Title

An Exploration of Transcriptional Regulation in the Human Malaria Parasite, Plasmodium falciparum

Author

Lu, Xueqing

Publication Date

2017

UNIVERSITY OF CALIFORNIA
RIVERSIDE

An Exploration of Transcriptional Regulation in the Human Malaria Parasite, *Plasmodium falciparum*

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

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

Xueqing Lu

December 2017

Dissertation Committee:
Dr. Karine Le Roch, Chairperson
Dr. Thomas Girke
Dr. Stefano Lonardi

The Dissertation of Xueqing Lu is approved:

____

____

____
Committee Chairperson

University of California, Riverside

Acknowledgements

First of all, I would like to express my gratitude for being a part of the Le Roch lab family. I especially thank Dr. Karine Le Roch for her continuous guidance, and Dr. Evelien Bunnik for being my mentor, who trained me and showed me how to be a scientist. Karine, I will never forget the hours you spent with me after my second year annual report, the Trader Joe goodies, the bubblies, the fancy cheese, and the la Galettes des rois that you brought for us over the years. Though the learning curve was steep in the beginning, your understanding and patience made the completion of work possible. Thank you for taking me into your lab and providing me with the environment and time to grow into an independent scientist. Evelien, I cannot thank you enough for having the patience to answer my countless questions, and teaching me all the computational and experimental techniques. Thank you for being my mentor and personal lab-wikipedia. Because of your time, critical feedbacks, visions and advice, I was able to finish this dissertation and become the bioinformatican that I am today.

I would like to express my sincere thanks to my Le Roch family members, especially Jacques Prudhomme, Gayani Batugedara, and the extension lab member/fantastic housemate, Hailey Choi. Thanks you Jacq for being the culture king of our lab and providing me with the parasites that are necessary to complete all my experiments. Thank you Gyni and Hailey for keeping the lab an enjoyable place. Thanks for "struggling" with me and updating me with all the "news", so I know I am never alone or being forever in my own bubble. Most importantly, thank you for being there whenever I felt blue or troubled. A big thank you for Hailey, my "life essential" housemate, who showered me with countless bake goods and delicious dinners during the past few years. You are always so helpful whenever there is a Maggie's dilemma. My journey to

finish graduate school and the work in this dissertation would never be fun and memorable without you all. It was my greatest pleasure to work with these great scientists.

I would also like to thank my wonderful dissertation committee members, Dr. Thomas Girke and Dr. Stefano Lonardi. Thank you Dr. Girke for helping me with coding issues and answering lists of computational questions. Thank you Dr. Lonardi for writing me the recommendation letters, which resulted in fellowships that allowed me to focus on my projects in the last year of graduate school. Because of all your help, I am able to finish my degree on time.

Lastly, I would like to acknowledge my family and friends. Thank you Sara Park, Pear, and Thais Choi for always have confidence in me and cheering me up whenever I am down. Thank you mom and dad for always being there for me and supporting every decision that I made. Because of you all, I had the courage and the dedication to keep moving forward and finish my doctoral work. Thanks for your care and love.

The text of this dissertation (or thesis), in part or in full, is a reprint of the material as it appears in "Analysis of Nucleosome Positioning Landscapes Enables Gene Discovery in the Human Malaria Parasite *Plasmodium falciparum,"* BMC Genomics (2015). The co-authors Xueqing Maggie Lu and Evelien M. Bunnik are equal contributors to this manuscript. Stefano Lonardi and Karine Le Roch directed and supervised the research, which forms the basis for this publication.

The text of this dissertation (or thesis), in part or in full, is a reprint of the material as it appears in "Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite *Plasmodium falciparum*," Nucleic Acids Research (2017). The co-author Xueqing Maggie Lu performed all computational analyses, generated GRO-seq datasets, participated in study design and drafted the manuscript. Evelien M. Bunnik and Karine Le Roch directed and supervised the research, which forms the basis for this publication.

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). The co-authors Gayani Batugedara drafted the majority of the manuscript. Xueqing Maggie Lu drafted partially of the manuscript. Karine Le Roch supervised the finalized the manuscript for publication.

Dedication

I dedicate this work to my family, especially my parents, who have nurtured me and offered enormous support during the past five years of my doctoral journey. You are the ones that taught me to never give up on my education and encouraged me to pursue my dreams with a passion.

献给最爱我的和我最爱的爸爸妈妈。

ABSTRACT OF THE DISSERTATION

An Exploration of Transcriptional Regulation in the Human Malaria Parasite, *Plasmodium falciparum*

by

Xueqing Lu

Genetics, Genomics, and Bioinformatics
University of California, Riverside, December 2017
Dr. Karine Le Roch, Chairperson

Malaria is one of the most lethal infectious diseases in many developing countries. Approximately half of the world's population is at risk of malaria transmission, and this number can be expected to grow as drug resistant strains continue to develop. Among the human infectious *Plasmodium* species, *Plasmodium falciparum* causes the most severe and lethal form of malaria. This parasite has an extreme AT-rich genome and a complex life cycle that is likely to be regulated by coordinate changes in gene expression. However, the mechanisms behind this fine-tuned gene expression and regulation system remain elusive. For instance, only a limited number of transcription factors have been identified. Recent studies suggest that epigenetic and post-transcriptional regulation may be used as alternative regulation strategies to compensate for the lack of transcription factors in this parasite. Therefore, in this dissertation work, we further explored the transcriptome, epigenome, and the proteome to better understand the transcriptional mechanisms in *P. falciparum*. In chapter 1, we demonstrated that genes are usually defined by unique nucleosomal features and that nucleosome landscape alone could be used to identify novel genes in organisms with a nucleotide bias. Next, we investigated nascent RNA expression profiles and observed that the majority of genes are transcribed at the trophozoite stage in response to the open chromatin structure of that stage. These results helped us link chromatin reorganization events to transcriptional activity and highlighted the importance of epigenetic and post-transcriptional regulation in this parasite. Therefore, in the latter two chapters, we further examined the proteasome and transcriptome isolated from both nuclear and cytoplasmic fractions to identify potential chromatin regulators. As a result, we identified a large number of chromatin-associated proteins and lncRNAs that are likely to have important roles in chromatin regulation and post-transcriptional and translational regulations. Collectively, data and results from these studies will become stepping-stones for future malaria studies and further assist the identification of promising anti-malarial drug targets.

Table of Contents

List of Figure

Lists of Tables

List of Supplemental Files

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: Enriched chromatin-associated proteins in the ChEP sample, P. falciparum nuclear proteome and the *in silico* analysis. (XLSX)

Supplemental File 3.5: Proteins associated with Structural Maintenance of Chromosomes Protein 3 (SMC3) during the IDC. (XLSX)

Introduction

Malaria

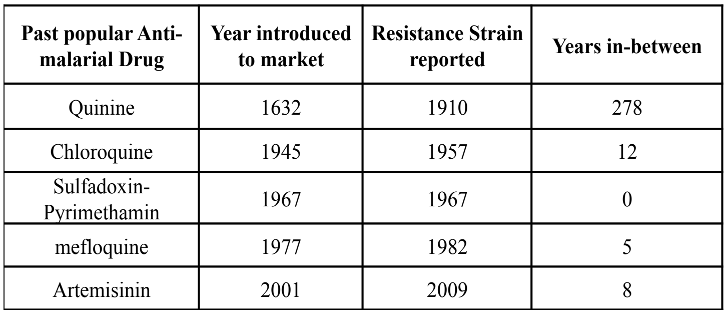
Malaria is one of the most life-threatening infectious diseases remaining in the world. In 2016, approximately half of the world population was at risk of malaria transmission with 212 million people infected and 429,000 people killed by the disease [1]. Malaria is caused by apicomplexan parasites from the genus *Plasmodium* and is transmitted by an infected female Anopheles mosquito. Through a mosquito bite, parasites are introduced into human host and cause the host to suffer from fevers, chills, and anemia. In severe cases, the large number of parasites accumulating in the blood stream may cause a blockage of the vessels that can result in coma or death of the host. Malaria-related incidences and mortality rates are especially high for infants, children under age of five, and pregnant women in developing countries, especially within the sub-Saharan Africa regions [1].

Currently, no licensed or commercially available vaccine has been announced for malaria, but one candidate, RTS,S/AS01 developed by the PATH Malaria Vaccine Initiative (MVI) for *P. falciparum* infection, is well advanced and is now in Phase 3 trial. As the safety and efficiency of this vaccine is still under study, vector control strategies, such insecticide-treated mosquito nets (ITNs) and indoor residual spraying (IRS) with insecticides, are recommended and are the most sufficient ways to prevent malaria.

The initial symptoms of malaria typically begin 8-25 days after infection and are similar to flu-like symptoms including headache, fever, shivering, and vomiting. Due to the non-specific presentation of malaria symptoms, malaria diagnosis requires a high degree of specification. For the past century, the most reliable technique for malaria diagnosis has been microscopic analysis at blood films. Such technique is labor intensive, time consuming, and often restricted by access to instruments and facilities. As an alternative, Rapid Diagnostic Tests (RDTs) have been implemented for diagnosing people with malaria at high parasitemia and at locations beyond the reach of microscopy; however, RDTs are unable to distinguish a new malaria infection from recent clearance of infection [4]. Therefore, more accurate and sensitive malaria diagnosing tools are still needed.

Malaria is a curable disease. Currently used anti-malarial drugs are chloroquine, mefloquine, doxycycline, atovaquone, proguanil hydrochloride, and artemisinins. The majority of these drugs are effective for treating malarias. However, due to the adaptable nature of the parasite and its vector, malaria parasite have inevitably evolved for nearly all effective drugs **(Table 1.1)**. Drug resistance strains have posed a growing problem in modern years. To prolong drug resistance, an artemisinin-based combination therapy (ACT) is given as treatment for resistant infections, but the rapid development of drug-resistant strains remains the biggest challenge for combating malaria.

Table I.1. Time frame of anti-malarial drugs and corresponding *Plasmodium falciparum* drug resistant strain development (Sources: World Health Organization, 2003, Medicines for Malaria Venture, 2015, and O'Brien et al., 2011 [3])



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

Plasmodium falciparum

There are five different *Plasmodium* species known for human infection: *P. falciparum, P. vivax, P. malariae, P. ovale, and P. knowlesi*. Among the five species, *P. falciparum* is responsible for the most severe form of malaria and causes about 90% of all malarial deaths. This parasite has a complex life cycle involving multiple hosts **(Figure 1.1)**. Through an infectious mosquito bite, parasite sporozoites are injected into the human blood stream. The parasites travel to the liver and invade liver cells [5, 6]. Inside the liver cells, sporozoites go through rapid DNA synthesis and mitosis to produce thousands of haploid daughter merozoites that are released into the blood stream and invade red blood cells (RBCs) [7]. After entering the RBCs, the parasites replicate asexually and develop through ring, trophozoite, and schizont stages [8]. This asexual cycle is also known as the clinically symptomatic intraerythrocytic developmental cycle (IDC). Each parasite produces 16-32 daughter cells, or merozoites, every 48 hours. These merozoites are then released into the blood stream and the replication cycle starts again. In general red cell infection cases, the host can destroy the infected cells with the body’s immune system by circulating the infected blood cells through the spleen and removing them using the lymphatic system. However,

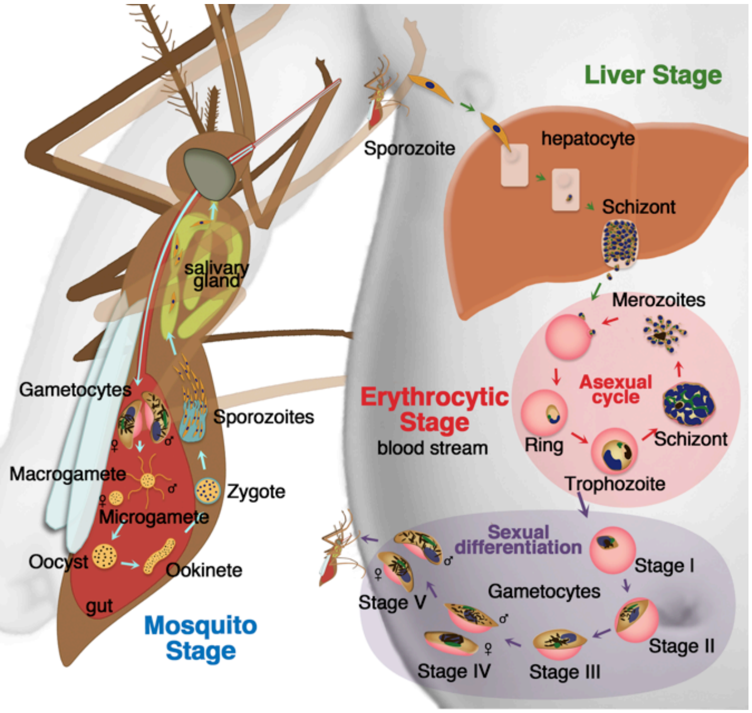


Figure I.1. Overview of the life cycle of *Plasmodium falciparum* in its human and mosquito hosts. (source: Biamonte MA, *et al*, 2013 [2])

in the case of malaria, the parasite releases adhesive proteins to the surface of the infected red cell and these help the cells stick to the microvasculature, therefore avoiding the splenic clearance cycle [9, 10]. Massive production of merozoites and the accumulation of infected red cells in the blood stream result in a blockage of the vascular system and can cause the host to suffer from headaches, diarrhea, and recurring fevers. In severe cases, it can lead to anemia, cerebral malaria, respiratory distress, and death of the host [11-13]. When growth conditions are not optimal for the parasite, some parasites switch from asexual reproduction to sexual differentiation and become gametocytes. Gametocytes circulate in the blood stream and can be picked up by mosquitos during a blood meal. Once inside the mosquito, the parasite undergoes the sexual reproduction cycle and eventually travels to the mosquito's salivary glands [14-17]. The parasite can then be re-introduced to a new human host during mosquito feeding and start the asexual replication cycle again. In general, the asexual forms of *P. falciparum* are responsible for disease and symptom development, while sexual forms are responsible for disease transmission.

The Genome of Plasmodium falciparum

The genome of *Plasmodium falciparum* was first published in 2002 using a whole chromosome shotgun sequencing approach [18]. It is the most eukaryotic AT-rich genome sequenced to date with an overall AT composition of 80.6% and 90% - 95% in the non-coding regions [18]. The size of the genome is approximately 23 megabases (Mb) consisting of 14 chromosomes. Currently, there are 5,772 predicted genes, in which 5,548 are predicted to be protein-encoding genes(version 34, PlasmoDB data base: http://plasmodb.org/plasmo/). The average gene density ranges from 1 gene per 4,300 to 1 gene per 4,800 base pairs, and nearly 50% of the genes contain introns. The average gene length, without introns, is approximately 2.3 kb and the average of exon size is about 950 basepairs [18]. The average intron length is about 178 basepairs and the average length of the intergenic region is about 1,700 basepairs [18]. The majority of the genes in the parasite was identified by gene-finding tools [18-20] and verified using experimental techniques including full-length cDNA, expressed sequence tag (EST), and mass spectrometry analysis [21-25]. However, due to low sequence similarity to proteins in other organisms, less than 60% of the genes are annotated. Moreover, no transposable elements, linker histone H1, RNA interference elements, or long tandem repeat ribosomal RNA (rRNA) arrays were identified in *P. falciparum's* genome. Instead, single 18S, 5.8S-26S rRNA units with different sequence variations were found distributed among different chromosomes [18]. Extensive size polymorphisms were also observed, especially at the telomeres regions. However, subtelomeric regions of the chromosome exhibit high degrees of conservation, which was hypothesized to be a result of promiscuous interchromosomal exchange[18]. Furthermore, many of these chromosome ends contain protein-encoding genes such as virulence, invasion, and gametocyte-specific genes. Long non-coding RNA telomere-associated repetitive elements transcripts (lncRNA-TAREs) have also been recently identified [26]. Beside the 14 main chromosomes, the parasite's genome also harbors a mitochondrial genome and an apicoplast genome, a plant-like plastid that is homologous to the chloroplasts [27-29]. The apicoplast genome is approximately 35 kilobases containing only 30 genes [27]. Though the exact role of the apicoplast is still unclear, it has been shown to be essential for parasite survival [30, 31], anabolic synthesis of fatty acids [32, 33], isoprenoid biosynthesis[34], and heme synthesis [35-37]. The sequenced genome provides a foundation for future studies on parasite's biology as well as for the search of new anti-malarial drug and vaccine targets, however, many functional elements of the genome remain undiscovered, especially with the AT content that challenged may widely used computational tools.

Transcriptional Regulation in Plasmodium falciparum

In a eukaryotic cell, genomic DNA is tightly wrapped around histone proteins and assembled as nucleosomes. These nucleosomes are then coiled and packaged together, resulting in a fiber also known as chromatin. Interactions between chromatin and protein complexes as well as the dynamics of nucleosome positioning and post-translational modifications (PTMs) of histone core proteins are of vital importance to the usage of DNA. The major step in gene transcription initiation is the recruitment of RNA polymerase II, along with other general transcription factors (TFIIs), to promoter regions to form the basal pre-initiation complex (PIC).

In *P. falciparum*, the nature and the contribution of mechanisms regulating gene expression are still poorly understood. Due to low sequence homology with other organisms, only a little more than half of the *plasmodium* genes have been successfully annotated, including RNA polymerase II and its subunits, 4 TATA-binding protein (TBP)-associated factors (TAFs), and 23 TFII components. Compare to organisms with similar genome sizes, a relatively low number of sequence-specific transcription factors and regulatory DNA elements such as enhancers and mediators were identified in *P. falciparum* [38-40]. Approximately 30 sequence-specific transcriptional factors have been found, and the majority belong to the apicomplexan-specific family (ApiAP2) [39, 41]. Although the specific role of some of these transcriptional factors are still unclear, many of the AP2 transcription factors are found throughout the parasite's genome and are believed to play an important role in stage-specific developments and gametogenesis. For example, it was well evident that AP2-G was important for gametocyte progression, AP2-Sp was responsible for sporozoite formation, and AP2-L was for liver-stage development [42-44]. With the lack of TAFs and sequence specific transcription factors, how transcription is initiated and regulated in this parasite remains uncertain and evolutionarily diverse mechanisms may be used as alternative solutions by the parasite.

Recent genome-wide nucleosome mapping studies in model organisms, such as yeast and human, have revealed consensus patterns in nucleosome organization, including lower nucleosome density at intergenic regions as compared to genic regions, a strong nucleosome-depleted region (NDR) near the promoter, and well-positioned nucleosomes (i.e., –1 and +1 nucleosomes) containing variant histone H2A.Z around the transcription start site (TSS) [45-48]. These findings suggest that specific positioning of nucleosomes, especially at promoter and transcription start or stop regions, largely contribute to transcriptional control by governing the access of components of the transcription initiation machinery to their binding sites. Furthermore, to ensure nucleosome dynamics and gene expression regulation, nucleosome components or the entire nucleosome may be repositioned, removed, or replaced through the action of ATP-dependent chromatin remodeling enzymes. In addition, post-translational modifications of histone proteins can have large effects on chromatin structure and gene activity. Besides precise positioning of nucleosomes, gene expression requires physical interaction between promoter regions and their distal regulatory elements, yet promoters and their regulatory elements are often linearly separated along the chromosome. To overcome this spatial constraint, chromatin loops are formed to bring together the regulatory elements and their promoters for gene activation.

Over the past decade, a series of molecular and genomic approaches have been developed (3C, 4C, 5C, Hi-C, etc.) to study the higher order organization of chromosomes by mapping interactions between genomic loci [49]. Hi-C analyses of mouse and human chromosome structures have revealed that eukaryotic genomes are organized into large blocks that show high levels of chromatin interactions within that region, but not with other loci in the genome. These regions, called topologically associated domains (TADs) [50, 51], are well defined by insulator proteins [52, 53] and are composed of many chromatin loops that have important functional roles in regulating gene expression. On a higher dimension, individual chromosomes organize and occupy distinct territories within the nucleus, and such organization is highly associated with gene density; gene-dense chromatin is usually enriched in the internal part of the nucleus, while gene-poor regions tend to locate toward the nuclear periphery [54-61].

In *Plasmodium falciparum*, recent studies have shown that the parasite's genome undergoes extensive remodeling events during different developmental stages of its cell cycle; the chromatin structure is relatively closed during the ring and schizont stages, but substantially opened during the trophozoite stage [62]. In addition, both western blot [63] and mass spectrometry analysis [64-66] has demonstrated that a large number of histones and nucleosomes[65] are depleted at the trophozoite stage. Through ultrastructual microscopic techniques, an increased number of nuclear pores as well as decomposition of chromatin near the nuclear envelope was observed at the trophozoite stage [27]. Collectively, these data suggest a de-condensed chromatin structure taking place at this cell cycle stage. In eukaryotes, a de-condensed chromatin or eurchromatin structures allows the transcriptional machinery to access of genomic DNA, thus allowing the genes to be transcribed at the given stage. In *P. falciparum*, a cascade of gene expression was observed across the parasite's asexual cell cycle when studying the mature RNA (mRNA) expression profiles and was poorly correlated with the binary activity of the chromatin organization events. This controversy is likely due to the fact that mRNA reflects not only transcriptional activity but also post-transcriptional activities, such as RNA degradation and mRNA stability. As chromatin organization events represent more transcriptional initiation regulation, mRNA expression may not be the best measurement for initial transcriptional activity. In chapter 3, I discuss a better understanding of the relationship between transcription activity and chromatin restructuring events using nascent RNA (newly synthesized RNA that are undergoing transcription). In summary, results from nascent RNA gene expression profiles have helped us to link the chromatin re-organization events to transcriptional activity and further highlighted the importance of epigenetic regulation and post-transcriptional regulation in this parasite.

It was well observed that, while the majority of the genome is maintained in an open chromatin environment during the trophozoite stage, a small subset of genes were controlled within highly condensed heterochromatin clusters located within subtelomeric regions and few internal loci. These regions are usually marked by heterochromatin protein 1 (PfHP1) and histone repression mark, H3K9me3 [62, 67-74]. An example of genes controlled by heterochromatin structure is the *P. falciparum* virulence genes (*var* genes). These *var* genes encode erythrocyte membrane proteins 1(PfEMP1s) that are expressed at the surface of parasite-infected red blood cells (RBCs) and act as both an atigen and adhesion protein. Success expression of PfEMP1s allows the infected RBCs attach to the blood vessel and escape from the host's immune system. Sixty *var* genes have been identified in the parasite's genome; however only one will be expressed and others remain suppressed. *Var* gene expression and regulation is critical for parasite survival. Previous studies showed that silent *var* genes are clustered within heterochromatin regions at the nuclear periphery marked by PfHP1 and H3K9me3 [62, 67-73] and an absent of PfHP1 protein resulted a disruption of the monoallelic var gene expression, resulted an arrestment in parasite growth [75, 76]. In addition, emerging studies showed that many chromatin-associated proteins such as histone deacetylases (PfSIR2A, PfSIR2B, PfHDA2) and histone lysine methltransferase (PfSET2) are also linked to *var* gene regulation. Manipulation of these proteins will either interrupt the silenced *var* gene cluster(s) or the loss of monoallelic *var* gene expression pattern.

Similar to other organisms, histone post-modification marks (PTMs) and nucleosome positioning also have very important roles in regulation of gene expression activity in P. *falciparum*. More than 232 different histone PTMs have been experimentally identified during the intraerythrocytic stages. Although some of the function of these PTMs have yet to be discovered, it is becoming clear that the epigenome in *P. falciparum* is in a highly dynamic state and is more diverged to other organisms than previously expected. Similar to other eukaryotes, main histones such as H2A, H2B, H3, and H4 along with histone variant H2A.z, H2Bv, H3.3, and CenH3 are found in *P. falciparum* by mass spectroscopy analysis [77, 78]. However, linker Histone H1 is absent [79] and H4K8 and H4K12 are the more favored acetylation sites in *P. falciparum* [77, 80]. H4k5ac, H4K8ac and H4K12ac are found more evently distributed throughout the IDC[79], while H3K4me3 and H3K9ac was found associate with the 5' regions of activated genes in schizonts but not in ring stage and H4K16ac and H4K9ac peaked at trophozoite and late schizont stages [71, 77]. In addition, various high through-put sequencing experiments have revealed that a majority of the parasite genome is likely to be covered by activating histone marks (H3K9ac and H3K4me3) as compared to silencing marks (H3K9me3 and H3K36me3) and further validates the transcriptional permissive euchroamtin state of the parasite. In addition, histone PTMs are also strongly associated with *var* gene regulation; H3K9me3 are found more abundantly around silenced *var* genes, while H3K9ac are found a the 5' flanking regions of activated *var* genes [74].

The nucleosome landscape of *P. falciparum* is slightly different to the nucleosome landscape of other eukaryotes. In the parasite's genome, the strong positioned nucleosomes are found at the start and stop of coding region instead of around the TSS. Moreover, because of the AT-richness of the genome that is relatively less flexible, thus uneasy to wrap around to the nucleosome core, the parasite evolved its H2A.Z and H2B.Z variations, which are better for AT-rich DNA binding and are found throughout the intergenic regions of the genome instead of being located at the +1 nucleosome in other eukaryotes [81, 82]. Although some controversy occurs, when the nucleosome occupancy profiles normalized by the number of parasite or nuclei, lower nucleosome occupancy is found during trophozite stage. Such observation has also been validated and confirmed by various experimental approaches including western blot, mass spectrometry, MNase-Seq, FAIRE-seq, and ChIP-seq [63-66, 83]. All these data further supported the genome re-organization events demonstrated by the 3D genome architecture model and purpose the idea that once chromatin opens up, nucleosomes are evicted giving raise to the massive transcriptional burst at the trophozoite stage, then chromatin are re-packed at a later stage limiting transcriptional activity and allowing only a small number of stage-specific genes to be expressed. It is hypothesized that this small number of stage-specific genes are more likely to be regulated by more traditional transcription regulators such as the sequence-specific transcription factor and histone PTMs.

In eukaryotic organisms, emerging evidences show that non-coding RNAs (ncRNAs) are also involved in transcriptional regulation and chromatin organization events. A well-known example is the lncRNA known as Xist. Xist mediates X-chromosome inactivation during zygotic development. Deposition of Xist on the X-chromosome recruits histone-modifying enzymes that place repressive histone marks such as H3K9 and H3K27 methyaltion for gene silencing and formation of heterochromatin structure. In addition, ncRNA have also been shown as mediators bridging together enhancers and promoters [84] or acted like enhancer element that activates transcription activity of neighboring coding genes [85]. In *Plasmodium*, hundreds of ncRNAs have been identified; however, only a small proportion of them have been annotated. Earlier studies show that some of these ncRNAs, generated from bidirectional promoter activities are localized in in the centormeric regions and are believed to be associated with maintenance, and organization of the centromeric chromatin [86]. Beside centromere regions, some of identified ncRNAs are found to be located at the subtelomeric regions and are involved in *var* gene regulation. Using DNA tiling microarray, Broadbent et al (2011) found 60 putative long non-coding RNAs, in which 22 are characterized as telomere-associated repetitive element transcripts (lncRNA-TARE) [26]. These lncRNA-TAREs are exclusively localized adjacent to upsB-type *var* genes and contain many SPE2 binding sites that are only otherwise found in *var* gene promter regions. SPE2 is a cis-acting element critical for *var* gene silencing and are known to be bound by transcription factor PfSIP2 [87]. It was hypothesis that these lncRNA-TAREs may be involving in PfSIP2 recruiting, thus associated with *var* gene regulation. Beside lncRNA-TAREs, other lncRNAs and 43 C/D and H/ACA-box subclasses of small nucelaolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs) have also been described and showed concordant patterns of expression across cell cycle for housekeeping gene and *var* gene regulation [88]. Additionally, many anti-sense transcripts initiated from *var* introns were observed. These antisense transcripts are believed to interfering with the monoallelic *var* gene expression by directly incorporate into chromatin [77, 89, 90]. Another type of antisense transcript is the natural antisense transcript (NAT). High levels of NATs have been found toward the 3' end of the open reading frames for more than 24% of the genes [91]. Though differentially expressed throughout the parasite's cell cycle, majority of NATs in *P. falciparum* is not strongly associated with their corresponding mRNA in term of gene expression levels suggesting these NATs are not likely to be involved in transcriptional interference or RNA-RNA formation process but maybe involved in epigenetic regulation [90-92]. In general, majority of the ncRNAs has a lack of sequence conservation, typically expressed at a low level, could be co-expressed with nearby coding genes, and are notably up or down regulated to module gene expression during the parasite's cell cycle [93, 94]. Compare to ncRNA biology studied in other eukaryotes, very little is known about ncRNA in *P. falciparum,* therefore, additional efforts are needed to further characterize and annotated for ncRNAs in *P. falciparum.*

Conclusion

Although malaria is a treatable disease, due to the quick adaptable nature of the parasite and its vector, malaria remains one of the most lethal infectious diseases in the world. Currently, there is no approved vaccine to prevent malaria-infection and because of the symptoms of malaria are so similar to the symptoms of the flu, many malaria infected patients failed to be properly diagnosis an early stage of the infection. In addition, the fast development of resistant strains to anti-malarial drugs became one of the biggest challenges for disease management. Though artemisinin-based combination therapy (ACT) is given as treatment for resistant infections, it could be only the matter of time for parasites to adapt to this therapy. Therefore, finding new drug targets as well as new anti-malarial tools are required. In addition, although a great global effort was made for parasite eradication and disease control, many missions were unable to complete due to the lack of funding and resources. For example, Global Technical Strategy for Malaria 2016–2030 was developed aiming to eliminate malaria and to reduce the global malaria mortality rates by 40% by 2020. However, due to the shortfall of funding and the incompletion of health systems, millions of people had limited access to life-saving tools and treatments. As a result, fewer than half of the targeted malaria-affected countries are being on track for 2020 milestone achievement [1]. To further combat the disease, long-term financial support, more completed health care systems, and an extensive research effort searching for parasite-specific drug target are still needed.

Among the five human-infection malaria parasites, *P. falciparum* is the one responsible for majority of deaths. This parasite has a complex life cycle and an extremely AT-rich genome that shared a low homology with other organisms. This low level of homology challenged many well-developed computational tools for gene identification and annotation. As compared to other eukaryotes with similar genome sizes, only a fraction of the transcription factors and regulators have been identified in P. falciparum. Due to this paucity of transcription factors, it is still unclear how this parasite manages its gene expression process to coordinate the different morphological changes in different cell cycles. Increasing evidences have shown that epigenetic regulations, such as changes of histone modification, nucleosome landscape, chromatin structure, and the organization of the three-dimensional genome architectures, may play critical roles in this fine-tuned gene expression system. The depletions of histones and nucleosome, as well as the three-dimensional genome structure models demonstrated that chromatin is opened at the trophozoite stage, and packed at the ring and schizont stage. In later chapter, I showed that this binary chromatin activity is well correlated with nascent RNA expression. Collectively, these data suggested that chromatin organization provides a basal control for genome-wide transcription activity and various epigenetic marks are vital for local heterochromatin maintenance, stage-specific gene expression, and regulation of certain antigenic variant genes. Beside epigenetic elements, long none coding RNAs have also been showed to play important roles in transcriptional-associated events including var gene regulation and chromatin structure maintenance. In the final chapter, I developed an experimental approach to better identify and understand the function of these lncRNAs in the human malaria parasite. Taken together, these advanced malarial studies bring up the possibility that transcription regulation in *P. falciparum* may largely differ from the ones in other eukaryotes; Instead of using a variety of transcription factors, this parasite relies heavily on epigenetic regulatory elements for transcription initiation, and the cascade of mature RNA expression pattern, observed by traditional RNA-seq, may be a result of both transcriptional and post-transcriptional regulation events. In the next chapters, I will further explore the parasite’s transcriptome and better characterize the mechanisms that are involved in the parasite’s gene expression system, especially at the transcriptional level.

Reference

1. WHO: **World Malaria Report. 2016.** 2016.

2. Biamonte MA, Wanner J, Le Roch KG: **Recent advances in malaria drug discovery.** *Bioorg Med Chem Lett* 2013, **23:**2829-2843.

3. O'Brien C, Henrich PP, Passi N, Fidock DA: **Recent clinical and molecular insights into emerging artemisinin resistance in Plasmodium falciparum.** *Curr Opin Infect Dis* 2011, **24:**570-577.

4. Maltha J, Guiraud I, Lompo P, Kabore B, Gillet P, Van Geet C, Tinto H, Jacobs J: **Accuracy of PfHRP2 versus Pf-pLDH antigen detection by malaria rapid diagnostic tests in hospitalized children in a seasonal hyperendemic malaria transmission area in Burkina Faso.** *Malar J* 2014, **13:**20.

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. Gerald N, Mahajan B, Kumar S: **Mitosis in the human malaria parasite Plasmodium falciparum.** *Eukaryot Cell* 2011, **10:**474-482.

8. Wright GJ, Rayner JC: **Plasmodium falciparum erythrocyte invasion: combining function with immune evasion.** *PLoS Pathog* 2014, **10:**e1003943.

9. Flick K, Chen Q: **var genes, PfEMP1 and the human host.** *Mol Biochem Parasitol* 2004, **134:**3-9.

10. Nery S, Deans AM, Mosobo M, Marsh K, Rowe JA, Conway DJ: **Expression of Plasmodium falciparum genes involved in erythrocyte invasion varies among isolates cultured directly from patients.** *Mol Biochem Parasitol* 2006, **149:**208-215.

