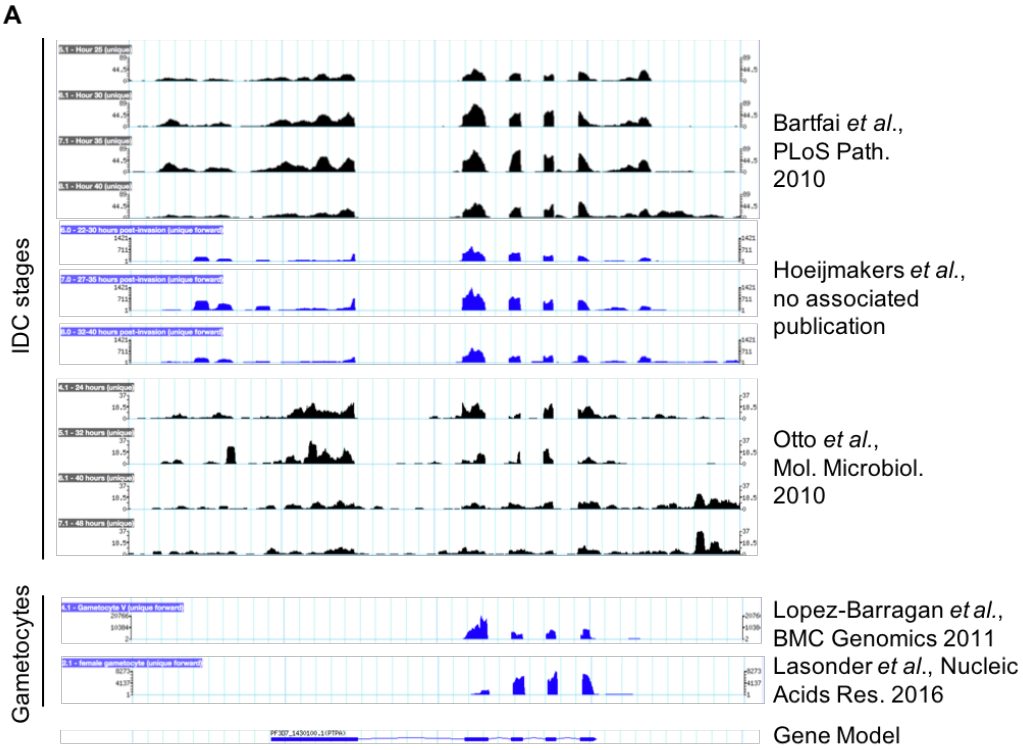
Supplementary Figure 1.13: MboI restriction site resolution contact count heatmaps of the region surrounding the location of the domain boundary in chr14. The domain boundary can be observed in both stage II/III and stage IV/V gametocytes, but is absent in other stages of the *P. falciparum* life cycle that were analyzed in this study.



Supplemental Figure 1.14: Confirmation of integrity of chromosome 14 around the domain boundary. (A) Schematic overview of PCR and Southern blot experiments. (B) PCR amplifications of genomic DNA in the region chr14:1,183,922-1,188,519 isolated from *P. falciparum* strains 3D7 (DNA isolated from mixed IDC stages) and NF54 (DNA isolated from mixed IDC stages and gametocyte stage), which spans the region containing the domain boundary. Both strains show bands of the expected size, demonstrating that chromosome 14 is intact in both parasite strains. (C) Detection of the 9,723 bp fragment released by DraIII digestion in both mixed IDC stage parasites (strain 3D7) and gametocyte stage parasites (strain NF54), further confirming that chromosome 14 is intact in both stages of the parasite life cycle.



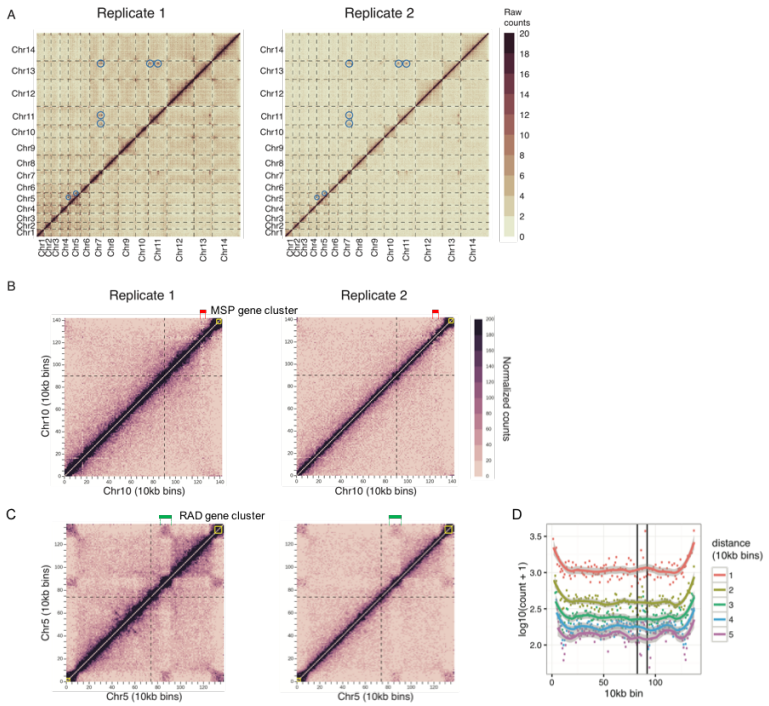
Supplemental Figure 1.15: Investigating the role of *pfap2-o3* on chromosome 14. (A) Gene model in the region around the location of the domain boundary on chromosome 14, close to the MboI restriction site at 1,187,169. *Pfap2-o3* (PF3D7_1429200) is located approximately 40 kb upstream of the domain boundary (red arrow). The region amplified by PCR as shown in panel A is indicated with a green bar. (B) Construction of transgenic *pbap2-o3-gfp P. berghei* parasites. Shown are schematic representations of the endogenous PBANKA_1015500 (*pbap2- o3*) locus (top), the GFP-tagging construct (middle) and the recombined PBANKA_1015500 locus following single cross-over recombination (bottom). Arrows 1 and 2 indicate PCR primers INT T219 and ol492 used to confirm successful integration in the PBANKA_1015500 locus following recombination. Arrows 1 and 3 indicate PCR primers INT T219 and T2192 used to control for the presence of the full-length template. (C) Results of the control PCRs on integration (left two lanes) and template integrity (right two lanes) of the PBANKA_1015500 locus in wild type (WT) and *pbap2-o3-gfp* (Tag) parasites.



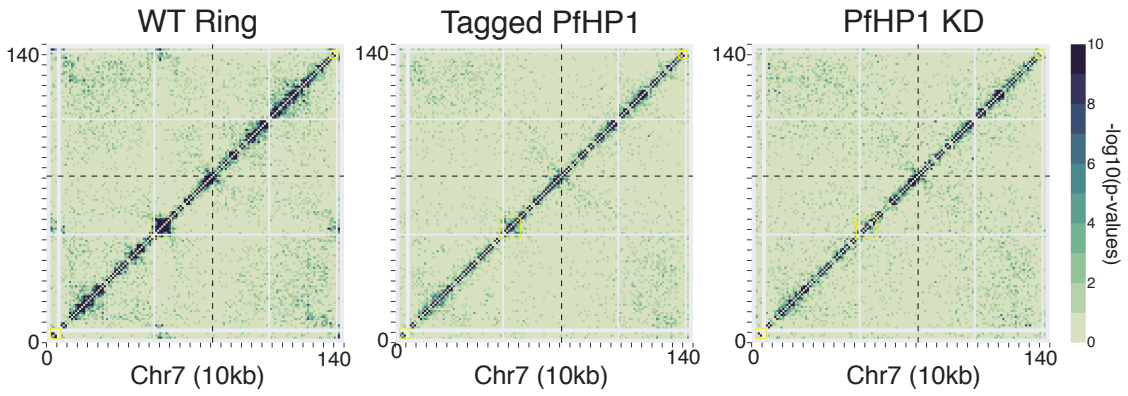
B

```
>PF3D7_1430100_INTRON1
GTACGTTTTTATTTATATATGAGTATATATTTGCATGTATATACACTTATGTTTTAATGCACATACACTT
ATACTTTAATGTACATACACTTATAGTTATGTATACTTGTATTGTAGTGTACATACACTTATGTTTTAAT
GTACATACACTTATATTTTAATGTACATACACTTATTTTAATATATGTATACTTGTATTGTAGTGTAT
ATACACTTATATTTTAATGTACATACACTTATATTTTAATGAACATACACTTATATTTTAATGTATATAC
ACCTACATTTTAATGTACATACACTTATAATTATGTATACTTGTATTGTAGTGTATGTACACTTATGTTT
TAAAGTATGTATACTTATATTTTAATGTAAGTACACTTGTATTTTAGTGTATATACAATATGTTTTAATG
TATGTATAGTTGTATTTTATTGTATGTACACTTAAGCTTTAATGTACATACACTAATGCCTTAATGTACA
TGCACTTAAGCTTTAATGTACATACACTAATGCCTTAATGTATATGCATTTATGCTTTAATGTACATACA
CTTATATTTTAATGTATGTGTGCTTGTGTTTTAATGTACATACACTTATGTTTTAATGTAAGTATATCTT
GGCTACATATATATATATATATATTTTTTTTTTTTTTCTTGTTTTTCAAAAATTTGCATGAACTG
TTCAG
```

Supplemental Figure 1.16: Expression and sequence analysis of PF3D7_1430100 (PTPA). (A) RNA-seq genome browser tracks from various publicly available data sets (obtained from PlasmoDB) showing the absence of exon 1 in the transcript expressed in gametocytes, while the variant expressed at the IDC stages contains all 5 exons. (B) The sequence of PF3D7_1430100 intron 1. A motif that is repeated a total of 12 times is highlighted in bold. Additional motifs can be found within the sequence.



Supplemental Figure 1.17: Characteristics of genome organization in *P. vivax* salivary gland sporozoites. (A) Strong interchromosomal contacts. (B) *P. vivax* contains a cluster of merozoite surface proteins (MSP, i.e. invasion genes) on chromosome 10 (1,225,905 - 1,264,358, red bar) that does not interact with any other loci on chr10. (C) A subset of Pv-fam-e genes on *P. vivax* chromosome 5 (825,000-920,000, green bar) interacts with (sub)-telomeric regions of chr5 and other chromosomes, similar to the behavior of internal *var* gene clusters in *P. falciparum*. (D) Absence of a translocation event at the RAD gene locus on chr5 in *P. vivax*. Shown are the number of contacts between each 10 kb bin and its neighboring 10 kb bins up to 5 bins distance. A translocation event would show a breakpoint in the plot, instead of the smooth curves that are observed. Vertical lines indicate the borders of the RAD gene locus.



Supplemental Figure 1.18: Tagging and depletion of PfHP1 result in a loss of *var* gene interactions. Fit-hi-c P-value matrices of chromosome 7 are shown for the wild-type *P. falciparum* ring stage, the ring-stage PfHP1-GFP-DD strain cultured in the presence of Shield-1 (tagged PfHP1) and the ring-stage PfHP1-DD-GFP strain cultured in the absence of Shield-1 (knockdown [KD] PfHP1). P-values were calculated using subsampled data that contained the same number of interactions for each condition. Note the loss of interactions between the internal *var* gene cluster and the subtelomeric *var* genes in both the tagged and the knockdown PfHP1 strain.

Supplemental Table 1.1: Number of contact counts after mapping and filtering out interactions between loci that are less than 1 kb apart.

Organism	Library	Contact counts
P. falciparum	Ring[1]	8,980,937
P. falciparum	Trophozoite[1]	6,540,198
P. falciparum	Schizont[1]	31,093,712
P. falciparum	Stage II/III gametocytes	55,427,998
P. falciparum	Stage IV/V gametocytes	14,171,256
P. falciparum	Sporozoites (combined)	7,097,155
P. falciparum	PfHP1-tagged strain	37,362,915
P. falciparum	PfHP1-knockdown strain	7,115,059
P. vivax	Sporozoites (combined)	15,931,694

[1]Libraries were generated as part of a previously published study (Ay *et al*., Genome Research, 2014).

Supplemental Table 1.2: Interchromosomal contact probability (ICP) and percentage of long- range contacts (PLRC) values for each Hi-C library generated in this study.

Organism	Library	ICP[1]	PLRC[2]
P. falciparum	Stage II/III gametocytes	0.45	32.39%
P. falciparum	Stage IV/V gametocytes	0.87	7.47%
P. falciparum	Sporozoites (replicate 1)	1.44	5.05%
P. falciparum	Sporozoites (replicate 2)	1.45	5.11%
P. falciparum	Sporozoites (combined)	1.44	5.07%
P. falciparum	PfHP1-tagged strain	1.64	10.63%
P. falciparum	PfHP1-knockdown strain	1.64	3.08%
P. vivax	Sporozoites (replicate 1)	1.75	5.62%
P. vivax	Sporozoites (replicate 2)	1.66	5.24%
P. vivax	Sporozoites (combined)	1.72	5.47%
P. falciparum[3]	**Trophozoite (not cross-linked)**	**7.82**	n.a.

1 ICP (inter-chromosomal contact probability index) is defined as (the number of inter-chromosomal interactions) / (the number of intra-chromosomal interactions above 1 kb distance).

2 Percentage of long-range contacts is calculated as (the number of interchromosomal contacts + the number of intrachromosomal contacts over 20 kb distance) / the number of reads mapped to the *Plasmodium* genome.

3 Control library generated without the formaldehyde cross-linking step of the Hi-C protocol from Ay *et al.*, Genome Research, 2014, included here for comparison.

Supplemental Table 1.3: Loci that show a two-fold or larger difference in contacts between two stages with a false discovery rate of 1%. Supplied as a separate xlsx file.

Supplemental Table 1.4: Significant interactions ($q<0.05$) between *pfap2* loci and virulence gene clusters. Supplied as a separate xlsx file.

Supplemental Table 1.5: The sum of Hi-C contacts for significant interactions (5% FDR) between 10 kb bins containing *pfap2* genes and 10 kb bins containing virulence genes.

Gene	Location	R	T	S	EG	LG	SPZ	Notes
PF3D7_0404100	chr4:224,779-231,733	7	11	6	-	-	7	n.a.
PF3D7_0420300[1]	chr4:917,990-928,411	33	114	55	9	55	12	n.a.
PF3D7_0611200	chr6:467,481-468,642	-	-	5	-	-	-	n.a.
PF3D7_0730300	chr7:1,297,459-1,301,454	-	-	15	-	-	15	*pfap2-I*
PF3D7_0802100	chr8:151,808-159,668	9	-	-	-	-	4	n.a.
PF3D7_0934400	chr9:1,349,790-1,350,392	13	-	-	-	-	-	n.a.
PF3D7_1139300	chr11:1,556,744-1,565,045	-	-	-	-	-	6	n.a.
PF3D7_1222400	chr12:890,581-898,257	-	58	56	-	-	-	n.a.
PF3D7_1222600	chr12:907,203-914,501	-	215	23	-	87	-	*pfap2-g*

R, ring; T, trophozoite; S, schizont; EG, early (stage II/III) gametocytes; LG, late (stage IV/V) gametocytes; SPZ, sporozoites; n.a., not applicable.

[1]Interactions between PF3D7_0420300 and the nearby internal virulence gene cluster (in the adjacent 10 kb bin) were not included in this table to exclude any interactions caused by physical constraints.

Supplemental Table 1.6: H3K9me3 ChIP-seq results. Supplied as a separate xlsx file.

Supplemental Table 1.7: Loci involved in long-range interactions in *P. vivax* sporozoites.

Chr	locus (kb)	start	end	gene	description	Pf homolog
3	185	184,691	187,491	PVX_000945	Apical sushi	PF3D7_0405900
3	285	277,713 284,048 287,913	280,254 287,041 291,841	PVX_000820 PVX_000815 PVX_000810	Flap endonuclease SIAP1 PLP1	PF3D7_0408500 PF3D7_0408600 PF3D7_0408700
7	1,085	1,080,758 1,085,365 1,088,721 1,089,698	1,083,176 1,087,252 1,089,632 1,090,833	PVX_099855 PVX_099860 PVX_099870 PVX_099875	Hypothetical Hypothetical Hypothetical Hypothetical	PF3D7_0928200 PF3D7_0928300 PF3D7_0928400 PF3D7_0928500
7	1,355	1,353,104 1,356,687	1,354,428 1,360,472	PVX_086945 PVX_086940	Exported Hypothetical	PF3D7_0935500 n.a.
7	1,385	1,380,491 1,383,396 1,388,280	1,381,592 1,383,959 1,389,756	PVX_086920 PVX_086915 PVX_086910	Hypothetical ETRAMP PHIST	n.a. n.a. n.a.
9	1,535	1,515,724	1,524,557	PVX_092570	ApiAP2 TF	PF3D7_1139300
11	175	169,826 172,656 175,850 179,797	171,495 174,464 176,663 183,156	PVX_115295 PVX_115290 PVX_115285 PVX_115280	Hypothetical SRP receptor NDP kinase Hypothetical	PF3D7_1366700 PF3D7_1366600 PF3D7_1366500 PF3D7_1366400
11	1,075 1,085	1072903 1080139	1073769 1086780	PVX_114310 PVX_114305	Ribonuclease H2 Tyr kinase	PF3D7_0623900 PF3D7_0623800
11	1,115	1,129,262	1,138,672	PVX_114260	ApiAP2 TF	PF3D7_0622900 (*pfap2-sp3)*
13	1,105	1,105,001	1,107,127	PVX_085325	Pv specific	n.a.
13	1,775	1,773,064	1,777,749	PVX_086035	ApiAP2 TF	PF3D7_1408200 (*pfap2-g2*)

Kb, kilobase; Pf, *Plasmodium falciparum*; n.a., not available.

Supplemental Table 1.8: Sequences of primers used for the generation of FISH probes.

Chr	Gene	Annotation	Locus (kb)	Primer sequence
8	PF3D7_0800300	*Var*	40-50	Forward: 5'-CGAAAGATAGTAGTGATGGT-3' Reverse: 5'-CACTTATGCATTTCCATCCA-3'
12	PF3D7_1222600	*pfap2-g*	90-92	Forward: 5'-ATGGATAATATGAATGCACCTA-3' Reverse: 5'-GTTGATAAATCACTAATAGCAC-3'

Supplemental Table 1.9: Sequences of primers used to validate that chr14 is physically intact around the domain boundary.

Chr	Location	Size	Primer sequence	Use
14	1,183,922 – 1,187,279	3,358	Forward: 5'- GTGTGTTAAATCCATTGATC -3' Reverse: 5'- GAAAGAATGTTGTTAAGCATCC -3'	PCR
14	1,184,647 – 1,187,744	3,098	Forward: 5'- GAGTAACTATAATATAGGTCC -3' Reverse: 5'- GCGCGATAAATATACACCACC -3'	PCR
14	1,185,204 – 1,188,066	2,863	Forward: 5'- GGTAATAGAGGATTTCAACA -3' Reverse: 5'- CGTGTACATATAAAGTGACATAC -3'	PCR
14	1,185,589 – 1,187,279	1,691	Forward: 5'- GTAGTGTACATACACTTATG -3' Reverse: 5'- GCGCGATAAATATACACCACC -3'	PCR
14	1,185,589 – 1,188,519	2,931	Forward: 5'- GTAGTGTACATACACTTATG -3' Reverse: 5'- CAATCCTCTATGTTTATCTACATC -3'	PCR
14	1,188,381 – 1,188,599	218	Forward: 5'- GAAACAATTTCCGATATATTTAACTCAACATAGA -3' Reverse: 5'- GAACTACCTGTGCCTCTCC -3'	Probe for Southern

Supplemental Movie 1.1: We recommend opening this file in VLC media player (http://www.videolan.org/vlc/). The video is also accessible via YouTube: https://youtu.be/fcccffs16FQ

The order of the stages shown in the video is as follows: ring, trophozoite, schizont, early gametocyte, late gametocyte and finally sporozoite.

CHAPTER 2: Comparative 3D Genome Organization in Apicomplexan Parasites

Evelien M. Bunnik[1], Aarthi Venkat[2,*], Jianlin Shao[2,#,*], Kathryn E. McGovern[3,$], Gayani Batugedara[4], Danielle Worth[3], Jacques Prudhomme[4], Stacey A. Lapp[5,6,7], Chiara Andolina[8,%], Leila S. Ross[9], Lauren Lawres[10], Declan Brady[11], Photini Sinnis[12], Francois Nosten[8], David A. Fidock[9], Emma H. Wilson[3], Rita Tewari[11], Mary R. Galinski[5,6,7], Choukri Ben Mamoun[10], Ferhat Ay[2,13,*], and Karine G. Le Roch[4,*]

[1] Department of Microbiology, Immunology & Molecular Genetics, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA.

[2] Division of Vaccine Discovery, La Jolla Institute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA.

[3] School of Medicine, Division of Biomedical Sciences, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA.

[4] Department of Molecular, Cell and Systems Biology, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA.

[5] International Center for Malaria Research, Education and Development, Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, 201 Dowman Drive, Atlanta, GA 30329, USA.

[6] Department of Medicine, Division of Infectious Diseases, Emory University, 1648 Pierce Dr NE, Atlanta, GA 30329, USA.

[7] Malaria Host-Pathogen Interaction Center, 954 Gatewood Road, Atlanta, GA 30329, USA.

[8] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine Research building, University of Oxford, Old Road campus, Roosevelt Drive, Headington, Oxford, OX3 7FZ, UK and Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Tak 63110, Thailand.

[9] Department of Microbiology and Immunology, Columbia University Medical Center, New York, NY 10032, USA.

[10] Department of Internal Medicine, Section of Infectious Diseases, Yale School of Medicine, 330 Cedar St, Boardman 110, New Haven, CT 06520, USA.

[11] School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK.

[12] Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, E5132, Baltimore, MD 21205, USA.

[13] School of Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92521, USA.

[#] Present address: Zhejiang Provincial Center for Cardio-Cerebro-Vascular Disease Control and Prevention, Zhejiang Hospital, Zhejiang Province, China
[$] Present address: BIO5 Institute, University of Arizona, Tucson, AZ, USA
[%] Present address: Radboud University Medical Centre, Nijmegen, The Netherlands.

*These authors contributed equally.

A version of this chapter has been submitted to *PNAS*, 2018.

Preface

Chromosomes within the eukaryotic cell nucleus are highly dynamic and adopt complex hierarchical structures. Understanding how this three-dimensional (3D) nuclear architecture affects gene regulation, cell cycle progression and disease pathogenesis are important biological questions in development and disease. We investigated the 3D organization of chromosomes in malaria parasites and related apicomplexan parasites to identify possible connections between genome architecture and pathogenicity. Genome organization was dominated by the clustering of *Plasmodium*-specific gene families in 3D space. In particular, the two most pathogenic human malaria parasites shared unique features in the organization of gene families involved in antigenic variation and immune escape. Related human parasites *Babesia microti* and *Toxoplasma gondii* that are less virulent lacked the correlation between gene expression and genome organization observed in human *Plasmodium* species. Our results suggest that genome organization in malaria parasites has been shaped by parasite-specific gene families and correlates with virulence. It is our hope that the identification of molecular components regulating these parasite-specific genes at the chromatin structure level will assist the identification of novel targets for future therapeutic strategies. Though I am not the primary contributor to the body of work presented in this chapter, these findings have helped establish the direction for my later projects and contribute to our understanding of the importance of nuclear organization in apicomplexan parasites. I specifically performed the in situ Hi-C experiments for *P. knowlesi*, *P. berghei*, *P. yoelii*, *B. microti* and *T. gondii*, and contributed to figures 2.2-2.4 and supplemental figures 2.1-2.3, 2.6- 2.9.

Abstract

The positioning of chromosomes in the nucleus of a eukaryotic cell is highly organized and has a complex and dynamic relationship with gene expression. In the human malaria parasite *Plasmodium falciparum*, the clustering of a family of virulence genes correlates with their coordinated silencing and has a strong influence on the overall organization of the genome. To identify conserved and species-specific principles of genome organization, we performed Hi-C experiments and generated 3D genome models for five *Plasmodium* species and two related apicomplexan parasites. *Plasmodium* species mainly showed clustering of centromeres, telomeres and virulence genes. In *P. falciparum*, the heterochromatic virulence gene cluster had a strong repressive effect on the surrounding nuclear space, while this was less pronounced in *P. vivax* and *P. berghei*, and absent in *P. yoelii*. In *P. knowlesi*, telomeres and virulence genes were more dispersed throughout the nucleus, but its 3D genome showed a strong correlation with gene expression. The *Babesia microti* genome showed a classical Rabl organization with colocalization of subtelomeric virulence genes, while the *Toxoplasma gondii* genome was dominated by clustering of the centromeres and lacked virulence gene clustering. Collectively, our results demonstrate that spatial genome organization in most *Plasmodium* species is constrained by the colocalization of virulence genes. *P. falciparum* and *P. knowlesi*, the only two *Plasmodium* species with gene families involved in antigenic variation, are unique in the effect of these genes on chromosome folding, indicating a potential link between genome organization and gene expression in more virulent pathogens.

Introduction

Apicomplexans are obligate intracellular parasites that can be highly pathogenic and are responsible for a wide range of diseases in humans and animals. The phylum consists of at least 6,000 species, with potentially many undiscovered members (1). Among apicomplexan parasites that infect humans, *Plasmodium* spp., the causative agents of malaria, have the highest health and economic impact. The most prevalent and deadly human malaria parasite is *P. falciparum*, responsible for an estimated 445,000 deaths per year (2). Other *Plasmodium* species that infect humans include *P. vivax* and *P. knowlesi*. *P. vivax* is widespread predominantly outside Africa, and *P. knowlesi* is a zoonosis in Southeast Asia. The natural hosts of *P. knowlesi* are long-tailed and pig-tailed macaques. However, transmission from monkeys to humans through a mosquito vector has been widely reported in Malaysia and can cause severe, potentially lethal disease (3). Other human-relevant apicomplexans include *Babesia microti* (4), the causative agent of human babesiosis, a malaria-like illness endemic in the US but with worldwide distribution, and *Toxoplasma gondii*, the causative agent of toxoplasmosis, an opportunistic infection commonly encountered among individuals with weakened immune systems (5).

During millions of years of co-evolution with their hosts, apicomplexan parasites have developed species-specific large multigene families that are involved in host-parasite interactions (6). These gene families are important for parasite survival, pathogenesis, virulence, and immune evasion. All *Plasmodium* species contain virulence genes that belong to the *Plasmodium* interspersed repeat (*pir*) superfamily. In addition, *P.*

falciparum and *P. knowlesi*, have evolved unique gene families that orchestrate these parasites to undergo antigenic variation, called *var* and *SICAvar*, respectively (7, 8). During the process of antigenic variation, the parasite can escape from host immune responses by changing which members of these large families of parasite antigens are expressed. The ability of these parasites to switch their antigenic profile correlates with their high virulence and persistence in the face of adaptive immune responses.

The biological functions of *pir* genes are largely unknown, but it has been suggested that they have many different roles in virulence, signaling, trafficking, protein folding, adhesion, and establishment of chronic infections (9-11). The availability of genome sequences for selected apicomplexan parasites has revealed the genomic landscape of these virulence gene families. Copy numbers for the virulence gene families typically range between 150 and 300 genes per organism, although there are some exceptions (for example, 980 *yir* genes in *P. yoelii* 17X) (6,12-14). These genes are located close to the telomere ends of most (sometimes all) of the 14 chromosomes of the various *Plasmodium* genomes. Similarly, the subtelomeric regions of *B. microti* chromosomes contain several small gene families encoding exported proteins. These proteins are targets of the antibody response in *B. microti*-infected humans and may be involved in antigenic variation (15-17). *T. gondii* also has multiple parasite-specific gene families involved in pathogenesis and immune evasion. Some of these are subtelomeric, but most are dispersed among the genome, either as individual genes, or in smaller and larger arrays (18). *T. gondii* does not use classic antigenic variation, although some of these *Toxoplasma*-specific gene

families may be involved in escape from immune responses (19).

First discovered in *P. knowlesi* (20-23), the P. falciparum *var* and *P. knowlesi SICAvar* gene families mediate antigenic variation and immune escape and may be one of the factors that make *P. knowlesi* and *P. falciparum* so lethal in humans. The *var* and *SICAvar* gene families encode *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) and Schizont Infected Cell Agglutination variant antigen, respectively, which are expressed on the surface of the infected red blood cell. As a result, these proteins are exposed to the host immune system and elicit strong antibody responses. Each parasite expresses a single PfEMP1 or a limited repertoire of SICAvar antigens (24, 25). Parasites can rapidly and efficiently escape from the host immune response by switching the gene variant that is expressed, resulting in successive cycles of antibody production and parasite escape (22, 26). In addition, PfEMP1 mediates adherence of infected red blood cells to the vasculature, which prevents clearance of the parasite by the spleen. Cytoadherence in vital organs causes tissue damage and is a major cause of pathology in *P. falciparum* malaria (27).

Similar to the *pir* genes, most *var* genes are located in the subtelomeric regions of all 14 *P. falciparum* chromosomes, although several chromosomes also harbor internal *var* genes. The *SICAvar* genes are distributed more evenly along the chromosomes, with only a few located in subtelomeric regions (7). One of the mechanisms involved in the complex network of clonal *var* gene expression is localization of these genes in

perinuclear clusters of heterochromatin (28, 29). In previous studies (30, 31), we observed that the requirement for *var* genes to come together in 3D space has a strong influence on the overall organization of the *P. falciparum* genome. Chromosomes with internal *var* gene clusters form loops to accommodate the perinuclear localization of all *var* genes. In addition, we observed a strong association between three- dimensional genome organization and gene expression.

Based on these observations, we hypothesized that gene families involved in antigenic variation need to be tightly regulated and therefore have a strong influence on genome organization. Organisms that do not undergo antigenic variation may thus have less stringent requirements with respect to the structure of their genome in the nucleus. To test this hypothesis, we studied the genome architecture of five different *Plasmodium* species parasites, two that are known to undergo antigenic variation (*P. falciparum* and *P. knowlesi*), and three that are not (*P. vivax*, *P. berghei*, and *P. yoelii*). In addition, we studied two related apicomplexan parasites (*B. microti* and *T. gondii*) to identify characteristics of genome organization that are specific to the *Plasmodium* genus. When considering genome-wide clustering of all virulence genes, we observed that all organisms studied here, with the exception of *T. gondii* and *P. knowlesi*, showed significant colocalization for these genes. In *P. knowlesi*, even though contact counts between *SICAvar* genes show a moderate enrichment compared to random, these were found to be scattered throughout the nucleus. However, *SICAvar* loci showed cross-shaped patterns in intra- chromosomal contact with enriched contact counts to other

SICAvars on the same chromosome similar to interaction patterns created by *P. falciparum var* genes, suggesting that colocalization of these genes is a conserved feature and may play role in mutually exclusive expression of *SICAvars*. *T. gondii* virulence genes are not located subtelomerically and did not cluster in 3D space, pointing towards differences in regulation of gene expression in this apicomplexan parasite.

Results

Profiling genome organization for apicomplexan parasites

We performed Hi-C experiments for five different apicomplexan parasites using the in situ Hi-C methodology (32) (see Methods): *Plasmodium knowlesi*, *P. yoelii*, *P. berghei*, *Toxoplasma gondii*, and *Babesia microti* (Table 2.1 and Figure 2.1). Hi-C data for *P. falciparum* and *P. vivax* were available from previous studies (30, 31). For all *Plasmodium* species except *P. vivax*, we performed the Hi-C experiment on human blood stage parasites. The only *P. vivax* sample available to this study was from mosquito salivary gland sporozoites, and we therefore also included *P. falciparum* sporozoites to allow comparisons between different developmental stages in different parasite species. Even though the global features of genome organization are comparable in *P. falciparum* trophozoites and gametocytes, we also included the *P. falciparum* gametocyte stage to allow a direct comparison with *P. berghei* gametocytes. *B. microti* tick isolate (LabS1) and clinical isolate (Bm1438) samples were obtained from non-synchronized blood stage infections in mice and from cultures that were predominantly at the last stage of intraerythrocytic development (tetrads). Finally, for *T. gondii* we included the rapidly

replicating tachyzoite stage and the more dormant bradyzoite stage, both cultured in vitro.

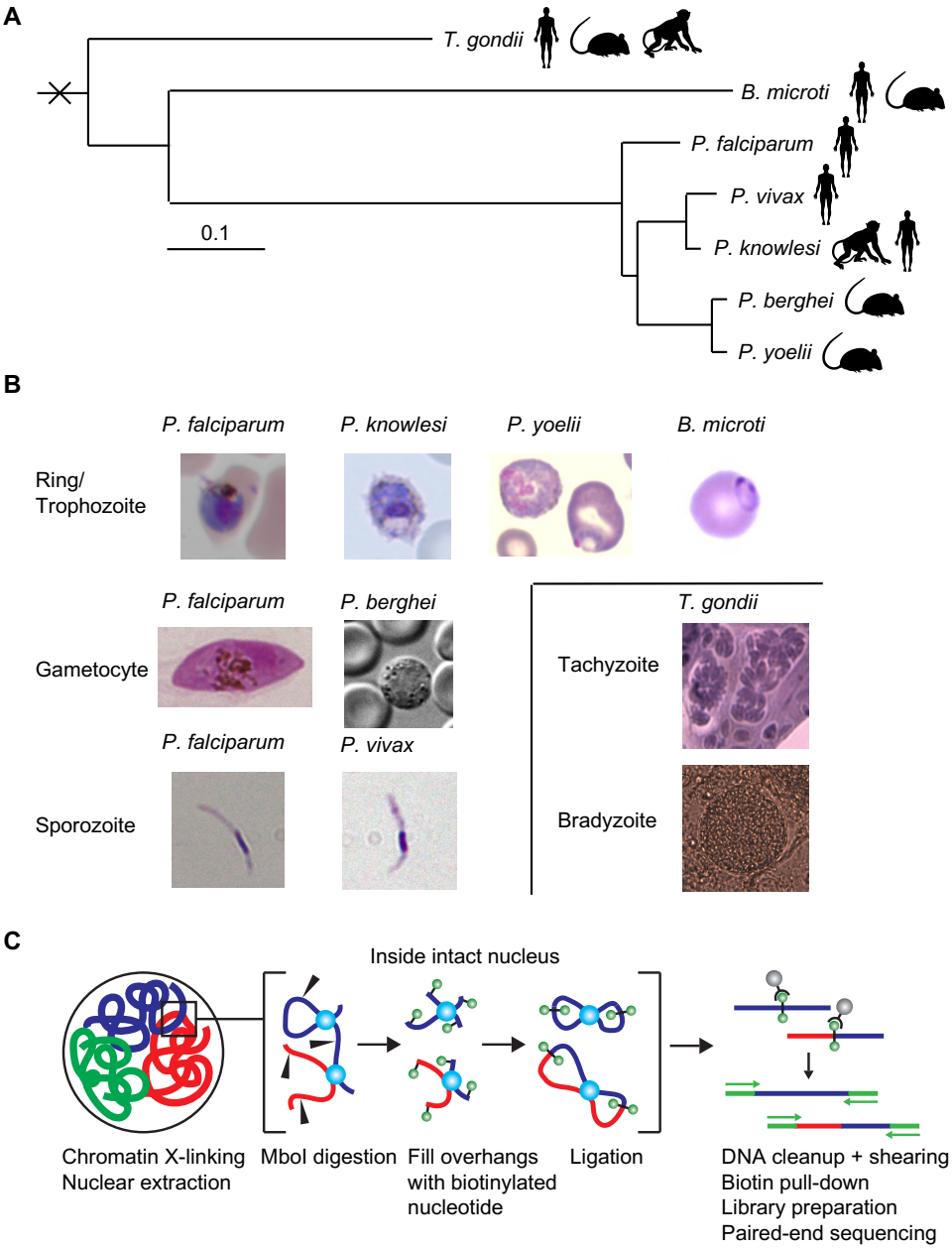


Figure 2.1: Overview of samples and protocol. A) Phylogenetic tree showing the genetic relationship between the seven different apicomplexan parasites used in this study (adapted from (61). B) Light microscopy images of the various parasites. C) Schematic representation of the in situ Hi-C protocol.

For several of these samples, we performed Hi-C experiments for two or more biological replicates or highly related strains (Table 2.1 and Materials and Methods), which showed a high degree of similarity as quantified by a Hi-C-specific reproducibility measure (33) (Supplementary Figure 2.1). For *P. falciparum,* we observed higher reproducibility among replicates of the same stage as compared to samples from different stages. Therefore, we combined the two gametocyte samples (early and late) as well as the two sporozoite replicates for subsequent analysis to obtain higher resolution. For *P. vivax*, the two sporozoite replicates showed very high reproducibility and were therefore also combined. For *P. berghei*, all three conditions (WT and Smc2/Smc4 mutants) were highly similar and, hence, were combined for further analysis. For *B. microti*, synchronous and asynchronous samples had higher reproducibility within compared to across these two groups regardless of whether the samples were from the tick or field isolates. Hence, we combined all synchronous samples together and similarly the asynchronous samples. After combining specimens with high correlations, we obtained a total of eleven samples from the various stages, strains, and conditions of the seven different apicomplexan parasites (Table 2.1).

For each sample, the Hi-C reads were processed (i.e., mapping, filtering, pairing and removing duplicates) using HiCPro package (34), resulting in a total of ~300 million unique valid read pairs out of more than 1.5 billion pairs sequenced (Supplementary Table 2.1). The observed intrachromosomal and interchromosomal contacts were aggregated into contact count matrices at 10-kb resolution and were then normalized using the ICE method to correct for experimental and technical biases (35) (normalized contact count heatmaps are accessible at http://apicomplexan3d.lji.org/). Next, we inferred a consensus 3D genome structure for each of the eleven samples using the negative binomial model from the PASTIS 3D modeling toolbox (36) (https://github.com/hiclib/pastis) (Figure 2.2 and Supplementary Fig. 2.2). For assessing the stability of the consensus structures, we used 100 random initializations of the 3D coordinates and clustered the resulting 100 structures for each sample (Supplementary Fig. 2.3). For most of our samples, we did not find any subset of 3D models that distinctly cluster with each other suggesting the resulting models are robust to differences in initialization. For *P. vivax,* which showed some sign of clustering, we sampled representative models from each potentially distinct cluster. Our comparison of 3D structures from the four most prominent clusters showed no striking differences in genome organization (Supplementary Fig. 2.4). Since we observed conservation of all of the main structural features among the different initializations in either case, we used a single representative model for each of our Hi-C samples.

Table 2.1: Description of organisms and source material included in this study.