11. Miller LH, Baruch DI, Marsh K, Doumbo OK: **The pathogenic basis of malaria.** *Nature* 2002, **415:**673-679.

12. Miller LH, Ackerman HC, Su XZ, Wellems TE: **Malaria biology and disease pathogenesis: insights for new treatments.** *Nat Med* 2013, **19:**156-167.

13. Marsh K, Forster D, Waruiru C, Mwangi I, Winstanley M, Marsh V, Newton C, Winstanley P, Warn P, Peshu N, et al.: **Indicators of life-threatening malaria in African children.** *N Engl J Med* 1995, **332:**1399-1404.

14. Liu Z, Miao J, Cui L: **Gametocytogenesis in malaria parasite: commitment, development and regulation.** *Future Microbiol* 2011, **6:**1351-1369.

15. Baker DA: **Malaria gametocytogenesis.** *Mol Biochem Parasitol* 2010, **172:**57-65.

16. Silvestrini F, Lasonder E, Olivieri A, Camarda G, van Schaijk B, Sanchez M, Younis Younis S, Sauerwein R, Alano P: **Protein export marks the early phase of gametocytogenesis of the human malaria parasite Plasmodium falciparum.** *Mol Cell Proteomics* 2010, **9:**1437-1448.

17. Kaushal DC, Carter R, Miller LH, Krishna G: **Gametocytogenesis by malaria parasites in continuous culture.** *Nature* 1980, **286:**490-492.

18. 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.

19. Hyman RW, Fung E, Conway A, Kurdi O, Mao J, Miranda M, Nakao B, Rowley D, Tamaki T, Wang F, Davis RW: **Sequence of Plasmodium falciparum chromosome 12.** *Nature* 2002, **419:**534-537.

20. Gardner MJ, Shallom SJ, Carlton JM, Salzberg SL, Nene V, Shoaibi A, Ciecko A, Lynn J, Rizzo M, Weaver B, et al: **Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14.** *Nature* 2002, **419:**531-534.

21. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ: **cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome.** *BMC Genomics* 2007, **8:**255.

22. 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.

23. 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.

24. 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.

25. Watanabe J, Sasaki M, Suzuki Y, Sugano S: **FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, Plasmodium falciparum.** *Nucleic Acids Res* 2001, **29:**70-71.

26. 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.

27. Wilson RJ, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW, Williamson DH: **Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum.** *J Mol Biol* 1996, **261:**155-172.

28. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS: **A plastid of probable green algal origin in Apicomplexan parasites.** *Science* 1997, **275:**1485-1489.

29. McFadden GI, Reith ME, Munholland J, Lang-Unnasch N: **Plastid in human parasites.** *Nature* 1996, **381:**482.

30. Fichera ME, Roos DS: **A plastid organelle as a drug target in apicomplexan parasites.** *Nature* 1997, **390:**407-409.

31. He CY, Striepen B, Pletcher CH, Murray JM, Roos DS: **Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of Toxoplasma gondii.** *J Biol Chem* 2001, **276:**28436-28442.

32. Surolia N, Surolia A: **Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum.** *Nat Med* 2001, **7:**167-173.

33. Waller RF, Keeling PJ, Donald RG, Striepen B, Handman E, Lang-Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI: **Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 1998, **95:**12352-12357.

34. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M, Turbachova I, Eberl M, Zeidler J, Lichtenthaler HK, et al: **Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.** *Science* 1999, **285:**1573-1576.

35. Sato S, Wilson RJ: **The genome of Plasmodium falciparum encodes an active delta-aminolevulinic acid dehydratase.** *Curr Genet* 2002, **40:**391-398.

36. van Dooren GG, Su V, D'Ombrain MC, McFadden GI: **Processing of an apicoplast leader sequence in Plasmodium falciparum and the identification of a putative leader cleavage enzyme.** *J Biol Chem* 2002, **277:**23612-23619.

37. Foth BJ, McFadden GI: **The apicoplast: a plastid in Plasmodium falciparum and other Apicomplexan parasites.** *Int Rev Cytol* 2003, **224:**57-110.

38. Batugedara G, Lu XM, Bunnik EM, Le Roch KG: **The Role of Chromatin Structure in Gene Regulation of the Human Malaria Parasite.** *Trends Parasitol* 2017, **33:**364-377.

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

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. Jiang C, Pugh BF: **A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome.** *Genome Biol* 2009, **10:**R109.

46. Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ: **Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes.** *Genome Res* 2009, **19:**967-977.

47. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD: **Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin.** *Cell* 2005, **123:**233-248.

48. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, Robert F, Gaudreau L: **Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning.** *PLoS Biol* 2005, **3:**e384.

49. de Wit E, de Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes Dev* 2012, **26:**11-24.

50. 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.

51. Nora EP, Dekker J, Heard E: **Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods?** *Bioessays* 2013, **35:**818-828.

52. Hou C, Li L, Qin ZS, Corces VG: **Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains.** *Mol Cell* 2012, **48:**471-484.

53. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, Cubenas-Potts C, Hu M, Lei EP, Bosco G, et al: **Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing.** *Mol Cell* 2015, **58:**216-231.

54. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Muller S, Eils R, Cremer C, Speicher MR, Cremer T: **Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes.** *PLoS Biol* 2005, **3:**e157.

55. Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA: **The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells.** *Hum Mol Genet* 2001, **10:**211-219.

56. Cremer T, Cremer M, Dietzel S, Muller S, Solovei I, Fakan S: **Chromosome territories--a functional nuclear landscape.** *Curr Opin Cell Biol* 2006, **18:**307-316.

57. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA: **Differences in the localization and morphology of chromosomes in the human nucleus.** *J Cell Biol* 1999, **145:**1119-1131.

58. Federico C, Scavo C, Cantarella CD, Motta S, Saccone S, Bernardi G: **Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates.** *Chromosoma* 2006, **115:**123-128.

59. Foster HA, Bridger JM: **The genome and the nucleus: a marriage made by evolution. Genome organisation and nuclear architecture.** *Chromosoma* 2005, **114:**212-229.

60. Zink D, Cremer T, Saffrich R, Fischer R, Trendelenburg MF, Ansorge W, Stelzer EH: **Structure and dynamics of human interphase chromosome territories in vivo.** *Hum Genet* 1998, **102:**241-251.

61. Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T: **Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions.** *Nat Rev Genet* 2007, **8:**104-115.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. Gupta AP, Bozdech Z: **Epigenetic landscapes underlining global patterns of gene expression in the human malaria parasite, Plasmodium falciparum.** *Int J Parasitol* 2017, **47:**399-407.

78. 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.

79. Miao J, Fan Q, Cui L, Li J, Li J, Cui L: **The malaria parasite Plasmodium falciparum histones: organization, expression, and acetylation.** *Gene* 2006, **369:**53-65.

80. 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.

81. 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.

82. 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.

83. 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.

84. Malik S, Roeder RG: **The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation.** *Nat Rev Genet* 2010, **11:**761-772.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. Vembar SS, Scherf A, Siegel TN: **Noncoding RNAs as emerging regulators of Plasmodium falciparum virulence gene expression.** *Curr Opin Microbiol* 2014, **20:**153-161.

91. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, Martins RM, Sismeiro O, Coppee JY, Scherf A: **Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum.** *BMC Genomics* 2014, **15:**150.

92. 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.

93. 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.

94. 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.

Chapter 1: Analysis of Nucleosome Positioning Landscapes Enables Gene Discovery in the Human Malaria Parasite *Plasmodium falciparum*

Xueqing Lu[1]*, Evelien M. Bunnik[1]*, Neeti Pokhriyal[2], Sara Nasseri[2], Stefano Lonardi[2], Karine G. Le Roch[1]

[1]Department of Cell Biology and Neuroscience, Institute for Integrative Genome Biology, Center for Disease Vector Research, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.

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

*These authors contributed equally to this work.

This chapter originally appeared as Lu et al. *BMC Genomics*, 2015 [1].

Abstract

Plasmodium falciparum, the deadliest malaria-causing parasite, has an extremely AT-rich (80.7%) genome. Because of high AT-content, sequence-based annotation of genes and functional elements remains challenging. In order to better understand the regulatory network controlling gene expression in the parasite, a more complete genome annotation as well as analysis tools adapted for AT-rich genomes are needed. Recent studies on genome-wide nucleosome positioning in eukaryotes have shown that nucleosome landscapes exhibit regular characteristic patterns at the 5'- and 3'-end of protein and non-protein coding genes. In addition, nucleosome depleted regions can be found near transcription start sites. These unique nucleosome landscape patterns may be exploited for the identification of novel genes. In this chapter, we proposed a computational approach to discover novel putative genes based exclusively on nucleosome positioning data in the AT-rich genome of *P. falciparum*. Using binary classifiers trained on nucleosome landscapes at the gene boundaries from two independent nucleosome positioning data sets, we were able to detect a total of 231 regions containing putative genes in the genome of *Plasmodium falciparum*, of which 67 highly confident genes were found in both data sets. Eighty-eight of these 231 newly predicted genes exhibited transcription signal in RNA-Seq data, indicative of active transcription. In addition, 20 out of 21 selected gene candidates were further validated by RT-PCR, and 28 out of the 231 genes showed significant matches using BLASTN against an expressed sequence tag (EST) database. Furthermore, 108 (47%) out of the 231 putative novel genes overlapped with previously identified but unannotated long non-coding RNAs. Collectively, these results provide experimental validation for 163 predicted genes (70.6%). Finally, 73 out of 231 genes were found to be potentially translated based on their signal in polysome-associated RNA-Seq representing transcripts that are actively being translated. These results clearly indicate that nucleosome positioning data contains sufficient information for

novel gene discovery. As distinct nucleosome landscapes around genes are found in many other eukaryotic organisms, this methodology could be used to characterize the transcriptome of any organism, especially when coupled with other DNA-based gene finding and experimental methods (e.g., RNA-Seq).

Introduction

As one of the world's most deadly infectious diseases, malaria is responsible for about 584,000 deaths annually, the vast majority of which are children under the age of five [2]. Currently, no approved vaccine is available for disease prevention, and the rapid development of parasite resistance to current antimalarial drugs is a major challenge for the control of malaria. Out of five human malaria parasite species, *Plasmodium falciparum* causes 90% of all malarial deaths [2]. *P. falciparum* has a complex life cycle involving multiple stages in two host organisms, humans and mosquitoes. This multi-stage life cycle is tightly regulated, presumably by strict control of stage-specific gene expression. However, the mechanisms regulating gene expression in *P. falciparum* are still poorly understood. In particular, relatively few specific transcription factors and regulatory elements have been identified [3, 4]. In addition, the annotation of protein coding and non-protein coding genes is incomplete. To facilitate our understanding of the parasite's life cycle and its regulatory mechanisms and thus assist the development of antimalarial drugs, a more accurately annotated genome is needed.

The draft of the annotated genome of *P. falciparum* was first published in 2002 [5]. *P. falciparum* has a relatively compact genome consisting of fourteen chromosomes with a total length of approximately 23 Mb [5]. The *P. falciparum* genome is the most AT-rich eukaryotic genome sequenced to date, with an overall AT-composition of 80.7%, rising to 90-95% in introns and intergenic regions [6]. Currently, 5,777 predicted protein coding genes have been reported (plasmoDB v26) and ~50% of these genes share little or no sequence similarity to genes or the encoded proteins in other organisms [5-7]. The average gene length in *P. falciparum* is 2.3 kb and the average length of intergenic regions is ~1.7 kb [8]. Both computational and evidence-based gene-finding methods have been applied to obtain gene annotations. Genome annotations of the reference strain 3D7 were performed *in silico* using software tools including Artemis, Genefinder, GlimmerM, and phat [9, 10]. Most of the predicted genes have been verified using various experimental techniques including full-length cDNA, expressed sequence tag (EST), and mass spectrometry analysis, among others [8, 11-14]. More comprehensive annotations of the parasite's gene structure and other functional elements have been possible since the advent of second-generation sequencing technology [7, 14-19].

Despite significant advances in the analysis of the parasite's genome, genome annotation in *P. falciparum* is still a work in progress. The AT-richness and the relative lack of sequence homology to other organisms hamper the application of sequence-based gene prediction tools and complicate the identification of functional DNA elements, such as protein-binding sites, promoters, or TATA-like boxes. In addition, as mentioned earlier, the parasite has a complicated multi-stage life cycle involving multiple hosts. Due to technical challenges, it is nearly impossible to capture the transcriptome at all different life cycle stages. We are therefore still in need of an improved genome annotation, as well as analysis tools capable of handling the parasite's AT-rich genome that will help us to better understand the regulatory mechanisms controlling gene expression in the parasite.

In mammalian genomes, a large number of non-coding transcripts have been identified based on chromatin signatures H3K4me3 and H3K36me3 [20]. This finding suggests that elements defining and bracing chromatin architecture may be used to assist the identification of undiscovered genes. In this study, we present a machine learning approach to predict genes in *P. falciparum* that is completely independent from the primary DNA sequence, but instead exploits the underlying chromatin structure and nucleosome landscape. The fundamental unit of chromatin is a nucleosome, a stretch of ~147 bp of DNA wrapped around a core of eight histone proteins. Nucleosomes are distributed non-uniformly around genes, and this distinct nucleosome landscape is known to play an important role in gene regulation. In particular, the core promoter is usually characterized by a nucleosome-depleted region that allows the binding of transcription factors and facilitates the assembly of the transcription preinitiation complex [21, 22]. Previous studies in our lab have highlighted several common and unique eukaryotic features of the *P. falciparum* nucleosome landscape. Similar to other eukaryotes, *Plasmodium*'s promoters and transcription start sites are relatively nucleosome depleted, and nucleosome occupancy is higher inside genes as compared to intergenic regions [23-25]. However, in contrast to the strongly positioned +1 nucleosome directly downstream of the transcription start site in other eukaryotes [21, 26-29], the most strongly positioned nucleosomes in *P. falciparum* are located at the start and end of the open reading frame [23, 24]. Based on these nucleosome landscape characteristics, we propose a novel method for gene detection using classifiers trained on nucleosome profiles of annotated genes. Other experimental methods used for gene detection, such as RNA-Seq or expressed sequence tags (EST), can be noisy, potentially resulting in false predictions. Therefore, our methodology may serve as a complementary approach for refining genome annotations, especially coupled with sequence-based gene predictions and other experimental methods.

Results

Building a classifier on nucleosome positioning profiles

In a previous study, our lab has used second-generation sequencing to generate high-resolution nucleosome positioning profiles for three different stages of *P. falciparum*'s asexual cycle [23]. This data set revealed a distinct nucleosome landscape around genes, with higher nucleosome occupancy inside genes, lower nucleosome coverage in intergenic regions, and strongly positioned nucleosomes at the gene boundaries (Figure 1.1A). In addition, as observed in other eukaryotic genomes [27, 28, 30], a nucleosome-depleted region was found immediately upstream of the transcription start site, which likely harbors the binding sites of transcription factors [23]. These observations were replicated using an independently generated *P. falciparum* nucleosome occupancy data set [23, 31] (Figure 1.1B). In this paper, we exploited this nucleosome landscape around genes to identify regions in the genome containing putative novel genes. To gain additional power for gene detection, we decided to predict the presence of novel genes using the two independently published nucleosome positioning data sets [23, 31]. For each data set, we summed the sequence coverage profile at each of the parasite's asexual stages into a single genome-wide nucleosome positioning data set. This resulted in a total of two combined profiles, namely i) profile B1 from Bunnik *et al*. [23] consisting of three asexual cycle time points, and ii) profile B2 from Bartfai *et al*. [31] consisting of four asexual cycle time points.

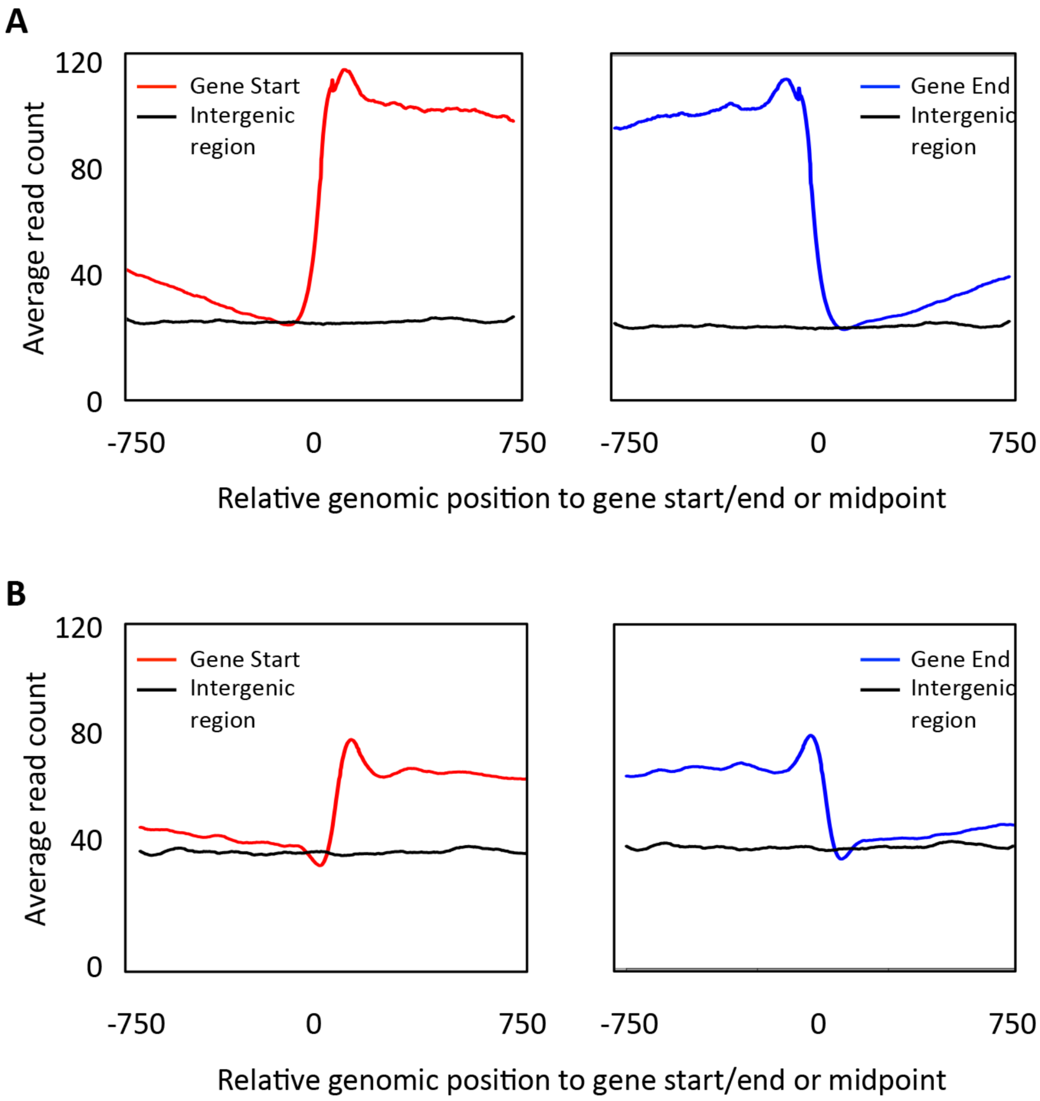


Figure 1.1. Nucleosome occupancy patterns in *P. falciparum*. A. Average sequence coverage profiles around the start (left panel) and the end (right panel) of genes (colored line), and in intergenic regions (black line) in the nucleosome occupancy data set from Bunnik *et al*. [23] (data set B1). **B.** Similar analysis for the nucleosome occupancy data set from Bartfai *et al*. [31] (data set B2). In all windows, the genomic position indicated on the x-axis is relative to the location of the gene start/end, or to the midpoint of intergenic windows.

From each of the combined nucleosome profiles, we extracted windows that either contained a gene start within its defined central region (positive class windows) or were completely derived from intergenic regions (negative class windows) (see Material and Methods; Supplemental Figure 1.1). We then used these positive and negative class windows to train a binary classifier (i.e., a support vector machine with RBF kernel) to recognize the general nucleosome occupancy pattern at gene start codons. The parameters of the classifier were optimized using cross validation (see Material and Methods; Supplemental Figure 1.2). In parallel, an independent classifier was trained on the nucleosome landscape at gene stop codons. Since we observed that nucleosome landscapes on the forward and reverse strands have slightly different characteristics, we independently optimized both strand-specific and non-strand-specific classifiers. All classifiers performed in very similar ways and optimized classifiers from both data sets gave total recall rates between 91 and 95% (Supplemental File 1.1).

These classifiers were then used on the nucleosome landscape of the whole *P. falciparum* genome to detect putative novel gene starts and ends. A sliding-window method was used to scan intergenic regions for the presence of predicted gene starts or gene ends. The classifier produced a confidence score between 0 and 1 for each prediction. A valid gene candidate was defined as a locus with a gene start and a gene end predicted using the same strand classifier with confidence scores above 0.7 and located within the same intergenic region (Figure 1.2A). No additional constraint on the distance between a predicted gene start and gene end was required, given the relatively short length of intergenic regions in the genome of *P. falciparum* (1.7 kb on average). A total of 298 final candidate regions with an average segment length of 1 kb were manually identified, of which 97 were detected using the B1 nucleosome positioning profile, and 201 were identified using the B2 nucleosome positioning profile (Supplemental File 1.1). Of the 298 candidate regions, 67 genes were identified in both B1 and B2 data sets with an average overlap in predicted gene region of 81%. This intersect between genes predicted by both data sets was highly statistically significant (*P* < 7.422e-66, calculated based on an hypergeometric distribution analysis [32]). Since overlapping regions may represent alternative splicing variants of the same gene, we merged overlapping regions using mergeBed (BEDtools [33]), resulting in a total of 231 unique regions harboring potential novel genes. All putative novel genes are uniformly distributed over the 14 chromosomes of the *P. falciparum* genome (Figure 1.2B).

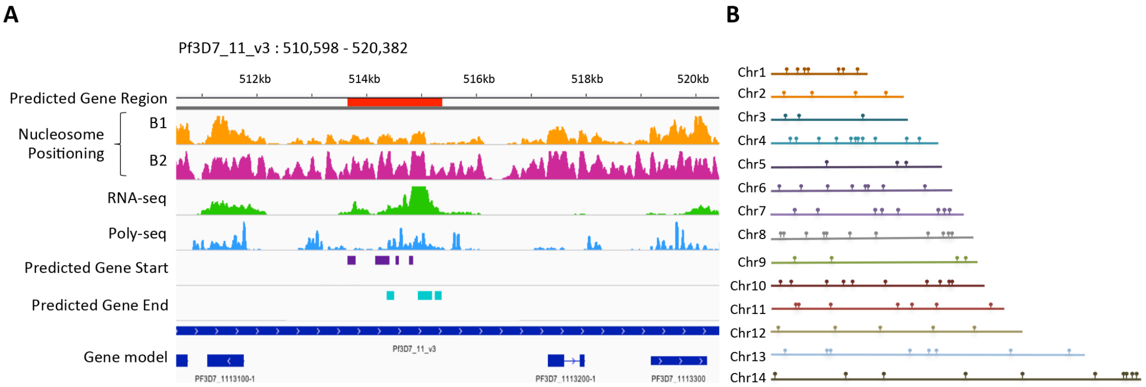


Figure 1.2. Characterization of regions containing putative novel genes. A. Genome browser view of an intergenic region containing a predicted gene region (Pf3D7_11_v3: 513,659 – 515,381, shown in red). Predicted gene starts and gene ends are indicated in purple and teal, respectively. This putative novel gene shows sequence coverage in both steady-state RNA-seq (green) and polysomal RNA-seq (blue) data sets. **B.** Random distribution of 97 regions predicted using classifiers trained on data set B1 across the 14 chromosomes of the *P. falciparum* genome.

Among these 231 predicted genes, 88 showed a signal (defined as an average of two or more reads per base) in a previously obtained RNA-Seq data set [19], which we considered strong evidence for the presence of a transcribed gene in this region (Supplemental File 1.1). On average, predicted gene regions are covered by eight reads per base, which is significantly higher than that the RNA-Seq coverage in intergenic regions of the same length (Table 1.1, *P* = 0.015, bootstrap Welch t-test with n = 100,000). In addition, 108 out of these 231 (47%) uniquely predicted regions overlap with previously identified long non-coding RNAs (lncRNAs), defined as non-coding transcripts larger than 200 bp that are not antisense or circular RNA [18, 34-37]. To further confirm transcriptional activity in the predicted gene regions, we designed a set of primers targeting 21 selected candidate regions. We were able to amplify 20 of the 21 targeted fragments from cDNA (Figure 1.3 and Supplemental File 1.2), suggesting that the majority of candidate genes may indeed be transcribed.

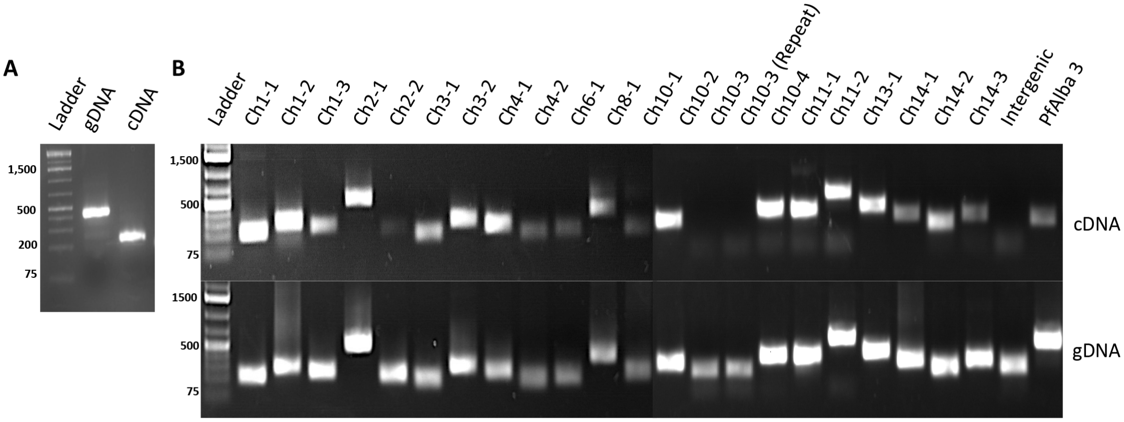


Figure 1.3. RT-PCR validation of 21 predicted novel genes. A. Amplification of a fragment of PfAlba3 (PF3D7_1006200) using genomic DNA (middle lane) or cDNA prepared from DNase-treated total RNA (right lane) as a template. Primers were designed on both sides of intron 1, yielding a 429 bp PCR product from genomic DNA and a 164 bp PCR product from cDNA. The presence of a single 164 bp PCR product amplified from cDNA confirms the absence of gDNA contamination. **B.** Out of our 231 novel candidate genes, we chose 21 regions for validation using reverse transcription polymerase chain reaction (RT-PCR). The top panel shows amplification products using DNase-treated cDNA as a template, while the bottom panel shows the control reactions using genomic DNA as a template. Of the 21 gene tested, we were able to amplify 20 of the predicted regions. As a control, we were unable to amplify a fragment of intergenic region that was not predicted to contain any genes (marked as "intergenic").

Characteristics of candidate novel P. falciparum genes

To further investigate the putative genes identified in this study, we compared several characteristics of the predicted regions with known coding and non-coding regions in the *P.*

falciparum genome. The average length of the predicted gene is 1,004 bp, which is similar to the average length of exons and lncRNAs in *P. falciparum*. The average GC-percentage for the predicted genes (16%) is lower than known coding genes (23%), but close to previously identified lncRNA regions (15%) and slightly higher than intergenic regions (13%) (Table 1.1 and Supplemental Figure 1.3A). Similarly, the average nucleosome occupancy in predicted gene regions ranged between that of known protein-coding genes and that of lncRNA genes (Supplemental Figure 1.3B-C). The nucleosome profiles at the predicted gene starts and gene ends recapitulate the nucleosome features observed in annotated genes, albeit at lower average nucleosome levels (Figure 1.4). Furthermore, the predicted novel genes have similar expression levels in steady-state mRNA-Seq [19] and polysome-associated mRNA-Seq [19] data sets as compared to lncRNA genes (Supplemental Figure 1.3D-E). Lastly, we examined the patterns of histone variant H2A.Z and histone marks H3K4me3 and H3K36me3. In *P. falciparum,* H2A.Z is almost exclusively found in nucleosomes located in intergenic regions [31], while H3K4me3 is enriched at the gene boundaries and H3K36me3 is enriched inside gene bodies [38] (Supplement Figure 1.4). We found that the average H2A.Z occupancy is higher in predicted genes than in annotated genes, and very similar to intergenic and previously identified lncRNA genes (Table 1.1 and Supplemental Figure 1.3F). In line with H3K36me3 being more abundant in coding regions as compared to noncoding regions in *P. falciparum*, we observed that the abundance of H3K36me3 in our predicted genes is in between that of coding and non-coding regions. In addition, H3K36me3 levels in our predicted genes are higher than in previously identified lncRNAs (Table 1.1 and Supplemental Figure 1.3G). Similar to the H3K36me3, H3K4me3 occupancy in our predicted genes is also found to be higher than in previous identified lncRNA and ranged between coding and non-coding regions (Table 1.1 and Supplemental Figure 1.3H).

Table 1.1. Characteristics of the 231 putative novel genes in comparison with annotated *P. falciparum* genes.

	Average RNA-seq coverage	**Average Poly-seq coverage**	**Average GC%**	**Average length (bp)**	**Avgerage Nuc coverage (B1)**	**Average Nuc coverage (B2)**	**Average H2A.z coverage**	**Average H3K36me3 coverage**	**Average H3K4me3 coverage**	***n***
Exon	58	30	27	949	35	77	20	41	84	14,795
Gene	75	34	23	2,494	37	69	17	50	83	5,680
Intergenic*	3	7	13	1,000	8	37	31	10	22	1,565
Published lncRNA [17, 33-36]	10	7	15	1,114	12	47	35	16	37	986
Predicted genes (B1)	13	4	18	934	22	62	29	25	63	97
Predicted genes (B2)	9	5	16	1,010	17	62	35	18	49	201
All predicted genes (B1+B2)	8	4	16	1,004	17	62	34	19	50	**231**

*Intergenic regions were defined as the middle 1 kb of all non-coding regions longer than 1,500 bp that do not overlap with annotated genes, predicted genes, or previously identified lncRNA [18, 34-37].

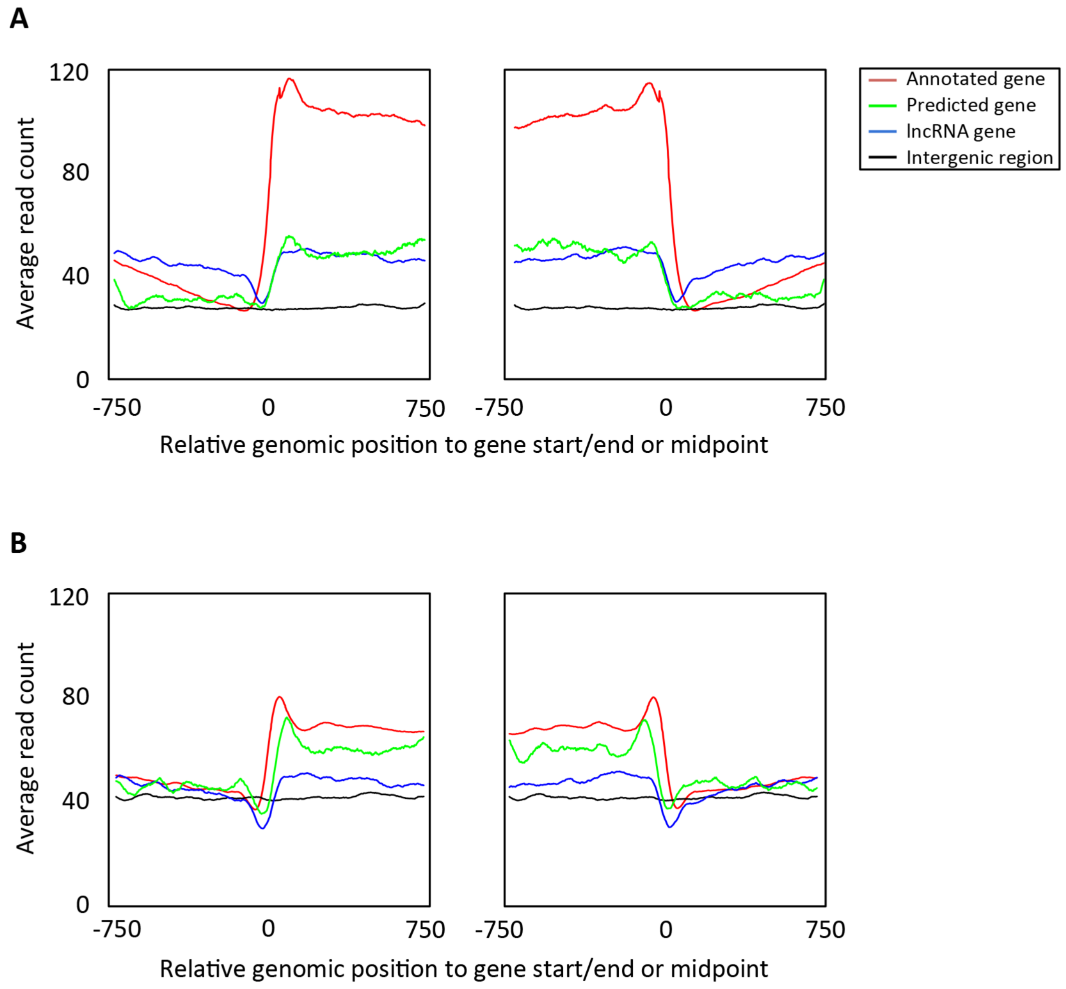


Figure 1.4. Nucleosome positioning profile around predicted genes. A. Average sequence coverage profiles around the start (left panel) and the end of genes (right panel) for annotated protein-coding genes (red), published lncRNA genes (blue) and predicted novel genes identified in this study (green) in the nucleosome occupancy data set from Bunnik *et al*. [23] (data set B1). The average nucleosome occupancy in intergenic regions is presented as a reference (black). **B.** Similar analysis for the nucleosome occupancy data set from Bartfai *et al*. [31] (data set B2). In all windows, the genomic position indicated on the x-axis is relative to the location of the gene start/end, or to the midpoint of intergenic windows.

As the majority of the predicted genes showed characteristics similar to those of lncRNA genes, we further classified our novel gene candidates into putatively protein-coding and non-protein-coding genes using a previously generated polysome-associated mRNA-Seq data set [19], which provides a snapshot of transcripts that are actively being translated. Using a cutoff of an average sequence depth of two reads across the entire predicted gene region, 73 out of 231 putative novel genes were found to be associated with polysomes and are thus potentially translated. For each predicted gene region, the longest open reading frame (ORF) was identified using ORF Finder with default values [39]. More putatively protein-coding gene candidates (6 out of 73 [8.2%]) than putatively non-coding gene candidates (3 out of 158 [0.2%]) contain an ORF longer than 100 amino acids (two-tailed Fisher's exact test, $P = 0.03$). In addition, putatively coding regions tend to have larger ORFs (average of 55 aa) than putatively non-coding regions (average of 48 aa, two-tailed Student's t-test, $P = 0.07$). On the other hand, the fraction of regions that does not contain an ORF larger than 30 amino acids is similar between both groups of gene candidates: 10 out of 73 (13.7%) of putative protein-coding regions versus 21 out of 158 (13.3%) of putative non-protein-coding regions (Supplemental File 1.1). However, 135 out of 231 novel genes have multiple non-overlapping ORFs on the same strand that could be exons belonging to a single gene. We could not find any evidence for splicing events in the RNA-seq data, although it should be mentioned that the sequence coverage in these regions is relatively low and may not allow the detection of such events.

Homology search

Comparative genomics is a powerful approach to gather evidence about putative genes. To find homologs of our putative novel genes, we aligned the predicted regions with known protein transcripts from the Uniprot-Trembl database using BLASTX [40-42]. Using stringent search settings (perfect matched length > 30% and e-value < 1e-5), no significant hits were found, suggesting that all of our predicted genes may be parasite-specific. This result is expected, since more than 50% of *P. falciparum* genes are unique to the parasite and majority of the putative genes identified in this study may be non-coding genes with low sequence conservation. Next, we searched against the reference RNA sequences (refseq_rna) database using the discontiguous MegaBLAST program of BLASTN that is tailored to more dissimilar sequences [41-43]. A similarity cutoff of e-value < 1e-5 resulted in two significant matches. One of the predicted gene regions (Pf3D7_14_v3:38,574-40,547) showed ~50% query coverage and more than 70% identity with approximately twenty of the *var* genes, while another putative gene region (Pf3D7_08_v3:1,288,505-1,289,391) showed more than 40% query coverage and 70% identity with ribosomal RNA sequences across protozoan species, including *Plasmodium vinckei vinckei*, *Theileria orientalis*, and *Babesia equi*. We also used BLAST to compare the candidate regions with a database of known expressed sequence tags (EST) and found 28 matches, the majority of which are derived from *P. falciparum* (Supplemental File 1.1). These findings provide independent evidence that our predicted regions might indeed contain novel genes.

Materials and Methods

Nucleosome positioning profiles

Nucleosome positioning profiles of the three main stages of *P. falciparum*'s asexual replication cycle were generated by micrococcal nuclease digestion of formaldehyde-crosslinked chromatin followed by chromatin immunoprecipitation using an antibody against histone H3. Nucleosome-bound DNA fragments were sequenced on the Illumina HiSeq platform as described in [23, 31]. Two *P. falciparum* 3D7 nucleosome positioning profiles were used in this study. Data set B1 from Bunnik *et al*. [23] consists of three asexual cycle time points (SRP026365), while data set

B2 from Bartfai *et al.* [31] consists of four asexual cycle time points (SRP003508). Reads were trimmed, mapped to *P. falciparum* 3D7 genome version 9.0 and were converted into coverage profiles by counting the number of sequence reads mapped at each nucleotide position as described in [23]. For each dataset, all coverage profiles were summed to generate a combined nucleosome profile *G* to be used as input data to train the classifier. The telomere and centromere regions display aberrant nucleosome coverage compared to the rest of the genome and were therefore removed from this data set.

By sliding a window of length *w* along the combined genome-wide profile *G* with a sliding step of h = 1 base pairs, we converted the input *G* into a set *D* of windows. Each window in *D* is a vector of length w, and each coordinate i of the vector represents the total number of mapped reads at location *i*. Inside each window, we defined a central region of length m, called *margin*. The total number of windows n is $\left\lceil\frac{(|G|-w+1)}{h}\right\rceil$, the coordinates of a window D_i $(i\ =\ 1,2,3\ldots n)$ is [a_i, b_i] = $[(i-1)h+1, (i-1)h+w]$ and the coordinates of the margin window D_i is $\left[ai+\frac{w}{2}-\ \frac{m}{2}, ai+\frac{w}{2}+\ \frac{m}{2}\right]$.

After extracting the windows, we assigned a label to each window depending on the presence or absence of a gene start or end, as defined below. Only the positive class and negative class windows were used to train the binary classifier for gene recognition. We defined a *negative class* as a window that does not overlap with any gene (intergenic windows), a *positive class* as a window that contains a gene start (or gene end for the detection of gene ends) inside the margin, and *other class* as a window that does not fall into the categories of positive or negative windows.

Cross Validation

The following section refers to the detection of gene starts. For gene ends, we used the same approach. To differentiate gene start sites and intergenic regions, a binary classifier was trained on positive class windows and negative class windows. Two randomly sampled data sets of windows were used interchangeably as training set or test set. One was sampled from windows of odd chromosomes, while the other was sampled from windows of even chromosomes. For each choice of parameter, we ran ten experiments. Odd chromosome windows were used as training and even chromosome windows were used for testing in the first five experiments, and vice versa for the other five experiments. All data was normalized with zero mean and unit variance. To evaluate the classifier's performance, we computed accuracy, precision and recall as described below.

$Accuracy = \frac{TP}{TP+FN}$

$Precision = \frac{TP}{TP+FP}$ (= specificity)

$Recall = \frac{TP}{TP+FN}$ (= sensitivity)

$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision+Recall}$

Recall and precision often show an inverse relationship, where it is possible to increase one at the cost of reducing the other. For our purpose of finding putative genes, the primary goal was to obtain the highest possible recall for both positive and negative classes.

Support Vector Machine Classifier

Support vector machine (SVM) is a family of binary classifiers than can learn from a training set to discriminate between positive and negative examples by finding a hyperplane that maximizes the margin [44]. To choose the best kernel for the SVM, we first used principal component analysis (PCA) to explore the relationship between the positive and negative classes, and then investigated different SVM kernels available from the Scikit-learn packages [45, 46]. Based on cross-validation experiments, we selected the RBF kernel and tuned the misclassification parameter *C* and the kernel parameter Υ using a two-dimensional grid search where C was chosen from the set $\{10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{6}, 10^{7}\}$, and Υ was chosen from the set $\{10^{-8}, 10^{-7}, 10^{-6}, \ldots, 10^{2}, 10^{3}\}$. All experiments were performed with 5-fold cross validation of 6,000 windows randomly sampled in equal quantities from both positive and negative class sets.

Training sample size, window size, and margin width

Using the optimized SVM-RBF hyperparameters, we tested how window size, training sample size, and margin width affect the performance of this classifier. We tested window size range from 500 bp to 2,000 bp with 500 base pair increments (Supplemental Figure 1.2A). We observed that short windows may not be able to capture enough context around the gene, while long windows resulted in increased computational cost and were problematic for the *P. falciparum* genome, where the average length of intergenic regions is 1,694 bases [5]. For margin width, we tested 25 bp, 50 bp and 100 bp (Supplemental Figure 1.2B). After testing different window sizes and margin widths in cross-validation experiments, we observed that the best recall rate is obtained using a window size of 1,500 bp and a margin width of 50 bp, which were selected as parameters for the final classifier.

In addition, we used cross-validation experiments to test the relationship between training sample size and the performance of the trained classifiers. We ran cross-validation experiments with training sizes of 2,000, 3,000, 6,000, 9,000, 12,000, and 18,000 windows (Supplemental Figure 1.2C). The results indicated that sample size does not have significant impact on the performance of the classifier, as long the sample is sufficiently large. We decided to use a training sample size of 6,000 windows with equal numbers derived from positive and negative classes, which achieves a good trade-off between computational cost and classifier performance. In contrast to this balanced training set, the vast majority of the windows in the genome are expected to be in the negative class. The imbalance in the test set should be reflected in the training set if the objective was to maximize the convex combination of precision and recall with the same weight. However, instead of optimizing precision, the main purpose of this study is to maximize the recall equally well for both positive and negative classes. The use of an imbalanced training set resulted in little change in recall, and we therefore used a balanced training set for this study. With these optimized parameters, we obtained average total recall rates of 0.94 for gene start classifiers trained on B1 data set, 0.92 for gene start classifiers trained on B2 data set, 0.94 for gene end classifiers trained on B1 data set and 0.93 for gene end classifiers trained on B2 data set (Supplemental File 1.1). The averaged total recall rate was 0.93 for all classifiers. The default confidence probability cutoff value for SVM classifier used here is 0.5. To increase the confidence of our gene prediction method, we tested different confidence probability cutoff values (0.6, 0.7, 0.8, 0.9) and observed that the number of predicted genes decreases as the cutoff value increases. We found that cutoff value of 0.7 gave the best trade-off between a reasonable number of predicted genes and a sufficiently high confidence in their prediction for both data sets B1 and B2.

Reverse transcription polymerase chain reaction (RT-PCR)

Twenty-one highly confident gene candidates that were predicted using both data set B1 and B2 were selected based on different combinations of RNA-Seq and polysomal RNA-Seq expression profiles (i.e. 3 genes showing high signals in both RNA-Seq and Poly-seq, 6 genes showing a high signal in only one of the profiles, and 12 genes showing low signals in both profiles). Total RNA was isolated from 10 ml of non-synchronous erythrocytic stage *P. falciparum* culture. To remove genomic DNA contamination, RNA samples were treated twice with 4 U DNase I (Life Technologies) per 10 µg of RNA for 30 minutes at 37°C. DNase I was inactivated by the addition of EDTA to a final concentration of 1 mM. 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 $MgCl_2$, 4 µl 0.1 M DTT, 2 µl 20 U/µl SuperaseIn and 1 µl 200 U/µl SuperScript III Reverse Transcriptase (all from 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. The absence of genomic DNA contamination was validated using a primer set targeting an intergenic region and a primer set targeting PfAlba3 (PF3D7_1006200) from inside exon 1 to within exon 2. Amplification of genomic DNA should give a product with a size of 429 bp including the intronic sequence, whereas amplification of cDNA should result in a fragment with a size of 164 bp. All 21 PCRs testing transcription activity of predicted genes were performed using 3 µl of the first-strand cDNA mixture with approximately 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 35 cycles. 5 µl of each PCR sample was used for agarose gel electrophoresis. For each primer set, PCR efficiency was tested using genomic DNA under the same amplification conditions as described above. All primer used for PCR validation are listed in Supplemental File 3.

Coverage plots and histone variant analysis

Sequence reads for ChIP-Seq experiments of *P. falciparum* nucleosome variant H2A.Z [31] (SRP003508) and histone marks H3K36me3 and H3K4me3 [38] (SRP022761) were downloaded and mapped to *P. falciparum* 3D7 genome version 9 using bowtie with default error rates. Coverage profiles for each time point were then generated using BEDtools [33]. For each histone variant, coverage profiles from different time points were summed to generate a combined profile. Sequence coverage for regions 750 bp before and after start and end codons of regions of interest were extracted from the summed coverage profiles. Averaged values for each relative position were then calculated and used to generate coverage plots using R.

Discussion and Conclusion

In this paper, we have used a machine learning approach for the detection of genes in the AT-rich genome of the human malaria parasite, *P. falciparum*, using exclusively nucleosome positioning data. Using classifiers trained on two independent nucleosome occupancy data sets, we detected a total of 231 putative novel genes. Eighty-eight of these 231 newly predicted genes exhibited transcription signal in RNA-Seq data and twenty out of 21 putative gene regions were validated by RT-PCR, indicating that our methodology is highly successful in identifying genes. Furthermore, of all putative gene regions identified using the nucleosome occupancy data set from Bunnik *et al*. [23], 69% were confirmed in the nucleosome positioning data set from Bartfai *et al*. [31], indicating that the classifiers trained on these two independently generated nucleosome landscapes are in good agreement. Collectively, our results demonstrate that local chromatin structure is sufficiently informative for genome annotation. Gene predictions based on nucleosome positioning datasets could thus be used to complement and augment sequence-based methodologies that are currently used for this purpose.

Based on the evidence we collected, it seems likely that many of the regions predicted here encode long non-coding RNAs. First, 108 of the predicted regions have been previously identified as lncRNA genes [18, 34-37]. Second, the sequence (GC-content) and nucleosome occupancy characteristics of the predicted regions are more similar to known lncRNAs than to protein-coding genes. Third, few of the predicted regions contain large ORFs. In other eukaryotic organisms, lncRNAs have been shown to be involved in the regulation of a multitude of cellular processes, one of which is regulation of gene expression by targeting general transcription factors and inducing chromatin remodeling [47-52]. In *P. falciparum*, identification and functional characterization of lncRNAs is ongoing. Most studies have focused on the identification of long non-coding telomeric end-associated transcripts that are similar to telomeric repeat-containing lncRNAs (TERRA) found in human and that are important for telomere maintenance [14, 35, 53]. Some of these lncRNAs contain binding sites for PfSIP2, a transcription factor specific to *Plasmodium* that is thought to be involved in regulation of *var* genes [35, 54]. This gene family is responsible for pathogenesis and immune evasion and most of its members are located in subtelomeric regions. These lncRNAs are likely to play important regulatory roles in *var* gene silencing by inducing heterochromatin formation, thus creating a repressive environment at the telomeric and subtelomeric ends [14, 35, 53, 55]. Additionally, lncRNAs have been implicated in various other processes, such as metabolic, biosynthetic and regulatory activities [14, 47, 56-59]. Our experimental results have expanded the list of putative lncRNAs in *P. falciparum*, and it will be of great interest to further validate and characterize these transcripts to understand their function in parasite biology.

Unfortunately, we were unable to use nucleosome positioning as a means to discover novel genes in the telomeric regions. Due to aberrant nucleosome positioning in the telomeric and centromeric regions compared to the rest of the genome, we had to exclude these regions from our gene predictions. The number of known lncRNAs derived from these regions is too small (n = 22) for accurate training of a separate classifier on these atypical parts of the genome.

In addition to putative lncRNAs, we also distinguished 73 regions that may contain protein-coding genes, based on the association of their transcripts with polysomes. The polysome profiling data set used in this study was obtained by separating polysomes on a sucrose gradient, followed by isolation and sequencing of mRNA in the polysome fractions [19]. This methodology provides a catalogue of transcripts that are actively being translated. However, it also captures polyadenylated transcripts that are merely associated with polysomes as regulatory elements, or that are present in ribonucleoprotein complexes that co-sediment with polysomes. Based on polysome profiling data alone, it is therefore impossible to determine whether a gene encodes a protein. Further study will be necessary to determine the translational status of the putative protein-coding genes identified in this study.

Beside protein-coding genes and genes encoding lncRNAs, a third option for regions identified in this study is to contain pseudogenes. For decades, pseudogenes have been considered non-functional or ‘junk’ DNA; however, the conserved sequence similarity between pseudogenes and coding genes suggests a selective maintenance of these non-coding elements. They may have an important biological role that has not yet been fully understood. In recent mammalian studies, transcripts of pseudogenes showed regulatory roles, largely through antisense mechanisms [60, 61]. Expressed pseudogenes have also been implicated in mRNA stability in transgene mouse mutants [62]. Similar regulatory pseudogenes may also be present in *P. falciparum*, in particular in predicted gene regions with homology to annotated genes as identified using BLAST searches.

As a selection criterion for the identification of regions containing putative novel genes, we used the presence of both a gene start and a gene end within the same intergenic region. However, we also identified regions with only a predicted gene start or a gene end, but not both. Often, the intergenic regions containing these single-end predictions do show sequence coverage in the steady state or polysomal RNA-Seq data sets. Possible explanations for such single-end predictions include the presence of genes coding for small transcripts that are difficult to capture using a nucleosome positioning dataset. Each nucleosome covers approximately 146 base pairs of DNA, raising the possibility that short genes do not show distinct nucleosome occupancy features. Alternatively, the nucleosome features at the other end of the predicted gene region may be irregular and therefore not meet the quality threshold for selection.

In this study, we have demonstrated that using a machine learning approach trained on the nucleosome landscape around genes, we were able to identify 231 putative genes, of which the majority showed evidence of expression in RT-PCR, EST, steady-state RNA-Seq, or polysomal RNA-Seq data sets in the malaria parasite, *P. falciparum*. A similar methodology could be used for predicting the location of transcription start sites (TSSs), since TSSs are generally marked by an upstream nucleosome-depleted region. Therefore, this approach may ultimately be useful to identify key regulatory elements and to complement other sequence-based genome annotation efforts, which will provide further insights into gene regulatory mechanisms in *P. falciparum*. Furthermore, similar machine learning approaches may also be applied to other organisms as long as a nucleosome-positioning data set is available and the nucleosome landscape around genes shows regular periodic characteristics.

Reference

1. Lu XM, Bunnik EM, Pokhriyal N, Nasseri S, Lonardi S, Le Roch KG: **Analysis of nucleosome positioning landscapes enables gene discovery in the human malaria parasite Plasmodium falciparum.** *BMC Genomics* 2015, **16:**1005.

2. WHO: **World Malaria Report. 2014.** http://www.who.int/malaria/publications/world_malaria_report_2014/report/en/**.** 2014.

3. 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.

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

5. 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.

6. Le Roch KG, Chung DW, Ponts N: **Genomics and integrated systems biology in Plasmodium falciparum: a path to malaria control and eradication.** *Parasite Immunol* 2012, **34:**50-60.

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. Watanabe J, Sasaki M, Suzuki Y, Sugano S: **FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, Plasmodium falciparum.** *Nucleic Acids Res* 2001, **29:**70-71.

9. Hyman RW, Fung E, Conway A, Kurdi O, Mao J, Miranda M, Nakao B, Rowley D, Tamaki T, Wang F, Davis RW: **Sequence of Plasmodium falciparum chromosome 12.** *Nature* 2002, **419:**534-537.

10. Gardner MJ, Shallom SJ, Carlton JM, Salzberg SL, Nene V, Shoaibi A, Ciecko A, Lynn J, Rizzo M, Weaver B, et al: **Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14.** *Nature* 2002, **419:**531-534.

11. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ: **cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome.** *BMC Genomics* 2007, **8:**255.

12. 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.

13. 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.

14. 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.

15. Ngwa CJ, Scheuermayer M, Mair GR, Kern S, Brugl T, Wirth CC, Aminake MN, Wiesner J, Fischer R, Vilcinskas A, Pradel G: **Changes in the transcriptome of the malaria parasite Plasmodium falciparum during the initial phase of transmission from the human to the mosquito.** *BMC Genomics* 2013, **14:**256.

16. Sorber K, Dimon MT, DeRisi JL: **RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts.** *Nucleic Acids Res* 2011, **39:**3820-3835.

17. Yamagishi J, Natori A, Tolba ME, Mongan AE, Sugimoto C, Katayama T, Kawashima S, Makalowski W, Maeda R, Eshita Y, et al: **Interactive transcriptome analysis of malaria patients and infecting Plasmodium falciparum.** *Genome Res* 2014, **24:**1433-1444.

18. 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.

19. 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.

20. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458:**223-227.

21. Jansen A, Verstrepen KJ: **Nucleosome positioning in Saccharomyces cerevisiae.** *Microbiol Mol Biol Rev* 2011, **75:**301-320.

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

23. 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.

24. 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.

25. 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.

26. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast.** *Nat Genet* 2007, **39:**1235-1244.

27. 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.

28. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics.** *Nat Rev Genet* 2009, **10:**161-172.

29. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature* 2009, **458:**362-366.

30. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K: **Dynamic regulation of nucleosome positioning in the human genome.** *Cell* 2008, **132:**887-898.

31. 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.

32. Nemates.org: http://nemates.org/MA/progs/overlap_stats.html. 2015.

33. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. Rombel IT, Sykes KF, Rayner S, Johnston SA: **ORF-FINDER: a vector for high-throughput gene identification.** *Gene* 2002, **282:**33-41.

40. UniProt-TrEMBL: http://www.ebi.ac.uk/uniprot.

41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

42. **Blast** [http://www.ncbi.nlm.nih.gov/BLAST/]

43. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al: **RefSeq: an update on mammalian reference sequences.** *Nucleic Acids Res* 2014, **42:**D756-763.

44. Chih-Wei Hsu C-cC, and Chih-Jen Lin: **A Practical Guide to Support Vector Classification.** Tech. Rep.; 2010.

45. Komura D, Nakamura H, Tsutsumi S, Aburatani H, Ihara S: **Multidimensional support vector machines for visualization of gene expression data.** *Bioinformatics* 2005, **21:**439-444.

46. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G: **Machine learning for neuroimaging with scikit-learn.** *Front Neuroinform* 2014, **8:**14.

47. Saxena A, Carninci P: **Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs.** *Bioessays* 2011, **33:**830-839.

48. Sanchez-Elsner T, Gou D, Kremmer E, Sauer F: **Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax.** *Science* 2006, **311:**1118-1123.

49. Rodriguez-Campos A, Azorin F: **RNA is an integral component of chromatin that contributes to its structural organization.** *PLoS One* 2007, **2:**e1182.

50. Goodrich JA, Kugel JF: **Non-coding-RNA regulators of RNA polymerase II transcription.** *Nat Rev Mol Cell Biol* 2006, **7:**612-616.

51. Bernstein E, Allis CD: **RNA meets chromatin.** *Genes Dev* 2005, **19:**1635-1655.

52. Lee JT: **Epigenetic regulation by long noncoding RNAs.** *Science* 2012, **338:**1435-1439.

53. Vembar SS, Scherf A, Siegel TN: **Noncoding RNAs as emerging regulators of Plasmodium falciparum virulence gene expression.** *Curr Opin Microbiol* 2014, **20:**153-161.

54. 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.

55. Bah A, Azzalin CM: **The telomeric transcriptome: from fission yeast to mammals.** *Int J Biochem Cell Biol* 2012, **44:**1055-1059.

56. Hung CL, Wang LY, Yu YL, Chen HW, Srivastava S, Petrovics G, Kung HJ: **A long noncoding RNA connects c-Myc to tumor metabolism.** *Proc Natl Acad Sci U S A* 2014, **111:**18697-18702.

57. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10:**155-159.

58. Vance KW, Sansom SN, Lee S, Chalei V, Kong L, Cooper SE, Oliver PL, Ponting CP: **The long non-coding RNA Paupar regulates the expression of both local and distal genes.** *EMBO J* 2014, **33:**296-311.

59. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81:**145-166.

60. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes.** *Nature* 2008, **453:**534-538.

61. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP: **A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.** *Nature* 2010, **465:**1033-1038.

62. Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A: **An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene.** *Nature* 2003, **423:**91-96.

Supplemental Information

Abbreviations

EST: expressed sequence tag; lncRNA: long non-coding RNA; ORF: open reading frame; refseq_rna: reference RNA sequence; TERRA: telomeric repeat-containing lncRNAs; TSS: transcription start sites; SVM: support vector machine; PCA: principal component analysis; RBF: radial basis function; RT-PCR: reverse transcription polymerase chain reaction.

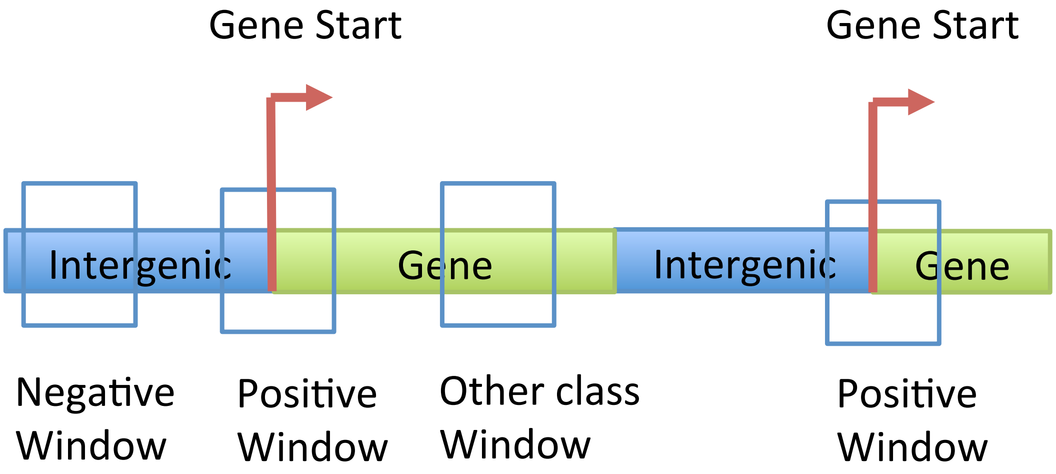
Author's contributions

XML performed all computational analyses, participated in study design and drafted the manuscript. EMB participated in design of the study, contributed to critical discussions and drafted the manuscript. NP participated in design of the study. SN assisted in computational analyses. SL and KGLR designed the study, supervised the project and helped drafting the manuscript. All authors have read and approved the final manuscript.

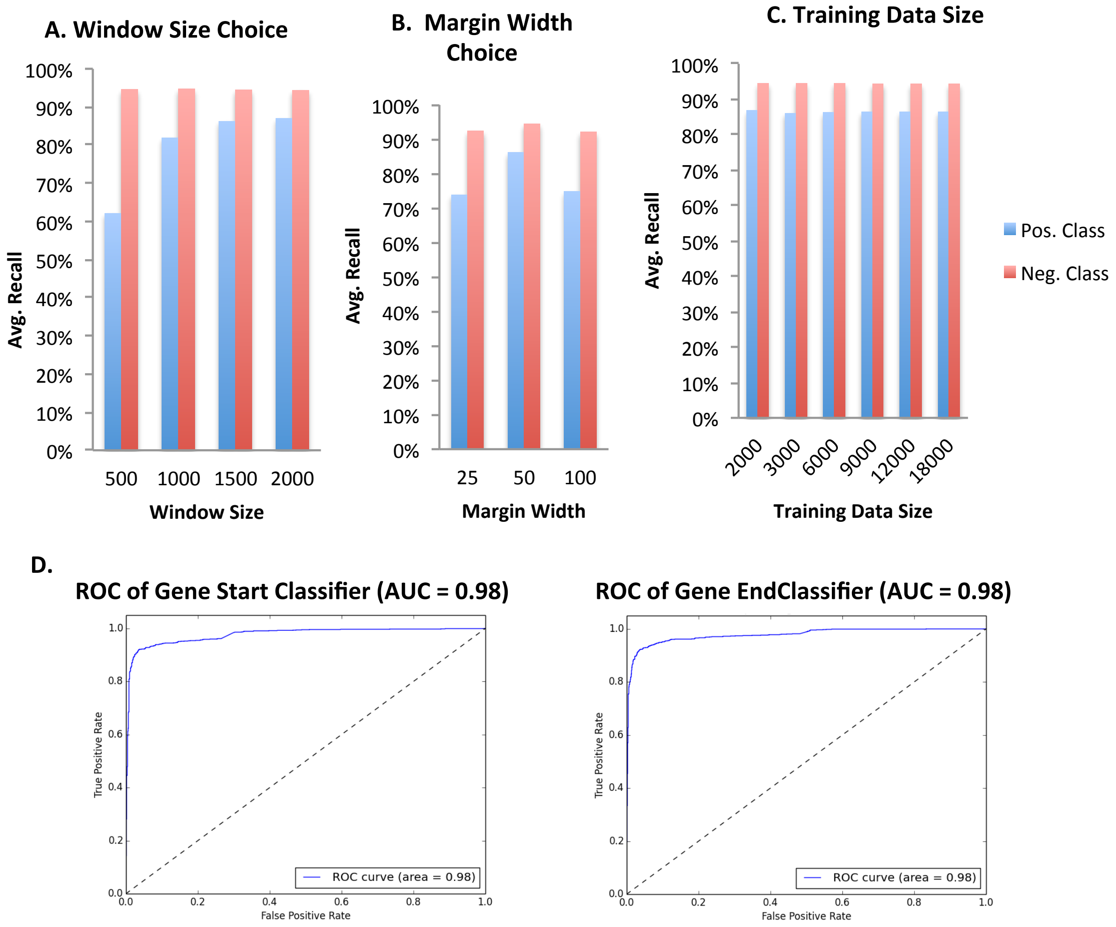
Acknowledgements

We thank Jacques Prudhomme for maintaining parasite cultures and assisting in experimental procedures. The following reagents were obtained through the MR4 as part of the BEI Resources Repository, NIAID, NIH: Plasmodium falciparum strains 3D7 (MRA-102) deposited by D.J. Carucci. This study was financially supported by the National Institutes of Health (grant R01 AI85077-01A1 to K.L.R. and S.L.) and the National Science Foundation (grant IIS-1302134 to S.L. and K.L.R.).

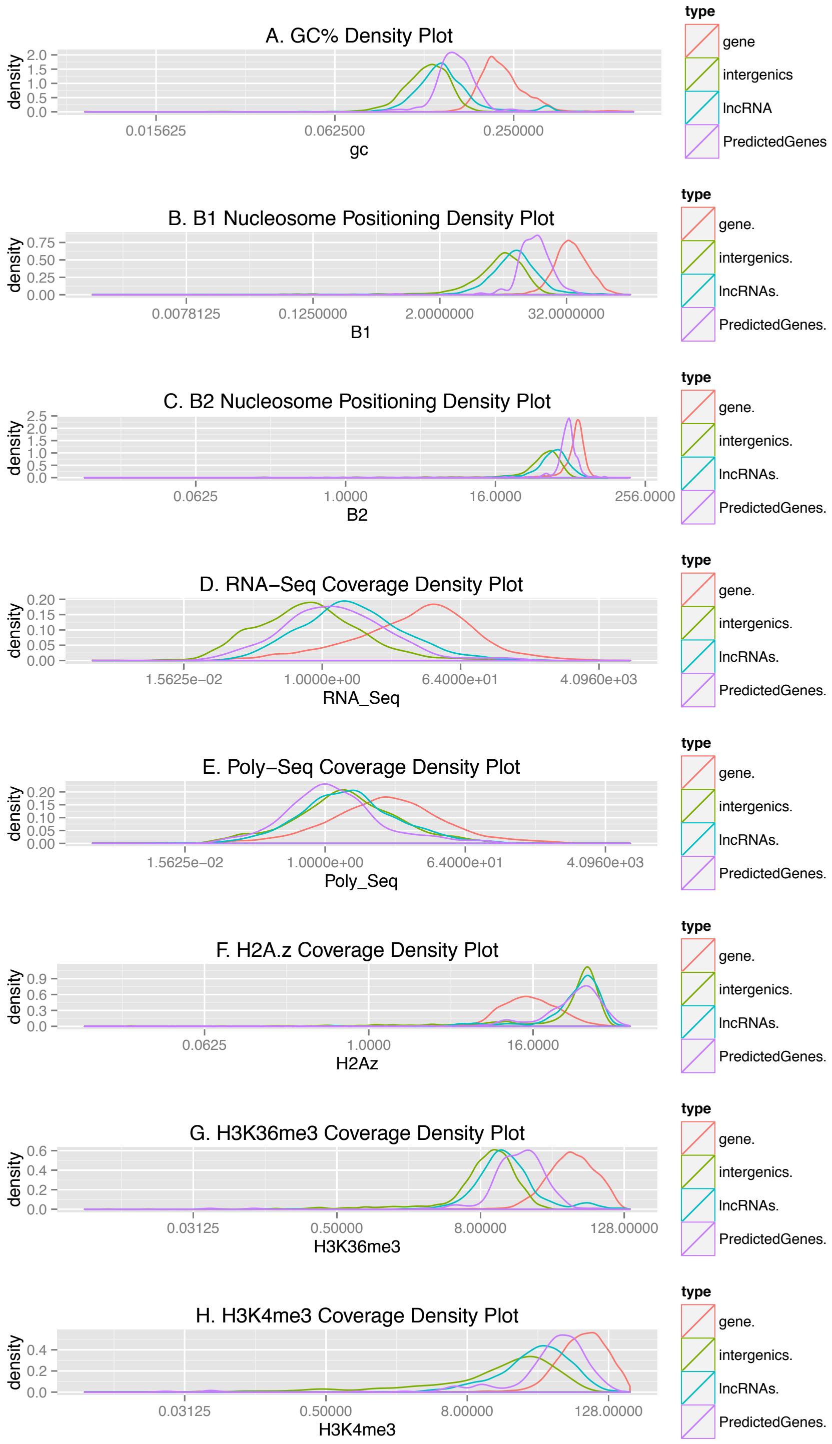
Supplemental Figures



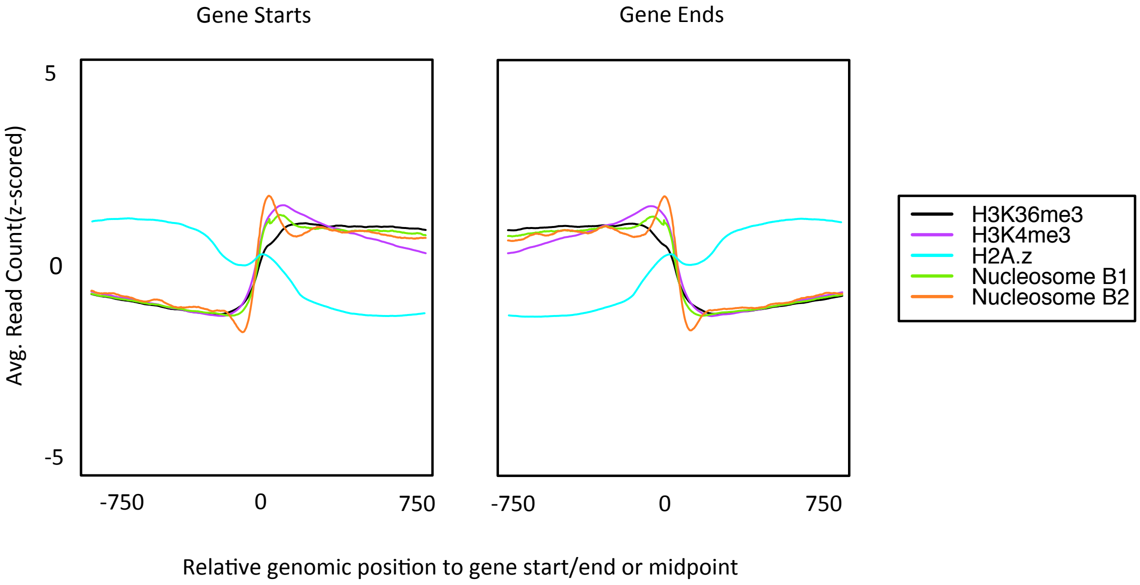
Supplemental Figure 1.1. Supervised machine learning approach for novel gene detection. Using a sliding window method, the genome-wide nucleosome positioning data set was converted into a set of subsequences (“windows”), where each window is a vector of length w, and each position is a numeric value representing the summed number of mapped reads. A label was then assigned to each of the windows based on the presence of a gene start. A binary classifier for gene start recognition was trained on gene start-containing windows (positive class) and intergenic windows (negative class) with support vector machine (SVM), RBF kernel. A similar approach was used to train a classifier for the detection of gene ends.