Species	Strain	Natural hosts	Source	Stage	Replicates/ strains
Plasmodium falciparum	Lab strain 3D7	Human/ Mosquito	In vitro culture	Trophozoites	1 sample
Plasmodium falciparum	Lab strain NF54	Human/ Mosquito	In vitro culture	Gametocytes	2 samples
Plasmodium falciparum	Lab strain NF54	Human/ Mosquito	*A. stephensi* mosquitoes	Sporozoites	2 biological replicates
Plasmodium vivax	Field strain	Human/ Mosquito	*A. cracens* mosquitoes	Sporozoites	2 biological replicates
Plasmodium knowlesi	Lab strain Pk1(A+)	Long-tailed macaque/ Mosquito	Rhesus macaques	Trophozoites	1 sample
Plasmodium berghei	Lab strain ANKA	Rodents/ Mosquito	Mouse	Gametocytes	3 strain variants
Plasmodium yoelii	Lab strain XNL	Rodents/ Mosquito	Mouse	Mixed blood stage	1 sample
Babesia microti	Tick isolate LabS1 + clinical isolate Bm1438	Warm-blooded animals/ ticks	Mouse	Mixed blood stage	2 strains, 1 sample each
Babesia microti	LabS1 + Bm1438	Warm-blooded animals/ ticks	Ex vivo mouse blood cultures	Synchronized tetrads	2 strains, 2 biological replicates each
Toxoplasma gondii	Lab strain ME49	Warm-blooded animals	In vitro culture	Tachyzoites	1 sample
Toxoplasma gondii	Lab strain ME49	Warm-blooded animals	In vitro culture	Bradyzoites	1 sample

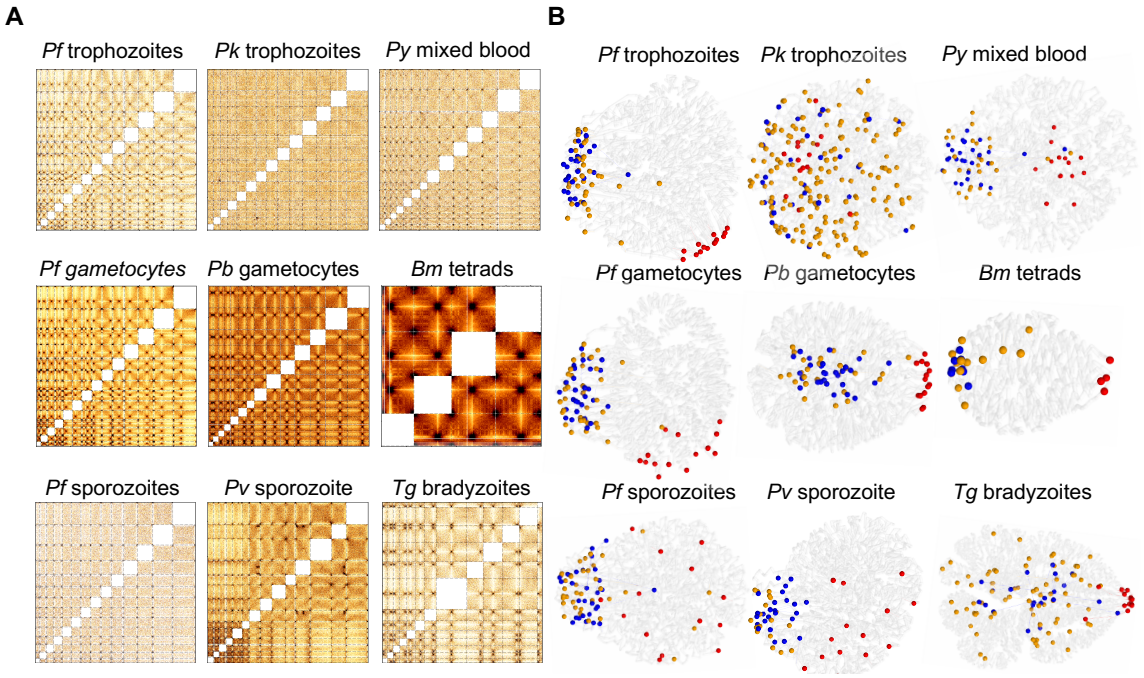


Figure 2.2: Hi-C data and 3D genome modeling. A) Normalized interchromosomal contact count heatmaps at 10-kb resolution. Chromosomes are lined up in numerical order starting with chr1 in the bottom left corner. Individual chromosomes are delineated by dashed lines. Intrachromosomal contacts are not displayed, hence the white squares along the diagonal of each heatmap. Larger versions of the interchromosomal heatmaps as well as all intrachromosomal heatmaps are accessible at http://apicomplexan3d.lji.org/. B) Representative 3D models for genomes of all organisms studied here. Chromosomes are shown as transparent white ribbons. Centromeres are indicated with red spheres, telomeres with blue spheres and virulence genes with orange spheres. Representative 3D models for asynchronous *B. microti* blood stages and *T. gondii* tachyzoites are shown in Supplementary Figure 2.2.

Detection of genome assembly problems and their correction with Hi-C data

Hi-C data has been used to detect translocation events or to improve genome assemblies based on Illumina and/or PacBio reads in many organisms, including *Arabidopsis thaliana, Aedes aegypti*, and human (37-39). Therefore, we first scanned our samples using a metric developed to detect genome assembly errors in Hi-C data (37) to avoid biases in our downstream analysis of 3D genome organization. Small misassemblies were observed in *P. yoelii* chr9 and in *B. microti* chr1 for the tetrad sample (Supplementary Fig. 2.5), but these are unlikely to influence our results. For *P. knowlesi*, we recently published a new assembly of the *P. knowlesi* genome based our Hi- C data in combination with new sequencing data generated using Illumina and Pacbio platforms (40). All analyses presented here were performed using the updated version of the *P. knowlesi* genome assembly. All other *Plasmodium* and *Babesia* samples were error-free. However, several issues were detected for *T. gondii* that were handled as described below.

The current genome assembly of *T. gondii* consists of 14 chromosomes, the same number as for *Plasmodium* species and close apicomplexan relatives, such as *Neospora caninum*. For 13 chromosomes, the location of the centromere has previously been identified by chromatin immunoprecipitation of centromeric and pericentromeric proteins (41, 42). For chromosome VIIb, the centromere has thus far remained elusive. In the interchromosomal heatmap using the original genome assembly with 14 chromosomes, it can be observed that chrVIIb and chrVIII have a higher number of interactions than any

other combination of chromosomes, and that the number of contacts is highest between the right telomere of chrVIIb and the left telomere of chrVIII (Figure 2.3A). These observations suggest that chrVIIb and chrVIII could be physically linked. After we computationally stitched chrVIIb and chrVIII together to create one large chromosome, the interchromosomal contact counts were at the expected levels, while no apparent discrepancies were observed in either the interchromosomal heatmap (Figure 2.3B) or the intrachromosomal heatmap of the combined chrVIIb and chrVIII (Figure 2.3C). This stitched chromosome showed a single centromere interaction with every other chromosome. In addition, we did not detect any signs of misassembly in the stitched chromosome (Figure 2.3D). The small signal observed in the tachyzoite sample is not at the junction of the stitched chrVIIb and chrVIII. Based on these observations, we propose that chrVIIb and chrVIII are in fact a single chromosome. This would explain the unusual interaction pattern in our original interchromosomal interaction plot, as well as the apparent absence of a centromere in chrVIIb. We have used this stitched chromosome in all of our analyses and refer to this chromosome as chrVIII.

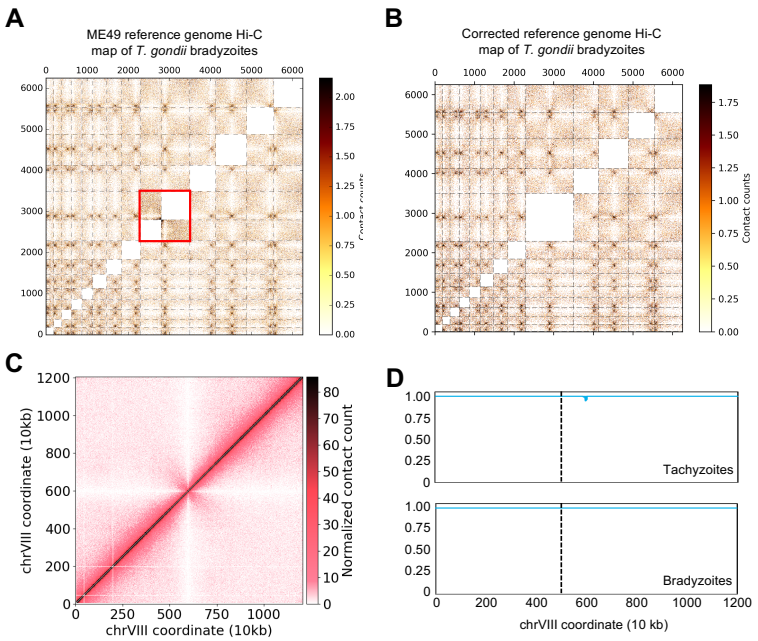


Figure 2.3: Correction of the *T. gondii* genome assembly by Hi-C data. A) Normalized interchromosomal contact count heatmap plotted using the current version of the *T. gondii* genome showing unusually high levels of contact counts between chrVIIb and chrVIII (inside the red box). B) Normalized interchromosomal contact count heatmap plotted using an updated version of the *T. gondii* genome in which chrVIIb and chrVIII have been stitched to form one large chrVIII. All interchromosomal contacts are at expected levels and the newly formed chromosome shows a single centromeric interaction with each other chromosome. C) The intrachromosomal contact count heatmap of the newly formed chromosome chrVIII, showing no discrepancies in contact counts along the chromosome or at the junction. D) Misassembly metric for the newly formed chrVIII, showing no signs of misassembly. The junction is indicated with a dashed line.

The misassembly metric detected several other issues in the *T. gondii* genome (Supplementary Fig. 2.5). Most prominently, chrXII showed an unusual pattern that is most likely caused by an inversion of a segment spanning from bin 272 to bin 499 (Supplementary Fig. 2.6A). The fraction of the parasite population harboring this inversion is approximately 10% in the tachyzoite sample and is close to 100% in the bradyzoite population. The tachyzoite and bradyzoite samples were generated using an unmodified and a transgenic ME49 strain, respectively. This result indicates that this inversion most likely arose spontaneously in the ME49 strain and can be selected during bottleneck events such as the generation of a transgenic strain. For the purpose of this study, the contact count patterns caused by this inversion were not considered a misassembly of the *T. gondii* genome.

Lastly, chrIX showed an unexpected signal around bin 500 (Supplementary Fig. 2.6B). Upon inspection of the contact counts in this region, we noted that bins 498 and 505 showed a much higher number of interactions than expected based on the surrounding bins, suggesting amplification of genomic sequences within these two bins. This effect was observed in both the tachyzoite and bradyzoite samples and may point towards an error in the *T. gondii* genome assembly. Since Hi-C does not provide sufficient resolution to resolve such potentially tandem array of repeats, we did not attempt to correct the reference genome for this region. Finally, smaller potential misassemblies were observed in chrIV and chrV.

Colocalization of functional elements and gene families in Plasmodium

In a previous study, we showed that during the blood stages in the human host, the *P. falciparum* genome has a more complex organization than the *Saccharomyces cerevisiae* genome, which is organized in a classical Rabl conformation. In *P. falciparum*, similar to yeast, we observed clustering of telomeres and centromeres on opposite sides of the nucleus but in addition we also observed domain-like structures, similar to multicellular organisms, surrounding genes involved in pathogenicity (30, 31). Clustering of telomeres and centromeres has previously been observed in budding and fission yeast (43, 44), plants, *Drosophila*, and recently also in mammalian cells using advanced single-cell Hi-C analysis and genome modeling (45). It can also clearly be observed in 3D models for the *P. falciparum* trophozoite and gametocyte stages (Figure 2.2B).

Here, we analyzed the organization of these genomic hallmarks in all analyzed apicomplexan parasites by testing for an enrichment in interactions between loci of interest, as well as for colocalization in our 3D models (Table 2.2). The centromeres interacted with each other in *P. knowlesi*, *P. berghei*, and *P. yoelii*, although the interchromosomal heatmaps showed that these interactions were relatively localized in *P. knowlesi* and *P. yoelii* (Supplementary Fig. 2.7). In agreement with this finding, the centromeres clustered in the 3D models of these organisms. As highlighted in a previous study (31), the centromeres in *P. vivax* salivary gland sporozoites showed more limited contacts compared to the blood stages of any of the *Plasmodium* species and these contacts did not involve any of the surrounding regions. Salivary gland sporozoites from

P. falciparum showed no clustering of centromeres, suggesting that the (near) complete loss of centromere interactions could be a general feature of the sporozoite stage. The telomeres of *P. yoelii* and *P. vivax* showed strong enrichment in contact counts, while the telomeres in *P. berghei* and *P. knowlesi* did not. However, the telomeres in *P. berghei* did come together in 3D space, although to a lesser extent than those in *P. yoelli* and *P. vivax*. All *Plasmodium* species that were analyzed in this study harbor virulence genes at the subtelomeric regions of nearly every chromosome. In line with clustering of the telomeres, we also observed colocalization of virulence genes in all of these organisms, with the exception of the *P. knowlesi SICAvar* genes (Table 2.1). The strong clustering of these genes was recapitulated in the 3D models of *P. falciparum*, *P. vivax*, *P. berghei* and *P. yoelii*. In conclusion, while we observed varying degrees of clustering of telomeres and centromeres, all *Plasmodium* genomes, except for *P. knowlesi*, showed colocalization of *pir* genes.

Table 2.2: Colocalization of centromeres, telomeres and virulence genes.

	Centromeres			Telomeres			Virulence genes		
Sample	Contact counts	3D distance	3D visual	Contact counts	3D distance	3D visual	Contact counts	3D distance	3D visual
Pf trophozoites	++	++	++	++	++	++	++	++	++
Pf gametocytes	++	++	+	++	++	++	++	++	++
Pf sporozoites	-	-	-	+	++	+	+	++	+
Pv sporozoites	++	+	+	++	++	+	++	++	+
Pk trophozoites	++	++	+	-	-	-	-	-	-
Pb gametocytes	++	++	++	-	+	+	+	++	+
Py mixed IDC	++	++	+	++	++	+	++	++	+
Bm mixed IDC	++	++	++	++	++	++	++	+	+
Bm tetrads	+	++	++	++	++	++	+	+	+
Tg tachyzoites	++	++	++	++	++	+	-	-	-
Tg bradyzoites	++	++	++	++	++	+	-	-	-

++ denotes $p < 0.001$ for contact counts and $p< 0.001$ as well as mean and median pairwise distance less than 0.7 of the respective value from randomized set of loci on 3D models; + denotes $p < 0.05$ for contact counts. Two different colocalization tests were used, one based on the statistical significance (46, 47) of contact counts between bins in the Hi-C data and the other based on the distance between bins in the 3D models. The

"3D visual" column denotes whether the listed functional elements are tightly clustered (++), somewhat clustered (+) or not clustered at all (-) as a result of inspecting the 3D models by eye and not by any statistical test.

Antigenic variation genes are associated with domain formation in Plasmodium species

In *P. knowlesi, SICAvar* genes are located in subtelomeric regions, but are also found scattered throughout the genome, either individually or in small groups of up to four genes. Due to the highly repetitive nature of their sequences, 31 of the *SICAvar* genes have low mappability. We were however unable to detect strong co-localization for the 136 remaining *SICAvar* genes. However, our co-localization test measures whether all loci in the genome cluster and does not pick up localized clusters that consist of only a limited subset of genes. Although Hi-C signals were weak for loci that consist of only one or a few genes, the contact count heatmaps showed additional interactions between large internal and subtelomeric *SICAvar* gene clusters, most clearly observed in chr4 (Figure 2.4 and Supplementary Fig. 2.8). This interaction is reminiscent of *var* gene interaction patterns observed in *P. falciparum* that give rise to domain-like structures (30). As in *P. falciparum,* chromosomal loops were also observed in *P. knowlesi*, but were absent in organisms without internal virulence genes (Figure 2.4 and Supplementary Fig. 2.8). These results suggest that subsets of *SICAvar* genes throughout the genome may interact with each other, similar to the *var* genes in *P. falciparum*. Such interactions may be crucial to coordinate mutually exclusive gene expression necessary for antigenic variation.

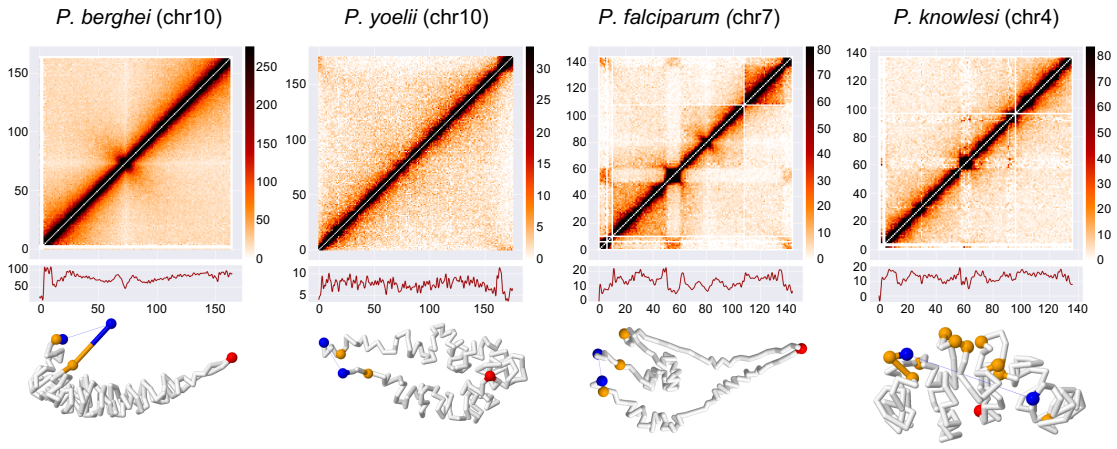


Figure 2.4: Formation of domain-like structures and chromosome loops by *var* and *SICAvar* genes. Top row: normalized intrachromosomal contact count heatmaps at 10-kb resolution for representative chromosomes, showing a canonical "X" shape for chromosomes of *P. berghei* and *P. yoelii*, and domain-like structures in chromosomes with internal *var* and *SICAvar* genes in *P. falciparum* and *P. knowlesi*, respectively. Middle row: quantification of insulation. For each bin, the average contact counts with the 10 upstream and the 10 downstream bins is plotted. Domains are characterized by a depletion of contact counts with surrounding bins. Bottom row: individual chromosome conformation extracted from the 3D model of the full genome. *P. berghei* and *P. yoelii* chromosomes show a folded structure anchored at the centromere, with both chromosome arms arranged in parallel. *P. falciparum* and *P. knowlesi* chromosomes show additional folding structures to bring virulence genes in close spatial proximity.

Genome organization in the apicomplexan parasites B. microti and T. gondii

Babesia microti has a relatively small genome with four chromosomes that showed strong interactions among the telomeres and the centromeres, resulting in a classic Rabl conformation (Figure 2.2B and Supplementary Fig. 2.2). Consequently, subtelomeric multi-gene families were strongly clustered, but additional virulence genes were localized away from the telomeric cluster. The 3D models of the asynchronous and the synchronous sample were very similar (Supplementary Fig. 2.2). However, the contact count heatmaps showed additional interaction patterns within chromosomes for the synchronized samples obtained at the tetrad stage, but not for the asynchronous samples (Supplementary Fig. 2.9). For chr3, these patterns may partially be caused by a virulence gene of the BMN1 family located at position 272,813 - 273,799. These results suggest that these interactions may not be maintained during the complete cell cycle and underscore the importance of using tightly synchronized cell populations to be able to observe transient interactions.

In agreement with a previous microscopy study (41), *T. gondii* centromeres showed strong interactions and colocalized in the 3D models of both the tachyzoite and the bradyzoite stage. The telomeres also interacted, but to a much lower extent as compared to the centromeres, and this interaction was not readily apparent from the 3D models (Supplementary Fig. 2.2). At the resolution used in our models, no significant differences were observed between tachyzoites and bradyzoites (Supplementary Fig. 2.2). Most virulence genes in *T. gondii* are not located in subtelomeric regions, but are found on

every chromosome arranged as single genes as well as smaller and larger arrays of up to 19 genes. As a consequence, virulence genes were scattered throughout the 3D model of the genome and we were unable to detect any significant colocalization pattern for these genes (Table 2.2).

Another difference between genome organization in *Plasmodium* species and *T. gondii* was observed for the strength of chromosome territories. The 3D models for *T. gondii* exhibited more territorialized chromosomes whereas *Plasmodium* chromosomes were more stretched out through the nucleus (see Supplementary Fig. 2.10). In order to quantify these differences directly from Hi-C data, we interrogated the relationship between distance and contact probability, which has been shown to depend on the arrangement of chromosomes within the nucleus (48, 49). However, this scaling relationship has been mainly used to compare different samples of the same organism or organisms with similar genome sizes. Here we adapted this analysis for comparing all our samples ranging from ~4.5Mb (*B. microti*) to ~70Mb (*T. gondii*) in size to each other as well as to human (32) and budding yeast genomes (50) as reference points. In order to do that, at 10kb resolution, we focused only on intra-chromosomal interactions within 1Mb distance and we computed the log2 fold change of the average contact for each genomic distance with respect to the overall average contact count of all possible pairs within 1Mb (Supplementary Fig. 2.11). Our results show that, similar to human genome organization, the contacts within 50kb to 500kb are highly enriched for both *T. gondii* samples compared to all other apicomplexan parasites and budding yeast. Interestingly, this

genomic distance range corresponds to topological domains (TADs) for human genome, which are known to be enriched for higher within- domain interactions. However, we have not seen clear TAD patterns in *T. gondii* genome, suggesting that this trend may be related to the larger chromosomes and genome size of *T. gondii* as compared to other apicomplexan parasites. Further analysis may be necessary to understand the relationship between genome size and chromosome territory formation in the presence and absence of TAD-like structures.

Relation between gene expression and genome organization

P. falciparum gene expression has been shown to be associated with the position of genes in the nucleus (30, 31). To determine the relation between gene expression and genome organization in other apicomplexan parasites, we binned the genes of each organism into 20 groups based on their distance from the centroid of all telomeres. For each bin, we calculated the average gene expression using stage-specific transcriptomics data sets and plotted these values against the average distances from the telomere centroid (**Figure** 2.5A). We also colored the 20 bins in the 3D models based on normalized average gene expression values (Figure 2.5B).

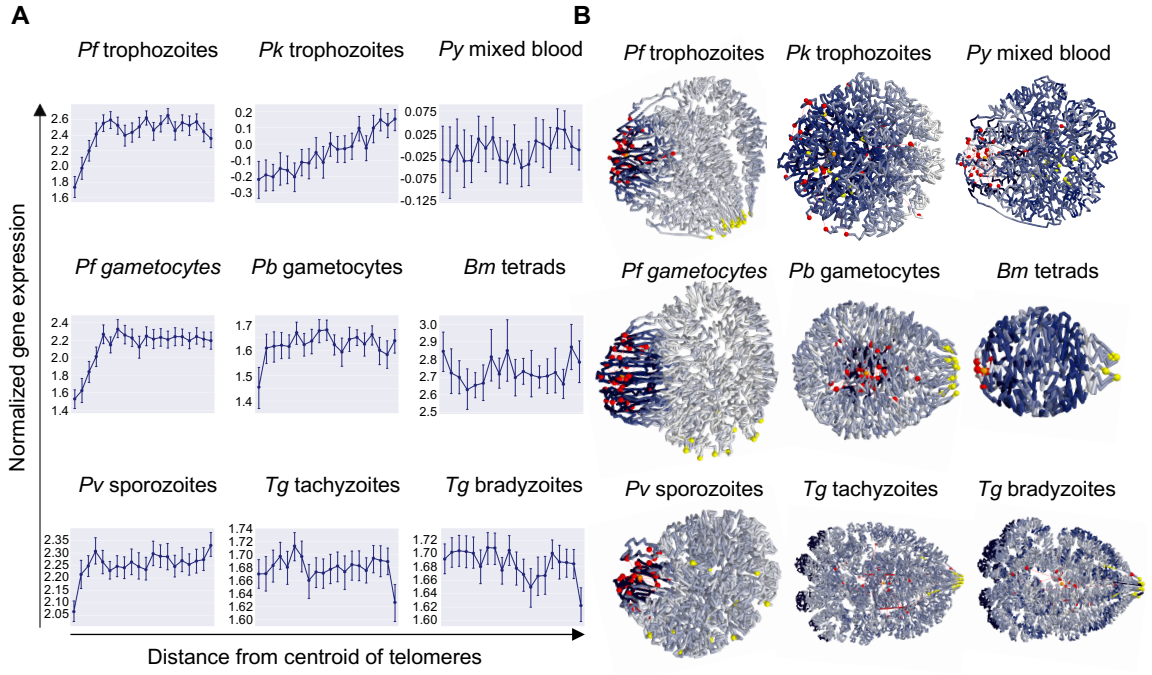


Figure 2.5: Correlation between genome organization and gene expression. A) Relation between gene expression and distance from the centroid of the telomeres. For each organism, genes were divided into 20 bins. For each bin, the average gene expression value was plotted. Error bars denote the range of expression values within each bin. B) Average gene expression values of each bin were projected onto the 3D models, using a color scale ranging from dark blue (low gene expression) to white (high gene expression). Centromeres are shown as yellow spheres, while telomeres are depicted as red spheres.

As expected based on previous work, *P. falciparum* trophozoites showed the lowest gene expression in the bin closest to the centroid of the telomere, and a gradient of increasing gene expression in the next three bins. The remaining 16 bins showed relatively comparable levels of gene expression. The results for *P. falciparum* gametocytes were almost identical. For *P. vivax* sporozoites and *P. berghei* gametocytes, a similar but much weaker pattern was observed, with reduced gene expression only in the bin closest to the telomere centroid. *P. yoelii* and *B. microti* did not show any relation between gene expression and 3D location relative to the telomeres. Surprisingly, the *P. knowlesi* trophozoite stage displayed a gene expression gradient across the entire genome, with the lowest gene expression close to the centroid of the telomere. In the absence of strong telomere and centromere clustering, the 3D model of *P. knowlesi* looked somewhat unorganized. However, the gene expression gradient observed here suggests that genome organization of this parasite is in fact strongly correlated with expression with highly expressed genes preferentially localizing on one side of the 3D genome and the telomeres localizing on the other side.

Finally, for the *T. gondii* samples, we observed a decrease in gene expression in the bin furthest away from the telomere centroid, in stark contrast to all other organisms analyzed here. This decrease was reproducible between the two stages. As can be observed in the 3D models, these genes are located at the nuclear periphery. In both tachyzoites and bradyzoites, the genes in this bin were strongly enriched for merozoite-specific gene expression (51) ($p<0.00001$, two-tailed Fisher's exact test), including 19 of 33 members

of *T. gondii* family A proteins. Collectively, these results suggest that the organization of *Plasmodium* genomes is to a large extent driven by virulence genes. On the other hand, *T. gondii* has adopted different ways to organize gene expression in relation to its nuclear architecture, although the nuclear periphery may also function as a site of gene silencing.

Discussion

From yeast to human cells, genome organization in eukaryotes has a tight relationship to gene expression. In particular, evidence is accumulating that the compartmentalized architecture of a cell nucleus is critical to many biological functions. Evolutionarily conserved mechanisms in the compartmentalization and function of yeast and mammalian genomes have been identified, but genome organization in these organisms also shows differences due to the fact that the nuclei of single cell organisms are typically 1000 times smaller than those of mammals. While nuclear architecture in single cell eukaryotes has been extensively studied in yeast, little is known about the genome organization of other unicellular organisms. We therefore investigated the genome architecture of various apicomplexan parasites. Our goal was to identify common features of genome organization and possible connections between genome architecture and pathogenicity.

In this study, we demonstrated that the Hi-C methodology can be employed to correct genome assemblies and study its 3D organization at the same time. While Hi-C experiments detect many long-range interactions, genomic segments that are physically

linked and in close proximity along the DNA strand are preferentially ligated to one another. This results in a highly reproducible relationship between genomic distance and contact probability. Using this property, we detected that chromosomes VIIb and VIII of the *T. gondii* genome are likely to be physically linked. Our proposed new genome assembly provides a coherent explanation of the apparent absence of a centromere in our original chrVIIb contact count matrix as well as the absence of centromeric protein TgCenH3 and pericentromeric protein TgChromo1 in chrVIIb in previous ChIP-on-chip analyses (41, 42). Our corrected *T. gondii* assembly has one centromere for each of the 13 chromosomes and the interaction between all of these centromeres is clearly visible in interchromosomal contact maps as well as 3D models.

Once genome assemblies were corrected, we focused our analysis on the identification of commonalities and differences in genome organization between members of the *Plasmodium* family, as well as related apicomplexan parasites. We previously showed (30) that although the genomes of *P. falciparum* and *Saccharomyces cerevisae* are similar in size, the *P. falciparum* genome has a more complex three-dimensional structure. In particular, we noted that spatial complexity was added by the requirement for virulence genes of the *var* family to colocalize.

The results from this study demonstrate that the organization of other *Plasmodium* genomes is also largely driven by their virulence genes. However, in contrast to *P. falciparum* and *P. knowlesi*, the rodent malaria species harbor virulence genes only in

subtelomeric regions. The organization of their genomes is therefore relatively simple and similar to fission and budding yeast, with clustering of pericentromeric and subtelomeric heterochromatin islands driving the overall structure. On the other hand, the internal localization of *var* and *SICAvar* genes necessitates the formation of chromosome loops to bring distal loci in close spatial proximity. Similarly, *P. vivax* chr5 harbors a locus that interacts with subtelomeric regions, generating similar domain-like patterns as compared to internal *var* gene islands (31). This locus contains genes of the *Pv-fam- e* family, encoding exported proteins involved in erythrocyte remodeling. The precise function of this gene family and the reasons for its association with subtelomeric heterochromatin remain to be discovered. The clustered organization of virulence genes allows for increased rates of recombination to generate additional diversity and coordination of gene expression. We can only speculate why certain genes in the human malaria parasites are located away from subtelomeric regions, but the requirements for domain formation and complex chromosome structure highlight that additional layers of control are necessary to orchestrate mutually exclusive gene expression.

The *pir* genes of *P. vivax* and the rodent malaria parasites are not subject to the epigenetically driven mutually exclusive expression that is observed for the *var* and *SICAvar* gene families, pointing towards additional differences in regulation between these gene families. While clonal expression of *var* and *SICAvar* genes is thought to enable escape of the parasite from host antibody responses, the expression of certain *pir* genes has been associated with the establishment of acute and chronic infections,

independent of adaptive immunity (9). This difference in virulence gene expression patterns between primate and rodent malaria parasites is reflected by the absence of several nuclear proteins in the genome of rodent malaria parasites. Of these, both the C-terminal extension of Rpb1 (52) and the histone methyltransferase PfSET2 (53, 54) have been shown to contribute to *var* gene regulation. These observations suggest that the various parasite lineages have evolved different strategies to survive in their respective hosts. One contributing factor could be the life span of the host. Parasites infecting long-lived primates may require escape from adaptive immune responses to establish chronic infections and ensure transmission to a susceptible host, while parasites infecting short-lived rodent hosts may benefit from more flexible expression of their virulence genes.

The unique features of *Plasmodium* genome organization as described above are highlighted by the analysis of *B. microti* and *T. gondii*. Both parasites show distinct differences in the structure of their genomes. The *B. microti* genome showed a classical Rabl organization, with colocalization of a subset of virulence genes, those located in subtelomeric regions, in 3D space. From this observation, we conclude that genome organization in *B. microti* is not as strongly involved in regulation of virulence gene expression as in *Plasmodium* parasites. In fact, we observed no association between genome structure and gene expression in general in this parasite. In *T. gondii*, virulence genes were scattered throughout the genome and did not colocalize in the nucleus. The only association between genome organization and gene expression that we could detect was the possible repression of stage-specific genes by the formation of perinuclear

heterochromatin. Our observation that tachyzoites and bradyzoites showed similar genome organization is in line with the relatively small differences in gene expression between these two life cycle stages, in particular when compared to gene expression profiles of merozoites and oocysts (55). However, it should be mentioned that the bradyzoites used here were obtained from in vitro culture and may not fully represent bradyzoites in a tissue cyst. Attempts to prepare Hi-C libraries from bradyzoites isolated from mouse brain tissue cysts were not successful.

The distinct organization of virulence genes suggests that *T. gondii* has adopted different mechanisms for gene regulation as compared to *Plasmodium* species, which may be related to its much broader host range and cell tropism. This is reflected in the larger number of ApiAP2 transcription factors encoded by the *T. gondii* genome (67 versus ~27 for *Plasmodium* species), which are used for both transcriptional activation and repression (56). During the evolution from the ancestor of Apicomplexan organisms, the *Plasmodium* lineage has lost many genes that were retained in *T. gondii* (57). *Plasmodium* species have thus evolved into highly specialized organisms that control their virulence through a highly restricted mechanism, whereas *T. gondii* has retained more properties of their free-living ancestor and is more flexible when it comes to gene regulation.

A limitation of this study is that Hi-C experiments were performed using millions of cells as input. The contact count matrices and 3D models therefore represent the average of a

population of possible genome organizations. In *P. falciparum*, various microscopy approaches have shown clustering of telomeres and virulence genes in 2 - 5 foci spread around the nuclear periphery (28, 29, 58-60). The nature of our data does not allow us to draw a conclusion about whether virulence genes are organized into a single large cluster as observed in our 3D model, or into multiple clusters that vary in *var* gene content among the parasite population.

Another limitation is the dependency of our 3D modeling on strong contact count signals to establish hallmarks of genome organization. In the 3D models, strongly interacting centromeres and telomeres (as seen for example in the trophozoite stage of *P. falciparum* and in *B. microti*) were organized as clusters located at the nuclear periphery. In organisms with weaker centromere interactions (*P. yoelii* and possibly *P. knowlesi*), the 3D models showed clustering of the centromeres in the center of the nucleus. Similarly, weaker telomere interactions (as seen in *P. berghei* and *T. gondii*) also formed clusters in the nuclear center. To further explore this behavior, we deleted the centromeres (all bins containing centromeric sequence) or the telomeres (40 kb region) from each genome, and generated 3D models of these modified genomes. Removal of one hallmark (centromeres or telomeres) resulted in distortion of the 3D structure and loss of clustering of the other hallmark (Supplementary Fig. 2.10). For *T. gondii*, FISH experiments have shown the colocalization of telomeres at the nuclear periphery (42). We therefore believe that the central location of centromere and telomere clusters may be an artifact of our modeling approach and consider it possible that these clusters are in fact located at the nuclear

periphery. For *P. knowlesi*, the telomeres were mostly not mappable and the absence of a signal from the telomeres may explain the unusual genome structure that we obtained for this parasite. Even though the centromeres showed localized interactions in the contact count heatmaps, the *P. knowlesi* 3D model displayed only weak centromere clustering. In the absence of data from telomeres due to mappability issues, the centromeres may be more dispersed in our 3D model than in reality. The mappable telomeres showed ~3-fold higher contact counts as compared to background, and it is therefore not unthinkable that the telomeres in fact cluster as well. Additional approaches, such as IFA or FISH may be necessary to confirm and further investigate our Hi-C based observations.

In conclusion, this study highlights the association between spatial organization of virulence gene families and gene expression in *Plasmodium* species. In *P. falciparum* and *P. knowlesi*, gene families involved in antigenic variation provide a potential link between genome organization and pathogenicity. In contrast, related human parasites *Babesia microti* and *Toxoplasma gondii* lack the correlation between gene expression and genome organization observed in human *Plasmodium* species. Our results emphasize the importance of 3D genome organization in eukaryotes and suggest that genome organization in malaria parasites has been shaped by parasite-specific gene families that affect virulence and clinical phenotypes. Identifying the molecular components regulating these parasite-specific genes at the chromatin structure level will assist the identification of new targets for novel therapeutic strategies.

Material and methods

In situ Hi-C procedure

Parasites were crosslinked in 1.25% formaldehyde in warm PBS for 25 min on a rocking platform in a total volume between 1 and 10 ml, depending on the number of parasites harvested. Glycine was added to a final concentration as 150 mM, followed by 15 minutes of incubation at 37°C and 15 minutes of incubation at 4°C, both steps on a rocking platform. The parasites were centrifuged at 660 x g for 20 min at 4°C, resuspended in 5 volumes of ice-cold PBS, and incubated for 10 min at 4°C on a rocking platform. Parasites were centrifuged at 660 x g for 15 min at 4°C, washed once in ice-cold PBS, and stored as a pellet at -80°C. To map the inter- and intrachromosomal contact counts, crosslinked parasites were subjected to the in situ Hi-C procedure (32), using MboI for restriction digests.

Hi-C data processing

For each sample, the Hi-C reads were processed (i.e., mapping, filtering, pairing and removing duplicates) using HiCPro package (34). The observed intrachromosomal and interchromosomal contacts were aggregated into contact count matrices at 10-kb resolution and were then normalized using the ICE method to correct for experimental and technical biases (35).

Reproducibility score for samples from same organisms

To compute the reproducibility scores for Hi-C data from the same organisms, we used

3DChromatin_ReplicateQC and method GenomeDISCO (33) with default parameters. GenomeDISCO employs graph diffusion and random walks for transformation, and compares the smoothed contact maps between pairs to estimate global similarity. The reproducibility scores in Supplementary Figure 2.1 are genome-wide concordance scores, averaged across all chromosomes.

Quantification of insulation of each locus from neighborhood

We computed a simple score to quantify the extent of insulation of each region from their nearest neighbors, which deviates from expected or average for regions of virulence genes in *P. falciparum* and several other parasites. This score is the average contact count from a region to its 10 upstream and 10 downstream mappable neighbors (100 kb on each side for 10kb resolution maps) and stretches of smaller scores (i.e., contact depletion) indicate insulated regions that give rise to a cross-shaped pattern in contact maps.

3D modeling and visualization

We inferred a consensus 3D genome structure for each organism using the negative binomial model from the PASTIS 3D modeling toolbox (36) (https://github.com/hiclib/pastis). For assessing the stability of the consensus structures, we used 100 random initializations of the 3D coordinates to generate a set of eleven 3D models for Hi-C data from various stages, strains and conditions of the seven different apicomplexan parasites. In order to check the robustness of inferred 3D models to initialization differences, we calculated a disparity score for each possible pair of 3D models (100 choose 2) by first performing Procrustes transformation to find the best

alignment and then computing the sum of the squares of the pointwise differences between the pair of structures. To determine stability of disparity, we clustered these disparity scores in clustermaps and observed very little variation among models with different initializations as well as conservation of all of the main structural features among these initializations (Supplementary Fig. 3). This allowed us to use only a single representative model for each of our Hi-C samples, which we visualized using Jmol: an open-source Java viewer for chemical structures in 3D (http://www.jmol.org/). PDB files and a manual to recreate the structures as displayed in Figure 2 is available at http://apicomplexan3d.lji.org/.

Colocalization tests on 3D distances

For each set of genes or functional annotations such as centromeres or telomeres, we characterized each locus/gene/annotation by including every 10kb bin that it overlaps with. For assessing the p-value of colocalization of loci within a given set, we computed the median pairwise distance for all pairs of loci within the set. Then, we randomly generated the same number of bins on the same chromosomes while preserving the genomic distance relationships for loci on each chromosome. The latter part is done by selecting an initial random locus and pairing it with an anchor locus in the real set and then choosing all other random loci that are at the same distance offset to the random anchor locus as compared to the distance of their counterparts and the real anchor locus. This approach could be considered a generalization of Witten-Noble colocalization test to 3D models instead of contact counts and to handle intra-chromosomal pairs as well as

inter. We performed the randomization a 100,000 times and computed the median pairwise distance between all pairs of loci for each random selection. This median is compared to the corresponding median from the real input set to compute the p-value of observing a random set of loci that are at least as proximal to each other as the real set. The smaller the p-value the more significantly colocalized the real set of loci.

Colocalization tests on contact maps

The colocalization test for contact counts was performed very similar to the corresponding test for 3D distances. The representation of loci by bins, the number of randomizations and the process to get a randomized set of loci was identical to the 3D distance case. The only difference, in this case, is that the mean contact counts were used for the calculation of p-value which correspond to the chance of observing at least as many counts among all pairs of loci in for a random set as for the real set of loci.

Distance scaling plots

The relationship between genomic distance between a pair of loci and the expected number of contacts for this pair was computed using Fit-Hi-C's equal occupancy binning of contact counts (prior to the spline fit) with up to 100 distance bins within 30kb to 1Mb interval. In order to account for differences across genome size, chromosome length and chromosome number for each parasite, we normalized the average number of contact counts per pair at each distance bin by the overall average for the whole distance range up to 1Mb. We then log2 transformed these values for y-axis. X-axis is simply the log10 of the genomic distance.

Correlation between gene expression and 3D models

Gene expression was represented as a function of distance to telomeres. To generate these plots, all genes are first sorted by increasing distance to the centroid of telomeres (x-axis). Then the distance was binned into 20 equal width quantiles and the log average expression value together with the range of values in the bin (y-axis) was plotted for genes in each quantile. For the coloring of 3D models, all 10kb bins are first sorted by increasing distance to the centroid of telomeres (x-axis) and these distances are also binned into 20 equal width quantiles. For each quantile, the representative expression value that is used as the color gradient in 3D visualization was computed as the overall average of the average value of gene expression for all 10kb bins in the quantile.

Data availability

Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number SRP151138. Hi-C data for *P. falciparum* trophozoites from a previous study are available from the NCBI Gene Expression Omnibus (GEO) under accession number GSE50199. Hi-C data for *P. falciparum* gametocytes and sporozoites, as well as *P. vivax* sporozoites from a previous study (31) are available from the NCBI Sequence Read Archive (SRA) under accession number SRP091967.

References

1. Adl SM, *et al.* (2007) Diversity, nomenclature, and taxonomy of protists. *Syst Biol* 56(4):684-689.

2. WHO (2017) The World Malaria Report. http://www.who.int/malaria/publications/world- malaria-report-2017/en/.

3. Cox-Singh J, *et al.* (2008) Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis* 46(2):165-171.

4. Beugnet F & Moreau Y (2015) Babesiosis. *Rev Sci Tech* 34(2):627-639.

5. Flegr J, Prandota J, Sovickova M, & Israili ZH (2014) Toxoplasmosis--a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. *PLoS One* 9(3):e90203.

6. Reid AJ (2015) Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. *Parasitology* 142 Suppl 1:S57-70.

7. Galinski MR, *et al.* (2018) Plasmodium knowlesi: a superb in vivo nonhuman primate model of antigenic variation in malaria. *Parasitology* 145(1):85-100.

8. Deitsch KW & Dzikowski R (2017) Variant Gene Expression and Antigenic Variation by Malaria Parasites. *Annu Rev Microbiol* 71:625-641.