Supplemental Figure 1.2. Optimization of classifier parameters trained on the positive strand of data set B1. Average recall rates for gene start detection from 10 cross-validation experiments for window size (A), margin width (B), and training data size (C). After comparing the recall rate for each parameter, the optimized classifier was trained using 6,000 windows of 1,500 bp with 50 bp margin width drawn in equal quantities from both positive and negative class. The ROC curves for optimized gene start and gene end classifiers are reported in (D). Results of optimization experiments for classifiers trained on the negative strand of data set B1 and classifiers trained on data set B2 were very similar and are therefore not shown. A detailed explanation of classifier optimization is presented in the Material and Methods section.



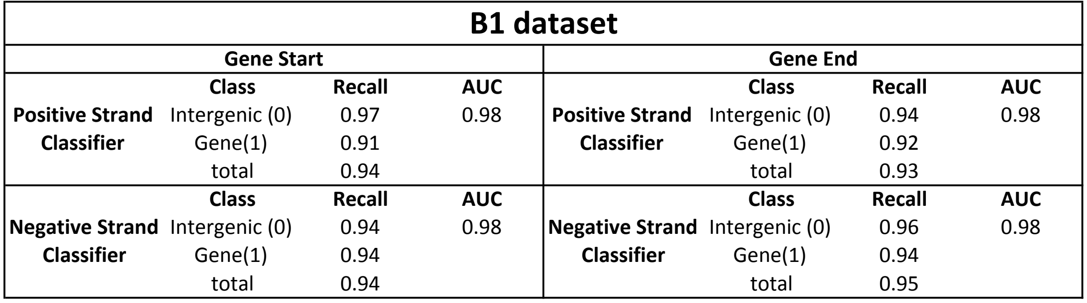
Supplemental Figure 1.3. Density plots of various characteristics of predicted gene regions versus intergenic regions and annotated coding and non-coding genes in *P. falciparum*.



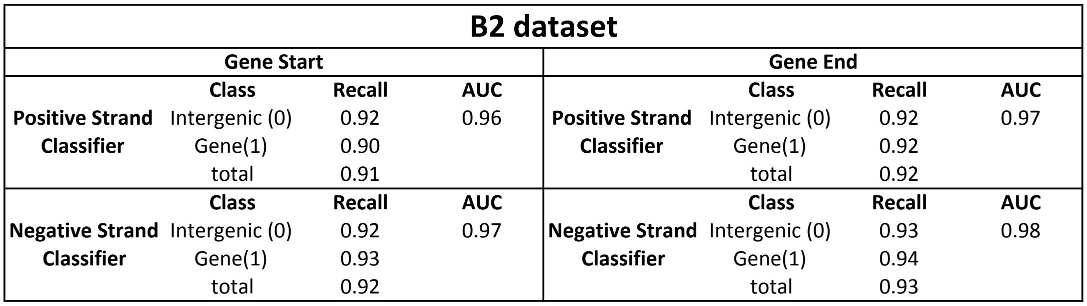
Supplemental Figure 1.4: Coverage profiles of histone variants around gene boundaries.

Supplemental Table

Supplemental Table 1.1: Classifier performance records.



B1 dataset							
Gene Start				Gene End			
	Class	**Recall**	**AUC**		**Class**	**Recall**	**AUC**
Positive Strand Classifier	Intergenic (0)	0.97	0.98	**Positive Strand Classifier**	Intergenic (0)	0.94	0.98
	Gene(1)	0.91			Gene(1)	0.92	
	total	0.94			total	0.93	
	Class	**Recall**	**AUC**		**Class**	**Recall**	**AUC**
Negative Strand Classifier	Intergenic (0)	0.94	0.98	**Negative Strand Classifier**	Intergenic (0)	0.96	0.98
	Gene(1)	0.94			Gene(1)	0.94	
	total	0.94			total	0.95	



B2 dataset							
Gene Start				Gene End			
	Class	**Recall**	**AUC**		**Class**	**Recall**	**AUC**
Positive Strand Classifier	Intergenic (0)	0.92	0.96	**Positive Strand Classifier**	Intergenic (0)	0.92	0.97
	Gene(1)	0.90			Gene(1)	0.92	
	total	0.91			total	0.92	
	Class	**Recall**	**AUC**		**Class**	**Recall**	**AUC**
Negative Strand Classifier	Intergenic (0)	0.92	0.97	**Negative Strand Classifier**	Intergenic (0)	0.93	0.98
	Gene(1)	0.93			Gene(1)	0.94	
	total	0.92			total	0.93	

Supplemental File

Supplemental File 1.1: List of predicted gene regions and their characteristics. (XLSX)

Supplemental File 1.2: Primers used for predicted gene validation. (XLSX)

Chapter 2: Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite *Plasmodium falciparum*

Xueqing Maggie Lu[1], Gayani Batugedara[1], Michael Lee[1], Jacques Prudhomme[1], Evelien M. Bunnik[1,2], and Karine G. Le Roch[1]

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

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

This chapter originally appeared as Lu et al. *NAR*, 2017 [1].

Abstract

Gene expression in *P. falciparum* is tightly regulated to ensure successful propagation of the parasite through the different stages of its complex life cycle. The earliest genome-wide transcriptomics studies in *P. falciparum* suggested a cascade of transcriptional activity over the course of the 48-hour intraerythrocytic developmental cycle (IDC). In recent years, this model of just-in-time transcription has been challenged by the finding that post-transcriptional regulation plays an important role in parasite gene expression. In this chapter, we set out to generate the first genome-wide nuclear run-on (GRO-seq) data set in *P. falciparum* to accurately determine the timing of transcription. Findings in this chapter indicate that a majority of genes is transcribed simultaneously during the IDC and that only a small subset of genes is subject to differential transcriptional regulation. RNA polymerase II is engaged with promoters of all genes prior to this transcriptional burst, suggesting that Pol II pausing plays a dominant role in gene regulation at the level of transcription. During gametocyte differentiation, the parasite stage that is transmitted to mosquitoes, the overall transcriptional program is surprisingly similar to the IDC, with the exception of relatively small subsets of genes that are either upregulated, such as motor genes, or downregulated, such as invasion genes. Results from this chapter suggest that further characterization of the molecular players that regulate stage-specific gene expression and Pol II pausing, in particular the kinases involved in this process, will contribute to our continuous search for novel targets of antimalarial drugs.

Introduction

As one of the world's deadliest infectious diseases, malaria is responsible for about 438,000 deaths annually, the vast majority of which occur among children under the age of five [2]. Of the five *Plasmodium* species that can cause malaria in humans, *P. falciparum* is responsible for the most severe form of malaria and causes about 90% of all malarial deaths [2]. Currently, no approved efficient protective vaccine is available for disease prevention, and the rapid development of parasite resistance to current antimalarial drugs is a major challenge for the control of malaria. Therefore, a better understanding of the parasite's biological system is required to identify novel drug targets and to further combat the disease.

P. falciparum has a complex life cycle involving multiple phases in its human and mosquito hosts. The stage responsible for clinical malaria is the intraerythrocytic developmental cycle (IDC). In this cycle, the parasite replicates asexually inside red blood cells and develops through ring, trophozoite and schizont stages to multiply into 16-32 daughter parasites [3]. During the IDC, environmental stress can induce sexual differentiation of parasites into male and female gametocytes. Gametocytes are morphological and functionally different from asexual parasites. Mature gametocytes, ingested by a mosquito, undergo sexual replication in the mosquito midgut and further develop into the salivary gland sporozoites that can be transmitted to a new human host. Formation and carriage of gametocytes is key to disease transmission.

This multi-stage life cycle of the parasite is highly fine-tuned, presumably by strict control of stage-specific gene expression. In eukaryotes, stage-specific regulation of gene expression can be a combined effect of transcriptional, post-transcriptional and translational control. In *P. falciparum*, the nature and the contribution of mechanisms regulating gene expression are still poorly understood. Compared to organisms with similar genome size, only one-third of the expected number of specific transcription factors (TFs) and few mediator subunits have been uncovered in the *P. falciparum* genome [4, 5]. Recently, an apicomplexan-specific family of proteins containing AP2 DNA binding domains (ApiAP2) has been identified in apicomplexan parasites as the major group of putative sequence-specific transcription factors [6-9]. While their number is relatively small (27 in the *P. falciparum* genome), they are likely to act as master regulators of transcription during parasite development. However, it remains unclear how such a limited number of TFs can generate complex patterns of gene expression in multiple life cycle stages.

Accumulating evidence suggests that *P. falciparum* uses chromatin structure as a basal control for transcriptional initiation. Genome architecture studies [10, 11] showed that chromatin is relatively closed during the ring and schizont stages, but opens substantially during the trophozoite stage, providing a transcriptional permissive state. This open-and-closed binary chromatin activity is also reflected in nucleosome occupancy studies [11-13] and histone abundance levels [14-16]. Nucleosome density is relatively low at the trophozoite stage, but maintained high at early ring and late schizont stages. In addition, studies using chromatin immunoprecipitation directed against RNA polymerase II (Pol II) and coupled to genomic DNA microarrays (ChIP-on-chip) indicate that Pol II is divided into a bi-phasic occupancy throughout the parasite IDC [17].

Controversially, such chromatin structure re-arrangement activity and Pol II profiling do not correlate well with previously observed complex patterns of gene expression profiles constructed from steady-state mRNA [18-22]. The cascade of gene expression observed at the steady-state mRNA level led to a “just-in-time” model suggesting mRNA is produced when the encoded protein is required during the cell cycle. However, comparative genomics analysis of steady state mRNA and protein profiles of different *P. falciparum* stages showed a significant delay between peak of mRNA and protein levels for 30 to 40% of the analyzed genes [14, 23], supporting a model of “just-in-time” translation for some mRNAs. Recently published novel genome-wide approaches that compared steady-state mRNA with polysome-associated mRNA or ribosomal occupancy of mRNAs have also provided a good indicator of active protein production and further validated the model of “just-in-time” translation for specific subsets of genes [19, 24]. In particular, this was true for proteins involved in remodeling of the erythrocyte just after parasite invasion.

The paucity of transcription factors [25], the lack of identified DNA regulatory elements [26] together with the weak correlation between chromatin-remodeling events, Pol II occupancy, and steady-state mRNA levels, suggested significant post-transcriptional mechanisms regulating the parasite development. However, when *Plasmodium* genes are exactly transcribed and how many of them are regulated at the post-transcriptional level to generate the cascade of steady-state mRNA observed throughout the parasite life cycle remains to be determined.

Results

Generation of nascent RNA profiles for the P. falciparum blood stages

In this study, we explored gene expression in *P. falciparum* at the initiation level using a modified global run-on sequencing (GRO-seq) methodology [27, 28] that specifically captures newly transcribed RNA (nascent RNA) in a genome-wide manner. An overview of the GRO-seq methodology is presented in Figure 2.1A. We have generated eight genome-wide nascent RNA profiles covering six asexual stages across the IDC, and early (stage II/III) and late (stage IV/V) gametocyte stages. To optimize the protocol, we determined that a minimum of 30 minutes incubation period was required to obtain sufficient nascent RNA from the *in vitro* transcription reaction (Supplemental Flie 2.1A). Several additional quality controls were implemented to further validate the GRO-seq methodology. First, experimental noise was measured by performing the nuclear run-on reaction in the presence of unmodified uridine (Supplemental Figure 2.1B). This negative control yielded extremely low amounts of RNA (Supplemental Figure 2.1B) and low genome coverage (<1-fold) after sequencing (Supplemental Figure 2.1C), confirming minimal DNA and non-nascent RNA contamination using GRO-seq methodology. Second, to confirm that the nuclear run-on reaction generates nascent RNA in a stage-specific manner, we performed our assay on tightly synchronized trophozoite-stage parasites and observed that TEX1 gene (trophozoite exported protein 1; PF3D7_0603400) was highly transcribed, while no signal was obtained for a sporozoite-specific gene, STP (putative serine/threonine protein kinase; PF3D7_0107600) (Supplemental Figure 2.1B). Finally, we generated two biological replicates for five of the IDC stages, which showed high Spearman correlation coefficients ranging from 0.88 to 0.95 (Supplemental Figure 2.1D), confirming the reproducibility of the GRO-seq methodology. Together, these results indicate that our experimental approach is efficient, reproducible, and has minimal background noise.

Upon sequencing of the GRO-seq libraries, we obtained between 670,456 to 11,903,568 mapped and filtered reads per stage after combining biological replicates, corresponding to 1.46 to 25.88 fold exome-coverage. To be able to directly compare gene expression levels between the various stages, we corrected for differences in the number of parasites used as input for the nuclear run-on reaction. This normalization was achieved by dividing the coverage read depth at each base pair by a stage-specific scaling factor that was based on, among others, the culture volume and parasitemia at the time point of harvest (see Material and Methods and Supplemental Flie 2.1).

A global picture of transcriptional activity in the blood stages

Overall, GRO-seq data revealed that transcriptional activity exhibited a bell curve shape during asexual cycle, from extremely low global transcriptional activity at the early ring stage, via slightly increased transcription at the late ring stage to a strong peak of transcription at the early and late trophozoite stages. Finally, transcriptional activity decreases again as the parasite progressed into the late schizont stage (Figure 2.1B, Supplemental Figure 2.2 and Supplemental Figure 2.3, and Supplemental Flie 2.2). This bell-curved pattern was validated using Immunofluorescence microscopy capturing RNA polymerase II abundance levels at single-cell resolution (Figure 2.1C). In addition, our GRO-seq data indicates that transcriptional activity was high in early gametocytes and subsequently decreased in late gametocytes (Figure 2.1B, Supplemental Figure 2.2 and Supplemental Figure 2.3, and Supplemental Flie 2.2). A total of 5,207 genes (99% of all protein-coding genes) were detected in at least one of the eight stages sampled in this study (Supplemental Flie 2.1), while 77 genes did not reach our threshold. These genes included genes expressed on the RBC surface (*var*, *surfin*) and genes expressed in other stages of the parasite life cycle, such as sporozoite invasion-associated protein 1 and liver specific protein 1 putative (LISP1).

Cluster analysis of transcriptional profiles across the P. falciparum asexual cell cycle

We identified a total of 5,187 genes expressed at any time point during the IDC, which were grouped into nine distinct clusters based on their nascent transcriptional profile across the IDC (Figure 2.1D and Supplemental Flie 2.1). The large majority of genes (n = 4,607; 89%) were most abundantly transcribed at the trophozoite stage, while 532 genes (10%) showed a high level of transcription at the schizont stage and 48 genes (1%) were most highly transcribed at the ring stage. We observed enrichment in Gene Ontology (GO) terms associated with host cell remodeling among the earliest transcribed genes (clusters A1 and A2) including several PHISTb and early transcribed membrane proteins (ETRAMPs) (Figure 2.1D, Supplemental Flie 2.1, and Supplemental Flie 2.3). Six out of 25 AP2 TFs detected during the IDC were most highly transcribed at these early stages, suggesting that these TF could be involved in driving this first wave of transcriptional activity. Genes that were most highly transcribed at the late trophozoite stage (cluster A3-A7) were associated with biological processes that are known to occur at this stage, such as translation and DNA replication. The earliest of these subsets of genes (cluster A3) included PfAlba1, as well as 16 putative RNA-binding proteins, and showed GO enrichment for "nucleic acid binding" and "RNA binding". Genes involved in pathogenesis associated with GO terms such as "cell adhesion molecule binding" and "infected host cell surface knob" were enriched at the early schizont stage (cluster A8). Finally, a relatively small number of genes (cluster A9) with strong enrichment for involvement in host cell invasion, such as merozoite surface proteins and rhoptry-associated proteins, were most abundantly transcribed at the late schizont stage. To validate these cluster assignments, we determined the relative transcript abundance for several genes detected at different stages of the IDC using semi-quantitative PCR. Ring, trophozoite, and schizont-stage genes all showed a good concordance between PCR results and the GRO-seq cluster analysis. In addition, a gene that did not pass our threshold for expression was also not detected by PCR (Supplemental Figure 2.4). Together, these results indicate that most genes are highly transcribed simultaneously at the trophozoite stage, while only a subset of genes is differentially regulated and transcribed either early in the cell cycle to enable the parasite to establish a hospitable environment inside the erythrocyte, or late in the cell cycle in preparation for merozoite egress and re-invasion.

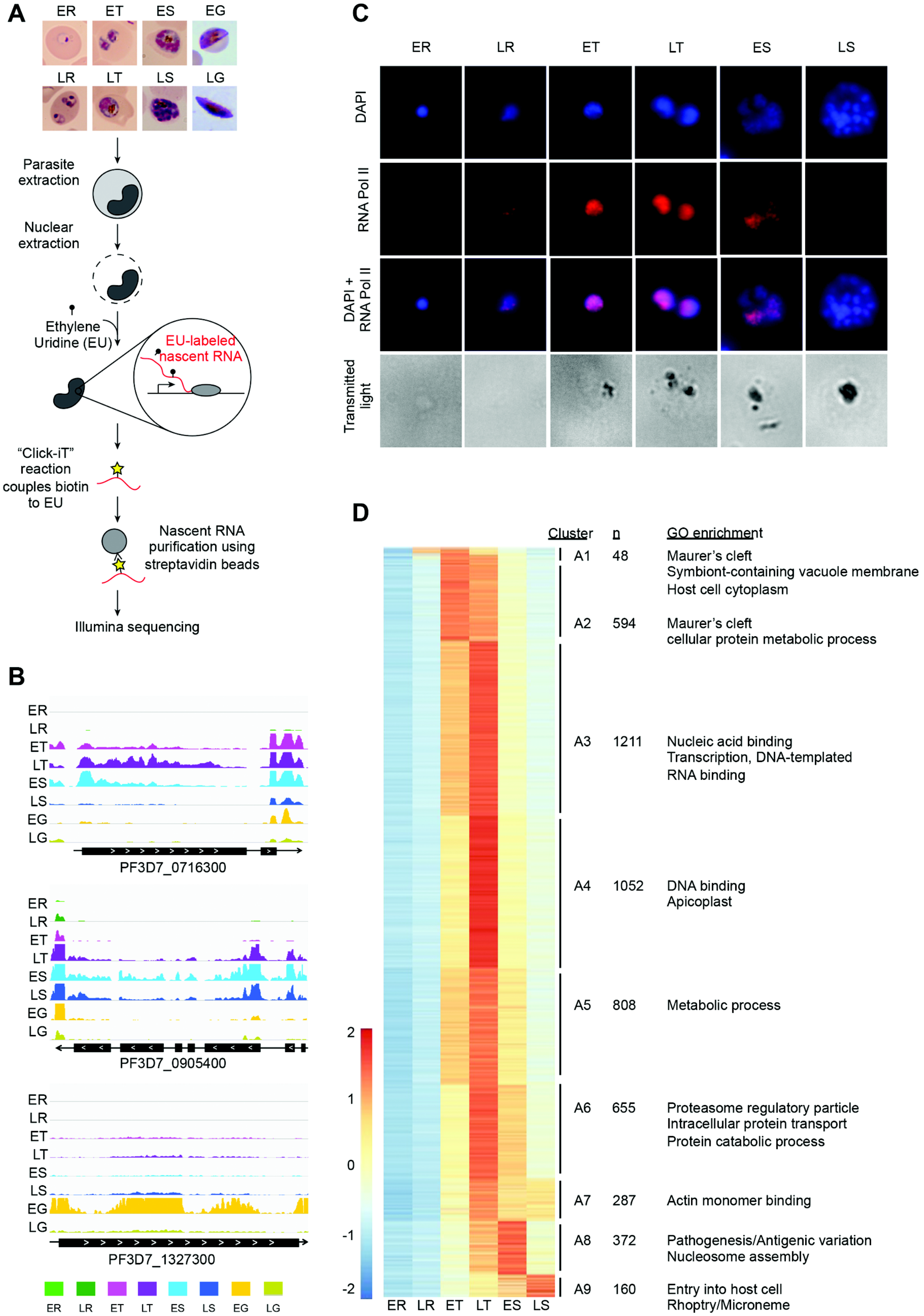

A
ER
ET
ES
EG
LR
LT
LS
LG
Parasite extraction
Nuclear extraction
Ethylene Uridine (EU)
EU-labeled nascent RNA
"Click-iT" reaction couples biotin to EU
Nascent RNA purification using streptavidin beads
Illumina sequencing
B
ER
LR
ET
LT
ES
LS
EG
LG
PF3D7_0716300
PF3D7_0905400
PF3D7_1327300
ER LR ET LT ES LS EG LG
C
ER
LR
ET
LT
ES
LS
DAPI
RNA Pol II
DAPI + RNA Pol II
Transmitted light
D
Cluster
n
GO enrichment
A1 48 Maurer's cleft
Symbiont-containing vacuole membrane
Host cell cytoplasm
A2 594 Maurer's cleft
cellular protein metabolic process
A3 1211 Nucleic acid binding
Transcription, DNA-templated
RNA binding
A4 1052 DNA binding
Apicoplast
A5 808 Metabolic process
A6 655 Proteasome regulatory particle
Intracellular protein transport
Protein catabolic process
A7 287 Actin monomer binding
A8 372 Pathogenesis/Antigenic variation
Nucleosome assembly
A9 160 Entry into host cell
Rhoptry/Microneme
2
1
0
-1
-2
ER LR ET LT ES LS

Figure 2.1. Global nuclear Run-On coupled to next-generation sequencing (GRO-seq) in the malaria parasite *P. falciparum*. (A) Schematic overview of the GRO-seq methodology. In brief, parasites were extracted from highly synchronized cultures followed by extraction of the nuclei. Transcription was allowed to take place for 30 minutes in the presence of 5-ethynyl uridine (EU). EU-labeled RNA was then purified and prepared for Illumina sequencing. **(B)** Genome browser view of normalized nascent RNA profiles that show the varying levels of transcriptional activity during the blood stages. The majority of genes were most highly transcribed at the trophozoite stage, and a representative gene (PF3D7_0716300) is shown in the top panel. A subset of genes was most highly transcribed at the schizont stage or gametocyte stages. An example of a gene that is highly transcribed at the schizont stage (PF3D7_0905400) is shown in the middle panel, whereas a gene with a profile of high transcription at the gametocyte stage (PF3D7_1327300) is shown in the bottom panel. **(C)** Immunofluorescence analysis showing RNA polymerase II activity at the asexual IDC stages. A strong signal was detected at the early and late trophozoite stages. At the early schizont stage, higher activity of RNA polymerase II was detected in some nuclei compared to others within a single parasite. No RNA polymerase II signal was observed at the early ring and late schizont stages. **(D)** A total of 5,221 genes were identified to be expressed during the IDC and were grouped into 9 clusters based on their expression patterns. A selection of enriched GO-terms is listed on the right of each cluster. ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont; EG, early gametocyte stage; LG, late gametocyte stage.

RNA polymerase II occupancy confirms nascent transcriptional profiles

To further validate our observation of widespread transcriptional activity in trophozoites and targeted transcription of mostly invasion-related genes in schizonts, we determined Pol II occupancy at the early ring, early trophozoite, and late schizont stages using chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq; Supplemental Figure 2.5, and Supplemental Flie 2.4). Pol II complexes were purified using an antibody that specifically binds to the C-terminal repeat domain with phosphorylated serines in position 2 (Ser2), a sign that Pol II is in a state of active transcription [29, 30]. Using semi-quantitative PCR, we verified enrichment of coding regions but not intergenic regions and low experimental noise in our ChIP procedure (Figure 2.2A).

Pol II occupancy at the early ring stage was extremely low as compared to other stages, in agreement with the lack of nascent RNA signal at this stage of the cell cycle (Figure 2.2B). Similar to the nascent gene expression profiles, the majority of genes showed the highest Pol II occupancy at the trophozoite stage, while a subset of genes was most highly occupied by Pol II at the schizont stage (Figure 2.2B). A metagenomic analysis showed that the average Pol II occupancy of genes with late schizont expression profiles in GRO-seq (cluster A9) is higher in schizonts as compared to genes in other clusters (Figure 2.2C). Finally, the expression patterns of the Pol II ChIP-seq and GRO-seq data sets are more similar than for either GRO-seq or Pol II ChIP-seq and a publicly available steady-state mRNA-seq data set [19] (Figure 2.2D). Together, these results validate our GRO-seq results and confirm that steady-state mRNA levels do not strictly reflect transcriptional activity, but may be subject to post-transcriptional processes, such as degradation and storage.

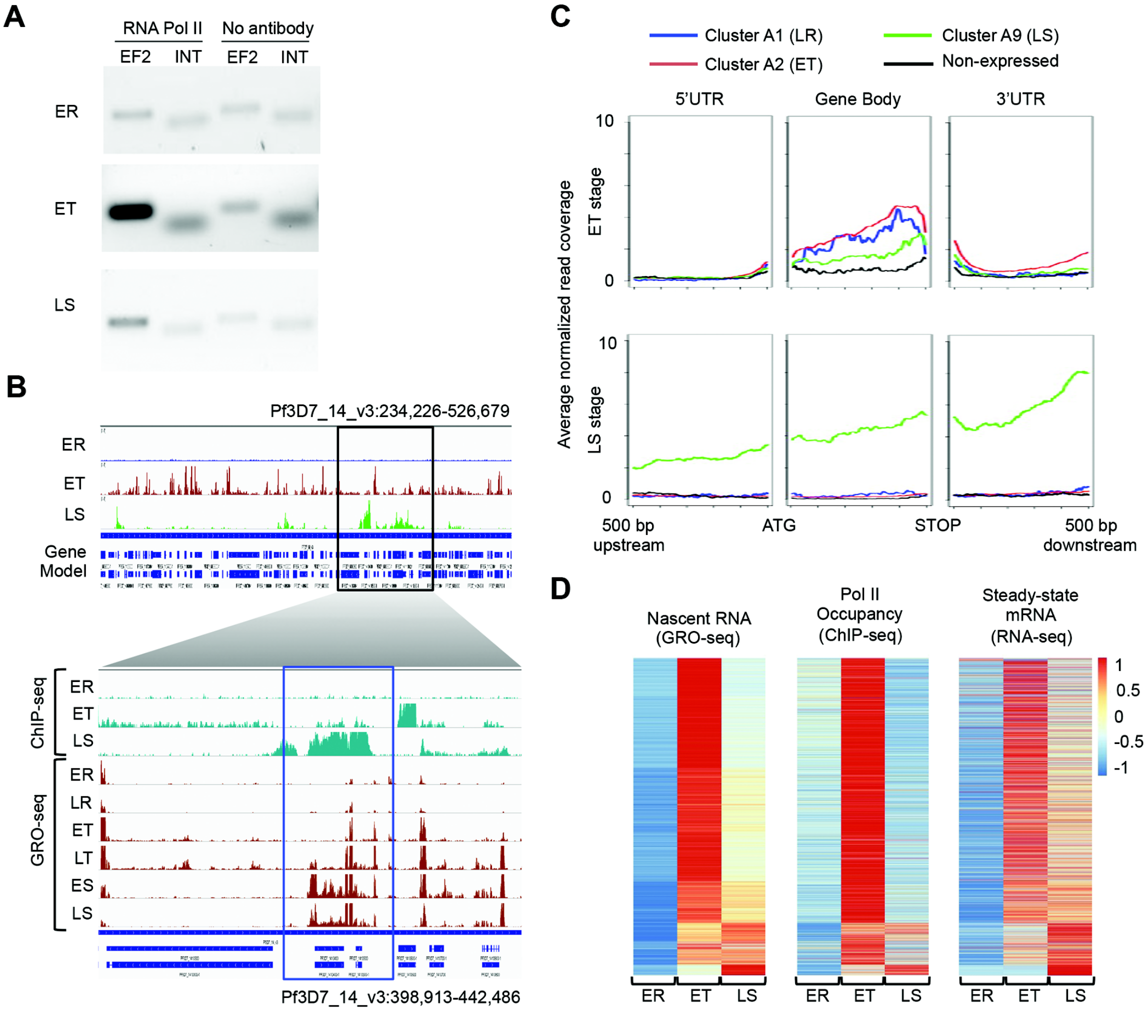


Figure 2.2. Confirmation of GRO-seq results by Pol II ChIP-seq. (A) Validation of Pol II ChIP using semi-quantitative PCR, showing enrichment of the coding region of elongation factor 2 (EF2) at the trophozoite and schizont stages as compared to an intergenic region (INT) and the negative (no antibody) control. **(B)** High level genome browser view (Pf3D7_14_v3:398,913-442,486) of normalized ChIP-seq data (top panel). The bottom panel shows a comparison between ChIP-seq (turquoise) and GRO-seq (dark red) data sets for a smaller region of chr14 as indicated by a black rectangle in the top panel. A rhoptry-associated membrane antigen gene (PF3D7_1410400, indicated by the blue rectangle) showed high expression at the schizont stage in both ChIP-seq and GRO-seq data sets. **(C)** Average Pol II ChIP-seq coverage plots of genes from GRO-seq ring-stage (A1), trophozoite-stage (A2), schizont-stage (A9) clusters, and non-expressed genes around gene start (ATG) and gene end. At the schizont stage, late schizont-stage genes (GRO-seq cluster A9) show higher Pol II occupancy in our ChIP-seq data set as compared to genes with other GRO-seq expression profiles, consistent with the GRO-seq results. **(D)** Comparison of gene expression profiles as observed in GRO-seq, Pol II Chip-seq, and RNA-seq [19] data sets, highlighting the discrepancies between transcriptional activity (measured by GRO-seq and ChIP-seq) and steady-state mRNA abundance. Genes (n = 4,888) are ranked in the same order in each heatmap. ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont.

Transcriptional activity in gametocytes

In early and late gametocytes, the most highly expressed genes encoded ribosomal proteins and proteins involved in movement and motor activity, while the bottom 20% of the genes were enriched for involvement in pathogenesis, erythrocyte remodeling, and antigenic variation (Supplemental Figure 2.3 and Supplemental Flie 2.2). To validate our results for the gametocyte stage, we examined 27 genes that have been shown to be essential for gametocytogenesis [7]. The majority of these genes (n=18, 67%) showed the highest transcriptional activity in the gametocyte stages, including the well-established gametocyte-specific makers *P. falciparum* gamete antigen 27/25 (PF3D7_1302100) and sexual stage-specific protein precursor Pfs16 (PF3D7_0406200) (Supplemental Figure 2.6). The remaining nine genes were most highly transcribed at one of the asexual stages, suggesting that the encoded proteins may play an essential role at the earlier stages of gametocytogenesis, but are not highly transcribed after early gametocyte differentiation. We also studied the transcriptional profiles for 686 homologs of *P. berghei* genes that were known to be transcribed at the gametocyte stage and are subject to translational repression by RNA-binding proteins DOZI and CITH [31, 32]. The majority of these genes were in the top 50% of transcriptional activity at either the early or the late gametocyte stage (n=537, 78%, p = 0.0001, Chi-square test), while 342 genes (50%, p = 0.0001, Chi-square test) were in the top 25% of transcriptional activity (Supplemental Flie 2.5), confirming that these genes are indeed active in gametocytes.

To further compare transcriptional activity between gametocyte stages and asexual stages, we calculated for each gene the fold change between the average nascent RNA abundance values of gametocyte stages and asexual stages, and subsequently divided genes into five groups based on this ratio (Supplemental Flie 2.1). Clusters B1 and B2 contain genes that show more than four-fold (n=403) or two-fold (n=1536) higher transcriptional activity in gametocytes than in the asexual stages, respectively. These groups showed enrichment for genes associated with motility (Figure 2.3 and Supplemental Flie 2.6), such as several genes encoding dynein subunits and actin-related proteins. Interestingly, many of these genes also showed transcriptional activity during the IDC, albeit at a lower level, but are not detected or present at much lower levels in steady-state mRNA [33] (Figure 2.3), suggesting that these transcripts may be degraded during the IDC when they are not needed. Cluster B3 contains a large group of genes (n=3107) for which transcriptional activity does not change by more than two-fold in the transition from IDC to gametocytes, indicating that overall, transcriptional programs of asexual parasites and gametocytes are not very different. This is also demonstrated by the relatively high correlation in GRO-seq data between the trophozoite and gametocyte stages (Spearman R 0.74 – 0.84, Supplemental Figure 2.1D). Finally, genes that were turned off in gametocytes as compared to the asexual stages were enriched for GO terms associated with pathogenesis and cell invasion, in line with our understanding of parasite biology (clusters B4 and B5; Figure 2.3).

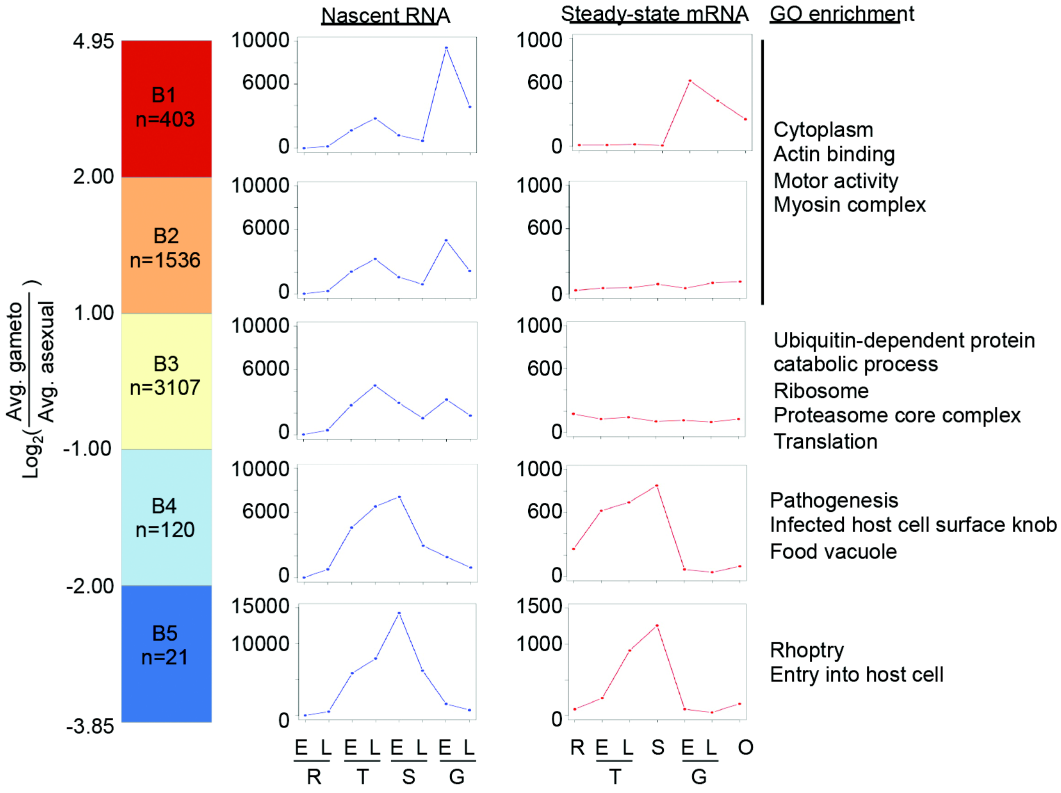


Figure 2.3. Differences in transcriptional profiles between asexual parasites and gametocytes. Genes were grouped based on the ratio between their average GRO-seq expression level in the asexual and in gametocytes. The average expression profiles for GRO-seq and RNA-seq [33] within each group are shown as blue and red lines, respectively. GO enrichment in each of the clusters is shown on the right. E, early; L, late; R, ring; T, trophozoite; S, schizont; G, gametocyte; O, ookinete.

Among the genes that were highly upregulated in gametocytes as compared to the IDC (clusters B1) were two AP2 transcription factors: the ookinete-specific transcription factor AP2-O and an AP2 TF with unknown function, PF3D7_1429200 (Supplemental Flie 2.1). Six out of eight CPW-WPC proteins that are involved in chromatin remodeling showed more than four-fold higher levels of transcription in gametocytes than in asexual stages. In addition, several mRNA-binding proteins were upregulated, including PUF1, PUF2, five RAP proteins with RNA-binding domains that are almost exclusively found in Apicomplexans, and putative RNA-binding protein PF3D7_0716000, which may be involved in posttranslational regulation and stabilization of more than 10% of the transcriptome that is known to occur in the transition from gametocytes to ookinetes [31, 32]. Finally, a relatively large fraction of genes in cluster B1 were conserved proteins with unknown function (n=199, 49.4%, p=0.0001, Fisher's exact test), indicating that we still lack a significant understanding of many of the parasite-specific processes that take place during sexual differentiation. To determine if a common transcription factor motif could be identified in promoters of the genes that were upregulated during gametocytogenesis, we performed a motif search on the 1,000 bp upstream of the annotated ATG of genes in clusters C1 and C2 using MEME. When performing the search on the total set of 1,939 genes, we identified motifs TGTDC and CATDCA, which both have overlap with previously identified AP2 TF binding motifs [34] (Supplemental Flie 2.7).

Nascent transcriptional activity and epigenetic landscape

In other eukaryotes, histone variants and activating histone post-translational modifications (PTMs) are associated with actively transcribed genes. In addition, nucleosome depletion around the transcription start site (TSS) has been shown to be associated with genes that are highly transcribed [11, 12, 35-38]. To further investigate mechanisms that contribute to regulation of transcriptional initiation, we therefore analyzed previously published H2A.Z, H3K9ac, H3K4me3 ChIP-seq datasets [39] and a nucleosome landscape MNase-seq dataset [11]. These data sets have previously been analyzed for correlation with steady-state mRNA abundance, but not with GRO-seq data. Similar to previous findings [39], no significant correlation was observed between transcriptional activity and H2A.z or H3K4me3 data sets (Supplemental Figure 2.7); however, a side-by-side comparison of GRO-seq and H3K9ac abundance heatmaps (Figure 2.4A) showed that genes with schizont-stage transcriptional activity tend to have higher H3K9ac marks at the later stage. Nucleosome occupancy in the 500bp upstream of the coding region is relatively high in ring and schizont-stage parasites, while global nucleosome depletion occurs in trophozoites and gametocytes (Figure 2.4A, Supplemental Figure 2.7, and Supplemental Flie 2.8), as described previously [11, 12]. Taken together, high levels of transcriptional activity at the trophozoite and gametocyte stages correlate well with an open chromatin structure. However, the subset of genes transcribed at the schizont stage does not seem to be regulated by the same changes in chromatin organization that occur at the trophozoite stage.

Nascent transcriptional activity and RNA polymerase II pausing

Recruitment of Pol II and the formation of the pre-initiation complex (PIC) are critical steps in gene activation and subject to strict regulation. However, evidence is emerging that Pol II can also be regulated at the level of early transcription elongation [27, 29, 40-45]. This elongation control is achieved by pausing of Pol II 30-50 nucleotides downstream of the promoter, and subsequently requires additional positive signals before elongation can be continued. Pol II pausing has previously been studied using GRO-seq data [27, 43, 46] and Pol II ChIP-seq data (mainly on Pol II with Serine 5 phosphorylated CTDs) [46, 47]. To find evidence for Pol II pausing in *P. falciparum*, we focused on the GRO-seq read coverage profiles in 5' UTR regions. In total, 60-70% of all GRO-seq reads mapped to intergenic regions, including the 5' UTRs (Supplemental Figure 2.8A). Similar to previously published GRO-seq datasets from other eukaryotes, we observed a peak in read coverage around the gene start (Supplemental Figure 2.8B) that was positively correlated with transcriptional activity (Supplemental Figure 2.8C). This pattern has been described to be the result of paused Pol II complexes that were activated during the nuclear run-on procedure. Next, we calculated a ratio, hereafter called pausing index, of sequence read density in the 5' UTR (defined as 500 bp upstream of ATG) to that in the gene body (500 bp downstream of ATG) for each gene at each developmental stage as described by

Core *et al.* [27]. We observed that the average of pausing indexes for all genes is highest at the ring stage, followed by a sharp decrease at the trophozoite stage and a small increase as the parasite enters the schizont stage (Figure 2.4B). These data suggest that Pol II is starting to be engaged with the genome in the ring stage, but is prevented from continuing transcriptional elongation until activation in the trophozoite stage, resulting in a global transcriptional burst. For genes with stage-specific transcription patterns, the Pol II pausing index was lowest in the stage with the highest transcriptional activity (Supplemental Figure 2.9A). The negative association between the Pol II pausing index and transcriptional activity are similar to observations in human cells and are believed to be indicative of transcriptional control, where the rate of Pol II engagement to the promoter is higher than the rate of Pol II entering its elongation phase.

In addition to the extended coverage at 5' UTR, we also observed a high read coverage at the 3' UTR regions, again similar to GRO-seq profiles from other eukaryotes (Supplemental Figure 2.9B). In yeast [48], Pol II pausing at the 3' UTR was found to be associated with splicing events. In *P. falciparum*, we observed that 3' UTR GRO-seq coverage was approximately 1.5-fold higher in multi-exon genes as compared to single exon genes at the late trophozoite stage (p= 1.283e-15, Mann-Whitney U test) (Supplemental Figure 2.9B). These results suggest that splicing may indeed contribute to Pol II pausing at the 3' UTR, but is unlikely to be the only event that triggers pausing of the transcriptional complex at this location.

Comparison between nascent RNA and steady-state mRNA abundance

To measure the degree in which transcriptional and post-transcriptional regulatory mechanisms contribute to global gene expression, we further clustered genes based on both nascent RNA and steady-state mRNA expression profiles, resulting in six distinct clusters (Figure 2.4C and Supplemental Flie 2.9). For this analysis, we only used the GRO-seq data that exactly matched the stages available in the steady-state mRNA data set [19]. Out of the 4,881 genes that changed in abundance during the cell cycle in both data sets, 2,503 genes (58%) showed nearly identical profiles for both data sets (clusters C2, C3, and C6, respectively). In contrast, cluster C4 contains 1,126 genes for which transcription peaked in late trophozoite stage, but that were continuously detected until the late schizont stage in steady-state mRNA, indicating that these transcripts undergo partial stabilization. This cluster showed enrichment for genes involved in protein metabolism (Supplemental Flie 2.9). In addition, cluster C5 (n=719) and C1 (n= 197) showed an even larger discrepancy between the moment of transcription and the time point of highest abundance in steady-state mRNA, suggesting that these transcripts are the subjects of strong post-transcriptional regulation. The small group of genes that was previously identified as ring-stage specific (cluster C1) enriched for involvement in erythrocyte remodeling also showed transcriptional activity at the trophozoite stage, suggesting that these transcripts might be transcribed later in the cell cycle and stored long-term until the next round of erythrocyte invasion.

For the clusters with genes that do not seem to be subject to post-transcriptional regulation and are most likely controlled at the level of transcription (clusters C2, C3, and C6), we mined the region upstream of the gene start for transcription factor binding motifs (Supplemental Flie 2.7). Motifs that almost exclusively contained A's or T's were identified for all clusters. Such stretches of AT could be reminiscent of the TATA box in higher eukaryotes, but were not further considered due to the high AT content of the *P. falciparum* genome. In clusters C2 and C3, we identified the short motifs STTC and SYTC, respectively. In addition, we observed enrichment for motifs GTG, GWG, and RTGT in clusters C3, C4, and C5, respectively. The reverse complements of these short motifs (CACACA and ACACAC) have previously been shown to be associated with DNA replication [49] and possibly interact with AP2 TFs PF3D7_0802100 and

PF3D7_1456000 [34]. In addition, genes in cluster C6 showed enrichment for motif GTGHA, which has previously been described as the binding sequence of AP2 TF PF3D7_1007700 and is associated with invasion genes [34, 49-51].

For genes in clusters showing evidence of post-transcriptional regulation (clusters C1, C4, and C5), we searched the regions upstream of the gene start and downstream of the gene stop for transcription factor or RNA-binding protein motifs. GTG and ACAC motifs were identified in the 500 bp downstream of the gene stop in cluster C5, similar to motifs identified in the 5'UTRs of clusters C3, C4, and C5.These results suggests that the GTG/CAC motif is very common and may be similar to frequently observed TATA stretches. Overall, the lack of specific motifs identified in our analyses emphasizes the need for developing novel experimental designs to discover mechanisms regulating transcripts at the transcriptional and post-transcriptional level.

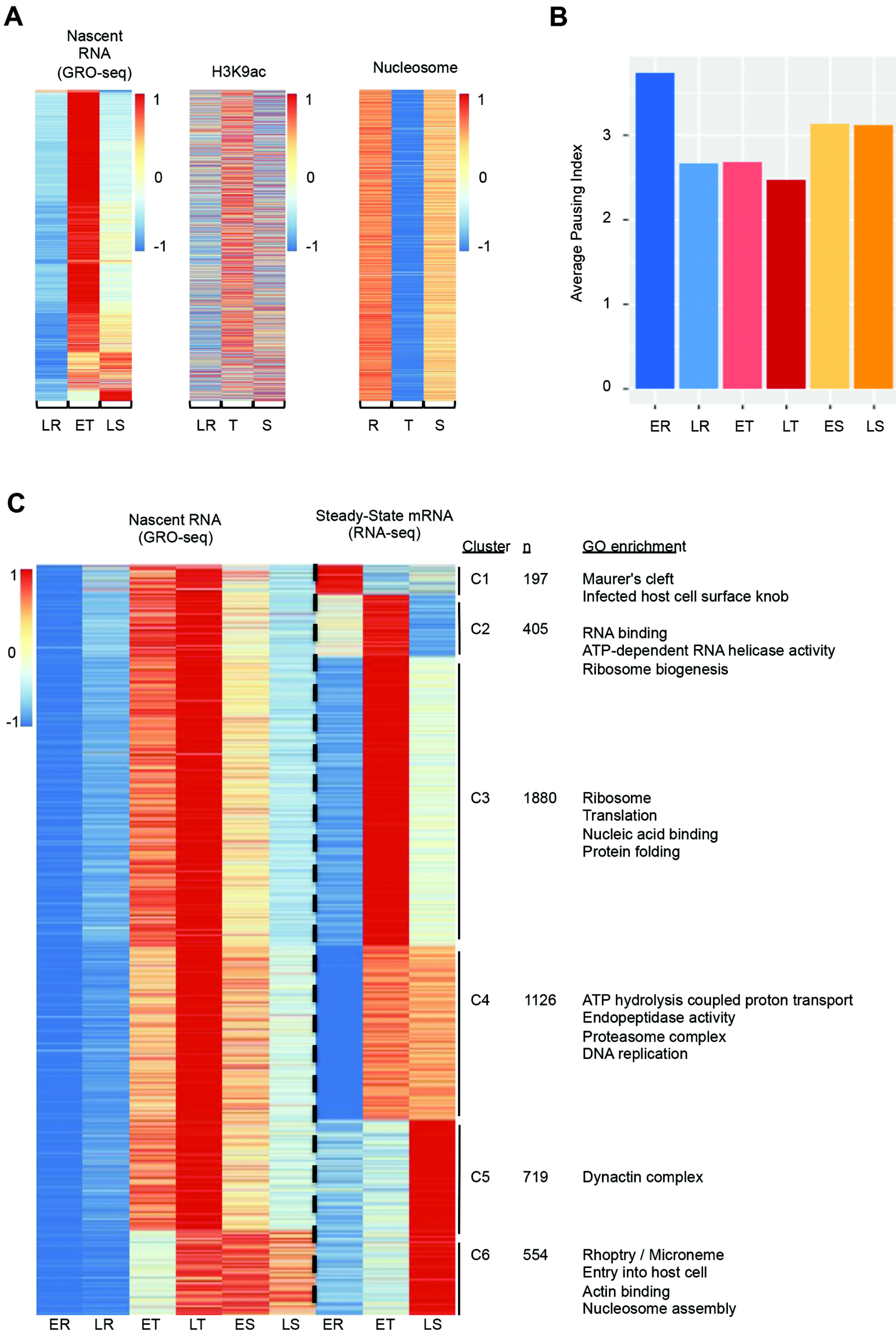

A
Nascent RNA (GRO-seq)
H3K9ac
Nucleosome
1
0
-1
LR ET LS
LR T S
R T S
B
Average Pausing Index
0
1
2
3
ER LR ET LT ES LS
ER
LR
ET
LT
ES
LS
C
Nascent RNA (GRO-seq)
Steady-State mRNA (RNA-seq)
1
0
-1
ER LR ET LT ES LS ER ET LS
Cluster n GO enrichment
C1 197 Maurer's cleft
Infected host cell surface knob
C2 405 RNA binding
ATP-dependent RNA helicase activity
Ribosome biogenesis
C3 1880 Ribosome
Translation
Nucleic acid binding
Protein folding
C4 1126 ATP hydrolysis coupled proton transport
Endopeptidase activity
Proteasome complex
DNA replication
C5 719 Dynactin complex
C6 554 Rhoptry / Microneme
Entry into host cell
Actin binding
Nucleosome assembly

Figure 2.4. Association of transcriptional activity with chromatin structure, Pol II pausing, and steady-state mRNA expression. (A) A comparison of transcriptional activity with H3K9ac abundance, and global nucleosome occupancy during the IDC. Genes were ranked according to transcriptional profile during the IDC in the same order as in Fig 1C. **(B)** Averaged Pol II pausing index for all genes in GRO-seq data at each stage. **(C)** Comparison of gene expression profiles in GRO-seq and RNA-seq. Nascent RNA and steady-state mRNA data sets were z-scored by gene individually and then clustered based on the combined data. Clusters C4, and C5 show large differences in the moment of peak transcript abundance between the two data sets, suggestive of strong post-transcriptional regulation. Enriched GO terms are indicated on the right.

Materials and Methods

Parasite culture

Parasite strain, *P. falciparum* 3D7, was cultured at approximately 8% parasitemia in human erythrocytes at 5% hematocrit in a total culture volume of 25 ml as described in [52]. To obtain highly synchronized cultures for the asexual stages, two 5% D-sorbitol treatments were performed eight hours apart at the ring stage. Parasites were collected every 6 hours covering early ring, late ring, early trophozoite, late trophozoite, early schizont, and late schizont stages. Giemsa-stained blood smears were used to assess parasite developmental stages. To obtain gametocyte-stage parasites, *P. falciparum* strain NF54 was cultured as described previously [53]. In brief, parasites were first synchronized using 5% sorbitol lysis buffer and diluted to reach 0.5% parasitemia at 8.3% hematocrit the following day. A reduction of culture media to 10 ml for three subsequent days was used as a way to stress the parasites and induce gametocyte production. Culture volume was then returned to 25 ml per flask. During the next five days, cultures were maintained by daily media exchange using media containing 10 ml of 50 mM N-acetyl glucosamine (NAG) to remove asexual stage parasites. Gametocyte cultures at 2% parasitemia were harvested 10 days (stage III) or 14 days (stage V) after the start of this procedure.

Nuclear isolation

Nuclear isolation was performed as described in [11, 54]. Parasite pellets were resuspended in 1 ml of nuclear extraction buffer (10 mM Tris-HCL pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 250 units of SUPERaseIn (Ambion), 10% glycerol, and 0.5% Igepal CA-360 (Sigma-Aldrich, St. Louis, MO)) and incubated on ice for 10 min. Parasites were then mechanically lysed by passing the suspension fifteen times through a 26G ½ inch needle. Nuclei were pelleted by centrifugation for 20 min at 2,500 x g at 4°C and resuspended in 1 ml of nuclear extraction buffer, followed by gently pipetting up and down 10 times. Nuclei were centrifuged for 20 min at 2,500 x g at 4°C and resuspended in 100 µl of storage solution (50 mM Tris-Cl pH 8, 5 mM $MgCl_2$, 0.1 mM EDTA, 40% glycerol, and 50 units of SUPERaseIn).

Nuclear Run-on Reaction

Nuclei (100 µl) were incubated with 600 µl of nuclear run-on reaction buffer (10 mM Tris-Cl pH 8.0, 5 mM $MgCl_2$, 1 mM DTT, 300 mM KCl, and 200 units of SUPERaseIn, 1% sarkosyl, 4 mM ATP, 1 mM CTP, 1 mM GTP, 200 mM Ethylene uridine (EU) (Click-it Nascent RNA Capture Kit, Thermo Fisher), 400 mM creatine phosphate, and 0.2 mg/mL creatine kinases) adopted and modified from [27, 28]. Reaction mixtures were incubated for 30 min at 37°C followed by nuclear RNA isolation.

Base hydrolysis of nuclear RNA

Base hydrolysis was performed as described in [27]. For each 20 µl of RNA, 5 µl of 1M NaOH is added and incubated for 15 minute on ice. The reaction was neutralized with 25 µl of 1M Tris-Cl pH 6.8. Fragmented RNA was precipitated by adding 4 µl glycogen, 75 µl 5 M ammonium acetate, and 700 µl 100% ethanol).

Nascent RNA purification and cDNA preparation

Nascent RNA was purified from total nuclear RNA samples using the Click-iT Nascent RNA Capture Kit (Thermo Fisher) according to the manufacturer's instructions. In brief, biotin-azide was attached to ethylene-groups of the EU-labeled RNA using click-it chemistry. The EU-labeled nascent RNA was purified using MyOne Streptavidin T1 magnetic Dynabeads (Life Technologies). The preparation of cDNA was performed using nascent RNA captured on the beads. cDNA synthesis reaction mix (6 µg of random hexamer (integrated DNA technologies, Coralville, IA, USA), 2 µg of anchored oligo $(dT)_{20}$ (Integrated DNA Technologies), 2 µl 10 mM dNTP mix (Life Technologies), and 14 µl Buffer J from Click-iT Nascent RNA Capture Kit (Thermo Fisher) in a total of 20 µl volume) was added to the beads and incubated for 10 min at 70°C, and then chilled on ice for 5 min . Next, a mix of 4 µl 10X RT buffer, 8 µl 20 mM $MgCl_2$, 4 µl 0.1 M DTT, 2 µl 20 U/µl SuperaseIn and 2 µl 200 U/µl SuperScript III Reverse Transcriptase (all from Life Technologies) was added to the mixture and incubated for 10 min at 25°C, 50 min at 50°C, and finally 5 min at 85°C for first strand cDNA synthesis. To digest RNA and release the first-strand cDNA, 2 µl 2 U/µl *E. coli* RNase H (Life Technologies) was added, followed by a 20 min incubation at 37°C. The beads then were removed using a magnet and first-strand cDNA was used for second-strand cDNA synthesis by adding 70 µl 5X nuclease-free water (Life Technologies), 30 µl second-strand buffer (Life Technologies), 3 µl 10 mM dNTP mix (Life Technologies), 4 µl 10 U/µl *E. coli* DNA Polymerase (NEB), and 1 µl 10 U/µl *E. coli*DNA ligase (NEB). The mixture was then incubated for 2 h at 16°C. Finally, double-stranded cDNA was purified using 1.8X Agencourt AMPure XP beads (Beckman Coulter). Validation PCRs were performed using the primers listed in Supplemental Flie 2.1.

Library preparation and sequencing

Libraries were prepared using the KAPA Biosystems Library Preparation Kit (KAPA Biosystems, Woburn, MA) according to the manufacturer’s instructions with the following modifications for the high AT-content of the *P. falciparum* genome: the libraries were amplified for 15 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 sequenced on the Illumina HiSeq2500 (Illumina, San Diego, CA) generating 50 bp paired-end sequence reads or the NextSeq500 generating 75 bp paired-end sequence reads.

Sequence mapping

The first ten bases and the last base or last 20 bases were systematically trimmed from 50 bp and 75 bp reads, respectively, using FastQ Trimmer from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Poly-A/T repeats and contaminating adaptor reads were removed using Scythe (https://github.com/ucdavis-bioinformatics/scythe)[55]. Reads containing bases with a quality score below 25 and Ns, reads that were unpaired, and reads shorter than 18 bases were also filtered using Sickle (https://github.com/najoshi/sickle) [56]. In addition, high quality single reads that lost their mate pair during read processing were kept and mapped as single-end reads in parallel with paired-end reads. All trimmed reads were first mapped to the human genome version hg19 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/) using Bowtie 2 [57], and all non-human reads were further mapped to *P. falciparum* 3D7 genome v13.0 (www.plasmoDB.org) using TopHat2 [58] allowing a maximum of one mismatch per read segment and a segment length of 18. Finally, reads that mapped to multiple locations in the genome (samtools v0.1.19), paired-end reads that were not properly paired (samtools v0.1.19), reads that were PCR duplicates (MarkDuplicates, Picard Tools v1.114), and reads that mapped to ribosomal RNA or transfer RNA were discarded from the final working reads.

Calculation of normalized gene expression values

Raw genome-wide coverage profiles were generated using BEDtools [59]. For each stage, numbers of mapped reads from both single-ended mapping and paired-end mapping were combined. For each gene, we then calculated the number of reads that mapped to its exons, and normalized these read counts by GC content and the sum of exon lengths using R package EDASeq [60]. Spearman correlations between biological replicates were calculated using the EDASeq normalized exon counts. For biological replicates that were highly correlated (Spearman

R > 0.85), bam files were merged using Samtools v0.1.19 [61] and the normalized exon count was recalculated. Genes with an average exon read count below 2 at all stages were considered not expressed and were removed from the data set. To accurately measure transcriptional activity at each stage, we normalized the exon read count to the amount of RNA yield per parasite. A stage-specific scaling factor was calculated for each library by dividing the total number of filtered reads by the amount of RNA extracted per 1 flask of parasite-infected culture. In addition, we corrected for differences in parasitemia by multiplying by a parasitemia factor that was calculated as the highest parasitemia in any stage divided by the parasitemia in the stage of interest. In order to compare libraries between stages, we re-standardized the normalized read counts of each library X relative to the normalized read counts of the library with the smallest number of filtered reads. All of these calculations were performed using the following equation:

$$\mathrm{scalingfactor}_{\mathrm{libraryX}} = \mathrm{parasitemiafactor} \times \frac{\mathrm{filteredreads}_{\mathrm{libraryX}} \div \mathrm{RNAyieldperflask}_{\mathrm{librayX}}}{\mathrm{filteredreads}_{\mathrm{smallestlibrary}} \div \mathrm{RNAyieldperflask}_{\mathrm{smallestlibrary}}}$$

The final abundance value of each gene was presented as the normalized exon read count per kilobase gene model divided by the scaling factor of that stage (Supplemental Flie 2.1). A gene with an abundance value below 15% of the median at all stages was considered not expressed and was discarded for further analysis.

Cluster and GO enrichment analysis

All genes that showed more than two-fold change in the normalized exon read count across the IDC stages were used for cluster analysis. For each gene, abundance values at the six different stages were z-scored, followed by k-means clustering in R (version 3.2.4) with a maximum of 1,000 iterations. The number of clusters used in this analysis was guided by the percent of variance captured (within group sum of squares). The optimal number of cluster was determined as the smallest number of clusters that captured at least 75% of the variance. For each cluster, gene ontology enrichment was performed using R package goseq [62]. All GO terms with a P-value < 0.001 were reported.

PCR validataion for GRO-seq cluster analysis

Four sets of primers were designed to amplify genes that are highly transcribed at the ring stage cluster (membrane associated histidine-rich protein, MAHRP1, PF3D7_1370300), trophozoite stage (trophozoite exported protein 1, TEX1, PF3D7_0603400), schizont stage (rhoptry-associated membrane antigen, RAMA, PF3D7_0707300), and a gene that did pass through filtration threshold (putative pyridine nucleotide transhydrogenase, PNT, PF3D7_1453500). PCR amplification was performed using cDNA library samples from early ring (ER), late ring (LR), late trophozoite (LT), and late schizont (LS) stage alone with a control sample with no DNA template (No Temp). As different amount of parasite was used for nascent RNA isolation at each stage, we first diluted each library sample parasite extracted from 2 culture flasks (approximately 20^9 parasites). All 4 PCRs were performed using 1 µl of the diluted cDNA library sample with approximately 10 pmole of both forward and reverse primers. DNA was incubated for 5 min at 95 °C, then 30 s at 98 °C, 30 s at 55 °C, 30 s at 62 °C for 35 cycles. 5 µl of each PCR sample was used for agarose gel electrophoresis. For each primer set, PCR efficiency was tested using genomic DNA under the same amplification conditions as described above. All primer used for PCR validation are listed in the Supplemental Flie 2.1.

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. The parasites were permeabilized with 0.1% Triton-X for 30 min at RT, followed by three washes 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). Slides were incubated for 1 hour at RT with anti-RNA polymerase II CTD phospho serine 2 antibody (Abcam ab5095; 1:250) followed by an incubation with donkey anti-rabbit dylight 550 antibody (Abcam ab98489; 1:500) for 1 hour at RT. Slides were washed with 1X PBS and mounted using Vectashield mounting medium with DAPI. Images were acquired using the Olympus BX40 epifluorescence microscope.

Chromatin immunoprecipitation

Synchronized parasite cultures were collected and subsequently lysed by incubating in 0.15% saponin for 10 min on ice. Parasites were centrifuged at 3,234 x g 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 3,234 x g at 4°C. Subsequently, parasites were crosslinked 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 crosslinking reaction, and incubated for 5 min at 37°C. Parasites were centrifuged for 5 min at 2,500 x g at 4°C, washed twice with cold PBS and stored at -80°C.

For chromatin immunoprecipitation, parasites were first 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 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (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 26 G ½ inch needle seven times. Parasite nuclei were centrifuged at 4°C for 20 min at 2,500 x g. 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, and phosphatase inhibitor cocktail). Chromatin was fragmented using the Covaris Ultra Sonicator

(S220) for 10 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 17,000 x g 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 and phosphatase inhibitor cocktail). Samples were precleared with Protein A Agarose beads to reduce non-specific background and incubated overnight at 4°C with 2 µg of anti-RNA pol II antibody (ab5095, Abcam). 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 was 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 $NaHCO_3$) for 15 min at RT. Samples were reverse crosslinked 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 1.8X Agencourt AMPure XP Beads (Beckman Coulter). Validation PCRs were performed using the primers listed in Supplemental Flie 2.4.

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 on the Illumina NextSeq500. Read coverage mapping to exonic regions were calculated for both positive and negative libraries, then normalized by dividing these numbers with million number of reads for each library. Finally, the signals obtained from the negative controls were subtracted from the ChIP-Seq library of the same stage.

Histone variants and nucleosome occupancy analysis

Sequence read files of MNase-digested chromatin (input), H2A.Z, H3K4me3, and H3K9ac ChIP-seq data sets (GSE23787) [39] and a nucleosome occupancy data set (SRP026365) [11] were downloaded from NCBI Sequence Read Archive. For H2A.Z, H3K4me3, K3K9ac, and input data sets, reads were mapped directly to *P. falciparum* 3D7 genome v13.0 (www.plasmoDB.org) using Bowtie 2 [57]. Non-uniquely mapped reads and PCR duplicates were discarded from final working reads. Coverage depth was first normalized to million mapped reads and was expressed as the ratio between sample and input. The nucleosome occupancy data set was mapped and normalized as described in the original publication [11]. The normalized read coverage in the 500 bp upstream of the annotated ATG was calculated and used to generate heatmaps using the command pheatmap in R.

Motif Identification

Two motif-discovery programs were used in this study to identify over-represented DNA motifs upstream of gene start site (ATG) of genes within each nascent cluster. A region of 1,000 bp upstream of the coding region was used for this search, based on the reported distribution of transcription start sites [63] (81% of TSS are located within 1,000 bp of the coding region). Genes that were located less than 1,000 bp from their 5’ neighboring gene were removed from the analysis. MEME-ChIP runs two de novo motif identification algorithms, MEME and DREME. The parameters used for MEME algorithm were minw=7, maxw=12 in zoops mode with E-value

<= 1e-01. The parameters used for DREME and CentriMo, a motif enrichment analysis algorithm included in the MEME-ChIP package, were the default parameters:-dreme-e 0.05 -centrimo-score 5.0 -centrimo-ethresh 10.

Discussion and Conclusion

Over the past few years, various studies have analyzed steady-state mRNA abundance throughout the blood stages of malaria parasites [18-21, 33]. However, steady-state mRNA is the product of transcription, stability, and degradation, and may therefore not accurately reflect transcriptional activity. Here, we present the first genome-wide study that specifically measures the timing and level of transcription during the *P. falciparum* asexual and sexual blood stages by performing global run-on sequencing. This data set is an invaluable asset towards a better understanding of gene regulation at the transcriptional and post-transcriptional levels during the life cycle of *P. falciparum*.