9. Brugat T, *et al.* (2017) Antibody-independent mechanisms regulate the establishment of chronic Plasmodium infection. *Nat Microbiol* 2:16276.

10. Yam XY, *et al.* (2016) Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of Plasmodium chabaudi indicates functional diversity. *Sci Rep* 6:23449.

11. Cunningham D, Lawton J, Jarra W, Preiser P, & Langhorne J (2010) The pir multigene family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* 170(2):65-73.

12. Gardner MJ, *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419(6906):498-511.

13. Otto TD, *et al.* (2014) A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* 12:86.

14. Carlton JM, *et al.* (2008) Comparative genomics of the neglected human malaria

parasite Plasmodium vivax. *Nature* 455(7214):757-763.

15. Lodes MJ*, et al.* (2000) Serological expression cloning of novel immunoreactive antigens of Babesia microti. *Infect Immun* 68(5):2783-2790.

16. Cornillot E*, et al.* (2013) Whole genome mapping and re-organization of the nuclear and mitochondrial genomes of Babesia microti isolates. *PLoS One* 8(9):e72657.

17. Silva JC*, et al.* (2016) Genome-wide diversity and gene expression profiling of Babesia microti isolates identify polymorphic genes that mediate host-pathogen interactions. *Sci Rep* 6:35284.

18. Lorenzi H*, et al.* (2016) Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nat Commun* 7:10147.

19. Lekutis C, Ferguson DJ, Grigg ME, Camps M, & Boothroyd JC (2001) Surface antigens of Toxoplasma gondii: variations on a theme. *Int J Parasitol* 31(12):1285-1292.

20. Howard RJ, Barnwell JW, & Kao V (1983) Antigenic variation of Plasmodium knowlesi malaria: identification of the variant antigen on infected erythrocytes. *Proc Natl Acad Sci U S A* 80(13):4129-4133.

21. Eaton MD (1938) The Agglutination of Plasmodium Knowlesi by Immune Serum. *J Exp Med* 67(6):857-870.

22. Brown KN & Brown IN (1965) Immunity to malaria: antigenic variation in chronic infections of Plasmodium knowlesi. *Nature* 208(5017):1286-1288.

23. al-Khedery B, Barnwell JW, & Galinski MR (1999) Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a P. knowlesi variant antigen. *Mol Cell* 3(2):131- 141.

24. Lapp SA*, et al.* (2013) Spleen-dependent regulation of antigenic variation in malaria parasites: Plasmodium knowlesi SICAvar expression profiles in splenic and asplenic hosts. *PLoS One* 8(10):e78014.

25. Chen Q*, et al.* (1998) Developmental selection of var gene expression in Plasmodium falciparum. *Nature* 394(6691):392-395.

26. Brown KN & Hills LA (1974) Antigenic variation and immunity to Plasmodium knowlesi: antibodies which induce antigenic variation and antibodies which

destroy parasites. *Trans R Soc Trop Med Hyg* 68(2):139-142.

27. Autino B, Corbett Y, Castelli F, & Taramelli D (2012) Pathogenesis of malaria in tissues and blood. *Mediterr J Hematol Infect Dis* 4(1):e2012061.

28. Freitas-Junior LH*, et al.* (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* 407(6807):1018-1022.

29. Lopez-Rubio JJ, Mancio-Silva L, & Scherf A (2009) Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell host & microbe* 5(2):179-190.

30. Ay F*, et al.* (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* 24(6):974-988.

31. Bunnik EM*, et al.* (2018) Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. *Nat Commun* 9(1):1910.

32. Rao SS*, et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665-1680.

33. Ursu O*, et al.* (2018) GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*.

34. Servant N*, et al.* (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 16:259.

35. Imakaev M*, et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* 9(10):999-1003.

36. Varoquaux N, Ay F, Noble WS, & Vert JP (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 30(12):i26-33.

37. Dudchenko O*, et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356(6333):92-95.

38. Jiao WB*, et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* 27(5):778-786.

39. Kaplan N & Dekker J (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* 31(12):1143-1147.

40. Lapp SA, *et al.* (2017) PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. *Parasitology* 145(1):1-14.

41. Brooks CF, *et al.* (2011) Toxoplasma gondii sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc Natl Acad Sci U S A* 108(9):3767-3772.

42. Gissot M, *et al.* (2012) Toxoplasma gondii chromodomain protein 1 binds to heterochromatin and colocalises with centromeres and telomeres at the nuclear periphery. *PLoS One* 7(3):e32671.

43. Duan Z, *et al.* (2010) A three-dimensional model of the yeast genome. *Nature* 465(7296):363- 367.

44. Tanizawa H, *et al.* (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research* 38(22):8164-8177.

45. Stevens TJ, *et al.* (2017) 3D structures of individual mammalian genomes studied by single- cell Hi-C. *Nature* 544(7648):59-64.

46. Ay F, Bailey TL, & Noble WS (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* 24(6):999-1011.

47. Witten DM & Noble WS (2012) On the assessment of statistical significance of three- dimensional colocalization of sets of genomic elements. *Nucleic acids research* 40(9):3849- 3855.

48. Lieberman-Aiden E, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.

49. Mirny LA (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* 19(1):37-51.

50. Eser U, *et al.* (2017) Form and function of topologically associating genomic domains in budding yeast. *Proc Natl Acad Sci U S A* 114(15):E3061-E3070.

51. Hehl AB, *et al.* (2015) Asexual expansion of Toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families

to attach, invade, and replicate within feline enterocytes. *BMC Genomics* 16:66.

52. Kishore SP, Perkins SL, Templeton TJ, & Deitsch KW (2009) An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. *J Mol Evol* 68(6):706-714.

53. Ukaegbu UE*, et al.* (2014) Recruitment of PfSET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in P. falciparum. *PLoS Pathog* 10(1):e1003854.

54. Jiang L*, et al.* (2013) PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* 499(7457):223-227.

55. Behnke MS, Zhang TP, Dubey JP, & Sibley LD (2014) Toxoplasma gondii merozoite gene expression analysis with comparison to the life cycle discloses a unique expression state during enteric development. *BMC Genomics* 15:350.

56. Radke JB*, et al.* (2018) Transcriptional repression by ApiAP2 factors is central to chronic toxoplasmosis. *PLoS Pathog* 14(5):e1007035.

57. Woo YH*, et al.* (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* 4:e06974.

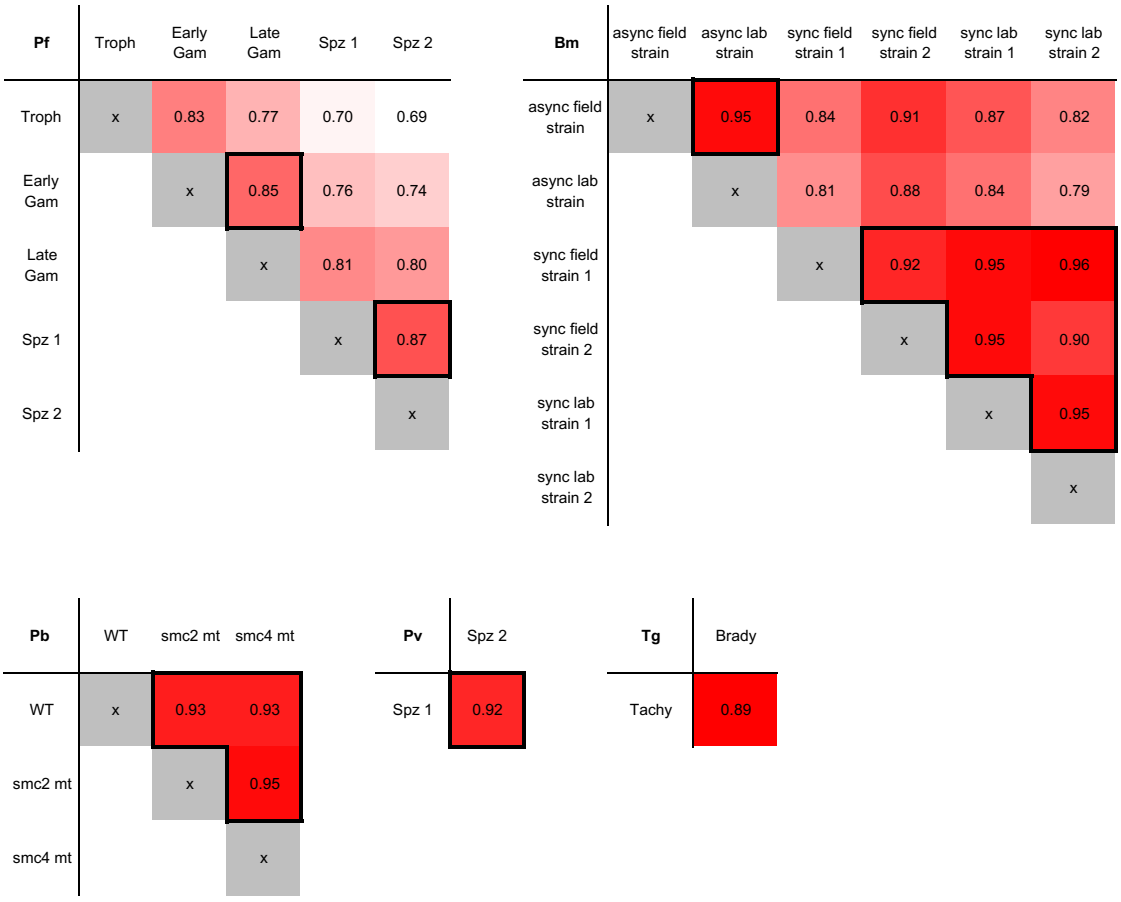
58. Brancucci NM*, et al.* (2014) Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host Microbe* 16(2):165-176.

59. Flueck C*, et al.* (2010) A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathogens* 6(2):e1000784.

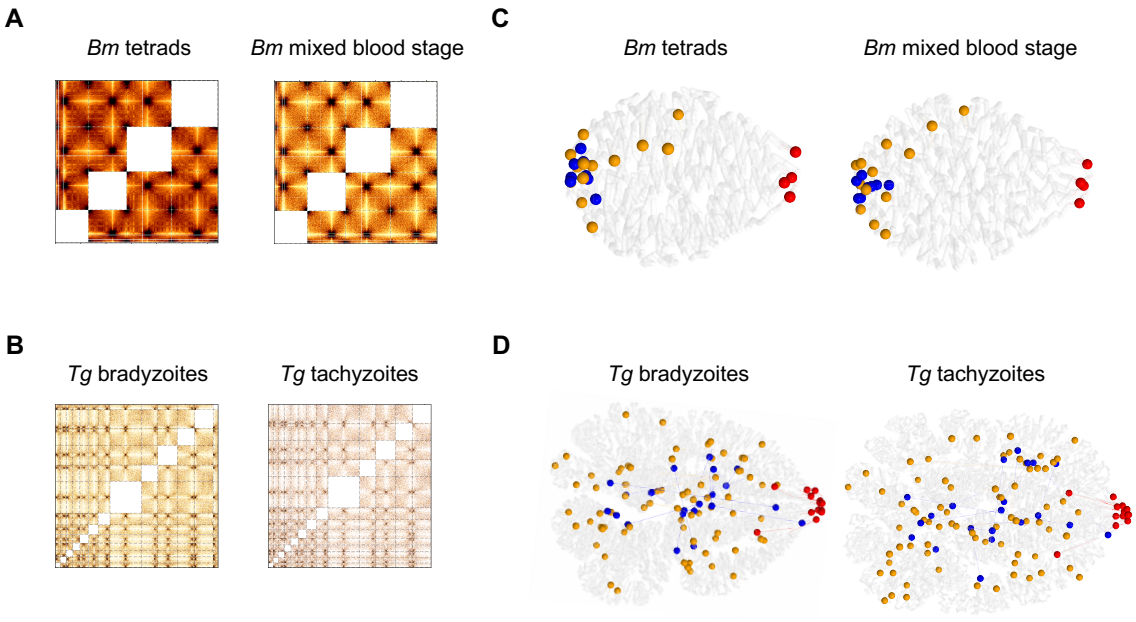
60. Ralph SA, Scheidig-Benatar C, & Scherf A (2005) Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. *Proc Natl Acad Sci U S A* 102(15):5414-5419.

61. Bensch S*, et al.* (2016) The Genome of Haemoproteus tartakovskyi and Its Relationship to Human Malaria Parasites. *Genome Biol Evol* 8(5):1361-1373.

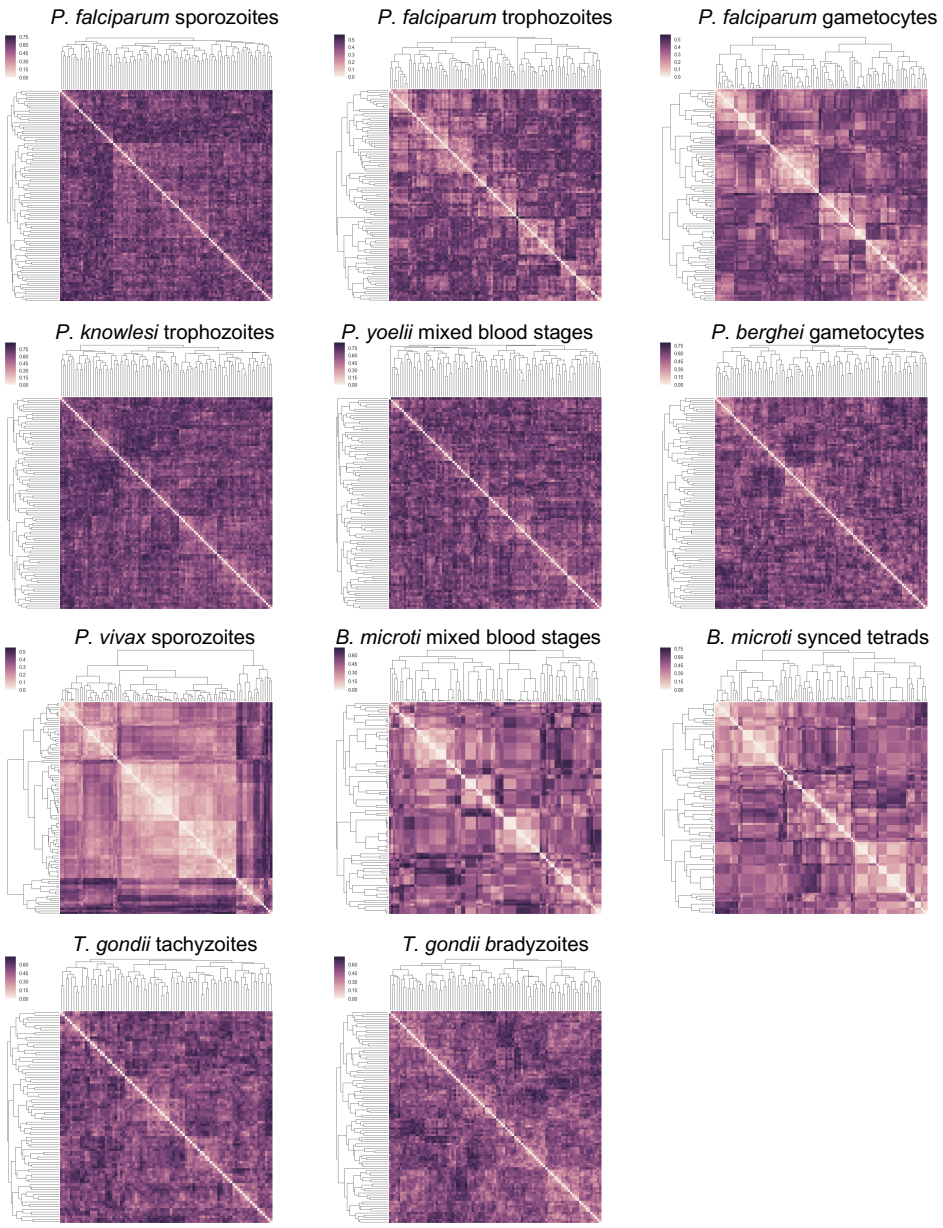
Supplemental Material



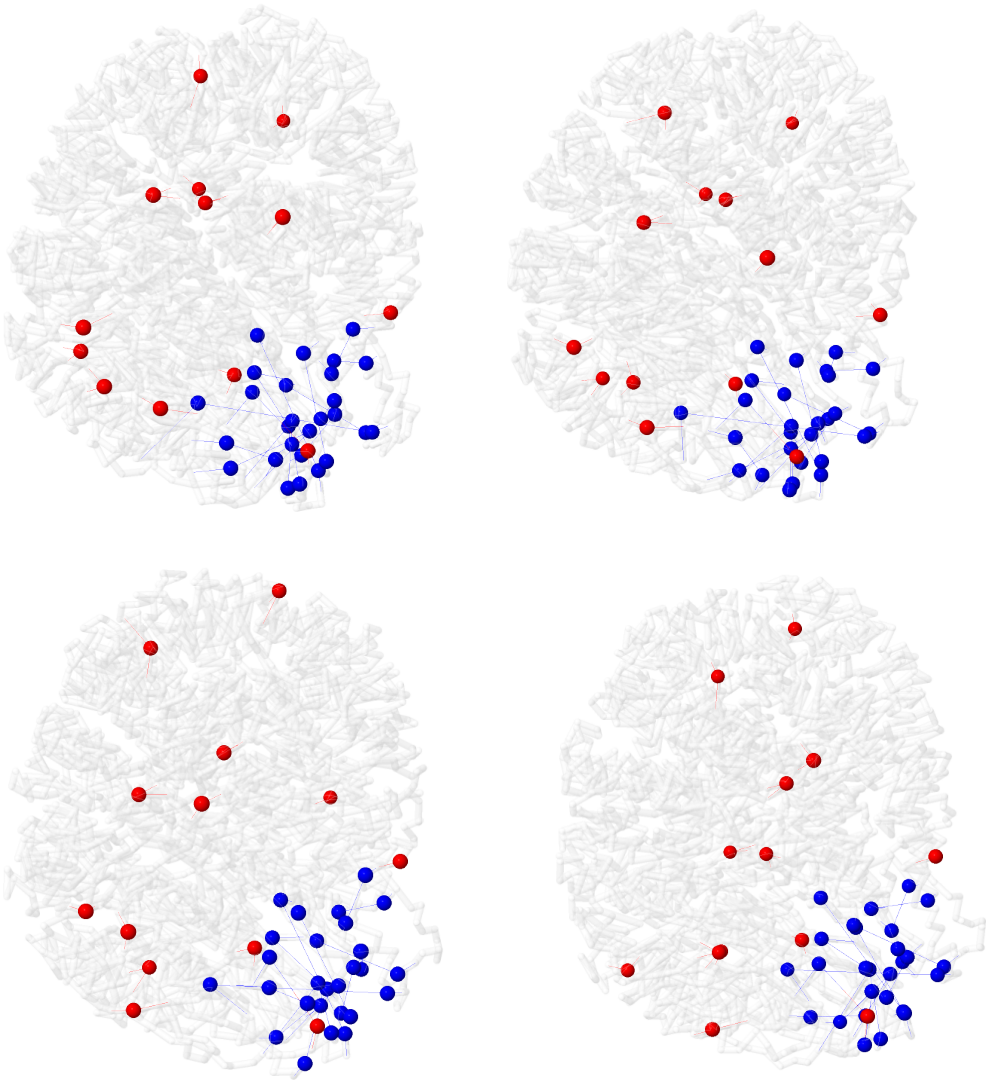
Supplementary Figure 2.1: Correlation between samples of the same organism. Concordance scores calculated using the package GenomeDISCO are shown for samples of the same organism from different stages, different strains, or biological replicates. Samples with boxed values were combined for downstream analysis.



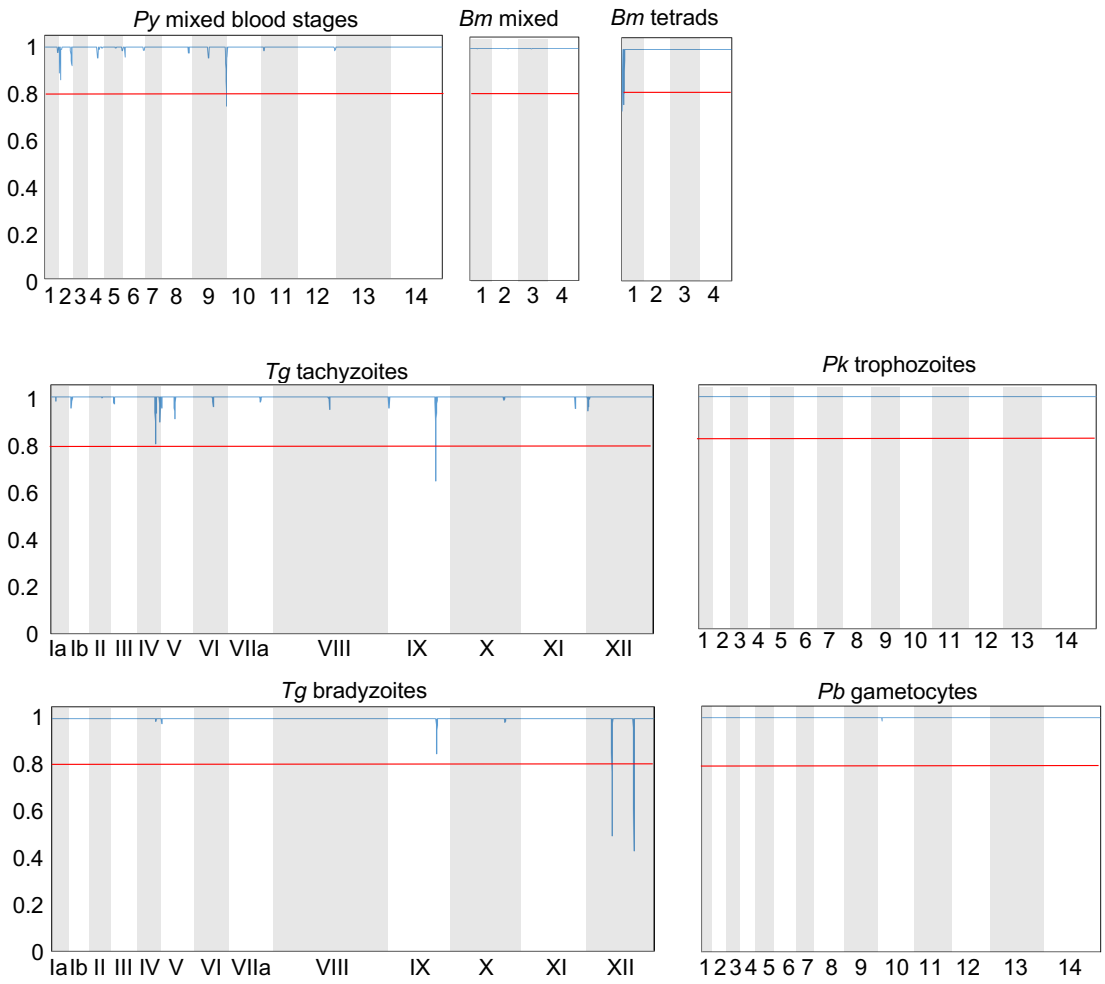
Supplementary Figure 2.2: Hi-C data and 3D genome modeling. A-B) Normalized interchromosomal contact count heatmaps at 10-kb resolution for *B. microti* (A) and *T. gondii* (B) samples. Chromosomes are lined up in numerical order starting with chr1 and chrIa, respectively, in the bottom left corner. Individual chromosomes are delineated by dashed lines. Intrachromosomal contacts are not displayed, hence the white squares along the diagonal of each heatmap. C-D) Representative 3D models for genomes of for *B. microti* (C) and *T. gondii* (D) samples. Chromosomes are shown as transparent white ribbons. Centromeres are indicated with red spheres, telomeres with blue spheres and virulence genes with orange spheres. Contact count patterns and genome organization are very similar for samples from the same organism.



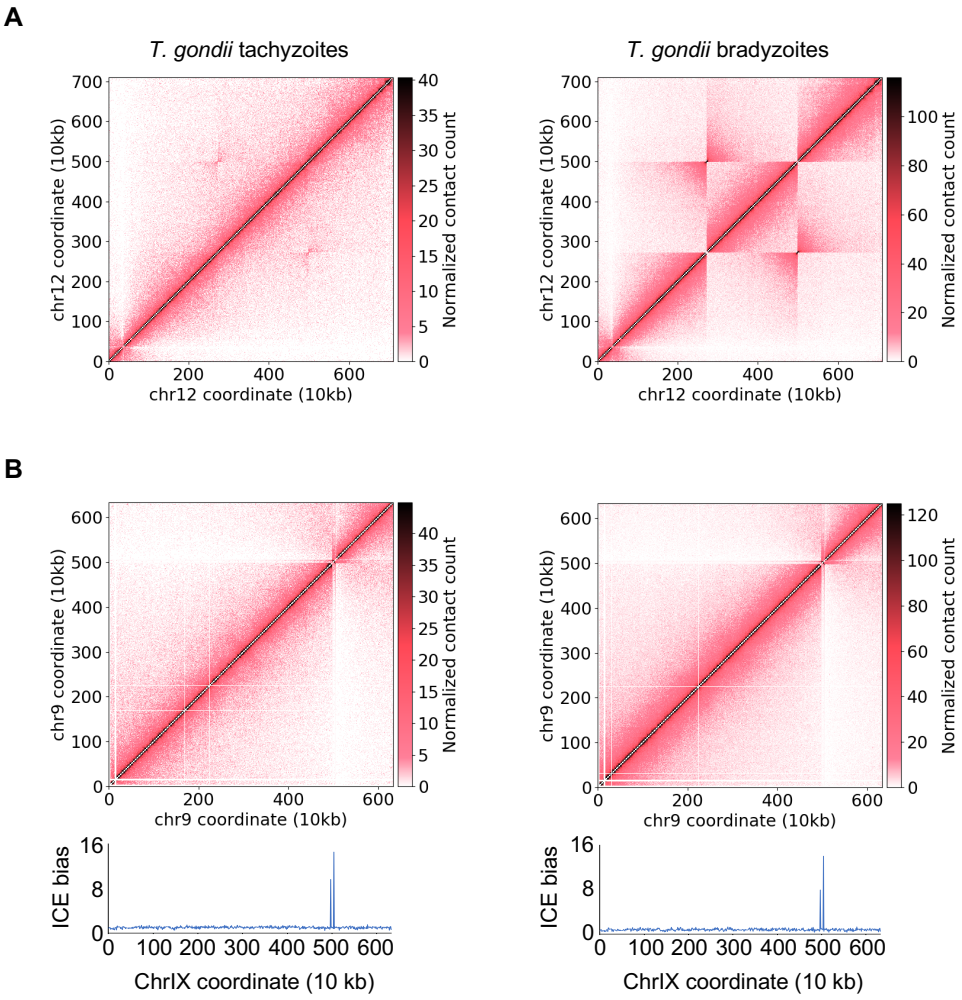
Supplementary Figure 2.3: Robustness of 3D models to random initializations of PASTIS. For each sample, 100 random initializations of the 3D coordinates were used to generate 3D models. The resulting 100 structures for each sample were transformed and clustered, and pairwise disparity scores were visualized in hierarchically clustered heatmaps.



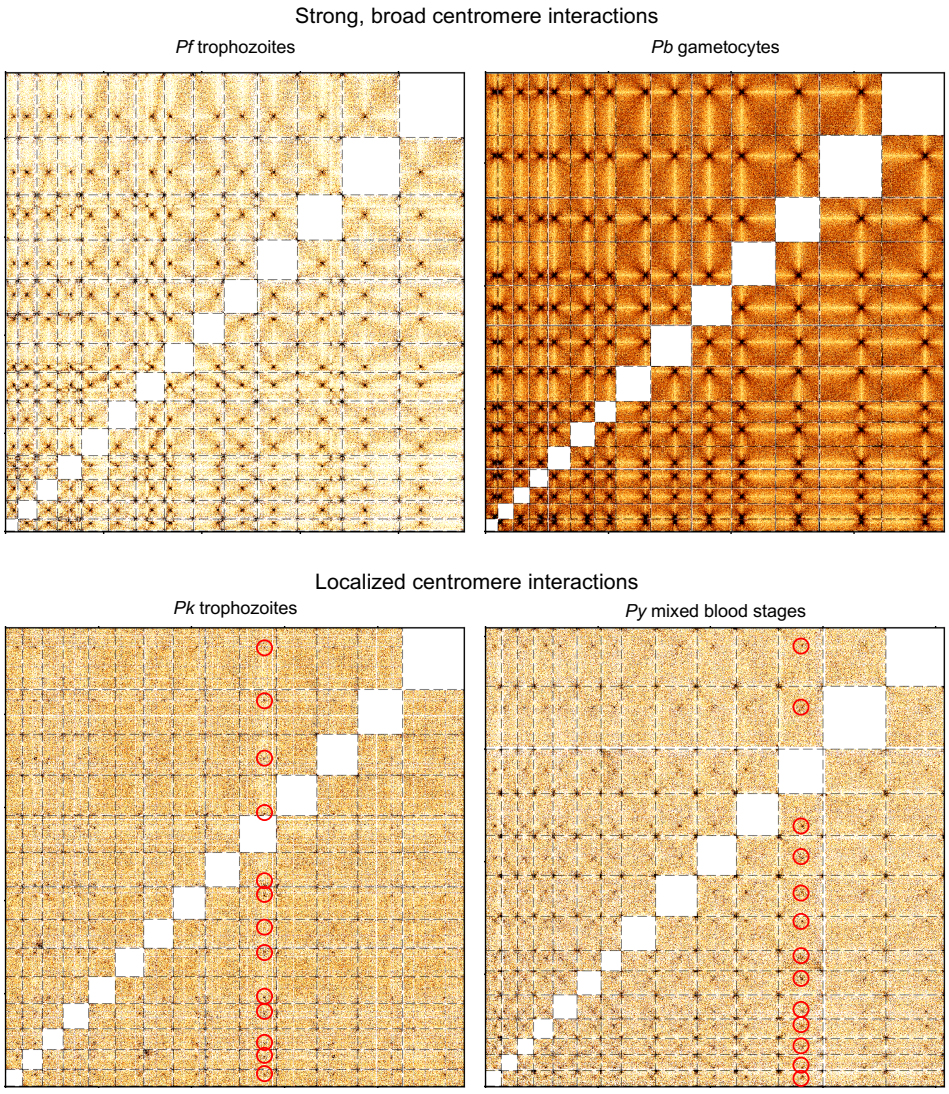
Supplementary Figure 2.4: Representative *P. vivax* 3D models from different initializations. For *P. vivax*, the 100 random initialization of 3D genome modeling resulted in four major clusters of genome structures. Representative models from each of these four clusters are shown here. The main hallmarks of genome organization are preserved in all four 3D models. Centromeres are indicated in red while telomeres are colored in blue.



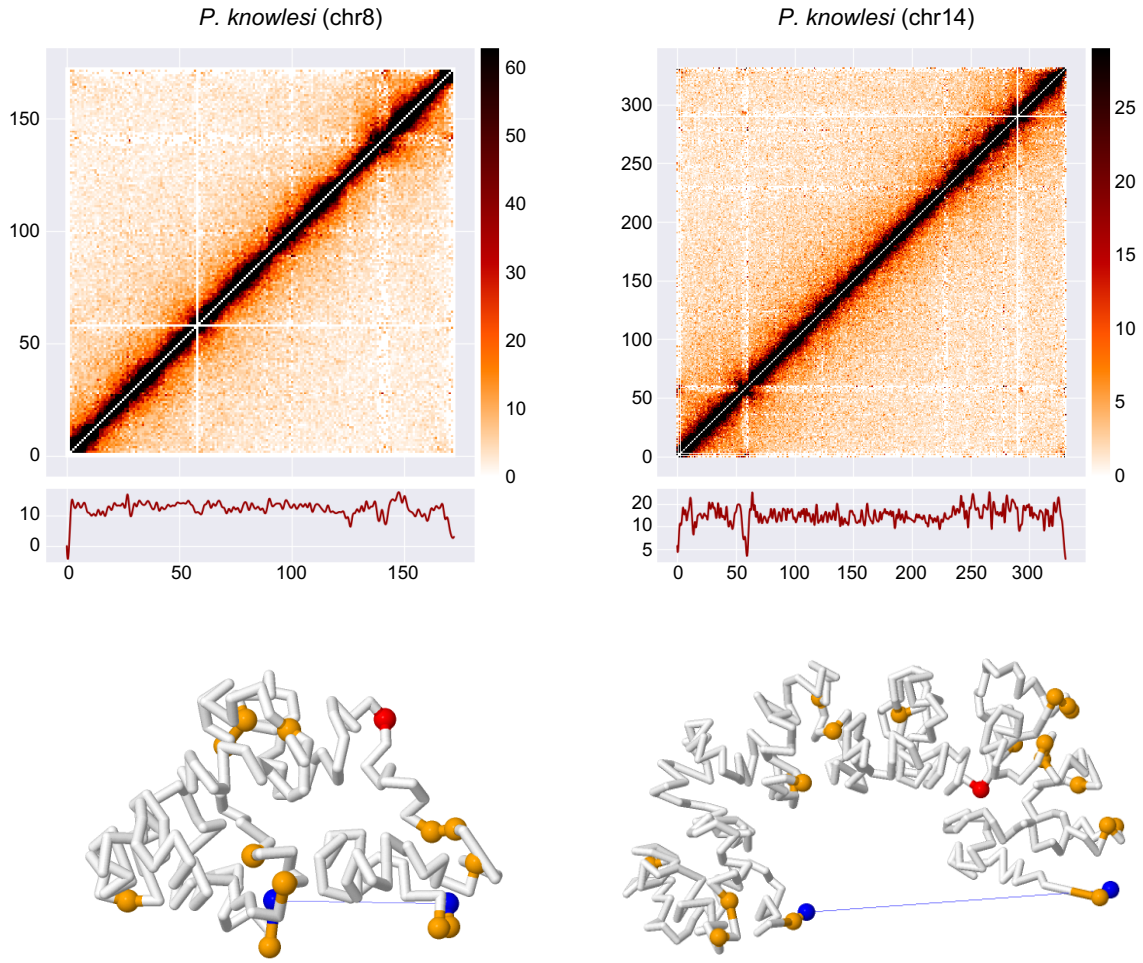
Supplementary Figure 2.5: Misassembly metrics for all samples used in this study. The misassembly metric $S_{observed}/S_{expected}$ was plotted for all chromosomes. To avoid artifacts induced by mappability issues, the score for bins with <50% mappability was set at 1. The threshold for misassembly issues was set at 0.8 and is highlighted with a red line. Chromosomes are indicated with grey and white shading. Results for *P. falciparum* and *P. vivax* samples are included in a recent publication (10).



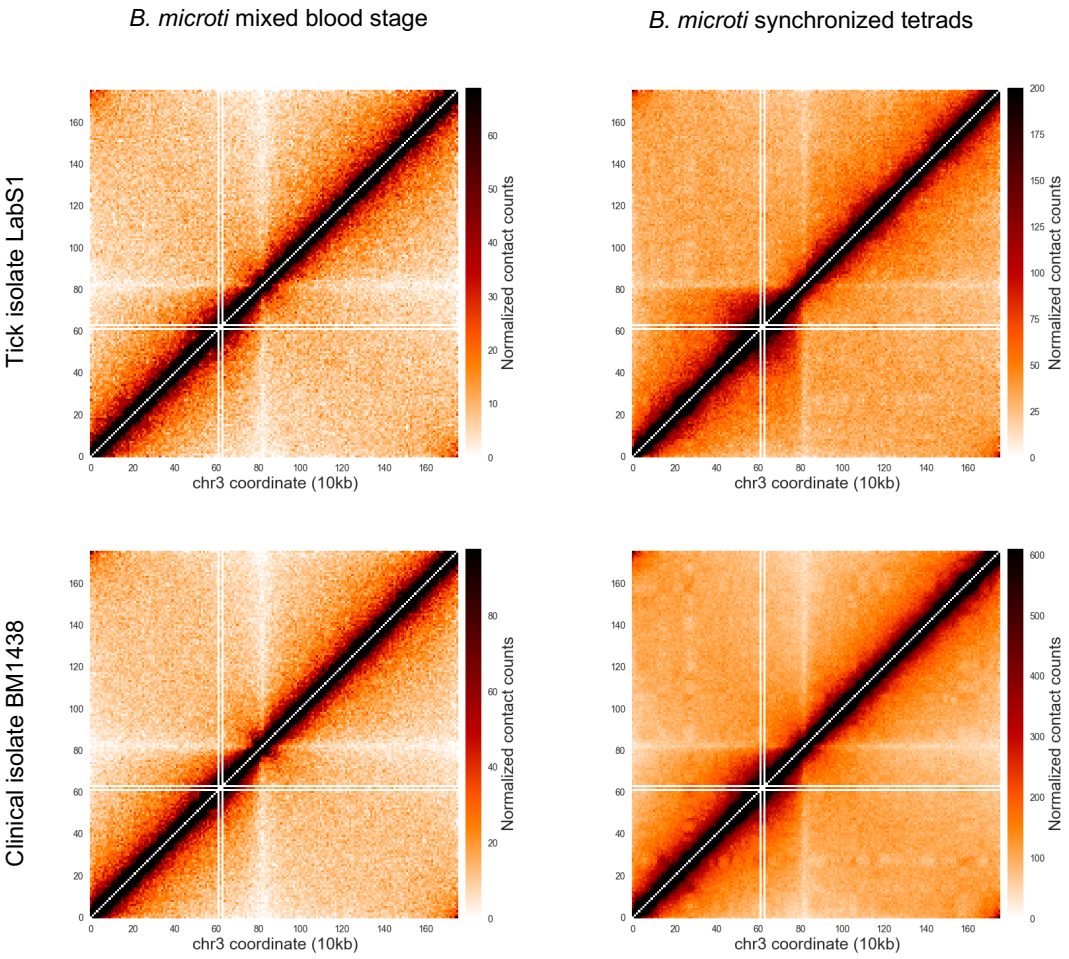
Supplementary Figure 2.6: Detection of inversion and duplication events in the *T. gondii* genome. A) Inversion in chrXII observed in ~10% of the population of tachyzoites (wild-type ME49 strain) and ~100% of the population of bradyzoites (transgenic ME49 strain) used in this study. B) Possible amplification of genome sequences in bins 498 and 505 of chrIX. The normalized contact 10-kb resolution heatmaps show an aberrant signal around bin 500 (top). The ICE bias tracks show a higher than expected number of interactions for bins 498 and 505 (bottom).



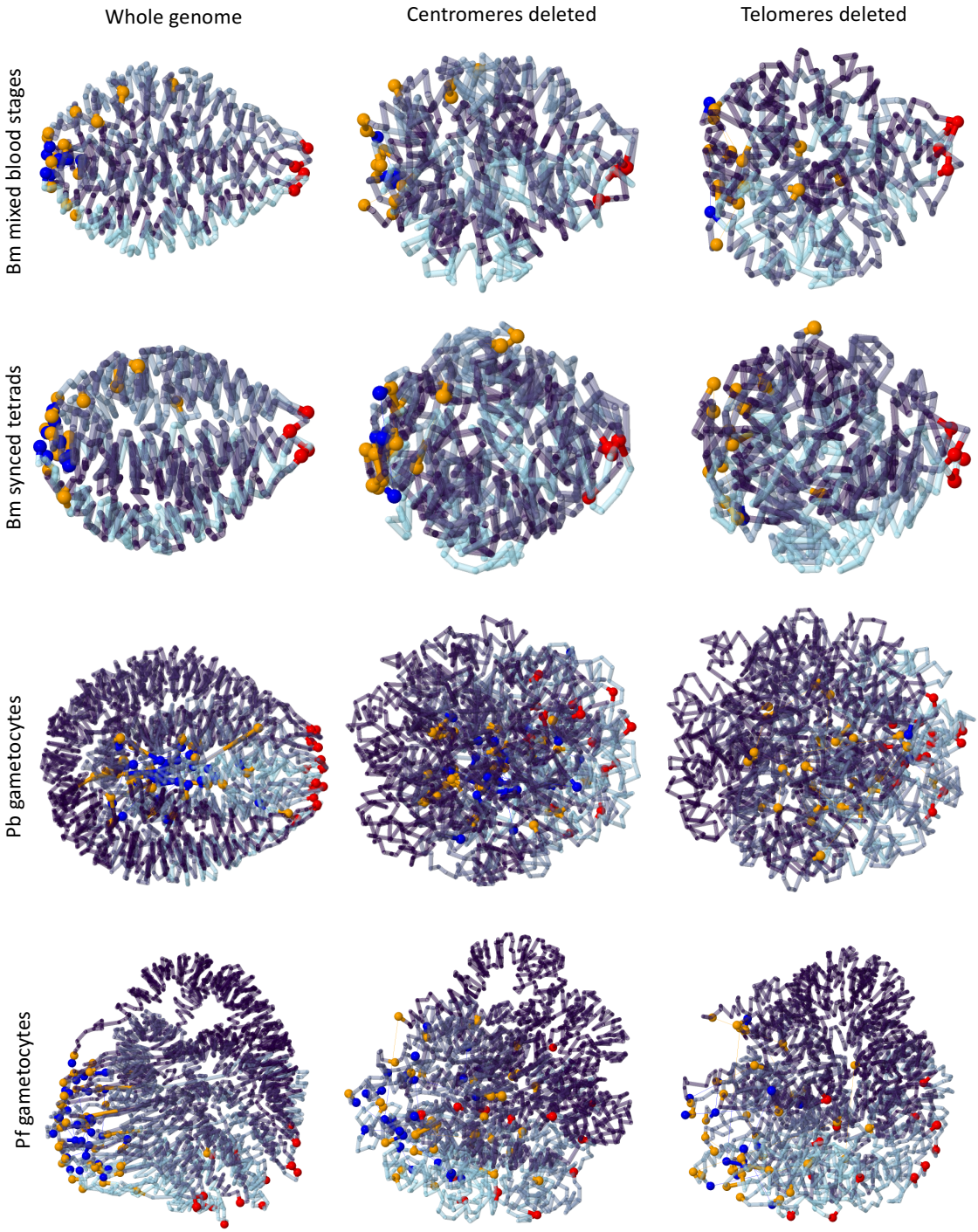
Supplementary Figure 2.7: Differences in centromere interaction patterns. Interchromosomal contact heatmaps of *P. falciparum* and *P. berghei* show strong centromere interactions that do not only involve the centromere itself, but also the surrounding regions (top row). In *P. knowlesi* and *P. yoelii*, centromere interactions are more localized (bottom row). Examples of centromeric interactions are indicated by red circles for one chromosome in each of these two organisms.