The results of our study suggest that transcriptional activity at the ring stage is limited to a small subset of genes encoding erythrocyte-remodeling proteins. However, at this stage of the life cycle, Pol II is already engaged with nearly every promoter in the genome, waiting for an activation signal that initiates a massive burst of transcription at the trophozoite stage. In line with this model, two general transcription factors, PfTBP and PfTFIIE, were previously shown to interact with both active and inactive promoters at the ring stage [64]. Once transcriptional elongation commences at the trophozoite stage, a large proportion of the genome is transcribed in agreement with a study that used the nuclear run-on methodology on individual *P. falciparum* genes [28]. This massive transcription event seems to be somewhat “leaky”, resulting in low-level transcription of genes that are specific for other life cycle stages of the parasite. Since these non-IDC genes are typically not detected in steady-state mRNA, we conclude that their transcripts may be quickly degraded. At the schizont stage, transcription is turned down, except for a subset of invasion-related genes that show upregulation of transcriptional activity, in agreement with a previous ChIP-on-ChIP analysis of Pol II [17]. Finally, as the parasite differentiates into a gametocyte, the transcriptional program remains largely unchanged as compared to the trophozoite stage with some exceptions, including invasion genes that are turned off and motility-related genes that are turned on.

The only genes for which we have been able to confirm an AP2 TF motif are the invasion genes, and together with ring-stage specific genes and virulence genes, these are the only genes that seem to be differentially regulated during the IDC. For the invasion genes, mechanisms that control gene expression include the binding of a specific transcript factor to the promoter region and the attachment of a bromodomain protein, PfBDP1, to acetylated histone H3 [65]. Virulence genes, in particular *var* genes are controlled by a combination of repressive histone modifications, long non-coding RNAs and localization away from the rest of the genome in perinuclear heterochromatin. In contrast, the massive transcriptional events at both the trophozoite and gametocyte stages are associated with a genome-wide depletion of nucleosomes [11, 13]. In addition, at the trophozoite stage, the promoters of the majority of genes are marked by activating histone PTM H3K9ac. Taken together, these data suggest that a large part of the genome is not regulated by classical eukaryotic mechanisms of transcription initiation that involve local chromatin changes and the presence of specific transcription factors that drive expression of a subset of genes. Instead, the majority of promoters are occupied by paused Pol II, activation of which coincides with genome-wide changes in chromatin structure, including nucleosome depletion [11, 13], increased chromosomal intermingling [11], nuclear expansion and an increase in the number of nuclear pores [66]. In this model of all-at-once transcription, there is no need for a large array of specific transcription factors and corresponding motifs, which may explain why a larger set of specific transcription factors has remained elusive in *Plasmodium* spp. to date. This lack of a need for finely tuned transcriptional activity may also explain how the parasite can quickly divide and form 16 – 32 daughter cells in a relatively short time frame (<12 hours).

The global Pol II pausing that takes place at the ring stage prior to massive transcriptional activity at the trophozoite stage may function as checkpoint before transcription elongation. In metazoans, the release of paused Pol II is mediated by phosphorylation of various proteins, including DRB sensitivity-inducing factor (DSIF, consisting of subunits SPT4 and SPT5), negative elongation factor (NELF), and the carboxyl terminal domain of the large subunit of Pol II at Ser2, by positive transcription elongation factor-b (P-TEFb) complex [67-69]. Inhibition of mammalian P-TEFb results a nearly complete block of transcription, suggesting that most active genes experience pausing events that require P-TEFb for elongation activation [43, 44, 47]. These results indicate that P-TEFb is a key regulator for transcription. In *Plasmodium*, many of the critical regulators, such as subunits of P-TEFb, DSIF subunits, and NELF, involved in Pol II pausing have been identified. The major P-TEFb subunits are cyclin-dependent kinase 9 (CDK9) and cyclin proteins (T1, T2 and K). Four cyclin genes have been described in *P. falciparum* (PfCYC1-4) [70, 71], of which only PfCYC4 shows homology to human cyclin T1, T2, and K. In addition, CDK9 is homologous to several parasite kinases, of which CDC2-related protein kinase 1 (PfCRK1) and protein kinase 5 (PfPK5) show the strongest similarity (Blastp E-value $<10^{-66}$). Studies in higher eukaryotes suggest that the nucleosome landscape, such as the positioning of the +1 nucleosome, could play a regulatory role in pausing by providing an energy barrier for elongating Pol II [72, 73]. The most strongly positioned nucleosomes in *P. falciparum* are at the start of the coding regions and could act as a barrier for RNA Pol II pausing. Furthermore, Pol II pausing and releasing have also been linked to nascent RNA hairpin structure, RNAs transcribed from enhancers [74, 75], promoter elements, and template DNA motifs, such as the downstream promoter element (DPE), TATA box, and GAGA motif [74, 76-82]. Unfortunately, due to low sequence homology and AT-richness of the *P. falciparum* genome, many regulatory mechanisms involved in RNA pol II regulation and pausing, such as enhancers, mediators, chromatin modifiers, and promoter elements, remain undefined. In addition, compared to mammalian polymerases that contain a C-terminal domain (CTD) with 52 identical heptad repeats, the *Plasmodium* CTD tail of Pol II displays wide variation in terms of length and composition [83, 84]. For example, primate malaria parasites, such as *P. knowlesi*, *P. vivax*, *P. falciparum*, and *P. cynomolgi* have an increased number of heptads with a high level of variability as compared to malaria parasites infect other species [85]. Additional work will be needed to truly understand how Pol II pausing is established and controlled in *Plasmodium* parasites. However, the present work established that while more classical regulatory mechanisms of transcription only control subsets of genes, such as invasion genes or *var* genes, the activation of paused Pol II complexes appears to be an essential genome-wide event during the IDC in *P. falciparum*. The identification of compounds that can specifically inhibit the activity of P-TEFb in the parasite will be a potentially powerful approach towards novel highly effective antimalarial drugs.

The large transcriptional activity that we observed here at the trophozoite and early schizont stages is in partial disagreement with the cascade of transcript abundance that has been observed in steady-state mRNA data sets [18-21, 33], but can be explained by a role for post-transcriptional gene regulation. In mature gametocytes, translational latency has been well documented and entails the temporary storage of hundreds of transcripts in ribonucleoprotein complexes of female gametocytes until translation once the parasite has developed into an ookinete inside the mosquito. Evidence is emerging that similar mechanisms control subsets of genes during the IDC. For example, Vembar *et al.* showed the targeted capture and stabilization of transcripts encoding invasion-related proteins by PfAlba1, followed by release and translation of these proteins at a later time point during the cell cycle [86]. Interestingly, our study shows that invasion-related genes are also regulated at the level of transcription initiation, suggesting that a single group of genes can be subject to regulation at multiple levels. We recently reported that *P. falciparum* encodes a relatively large number of RNA-binding proteins, and that many of the known translational regulators that act during other stages of the life cycle, such as DOZI and CITH, are also associated with mRNA during the IDC [87]. These results are indicative of widespread control of gene expression at the post-transcriptional level. In addition, post-transcriptional gene regulation could also explain how the similar transcriptional profiles in trophozoites and gametocytes can give rise to widely different cell types.

This study reveals for the first time the transcriptional activity of genes during the intraerythrocytic developmental cycle and gametocyte differentiation. Our main findings are that (1) most genes are actively transcribed at the trophozoite stage, (2) the transcriptional profile of gametocytes is surprisingly similar to trophozoites with the exception of downregulation of invasion genes and upregulation of genes related to motor activity, and (3) Pol II pausing acts as major control mechanism during the IDC, halting transcriptional elongation in the ring stage and once lifted, giving rise to the transcriptional burst at the trophozoite stage. Together, these results provide a much-needed increase in our understanding of *P. falciparum* biology and suggest that proteins involved in transcriptional elongation may be highly effective targets for anti-malarial therapy.

Reference

1. 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.

2. WHO: **World Malaria Report. 2015.** http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/**.** 2015.

3. Wright GJ, Rayner JC: **Plasmodium falciparum erythrocyte invasion: combining function with immune evasion.** *PLoS Pathog* 2014, **10:**e1003943.

4. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, Waters AP, Cowman AF, Crabb BS, de Koning-Ward TF: **Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites.** *Nucleic Acids Res* 2009, **37:**3788-3798.

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

6. Painter HJ, Campbell TL, Llinas M: **The Apicomplexan AP2 family: integral factors regulating Plasmodium development.** *Mol Biochem Parasitol* 2011, **176:**1-7.

7. Josling GA, Llinas M: **Sexual development in Plasmodium parasites: knowing when it's time to commit.** *Nat Rev Microbiol* 2015, **13:**573-587.

8. 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.

9. 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.

10. Ay F, Bunnik EM, Varoquaux N, Vert JP, Noble WS, Le Roch KG: **Multiple dimensions of epigenetic gene regulation in the malaria parasite Plasmodium falciparum: Gene regulation via histone modifications, nucleosome positioning and nuclear architecture in P. falciparum.** *Bioessays* 2015, **37:**182-194.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. Saraf A, Cervantes S, Bunnik EM, Ponts N, Sardiu ME, Chung DD, 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.

17. Rai R, Zhu L, Chen H, Gupta AP, Sze SK, Zheng J, Ruedl C, Bozdech Z, Featherstone M: **Genome-wide analysis in Plasmodium falciparum reveals early and late phases of RNA polymerase II occupancy during the infectious cycle.** *BMC Genomics* 2014, **15:**959.

18. 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.

19. 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.

20. 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.

21. 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.

22. Sorber K, Dimon MT, DeRisi JL: **RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts.** *Nucleic Acids Res* 2011, **39:**3820-3835.

23. Foth BJ, Zhang N, Chaal BK, Sze SK, Preiser PR, Bozdech Z: **Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite Plasmodium falciparum.** *Mol Cell Proteomics* 2011, **10:**M110 006411.

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

25. 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.

26. 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.

27. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.** *Science* 2008, **322:**1845-1848.

28. Sims JS, Militello KT, Sims PA, Patel VP, Kasper JM, Wirth DF: **Patterns of gene-specific and total transcriptional activity during the Plasmodium falciparum intraerythrocytic developmental cycle.** *Eukaryot Cell* 2009, **8:**327-338.

29. Jonkers I, Lis JT: **Getting up to speed with transcription elongation by RNA polymerase II.** *Nat Rev Mol Cell Biol* 2015, **16:**167-177.

30. Bowman EA, Kelly WG: **RNA polymerase II transcription elongation and Pol II CTD Ser2 phosphorylation: A tail of two kinases.** *Nucleus* 2014, **5:**224-236.

31. Guerreiro A, Deligianni E, Santos JM, Silva PA, Louis C, Pain A, Janse CJ, Franke-Fayard B, Carret CK, Siden-Kiamos I, Mair GR: **Genome-wide RIP-Chip analysis of translational repressor-bound mRNAs in the Plasmodium gametocyte.** *Genome Biol* 2014, **15:**493.

32. 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.

33. 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.

34. 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.

35. Mellor J: **Dynamic nucleosomes and gene transcription.** *Trends Genet* 2006, **22:**320-329.

36. Nocetti N, Whitehouse I: **Nucleosome repositioning underlies dynamic gene expression.** *Genes Dev* 2016, **30:**660-672.

37. Voss TC, Hager GL: **Dynamic regulation of transcriptional states by chromatin and transcription factors.** *Nat Rev Genet* 2014, **15:**69-81.

38. 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.

39. 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.

40. Adelman K, Lis JT: **Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.** *Nat Rev Genet* 2012, **13:**720-731.

41. Gaertner B, Zeitlinger J: **RNA polymerase II pausing during development.** *Development* 2014, **141:**1179-1183.

42. Kwak H, Fuda NJ, Core LJ, Lis JT: **Precise maps of RNA polymerase reveal how promoters direct initiation and pausing.** *Science* 2013, **339:**950-953.

43. Jonkers I, Kwak H, Lis JT: **Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons.** *Elife* 2014, **3:**e02407.

44. Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC, Adelman K: **Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals.** *Mol Cell* 2013, **52:**517-528.

45. Min IM, Waterfall JJ, Core LJ, Munroe RJ, Schimenti J, Lis JT: **Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells.** *Genes Dev* 2011, **25:**742-754.

46. Williams LH, Fromm G, Gokey NG, Henriques T, Muse GW, Burkholder A, Fargo DC, Hu G, Adelman K: **Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks.** *Mol Cell* 2015, **58:**311-322.

47. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA: **c-Myc regulates transcriptional pause release.** *Cell* 2010, **141:**432-445.

48. Carrillo Oesterreich F, Preibisch S, Neugebauer KM: **Global analysis of nascent RNA reveals transcriptional pausing in terminal exons.** *Mol Cell* 2010, **40:**571-581.

49. 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.

50. Iengar P, Joshi NV: **Identification of putative regulatory motifs in the upstream regions of co-expressed functional groups of genes in Plasmodium falciparum.** *BMC Genomics* 2009, **10:**18.

51. Essien K, Stoeckert CJ, Jr.: **Conservation and divergence of known apicomplexan transcriptional regulons.** *BMC Genomics* 2010, **11:**147.

52. Trager W, Jensen JB: **Human malaria parasites in continuous culture.** *Science* 1976, **193:**673-675.

53. Ifediba T, Vanderberg JP: **Complete in vitro maturation of Plasmodium falciparum gametocytes.** *Nature* 1981, **294:**364-366.

54. Core LJ, Lis JT: **Transcription regulation through promoter-proximal pausing of RNA polymerase II.** *Science* 2008, **319:**1791-1792.

55. Buffalo V: **Scythe - A very simple adapter trimmer** pp. Scythe - A very simple adapter trimmer 2011:Scythe - A very simple adapter trimmer

56. Joshi NA FJ: **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files** pp. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files 2011:Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

57. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

58. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.

59. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

60. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-Seq data.** *BMC Bioinformatics* 2011, **12:**480.

61. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

62. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol* 2010, **11:**R14.

63. Gissot M, Refour P, Briquet S, Boschet C, Coupe S, Mazier D, Vaquero C: **Transcriptome of 3D7 and its gametocyte-less derivative F12 Plasmodium falciparum clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors.** *Gene* 2004, **341:**267-277.

64. Gopalakrishnan AM, Nyindodo LA, Ross Fergus M, Lopez-Estrano C: **Plasmodium falciparum: Preinitiation complex occupancy of active and inactive promoters during erythrocytic stage.** *Exp Parasitol* 2009, **121:**46-54.

65. 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.

66. 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.

67. Peterlin BM, Price DH: **Controlling the elongation phase of transcription with P-TEFb.** *Mol Cell* 2006, **23:**297-305.

68. Zhou Q, Li T, Price DH: **RNA polymerase II elongation control.** *Annu Rev Biochem* 2012, **81:**119-143.

69. Lis JT, Mason P, Peng J, Price DH, Werner J: **P-TEFb kinase recruitment and function at heat shock loci.** *Genes Dev* 2000, **14:**792-803.

70. Le Roch K, Sestier C, Dorin D, Waters N, Kappes B, Chakrabarti D, Meijer L, Doerig C: **Activation of a Plasmodium falciparum cdc2-related kinase by heterologous p25 and cyclin H. Functional characterization of a P. falciparum cyclin homologue.** *J Biol Chem* 2000, **275:**8952-8958.

71. Merckx A, Le Roch K, Nivez MP, Dorin D, Alano P, Gutierrez GJ, Nebreda AR, Goldring D, Whittle C, Patterson S, et al: **Identification and initial characterization of three novel cyclin-related proteins of the human malaria parasite Plasmodium falciparum.** *J Biol Chem* 2003, **278:**39839-39850.

72. Li J, Gilmour DS: **Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor.** *EMBO J* 2013, **32:**1829-1841.

73. Weber CM, Ramachandran S, Henikoff S: **Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase.** *Mol Cell* 2014, **53:**819-830.

74. Liu X, Kraus WL, Bai X: **Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways.** *Trends Biochem Sci* 2015, **40:**516-525.

75. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EE: **Enhancer loops appear stable during development and are associated with paused polymerase.** *Nature* 2014, **512:**96-100.

76. Artsimovitch I, Landick R: **Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals.** *Proc Natl Acad Sci U S A* 2000, **97:**7090-7095.

77. Toulme F, Mosrin-Huaman C, Artsimovitch I, Rahmouni AR: **Transcriptional pausing in vivo: a nascent RNA hairpin restricts lateral movements of RNA polymerase in both forward and reverse directions.** *J Mol Biol* 2005, **351:**39-51.

78. Gaertner B, Johnston J, Chen K, Wallaschek N, Paulson A, Garruss AS, Gaudenz K, De Kumar B, Krumlauf R, Zeitlinger J: **Poised RNA polymerase II changes over developmental time and prepares genes for future expression.** *Cell Rep* 2012, **2:**1670-1683.

79. Amir-Zilberstein L, Ainbinder E, Toube L, Yamaguchi Y, Handa H, Dikstein R: **Differential regulation of NF-kappaB by elongation factors is determined by core promoter type.** *Mol Cell Biol* 2007, **27:**5246-5259.

80. Hendrix DA, Hong JW, Zeitlinger J, Rokhsar DS, Levine MS: **Promoter elements associated with RNA Pol II stalling in the Drosophila embryo.** *Proc Natl Acad Sci U S A* 2008, **105:**7762-7767.

81. Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K: **Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila.** *Science* 2010, **327:**335-338.

82. Lee C, Li X, Hechmer A, Eisen M, Biggin MD, Venters BJ, Jiang C, Li J, Pugh BF, Gilmour DS: **NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila.** *Mol Cell Biol* 2008, **28:**3290-3300.

83. Kishore SP, Perkins SL, Templeton TJ, Deitsch KW: **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* 2009, **68:**706-714.

84. Ukaegbu UE, Deitsch KW: **The Emerging Role for RNA Polymerase II in Regulating Virulence Gene Expression in Malaria Parasites.** *PLoS Pathog* 2015, **11:**e1004926.

85. Yamaguchi Y, Inukai N, Narita T, Wada T, Handa H: **Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA.** *Mol Cell Biol* 2002, **22:**2918-2927.

86. 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.

87. 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.

Supplemental Information

List of abbreviations

IDC: intraerythrocytic developmental cycle; TF: transcription factor; Pol II: RNA polymerase II; GRO-seq: global run-on sequencing; ChIP-seq: Chromatin Immunoprecipitation Sequencing; GO: Gene Ontology; PIC: Pre-initiation complex; CDS: Coding sequence; UTR: Untranslated regions

Availability of data and material

GRO-seq and ChIP-seq datasets generated and analyzed during the current study are available in the NCBI Gene Expression Omnibus (GEO) under accession number GSE85478. Supplemental materials and methods are included as additional files.

Funding

This study was financially supported by the National Institutes of Health (grants R01 AI85077-01A1 and R01 AI06775-01 to KGLR), the National Science Foundation (grant IIS-1302134 to KGLR), the University of California, Riverside (NIFA-Hatch-225935 to KGLR) and The University of Texas Health Science Center at San Antonio (EMB). The funding bodies had no role in the design of the study, in collection, analysis, and interpretation of data, or in writing the manuscript.

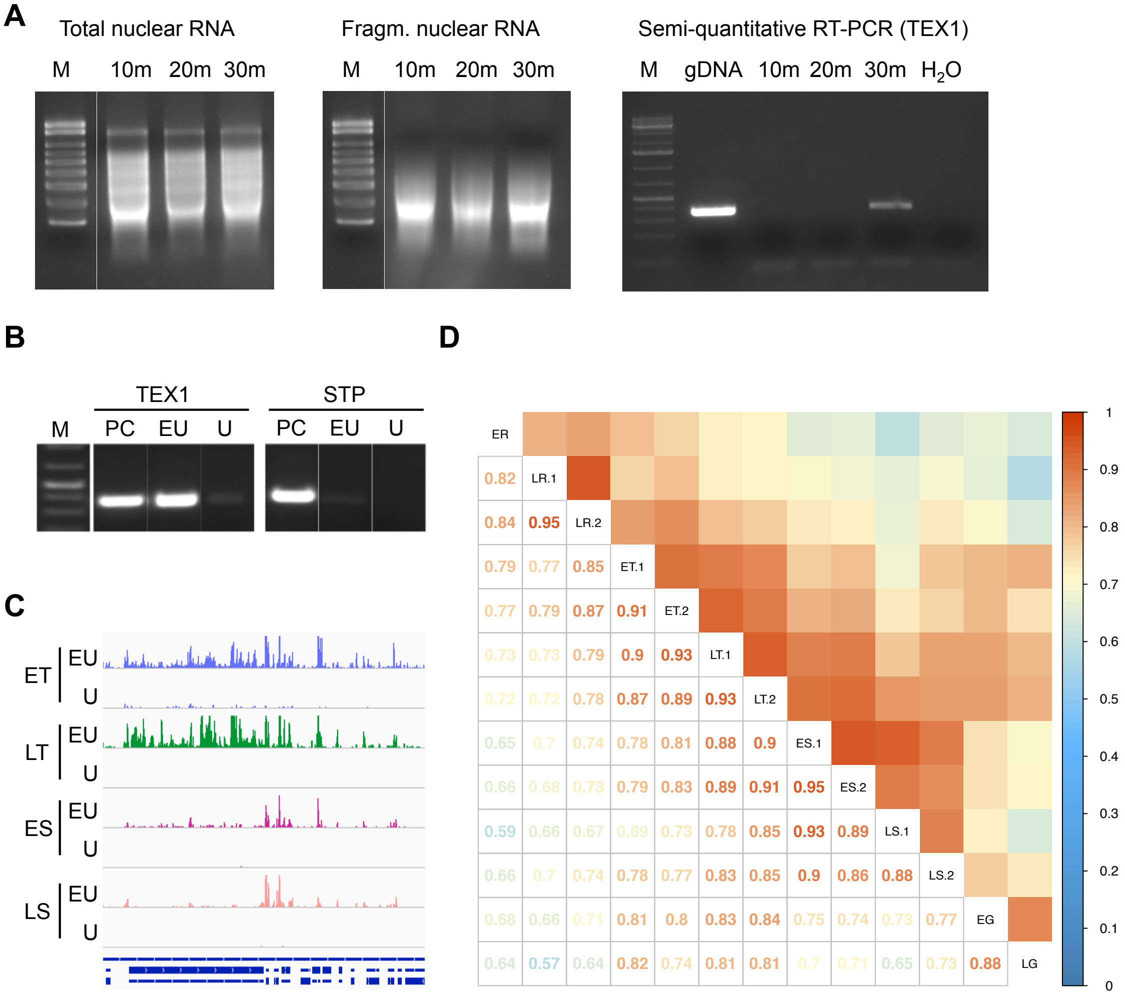
Acknowledgements

The following reagents were obtained through the MR4 as part of the BEI Resources Repository, NIAID, NIH: Plasmodium falciparum strains 3D7 (MRA-102) deposited by D.J. Carucci and NF54 (MRA-1000) deposited by Megan Dowler, Walter Reed Army Institute of Research.

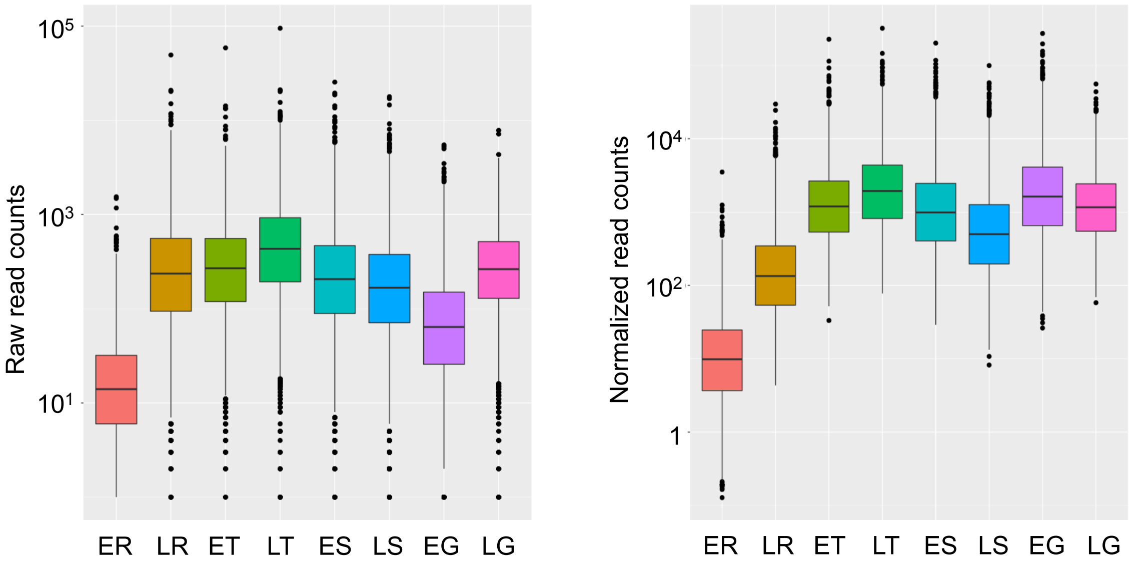
Authors' contributions

XML performed all computational analyses, generated GRO-seq datasets, participated in study design and drafted the manuscript. GB generated ChIP-seq datasets and assisted in experimental procedures; ML assisted in experimental procedures; JP maintained *P. falciparum* cultures and assisted in experimental procedures; EMB and KGLR participated in design of the study, critical discussion, supervised the project, and drafted the manuscript; KGLR was responsible for funding acquisition. All authors read and approved the final manuscript.

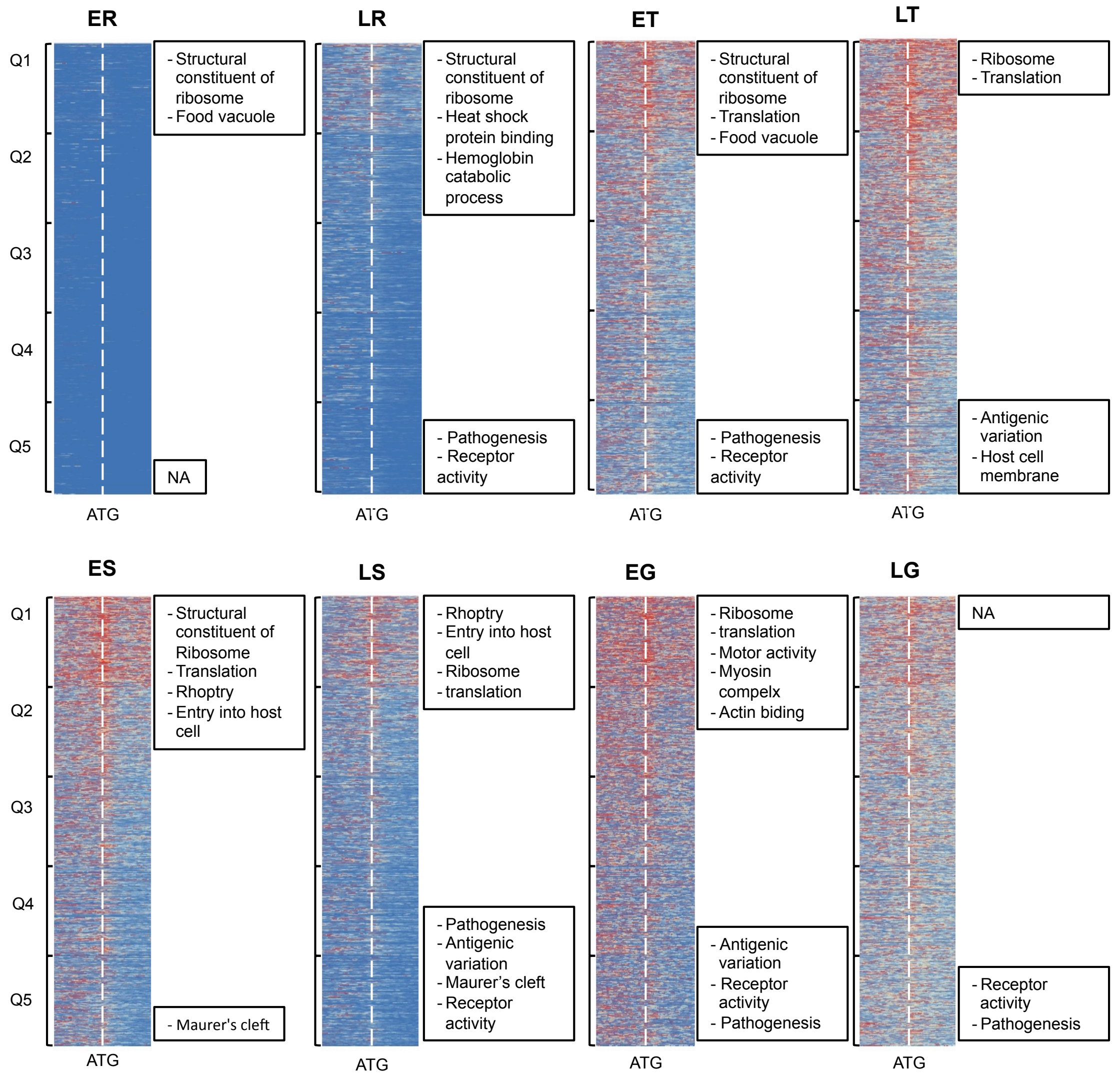
Supplemental Figures



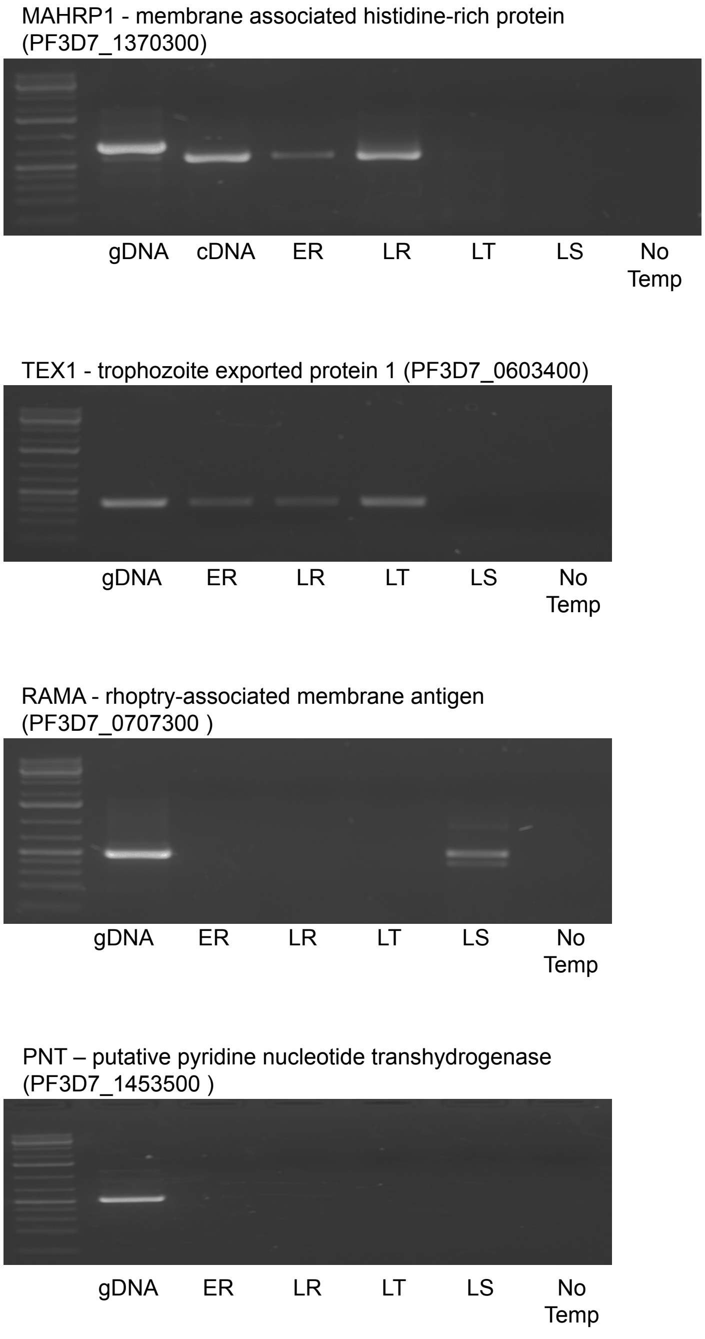
Supplemental Figure 2.1: Optimization and validation of GRO-seq methodology. (A) Optimization of the incubation times for the nuclear run-on reaction using late-stage trophozoites. Nuclear run-on was performed for 10, 20, or 30 minutes, followed by isolation of total nuclear RNA (left panel), fragmentation of total nuclear RNA (middle panel), and semi-quantitative RT-PCR on the TEX1 gene that is known to be expressed during the trophozoite stage (PF3D7_0603400; right panel). (B) Semi-quantitative PCR on early trophozoites showed strong enrichment of TEX1, while the sporozoite-specific gene STP (PF3D7_0107600) was not detected. For each gene, a positive PCR control (PC, using gDNA) is shown, as well as a regular nuclear run-on sample obtained using labeled uridine (EU) and a negative control nuclear run-on sample obtained using unlabeled uridine (U). Minimal signal of either gene was observed in the negative control samples. (C) IGV genome browser view of GRO-seq samples (EU) and their corresponding control samples (U), at early (E) and late (L) trophozoite (T) and schizont (S) stages, demonstrating the absence of signal in the negative controls. (D) Reproducibility of GRO-seq methodology in *P. falciparum*. Spearman correlations between gene abundance values of all GRO-seq samples were calculated before normalization. Note the relatively high correlation between trophozoite and gametocyte samples. M, marker; NC, no template control.



Supplemental Figure 2.2. The distribution of gene transcription values at all stages for raw (left panel) and normalized data (right panel).



Supplemental Figure 2.3. GRO-Seq expression profiles during the blood stages. For each stage, genes were sorted based on gene expression level from highest to lowest, and were divided into five equally sized groups. GO enrichment results are presented for the top and the bottom group from each stage (see Supplemental File 2.2 for full GO enrichment results).



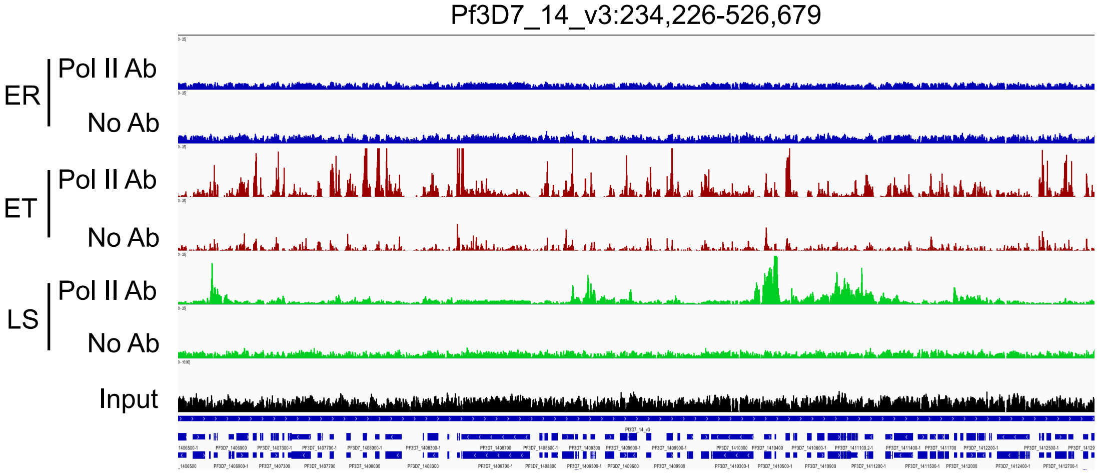
MAHRP1 - membrane associated histidine-rich protein (PF3D7_1370300)

TEX1 - trophozoite exported protein 1 (PF3D7_0603400)

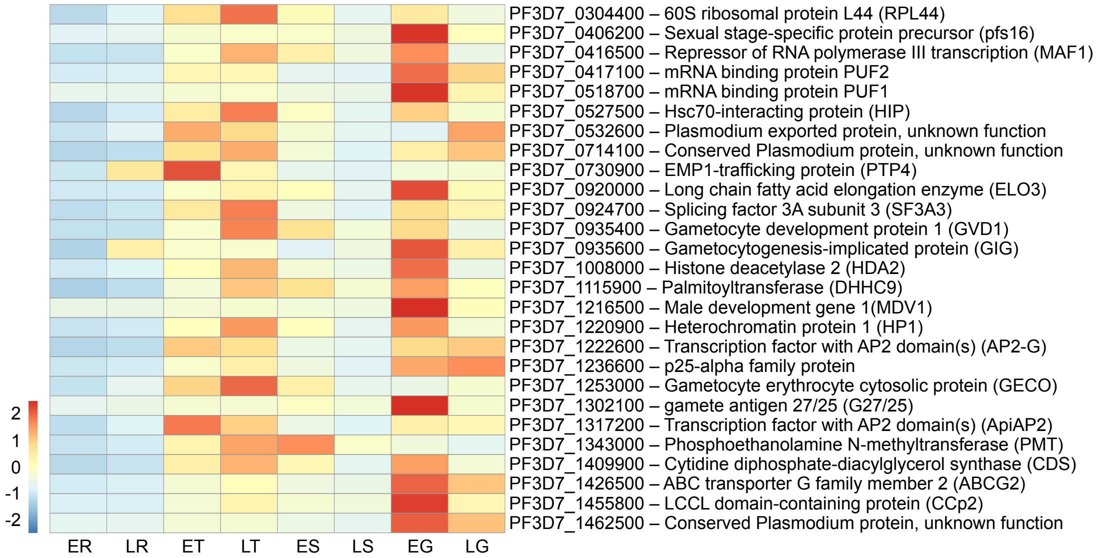
RAMA - rhoptry-associated membrane antigen (PF3D7_0707300)

PNT – putative pyridine nucleotide transhydrogenase (PF3D7_1453500)

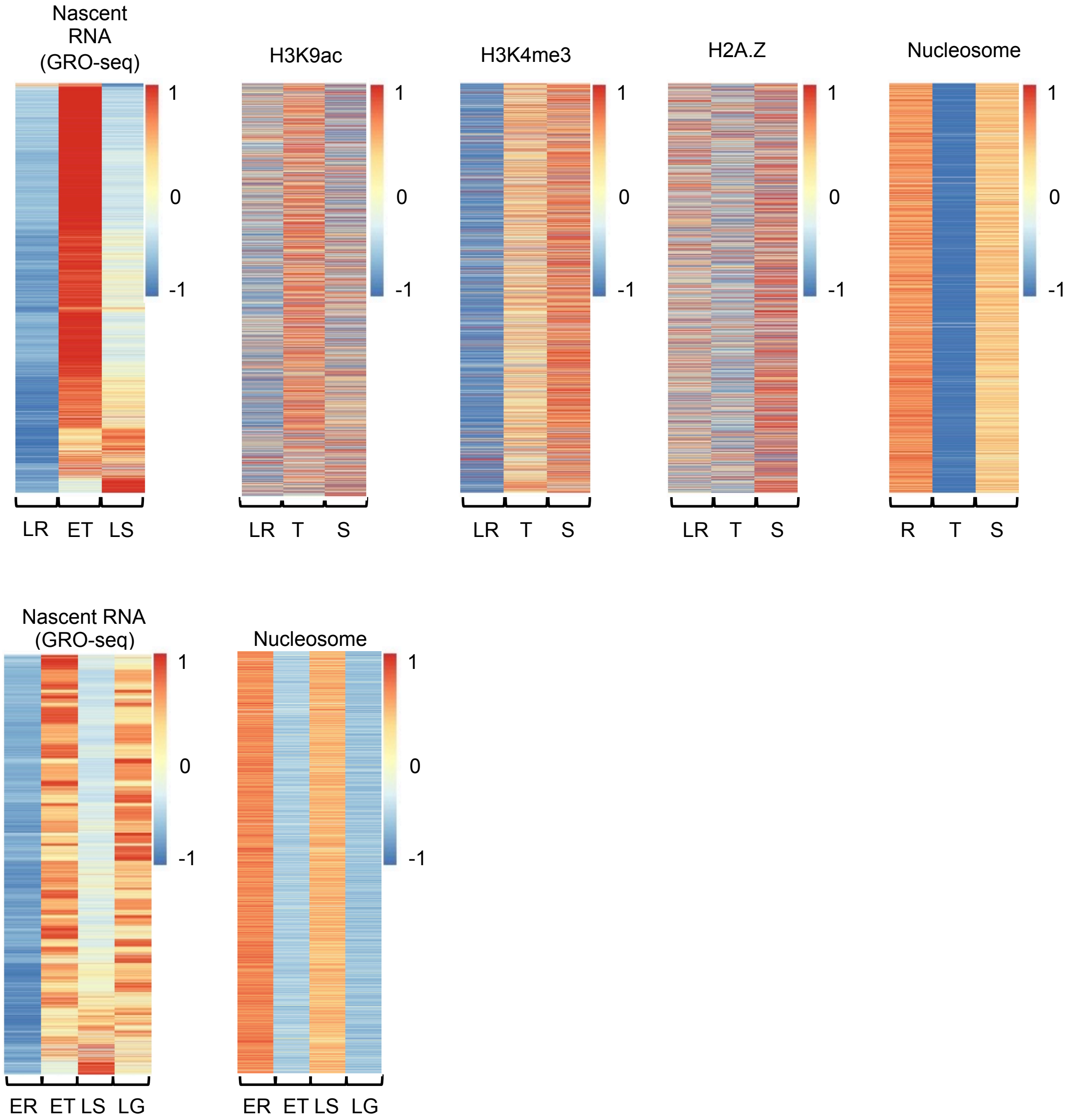
Supplemental Figure 2.4. PCR validation for GRO-seq cluster analysis. Four sets of primers were designed to amplify genes that were present in cluster A1 (ring stage expression - MAHPR1), in cluster A3 (trophozoite stage expression - TEX1), in cluster A9 (schizont stage expression - RAMA), and a gene that did not pass our threshold for expression (PNT). PCRs were performed using cDNA library samples from early ring (ER), late ring (LR), late trophozoite (LT), and late schizont (LS) stages, and were diluted to the amount of nascent RNA isolated from 20 x 10^9 parasites. For each gene, the PCR intensities were highest at the stage corresponding to their assigned GRO-seq cluster. No amplification was observed for the non-detected gene. gDNA, genomic DNA control; no temp, no template control.



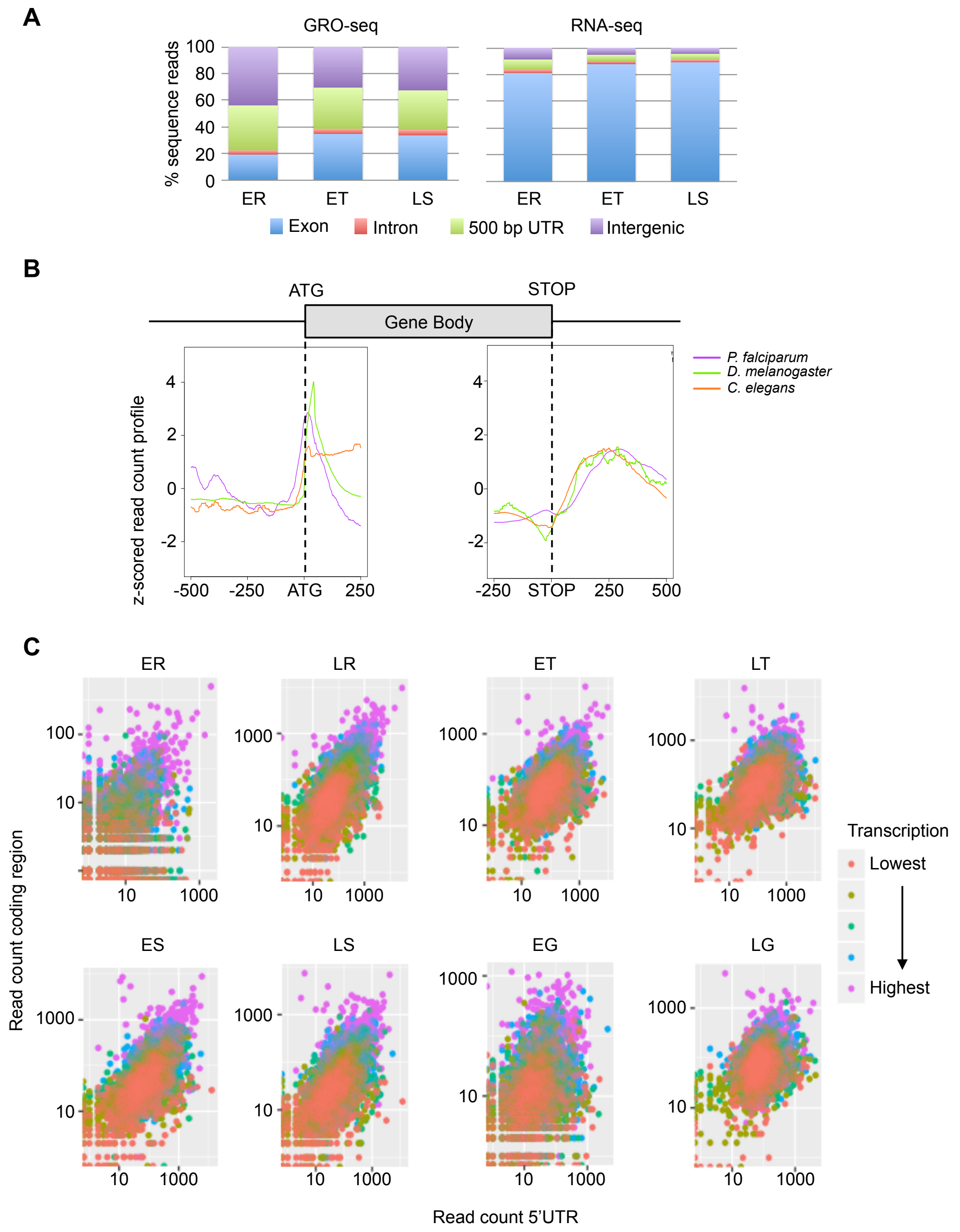
Supplemental Figure 2.5. Genome browser view of Pol II ChIP-seq data. Shown are the Pol II Ab tracks for early ring, early trophozoite, and late schizont stages with the corresponding no antibody controls. The bottom track shows sheared chromatin that was used as an input for ChIP. ER, early ring; ET early trophozoite; LS, late schizont.



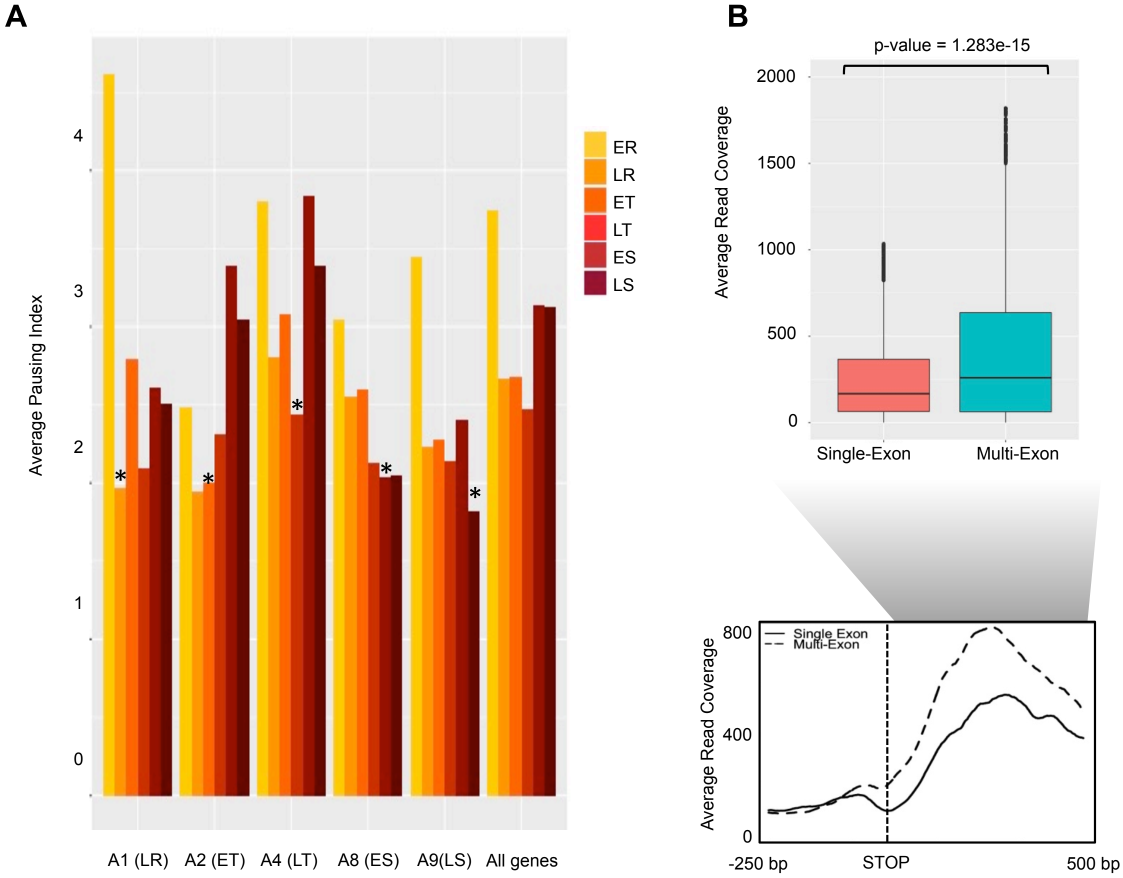
Supplemental Figure 2.6. GRO-Seq expression profiles of 27 genes that have been shown to be essential for gametocytegenesis. ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont; EG, early gametocyte stage; LG, late gametocyte stage.



Supplemental Figure 2.7. Association of transcriptional activity with chromatin structure. A comparison of transcriptional activity with relative H2A.Z, H3K4me3, and H3K9ac abundance (all three data sets from Bartfai *et al.* 2010), and global nucleosome occupancy (Bunnik *et al.*, 2014) during the IDC and gametocyte stages. Gametocyte data was not available for H2A.Z and the histone PTMs. Genes were ranked according to their transcriptional profile during the IDC in the same order as in Fig 1C. ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont; LG, Late gametocyte stage.



Supplemental Figure 2.8. Read coverage distribution. (A) Read coverage distribution over exons, introns, untranslated regions (UTRs, defined as 500bp upstream and downstream of annotated start and stop codons), and intergenic regions. (B) Average GRO-Seq coverage profiles at 5' UTRs and 3' UTRs in *Plasmodium falciparum* (blue), *Drosophila melanogaster* (green), and *Caenorhabditis elegans* (red). (C) The read counts in the 500 bp upstream of the ATG (5'UTR) plotted against the read counts in the 500 bp downstream of the ATG (coding region). The pol II pausing index is the ratio of these two numbers. Genes were sorted based on GRO-Seq signal, divided into five equal groups (see Supplementary Fig. 3) and are color-coded per group based on transcriptional activity.



Supplemental Figure 2.9. Association of transcriptional activity and Pol II pausing. (A) Pol II pausing index for all genes in our GRO-seq data set and for clusters of genes with stage-specific expression profiles. The Pol II pausing index corresponding to the stage of gene expression is indicated with an asterisk. (B) Average GRO-seq landscape at the 3’ UTRs of single-exon genes and multi-exon genes at the late trophozoite stage. A significant difference was observed for the average read coverage between these two groups of genes (P = 1.283e-15, Mann-Whitney U test). Outliers, defined as values more than 1.5 interquartile range from the median, were removed from the data set for plotting purposes. ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont.

Supplemental Files

Supplemental Flie 2.1: GRO-seq analysis associated information including raw and normalized exon counts for all genes, cluster information, and library mapping statistics. (XLSX)

Supplemental Flie 2.2: Enriched GO terms for GRO-seq analysis associated with Supplemental Figure 2.3. (XLSX)

Supplemental Flie 2.3: Enriched GO terms for GRO-seq gene expression and clustering analysis associated with Figure 2.1D. (XLSX)

Supplemental Flie 2.4: Raw and normalized exon counts for all genes and library normalization information associated with Pol II ChIP-seq data analysis. (XLSX)

Supplemental Flie 2.5: CITH and DOZI analysis associated information. (XLSX)

Supplemental Flie 2.6: Enriched GO terms for gametocyte transcriptional activity analysis associated with Figure 2.3. (XLSX)

Supplemental Flie 2.7: Data associated with motif analysis. (XLSX)

Supplemental Flie 2.8: Raw and normalized read counts at the 5' untranslated region for epigenetic landscape analysis associated with Supplemental Figure 7. (XLSX)

Supplemental Flie 2.9: Data associated with GRO-seq and RNA-seq comparison analysis. (XLSX)

Supplemental Flie 2.10: Primer information associated with semi-quantitative PCR validation. (XLSX)

Chapter 3: The chromatin-bound proteome of the human malaria parasite, *Plasmodium falciparum*

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

[1]Department of Cell Biology and Neuroscience, 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

*These authors contributed equally to this work.

Abstract

Chromatin proteins mediate fundamental cellular processes such as gene expression, DNA replication, DNA repair and maintain integrity of the genome. Recently, it has become evident that *Plasmodium falciparum*, the causative agent of malaria, displays limited tight transcriptional control of gene expression. Accumulating evidence suggests that parasite chromatin is highly structured at the three-dimentional (3D) level, and this structure potentially provides an epigenetic mechanism to regulate gene expression. To gather insights into how parasite 3D nuclear structure is being maintained, we undertook complementary computational, comparative genomics and experimental approaches to identify and characterize chromatin-associated proteins (CAPs) in *P. falciparum*. Over a 1000 CAPs are identified by hidden Markov model and NCBI RPS-BLAST searches, of which the abundance of CAPs in the parasite proteome is similar to other apicomplexan parasites but slightly higher than kinetoplastids. Several chromatin-associated domains (CADs) are enriched in apicomplexan parasites and plant species specifically. Using a novel methodology aimed at enriching for chromatin-bound proteome, we experimentally captured 987 CAPs during the blood stages, 397 of which overlaps with the in sillico identified candidate CAPs. Among the experimentally validated CAPs are many characterized chromatin regulators such as histone-modifying enzymes, parasite-specific transcription factors, DNA repair proteins, chromatin-assembly factors and many parasite proteins of unknown function. Finally, we validate several of our candidate proteins including a CROWDED-like NUCLEI (CRWN) protein, a plant nuclei protein that is functionally analogous to the animal nuclear lamina, using standard cellular and molecular approaches. Collectively, our results provide the most comprehensive overview of CAPs in *P. falciparum*. A better understanding of these CAPs will not only provide a complete picture of the complex molecular components that regulate

chromatin structure and genome architecture in the parasite, but will also assist the identification of novel targets for therapeutic strategies.

Introduction

The human malaria parasite, one of the deadliest infectious agents in the world, still contributes significantly to the global burden of disease. In 2015, an estimated 214 million cases of infection and 438,000 malaria-related deaths were reported [1], a majority of which are caused by the most lethal human malaria parasite, *Plasmodium falciparum*. Despite continued efforts to prevent malaria infections, treatment of affected individuals still remains one of the primary means of reducing malaria mortality. Given the absence of an FDA-approved vaccine and parasite resistance to all current antimalarial drugs [2, 3], there is a desperate need for new therapeutic approaches.

One promising strategy towards the development of novel and effective antimalarial compounds is to gain a better understanding of mechanisms regulating gene expression in the parasite. Since the publication of the *P. falciparum* genome in 2002 [4], researchers have attempted to explore the transcriptional machinery of the parasite in detail. The distinct developmental stages of the *P. falciparum* life cycle are characterized by coordinated changes in gene expression [5-7]. However, a surprisingly low number of specific transcription factors have been identified in the parasite genome [8, 9] and in particular, only a few stage-specific transcription factors have been validated [10-14]. 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 [15-19], translational and post-translational regulation [15, 20, 21] as well as change of chromatin structure, may control the expression of the predicted 6,372 genes in the malaria parasite.

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 [22-24]. Using the latest Hi-C methodology, our lab has determined the three-dimensional (3D) nuclear architecture of *P. falciparum* throughout its life cycle [25]. 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 [26], 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 chromosome 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 yeast [27, 28]. 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 stevor* genes) add a striking complexity and act as structural elements that shape whole genome architecture [25].

Architectural proteins involved in maintenance of chromatin structure have been studied in organisms ranging from yeast to human [29]. Among these proteins are RNA polymerase III-associated factor TFIIIC, cohesins, condensins and CCCTC-binding factor (CTCF) [29-32]. CTCF is an insulator protein conserved in vertebrates that is enriched at chromosome domain boundaries and interacts with the nuclear lamina [33]. Some of these components have homologues in the *P. falciparum* genome but only a few have been characterized at the functional level. Furthermore, many of conserved chromatin architectural proteins or chromatin-associated proteins (CAPs), involved in chromatin maintenance (e.g. lamina proteins) are missing in the parasite genome [34]. 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 [33, 35]. While absent in the malaria parasite, these proteins are likely to have distant homologues in *P. falciparum* 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 of model organisms, their importance in the development and virulence of *Plasmodium* has recently been appreciated for a small number of candidates [36-39], but a large number of them still need to be identified and characterized at the functional level. Given the potential roles of CAPs in almost all aspect of parasite biology, we performed a comprehensive computational and comparative genomics approach to generate an extended atlas of chromatin associated proteins in *P. falciparum*. Using a set of advance bioinformatics tools, we identified 1,190 of well-defined and putative CAPs in the parasite genome from which 162 proteins (13.6%) have been previously described as having chromatin-related functions [40]. We provide functional annotation based on homology, domain organization, domain clustering and expression patterns analysis. In addition, we developed an unbiased chromatin proteomics approach termed Chromatin Enrichment for Proteomics (ChEP) to experimentally validate some of our candidates. ChEP has been successfully used to identify chromatin-bound molecules and predict their function and regulation in a number of organisms [41-43]. Furthermore, we validated several of our candidate proteins including a CROWDED-like NUCLEI (CRWN) protein using standard cellular and molecular approaches. CRWN proteins are present in plant nuclei and resemble animal and fungal lamina but the machinery and processes that underlie these proteins appear to be evolutionarily distinct from their animal counterparts [44, 45]. In plant, CRWN proteins are essential for viability and play diverse roles in both heterochromatin organization and the control of nuclear morphology [46]. Identification and characterization of these homologues in *Plasmodium* could reveal novel exciting targets for drug discovery. Altogether, our results validate that mechanisms regulating chromatin structure in the parasite are most likely complex but a better understanding of these CAPs will not only provide a comprehensive picture of the complex molecular components that control chromatin organization and genome architecture of this deadly parasite, but will also assist the identification of novel targets for therapeutic strategies.

Results

In silico identification and classification of chromatin-associated proteins

To study and characterize chromatin-associated proteins (CAPs) in *P. falciparum*, we first identified and characterized *Plasmodium* proteins that contain domains with nuclear or chromatin functions. Hereafter, we call these domains chromatin-associated domains (CADs). To obtain a list of all possible 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. Next, we searched the *Plasmodium* proteome for all possible domains using both hmmscan (HMMER v3.16) and NCBI Reversed Position Specific BLAST (RPS-BLAST). We then searched for 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) in *P. falciparum*. 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 (Additional File 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 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 [40], 877 have non-chromatin related annotation and 151 proteins (12.7%) are unknown proteins where functions have yet to be discovered.

To have a better understanding of the *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 associated with cell cycle progression, chromatin remodeling, DNA binding, transcription regulation, or other nuclear activity (total 77 members). Transcription or mRNA processing-associated RNA-binding domains (73 members), catalytic domain of the Dual-specificity protein kinases (64 members), DEAD box helicase domains (63 members), WD40

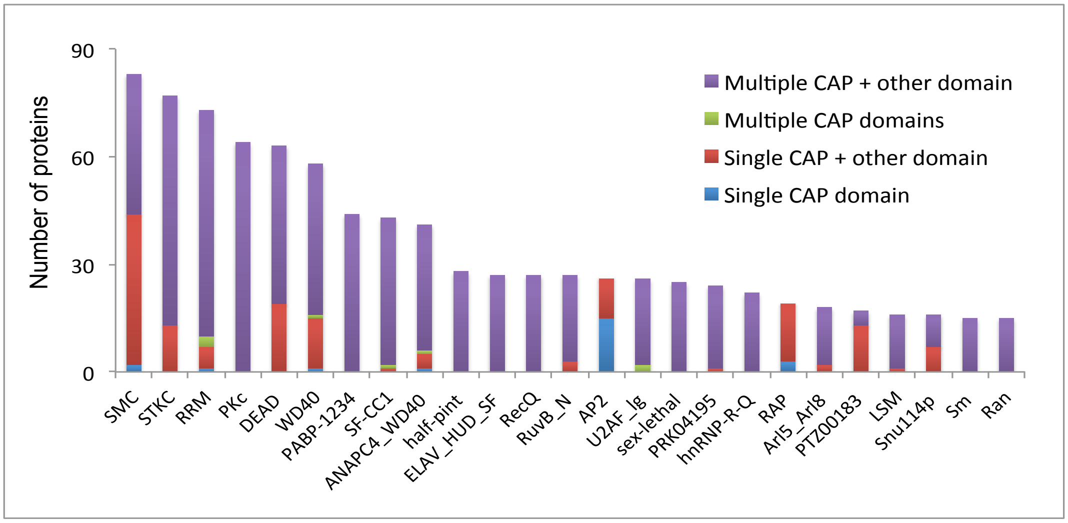


Figure 3.1. Characterization of chromatin-associated domains (CADs) that were found in eight or more candidate CAPs.

domains (58 members), polyadenylate binding domains (44 members), and splicing factor, CC1-like domains (43 members) were also found to be abundant in the parasite's genome along with other domains such as the anaphase-promoting complex unit, AP2 transcription factor domains, small nuclear ribonucleoprotein domains, and GTP-binding nuclear protein domains (Figure 3.1). 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. This finding suggests that these CAPs may likely have multiple functional roles in the biology of the parasite (Figure 3.1). To explore the potential function of these chromatin-associated candidate proteins, we further categorized these proteins based on their functionality using protein descriptions (Figure 3.2). 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

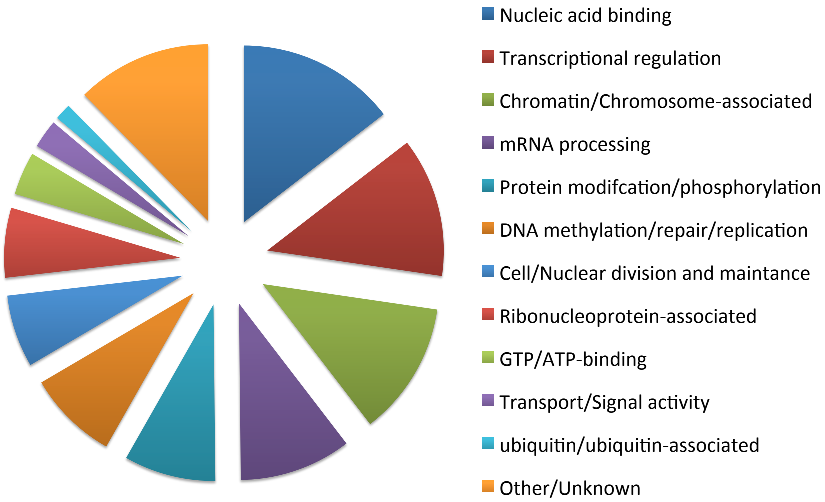


Figure 3.2. Classification of candidate chromatin-associated proteins (CAP) based on their annotation or associated domains.

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, protein involved in splicing, 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%), ubiquitin or ubiquitin-associated proteins (n=18, 1.5%). About 150 proteins (12.4%) were found to be associated with other cellular functions that are non-chromatin related. Some example proteins belonging to this group are 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 majority of these chromatin-associated candidate proteins are likely to be involved in transcription related events (Figure 3.3).

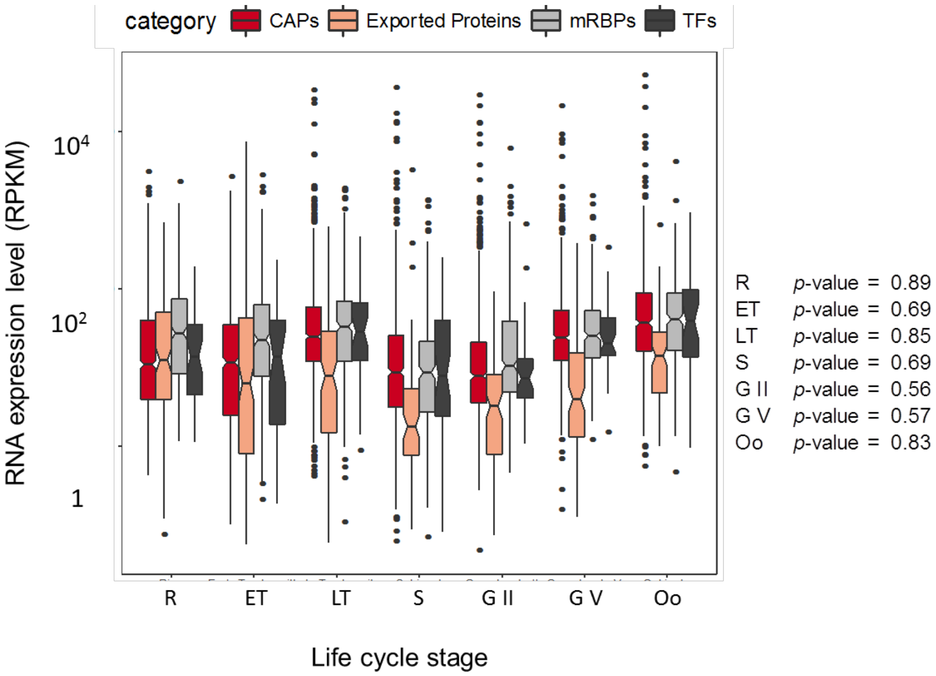


Figure 3.3. Gene expression comparison between candidate CAPs and other classes of proteins during different developmental stages of the parasite life cycle. Statistical significant differences in average expression levels between CAPs and transcription factors are indicated in *p*-value. (R) ring, (ET) early trophozoite, (LT) late trophozoite, (S) schizont, (G II) Early gametocytes at stage II, (G V) late gametocytes at stage V, (Ook) ookinete

Comparative analysis of chromatin associated proteins

To better understand the roles of CAPs in parasite biology, we next performed a genomic comparative analysis of CAPs among a variety of organisms, including two additional apicomplexan parasites (*Plasmodium vivax* and *Toxoplasma gondii*), euglenid parasites (*Trypanosoma brucei, Trypanosoma cruzii,* 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, the amount of CAPs in *Plasmodium falciparum* was relatively similar to number of CAPs in other apicomplexan parasites (~22% of full proteome), slightly higher than those in euglenid parasites, *S. pombe,* and *C. elegans*, but lower than the number of CAPs in the *S. cerevisiae* and higher eukaryotes (Figure 3.4). This suggests that *P. falciparum* chromatin structure is more complex and is regulated via more CAPs as compared to *T. brucei, T. cruzi, L. major, S. pombe* and *C. elegans*, possibly due to the structural complexity added by virulence genes [25]. However, in higher eukaryotes, hierarchical chromatin elements such as compartments, topologically associating domains (TADs) and insulated domains have been described [47], and as a result might be regulated via more CAPs as compared to *P. falciparum*.