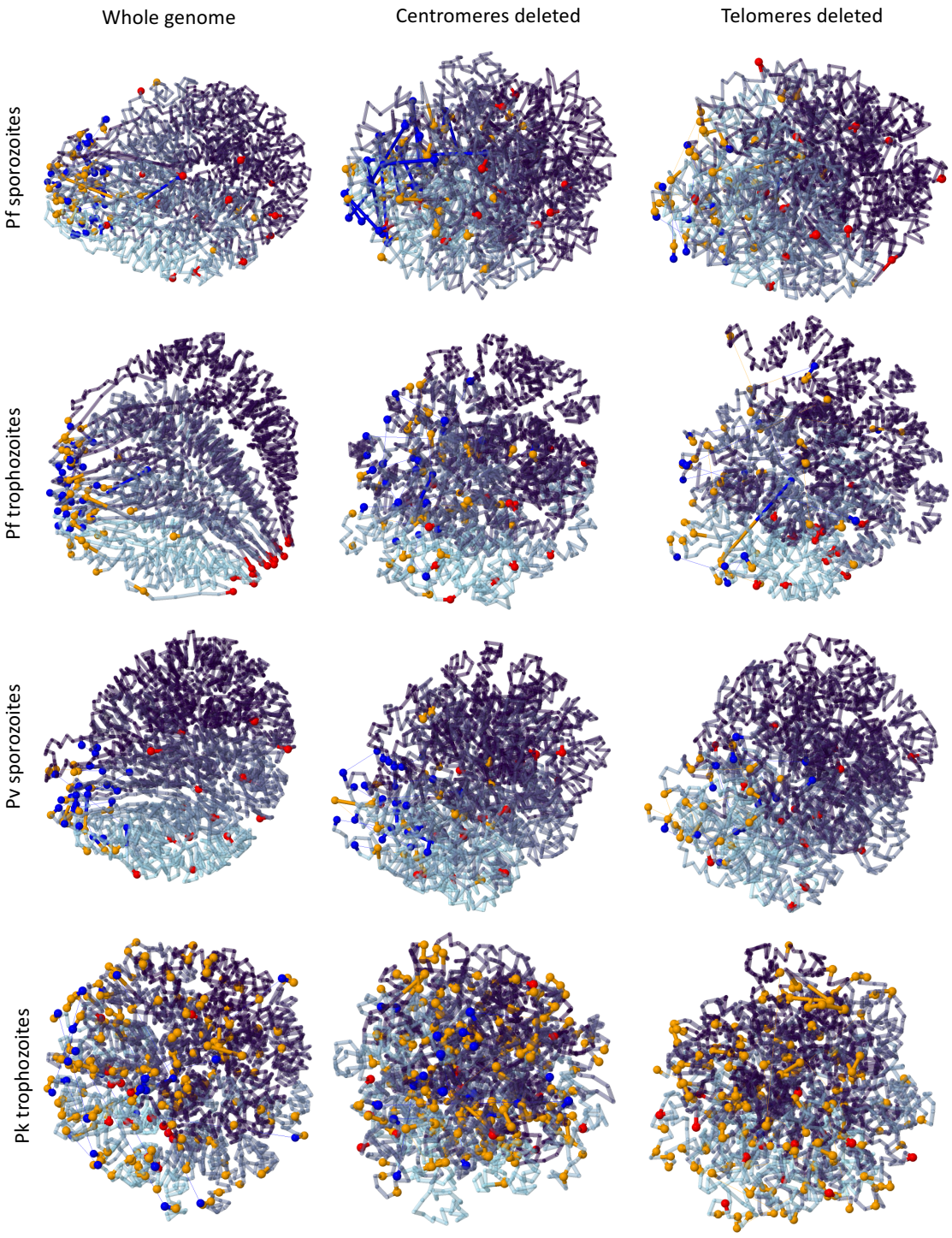


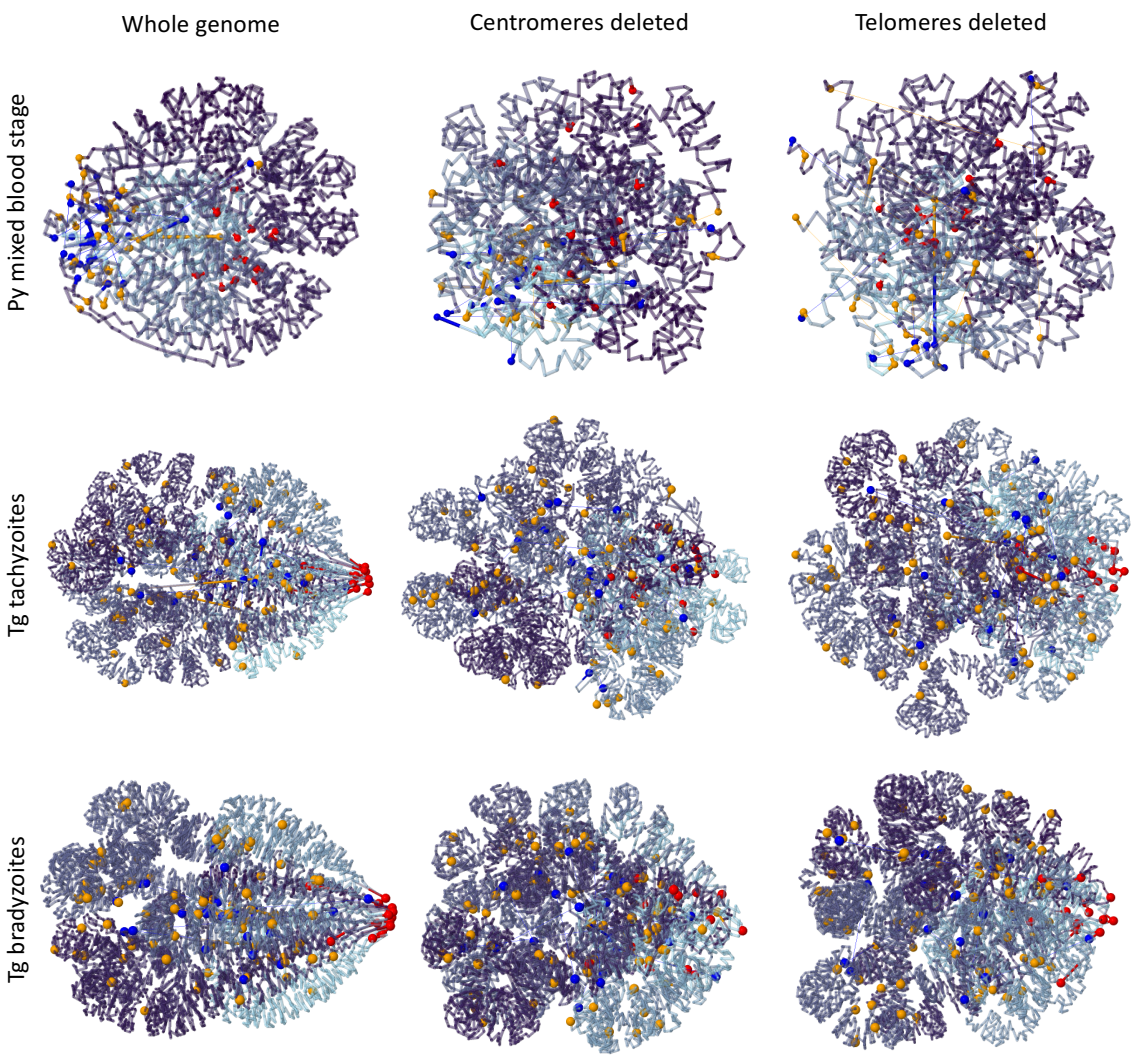
Supplementary Figure 2.8: Formation of domain-like structures and chromosome loops by *SICAvar* genes. Top row: normalized intrachromosomal contact count heatmaps at 10-kb resolution for *P. knowlesi* chr8 and chr14. Middle row: quantification of domain formation. For each bin, the average contact counts with the 10 upstream and the 10 downstream bins is plotted. Domains are characterized by a depletion of contact counts with surrounding bins. Bottom row: individual chromosome conformation extracted from the 3D model of the full genome.



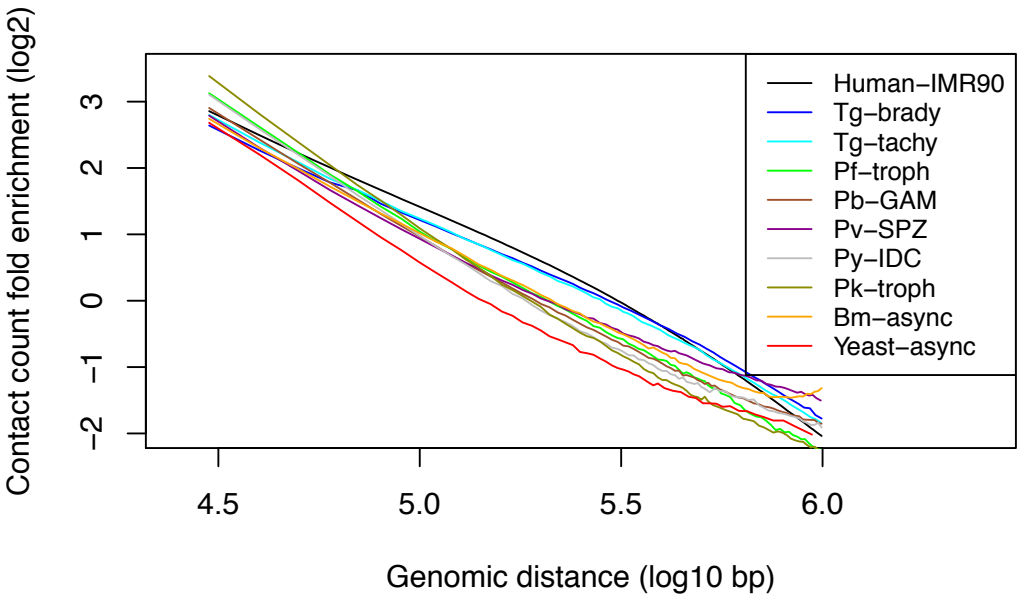
Supplementary Figure 2.9: Comparison between Hi-C data obtained from nonsynchronous and synchronous *B. microti* cultures. Normalized contact heatmaps of chromosome 3 at 10-kb resolution obtained from mixed blood stage culture (left) and from synchronized tetrads (right). The synchronized sample shows contact count patterns that are not visible in the nonsynchronous sample, emphasizing the requirement for the use of tightly synchronized samples to detect stage-specific, cell-cycle dependent or otherwise transient interactions.


Whole genome
Centromeres deleted
Telomeres deleted
Bm mixed blood stages
Bm synced tetrads
Pb gametocytes
Pf gametocytes


Whole genome
Centromeres deleted
Telomeres deleted
Pf sporozoites
Pf trophozoites
Pv sporozoites
Pk trophozoites



Supplementary Figure 2.10: Effect of deleting hallmarks of genome organization on the 3D model. For each sample, 3D models are shown for the full-length genome (left), the genome in which all bins containing centromeric sequence were deleted (center), and the genome in which 4 bins on all chromosome ends were deleted (right).



Supplementary Figure 2.11: Contact count fold enrichment as a function of genomic distance for all organisms included in this study. As references, data from yeast and human Hi-C experiments have also been included. The relationship between contact count probability and genomic distance in *T. gondii* is more similar to the human genome than to other apicomplexan parasites.

Supplementary Table 2.1: Numbers of sequence reads generated in Hi-C experiments and valid interaction pairs retained for downstream analyses.

Merged Sample	Strains/Conditions/Replicates	Raw paired-end reads	Aligned Pairs	Unique Valid Pairs	Percentage of valid pairs		Publication
Babesia microti mixed blood stage	async field strain	49,853,688	11,307,312	2,875,817	5.77%	*	This work
	async lab strain	70,606,554	22,106,352	4,427,842	6.27%	*	
	TOTAL	**120,460,242**	**33,413,664**	**7,303,659**	6.06%	*	
Babesia microti synchronized schizonts	sync field strain 1	24,031,449	13,160,646	9,443,749	39.30%		This work
	sync field strain 2	29,384,425	17,111,462	12,977,882	44.17%		
	sync lab strain 1	68,824,017	40,499,550	29,764,580	43.25%		
	sync lab strain 2	60,108,185	33,887,411	21,601,639	35.94%		
	TOTAL	**182,348,076**	**104,659,069**	**73,787,850**	40.47%		
Plasmodium berghei gametocytes	smc2 mut	82,330,220	45,142,471	25,598,041	31.09%		This work
	smc4 mut	37,336,346	16,429,012	9,565,745	25.62%		
	WT	48,635,834	27,470,015	10,110,737	20.79%		
	TOTAL	**168,302,400**	**89,041,498**	**45,274,523**	26.90%		
Plasmodium falciparum gametocytes	Early gametocytes	150,070,973	100,049,648	61,713,026	41.12%		Bunnik et al *Nature Communications* 2018
	Late gametocytes	75,338,560	46,265,208	17,594,303	23.35%		
	TOTAL	**225,409,533**	**146,314,856**	**79,307,329**	35.18%		
Plasmodium falciparum sporozoites	Sporozoites 1	130,539,746	15,697,495	1,740,915	1.33%	*	Bunnik et al *Nature Communications* 2018
	Sporozoites 2	51,856,178	7,652,584	925,159	1.78%	*	
	TOTAL	**182,395,924**	**23,350,079**	**2,666,074**	1.46%	*	
P. falciparum trophozoites	**TOTAL**	**60,567,770**	**25,123,926**	**6,968,959**	11.51%		Ay et al *Genome Research* 2014
P. knowlesi trophozoites	**TOTAL**	**33,857,803**	**15,882,302**	**11,243,440**	33.21%		Lapp et al *Parasitology* 2017
Plasmodium vivax sporozoites	Sporozoites 1	138,095,734	25,859,628	9,976,858	7.22%	*	Bunnik et al *Nature Communications* 2018
	Sporozoites 2	160,681,970	15,860,169	6,145,121	3.82%	*	
	TOTAL	**298,777,704**	**41,719,797**	**16,121,979**	5.40%	*	
P. yoelii mixed blood stage	PanK1 mut	58,044,078	18,912,344	5,533,378	9.53%	*	This work
T. gondii bradyzoites	**TOTAL**	**117,710,987**	**54,058,885**	**40,176,192**	34.13%		This work
T. gondii tachyzoites	**TOTAL**	**33,116,669**	**19,060,530**	**13,075,539**	39.48%		This work

* low percentages in these samples are due to host DNA contamination

CHAPTER 3: Comparative Analysis of Chromatin-Associated Proteins in Apicomplexans and Characterization of a Plant-Like Nuclear Lamina Protein in *Plasmodium*

Gayani Batugedara[1,4], Xueqing M. Lu[1,4], Anita Saraf[2], Anthony Cort[1], Steven Abel[1], Jacques Prudhomme[1], Laurence Florens[2], Evelien M. Bunnik[3] and Karine G. Le Roch[1*]

[1]Department of Molecular Cell and Systems Biology, University of California Riverside, Riverside, CA 92521, USA

[2]Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO, USA

[3]Department of Microbiology, Immunology & Molecular Genetics, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

[4] These authors contributed equally to this work.

A version of this chapter was submitted to *Cell Reports*, 2018.

Preface

From yeast to human cells, genome organization in eukaryotes has a tight relationship with gene expression. As outlined in the previous two chapters, it is now becoming increasingly apparent that apicomplexan parasite nuclear architecture and chromosome dynamics are quite complex and are likely regulated by many proteins. Therefore, to understand global chromatin organization within the nucleus and its affect on parasite development, it is critical to identify and characterize the components that regulate these processes. To investigate parasite proteins and protein complexes maintaining and regulating nuclear architecture, we undertook comparative genomics analysis using twelve distinct eukaryotic genomes. We identified conserved and apicomplexan parasite-specific chromatin-associated domains (CADs) and proteins (CAPs). We next use complementary methods to experimentally capture chromatin-bound proteins and in particular validate and characterize two candidate CAPs in *Plasmodium falciparum*. Identification of parasite-specific proteins that are critical for regulating genome structure could serve as potential drug targets that can disrupt parasite development with high specificity and low toxicity to the host.

Abstract

Proteins interacting with DNA are known to mediate fundamental processes such as gene expression, DNA replication, and maintenance of genome integrity. Accumulating evidence suggests that the chromatin of apicomplexan parasites such as the human malaria parasite, *Plasmodium falciparum*, are highly structured at the three-dimensional level, and this structure provides an epigenetic mechanism for gene expression regulation. To investigate how parasite nuclear architecture is being maintained and regulated, we undertook comparative genomics analysis using twelve distinct eukaryotic genomes. We identified conserved and apicomplexan parasite-specific chromatin-associated domains (CADs) and proteins (CAPs). Using a mass spectrometry approach that specifically enriches for chromatin-bound proteins, we experimentally capture 987 CAPs during the *P. falciparum* erythrocytic stages. We further characterize, at the cellular and molecular levels, two of our candidate proteins including a homolog of the CROWDED-like NUCLEI (CRWN) protein, a plant-like protein that is functionally analogous to animal nuclear lamina. Collectively, our results provide the most comprehensive overview of CAPs in apicomplexan parasites, and contribute significantly to our understanding of the complex molecular components regulating chromatin structure and genome architecture in these deadly parasites.

Introduction

Apicomplexans are obligate protozoan parasites that are responsible for a wide range of diseases in humans and animals. Among apicomplexan parasites that infect humans,

Plasmodium spp., the causative agents of malaria, have the largest health and economic impact. While the most prevalent and deadly human malaria parasite, *P. falciparum,* is responsible for an estimated 445,000 deaths per year [1], *P. vivax* and *P. knowlesi* also infect humans. Other apicomplexan parasites relevant to humans include *Babesia microti* [2], the causative agent of human babesiosis, a malaria-like illness endemic in the US but with worldwide distribution, and *Toxoplasma gondii*, the causative agent of toxoplasmosis, an opportunistic parasite that cause infections in immunocompromised individuals [3]. The kinetoplastid parasites, another class of human-relevant protozoan pathogens also contributing to the global burden of disease includes *Trypanosoma brucei* (causing African sleeping sickness, *Trypanosoma cruzi* (causing Chagas disease), and *Leishmania* spp (causing leishmaniasis) [4-6].

Despite continued efforts to prevent parasitic infections, treatment of affected individuals still remains one of the primary means of reducing parasitic mortality and morbidity. Given the absence of an FDA-approved vaccine and parasite resistance to most current antiparasitic drugs [7, 8], there is a desperate need for new therapeutic approaches. One promising strategy towards the development of novel and effective antiparasitic compounds is to gain a better understanding of mechanisms regulating gene expression in these parasites. Since the publication of the first parasite genomes such as the *P. falciparum* genome that was published over 15 years ago [9], researchers have attempted to explore the transcriptional machinery of parasites in detail. The distinct developmental stages of the parasite life cycles are characterized by coordinated changes in gene

expression [10-16]. However, a surprisingly low number of specific transcription factors have been identified in their genomes [17-19] and in particular, only a few stage-specific transcription factors have been validated in *Plasmodium* spp or *T. gondii* [20-26]. Therefore, the coordinated cascade of transcripts observed throughout the parasite life cycles is unlikely to be regulated only by this limited collection of specific transcription factors, which suggests that additional components and mechanisms, such as post-transcriptional [27-31], translational and post-translational regulation [27, 32, 33] as well as change in chromatin structure, may control the expression of the predicted thousands of genes in apicomplexan parasites.

Recently, several groups, including ours, have developed chromosome conformation capture (3C) coupled to next generation sequencing methods (called Hi-C) as a way of understanding spatial organization of the nucleus and its role in regulating biological processes [34-36]. Using the latest Hi-C methodology, our lab has determined the three-dimensional (3D) nuclear architecture of *P. falciparum* throughout its life cycle [37, 38]. Our work showed that parasite chromatin loosens following the invasion of a red blood cell allowing for gene expression, and re-packs prior to the next cycle. Additionally, Western blot and mass spectrometry analyses show a significant depletion of all histone proteins at the trophozoite stage [39], supporting that a significant amount of transcriptional activity happens during the trophozoite stage. This suggests that changes in chromatin structure may control, at least partially, gene expression and parasite development. Additionally, our Hi-C results demonstrate that the parasite nucleus is

highly organized. In particular, telomere ends of the chromosomes cluster together in heterochromatin area(s) in close proximity to the nuclear membrane while the centromeres cluster at the opposite of a large heterochromatin cluster, much like the genome organization observed in the similarly sized budding and fission yeasts [40, 41]. However, the parasite genome exhibits a higher degree of organization than the budding yeast genome as genes involved in immune evasion (e.g., *var, rifin* and *stevor* genes) add a striking complexity and act as structural elements that shape global genome architecture [37]. Very recently, we performed additional Hi-C experiments to generate 3D genome models for four other *Plasmodium* species and two related apicomplexan parasites including *T. gondii* and *B. microti*. We demonstrated that spatial genome organization in apicomplexan parasites are often constrained by the colocalization of virulence genes that have a unique effect on chromosome folding. We also identified a potential link between genome organization and gene expression in more virulent pathogens (Bunnik et al., manuscript submitted). Based on these observations, we hypothesize that architectural proteins that interact with chromatin and have a strong influence on genome organization may represent novel targets for antiparasitic interventions.

Architectural proteins involved in maintenance of chromatin structure have been studied in organisms ranging from yeast to human [42]. Among these proteins are RNA polymerase III-associated factor TFIIIC, cohesin, condensin, and CCCTC-binding factor (CTCF) [42-45]. CTCF is an insulator protein conserved in vertebrates that is enriched at chromosome domain boundaries and interacts with the nuclear lamina [46]. Some of

these components have homologs in the parasite genomes but only a few have been characterized at the functional level. Furthermore, many conserved chromatin architectural proteins or chromatin-associated proteins (CAPs) involved in chromatin maintenance (e.g. lamina proteins) are missing in parasite genomes [47]. As an example, lamina proteins in metazoans are essential for many nuclear functions including nuclear shape maintenance and architecture, chromatin organization, DNA replication, transcription, and cell cycle progression [46, 48]. While absent in apicomplexan parasites, these proteins are likely to have distant homologs in the parasite genomes as they are critical for nuclear membrane organization and chromatin structure regulation.

Although most of our understanding of proteins involved in chromatin structure and their functions comes from studies on model organisms, their importance in the development and virulence of apicomplexan parasites including *Plasmodium* has recently been appreciated for a small number of candidates [49-52]. Yet a large number of these proteins still need to be identified and functionally characterized. Given the potential roles of CAPs in almost all aspects of parasite biology, we performed a comprehensive computational and comparative genomics approach to generate an extended atlas of chromatin-associated domains (CADs) in twelve eukaryotic organisms and have identified conserved, as well as apicomplexan parasite-specific, euglenid parasite-specific, unicellular yeast-specific and several multicellular organism-specific CADs. We provide functional annotations based on homology, domain organization, domain clustering, and expression pattern analysis. More specifically, using a set of advanced

bioinformatics tools, we identified 1,190 well-defined and putative CAPs in the *P. falciparum* genome, of which 162 proteins (13.6%) have been previously described as having chromatin-related functions [53]. In addition, we employed an unbiased chromatin proteomics approach termed Chromatin Enrichment for Proteomics (ChEP) to experimentally validate some of our candidate CAPs. ChEP has been successfully used to identify chromatin-bound molecules and predict their function and regulation in a number of organisms [54-56]. We further validated two of our candidate proteins including a CROWDED-like NUCLEI (CRWN) protein using standard cellular and molecular approaches. CRWN proteins, present in plant nuclei, resemble animal and fungal lamina. However, the machinery and processes in which these proteins participate appear to be evolutionarily distinct from their animal counterparts [57, 58]. In plants, CRWN proteins are essential for viability and play diverse roles in both heterochromatin organization and the control of nuclear morphology [59]. Identification and characterization of these homologs in apicomplexan parasites could reveal novel and exciting targets for drug discovery. Altogether, our results validate that while mechanisms regulating chromatin structure in apicomplexan parasites are most likely complex, some of our newly identified candidates have plant-like features that could be specifically targeted by new antimalarial strategies. A better understanding of these CAPs will not only provide a comprehensive view of the complex molecular components that control chromatin organization and genome architecture in these deadly parasites, but will also assist the identification of novel targets for therapeutic strategies.

Results

Comparative analysis of chromatin-associated domains in apicomplexan parasites and other eukaryotes

To obtain a list of all possible domains associated with chromatin-bound proteins, hereafter referred to as chromatin-associated domains (CADs), we first filtered NCBI Conserved Domain Database and Pfam database for domains with chromatin-related cellular functions including heterochromatin regulation, chromosome organization, nucleic acid binding, and histone modifications. A total of 3,870 CADs was found regardless of their organism sources. We next performed a genomic comparative analysis of CAPs among a variety of organisms, including three apicomplexan parasites (*Plasmodium falciparum*, *Plasmodium vivax* and *Toxoplasma gondii*), three euglenid parasites (*Trypanosoma brucei, Trypanosoma cruzi,* and *Leishmania major*), two unicellular organisms (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), and four multicellular organisms (*Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, and Arabidopsis thaliana*). Since not all genomes have been annotated at the same level, manual curation of the CAD list (n = 3,870) was avoided to eliminate bias and to ensure a fair comparison between organisms. Therefore, we systematically performed HMM searches on the proteomes of the above organisms to find proteins that contained any of the 3,870 CADs. Generally, we observed relatively similar numbers of CAPs in apicomplexan parasites (~22% of the full proteome), slightly lower number of CAPs in euglenid parasites, and a higher number of CAPs in *S. cerevisiae* and higher eukaryotes (Figure 3.1A). The relatively low number of CAPs identified in euglenids is

most likely due to distinct modes of gene regulation in these parasites [60], as control of protein expression happens almost exclusively at the post-transcriptional level. As a whole, our data underline that chromatin structure in apicomplexans is complex, possibly due to the intricacies of molecular mechanisms regulating transcription at the chromatin level in these parasites. The presence of large families of virulence genes in apicomplexa may also add some additional complexity [37]. In higher eukaryotes, increasing amounts of hierarchical chromatin elements such as compartments, topologically associating domains (TADs) and insulated domains have been described [61]. It is therefore not surprising that these more complex eukaryotes require more CAPs to regulate their chromatin structure.

To identify functional differences in chromatin-associated processes, the candidate CADs were clustered based on their relative abundance in all investigated species (Figure 3.1B). Each cluster was then analyzed for domain-associated Gene Ontology (GO) enrichment. Twelve distinct clusters were obtained. Clusters 1-3 contained CADs that are relatively abundant in apicomplexans, of which the domains in cluster 1 were almost exclusively enriched in apicomplexan parasites. These domains showed enrichment for GO terms associated with nucleic acid binding and specifically AP2 domain-containing transcription factors (Figure 3.1C). While highly abundant in apicomplexans, AP2 domain containing proteins were also abundant in plant species. AP2 family of transcription factors play an essential role in floral development in *A. thaliana* [62] and in *P. falciparum,* proteins containing AP2 binding domains (ApiAp2) have been identified

as sequence-specific transcription factors [63] and are believed to be master regulators of transcription during parasite development [20, 25]. The enrichment of AP2 domains in apicomplexan parasites and *A. thaliana* in our cluster analysis further validates our classification methods. Cluster 1 also includes the PHD_OBE1_like domain that is present in *A. thaliana* PHD finger proteins (Figure 3.1C). This domain enriched in plant species, is also conserved in *Plasmodium* species, which highlights presence of plant-like protein domains in apicomplexan parasites. In addition, cluster 1 harbored the RCC1 domain (Figure 3.1C), which is found in chromosome condensation regulating proteins. While these proteins are conserved among unicellular and multicellular organisms, a highly divergent ortholog of the Regulator of Chromosome Condensation 1 (RCC1) that is critical for parasite pathogenesis has been identified in apicomplexan parasites [64]. These atypical apicomplexan RCC1 proteins show different arrangements of RCC1 domains compared to their higher eukaryotic RCC1 orthologs. Cluster 2 contained CADs that were abundant in all unicellular organisms. This cluster showed enrichment for GO terms associated with DNA-binding and polymerase activity. The PRK09603 domain enriched in this cluster is found in bifunctional DNA-directed RNA polymerase II proteins (Figure 3.1C). This protein is found in many prokaryotic members and is the single type of RNA polymerase that performs transcription in bacteria [65]. In line with its abundance in bacteria, this domain seems to also be conserved in unicellular eukaryotes. Cluster 3 contained CADs that were enriched in apicomplexan species and yeast but not in euglenid parasites. Enriched domains (LSM, PRK00737 and RRM2_SRSF4) are found in proteins involved with RNA processing and splicing (Figure

3.1C). Unlike their unicellular counterparts, euglenid parasites transcribe their protein-coding genes into polycistronic RNAs [60] and processes the RNA through a special mechanism termed trans-splicing where exons from two different primary transcripts are ligated [66]. This suggests that RNA processing and splicing proteins in *Trypanosomes* and *Leishmania* are divergent from those proteins found in apicomplexans and yeast where the primary mechanism of RNA processing is via cis-splicing [67].

Chromatin-associated domains in cluster 4 were abundant in all twelve organisms. One of the major domains enriched in this cluster was the SMC N-terminal domain (Figure 3.1C). SMC domain-containing proteins are a large family of ATPases that play a role in many aspects of chromosome organization [68]. Different SMC subunits makeup cohesin and condensin complexes, and these proteins play an essential role in chromosome assembly and segregation [69, 70]. In particular, condensin promotes chromosome compaction [69], while cohesin facilitates sister chromatid separation during mitosis and meiosis [70]. Enrichment of this domain across many eukaryotes including apicomplexans, kinetoplastids, yeast, and vertebrates, suggests that proteins containing SMC domains are highly conserved and are important for maintaining and regulating chromatin structure in a wide variety of organisms. Among cluster 5 were CADs mostly abundant in kinetoplastids. This cluster harbored the domain PSP1, which was originally observed in yeast and was reported to be involved in suppressing mutations in the DNA polymerase alpha subunit (Figure 3.1C) [71]. The PSP1 motif has been found to be conserved at the C-terminal end of *Crithidia fasciculata* cycling sequence binding

proteins (CSBP), which binds to sequence elements present in mRNAs that accumulate during the cell cycle [72]. Homologs of CSBP proteins were found only among the kinetoplastids, however whether these proteins share a functional relationship with the yeast PSP1-containing proteins has yet to be determined. Cluster 11 contained CADs most abundant in non-protozoan organisms. This cluster represented GO terms associated with chromosome and chromatin structure. A representative domain in this cluster, NHP6B/HMG, is found in High-Mobility Group B proteins (HMGBs) (Figure 3.1C). HMGBs are highly abundant DNA-binding proteins that are involved in many nuclear functions including chromatin remodeling, transcription, recombination and DNA repair [73, 74]. The C-terminal acidic tail typical of metazoan HMGBs [75, 76], which regulates the DNA-binding characteristics of the HMGB-box domains, is missing from most unicellular organisms [77, 78]. This suggests that protozoan HMGBs might have additional sequence characteristics that enable HMGBs to bind DNA, which are absent from higher eukaryotes.

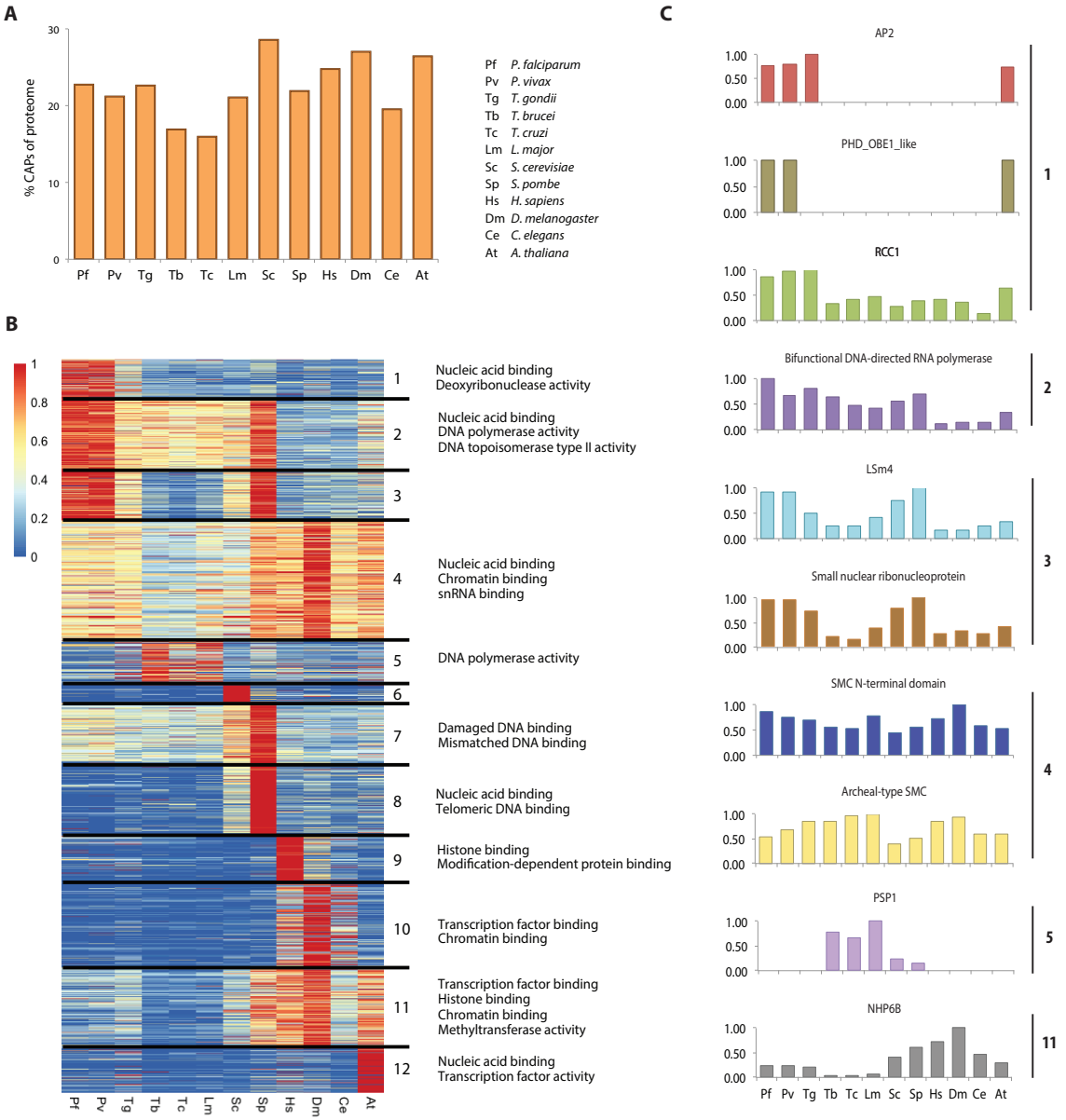

A
% CAPs of proteome
30
20
10
0
Pf Pv Tg Tb Tc Lm Sc Sp Hs Dm Ce At
Pf P. falciparum
Pv P. vivax
Tg T. gondii
Tb T. brucei
Tc T. cruzi
Lm L. major
Sc S. cerevisiae
Sp S. pombe
Hs H. sapiens
Dm D. melanogaster
Ce C. elegans
At A. thaliana
B
1
0.8
0.6
0.4
0.2
0
1 Nucleic acid binding
Deoxyribonuclease activity
2 Nucleic acid binding
DNA polymerase activity
DNA topoisomerase type II activity
3
4 Nucleic acid binding
Chromatin binding
snRNA binding
5 DNA polymerase activity
6
7 Damaged DNA binding
Mismatched DNA binding
8 Nucleic acid binding
Telomeric DNA binding
9 Histone binding
Modification-dependent protein binding
10 Transcription factor binding
Chromatin binding
11 Transcription factor binding
Histone binding
Chromatin binding
Methyltransferase activity
12 Nucleic acid binding
Transcription factor activity
Pf Pv Tg Tb Tc Lm Sc Sp Hs Dm Ce At
C
AP2
PHD_OBE1_like
RCC1
1
Bifunctional DNA-directed RNA polymerase
2
LSm4
Small nuclear ribonucleoprotein
3
SMC N-terminal domain
Archeal-type SMC
4
PSP1
5
NHP6B
11
1.00
0.50
0.00
Pf Pv Tg Tb Tc Lm Sc Sp Hs Dm Ce At

Figure 3.1: Relative abundance of chromatin-binding domains in apicomplexan parasites compared to other eukaryotes. (A) Relative abundance of CAPs in the full proteome of various organisms (B) k-means clustering of the relative abundance of CADs among 12 organisms. The CAD abundance was first normalized for each organism by proteome size and then scaled to the CAD frequency with the highest relative abundance of that CAD. A subset of the Gene Ontology (GO) enriched terms associated with the Pfam domains (false discovery rate, FDR<0.01) for each cluster are shown on the right. (C) Selection of CADs and their relative abundance among all 12 organisms.

In silico identification and classification of chromatin-associated proteins in P. falciparum

To study and characterize CAPs in *P. falciparum*, we next searched the *Plasmodium* proteome for all known protein domains (chromatin and non-chromatin related) using both hmmscan (HMMER v3.16) and NCBI Reversed Position Specific BLAST (RPS-BLAST). We then selected all parasite proteins containing any of the 3,870 CADs. As a result, we identified a total of 1,114 candidate CAPs (20.1% of *P. falciparum* proteome, n= 5548) covering 1,629 unique CADs (42% of total CADs). Out of these 1,114 candidate CAPs, 460 proteins were identified using RPS-BLAST, 82 proteins were identified using hmmscan, and 572 proteins were identified using both approaches (Figure 3.2A, Supplemental file 3.1). Additionally, 76 *Plasmodium* proteins that lacked any of the CADs, but have chromatin-associated functions based on their protein

annotation, were manually added to the final chromatin-associated protein candidates list. Among the final list of 1,190 candidate CAPs, 162 proteins (13.6%) have been previously described as having chromatin-related functions in the parasite [53], 877 (73.7%) have non-chromatin related annotation and 151 proteins (12.7%) are unknown proteins for which functions have yet to be discovered.

To better define *Plasmodium* CAPs, we further characterized the chromatin-associated domains that they carried. The most abundant CADs were structural maintenance of chromosome domains (SMC) (83 members) and domains from the serine/threonine kinase catalytic family (STKC) associated with cell cycle progression, chromatin remodeling, DNA binding, transcription regulation, or other nuclear activities (total 77 members). Transcription or mRNA processing-associated RNA-binding domains (RRM; 73 members), catalytic domain of the dual-specificity protein kinases (PKc; 64 members), DEAD box helicase domains (63 members), WD40 domains (58 members), polyadenylate binding domains (PABP; 44 members), and splicing factor, CC1-like domains (SFD-CC1; 43 members) were also found to be abundant in the parasite's genome along with other domains such as the anaphase-promoting complex unit (ANAPC4), AP2 transcription factor domains, small nuclear ribonucleoprotein domains, and GTP-binding nuclear protein domains (Figure 3.2B). When investigating the structural features of these highly abundant CAPs (domains present in 15 or more candidate proteins), we observed that many of these CAD-containing proteins consist of either a single CAD in combination with non-chromatin-related domains or multiple

CADs in combination with non-chromatin-related domains (Figure 3.2B). In other words, CADs were rarely observed to be the uniquely defining domain(s) of a protein. This finding suggests that CAPs may likely have multiple functional roles in the biology of the parasite.

To explore the potential function of these chromatin-associated candidate proteins, we further categorized these proteins based on their function using gene annotations obtained from PlasmoDB (Figure 3.2C). We found that a large number of the proteins are likely to be nucleic acid binding proteins (n=172, 14.5%) or proteins involved in transcriptional regulation (n=151, 12.8%). Among these protein candidates are high mobility group B1-B4 proteins (PF3D7_1202900, PF3D7_0817900, PF3D7_1205800, and PF3D7_1359200), proteins that form the transcription initiation factor TFIID subunit (PF3D7_0934100, PF3D7_0522200, and PF3D7_0929000), and known transcriptional regulators such as Sir2A/B proteins (PF3D7_1328800 and PF3D7_1451400) and transcriptional coactivator ADA2 (PF3D7_1014600). Another large group of the proteins are found to be structurally or functionally related to chromatin and chromosome structure (n=146, 12.3%). These proteins include histones, histone modification proteins, nucleosome assembly proteins, chromatin remodeling proteins, and chromosomal structural proteins. RNA processing proteins (n=122, 10.3%), such as RNA polymerase subunits, proteins involved in splicing, and cleavage and polyadenylation proteins were also abundantly found in our list. Furthermore, proteins involved with protein modification (n=100, 8.4%), DNA methylation, replication and repairs (n=99, 8.3%), cell

or nuclear division (n = 79, 6.6%), and ribonucleoprotein or ribosome-associated proteins (n=76, 6.4%) were also reported. A relatively smaller portion of proteins were found to be associated with GTP/ATP binding (n=47, 3.9%), protein transportation or signaling activity (n= 31, 2.6%), and the ubiquitin pathway (n=18, 1.5%). About 150 proteins (12.4%) were found to contain domains that are typically present in a variety of proteins that may or may not be related to chromatin. Some example proteins belonging to this group were zinc finger proteins, WD repeat-containing proteins, and ion-binding proteins. Lastly, we looked into the overall gene expression of the identified candidate CAPs. We observed that the expression level of candidate CAPs are similar to the expression level of transcription factors suggesting that the majority of these chromatin-associated candidate proteins are likely to be important for facilitating transcription in the parasite (Figure 3.2D).

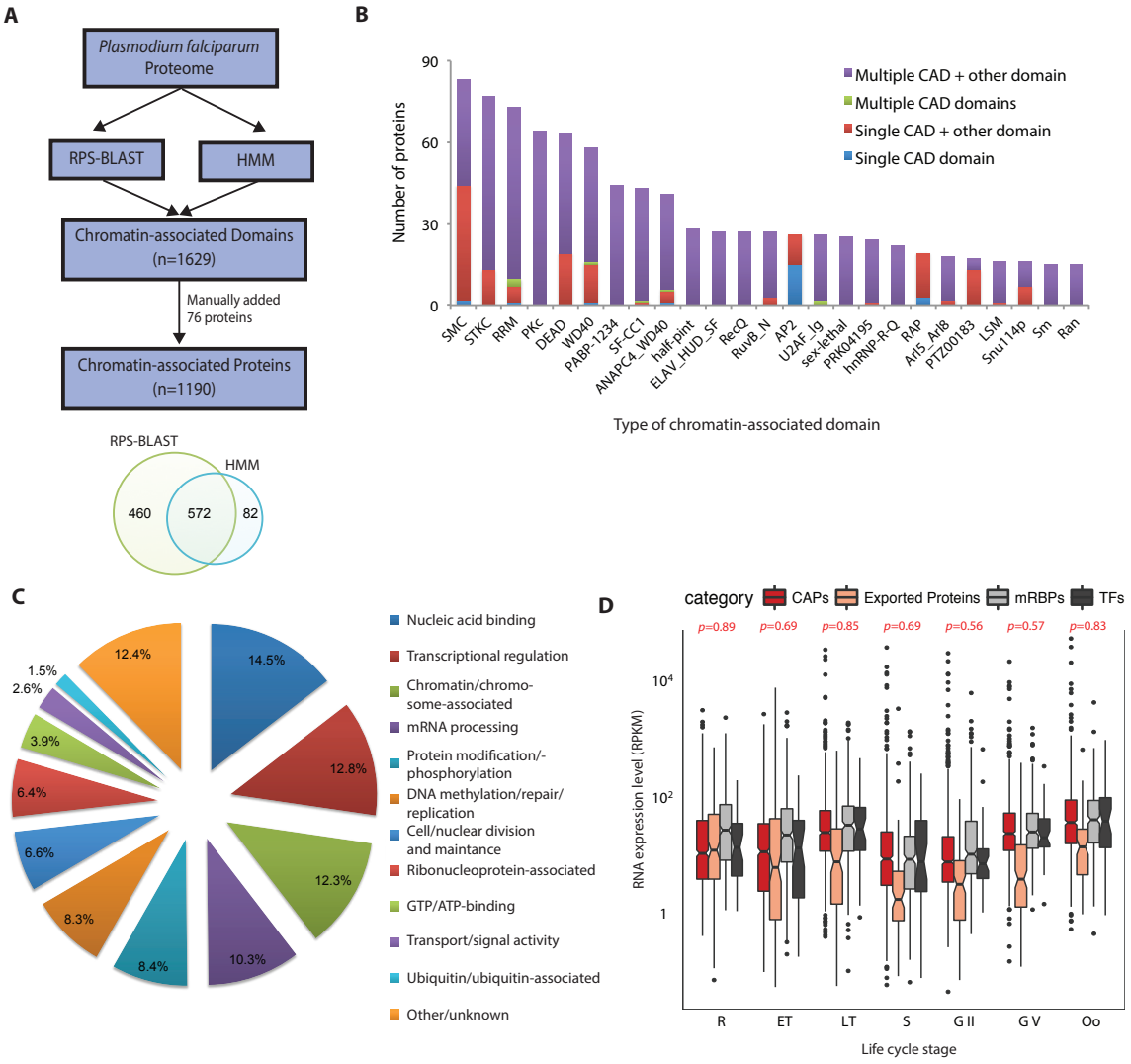

A
Plasmodium falciparum Proteome
RPS-BLAST
HMM
Chromatin-associated Domains (n=1629)
Manually added 76 proteins
Chromatin-associated Proteins (n=1190)
RPS-BLAST
HMM
460
572
82
B
Number of proteins
90
60
30
0
Multiple CAD + other domain
Multiple CAD domains
Single CAD + other domain
Single CAD domain
SMC
STKC
RRM
PKc
DEAD
WD40
PABP-1234
SF-CC1
ANAPC4_WD40
half-pint
ELAV_HUD_SF
RecQ
Ruv8_N
AP2
U2AF_lg
sex-lethal
PRK04195
hnRNP-R-Q
RAP
Arf5_Arl8
PTZ00183
LSM
Snu114p
Sm
Ran
Type of chromatin-associated domain
C
14.5%
12.8%
12.3%
10.3%
8.4%
8.3%
6.6%
6.4%
3.9%
2.6%
1.5%
12.4%
Nucleic acid binding
Transcriptional regulation
Chromatin/chromo-some-associated
mRNA processing
Protein modification/-phosphorylation
DNA methylation/repair/replication
Cell/nuclear division and maintance
Ribonucleoprotein-associated
GTP/ATP-binding
Transport/signal activity
Ubiquitin/ubiquitin-associated
Other/unknown
D
category
CAPs
Exported Proteins
mRBPs
TFs
p=0.89
p=0.69
p=0.85
p=0.69
p=0.56
p=0.57
p=0.83
RNA expression level (RPKM)
10⁴
10²
1
R
ET
LT
S
G II
G V
Oo
Life cycle stage

Figure 3.2: Overview of chromatin-associated proteins in *Plasmodium falciparum*. (A) In silico methodology. A total of 1190 proteins (20.1% of proteome) in *P. falciparum* were predicted to be chromatin-associated proteins, covering 1629 unique chromatin associated domains. (B) Characterization of chromatin-associating domains (CADs) that were found in eight or more candidate CAPs. (C) Classification of candidate chromatin-associated proteins (CAPs) based on their annotation or associated domains. (D) Gene expression comparison between candidate CAPs and other classes of proteins during different developmental stages of the parasite life cycle. P-value for comparing the statistically significant differences in average expression levels between candidate CAPs and transcription factors are reported. (R) ring, (ET) early trophozoite, (LT) late trophozoite, (S) schizont, (G II) early gametocytes at stage II, (G V) late gametocytes at stage V, (Ook) ookinete.

Experimental validation of chromatin-associated proteins in P. falciparum

To validate our in silico identified candidate CAPs, we next used a method designed to isolate, in a genome-wide manner, all proteins associated with chromatin. This methodology, termed Chromatin Enrichment for Proteomics (ChEP) (Figure 3.3A), was adapted from published studies on human and mouse cell lines [55]. Briefly, parasites were extracted at the ring, trophozoite, or schizont stages and cross-linked with formaldehyde to preserve protein-nucleic acid interactions. Optimization experiments showed that longer cross-linking time at a higher temperature, compared to the standard chromatin immunoprecipiation conditions (10 mins at room temperature), was necessary

to obtain sufficient cross-linking between DNA and proteins for the ChEP methodology (data not shown). Parasite nuclei were then extracted in the presence of RNase A to avoid enrichment of proteins associated with nascent RNA rather than directly with chromatin. Non-cross-linked proteins were washed away using a highly denaturing buffer. As a negative control, we isolated proteins from the cytoplasmic fraction. We observed a clear enrichment by Western blot analysis, of the nuclear marker histone H3 in the nuclear fractions and cytoplasmic marker aldolase in the cytoplasmic fractions following the ChEP protocol (Figure 3.3B).

Proteins isolated from the parasite nucleus following the ChEP methodology as well as the cytoplasmic fraction were analyzed using multidimensional protein identification technology (MudPIT). Two biological replicates, as well as two technical replicates were analyzed for each intraerythrocytic stage. We identified a total of 940, 934, and 1,016 proteins in the nuclear ChEP fraction at the ring, trophozoite and schizont stages, respectively (Figure 3.3C and Supplemental file 3.2). We then compared the nuclear ChEP proteins to the control cytoplasmic proteins to identify proteins enriched in the ChEP samples. We identified 468, 494, and 639 proteins that were detected at $\geq$ 2-fold enrichment in the nuclear fraction at the ring, trophozoite and schizont stages, respectively (Figure 3.3D and Supplemental file 3.3). The experimentally detected candidate CAPs enriched by $\geq$ 2-fold abundance were compared to the computationally detected CAPs. A total of 469 candidate CAPs that were captured by the MudPIT

analysis validated 39% of the CAP candidates identified computationally (Supplemental figure 3.1).

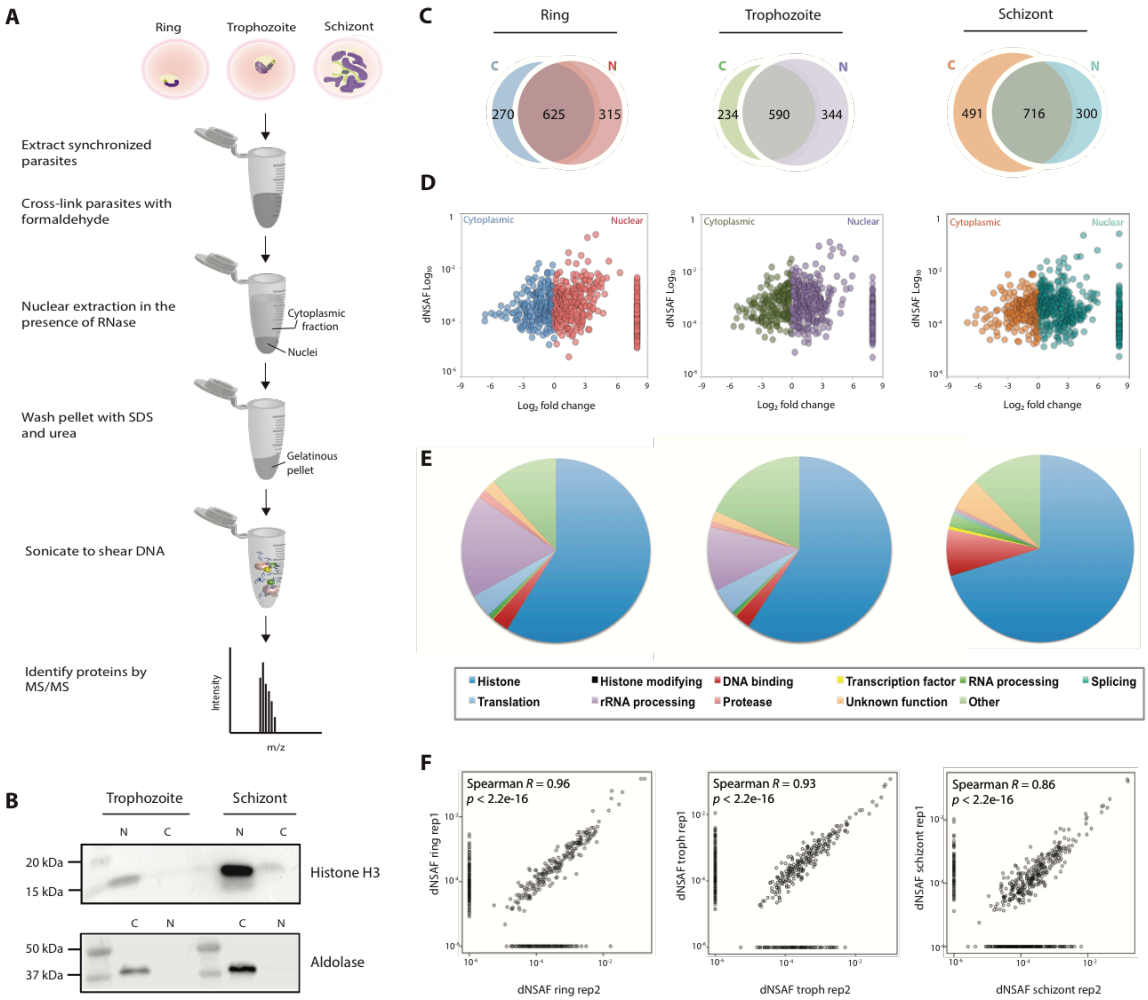


Figure 3.3: Chromatin enrichment for proteomics (ChEP). (A) Outline of the ChEP procedure. (B) Validation of protein enrichment in the nuclear fraction from ChEP by Western blot analysis (top gel: lane 1- marker; 2- trophozoite nuclear fraction; 3- trophozoite cytoplasmic fraction; 4- blank; 5- schizont nuclear fraction; 6- schizont cytoplasmic fraction, bottom gel: lane 1- marker; lane 2- trophozoite cytoplasmic fraction; lane 3- trophozoite nuclear fraction; lane 4- marker; lane 5- schizont cytoplasmic fraction; lane 6- schizont nuclear fraction). Western blots show an enrichment for histone H3 in the nuclear fractions and enrichment of aldolase in the cytoplasmic fractions following the ChEP protocol. (C) Number of proteins identified in the nuclear and cytoplasmic fractions of the ChEP sample at ring, trophozoite, and schizont stages. (D) Semi-quantitative proteomic analysis of the ChEP samples demonstrating that ChEP enriches for chromatin-associated proteins in the nuclear fraction. (E) Proteins that are enriched 2-fold or more in the nuclear fraction are classified according to their function. (F) Correlation in protein abundance between replicate experiments.

The ChEP-identified CAPs showed strong enrichment for GO terms associated with typical chromatin-associated processes such as histone and histone modifying, DNA binding, transcription, RNA processing and splicing (Figure 3.3E). Proteins functioning in translation-related processes were also enriched in the nuclear ChEP sample, which points to the existence of nuclear translation in the parasite [79, 80]. Additionally, ribosomal RNA (rRNA) processing proteins were enriched in the ChEP sample and we

observed a larger number of these proteins at the ring and trophozoite stages (18% at ring and 11% at trophozoite stage, respectively) compared to the late schizont stage (4%). Ribosome biogenesis takes place in the nucleus [81], and considering the biology of the parasite, the majority of ribosomes will need to be assembled in preparation for the higher levels of translation that takes place at the later trophozoite and schizont stages. By adapting this novel ChEP protocol, we have also identified a large number of proteins with unknown function as likely interacting with chromatin (2% at ring and trophozoite stages and 5% schizont stage). In general, proteins detected in a given ring, trophozoite and schizont sample were also detected in its matching biological replicate. Additionally, calculation of Spearman rank coefficients showed that relative protein abundance correlated strongly between matching replicates (Spearman R = 0.96 at ring, 0.93 at trophozoite and 0.86 at schizont stages; Figure 3.3F). Furthermore, CAPs with higher relative abundance levels were more likely to be detected in the replicate experiments. This clearly demonstrates the reproducibility of our ChEP coupled to mass spectrometry methodology.

Functional validation of candidate chromatin-associated proteins in P. falciparum

Putative chromatin-bound protein candidates enriched using the ChEP methodology with high reproducibility were searched for the existence of even distantly related homologs using PSI-Blast HHPred [82]. Candidate proteins with domain homology for chromatin components were selected for further molecular and cellular characterization. To this end, proteins in the ChEP enriched fraction and annotated as *Plasmodium* proteins of

unknown function were BLASTed against protein domains known to be involved in nuclear function in metazoans, eukaryotic pathogens or plants such as nuclear lamina or lamina-like proteins, cohesin, condensin, CTCF insulator or insulator-like proteins. Our analysis identified two putative homologs (PF3D7_1325400 and PF3D7_1126700) of coiled-coil proteins that are among the nuclear matrix constituent proteins found in plants. In *A. thaliana,* these proteins are encoded by *CRWN* genes [83]. PF3D7_1126700 was more abundant in the ChEP sample at the schizont stage (dNSAF = 0.0011) compared to PF3D7_1325400 (dNSAF = 0.0004). However, PF3D7_1325400 was identified with higher confidence in the BLAST search (E-value = 0.01) and was therefore selected for further analysis. Hereafter, PF3D7_1325400 will be referred to as 'CRWN-like' protein. A second protein, PF3D7_0414000, annotated as structural maintenance of chromosome 3 (SMC3), was also selected for further validation. SMC3, a subunit of the cohesin complex, although annotated as such, has not yet been characterized in *P. falciparum*.

Custom antibodies were generated for each protein by designing peptide antigens targeting the C-terminal end of each protein (see methods). To validate these antibodies we first performed Western blots using nuclear and cytoplasmic protein lysates from mixed-stage *Plasmodium* parasites. We observed a clear enrichment of SMC3 and CRWN-like proteins in the nuclear fraction (Figure 3.4A, 3.4B). Our results validate the use of these custom antibodies to detect the *P. falciparum* CRWN-like (~419 kDa) and SMC3 (~140 kDa) proteins.

We further investigated the subcellular localization of the CRWN-like and SMC3 proteins in intraerythrocytic parasites using immunofluorescence assays (IFA). A single focus per nucleus was observed for the SMC3 protein at ring, trophozoite and schizont stages (Figure 3.4C). At all three asexual stages, the CRWN-like protein localized to the nuclear compartment (Figure 3.4D). In particular, we observed a single focus per nucleus at the ring and schizont stages. At the trophozoite stage, the number of foci varied, in line with the increased level of DNA replication and nuclear expansion that takes place during this stage. In *A. thaliana*, CRWN proteins localize to the nuclear periphery and play a role in regulating heterochromatin environments in the nucleus [83]. It is possible that the CRWN-like protein in *Plasmodium* is similarly localizing to the heterochromatin regions of the nucleus.

To specifically localize the SMC3 and CRWN-like proteins within the parasite nucleus, we next performed double immunofluorescence staining using a commercially available anti-H3K9me3 antibody (Millipore; 07-442-AF488). As all the antibodies were raised in the same host species, we performed sequential immunofluorescence staining to avoid cross reactivity between antibodies. Briefly, fixed and permeabilized parasites were first incubated with the custom anti-SMC3 or anti-CRWN antibodies followed by the corresponding fluorescent secondary antibody (see methods). After extensive washing, the parasites were incubated with the fluorescently conjugated anti-H3K9me3 antibody. At the ring and schizont stages, the SMC3 proteins localize to the nuclear periphery away

from the heterochromatin regions marked by H3K9me3 (Figure 3.4E). However, co-localization of H3K9me3 and CRWN-like proteins can be observed, although the localization of CRWN-like proteins appear to extend to regions adjacent to heterochromatin regions of the nucleus as well (Figure 3.4F). Further experiments will be needed to validate the exact function of this CRWN-like protein in heterochromatin maintenance during parasite development.

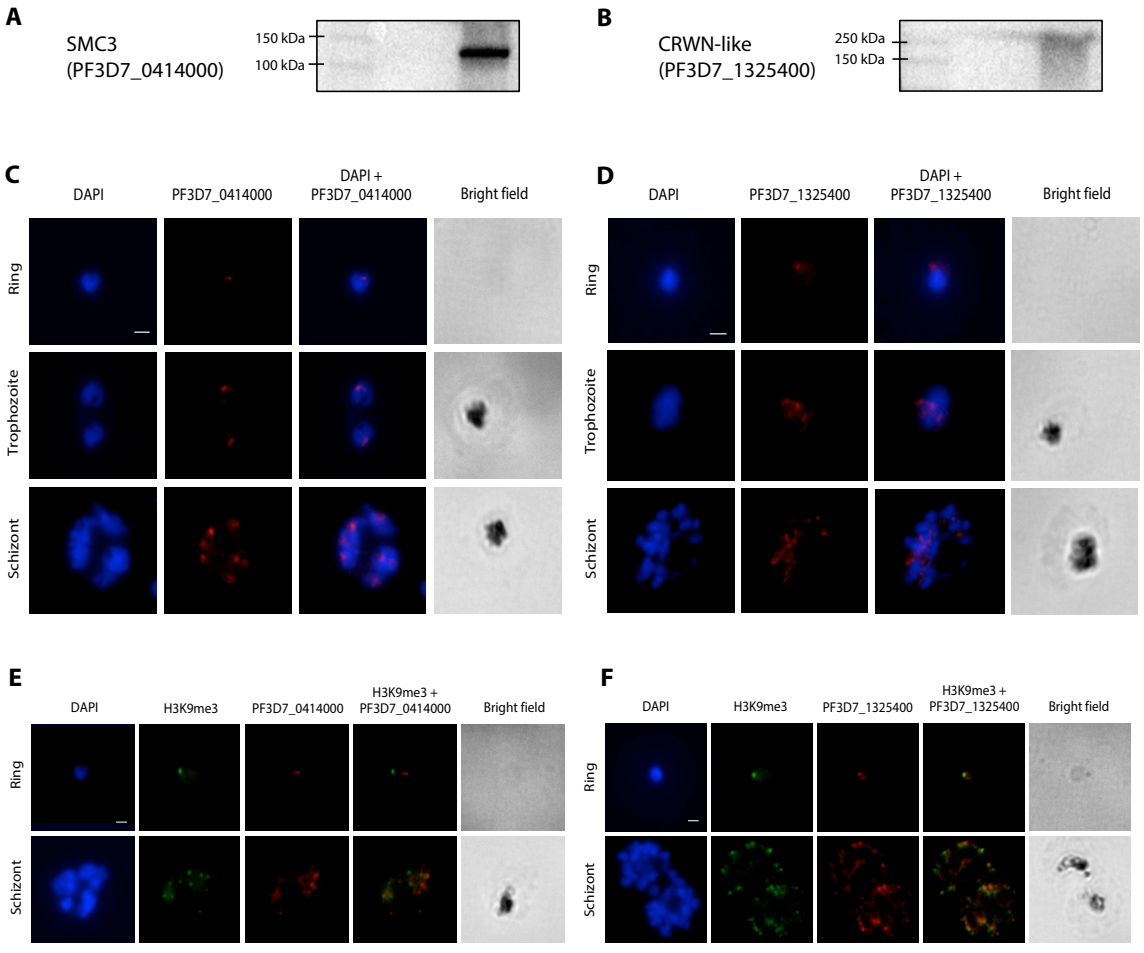


Figure 3.4: Experimental validation of candidate CAPs. Western blots show enrichment of SMC3 (A) and CRWN-like (B) proteins in the nuclear fraction (lane 1- marker; lane 2- cytoplasmic protein lysate; lane 3- nuclear protein lysate). (C) Subcellular localization of candidate chromatin-associated SMC3 protein (PF3D7_0414000) during the asexual life stages of the parasite. (D) Subcellular localization of candidate chromatin-associated CRWN-like protein (PF3D7_1325400) during the asexual life stages of the parasite. (E) Immunofluorescence analysis showing the localization of SMC3 at the periphery of the parasite nucleus away from the heterochromatin regions marked by H3K9me3. (F) Colocalization of CRWN-like proteins to the heterochromatin regions of the parasite nucleus marked by H3K9me3. Scale bar indicates 2 μm.

Protein interaction study

To investigate parasite-specific molecular components interacting with SMC3 and CRWN-like proteins, we performed co-immunoprecipitation experiments using the custom generated antibodies. Briefly, mixed-stage parasite protein lysates were incubated with anti-SMC3 or anti-CRWN-like custom antibodies and the antibody-protein complexes were collected using magnetic beads. In duplicate experiments, proteins interacting with SMC3 or CRWN-like proteins were analyzed using MudPIT. A total of 62 proteins were detected at > 2-fold higher abundance to be interacting with SMC3 in the parasite compared to the no-antibody negative control (Supplemental figure 3.2, Supplemental file 3.4). These proteins enriched for GO terms associated with DNA- and RNA-binding such as transcription elongation factor, RNA-binding protein, and

nucleosome assembly protein, none of which were detected in negative control immunoprecipitations. Additionally, we successfully recovered SMC3 and another subunit of the cohesin complex, SMC1 (PF3D7_1130700), which validates our methodology. However, using the anti-CRWN-like antibody, we were unable to immunoprecipitate the CRWN-like protein and its binding partners, indicating that the antibody-protein interaction was too weak for immunoprecipitation of a large protein (~419 kDa). Alternative tagging strategies will be needed to identify interacting partners of the CRWN-like protein in the parasite.

Genomic distribution of SMC3

In order to determine the genome-wide distribution of SMC3, we next performed ChIP-seq experiments. In duplicate experiments, trophozoite stage parasites were cross-linked with formaldehyde. Sonicated chromatin was incubated with the anti-SMC3 antibody and the resulting DNA-protein-antibody complexes were collected using Agarose beads. Purified DNA fragments were sequenced using next-generation sequencing technology. A no-antibody sample was used as a negative control. In trophozoites, SMC3 marking was restricted to the centromere region on all 14 chromosomes (Figure 3.5). Interestingly, on chromosome 7, SMC3 localized to a region slightly outside of the centromere location. Pearson correlation between biological replicates (R= 0.959) confirmed the reproducibility of our ChIP-seq experiments. Cohesin consists of four protein subunits (SMC1, SMC3, SCC1, and SCC3) and the enrichment of this complex in genomic locations exists in all eukaryotes. In mammalian cells, cohesin sites are found near

transcription start sites and co-localizing with CTCF, where they play multiple roles in chromatin organization [84, 85]. In yeast, cohesin localizes to centromeres and extends to nearby pericentromeric regions [86, 87]. Preferential loading of cohesin at centromeres is a kinetochore-dependent process [88]. The parasite SMC3 distribution during the trophozoite stage resembles the yeast cohesin occupancy. At the trophozoite stage the parasite prepares for mitosis and our results suggest that cohesin has a possible role in sister chromatid separation and cell cycle regulation at this developmental time point. However, in comparison with the yeast cohesin distribution, the parasite SMC3 occupancy does not extend to nearby pericentromeric regions, which suggests that the SMC3 subunit in particular might be important for sister chromatid cohesion.

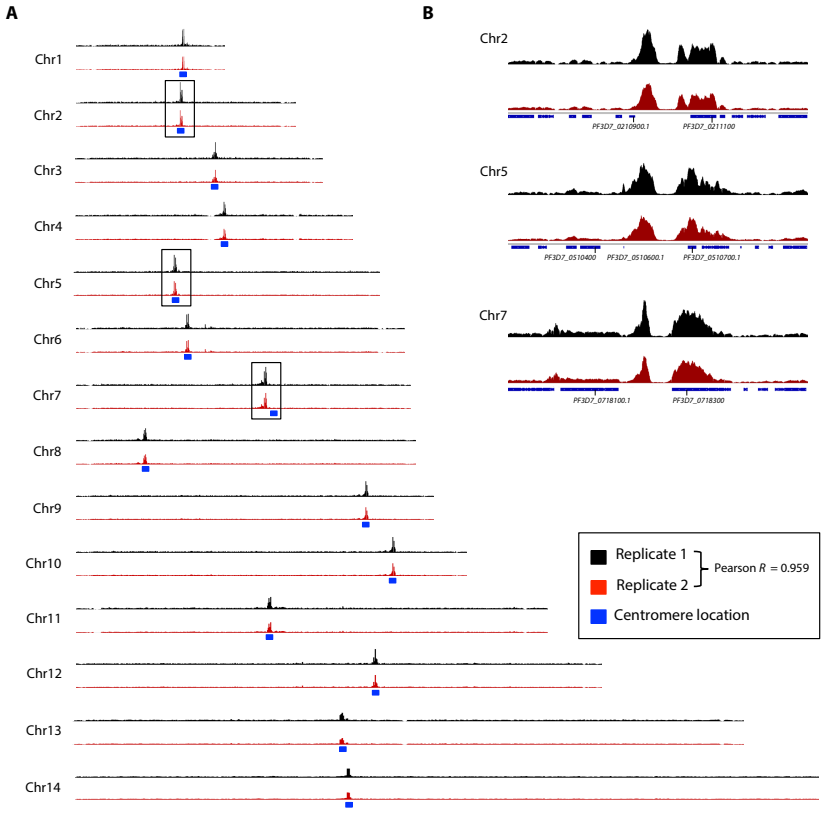


Figure 3.5: ChIP-seq analysis showing genome-wide distribution of SMC3 in trophozoites. (A) SMC3 distribution across all fourteen chromosomes. The blue box indicates the location of the centromere on each chromosome. The regions depicted in panel B are indicated with black boxes. (B) Zoomed in regions on chromosomes 2, 5 and 7 depicting SMC3 distribution.

Discussion

Increasing evidence points towards genome architecture and chromatin structure regulation playing an important role in gene expression throughout the life cycle of apicomplexan parasites [50, 89-92] (Bunnik et al., manuscript submitted). To better understand how the three-dimensional structure of the genome is being maintained, it is vital to identify proteins and protein-complexes that associate with chromatin throughout parasite development. Although snapshots of the apicomplexan proteomes have been generated [39, 93-97], no such comparative and complementary approaches have been performed to generate an accurate view of chromatin-associated domains (CADs) in apicomplexan parasites and specifically chromatin-associated proteins (CAPs) in *P. falciparum*.

In an unbiased comparison with other eukaryotic organisms, we observed that apicomplexan parasites encode a relatively large number of CAPs. Some of these candidate CAPs contain domains that are found almost exclusively in apicomplexan parasites (Figure 3.1B). While the exact function of these proteins will have to be

validated at the molecular level, this finding in all likelihood reflects the importance of chromatin-associated proteins for parasite biology.

By searching the *P. falciparum* proteome using a large collection of Pfam (HMM) and NCBI (RPS-BLAST) CADs, we have attempted to identify all *Plasmodium* CAPs. Since the *P. falciparum* genome is relatively distant from more traditional model organisms, we used less-stringent parameters for the HMM search to be able to identify CAPs. In addition, we have tried to account for false positive hits by using information from the current genome annotation to filter our initial broad search to proteins that specifically interact with chromatin (see methods). Among the final list of 1,190 candidate CAPs, only 162 proteins (13.6%) have been previously described as having chromatin-related functions in the parasite [53].

Out of the 1,190 in silico-identified *P. falciparum* proteins that contain a CAD, 469 proteins (39%) were experimentally confirmed via our Chromatin Enrichment for Proteomics (ChEP) approach. Proteins that were not identified using our ChEP methodology may only be transiently expressed or may have low expression levels and are thus difficult to detect by mass spectrometry. It is also important to note that the presence of a protein in the ChEP sample is not sufficient to conclude that it has a function in chromatin structure, since a number of proteins with no expected chromatin function can be found in our experiment. However, the preservation of *in vivo* chromatin characteristics through cross-linking is vital for studying chromatin-associated processes

by proteomics. Thus, by implementing the ChEP methodology, we have attempted to enrich for chromatin-bound factors by minimizing the loss of transiently bound factors and reducing the risk of purification artifacts that can be introduced following cell lysis. A previous study explored the nuclear proteome of *P. falciparum* during the asexual developmental cycle [96]. It is important to note that the methodology used to identify all proteins localizing to the parasite nucleus may skew the analysis towards preferentially identifying subsets of proteins, i.e. proteins that remain associated with chromatin even in the presence of detergents or proteins soluble at certain salt concentrations. In contrast, ChEP enriches for chromatin-bound factors not on the basis of solubility but rather on the basis of 'cross-linkability'. For example, many proteins with non-chromatin related functions such as transporter activity were enriched in the nuclear proteome dataset and were not identified using our ChEP methodology. However, many DNA- and RNA-binding proteins such as high mobility group protein B2, putative structural maintenance of chromosome proteins and the putative CRWN-like protein that might be an integral part of the parasite nucleus were identified in greater abundance and were significantly enriched in our ChEP sample as compared to the nuclear proteome dataset. Taken together, these results demonstrate the power of this novel methodology to identify chromatin-bound components in the parasite in an unbiased manner.

Cohesin and condensin protein complexes, composed of SMC subunits, were enriched in our ChEP samples and are an integral part of the eukaryotic nucleus. While SMC proteins are annotated in *P. falciparum*, further characterization of these proteins is lacking. Here,

we have explored the expression, localization, and genome-wide distribution of the SMC3 protein in the parasite. Using immunofluorescence, we observed a single SMC3 focus at the ring, trophozoite, and schizont stages (Figure 3.4C). This result was validated by our ChIP-seq analysis showing the distribution of SMC3 at the trophozoite stage to be confined to the centromeric regions on all chromosomes (Figure 3.5). According to previously published *P. falciparum* nuclear architecture data [37], the centromeres of all chromosomes cluster together near the periphery of the parasite nucleus and therefore proteins, such as SMC3, localizing to the centromeric regions of chromosomes would appear as a single focus in immunofluorescence experiments. Double immunofluorescence staining further confirmed that SMC3 proteins localize to the periphery of the parasite nucleus away from the heterochromatin regions of the nucleus (Figure 3.4E). Additional mechanistic insight into how this protein functions in the parasite is lacking and warrants further investigation.

Plant-related SMC domain-containing proteins, Crowded Nuclei (CRWN) proteins, are not as widely conserved in other eukaryotes. In *A. thaliana,* CRWN proteins are among the coiled-coil proteins belonging to the Nuclear Matrix Constituent Protein (NMCP) family of proteins and were originally identified as residing at the nuclear periphery in carrots [98]. Previous studies have demonstrated the importance of CRWN proteins in plant viability as evident by the inability to recover mutants with disruptions in *CRWN* genes [83]. Additionally, mutants deficient in CRWN proteins exhibit altered nuclear organization including reduced nuclear size, abnormal nuclear shape, and

heterochromatin organization. The coiled-coil domain and nuclear periphery localization suggest that these NMCP-related proteins might be functionally analogous to components of the animal nuclear lamina [58]. Despite the critical role in providing structure to the metazoan nucleus, lamina proteins have not been identified in plants or unicellular eukaryotes. While lamina-like protein (NUP-1) has been detected in kinetoplastids [99], lamina-like proteins have not been detected in *Plasmodium* species [47]. Here, we identify and localize, for the first time, a possible CRWN-like protein in *P. falciparum* that might be an integral part of the parasite nucleus. This CRWN-like protein localizes to a single focus inside the nucleus at ring and schizont stages (Figure 4D). Additionally, using double immunofluorescence staining, we observed that the CRWN-like protein co-localizes to heterochromatin and nearby regions of the parasite nucleus marked by the repressive histone mark H3K9me3 (Figure 3.4F). It is possible that this protein regulates heterochromatin regions in the nucleus, much like what has been observed in plant species [83]. A recent high-throughput transposon insertional mutagenesis study was performed to distinguish essential and dispensable genes in the *P. falciparum* genome [100]. According to this study, both genes identified in our analysis as putative homologs of the plant coiled-coil proteins (PF3D7_1325400 and PF3D7_1126700) were classified as non-essential genes. However, it is still possible that both putative CRWN protein homologs are an integral part of the parasite nucleus and one CRWN-like protein compensates for the loss of the other. A previous study exploring *P. falciparum* invasion pathways made a similar discovery where the loss of one family of invasion genes resulted in the activation of a separate invasion gene family [101]. Furthermore, in *A.*

thaliana that harbors four *CRWN* genes, quadruple *CRWN* mutants are non-viable, indicating that CRWN proteins participate in essential processes, but single, double and even some triple mutants are viable, indicating a degree of complementation between different *CRWN* genes [83]. Therefore, while it might be possible to make viable single *CRWN*-like mutants in the parasite, attempting to disrupt multiple *CRWN*-like genes will likely provide more information about the essentiality of these nuclear proteins. Further characterization of CRWN-like proteins in *Plasmodium* could improve our understanding of telomere and antigenic variation gene clustering at the nuclear periphery. More importantly, such novel plant-related proteins that play an important role in parasite nuclear organization can serve as ideal drug targets that can disrupt the parasite 3D nuclear structure with high specificity and low toxicity to the host.

This study presents the most comprehensive overview of chromatin-associated proteins in apicomplexan parasites to date. We have computationally identified chromatin-binding proteins based on the presence of chromatin-binding domains and further classified these candidate proteins into functional categories. We have also provided experimental evidence for CAPs during *P. falciparum* development using a new methodology termed Chromatin Enrichment for Proteomics (ChEP). We have further validated cellular localization and expression for two candidate chromatin-bound proteins. The function of many CAPs is still unknown and further characterization of CAPs is needed to increase our understanding of parasite biology. It is likely that our results will not only boost our

understanding of chromatin structure and chromatin-based processes, but will also help to identify key players in pathogenesis and gene regulation in parasites.

Methods

Chromatin-associated domain search

Protein sequences were obtained from the following sources: PlasmoDB version 29.0 (*P. falciparum* strain 3D7), PlasmoDB version 29.0 (*P. vivax* strain Sal I), ToxoDB version 24.0 (*T. gondii* strain ME49), TriTrypDB version 24.0 (*T. brucei* strain TREU927, *T. cruzi* strain CL Brener Esmeraldo-like, and *L. major* strain Friedlin), Saccharomyces Genome Database (*S. cerevisiae* strain S288C genome assembly R64-2-1, PomBase (*S. pombe* downloaded on 25 June 2015), Araport (*A. thaliana* 11 downloaded on 10 Jan 2017), Ensembl release 80 (*H. sapiens* genome assembly GRCh38.p2, *C. elegans* genome assembly WBcel235, and *D. melanogaster* genome assembly BDGP6).

Protein sequences were first searched for the presence of Pfam HMM profiles (Pfam version 30.0) using the function hmmscan of the HMMER software package (version 3.16, February 2015) as described in [29]. Domains were also searched independently using NCBI Reversed Position Specific BLAST (RPS-BLAST version 2.6.0) against NCBI conserved domain database (pre-calculated PSSMS originating from Cdd from various alignment collections version 3.16). If multiple RPS-BLAST hits were reported for the same conserved domain, only the one with the highest percent identity was maintained for each protein. An e-value of 0.001 was used for both approaches, and if a

protein had multiple isoforms, only the first isoform was kept. Chromatin associated protein candidates were filtered using 3,870 pre-filtered chromatin associated domains. The list of chromatin-associated domains was generated based on domain annotation found on NCBI conserved domain database as well as pfam domain database. To obtain such list, we first searched the pfam database using keywords that are known to be related to nucleus or chromatin regulation such as nucleoporin, nuclear pore complex, chromatin remodeling, histone modification, etc (see Supplemental file 3.1 for details). Next, we further selected chromatin-associated domains in the resulting list base on their annotation. A similar approach was used to identify chromatin-associated domains within the NCBI conserved domains listed in the cddid_all.tbl file. Next, both chromatin-associated domain lists were combined and carefully curated manually. Domains without clear definition of chromatin or nuclear related functions were excluded from the final list. Finally, pfam domain identifiers from hmmscan result were converted into NCBI PSSMS identifiers, and result lists from both approaches were merged. To obtain the final list chromatin-associated candidate proteins in *P. falciparum*, both manual curation and a list of exported proteins with all proteins with an Export Prediction (ExportPred) score above 5, as well as proteins with an PEXEL or HT motif for export to the red blood cell membranes (downloaded from PlasmoDB) was used to rule out potential false positive proteins, as these proteins are more likely to be exported into the red blood cell than to be transported into the nucleus.

Barplot comparison of CAPs

For each organism, both hmmscan and RPS-BLAST approaches were used and merged as previously described. Since not all genomes have been annotated at the same level, manual curation was avoided to eliminate bias and to ensure a fair comparison between organisms; therefore, we systematically calculated the number of protein containing any of the filtered chromatin-associated domains (n = 3,870) irrespective of protein annotation. For each organism, the calculated value was then corrected by the proteome size and expressed as the percentage of chromatin-associated proteins in the full proteome of that organism.

Domain Heatmap

For each chromatin-associated domain presented in any of the 11 organisms, we first calculated its abundance in all organisms. Next, the abundance value was corrected by genome size and expressed as the number per 10,000 genes. The relative abundance values were scaled to the domain frequency in the organism with the highest relative abundance of that domain. Finally, all chromatin-associated domains (n= 2,867) obtained in at least one of the organisms were clustered using k-mean clustering algorithm with a maximum of 1000 iterations (R v3.31). The number of clusters was selected based on percentage of variance captured, in which a minimum of 60% variance was required and an increase in number of cluster did not capture an additional 2% of the variance. Domains associated GO enrichment analysis was performed with dcGO

(http://supfam.org/SUPERFAMILY/dcGO/index.html) with default parameters and pfam domain IDs.

Protein classification

Candidate proteins were classified based on their general function using existing annotations and known functions of homologs in other species from various sources including PlasmoDB, UniProt, and NCBI gene database. Proteins with no annotation details were classified based on their domain functionality.

Gene expression analysis for Plasmodium CAPs

The gene expression profiles and boxplots were generated using steady-state mRNA expression profile downloaded from plasmodb.org. The expression profiles were pre-processed using standardized pipelines and are RPKM-transformed. For boxplots, gene groups were generated based on gene annotation. The list of RNA binding proteins was obtained from [29].

Parasite cultures

The *P. falciparum* strain 3D7 was cultured in human O+ erythrocytes at 5% hematocrit as previously described [102]. Cultures were synchronized at ring stage with 5% (w/v) D-sorbitol treatments [103]. Parasite cultures (8% parasitemia in 5% hematocrit) were harvested 48 hours after the first sorbitol treatment (ring stage) and 18 hours (trophozoite stage) and 36 hours thereafter (schizont stage).

Chromatin enrichment for proteomics (ChEP)

Chromatin-associated proteins were isolated at different stages of the parasite erythrocytic cycle (early ring, early trophozoite, and late schizont stages) using a protocol adapted from [55]. Briefly, synchronized parasites were cross-linked with 1% formaldehyde for 15 min at 37°C. Cross-linking was quenched by adding 0.125 M glycine for 5 min at room temperature. The parasites were then washed with phosphate-buffered saline (PBS), incubated in nuclear extraction buffer (10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 0.5 mM AEBSF, protease inhibitor cocktail (Roche) and 0.25% Igepal) for 30 min and needle sheared using a 25-gauge needle. Extracted nuclei were spun at 1,300 rcf for 20 min at 4°C. The nuclear pellet was incubated in nuclear extraction buffer containing 200 ug/mL RNase A for 15 mins at 37°C followed by two washes with 1x PBS. Nuclei were further washed with highly denaturing extraction buffers containing 4% SDS and 8M urea to wash away non-cross-linked proteins. Chromatin was solubilized and genomic DNA was sheared by sonication (Covaris; 5% duty cycle, 140 intensity peak incident power, 200 cycles per burst).

As a negative control, protein from the cytoplasmic fractions of early ring, early trophozoite and late schizont stage parasites were also extracted. For the isolation of cytoplasmic fractions, synchronized parasite cultures were collected and subsequently lysed by incubating in 0.15% saponin for 10 min on ice. Parasites were centrifuged at 1,500 rcf for 10 min at 4°C, and washed three times with PBS. For each wash, parasites

were resuspended in cold PBS and centrifuged for 10 min at 1,500 rcf at 4°C. After the last wash, parasites were resuspended in PBS, transferred to a microcentrifuge tube and centrifuged for 5 min at 2100 rcf at 4°C. Subsequently, the parasite pellet was resuspended in 1.5X volume of cytoplasmic lysis buffer (0.65% Igepal CA-360 (Sigma-Aldrich), 10 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 2 mM 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF), and EDTA-free protease inhibitor cocktail (Roche) and lysed by passing through a 26G ½ inch needle fifteen times. Parasite nuclei were centrifuged at 14,500 rcf for 15 min at 4°C and the supernatant containing the cytoplasmic extract was collected.

Custom antibody generation

Custom peptide antibodies were designed to target the C-terminal domain of 2 proteins: PF3D7_1325400 and PF3D7_0414000 (Thermo Fisher Scientific). For PF3D7_1325400, a 17-amino acid peptide (sequence: KEANKNIKLLQKYNKKM) and for PF3D7_0414000, a 18-amino acid peptide (sequence: KNEAYEIISIEEKHALEN) were used to immunize two rabbits. Antisera from day 72 post-immunization were collected and affinity-purified to obtain antibodies specifically targeting the proteins of interest.

Immunofluorescence microscopy

P. falciparum asexual stage parasites were fixed onto slides using 4% paraformaldehyde for 30 min at RT. Slides were washed three times using 1x PBS. Parasites were permeabilized with 0.2% Triton-X for 30 min at RT, followed by a wash step with 1x

PBS. Samples were blocked overnight at 4°C in IFA buffer (2% BSA, 0.05% Tween-20, 100 mM glycine, 3 mM EDTA, 150 mM NaCl and 1x PBS). Cells were incubated with custom anti-SMC3 (Thermo Fisher; 1:500) or anti-CRWN (Thermo Fisher; 1:500) antibodies for 1 hr at RT followed by anti-rabbit Alexa Fluor 488 (Life Technologies A11008; 1:500) secondary antibody for 1 hr at RT. Slides were mounted in Vectashield mounting medium with DAPI. Images were acquired using the Olympus BX40 epifluorescence microscope.

For the sequential double-staining immunofluorescence methodology, *P. falciparum* asexual stage parasites were fixed, permeabilized and blocked as described above. Parasites were incubated with custom anti-SMC3 or anti-CRWN antibodies for 1 hr at RT followed by anti-rabbit DyLight 550 (Abcam ab98489; 1:500) secondary antibody for 1 hr at RT. Slides were washed 7x using 1x PBS/0.01% Tween-20 to remove any unbound secondary antibody and incubated with Anti-H3K9me3 antibody, Alexa Fluor 488 conjugate (Millipore 07-442-AF488; 1:100) for 1 hr at RT. Slides were mounted in Vectashield mounting medium with DAPI. Images were acquired using the Olympus BX40 epifluorescence microscope.

Western blot analysis

Mixed-stage 3D7 *P. falciparum* parasite cultures were collected and lysed using 0.15% saponin for 10 min on ice. After subsequent washes, the parasite pellet was resuspended in 1.5X volume of cytoplasmic lysis buffer (0.65% Igepal CA-360 (Sigma-Aldrich), 10

mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 2 mM AEBSF, and EDTA-free protease inhibitor cocktail (Roche)) and lysed by passing through a 26G ½ inch needle fifteen times. Parasite nuclei were centrifuged at 14,500 rcf for 15 min at 4°C and the supernatant containing the cytoplasmic extract was collected. To extract proteins from the parasite nucleus, the nuclear pellet was resuspend in 1 ml of shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris pH 7.5, protease inhibitors, phosphatase inhibitors), lysed by passing through a 26 G ½ inch needle seven times, and sonicated seven times 10 seconds on/30 seconds off using a probe sonicator. Extracted nuclear protein lysates were incubated for 10 mins at room temperature with DNase I to remove DNA and centrifuged for 10 mins at 14,500 rcf to remove cell debris.

Twenty micrograms of parasite cytoplasmic and nuclear protein lysates were diluted 1:1 in 2X laemmli buffer and heated at 95°C for 10 mins. The protein lysates there then loaded on an Any-KD SDS-PAGE gel (Bio-rad) and run for 1 hour at 125 V. Proteins were transferred to a PVDF membrane for 1 hour at 18 V, stained using commercial antibodies anti-histone H3 (Abcam ab1791, 1:3000), anti-Plasmodium aldolase (Abcam ab207494, 1:1000) or custom antibodies generated against PF3D7_1325400 (Thermo Fisher, 1:100) and PF3D7_0414000 (Thermo Fisher, 1:100) followed by an incubation with secondary antibody, Goat Anti-Rabbit IgG HRP Conjugate (Bio-Rad, 1:10,000). The membranes were visualized using the Bio-Rad Chemidoc MP Gel Imager.

Protein pull-down assay

Mixed-stage 3D7 *P. falciparum* parasite cultures were collected and lysed using 0.15% saponin for 10 min on ice. After subsequent washes, the parasite pellet was resuspended in 2.5X volume of IP buffer (0.65% Igepal CA-360 (Sigma-Aldrich), 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM EDTA, 1% Triton-X, 1 mM AEBSF, 5 μM E-64 and EDTA-free protease inhibitor cocktail (Roche)) and lysed by passing through a 26G ½ inch needle ten times and sonicated 7 times 10 seconds on/30 seconds off using a probe sonicator. Extracted nuclear protein lysates were incubated for 10 mins at room temperature with DNase I to remove DNA and centrifuged for 10 mins at 14,500 rcf to remove cell debris.

Washed Protein A magnetic beads (Pure Proteome) were added to the protein sample and incubated for 1 hour at 4°C to preclear the lysate. Precleared lysate was transfered to a new microcentrifuge tube and split equally for the antibody and no antibody control. The anti-SMC3 or anti-CRWN custom antibodies were added at a 1:50 ratio and incubated overnight at 4°C. The negative control with no antibody was also incubated overnight. Antibody-protein complexes were recovered using Protein A magnetic beads (Pure Proteome), followed by extensive washes with wash buffer A (1% Triton-X, 1 mM EDTA in 1X PBS), wash buffer B (wash buffer A, 0.5 M NaCl) and wash buffer C (1 mM EDTA, 1X PBS). Proteins were eluted using 0.1 M glycine, pH 2.8 and the eluent was neutralized using 2 M Tris-HCl, pH 8.0.

Chromatin immunoprecipitation

Synchronized parasite cultures were collected at the early trophozoite stage and subsequently lysed by incubating in 0.15% saponin for 10 min on ice. Parasites were centrifuged at 1,500 rcf for 10 min at 4°C, and washed three times with PBS. For each wash, parasites were resuspended in cold PBS and centrifuged for 10 min at 1,500 rcf at 4°C. Subsequently, parasites were cross-linked for 10 min with 1% formaldehyde in PBS at 37°C. Glycine was added to a final concentration of 0.125 M to quench the cross-linking reaction, and incubated for 5 min at 37°C. Parasites were centrifuged for 5 min at 2,100 rcf at 4°C, washed twice with cold PBS and stored at -80°C.

Parasites were incubated on ice in nuclear extraction buffer (10 mM HEPES, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 0.5 mM AEBSF, EDTA-free protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Roche)). After 30 min, Igepal CA-360 (Sigma-Aldrich) was added to a final concentration of 0.25% and the parasites were lysed by passing the suspension through a 26G ½ inch needle seven times. Parasite nuclei were centrifuged at 4°C for 20 min at 2,100 rcf. Parasite nuclei were resuspended in shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris HCl pH 7.5, EDTA-free protease inhibitor cocktail (Roche), and phosphatase inhibitor cocktail (Roche)). Chromatin was fragmented using the Covaris Ultra Sonicator (S220) for 8 min with the following settings; 5% duty cycle, 140 intensity peak incident power, 200 cycles

per burst. To remove insoluble material, samples were centrifuged for 10 min at 16,800 rcf at 4°C.

Fragmented chromatin was diluted 1:1 in ChIP dilution buffer (30 mM Tris-HCl pH 8, 3 mM EDTA, 0.1% SDS, 300 mM NaCl, 1.8% Triton X-100, EDTA-free protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Roche)). Samples were precleared with Protein A Agarose beads to reduce non-specific background and incubated overnight at 4°C with 2 μg of custom anti-SMC3 antibody (Thermo Fisher Scientific). A sample with no antibody was also incubated overnight at 4°C to be used as the negative control. Antibody-protein complexes were recovered using Protein A Agarose beads, followed by extensive washes with low salt immune complex wash buffer, high salt immune complex wash buffer, LiCl immune complex wash buffer and TE buffer. Chromatin was eluted from the beads by incubating twice with freshly prepared elution buffer (1% SDS, 0.1 M NaHCO3) for 15 min at RT. Samples were reverse cross-linked overnight at 45°C by adding NaCl to a final concentration of 0.5 M. RNase A (Life Technologies) was added to the samples and incubated for 30 min at 37°C followed by a 2 h incubation at 45°C with the addition of EDTA (final concentration 8 mM), Tris-HCl pH 7 (final concentration 33 mM) and proteinase K (final concentration 66 μg/mL; New England Biolabs). DNA was extracted by phenol:chloroform:isoamylalcohol and ethanol precipitation. Extracted DNA was purified using Agencourt AMPure XP Beads (Beckman Coulter).

Libraries from the ChIP samples were prepared using the KAPA Library Preparation Kit (KAPA Biosystems). Libraries were amplified for a total of 12 PCR cycles (12 cycles of [15 s at 98°C, 30 s at 55°C, 30 s at 62°C]) using the KAPA HiFi HotStart Ready Mix (KAPA Biosystems). Libraries were sequenced with a NextSeq500 DNA sequencer (Illumina). Raw read quality was first analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the last base were removed using Trimmomatic. Any base with a quality score below 25 was trimmed using Sickle (https://github.com/najoshi/sickle). Trimmed reads were then mapped to *P. falciparum* genome (v38) using Bowtie2 (v2.3.4.1) [104]. Uniquely mapped reads were further filtered, resulting in a total of $25x10^6$ reads for SMC3 ChIP-seq replicate 1, $17.9x10^6$ reads for SMC3 ChIP-seq replicate 2 and $16.8x10^6$ reads for the no-antibody negative control libraries. Read coverage per nucleotide was determined using BEDTools. Both positive and negative libraries were then normalized by dividing through the numbers of million mapped reads. For each nucleotide, the signal from the negative control library was then subtracted from the SMC3 ChIP-seq libraries and any negative value was replaced with a zero. Genome browser tracks were generated and viewed using the Integrative Genomic Viewer (IGV) by Broad institute. Centromere locations were obtained from [105].

Multidimensional Protein Identification Technology (MudPIT)

Proteins were precipitated with 20% trichloroacetic acid (TCA) and the resulting pellet was washed once with 10% TCA and twice with cold acetone. About 50 µg of the TCA-

precipitated protein pellet was solubilized using Tris-HCl pH 8.5 and 8 M urea, followed by addition of TCEP (Tris(2-carboxyethyl)pho- sphine hydrochloride; Pierce) and CAM (chloroacetamide; Sigma) were added to a final concentration of 5 mM and 10 mM, respectively. The protein samples were digested using Endoproteinase Lys-C at 1:100 w/w (Roche) at 37 °C overnight. The samples were brought to a final concentration of 2 M urea and 2 mM CaCl2 and a second digestion was performed overnight at 37 °C using trypsin (Promega) at 1:100 w/w. The reactions were stopped using formic acid (5% final). The samples were loaded on a split-triple-phase fused-silica micro-capillary column and placed in-line with a linear ion trap mass spectrometer (LTQ) (Thermo Scientific), coupled with a Quaternary Agilent 1100 Series HPLC system. A fully automated 10-step chromatography run (for a total of 20 h) was carried out, as described in [106]. Each full MS scan (400–1600 m/z) was followed by five data-dependent MS/MS scans. The number of the micro scans was set to 1 both for MS and MS/MS. The dynamic exclusion settings used were as follows: repeat count 2; repeat duration 30 s; exclusion list size 500 and exclusion duration 120 s, while the minimum signal threshold was set to 100. The MS/MS data set was searched using ProLuCID (v. 1.3.3) [107] against a database consisting of 5,538 *P. falciparum* non-redundant proteins (PlasmoDB 9.1), 34,521 *Homo sapiens* non-redundant proteins (downloaded from NCBI 08-27-2012), 177 usual contaminants (such as human keratins, IgGs, and proteolytic enzymes), and, to estimate false discovery rates (FDRs), 36,179 randomized amino acid sequences derived from each non-redundant protein entry. To account for alkylation by CAM, 57 Da were added statically to the cysteine residues. To account for the oxidation of

methionine residues to methionine sulfoxide (which can occur as an artifact during sample processing), 16 Da were added as a differential modification to the methionine residue. Peptide/spectrum matches were sorted and selected using DTASelect/CONTRAST [108]. Proteins had to be detected by one peptide with two independent spectra, leading to average FDRs at the protein and spectral levels. To estimate relative protein levels and to account for peptides shared between proteins, normalized spectral abundance factors (dNSAFs) were calculated for each detected protein, as described in [109].

MudPIT data analysis

Two biological replicates with two technical replicates each were prepared for ChEP and cytoplasmic control samples at the ring, trophozoite, and schizont stages. Enrichment for chromatin-associated proteins in each individual experiment was defined as detection of two or more spectra of that protein in the ChEP sample and a greater than or equal to two-fold higher normalized abundance factor (dNSAF) as compared to the control cytoplasmic sample. The lists of all detected proteins with individual peptide/spectral counts are provided in Supplemental file 3.2.

References

1. WHO: **The World Malaria Report.** http://www.who.int/malaria/publications/worldmalaria-report-2017/en/**.** 2017.

2. Beugnet F, Moreau Y: **Babesiosis.** *Rev Sci Tech* 2015, **34:**627-639.

3. Flegr J, Prandota J, Sovickova M, Israili ZH: **Toxoplasmosis--a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries.** *PLoS One* 2014, **9:**e90203.

4. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M, Team WHOLC: **Leishmaniasis worldwide and global estimates of its incidence.** *PLoS One* 2012, **7:**e35671.

5. Bern C: **Chagas' Disease.** *N Engl J Med* 2015, **373:**456-466.

6. Kennedy PG: **Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness).** *Lancet Neurol* 2013, **12:**186-194.

7. Sibley CH: **Understanding drug resistance in malaria parasites: basic science for public health.** *Mol Biochem Parasitol* 2014, **195:**107-114.

8. Takala-Harrison S, Jacob CG, Arze C, Cummings MP, Silva JC, Dondorp AM, Fukuda MM, Hien TT, Mayxay M, Noedl H, et al: **Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia.** *J Infect Dis* 2015, **211:**670-679.

9. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419:**498-511.

10. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1:**E5.

11. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, Prudhomme J, Florens L, Le Roch KG: **Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum.** *Genome Biol* 2013, **14:**R128.

12. Lapp SA, Mok S, Zhu L, Wu H, Preiser PR, Bozdech Z, Galinski MR: **Plasmodium knowlesi gene expression differs in ex vivo compared to in vitro blood-stage cultures.** *Malar J* 2015, **14:**110.

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

14. Otto TD, Bohme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, Religa AA, Robertson L, Sanders M, Ogun SA, et al: **A comprehensive evaluation of rodent malaria parasite genomes and gene expression.** *BMC Biol* 2014, **12:**86.

15. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, White MW: **The transcriptome of Toxoplasma gondii.** *BMC Biol* 2005, **3:**26.

16. Sacci JB, Jr., Ribeiro JM, Huang F, Alam U, Russell JA, Blair PL, Witney A, Carucci DJ, Azad AF, Aguiar JC: **Transcriptional analysis of in vivo Plasmodium yoelii liver stage gene expression.** *Mol Biochem Parasitol* 2005, **142:**177-183.

17. Balaji S, Babu MM, Iyer LM, Aravind L: **Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains.** *Nucleic Acids Res* 2005, **33:**3994-4006.

18. Coulson RM, Hall N, Ouzounis CA: **Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum.** *Genome Res* 2004, **14:**1548-1554.

19. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, Llinas M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proc Natl Acad Sci U S A* 2008, **105:**8393-8398.

20. Iwanaga S, Kaneko I, Kato T, Yuda M: **Identification of an AP2-family protein that is critical for malaria liver stage development.** *PLoS One* 2012, **7:**e47557.

21. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, Williams AE, Drought LG, Kwiatkowski DP, Baker DA, et al: **A transcriptional switch underlies commitment to sexual development in malaria parasites.** *Nature* 2014, **507:**248-252.

22. Lesage KM, Huot L, Mouveaux T, Courjol F, Saliou JM, Gissot M: **Cooperative binding of ApiAP2 transcription factors is crucial for the expression of virulence genes in Toxoplasma gondii.** *Nucleic Acids Res* 2018, **46:**6057-6068.

23. Radke JB, Worth D, Hong D, Huang S, Sullivan WJ, Jr., Wilson EH, White MW: **Transcriptional repression by ApiAP2 factors is central to chronic toxoplasmosis.** *PLoS Pathog* 2018, **14:**e1007035.

24. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, Religa AA, Bushell E, Graham AL, Cameron R, et al: **A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium.** *Nature* 2014, **507:**253-257.

25. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I: **Transcription factor AP2-Sp and its target genes in malarial sporozoites.** *Mol Microbiol* 2010, **75:**854-863.

26. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, Waters AP, Kato T, Kaneko I: **Identification of a transcription factor in the mosquito-invasive stage of malaria parasites.** *Mol Microbiol* 2009, **71:**1402-1414.

27. Kirchner S, Power BJ, Waters AP: **Recent advances in malaria genomics and epigenomics.** *Genome Med* 2016, **8:**92.

28. Balu B, Maher SP, Pance A, Chauhan C, Naumov AV, Andrews RM, Ellis PD, Khan SM, Lin JW, Janse CJ, et al: **CCR4-associated factor 1 coordinates the expression of Plasmodium falciparum egress and invasion proteins.** *Eukaryot Cell* 2011, **10:**1257-1263.

29. Bunnik EM, Batugedara G, Saraf A, Prudhomme J, Florens L, Le Roch KG: **The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum.** *Genome Biol* 2016, **17:**147.

30. Vembar SS, Macpherson CR, Sismeiro O, Coppee JY, Scherf A: **The PfAlba1 RNA-binding protein is an important regulator of translational timing in Plasmodium falciparum blood stages.** *Genome Biol* 2015, **16:**212.

31. Eshar S, Altenhofen L, Rabner A, Ross P, Fastman Y, Mandel-Gutfreund Y, Karni R, Llinas M, Dzikowski R: **PfSR1 controls alternative splicing and steady-state RNA levels in Plasmodium falciparum through preferential recognition of specific RNA motifs.** *Mol Microbiol* 2015, **96:**1283-1297.

32. Caro F, Ahyong V, Betegon M, DeRisi JL: **Genome-wide regulatory dynamics of translation in the Plasmodium falciparum asexual blood stages.** *Elife* 2014, **3**.

33. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z: **Quantitative protein expression profiling reveals extensive post-transcriptional regulation and**

post-translational modifications in schizont-stage malaria parasites. *Genome Biol* 2008, **9:**R177.

34. Dekker J: **Gene regulation in the third dimension.** *Science* 2008, **319:**1793-1794.

35. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295:**1306-1311.

36. Fudenberg G, Getz G, Meyerson M, Mirny LA: **High order chromatin architecture shapes the landscape of chromosomal alterations in cancer.** *Nat Biotechnol* 2011, **29:**1109-1113.

37. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: **Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression.** *Genome Res* 2014, **24:**974-988.

38. Bunnik EM, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Cort A, Shi L, Andolina C, Ross LS, Brady D, et al: **Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages.** *Nat Commun* 2018, **9:**1910.

39. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, et al: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.** *Genome Res* 2004, **14:**2308-2318.

40. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465:**363-367.

41. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K: **Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation.** *Nucleic Acids Res* 2010, **38:**8164-8177.

42. Ong CT, Corces VG: **CTCF: an architectural protein bridging genome topology and function.** *Nat Rev Genet* 2014, **15:**234-246.

43. Hiraga S, Botsios S, Donze D, Donaldson AD: **TFIIIC localizes budding yeast ETC sites to the nuclear periphery.** *Mol Biol Cell* 2012, **23:**2741-2754.

44. Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, Struhl K: **Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells.** *Nat Struct Mol Biol* 2010, **17:**635-640.

45. D'Ambrosio C, Schmidt CK, Katou Y, Kelly G, Itoh T, Shirahige K, Uhlmann F: **Identification of cis-acting sites for condensin loading onto budding yeast chromosomes.** *Genes Dev* 2008, **22:**2215-2227.

46. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B: **Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.** *Nature* 2008, **453:**948-951.

47. McCulloch R, Navarro M: **The protozoan nucleus.** *Mol Biochem Parasitol* 2016, **209:**76-87.

48. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Graf S, Flicek P, Kerkhoven RM, van Lohuizen M, et al: **Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.** *Mol Cell* 2010, **38:**603-613.

49. Brancucci NM, Bertschi NL, Zhu L, Niederwieser I, Chin WH, Wampfler R, Freymond C, Rottmann M, Felger I, Bozdech Z, Voss TS: **Heterochromatin protein 1 secures survival and transmission of malaria parasites.** *Cell Host Microbe* 2014, **16:**165-176.

50. Volz JC, Bartfai R, Petter M, Langer C, Josling GA, Tsuboi T, Schwach F, Baum J, Rayner JC, Stunnenberg HG, et al: **PfSET10, a Plasmodium falciparum methyltransferase, maintains the active var gene in a poised state during parasite division.** *Cell Host Microbe* 2012, **11:**7-18.

51. Malmquist NA, Moss TA, Mecheri S, Scherf A, Fuchter MJ: **Small-molecule histone methyltransferase inhibitors display rapid antimalarial activity against all blood stage forms in Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 2012, **109:**16708-16713.

52. Malmquist NA, Sundriyal S, Caron J, Chen P, Witkowski B, Menard D, Suwanarusk R, Renia L, Nosten F, Jimenez-Diaz MB, et al: **Histone methyltransferase inhibitors are orally bioavailable, fast-acting molecules with activity against different species causing malaria in humans.** *Antimicrob Agents Chemother* 2015, **59:**950-959.

53. Bischoff E, Vaquero C: **In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the**

erythrocytic development of Plasmodium falciparum. *BMC Genomics* 2010, **11:**34.

54. Fujita T, Fujii H: **Direct identification of insulator components by insertional chromatin immunoprecipitation.** *PLoS One* 2011, **6:**e26109.

55. Kustatscher G, Hegarat N, Wills KL, Furlan C, Bukowski-Wills JC, Hochegger H, Rappsilber J: **Proteomics of a fuzzy organelle: interphase chromatin.** *EMBO J* 2014, **33:**648-664.

56. Franklin S, Chen H, Mitchell-Jordan S, Ren S, Wang Y, Vondriska TM: **Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies high mobility group protein B2 as a regulator of hypertrophic growth.** *Mol Cell Proteomics* 2012, **11:**M111 014258.

57. Ciska M, Moreno Diaz de la Espina S: **The intriguing plant nuclear lamina.** *Front Plant Sci* 2014, **5:**166.

58. Ciska M, Masuda K, Moreno Diaz de la Espina S: **Lamin-like analogues in plants: the characterization of NMCP1 in Allium cepa.** *J Exp Bot* 2013, **64:**1553-1564.

59. Ong NH, Purcell TL, Roch-Levecq AC, Wang D, Isidro MA, Bottos KM, Heichel CW, Schanzlin DJ: **Epithelial healing and visual outcomes of patients using omega-3 oral nutritional supplements before and after photorefractive keratectomy: a pilot study.** *Cornea* 2013, **32:**761-765.

60. Palenchar JB, Bellofatto V: **Gene transcription in trypanosomes.** *Mol Biochem Parasitol* 2006, **146:**135-141.

61. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376-380.

62. Jofuku KD, den Boer BG, Van Montagu M, Okamuro JK: **Control of Arabidopsis flower and seed development by the homeotic gene APETALA2.** *Plant Cell* 1994, **6:**1211-1225.

63. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M: **Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite.** *PLoS Pathog* 2010, **6:**e1001165.

64. Frankel MB, Mordue DG, Knoll LJ: **Discovery of parasite virulence genes reveals a unique regulator of chromosome condensation 1 ortholog critical for efficient nuclear trafficking.** *Proc Natl Acad Sci U S A* 2007, **104:**10181-10186.

65. Trinh V, Langelier MF, Archambault J, Coulombe B: **Structural perspective on mutations affecting the function of multisubunit RNA polymerases.** *Microbiol Mol Biol Rev* 2006, **70:**12-36.

66. Liang XH, Haritan A, Uliel S, Michaeli S: **trans and cis splicing in trypanosomatids: mechanism, factors, and regulation.** *Eukaryot Cell* 2003, **2:**830-840.

67. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: **Transcription-mediated gene fusion in the human genome.** *Genome Res* 2006, **16:**30-36.

68. Haering CH, Lowe J, Hochwagen A, Nasmyth K: **Molecular architecture of SMC proteins and the yeast cohesin complex.** *Mol Cell* 2002, **9:**773-788.

69. Strunnikov AV, Jessberger R: **Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions.** *Eur J Biochem* 1999, **263:**6-13.

70. Freeman L, Aragon-Alcaide L, Strunnikov A: **The condensin complex governs chromosome condensation and mitotic transmission of rDNA.** *J Cell Biol* 2000, **149:**811-824.

71. Formosa T, Nittis T: **Suppressors of the temperature sensitivity of DNA polymerase alpha mutations in Saccharomyces cerevisiae.** *Mol Gen Genet* 1998, **257:**461-468.

72. Mittra B, Ray DS: **Presence of a poly(A) binding protein and two proteins with cell cycle-dependent phosphorylation in Crithidia fasciculata mRNA cycling sequence binding protein II.** *Eukaryot Cell* 2004, **3:**1185-1197.

73. Stros M: **HMGB proteins: interactions with DNA and chromatin.** *Biochim Biophys Acta* 2010, **1799:**101-113.

74. Bianchi ME, Agresti A: **HMG proteins: dynamic players in gene regulation and differentiation.** *Curr Opin Genet Dev* 2005, **15:**496-506.

75. Sessa L, Bianchi ME: **The evolution of High Mobility Group Box (HMGB) chromatin proteins in multicellular animals.** *Gene* 2007, **387:**133-140.

76. Kawase T, Sato K, Ueda T, Yoshida M: **Distinct domains in HMGB1 are involved in specific intramolecular and nucleosomal interactions.** *Biochemistry* 2008, **47:**13991-13996.

77. Abhyankar MM, Hochreiter AE, Hershey J, Evans C, Zhang Y, Crasta O, Sobral BW, Mann BJ, Petri WA, Jr., Gilchrist CA: **Characterization of an Entamoeba histolytica high-mobility-group box protein induced during intestinal infection.** *Eukaryot Cell* 2008, **7:**1565-1572.

78. Gnanasekar M, Velusamy R, He YX, Ramaswamy K: **Cloning and characterization of a high mobility group box 1 (HMGB1) homologue protein from Schistosoma mansoni.** *Mol Biochem Parasitol* 2006, **145:**137-146.

79. Iborra FJ, Jackson DA, Cook PR: **The case for nuclear translation.** *J Cell Sci* 2004, **117:**5713-5720.

80. Iborra FJ, Jackson DA, Cook PR: **Coupled transcription and translation within nuclei of mammalian cells.** *Science* 2001, **293:**1139-1142.

81. Fromont-Racine M, Senger B, Saveanu C, Fasiolo F: **Ribosome assembly in eukaryotes.** *Gene* 2003, **313:**17-42.

82. Soding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33:**W244-248.

83. Wang H, Dittmer TA, Richards EJ: **Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization.** *BMC Plant Biol* 2013, **13:**200.

84. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al: **Cohesins functionally associate with CTCF on mammalian chromosome arms.** *Cell* 2008, **132:**422-433.

85. Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A: **CTCF physically links cohesin to chromatin.** *Proc Natl Acad Sci U S A* 2008, **105:**8309-8314.

86. Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, Koshland DE, DeRisi JL, Gerton JL: **Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae.** *PLoS Biol* 2004, **2:**E259.

87. Lengronne A, Katou Y, Mori S, Yokobayashi S, Kelly GP, Itoh T, Watanabe Y, Shirahige K, Uhlmann F: **Cohesin relocation from sites of chromosomal loading to places of convergent transcription.** *Nature* 2004, **430:**573-578.

88. Weber SA, Gerton JL, Polancic JE, DeRisi JL, Koshland D, Megee PC: **The kinetochore is an enhancer of pericentric cohesin binding.** *PLoS Biol* 2004, **2:**E260.

89. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, Thompson JK, Freitas-Junior LH, Scherf A, Crabb BS, Cowman AF: **Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum.** *Cell* 2005, **121:**13-24.

90. Dzikowski R, Li F, Amulic B, Eisberg A, Frank M, Patel S, Wellems TE, Deitsch KW: **Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites.** *EMBO Rep* 2007, **8:**959-965.

91. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG: **Nucleosome landscape and control of transcription in the human malaria parasite.** *Genome Res* 2010, **20:**228-238.

92. Tonkin CJ, Carret CK, Duraisingh MT, Voss TS, Ralph SA, Hommel M, Duffy MF, Silva LM, Scherf A, Ivens A, et al: **Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum.** *PLoS Biol* 2009, **7:**e84.

93. nderson DC, Lapp SA, Akinyi S, Meyer EV, Barnwell JW, Korir-Morrison C, Galinski MR: **Plasmodium vivax trophozoite-stage proteomes.** *J Proteomics* 2015, **115:**157-176.A

94. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, et al: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419:**520-526.

95. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, Mann M: **Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry.** *Nature* 2002, **419:**537-542.

96. Oehring SC, Woodcroft BJ, Moes S, Wetzel J, Dietz O, Pulfer A, Dekiwadia C, Maeser P, Flueck C, Witmer K, et al: **Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum.** *Genome Biol* 2012, **13:**R108.

97. Wang ZX, Zhou CX, Elsheikha HM, He S, Zhou DH, Zhu XQ: **Proteomic Differences between Developmental Stages of Toxoplasma gondii Revealed by iTRAQ-Based Quantitative Proteomics.** *Front Microbiol* 2017, **8:**985.

98. Mochizuki R, Tsugama D, Yamazaki M, Fujino K, Masuda K: **Identification of candidates for interacting partners of the tail domain of DcNMCP1, a major component of the Daucus carota nuclear lamina-like structure.** *Nucleus* 2017, **8:**312-322.

99. DuBois KN, Alsford S, Holden JM, Buisson J, Swiderski M, Bart JM, Ratushny AV, Wan Y, Bastin P, Barry JD, et al: **NUP-1 Is a large coiled-coil nucleoskeletal protein in trypanosomes with lamin-like functions.** *PLoS Biol* 2012, **10:**e1001287.

100. Zhang M, Wang C, Otto TD, Oberstaller J, Liao X, Adapa SR, Udenze K, Bronner IF, Casandra D, Mayho M, et al: **Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis.** *Science* 2018, **360**.

101. Stubbs J, Simpson KM, Triglia T, Plouffe D, Tonkin CJ, Duraisingh MT, Maier AG, Winzeler EA, Cowman AF: **Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes.** *Science* 2005, **309:**1384-1387.

102. Trager W, Jensen JB: **Human malaria parasites in continuous culture. 1976.** *J Parasitol* 2005, **91:**484-486.

103. Lambros C, Vanderberg JP: **Synchronization of Plasmodium falciparum erythrocytic stages in culture.** *J Parasitol* 1979, **65:**418-420.

104. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

105. Hoeijmakers WA, Flueck C, Francoijs KJ, Smits AH, Wetzel J, Volz JC, Cowman AF, Voss T, Stunnenberg HG, Bartfai R: **Plasmodium falciparum centromeres display a unique epigenetic makeup and cluster prior to and during schizogony.** *Cell Microbiol* 2012, **14:**1391-1401.

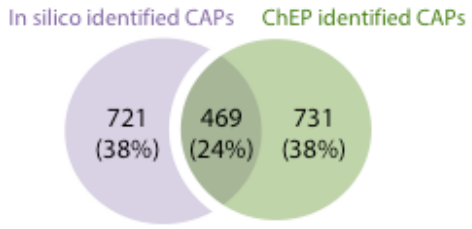
106. Florens L, Washburn MP: **Proteomic analysis by multidimensional protein identification technology.** *Methods Mol Biol* 2006, **328:**159-175.

107. Xu T, Park SK, Venable JD, Wohlschlegel JA, Diedrich JK, Cociorva D, Lu B, Liao L, Hewel J, Han X, et al: **ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity.** *J Proteomics* 2015, **129:**16-24.

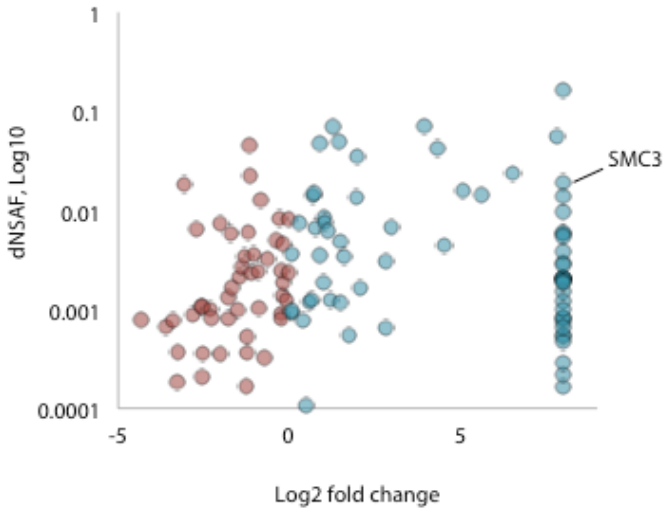
108. Tabb DL, McDonald WH, Yates JR, 3rd: **DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics.** *J Proteome Res* 2002, **1:**21-26.

109. Zhang Y, Wen Z, Washburn MP, Florens L: **Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins.** *Anal Chem* 2010, **82:**2272-2281.

Supplemental Material



Supplemental figure 3.1: Comparison of computationally identified candidate CAPs with the ChEP enriched CAPs.



Supplemental figure 3.2: Fold enrichment of putative proteins interacting with SMC3 in the parasite. Proteins enriched in the immunoprecipitated sample with the SMC3 antibody are indicated in *blue* and proteins enriched in the negative control are indicated in *red*.

Supplemental file 3.1: Computation domain prediction experiment associated tables. (XLSX)

Supplemental file 3.2: Experimentally captured chromatin-associated proteins at the ring, trophozoite and schizont stages. (XLSX)

Supplemental file 3.3: List of chromatin-associated proteins enriched by ≥ 2-fold abundance in the nuclear fraction at the ring, trophozoite and schizont stages. (XLSX)

Supplemental file 3.4: Proteins associated with Structural Maintenance of Chromosomes Protein 3 (SMC3) during the IDC. (XLSX)

CHAPTER 4: The Role of LncRNAs in Malaria Parasites: Deciphering the Non-Coding Code of Pathogenicity and Sexual Differentiation

Gayani Batugedara[1+], Xueqing M. Lu[2+], Steven Abel[1], Desiree Williams[1], Tina Wang[1], Anthony Cort[1], Jacques Prudhomme[1] and Karine G. Le Roch[1*]

[1]Department of Molecular Cell and Systems Biology, University of California Riverside, Riverside, CA 92521, USA

[+] These authors contributed equally to this work.

Preface

During its life cycle, the malaria parasite develops through distinct parasitic stages that are characterized by remarkable changes in steady state mRNA levels. This coordinated cascade of transcripts is unlikely to be tightly regulated by the surprisingly small number of identified specific transcription factors (only one third of what is expected for its genome size). This suggests that additional components and mechanisms control the expression of the predicted 6,372 genes in the parasite. In the past years, it has become apparent that RNA, itself the product of transcription, is an important regulator of transcriptional regulation. In particular, long non-coding RNAs (lncRNAs), which are numerous in eukaryotes, have been identified as vital modulators of gene expression. The regulatory roles of lncRNAs in parasite biology are only now starting to emerge. To investigate the role of lncRNAs in *P. falciparum* we explored the distribution of lncRNAs in nuclear and cytoplasmic subcellular locations. Subcellular localization and stage-specific expression of several putative lncRNAs was validated at single-cell resolution using RNA fluorescence in situ hybridization (RNA-FISH) technology. More importantly, the genome-wide occupancy of several candidate nuclear lncRNAs was explored using Chromatin Isolation by RNA Purification followed by deep sequencing (ChIRP-seq) technology, a methodology that specifically captures lncRNA:chromatin interactions. In addition to generating a highly curated catalog of nuclear and cytoplasmic lncRNAs in the parasite, this work highlights the regulatory capacities of several candidate lncRNAs and opens up new avenues for targeted approaches towards the development of new antimalarial therapies.

Abstract

Plasmodium falciparum has a complex life cycle that requires coordinated gene expression regulation to allow host cell invasion, transmission and immune evasion. However, this cascade of transcripts is unlikely to be regulated by the limited number of identified parasite-specific transcription factors. Increasing evidence now suggests a major role for epigenetic mechanisms in gene expression in the parasite. In eukaryotes, many lncRNAs have been identified and are shown to be pivotal regulators of genome structure and gene expression. To investigate the regulatory roles of lncRNAs in *P. falciparum* we explored the intergenic lncRNA distribution in nuclear and cytoplasmic subcellular locations. Together with the assistance of our recently generated nascent RNA expression profiles, we identified a total of 1,094 lncRNAs, of which 64% were identified as novel lncRNAs in *P. falciparum*. The subcellular localization and stage-specific expression of several putative lncRNAs was validated using RNA fluorescence in situ hybridization (RNA-FISH). Additionally, the genome-wide occupancy of several candidate nuclear lncRNAs was explored using Chromatin Isolation by RNA Purification (ChIRP) methodology. ChIRP-seq of candidate lncRNAs revealed that lncRNA occupancy sites within the parasite genome are focal and sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis, erythrocyte remodeling, and regulation of sexual differentiation. It is also likely that the cytoplasmic enriched lncRNAs are involved in post-transcriptional and translational regulation. Further characterization of these 1000+ lncRNAs would be needed to understand their exact function. Our findings bring a new level of insight into

the role of lncRNAs in genome structure, pathogenicity, gene regulation and sexual differentiation and opens up new avenues for targeted approaches towards therapeutic strategies in the deadly malaria parasite.

Introduction

Malaria, a mosquito-borne infectious disease, is caused by protozoan parasites of the genus *Plasmodium*. Among the human-infecting species, *Plasmodium falciparum* is the most prevalent and deadly, with an estimated 438,000 deaths per year [1]. The parasite has a complex life cycle involving multiple biological stages in both human and mosquito hosts. This multi-stage developmental cycle is tightly regulated by coordinated changes in gene expression, but the exact mechanisms regulating these events are largely unknown.

Compared to other eukaryotes with a similar genome size, *P. falciparum* has an extremely AT-rich genome and a relatively low number of sequence-specific transcription factors (TFs), approximately two-thirds of the TFs expected based on the size of the genome. Our understanding of the regulation of these TFs, and how various TFs could act together to organize transcriptional networks is still very limited. Recently generated nascent RNA expression profiles [2], as well as single cell sequencing [3] revealed that a majority of the genes in the parasite are transcribed during the trophozoite and schizont stages and that the cascade of gene expression observed using messenger RNA (mRNA) is likely the result of a combination of transcriptional [3-6] and post-

transcription regulatory events [7-10]. Additionally, chromosome conformation capture methods (Hi-C) suggest that the three-dimensional (3D) genome structure of *P. falciparum* throughout its life cycle is strongly connected with transcriptional activity of specific gene families [11]. Therefore, understanding the mechanisms regulating the parasite replication cycle inside the red blood cell (RBC) is important for identifying key factors that could be targets for new drug therapies. A number of studies have highlighted the parasite's ability to escape the host immune response by expressing variants of antigens on the RBC surface [12, 13]. To date, several multi-gene families such as *var*, *rifin* and *stevor*, which encode for *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), RIFIN and STEVOR, respectively, have been identified as key regulators of antigenic variation. However, there have been few reports regarding the role of non-protein coding transcripts in the regulation of parasite virulence [14, 15].

With advances in biotechnology and next generation sequencing technologies, huge strides have been made in genomics studies revealing that the transcriptome of an organism is much larger than what we once expected. In eukaryotes spanning from yeast to human, many non-coding RNAs (ncRNAs) have been recently recognized as key regulators of chromatin states and gene expression [16-18]. One class of ncRNAs, the long noncoding RNAs (lncRNAs), are defined as none protein coding RNA molecules which are $\geq$ 200 nucleotides in length. Many lncRNAs share features with mature messenger RNAs (mRNAs) including 5' caps, polyadenylated tails, and introns. In addition, lncRNAs are often expressed and functionally associated in a cell-type specific

manner. LncRNAs enriched in the nuclear fraction often associate with regulation of epigenetics and transcription [19-22], while lncRNAs enriched in the cytoplasm are associated with mRNA processing, post-transcriptional regulation, translational regulation, and cellular signaling process [23-25]. A well-studied example of a lncRNA that regulates gene expression is Xist (X inactive specific transcript), which mediates X-chromosome inactivation during zygotic development [26]. Deposition of Xist on the X chromosome recruits histone-modifying enzymes that place repressive histone marks, such as H3K9 and H3K27 methylation, leading to gene silencing and the formation of heterochromatin. As another example, long telomeric repeat-containing lncRNAs (TERRA) have been recently identified as a major component of telomeric heterochromatin [27, 28].

To date, several studies have explored ncRNAs in *P. falciparum* using different techniques [29-31]. Specifically, ncRNAs have been linked to regulation of virulence genes [14, 32, 33]. More recently, a novel family of twenty-two lncRNAs transcribed from the telomere-associated repetitive elements (TAREs) was identified in the parasite [30, 33, 34]. The TARE-lncRNAs show functional similarities to the eukaryotic family of non-coding RNAs involved in telomere and heterochromatin maintenance [35]. However, chromatin occupancy sites of most lncRNAs are not known, therefore, the regulatory roles of a large portion of the non-coding transcriptome of *P. falciparum* still remain a mystery.

To investigate the regulatory roles of lncRNAs in *P. falciparum* we explored the intergenic lncRNA distribution separately in nuclear and cytoplasmic subcellular locations. Together with the assistance of our recently generated nascent RNA expression profiles [2], we identified a total of 1,094 lncRNAs, of which 64% were identified as novel lncRNAs in *P. falciparum*. We further validated the subcellular localization and stage-specific expression of several putative lncRNAs using RNA fluorescence in situ hybridization (RNA-FISH). Additionally, the genome-wide occupancy of several candidate nuclear lncRNAs was explored using Chromatin Isolation by RNA Purification (ChIRP) methodology, a method that specifically captures lncRNA:chromatin interactions with high sensitivity and low background. ChIRP-seq of candidate lncRNAs revealed that lncRNA occupancy sites within the parasite genome are sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis, remodeling of RBC, and regulation of sexual differentiation. Collectively, our results provide crucial information regarding the role of lncRNAs in gene expression in the malaria parasite.

Results

Identification of lncRNAs

In order to comprehensively identify lncRNA populations in *P. falciparum* we extracted total RNA from both nuclear and cytoplasmic fractions using synchronized parasite cultures at early ring, early trophozoite, late schizont, and gametocyte stages. The samples collected here allows for gene expression profiling during the critical processes

of parasite egress, invasion and sexual differentiation. In brief, extracted parasites were subjected to a modified cell fractionation procedure described in PARIS kit (ThermoFisher) (see methods). Successful isolation of both subcellular fractions was validated using western blot with an anti-histone H3 antibody as a nuclear marker, and an anti-aldolase antibody as a cytoplasmic marker (Figure 4.1B). After separation of nuclear material from the cytoplasmic material, total RNA and subsequence polyadenylated mRNA was isolated from both fractions. Strand-specific libraries were then prepared and sequenced (see methods for details). For verification, spearman correlations in gene expression levels were calculated among nuclear samples, cytoplasmic samples, and a previous published steady-state total mRNA dataset generated in our lab [36] (Supplemental figure 4.2). Once validated, a computational pipeline was implemented for the identification of lncRNAs. Briefly, all nuclear and cytoplasmic RNA libraries were merged into one single file, then assembled into nuclear and cytosol transcriptome independently using cufflinks. Subsequently, transcripts were filtered based on length, expression level, presence of primary transcript from our previously published GRO-seq dataset [2], and sequence coding potential (Figure 4.1A). To specifically identify lncRNA candidates within the intergenic regions, we removed any predicted transcripts that overlap with annotated genes. Our goal was to select transcripts that are ≥ 200bp in length, consistently expressed in both published nascent RNA and steady-state RNA expression profiles, and that are likely to be non protein-encoding genes. As a result, we identified a total of 1,094 intergenic lncRNAs in *P. falciparum* irrespective of the developmental stage. Three hundred ninety-five lncRNAs (36%) overlapped with

previously identified intergenic lncRNAs in [29, 37], and 699 lncRNAs were identified as novel in *P. falciparum* (Figure 4.1C).

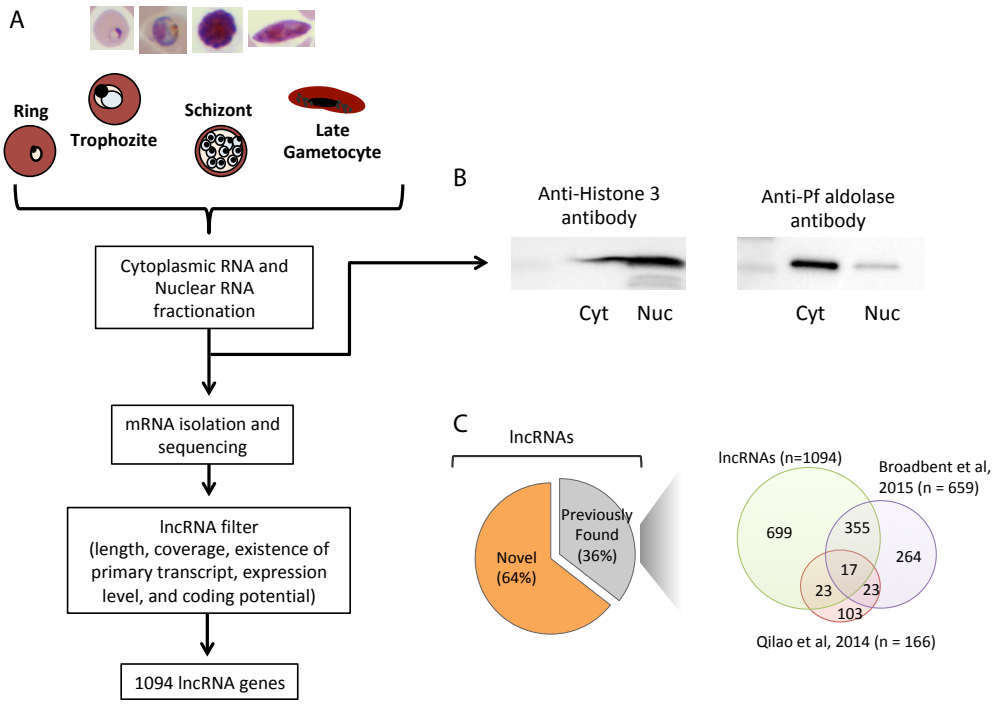


Figure 4.1: Nuclear and cytoplasmic lncRNA identification. (A) A general overview of the lncRNA identification pipeline. (B) Validation of cell fractionation efficiency using anti-histone H3 and anti-aldolase as nuclear and cytoplasmic markers. (C) Comparison of lncRNA candidates with lncRNAs identified from previous publications.

Length, GC content, and RNA stability of cytoplasmic and nuclear lncRNAs

Next, we categorized our candidate lncRNAs into nuclear lncRNAs, cytoplasmic lncRNAs, or indistinguishable lncRNAs that are equally distributed in both fractions.

Among the total identified 1,094 lncRNAs, 574 lncRNAs (52%) were enriched in the nuclear fraction, 290 lncRNAs (27%) were enriched in the cytoplasmic fraction, and 230 lncRNAs (21%) showed similar distribution between both subcellular fractions (Figure 4.2A). Further, we explored the physical properties of lncRNAs. We observed that lncRNAs are in general shorter in length and less GC rich as compare to protein-encoding mRNAs (Figure 4.2B and C). Using total steady-state mRNA expression profiles and nascent RNA expression profiles, we then estimated the expression levels and stability of the lncRNAs. RNA stability was calculated as the ratio between steady-state mRNA expression levels over nascent RNA expression levels. We discovered that, although the overall cell cycle gene expression pattern of the lncRNAs is similar to the expression pattern of coding mRNAs, lncRNAs are less abundant and less stable than coding mRNAs; nuclear IncRNAs are particularly low expressed and unstable as compared to the other two groups of IncRNAs (Figure 4.2D). These observations are consistent with previous lncRNA annotation studies in human breast cancer cells [38] and noncoding RNA stability studies in mammalian genomes [39]. Our results suggest that the low expression level and the low stability of these lncRNAs may be the reason why they failed to be detected in previous identification attempts. By taking advantage of primary transcripts detected in our GRO-seq dataset, we significantly improved the sensitivity of lncRNA detection, especially for those localized in the nuclear fraction and expressed at a lower level.

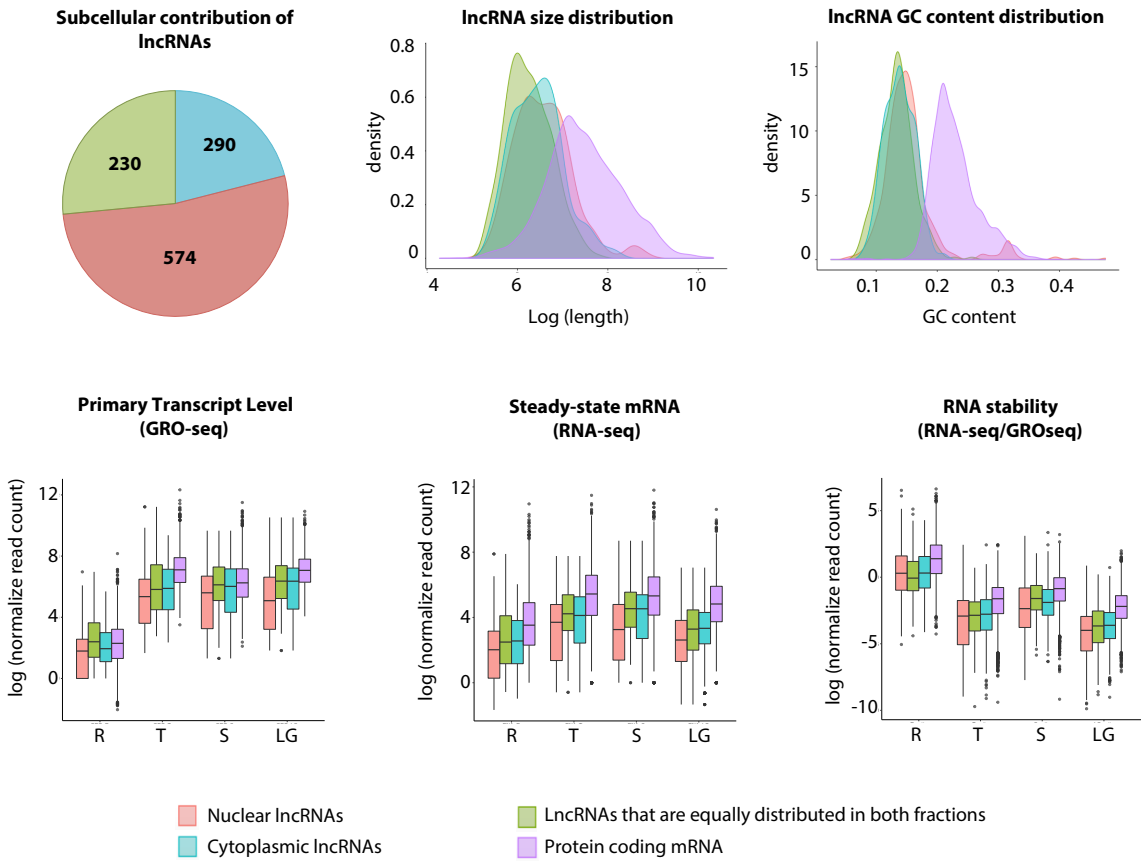


Figure 4.2: Candidate lncRNA categorization. (A) A total of 1,094 lncRNA candidates was identified, covering 574 nuclear enriched lncRNAs, 290 cytoplasmic enriched lncRNAs, and 230 lncRNAs found in both fractions. Density plot of the size (B) and GC content (C) of lncRNA candidates and annotated protein encoding mRNAs. Expression levels of primary transcripts (left), steady-state mRNA (middle), and relative stability (right) of lncRNA candidates and annotated protein encoding mRNAs.

Stage Specific Expression and Epigenetic landscape of Cytosolic and Nuclear lncRNAs

As lncRNAs often exhibit specific expression patterns in other eukaryotes, we investigated the stage specificity of identified candidate lncRNAs across the cell cycle. Using k-mean clustering, we were able to group lncRNAs into 7 distinct clusters (Figure 4.3A). Generally, nearly all lncRNAs showed a strong coordinated cascade throughout the parasite's cell cycle. A large fraction of the lncRNAs was highly expressed at ring and schizont stages as compared to the trophozoite and gametocyte stages (Figure 4.3B). Cluster 1 contains lncRNAs that are more abundantly expressed in the nuclear fraction of ring stage parasites and are also moderately expressed in the nuclear fraction of schizont stage parasites. LncRNAs representative of this cluster are the lncRNA-TAREs. We observed that all lncRNA-TAREs identified in this study are clustered into this group with an average expression of 2.01 log two-fold change of nuclear to cytoplasmic ratio (Figure 4.3C and Figure 4.3D). This finding validates our approach and suggests that lncRNAs in this cluster may contribute to the maintenance and regulation of chromatin structure and *var* gene regulation. Approximately 40% of the identified lncRNAs are more abundantly found in either the nuclear or cytoplasmic fraction at the schizont stage (cluster 5 and 6), after DNA replication and the peak of transcriptional activity observed at the trophozoite stage. We observed a few lncRNAs that are solely expressed during the asexual cycle with distinct changes in heterochromatin marks (Figure 4.3D). The presence of some of these lncRNAs was confirmed using reverse transcriptase PCR (RT-PCR) as demonstrated in Supplemental Figure 4.1. Based on clustering analysis, we also found that 19% of the lncRNAs are more exclusively expressed at a high level at the

gametocyte stage (cluster 7). Interestingly, two unique lncRNAs in this cluster were identified in a previous study to be located within heterochromatin regions marked by repressive histone marks H3K9me3 at the trophozoite stage [11] (Figure 4.3E and Figure 4.3F). At the gametocyte stage however, the H3K9me3 was lost. Additionally, both lncRNAs are transcribed from regions adjacent to gametocyte-specific genes. These results suggest that transcripts upregulated in gametocytes could possibly be essential for gametocyte development.

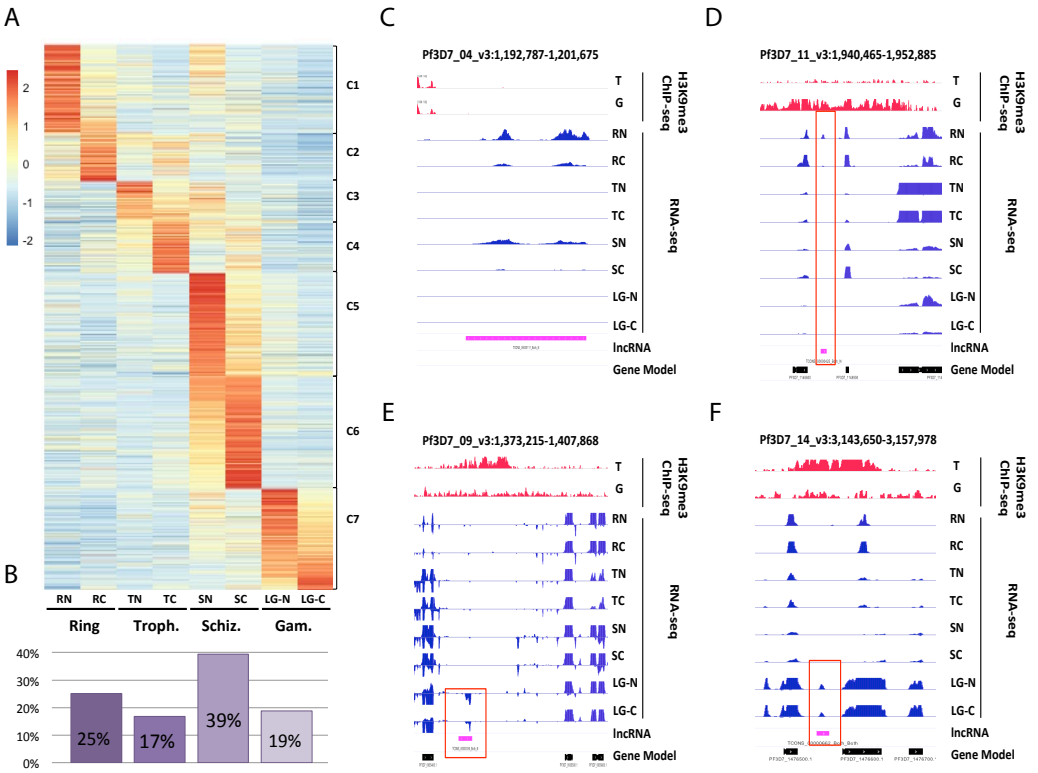


Figure 4.3: Gene expression patterns of lncRNAs. (A) lncRNAs are grouped into 7 clusters based on their cell cycle expression patterns. (B) Percentage of lncRNAs that are highly expressed in each subcellular fraction at ring, trophozoite, schizont, and late gametocyte stages. Genome browser views of H3K9me3 ChIP-seq and RNA-seq datasets for one identified lncRNA-TARE located on the right arm of chromosome 4 (C), one asexual-specific lncRNA (D), and two gametocyte-specific lncRNAs located in intergenic regions of chromosome 9 (E) and 14 (F).

Validation of lncRNA localization and stage-specific expression

To validate the localization of candidate lncRNAs, we utilized RNA fluorescence in situ hybridization (RNA-FISH). Two candidate lncRNAs enriched in the cytoplasmic fraction and 5 lncRNAs enriched in the nuclear fraction were used for RNA-FISH (Figure 4.4). Briefly, mixed stage parasites were fixed and hybridized to fluorescently labeled ~200-300 nucleotide antisense RNA probes (see methods for details). The hybridization images clearly demonstrate that the cytoplasmic lncRNAs are localized outside the DAPI-stained genomic DNA, while the nuclear lncRNAs localize to distinct foci within the DAPI-stained nuclei (Figure 4.4). Additionally, using RNA-FISH, we were also able to validate the stage-specific expression of candidate lncRNAs. Specifically, lncRNA-267 and lncRNA-13 were expressed at the ring and trophozoite stages; lncRNA-178 was expressed at the trophozoite and schizont stages; lncRNA-643 was expressed at the schizont stage only and lncRNA-TARE4 was expressed at all three asexual stages (Figure 4.4). LncRNA-ch9 and lncRNA-ch14 were only expressed at the gametocyte stage. These

results highlight that, similar to protein coding transcripts, these candidate lncRNAs are also developmentally regulated.

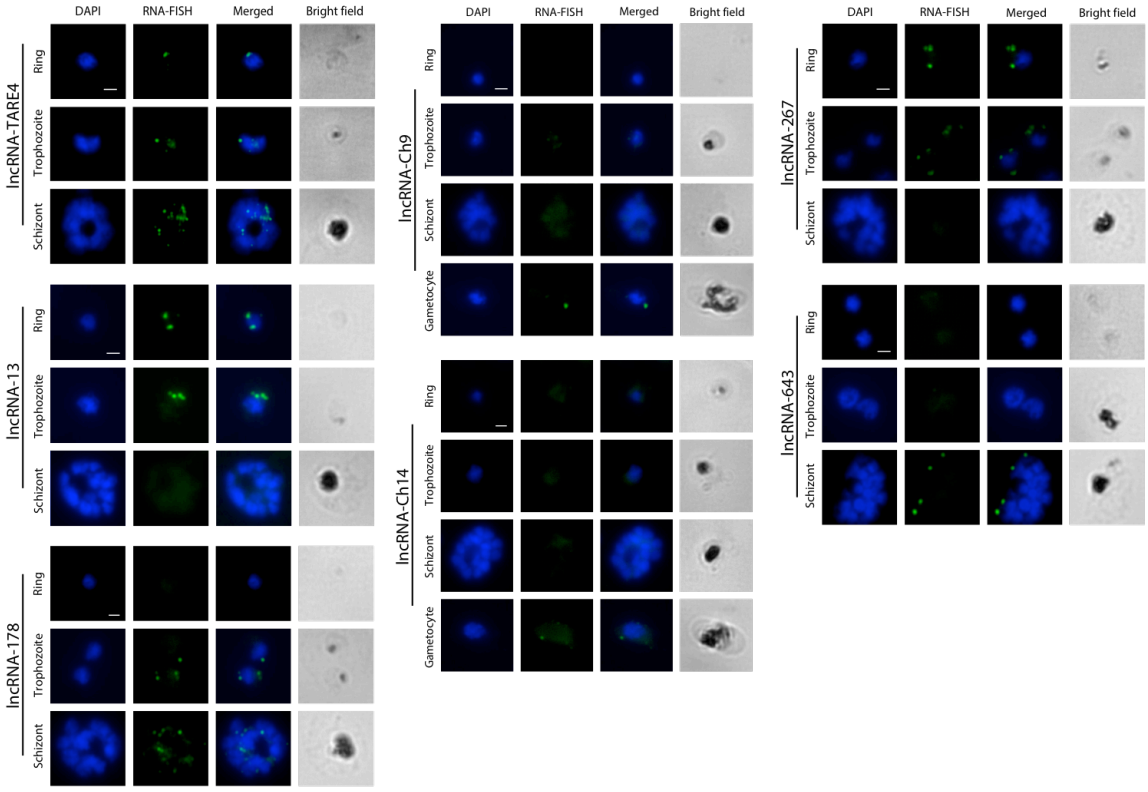


Figure 4.4: RNA-FISH experiments show localization of several candidate lncRNAs. Nuclei are stained with DAPI. Nuclear lncRNAs (lncRNA-TARE4, lncRNA-13 and lncRNA-178, lncRNA-ch9 and lncRNA-ch14) colocalize with DAPI. Cytoplasmic lncRNAs (lncRNA-267 and lncRNA-643) do not colocalize with the DAPI stained nuclei. Scale bar indicates 2µm.

Genomics maps of RNA-chromatin interactions

To explore the roles of lncRNAs in chromatin regulation, we must first identify occupancy sites of lncRNAs within the genome. For an unbiased high-throughput discovery of RNA-bound DNA in *P. falciparum*, we adapted a method termed Chromatin Isolation by RNA Purification (ChIRP) (Figure 4.5A) [40, 41]. Briefly, synchronized parasites were extracted and crosslinked. Parasite nuclei were then extracted and chromatin was solubilized and sonicated. Biotinylated anti-sense oligonucleotides that tile the RNA of interest were hybridized to target RNAs and isolated using magnetic beads. Purified DNA fragments were sequenced using next-generation sequencing technology. An input control was used to normalize the signal from ChIRP enrichment. To validate the ChIRP protocol, we analyzed the RNA fraction to validate the specificity of the biotinylated oligonucleotides to target the RNA of interest. RT-PCR results confirm that the TARE4 probes retrieve the lncRNA-TARE4 and the control tRNA ligase probes retrieve the tRNA ligase RNA respectively (Figure 4.5B). Neither RNA was retrieved in the negative control that was incubated with no probes. These results confirm that the biotinylated probes target the RNA of interest with specificity.

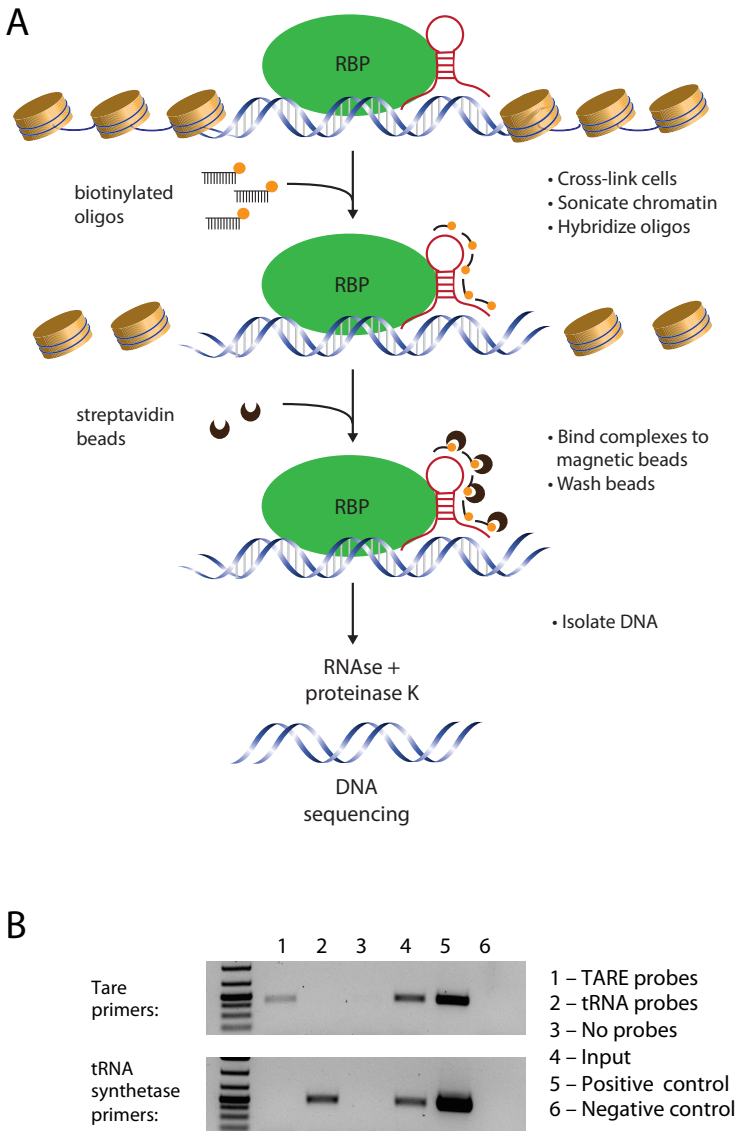


Figure 4.5: Chromatin Isolation by RNA Purification (ChIRP). (A) Schematic representation of the ChIRP methodology. (B) RT-PCR following the ChIRP protocol validates the specificity of the biotinylated antisense probes.

LncRNAs transcribed from the telomere regions of chromosomes have been implicated in *var* gene regulation [14, 32, 33]. Here, for the first time, we identify genomic occupancy sites of lncRNA-TARE4, a lncRNA transcribed from the telomere region on chromosome 4. LncRNA-TARE4 binding sites occur on multiple chromosomes and are enriched in subtelomeric regions as well as genomic regions around several antigenic variation genes belonging to the *var* and *rifin* gene families (Figure 4.6 A-B). Using ChIRP analysis, we identified over 200 lncRNA-TARE4 binding sites in the parasite genome, which represents a large resource to study potential functions of lncRNA-TAREs in regulating antigenic variation and heterochromatin maintenance. Additionally, given the already existing evidence for the role of lncRNAs in *var* gene regulation, our lncRNA-TARE4 ChIRP results validate our methodology and highlight the effectiveness of ChIRP for exploring other lncRNA:chromatin relationships within the parasite.

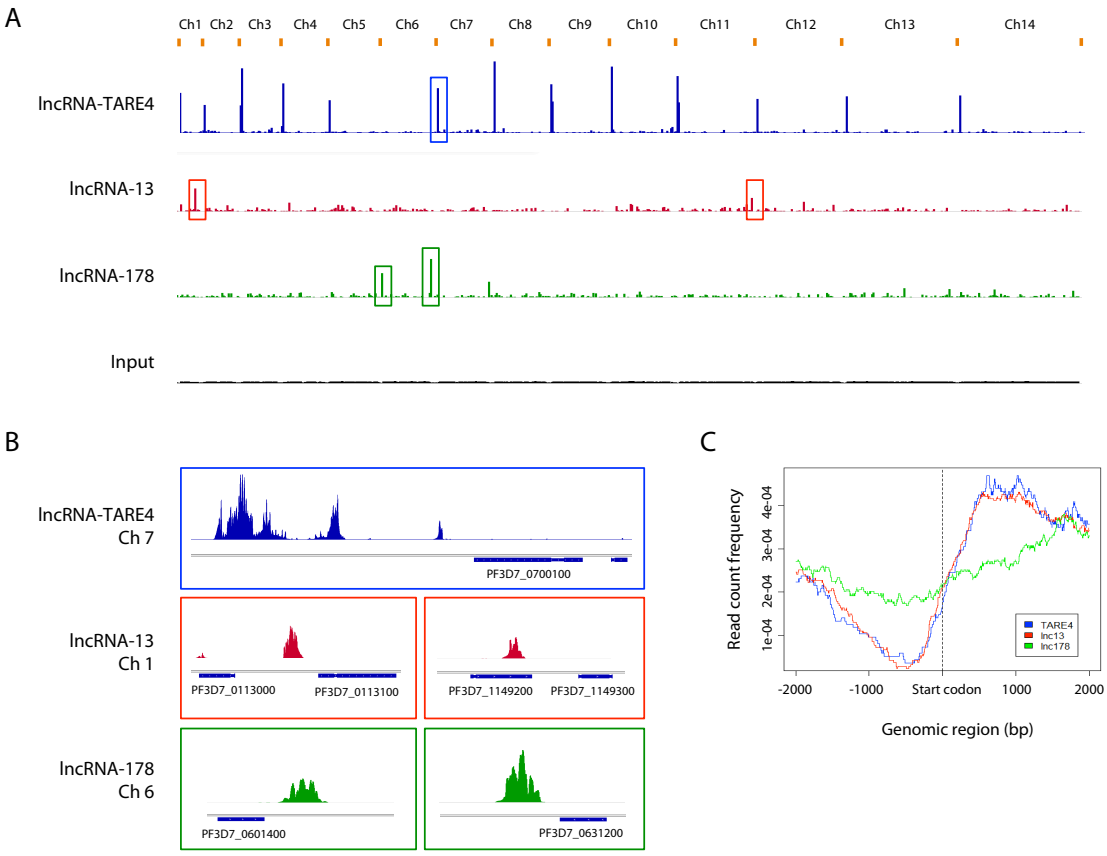


Figure 4.6: ChIRP-seq reveals candidate lncRNA binding sites. (A) Genome-wide binding sites of lncRNA TARE-4, lncRNA-13 and lncRNA-178. The regions depicted in panel B are indicated in colored boxes (B) Zoomed in regions of the most significant peaks that correspond to lncRNA binding sites. (C) LncRNA occupancy patterns around the start codon.

We next explored the genomic binding sites of two lncRNAs, lncRNA-13 and lncRNA-178, identified as enriched in the nuclear fraction and validated using RNA-FISH. LncRNA-13, transcribed from a region on chromosome 1 (PF3D7_01: 493,112-493,906)

is highly expressed at the trophozoite stage (Supplemental file 4.1). ChIRP analysis reveals that lncRNA-13 binding sites at the trophozoite stage are enriched around surface antigen genes, namely PF3D7_0113100 (SURFIN4.1) and PF3D7_1149200 (RESA, ring-infected erythrocyte surface antigen) (Figure 4.6B). Both SURFIN and RESA family of proteins have been implicated in erythrocyte invasion-related processes and are transcribed in mature-stage parasites [42, 43]. Given that a trophozoite stage lncRNA was identified adjacent to surface antigen genes which are transcribed at the schizont stage, it is tempting to speculate that the lncRNA-13 might be repressing the transcription of these surface antigen genes until such time when they need to be expressed. LncRNA-178 is transcribed from a region on chromosome 6 (PF3D7_06: 53,758-54,745) and is highly expressed at the schizont stage (Supplemental file 4.1). ChIRP-seq results reveal that lncRNA-178 binding sites are most enriched around erythrocyte membrane protein 1 (PfEMP1) genes (Figure 4.6B). According to our recently generated nascent RNA data, these genes are transcribed at the trophozoite stage. Much like in the case of lncRNA-13, there seems to be an inverse correlation of expression between lncRNA-178 and its target genes. Therefore, it is possible that the lncRNA-178 binds to and represses the transcription of said genes until they need to be expressed. However, additional ChIRP experiments, conducted at other time points, are needed to confirm the regulatory capacity of both lncRNA-13 and lncRNA-178.

By aligning ChIRP signals for lncRNA-TARE4, lncRNA-13 and lncRNA-178 across all bound regions, we discovered that the occupancy of lncRNAs was enriched upstream of

the start codon and within the gene body (Figure 4.6C). This pattern provides additional support for a notion that the lncRNAs explored here might have a role in recruiting proteins and protein complexes to bind to regulatory regions of genes, thereby regulating their transcription. Collectively, these experiments suggest that lncRNAs in the parasite could be surprisingly like sequence-specific transcription factors in regulating chromatin states and gene expression.

Discussion

Many lncRNAs are now recognized as potential regulators of chromatin structure and gene expression in eukaryotes. The extent of lncRNA regulation in *Plasmodium* is only now starting to emerge. The work described here presents the first global detection of lncRNAs from different subcellular locations throughout the cell cycle of *P. falciparum*. Using both experimental and computational pipelines, we identified 1,094 lncRNAs covering 290 cytoplasmic enriched, 574 nuclear enriched, and 230 indistinguishable lncRNAs that were localized in both fractions. In the last decade, many lncRNAs have been discovered with diverse cellular functions outside of the nucleus. These types of lncRNAs have been reported to interact with ribosomes [24], and are often associated with post-transcriptional and translational processes [23]. Some cytoplasmic lncRNAs, such as half-STAU1-binding site RNAs (1/2-sbsRNAs) [44, 45] and growth arrested DNA-damage inducible gene 7 (gadd7) [46], are shown to alter the stability of mRNA, while others, including lncRNA-p21 [47] and AS UCHl1 [48] are shown to either promote or repress translational processes. The dataset generated in this study provides

the first comprehensive global view of cytoplasmic lncRNAs expressed across the parasite's cell cycle. Our data suggest that cytoplasmic lncRNAs are also coordinately expressed but are less abundant as compared to the number of nuclear lncRNAs in the parasite. In addition, we observed that a small group of cytoplasmic lncRNAs is highly expressed at the trophozoite stage, the stage where a large proportion of genes are transcribed [2]. Though more in-depth studies will be required to confirm the functions of these trophozoite-expressed cytoplasmic lncRNAs, it is possible that some of these lncRNAs are involved in mRNA stability, alternative splicing, or translational regulation.

By utilizing recently published nascent RNA expression profiles (GRO-seq [2]), we were able to significantly improve the sensitivity of lncRNA detection, especially for the identification of nuclear lncRNAs. This study identified 699 novel lncRNAs in the parasite. More than 300 of these newly identified lncRNAs were enriched in the nuclear fraction. In other eukaryotes, functions of nuclear lncRNAs have been determined as either directly promoting or repressing gene expression activity [49, 50], guiding or enhancing the functions of regulatory proteins [20, 50-53], or assisting the alteration of chromatin structures by shaping 3D genome organization [21, 54, 55]. Some of the well-characterized nuclear lncRNAs, such as Xist [56], Firre [57], and Neat [58], were shown to be particularly important for nuclear organization and chromatin conformation changes. In *P. falciparum*, emerging evidence has shown that chromatin structure and genome organization are of vital importance for the parasite's gene expression and regulation system. Therefore, identification and characterization of nuclear enriched

lncRNAs may help us to uncover chromatin-associated regulators in the parasite.

In our present work, we observed that a large number of lncRNAs, including the lncRNA-TAREs, are highly abundant at the ring and schizont stages. This finding suggests that some of these lncRNAs (cluster 1, Figure 4.2A) are likely to be involved in heterochromatin maintenance or chromatin structure re-organization events, as Hi-C experiments show that chromatin structure is compacted at the ring and schizont stages [59]. Additionally, ChIRP experiments mapping genome-wide binding sites of lncRNAs revealed that lncRNA-TARE4 binds to subtelomeric regions on multiple chromosomes as well as regulatory regions around genes involved in pathogenesis and immune evasion. Previous reports showed that subtelomeric regions as well as virulence gene families cluster in perinuclear heterochromatin. Therefore, it is possible that lncRNA-TARE4 interacts with or recruits histone-modifying complexes to targeted regions in order to maintain them in a heterochromatin state, much like the case of X chromosome inactivation regulated via lncRNA Xist [56].

Genomic occupancy of two other lncRNAs explored here, lncRNA-13 and lncRNA-178, suggest that these lncRNAs bind to regions around different surface antigens (Figure 4.6B). In both cases, an inverse relationship was observed between the lncRNA expression and the expression of genes around the lncRNA occupancy sites. Given already existing evidence for lncRNA-associated epigenetic modification and transcriptional regulation in other eukaryotic systems [60-62], it is possible that lncRNA-

13 and -178 in the parasite are responsible for coordinated recruitment of distinct repressing histone-modifying complexes to target loci. Additionally, we discovered that the occupancy of these lncRNAs was enriched upstream of the start codon (Figure 4.6C). This pattern of lncRNA occupancy provides additional support for the idea that the lncRNAs explored here might have a role in recruiting protein complexes to promoter regions of target genes in order to repress transcription, either by preventing the formation of the pre-initiation complex or recruiting histone modifiers. However, while additional experiments are needed to confirm the roles of these nuclear lncRNAs in the parasite, using ChIRP-seq, we demonstrate that genome-wide collections of RNA binding sites can be used to discover the DNA sequence motifs enriched by lncRNAs. These findings signify the existence of lncRNA target sites in the genome, an entirely new class of regulatory elements that could be essential for transcriptional regulation in the malaria parasite.

Compared to the progress made in understanding lncRNA biology in higher eukaryotes, the field of lncRNA in *Plasmodium* is still young, yet full of potential. Analysis of promoter and gene body regions with available histone modification datasets (H3K9me3, H3K36me3, H3K9ac) are still needed for further annotation of these candidate lncRNAs. In addition, understanding how lncRNAs may contribute to cell cycle progression and sexual differentiation in the parasite is still a work in progress. However, from recent studies conducted in model eukaryotes it is clear that lncRNAs represent a new paradigm in chromatin remodeling and genome regulation. Therefore, this newly generated dataset

will not only assist future lncRNA studies in the malaria parasite, but will also help in identifying parasite-specific gene expression regulators that can ultimately be used as new anti-malarial drug targets.

Materials and Methods

Parasite culture

P. falciparum 3D7 strain at ~ 8% parasitemia was cultured in human erythrocytes at 5% hematocrit in 25 mL of culture as previously described in [63]. Two synchronization steps were performed with 5% D-sorbitol treatments at ring stage with eight hours apart. Parasites were collected at early ring, early trophozoite, and late schizont stages. Parasite developmental stages were assessed using Giemsa-stained blood smears (Supplemental Figure 4.1). Gametocyte parasites were induced from the *P. falciparum* NF54 strain and were harvested 15 days (stage IV to V) after the induction procedure as previously described in [64].

Nuclear and cytosolic RNA isolation

Highly synchronized parasites were first extracted using 0.15% saponin solution followed by centrifugation at 1500 x g for 10 mins at 4°C. Parasite pellets were then washed twice with ice cold PBS and re-collected at 1500 x g. Parasite pellets were resuspended in 500 uL ice cold Cell Fractionation Buffer (PARIS kit, ThermoFisher; AM1921) with 10 uL of RNAase Inhibitor (SUPERaseIn 20U/uL, Invitrogen; AM2694) and incubated on ice for 10 minutes. Samples were centrifuged at 500 x g for 5 mins at 4°C. After centrifugation,

the supernatant containing the cytoplasmic fraction was collected. Nuclei were resuspended in 500 uL Cell fractionation buffer and 15 uL RNAse Inhibitor as described above. To obtain a more purified nuclear fraction, the pellet was syringed with a 26G inch needle five times. The sample was incubate on ice for 10 mins and centrifuged at 500 x g for for 5 mins at 4°C. The nuclear pellet was resuspended in 500 uL of ice cold Cell Disruption Buffer (PARIS kit, ThermoFisher; AM1921). For both cytoplasmic and nuclear fractions, RNA was isolated by adding 5 volumes of Trizol LS Reagent (Life Technologies, Carlsbad, CA, USA) followed by a 5 min incubation at 37°C. RNA was then isolated according to manufacturer's instructions. DNA-free DNA removal kit (ThermoFisher; AM1906) was used to remove potential genomic DNA contamination according to manufacturer's instruction, and the absence of genomic DNA was confirmed by performing a 40-cycle PCR on the PfALBA3 gene using 200 to 500 ng input RNA.

mRNA isolation and library preparation

Messenger RNA was purified from total cytoplasmic and nuclear RNA samples using NEBNext Poly(A) nRNA Magnetic Isolation module (NEB; E7490S) with manufacturer's instructions. Once mRNA was isolated, strand-specific RNA-seq libraries were prepared using NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB; E7420S) with library amplification specifically modified to accommodate the high AT content of *P. falciparum* genome: libraries were amplified for a total of 12 PCR cycles (45 s at 98°C followed by 15 cycles of [15 s at 98°C, 30 s at 55°C, 30 s at 62°C], 5 min 62°C). Libraries were then sequenced on Illumina NExtSeq500 generating 75 bp

paired-end sequence reads.

Sequence Mapping

After sequencing, the quality of raw reads was analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The first 15 bases and the last base was trimmed. Contaminating adaptor reads, reads that were unpaired, bases below 28 and Ns, and reads shorter than 18 bases were also filtered using Sickle (https://github.com/najoshi/sickle). All trimmed reads were then mapped to *P. falciparum* genome (v34) using HISAT2 with the following parameters: –t, -- downstream-transcriptomeassembly, --max-intronlen 3000, --no-discordant, --summary-file, --known-splicesite-infile, --rna-strandness RF, and --novel-splicesite-outfile. After mapping, we removed all reads that were not uniquely mapped, not properly paired (samtools v 0.1.19-44428cd), and are likely to be PCR duplicates (Picard tools v1.78). The final number of working reads for each library is listed in Supplemental Table S4.1. For genome browser tracks, read coverage per nucleotide was first determined using BEDTools and normalized per million mapped reads.

Transcriptome assembly and lncRNA identification

To identify lncRNAs in the nuclear and cytoplasmic fractions, we first merged all nuclear libraries and cytoplasmic liberties into two sets: one nuclear library dataset and one cytoplasmic library dataset. Then cufflinks (v2.1.1) was used for transcriptome assembly with the following parameters: -p 8 -b PlasmoDB-34_Pfalciparum3D7_Genome.fasta -M

PlasmoDB-34_Pfalciparum3D7.gff --library-type first strand -I 5000. After obtaining the assembled transcripts, a minimal read coverage threshold was applied. Transcript abundance calculation was used during Cufflink assembly, and any transcripts with a minimal read coverage below 5 and a FPKM value below 1 were removed. In addition, any transcript with a size shorter than 200 bp was also excluded from down stream analysis. Next, for all remaining transcripts, we calculated its primary transcription level using GRO-seq datasets (GSE85478) from [2] and removed any transcripts that had a read coverage below the 15% of median expression level of all protein encoding genes. After filtering out transcripts with no primary transcription levels, we then removed transcripts overlapped with any annotated gene regions and focused solely on long intergenic non-coding RNA candidates. For those lncRNA that remained, we then calculated their protein potential using Coding Potential Calculator (http://cpc.cbi.pku.edu.cn/). Any lncRNA that was predicted to be coding or weak non-coding RNA with a coding potential score above -1 was removed from our final lncRNA candidate list. To assign cellular locations, log two ratios of total nuclear fraction over total cytoplasmic fractions were calculated. lncRNAs with a ratio above 0.25 are classified as nuclear lncRNA, lncRNAs with a ratio below -0.25 are classified as cytoplasmic lncRNAs, and lncRNA with a ratio between -0.25 and 0.25 are classified as lncRNAs present equally in both fractions.

Overlap between previous intergenic lncRNAs

Overlapping regions between lncRNA candidates and previously identified intergenic

lncRNAs are identified using BEDTools v2.25.0 with at least 25% overlapping between the two fragments (-r -f 0.50).

Estimation of transcript stability

Read coverage values were calculated from total steady-state mRNA datasets (SRP026367, SRS417027, SRS417268, SRS417269) using BEDTools v2.25.0. The read counts are then normalized as described in the original publication, and ratios between RNA-seq and GRO-seq coverage values were calculated for each lncRNA and gene. This ratio reflects the relative abundance of the mature RNA transcript over its corresponding primary transcript and is a simple but convenient measurement for transcript stability.

Western Blot

Mixed-stage parasites were collected as described above. Parasite pellets were gently resuspended in 500 uL of ice cold Cell Fractionation Buffer (PARIS kit, ThermoFisher; AM1921) and 50 uL of 10X EDTA-free Protease inhibitor (cOmplete Tablets, Mini EDTA-free, EASY pack, Roche; 05 892 791 001). Solution was incubated on ice for 10 mins and the sample was centrifuged for 5 mins at 4°C and 500 x g. The supernatant containing cytoplasmic fraction was collected carefully and the nuclear pellet was resuspended in 500 uL Cell Fractionation Buffer followed by needle-lysis 5x using 26 G inch needle. Nuclei were collected again at 4°C and 500 x g. The supernatant was discarded and the nuclei pellet in 500 uL of Cell Disruption Buffer (PARIS kit, ThermoFisher; AM1921) and incubated on ice for 10 minutes. The nuclear fraction was

then sonicated 7x with 10 seconds on/30 seconds off using a probe sonicator. Extracted nuclear protein lysates were incubated for 10 mins at room temperature and centrifuged for 2 mins at 10,000 x g to remove cell debris. Seven micrograms of parasite cytoplasmic and nuclear protein lysates were diluted in 2X laemmli buffer at a 1:1 ratio followed by heating at 95°C for 10 mins. Protein lysates are then loaded on an Any-KD SDS-PAGE gel (Bio-rad) and run for 1 hour at 125 V. Proteins were transferred to a PVDF membrane for 1 hr at 18 V, then stained using commercial antibodies generate against histone H3 (1: 3,000 dilution, Abcam; ab1791) and PfAldolase (1:1,000 dilution, Abcam; ab207494), and secondary antibody, Goat Anti-Rabbit IgG HRP Conjugate (1:25,000 dilution, Bio-Rad; 1706515). Membranes were visualized using the Bio-Rad ChemidDoc MP Gel Imager.

Reverse transcriptase PCR

Total RNA was isolated from 10 mL of mixed-stage asexual *P. falciparum* culture and 25 mL of late gametocyte stage culture. Total RNA quality was checked on an agarose gel and genomic DNA contamination were removed using DNA-free DNA removal kit (ThermoFisher; AM1906) according to manufacturer's instructions. The absence of genomic DNA was validated using a primer set targeting an intergenic region within PfAlba3 (PF3D7_1006200). Approximately 1 µg of DNase I treated RNA from each sample was used in a 35-cycle PCR reaction to confirm the absent of genomic DNA contamination. DNase-treated total RNA was then mixed with 0.1 µg of random hexamers, 0.6 µg of oligo-dT(20), and 2 µL 10 mM dNTP mix (Life Technologies) in

total volume of 10 µL, incubated for 10 minutes at 70°C and then chilled on ice for 5 minutes. This mixture was added to a solution containing 4 µL 10X RT buffer, 8 µL 20 mM MgCl2, 4 µL 0.1 M DTT, 2 µL 20U/µl SuperaseIn and 1 µL 200 U/µL SuperScript III Reverse Transcriptase (Life Technologies). First-strand cDNA was synthesized by incubating the sample for 10 minutes at 25°C, 50 minutes at 50°C, and finally 5 minutes at 85°C. First strand cDNA is then mixed with 70 µL of nuclease free water, 30 µL 5x second-strand buffer (Life Technologies), 3 µL 10 mM dNTP mix (Life Technologies), 4 µL 10 U/µl *E. coli* DNA Polymerase (NEB), 1 µL 10 U/µL E. coli DNA ligase (NEB) and 1 µL 2 U/µL *E. coli* RNase H (Life Technologies). Samples were incubated for 2 h at 16°C and double stranded cDNA was purified using AMPure XP beads (Beckman Coulter). For testing transcription activity of predicted genes, 450 ng of double stranded cDNA was mixed with 10 pmole of both forward and reverse primers. DNA was incubated for 5 minutes at 95°C, then 30s at 98°C, 30s at 55°C, 30s at 62°C for 25 cycles. All primers used for PCR validation are listed in Supplemental File 4.1.

Chromatin Isolation by RNA purification (ChIRP)

Synchronized parasite cultures were collected and incubated in 0.15% saponin for 10 min on ice to lyse red blood cells. Parasites were centrifuged at 3234 x g for 10 mins at 4°C and subsequently washed three times with PBS by resuspending in cold PBS and centrifuging for 10 mins at 3234 x g at 4°C. Parasites were cross-linked for 15 mins at RT with 1% glutaraldehyde. Cross-linking was quenched by adding glycine to a final concentration of 0.125 M and incubating for 5 mins at 37°C. Parasites were centrifuged

at 2500 x g for 5 mins at 4°C, washed three times with cold PBS and stored at -80°C.

To extract nuclei, parasite were first incubated on in nuclear extraction buffer (10 mM HEPES, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 0.5 mM 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF), EDTA-free protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Roche)) on ice. After 30 mins, Igepal CA-360 (Sigma-Aldrich) was added to a final concentration of 0.25% and needle sheared seven times by passing through a 26 G ½ needle. Parasite nuclei were centrifuged at 2500 x g for 20 mins at 4°C and resuspended in shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris–HCl pH 7.5, EDTA-free protease inhibitor cocktail and phosphatase inhibitor cocktail). Chromatin was fragmented using the Covaris Ultra Sonicator (S220) to obtain 100-500 bp DNA fragments with the following settings: 5% duty cycle, 140 intensity incident power, 200 cycles per burst. Sonicated samples were centrifuged for 10 mins at 17000 x g at 4°C to remove insoluble material.

Fragmented chromatin was precleared using Dynabeads MyOne Streptavidin T1 (Thermo Fisher) by incubating for 30 mins at 37°C to reduce non-specific background. Per ChIRP sample using 1 mL of lysate, 10 uL each was removed for the RNA input and DNA input, respectively. Each sample was diluted in 2x volume of hybridization buffer (750mM NaCl, 1% SDS, 50mM Tris-Cl pH 7.5, 1mM EDTA, 15% formamide, 0.0005x volume of AEBSF, 0.01x volume of Superase-in (Ambion) and 0.01x volume of protease inhibitor cocktail). ChIRP probes used for each lncRNA (see Table X) were pooled,

heated at 85°C for 3 mins and cooled on ice. ChIRP probes were added to each sample (2 uL of 100 uM pooled probes per sample) and incubated at 37°C with end-to-end rotation for 4 hours. Prior to completion of hybridization, Dynabeads MyOne Streptavidin T1 beads were washed three times on a magnet stand using lysis buffer (50mM Tris-Cl pH 7, 10mM EDTA, 1% SDS). After the hybridization, 100 uL of washed T1 beads were added to each tube and incubated for 30 mins at 37°C. Beads were washed with wash buffer (2x SSC, 0.5% SDS, 0.005x volume of AEBSF) and split evenly for isolation of DNA and RNA fractions.

For RNA isolation, the RNA input and beads were resuspended in RNA elution buffer (100mM NaCl, 10mM Tris-HCl pH7.0, 1mM EDTA, 0.5% SDS, 1mg/mL Proteinase K), incubated at 50°C for 45 mins and boiled at 95°C for 15 mins.

Libraries from the ChIRP samples were prepared using the KAPA Library Preparation Kit (KAPA Biosystems). Libraries were amplified for a total of 12 PCR cycles (12 cycles of [15 s at 98°C, 30 s at 55°C, 30 s at 62°C]) using the KAPA HiFi HotStart Ready Mix (KAPA Biosystems). Libraries were sequenced with a NextSeq500 DNA sequencer (Illumina). Raw read quality was first analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the last base were removed using Trimmomatic. Any base with a quality score below 25 was trimmed using Sickle (https://github.com/najoshi/sickle). Trimmed reads were then mapped to *P. falciparum* genome (v38) using Bowtie2 (v2.3.4.1). Uniquely mapped reads were further filtered.

Read coverage per nucleotide was determined using BEDTools. All libraries, including the input, were then normalized by dividing through the numbers of million mapped reads. For each nucleotide, the signal from the input library was then subtracted from each of the ChIRP-seq libraries and any negative value was replaced with a zero. Genome browser tracks were generated and viewed using the Integrative Genomic Viewer (IGV) by Broad institute.

RNA in situ hybridization (RNA-FISH)

RNA FISH was performed with slight modifications as described by [Mancio-Silva, Malaria 2012] on mixed-stage asexual and gametocyte stage parasites. Antisense RNA probes for seven nuclear lncRNAs; -TARE4, -178, -13, -1494, -271, -ch9, -ch14 and two cytoplasmic lncRNAs; -267, -643, were labeled by in vitro transcription in the presence of fluorescein. RNA FISH was also performed using sense RNA probes as controls. Briefly, fixed and permeabilized parasites were incubated with RNA probes overnight at 37°C. Parasites were washed with 2x SSC three times for 15 mins each at 45°C followed by one wash with 1x PBS for 5 mins at room temperature. The slides were mounted in Vectashield mounting medium with DAPI and visualized using the Olympus BX40 epifluorescence microscope.

References

1. WHO: **The World Malaria Report.** http://www.who.int/malaria/publications/worldmalaria-report-2017/en/**.** 2017.

2. Lu XM, Batugedara G, Lee M, Prudhomme J, Bunnik EM, Le Roch KG: **Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite Plasmodium falciparum.** *Nucleic Acids Res* 2017, **45:**7825-7840.

3. Reid AJ, Talman AM, Bennett HM, Gomes AR, Sanders MJ, Illingworth CJR, Billker O, Berriman M, Lawniczak MK: **Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites.** *Elife* 2018, **7**.

4. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, Llinas M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proc Natl Acad Sci U S A* 2008, **105:**8393-8398.

5. Gomez-Diaz E, Yerbanga RS, Lefevre T, Cohuet A, Rowley MJ, Ouedraogo JB, Corces VG: **Epigenetic regulation of Plasmodium falciparum clonally variant gene expression during development in Anopheles gambiae.** *Sci Rep* 2017, **7:**40655.

6. Gupta AP, Chin WH, Zhu L, Mok S, Luah YH, Lim EH, Bozdech Z: **Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum.** *PLoS Pathog* 2013, **9:**e1003170.

7. Bunnik EM, Batugedara G, Saraf A, Prudhomme J, Florens L, Le Roch KG: **The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum.** *Genome Biol* 2016, **17:**147.

8. Lacsina JR, LaMonte G, Nicchitta CV, Chi JT: **Polysome profiling of the malaria parasite Plasmodium falciparum.** *Mol Biochem Parasitol* 2011, **179:**42-46.

9. Mair GR, Braks JA, Garver LS, Wiegant JC, Hall N, Dirks RW, Khan SM, Dimopoulos G, Janse CJ, Waters AP: **Regulation of sexual development of Plasmodium by translational repression.** *Science* 2006, **313:**667-669.

10. Shock JL, Fischer KF, DeRisi JL: **Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle.** *Genome Biol* 2007, **8:**R134.

11. Bunnik EM, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Cort A, Shi L, Andolina C, Ross LS, Brady D, et al: **Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages.** *Nat Commun* 2018, **9:**1910.

12. Miller LH, Good MF, Milon G: **Malaria pathogenesis.** *Science* 1994, **264:**1878-1883.

13. Scherf A, Lopez-Rubio JJ, Riviere L: **Antigenic variation in Plasmodium falciparum.** *Annu Rev Microbiol* 2008, **62:**445-470.

14. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW: **Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum.** *RNA* 2009, **15:**116-127.

15. Li F, Sonbuchner L, Kyes SA, Epp C, Deitsch KW: **Nuclear non-coding RNAs are transcribed from the centromeres of Plasmodium falciparum and are associated with centromeric chromatin.** *J Biol Chem* 2008, **283:**5692-5698.

16. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.** *Cell* 2007, **129:**1311-1323.

17. Corcoran AE: **The epigenetic role of non-coding RNA transcription and nuclear organization in immunoglobulin repertoire generation.** *Semin Immunol* 2010, **22:**353-361.

18. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, et al: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.** *Nature* 2010, **464:**1071-1076.

19. Ransohoff JD, Wei Y, Khavari PA: **The functions and unique features of long intergenic non-coding RNA.** *Nat Rev Mol Cell Biol* 2018, **19:**143-157.

20. Engreitz JM, Ollikainen N, Guttman M: **Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression.** *Nat Rev Mol Cell Biol* 2016, **17:**756-770.

21. Quinodoz S, Guttman M: **Long noncoding RNAs: an emerging link between gene regulation and nuclear organization.** *Trends Cell Biol* 2014, **24:**651-663.

22. Nakagawa S, Kageyama Y: **Nuclear lncRNAs as epigenetic regulators-beyond skepticism.** *Biochim Biophys Acta* 2014, **1839:**215-222.

23. Rashid F, Shah A, Shan G: **Long Non-coding RNAs in the Cytoplasm.** *Genomics Proteomics Bioinformatics* 2016, **14:**73-80.

24. Carlevaro-Fita J, Rahim A, Guigo R, Vardy LA, Johnson R: **Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells.** *RNA* 2016, **22:**867-882.

25. Tichon A, Gil N, Lubelsky Y, Havkin Solomon T, Lemze D, Itzkovitz S, Stern-Ginossar N, Ulitsky I: **A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells.** *Nat Commun* 2016, **7:**12209.

26. Maclary E, Hinten M, Harris C, Kalantry S: **Long nonoding RNAs in the X-inactivation center.** *Chromosome Res* 2013, **21:**601-614.

27. Schoeftner S, Blasco MA: **Chromatin regulation and non-coding RNAs at mammalian telomeres.** *Semin Cell Dev Biol* 2010, **21:**186-193.

28. Feuerhahn S, Iglesias N, Panza A, Porro A, Lingner J: **TERRA biogenesis, turnover and implications for function.** *FEBS Lett* 2010, **584:**3812-3818.

29. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC: **Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA.** *BMC Genomics* 2015, **16:**454.

30. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, Ribacke U, Volkman S, Duraisingh M, Wirth D, Sabeti PC, Rinn JL: **A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs.** *Genome Biol* 2011, **12:**R56.

31. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, et al: **Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum.** *Genome Res* 2008, **18:**281-292.

32. Amit-Avraham I, Pozner G, Eshar S, Fastman Y, Kolevzon N, Yavin E, Dzikowski R: **Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 2015, **112:**E982-991.

33. Sierra-Miranda M, Delgadillo DM, Mancio-Silva L, Vargas M, Villegas-Sepulveda N, Martinez-Calvillo S, Scherf A, Hernandez-Rivas R: **Two long non-coding RNAs generated from subtelomeric regions accumulate in a novel perinuclear compartment in Plasmodium falciparum.** *Mol Biochem Parasitol* 2012, **185:**36-47.

34. Raabe CA, Sanchez CP, Randau G, Robeck T, Skryabin BV, Chinni SV, Kube M, Reinhardt R, Ng GH, Manickam R, et al: **A global view of the nonprotein-coding transcriptome in Plasmodium falciparum.** *Nucleic Acids Res* 2010, **38:**608-617.

35. Luke B, Lingner J: **TERRA: telomeric repeat-containing RNA.** *EMBO J* 2009, **28:**2503-2510.

36. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, Le Roch KG: **DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum.** *BMC Genomics* 2014, **15:**347.

37. Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, Xu L, Li X, Zhao Y, Zheng H, et al: **Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data.** *Parasitol Res* 2014, **113:**1269-1281.

38. Sun M, Gadad SS, Kim DS, Kraus WL: **Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells.** *Mol Cell* 2015, **59:**698-711.

39. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS: **Genome-wide analysis of long noncoding RNA stability.** *Genome Res* 2012, **22:**885-898.

40. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY: **Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.** *Mol Cell* 2011, **44:**667-678.

41. Quinn JJ, Ilik IA, Qu K, Georgiev P, Chu C, Akhtar A, Chang HY: **Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification.** *Nat Biotechnol* 2014, **32:**933-940.

42. Winter G, Kawai S, Haeggstrom M, Kaneko O, von Euler A, Kawazu S, Palm D, Fernandez V, Wahlgren M: **SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes.** *J Exp Med* 2005, **201:**1853-1863.

43. Pei X, Guo X, Coppel R, Bhattacharjee S, Haldar K, Gratzer W, Mohandas N, An X: **The ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum stabilizes spectrin tetramers and suppresses further invasion.** *Blood* 2007, **110:**1036-1042.

44. Gong C, Maquat LE: **lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.** *Nature* 2011, **470:**284-288.

45. Kim YK, Furic L, Parisien M, Major F, DesGroseillers L, Maquat LE: **Staufen1 regulates diverse classes of mammalian transcripts.** *EMBO J* 2007, **26:**2670-2681.

46. Hollander MC, Alamo I, Fornace AJ, Jr.: **A novel DNA damage-inducible transcript, gadd7, inhibits cell growth, but lacks a protein product.** *Nucleic Acids Res* 1996, **24:**1589-1593.

47. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M: **LincRNA-p21 suppresses target mRNA translation.** *Mol Cell* 2012, **47:**648-655.

48. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al: **Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.** *Nature* 2012, **491:**454-457.

49. Guil S, Esteller M: **Cis-acting noncoding RNAs: friends and foes.** *Nat Struct Mol Biol* 2012, **19:**1068-1075.

50. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al: **Long noncoding RNAs with enhancer-like function in human cells.** *Cell* 2010, **143:**46-58.

51. Ng SY, Bogu GK, Soh BS, Stanton LW: **The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis.** *Mol Cell* 2013, **51:**349-359.

52. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, et al: **The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex.** *Nat Genet* 2013, **45:**1392-1398.

53. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY: **Long noncoding RNA as modular scaffold of histone modification complexes.** *Science* 2010, **329:**689-693.

54. Mele M, Rinn JL: **"Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription.** *Mol Cell* 2016, **62:**657-664.

55. Rinn J, Guttman M: **RNA Function. RNA and dynamic nuclear organization.** *Science* 2014, **345:**1240-1241.

56. Cerase A, Pintacuda G, Tattermusch A, Avner P: **Xist localization and function: new insights from multiple levels.** *Genome Biol* 2015, **16:**166.

57. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, et al: **Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.** *Nat Struct Mol Biol* 2014, **21:**198-206.

58. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB: **An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.** *Mol Cell* 2009, **33:**717-726.

59. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: **Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression.** *Genome Res* 2014, **24:**974-988.

60. Hanly DJ, Esteller M, Berdasco M: **Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer?** *Philos Trans R Soc Lond B Biol Sci* 2018, **373**.

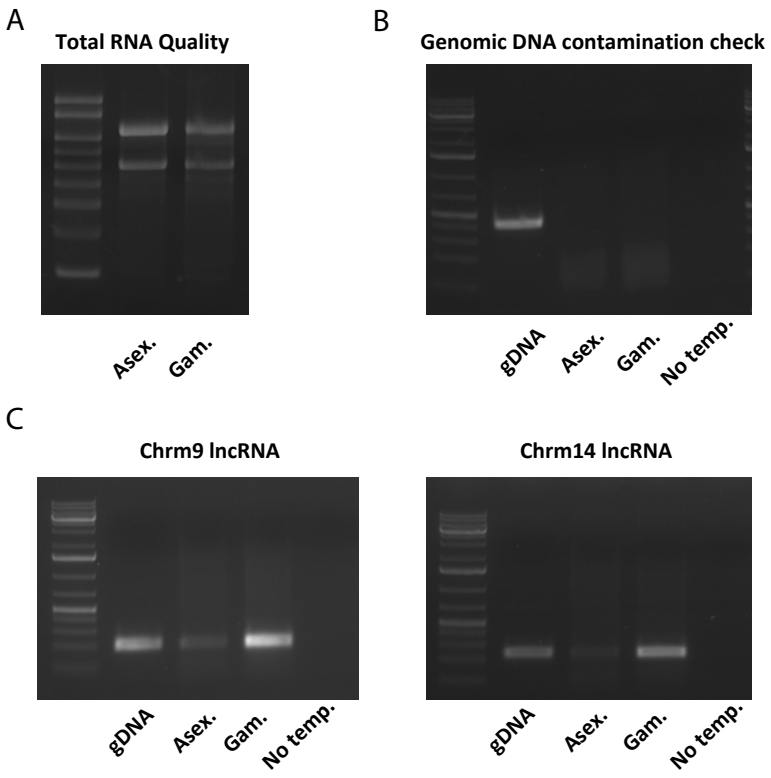
61. Vance KW, Ponting CP: **Transcriptional regulatory functions of nuclear long noncoding RNAs.** *Trends Genet* 2014, **30:**348-355.

62. Wierzbicki AT: **The role of long non-coding RNA in transcriptional gene silencing.** *Curr Opin Plant Biol* 2012, **15:**517-522.

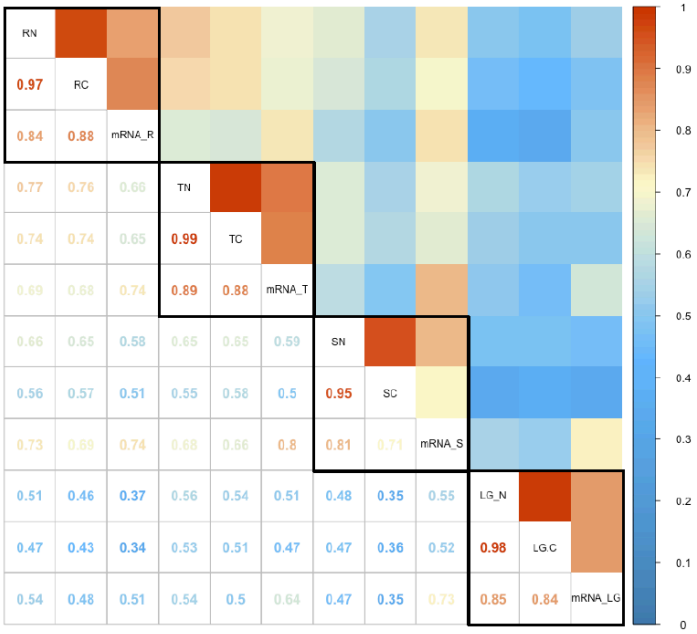
63. Trager W, Jensen JB: **Human malaria parasites in continuous culture. 1976.** *J Parasitol* 2005, **91:**484-486.

64. Ifediba T, Vanderberg JP: **Complete in vitro maturation of Plasmodium falciparum gametocytes.** *Nature* 1981, **294:**364-366.

Supplemental Material



Supplemental figure 4.1: RT-PCR validation of selected lncRNAs. (A) Total RNA was extracted from both asexual and gametocyte stage parasites. RNA quality was validated on an agarose gel. (B) Genomic DNA was removed and verified using RT-PCR with primers designed to amplify a fragment of the PfAlba3 gene (PF3D7_1006200). Primers were designed on both sides of intron 1, yielding a 429 bp PCR product from genomic DNA and a 164 bp PCR product from cDNA. The absence of PCR product amplified from RNA confirms the absence of gDNA contamination. (C) RT-PCR validation of two selected lncRNA that are most abundantly expressed at the gametocyte stages.



Supplemental figure 4.2: Correlation analysis. Spearman correlations in gene expression levels among nuclear fractions, cytoplasmic fractions, and steady-state mRNA across *P. falciparum* cell cycle.

Supplemental file 4.1: List of identified lncRNAs (XLSX)

CONCLUSIONS

Malaria is one of the deadliest infectious diseases worldwide, and is caused by apicomplexan parasites of the *Plasmodium* species. Among the species that infect humans; *P. falciparum*, the most prevalent and deadly human malaria parasite, is responsible for approximately 445,000 malaria-related deaths per year, most of which occur in sub-Saharan Africa [1]. *P. vivax* is also responsible for significant disease, mostly in Southeast Asia. Children under the age of five, and pregnant women are most susceptible to the disease. Despite the billions of dollars invested each year into reducing the spread of malaria and treating the infected, efforts to eradicate this disease have been largely unsuccessful.

Plasmodium parasites have a complex life cycle. The asexual stages of the parasite (ring, trophozoite and schizont) are responsible for pathogenesis, while the sexual stages are responsible for disease transmission. Understanding how the transitions between the various life cycle stages of the *Plasmodium* parasite are regulated remains an important goal in malaria research. These developmental stages are characterized by changes in gene expression as well as changes in global chromatin structure [2-5]. Chromatin is a complex of DNA and proteins that makeup chromosomes within a cell so that DNA can be packaged to fit inside the nucleus. We now know that decrypting the information encoded in the linear DNA sequence is not sufficient to understand its function. Therefore, we need to understand how genetic information is organized and regulated at the three-dimensional (3D) level.

P. falciparum uses a combination of different epigenetic mechanisms to regulate its gene expression. However, our understanding of the parasite epigenome is far from complete. Although most chromatin modifications used by the parasite are also common to other eukaryotes, several features of chromatin regulation are unique to *P. falciparum*. Exploring the underlying regulatory mechanisms of how parasite chromatin structure is established and maintained could lead to identification of specific molecules (i.e proteins and lncRNAs) important for chromatin regulation in the malaria parasite.

At the epigenetic level, the *Plasmodium* genome architecture points towards a binary structure, with the majority of the genome existing as transcriptionally permissive euchromatin and a small subset of genes present in a transcriptionally silent heterochromatin state [2, 6]. This heterochromatin cluster is localized at the periphery of the nucleus and is characterized by high levels of H3K9me3 and H3K36me3 histone marks, PfHP1 and high nucleosome density [7, 8]. The euchromatin environment harbors active genes, including the single active *var* gene at the nuclear periphery, and is characterized by high levels of H3K4me3 and H3K9ac histone modifications. During the asexual cycle, the parasite nucleus and chromatin undergo drastic remodeling to accommodate the high transcriptional activity at the trophozoite stage [2]. Chromatin structure remains relatively compact during the ring and schizont stages, but opens substantially during the trophozoite stage. This open-and-close chromatin structure is also reflected at the nucleosome landscape and global histone levels [9-12]. As the parasite

transitions from the trophozoite stage to the schizont stage, the changes in nuclear architecture are reversed by reassembling nucleosomes, increasing global histone levels and compacting the genome. In the transition from trophozoite to schizont stage, the DNA is replicated, and the nucleus is divided into multiple daughter nuclei. Collectively, these observations suggest that the majority of the parasite genome is regulated via genome-wide changes in chromatin structure, while a small subset of genes are regulated by classical transcriptional regulation mechanisms, such as changes in local chromatin structure and specific transcription factors. Exploration of regulatory mechanisms that regulate large chromatin rearrangements throughout the parasite life cycle could enable the discovery of molecules that can target parasite development with high specificity.

Previously, we assessed genome organization at the ring, trophozoite, and schizont stages of the IDC in *P. falciparum* using Hi-C experiments (chromosome conformation capture coupled with next-generation sequencing) [2]. We observed a strong association between genome architecture and gene expression, suggesting that the 3D organization of the genome is very important for gene regulation. In the first chapter, we analyze genome organization in the *P. falciparum* and *P. vivax* transmission stages. Major changes occur in the localization and interactions of genes involved in pathogenesis and immune evasion, host cell invasion, sexual differentiation, and master regulation of gene expression. Furthermore, we observe reorganization of subtelomeric heterochromatin around genes involved in host cell remodeling. Depletion of heterochromatin protein 1 (PfHP1) resulted in loss of interactions between virulence genes, confirming that PfHP1

is essential for maintenance of the repressive center. In the second chapter we further compare genome organization of five other *Plasmodium* species and two related apicomplexan species. Genome organization was dominated by the clustering of *Plasmodium*-specific gene families in 3D space. In particular, the two most pathogenic human malaria parasites shared unique features in the organization of gene families involved in antigenic variation and immune escape. Related human parasites *Babesia microti* and *Toxoplasma gondii* that are less virulent lacked the correlation between gene expression and genome organization observed in human *Plasmodium* species. Our results provide important novel insights into the connection between genome organization, heterochromatin, and stage-specific gene expression.

As chromatin structure and genome architecture is particularly important for global transcriptional regulation in malaria parasites, chromatin-associated regulators may be promising targets for anti-malarial therapies. Therefore, in the later chapters, we focused on the identification of potential chromatin-associated regulators such as proteins and lncRNAs. By carefully surveying the parasite proteome, we were able to generate the first and most up-to-date comprehensive overview of the *Plasmodium* chromatin-associated proteome. We have further validated the cellular localization and expression for two candidate chromatin-associated proteins (CAPs). The function of many CAPs is still unknown and further characterization of CAPs is needed to increase our understanding of parasite biology thus providing helpful insights to better understand gene expression regulation in the parasite. Apart from proteins, long non-coding RNAs

(lncRNAs) have also been shown to have important roles in both chromatin structure regulation and gene expression. Therefore, in the last chapter of this dissertation, we performed genome-wide identifications of both nuclear and cytoplasmic lncRNAs within the *P. falciparum* genome. As a result, we identified over a 1000 lncRNAs that are differentially expressed, not only between subcellular locations but also across the parasite's life cycle. We further explore the regulatory roles of several candidate lncRNAs by investigating their genome-wide occupancy sites. Using a novel methodology adapted to the parasite genome, our results revealed that selected lncRNA occupancy sites within the parasite genome are focal and sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis, erythrocyte remodeling, and regulation of sexual differentiation. Our findings bring a new level of insight into the role of lncRNAs in chromatin structure and genome organization in the parasite.

Collectively, using genome-wide datasets from these novel methodologies, we hope to identify not only components that regulate chromatin structure, but also key regulators of *P. falciparum* gene expression. Parasite-specific protein and lncRNA candidates that can be validated at the functional level will be ideal as drug targets that disrupt the parasite 3D nuclear structure. It is likely that our results will not only improve our understanding of chromatin structure and genome biology but will also help to identify key players in pathogenesis and sexual differentiation in malaria parasites.

Future Perspectives

An overwhelming amount of evidence now points towards epigenetic landscape and nuclear architecture regulating gene expression in eukaryotes. The human malaria parasite, *Plasmodium falciparum* uses a combination of different epigenetic mechanisms to regulate its gene expression. However, our understanding of the parasite epigenome is not yet complete. At the epigenetic level, a majority of the parasite genome exists in a transcriptionally permissive euchromatin state. Additionally, a small subclass of genes is harbored in a transcriptionally repressive heterochromatin state, and the maintenance and regulation of this repressive heterochromatin environment within the parasite nucleus is essential for parasite survival. Collectively, our studies investigate the three-dimensional chromosomal organization in malaria parasites and highlight the possible connections between genome architecture and pathogenicity. Given the impact of chromatin structure and genome organization on parasite gene expression, it is important to identify molecular components that maintain and regulate these processes. In the past decade, several inhibitors targeting histone deacetylases (HDACs) and histone methyltransferases (HMTs) have been shown to interfere with parasite growth and survival. In addition, conditional deletions of heterochromatin protein 1 (PfHP1) or histone deacetylase 2 (PfHDA2) have been shown to cause developmental arrest in blood-stage parasites. While targeting histone modifying enzymes and proteins such as PfHP1, which are vital for chromatin structure maintenance, are promising targets for antimalarial therapies, many of these proteins are well conserved among eukaryotic organisms and when targeted for antimalarial therapies could be toxic to the human host as well. Therefore, it

is important to identify parasite-specific components that could distrust chromatin structure and genome organization with high specificity and low toxicity to the host. Several potential chromatin-associated proteins and long non-coding RNAs that are parasite-specific and are possible regulators of genome architecture and gene expression in the parasite have been identified in this body of work. It is my hope that molecular regulators that can be validated at the functional level will provide possible new targets for novel antimalarial therapies.

References

1. WHO: **The World Malaria Report.** http://www.who.int/malaria/publications/worldmalaria-report-2017/en/**.** 2017.

2. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: **Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression.** *Genome Res* 2014, **24:**974-988.

3. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1:**E5.

4. Bunnik EM, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Cort A, Shi L, Andolina C, Ross LS, Brady D, et al: **Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages.** *Nat Commun* 2018, **9:**1910.

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

6. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A: **Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum.** *Nature* 2000, **407:**1018-1022.

7. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, Rojas-Meza AP, Mancio-Silva L, Leal-Silvestre RJ, Gontijo AM, Shorte S, Scherf A: **Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites.** *Cell* 2005, **121:**25-36.

8. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AM, Cohen AM, Janssen-Megens EM, van de Vegte-Bolmer M, Selzer RR, Iniguez AL, et al: **Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 2009, **106:**9655-9660.

9. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, Le Roch KG: **DNA-encoded nucleosome occupancy is associated with transcription**

levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics* 2014, **15:**347.

10. Ponts N, Harris EY, Lonardi S, Le Roch KG: **Nucleosome occupancy at transcription start sites in the human malaria parasite: a hard-wired evolution of virulence?** *Infect Genet Evol* 2011, **11:**716-724.

11. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG: **Nucleosome landscape and control of transcription in the human malaria parasite.** *Genome Res* 2010, **20:**228-238.

12. Westenberger SJ, Cui L, Dharia N, Winzeler E, Cui L: **Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes.** *BMC Genomics* 2009, **10:**610.