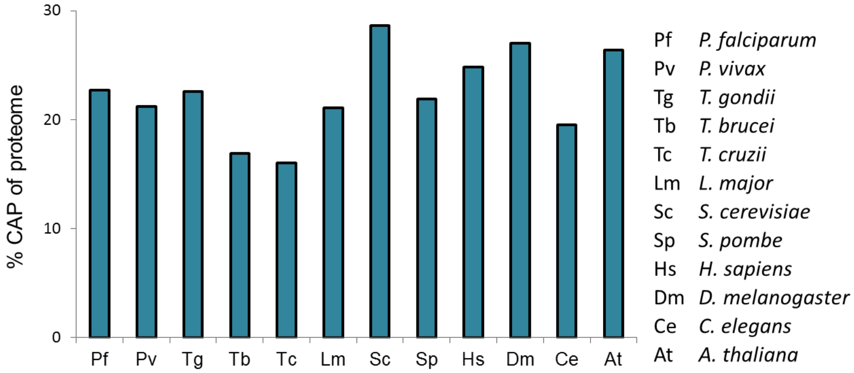


Figure 3.4. Relative abundance of CAPs in the full proteome of various organisms.

To identify functional differences in chromatin-associated processes, the CADs were clustered based on their relative abundance in all investigated species (Figure 3.5). Each cluster was then analyzed for domain-associated Gene Ontology (GO) enrichment. Twelve distinct clusters were obtained. Clusters 1-3 contain CADs that are relatively abundant in apicomplexans, of which the domains in cluster 1 are almost exclusively enriched in *Plasmodium* species. These domains show enrichment for GO terms associated with nucleic acid binding and specifically AP2 domain-containing transcription factors (Figure 3.6). While highly abundant in apicomplexans, AP2 domain containing proteins are also abundant in plant species. AP2 family of transcription factors play an essential role in floral development in *A. thaliana* [48] and in *P. falciparum,* proteins containing AP2 binding domains (ApiAp2) have been identified as sequence-specific transcription factors [49] and are believed to be master regulators of transcription during parasite development [10, 13]. The enrichment of AP2 domains in apicomplexan parasites and *A. thaliana* in our expression analysis further validates our classification methods. In addition, cluster 1 harbors the RCC1 domain (Figure 3.6), 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 was identified in apicomplexan parasites [50]. These atypical apicomplexan RCC1 proteins show different arrangements of RCC1 domains compared to their higher eukaryotic RCC1 orthologs. Cluster 2 contains CADs that are abundant in all unicellular organisms. This cluster shows 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 3.6). This protein is found in many prokaryotic members and is the single type RNA polymerase that performs transcription in bacteria [51]. Given the abundance in bacteria, this domain seems to also be conserved in unicellular eukaryotes. Cluster 3 contains CADs that are 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 3.6). Unlike their unicellular counterparts, euglenid parasites transcribe their protein-coding genes into polycistronic RNAs [52] and processes the RNA through a special mechanism termed trans-splicing where exons from two different primary transcripts are ligated [53]. 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 [54].

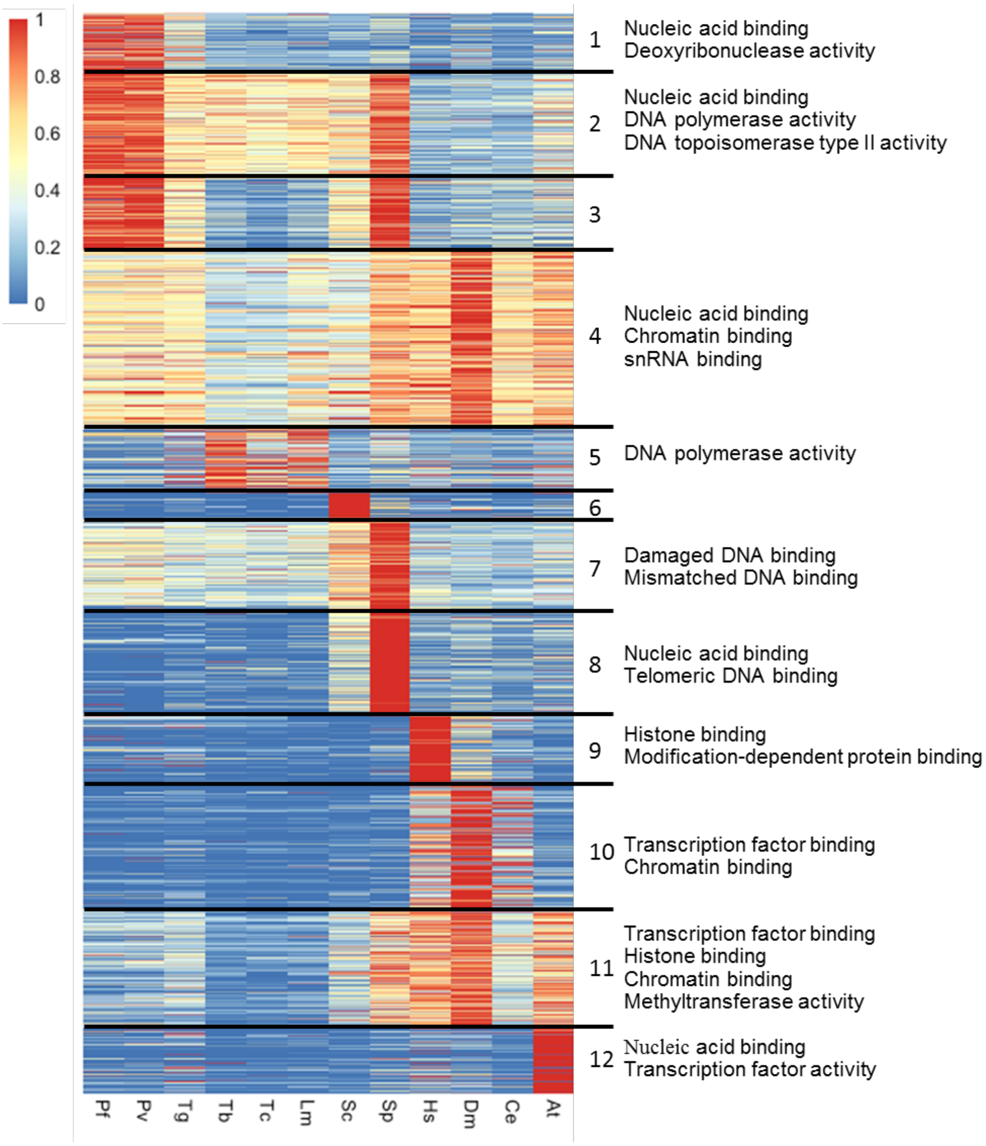


Figure 3.5. *k*-means clustering of the relative abundance of CADs among 12 organisms. The CAD abundance was first normalized for each organism by proteome size and then scaled to the CAD frequency with the highest relative abundance of that CAD. A subset of the Gene Ontology (GO) enriched terms associated with the Pfam domains (false discovery rate, FDR<0.01) for each cluster are shown on the right.

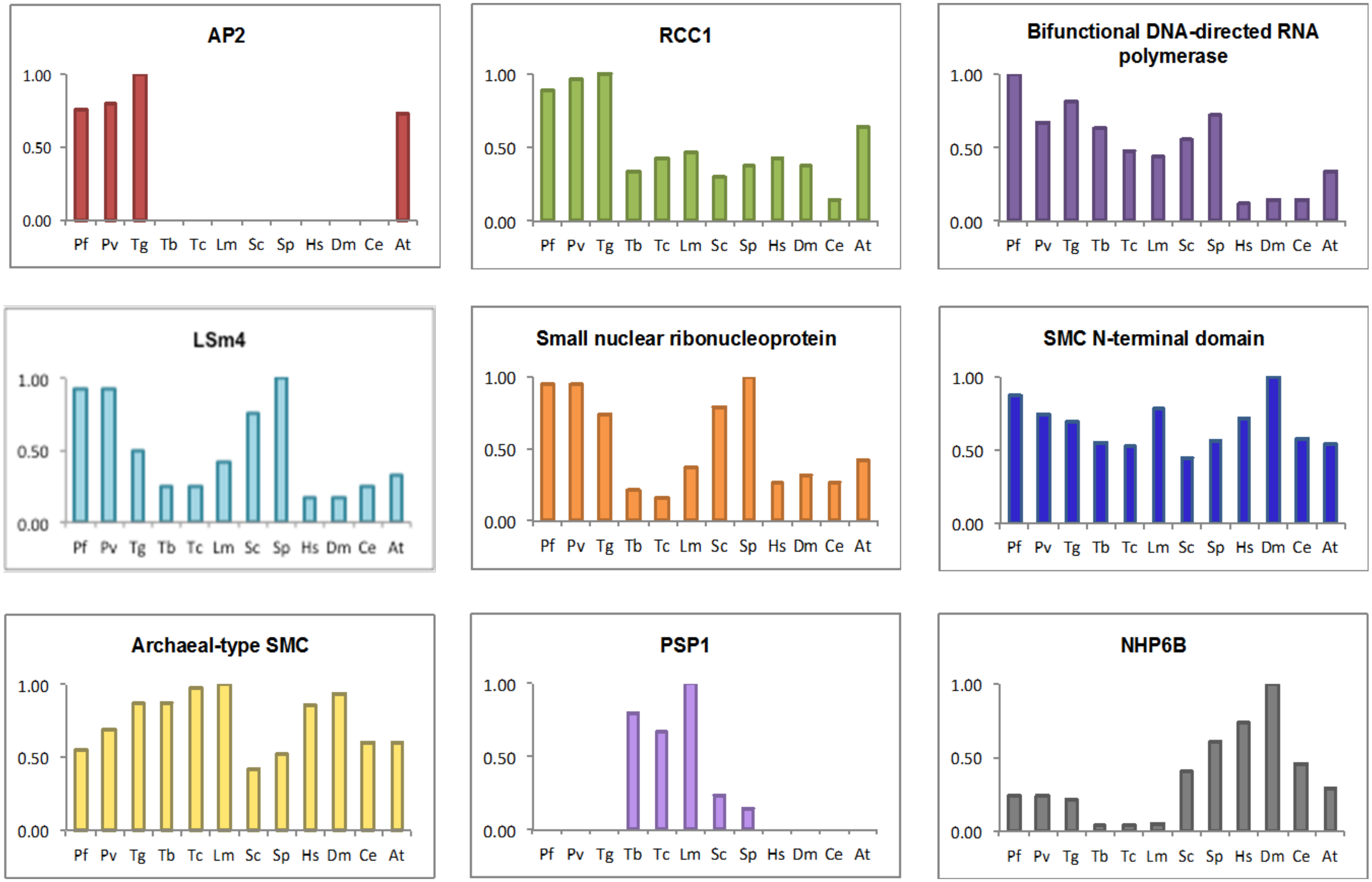


Figure 3.6. Selection of CADs and their relative abundance among all 12 organisms.

Chromatin-associated domains in cluster 4 are abundant in all twelve organisms discussed. One of the major domains enriched in this cluster is the SMC N-terminal domain (Figure 3.6). SMC domain-containing proteins are a large family of ATPases that play a role in many aspects of chromosome organization [55]. Different SMC subunits makeup cohesin and condensin complexes, and these proteins play an essential role in chromosome assembly and segregation [56, 57]. In particular, condensin promotes chromosome compaction [56], while cohesin facilitates sister chromatid separation during mitosis and meiosis [57]. 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 are CADs mostly abundant in kinetoplastids. This cluster harbors 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 3.6) [58]. 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 [59]. Homologues of CSBP proteins were found only among the kinetoplastids, however whether these proteins share a functional relationship with the yeast PSP1-containing proteins have yet to be determined. Cluster 11, contains CADs most abundant in non-protozoan organisms. This cluster represents 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 3.6). HMGBs are highly abundant DNA-binding proteins that are involved in many nuclear functions including chromatin remodeling, transcription, recombination and DNA repair [60, 61]. The C-terminal acidic tail typical of metazoan HMGBs [62, 63], which regulates the DNA-binding characteristics of the HMGB-box domains, is missing from most unicellular organisms [64, 65]. This suggests that protozoan HMGBs might have additional sequence characteristics that enable HMGBs to bind DNA, which are absent from higher eukaryotes.

Experimental validation of chromatin-associated proteins

To validate our in silico identification of chromatin-bound proteins, we next performed a method designed to isolate, in a genome-wide manner, all proteins associated with chromatin. This methodology, termed Chromatin Enrichment for Proteomics (ChEP) (Figure 3.7A) was adapted from published studies on human and mouse cell lines [42]. 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 a longer cross-linking time, at a higher 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. Under these conditions, we observed a clear enrichment, using western blot analysis, of nuclear proteins histone H3 and RNA polymerase II in the nuclear fraction (Figure 3.7B). As a negative control, we isolated proteins from the cytoplasmic fraction.

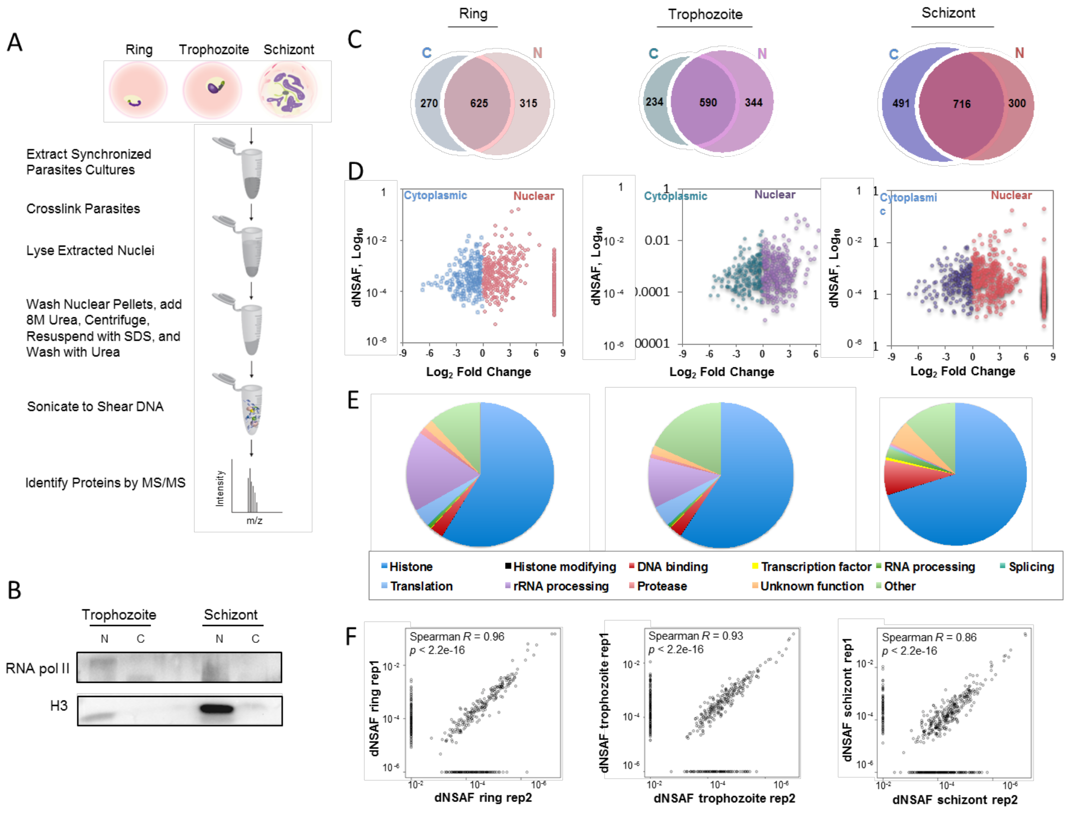


Figure 3.7. Chromatin enrichment for proteomics (ChEP). A) Outline of the ChEP procedure. B) Validation of protein enrichment in the nuclear fraction from ChEP by western blot analysis (1- trophozoite nuclear fraction, 2- trophozoite cytoplasmic fraction, 3- schizont nuclear fraction, 4- schizont cytoplasmic fraction). Western blots show an enrichment for RNA polymerase II and histone H3 in the nuclear fraction compared to the cytoplasmic fraction. 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) Proteomic analysis of the nuclear ChEP sample. 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.

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 performed for each intraerythrocytic stage. We identified a total of 940, 934 and 1,016 proteins at the ring, trophozoite and schizont stages respectively in the nuclear ChEP fraction (Figure 3.7C and Supplemental File 3.2). We then compared the nuclear ChEP proteins to the control cytoplasmic proteins to identify proteins enriched in the ChEP sample. We identified 370, 611 and 706 proteins at the ring, trophozoite and schizont stages respectively, that are detected at ≥ 2-fold enrichment in the nuclear fraction as compared to the cytoplasmic sample (Figure 3.7C, Supplemental File 3.3).

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 3.7E, Supplemental File 3.3). 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 [66, 67]. 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 [68], and considering the biology of the parasite, a majority of the 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). On average, proteins detected in a given ring, trophozoite and schizont sample was also detected in its matching biological replicate. Additionally, calculation of spearman rank coefficients showed that each sample correlated strongly with its matching replicate (Spearman R = 0.96 at ring, 0.93 at trophozoite and 0.86 at schizont stages; Figure 3.7F). 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 and mass spectrometry methodology.

A recent high-throughput proteomic analysis explored the nuclear proteome during the *P. falciparum* intra-erythrocytic developmental stages [69]. To validate the ability of our ChEP methodology to specifically enrich chromatin-bound proteins, we compared our dataset to the *P. falciparum* nuclear proteome. In order to perform an unbiased comparison, we filtered both our dataset and the Oehring [69] dataset to identify proteins that were detected with ≥ 2-fold abundance (based on uniquely detected peptide counts) in the nuclear fraction as compared to the control cytoplasmic fraction. For both datasets, biological replicates were merged and the uniquely detected peptide counts for each protein was averaged. The enriched proteins at ring, trophozoite and schizont stages were merged and duplicate protein IDs were removed. The final filtered protein list used for comparison included 1200 proteins from the ChEP experiment and 909 proteins from complete nuclear proteome. A total of 490 (30%) protein candidates from the complete nuclear proteome were identified in our ChEP experiment (Supplemental Figure 3.2). These proteins enriched for GO terms associated with nucleic acid binding, as well as transcriptional and translational processes (Supplemental File 3.4). A total of 419 (26%) proteins from the nuclear proteome list, enriching for GO terms associated with substrate-specific and transmembrane transporter activities, were not enriched in our ChEP list. In total, 710 (44%) proteins identified from the ChEP methodology were not enriched in the nuclear proteome analysis. These proteins most highly enriched for GO terms associated with DNA- and RNA-binding. Additionally, out of the 397 proteins that were enriched in both our in silico and ChEP analysis (Supplemental Figure 3.1B), 209 (53%) proteins were enriched in the nuclear proteome. Taken together, these results validate the high specificity of the ChEP methodology and our ability to detect and enrich for chromatin-associated protein candidates in an unbiased manner.

The experimentally detected candidate CAPs enriched by ≥ 2-fold abundance were compared to the computationally detected CAPs. A total of 397 candidate CAPs that were captured by the

MudPIT analysis validated 40% of the CAP candidates identified in our HMM search and represent 33% of all computationally detected candidate CAPs (Supplemental Figure 3.1B). This group of proteins validate by both our *in silico* and experimental methods, are most likely to be involved in maintaining or regulating chromatin structure in the parasite. The exact function of these novel proteins in chromatin-related processes will need to be validated, but validation of one such protein candidate below highlights their potential as new therapeutic targets.

Functional validation of candidate chromatin-associated proteins

Putative chromatin-bound protein candidates enriched using the ChEP methodology with high reproducibility was searched for the existence of even distantly related homologs using PSI-Blast HHPred [70]. Candidate proteins with domain homology for chromatin components was 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, condensing, CTCF insulator or insulator-like proteins. Our analysis identified 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 [71]. 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 (E-value = 0.01) and was used 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 used 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. Using western blot, we observed a clear enrichment of CRWN-like and SMC3 proteins in the nuclear fraction (Figure 3.8C). Our results validate the use of these custom antibodies to detect the *P.falciparum* CRWN-like (~300 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 was observed for SMC3 protein at trophozoite and schizont stages (Figure 3.8A, right panel). We were unable to detect a fluorescence signal for SMC3 at the ring stage, possibly due to nuclear compaction at this stage. At all three asexual stages, the CRWN-like protein localized to the nuclear compartment (Figure 3.8A, left panel). In particular, we observed a single focus per nucleus at the ring and schizont stages (Figure 3.8A, left panel). 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 [71]. It is possible that the CRWN-like protein in *Plasmodium* is similarly localizing to the heterochromatin region of the nucleus. Further experiments will be needed to validate the exact function of this CRWN-like protein during parasite development.

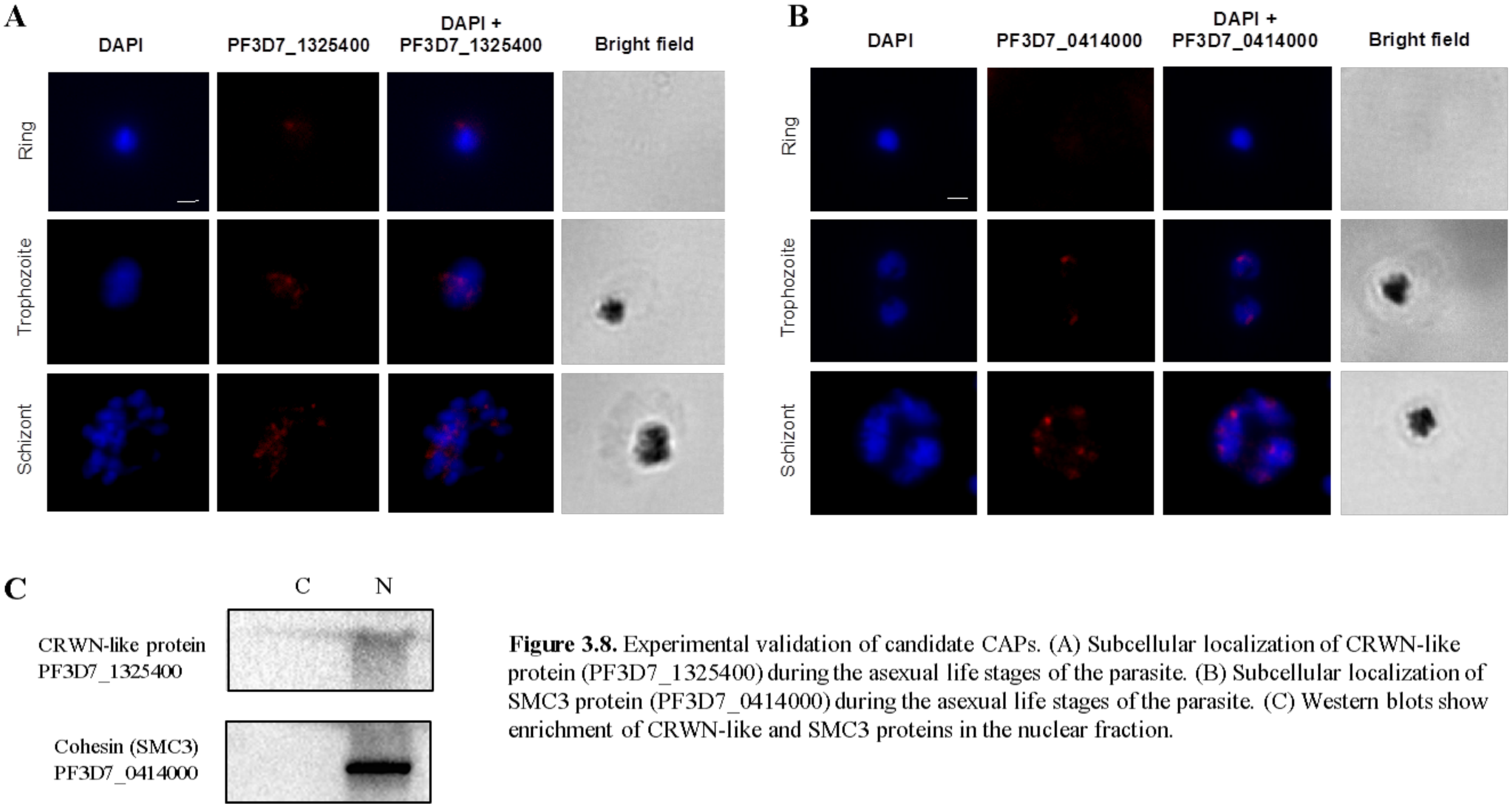


Figure 3.8. Experimental validation of candidate CAPs. (A) Subcellular localization of CRWN-like protein (PF3D7_1325400) during the asexual life stages of the parasite. (B) Subcellular localization of SMC3 protein (PF3D7_0414000) during the asexual life stages of the parasite. (C) Western blots show enrichment of CRWN-like and SMC3 proteins in the nuclear fraction.

Protein interaction study

To investigate parasite-specific molecular components interacting with SMC3 and CRWN-like proteins, we performed immunoprecipitation experiments using the custom generated antibodies. Briefly, mixed-stage parasite protein lysates were subjected to Dnase I and heparin sulfate treatment to solubilize chromatin-associated complexes. The solubilized protein fraction was 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 45 proteins were detected with ≥2-fold higher abundance to be interacting with SMC3 in the parasite (Supplemental Figure3.3, Supplemental File 3.5). These proteins enriched for GO terms associated with DNA- and RNA-binding such as transcription factor with AP2 domain (fold change = infinity), HMGB1 (fold change = infinity), and nucleosome assembly protein (fold change = infinity). Additionally, we successfully recovered SMC3 (fold change = infinity) and another subunit of the cohesin complex, SMC1 (PF3D7_1130700, fold change = infinity), 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 the immunoprecipitation methodology. Alternative tagging strategies will be needed to identified interacting partners of CRWN-like protein in the parasite.

Genomic distribution of our candidate proteins

In order to determine the genome-wide distribution of SMC3 and CRWN-like proteins, we next performed ChIP-seq experiments. Briefly, trophozoite stage parasites were cross-linked with formaldehyde. Sonicated chromatin was incubated with anti-SMC3 and anti-CRWN-like antibodies 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 3.9). 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 [72, 73]. In yeast, cohesin localizes to centromeres and extends to nearby pericentromeric regions [74, 75]. Preferential loading of cohesin at centromeres is a kinetochore-dependent process [76]. 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 during 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.

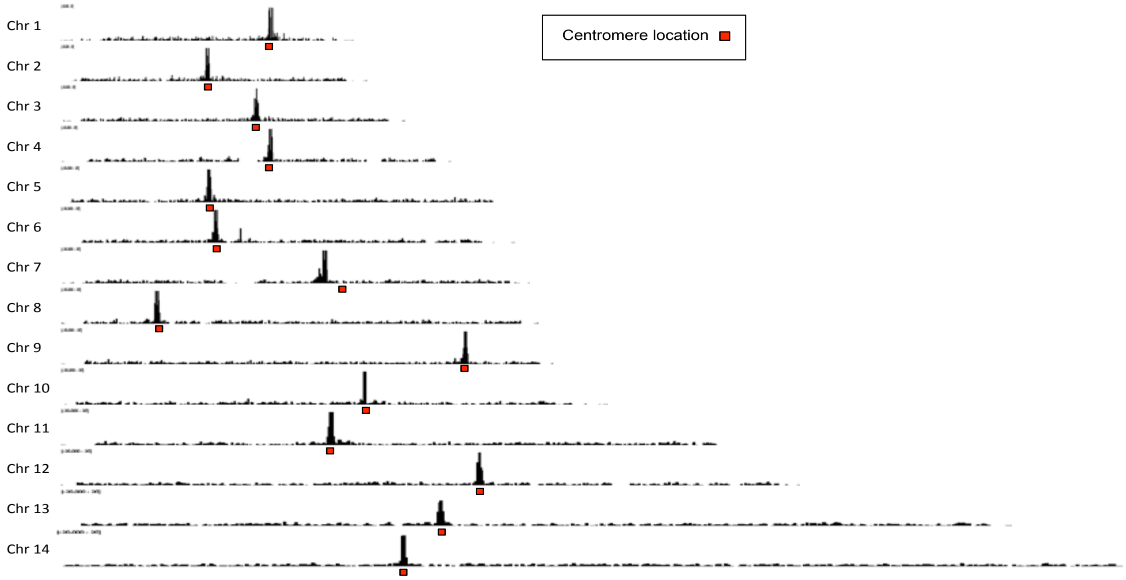


Figure 3.9: ChIP-seq analysis showing genome-wide distribution of SMC3 in trophozoites. The red box indicates the location of the centromere on each chromosome.

Materials and 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 [17]. 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). For each protein, if multiple RPS-BLAST hits were reported for the same conserved domain, only the one with the highest percent identity is maintained. An e-value of 0.001 was used for both approaches, and if a protein has multiple isoform, only the first isoform is 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, and etc. (see Additional for details). Next, we further selected chromatin-associated domains in the resulting list base on their annotation. Similarly approach was used to identify chromatin associated domains within the NCBI conserved domains listed cddid_all.tbl file. Next, both chromatin-associated domain lists were combined and carefully curated by person. Domains without a clear defined 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 chromatin associated candidate proteins in *plasmodium falciparum*, both manual-curating and a list of exported proteins that includes 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 the potential false positive proteins, as these proteins are more likely to be exported into the red blood cell than to be exported into the nucleus.

Protein classification

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

Gene expression analysis for plasmodium CAPs

The gene expression profiles and boxplot were generated using steady-state mRNA expression profile downloaded at plasmodb.org. The expression profiles were pre-processed using standardized pipelines and are RPKM transformed. For boxplot, gene groups were generated based on gene annotation and the list of RNA binding proteins were obtained from [17].

Barplot comparison of CAPs

For each organism, both hmmscan and RPS-BLAST approaches were used and merged as previous 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 to protein annotation. For each organism, the calculated value is then corrected by the proteome size and expressed as the percentage of chromatin-associated protein 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 is corrected by genome size and expressed as the number per 10,000 genes. Any domains with a value of zero after genome size correction were removed. The remaining relative abundance values are 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 was 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, in which a minimum of 60% variance is required and an increase in number of cluster did not capture at 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.

Parasite cultures

The *P. falciparum* strain 3D7 was cultured in human O+ erythrocytes at 5% hematocrit as previously described [77]. Cultures were synchronized at ring stage with 5% (w/v) D-sorbitol treatments [78]. 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 early schizont stages) using a protocol adapted from [42].

Briefly, synchronized parasites were crosslinked with 1% formaldehyde for 15 min at 37°C. Crosslinking 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 4000 rpm for 20 min at 4°C.

The nuclear pellet was washed with highly denaturing extraction buffers containing 4% SDS and 8M urea to wash away non-crosslinked proteins. Chromatin was solubilized and genomic DNA was sheared by sonication (Covaris). As a negative control, protein from the cytoplasmic fractions of the early trophozoite and late schizont parasites were 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 4,200 rpm 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 4,200 rpm at 4°C. After the last wash, parasites were resuspended in PBS, transferred to a microcentrifuge tube and centrifuged for 5 min at 5,000 rpm 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 13,000 rpm 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) was used to immunize two rabbits. Antisera from day 72 post-immunization was collected and affinity purified to purify antibodies specifically targeting the protein of interest.

Immunofluorescence microscopy

P. falciparum asexual and sexual stage parasites were fixed onto slides using 4% paraformaldehyde for 30 min at RT. Slides were washed three times using 1x PBS. The parasites were permeabilized with 0.1% 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 anti-Histone H3 antibody (ab8898 (Abcam) for gametocyte samples; 1:500 and 07-442 (Millipore) for asexual stage samples; 1:500) for 1 hr at RT followed by anti-rabbit Alexa Fluor 488 (Life Technologies A11008; 1:500). Slides were mounted in Vectashield mounting medium with DAPI. Images were acquired using the Olympus BX40 epifluorescence microscope or the Leica SP5 confocal 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 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 13,000 rpm 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 13,000 rpm 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 custom antibodies generate against PF3D7_1325400 (Thermo Fisher, 1:100) and PF3D7_0414000 (Thermo Fisher, 1:100) and Goat Anti-Rabbit IgG HRP Conjugate (Bio-Rad, 1:10,000). The membranes were visualized using the Bio-Rad ChemidDoc 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 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (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 13,000 rpm 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 collected to a new microcentrifuge tube and split equally for the antibody and no antibody control. The SMC3 custom antibody was added at a 1:50 ratio the antibody tube 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 4,200 rpm 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 4,200 rpm at 4°C. Subsequently, parasites were crosslinked 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 crosslinking reaction, and incubated for 5 min at 37°C. Parasites were centrifuged for 5 min at 5,000 rpm 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 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (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 26 G ½ inch needle seven times. Parasite nuclei were centrifuged at 4°C for 20 min at 5,000 rpm. 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 14,000 rpm 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). 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 crosslinked 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 reads quality was first analyzed using FastQC(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and the first 15 bp and the last base were removed. Any base with a quality score below 25 was trimmed using Sickle (https://github.com/najoshi/sickle). Trimmed reads are then mapped to *P. falciparum* genome (v34) using Bowtie2 (v2.2.2) [79]. Uniquely mapped reads were further filtered, and Read coverage per nucleotide was first determined using BEDTools. The negative control library was obtained from NCBI GEO database under accession number, GSE85478. The negative control represented a no antibody control from the early trophozoite stage as described previously [80]. Both libraries were then normalized by dividing numbers of million mapped reads and signals from negative control library was subtracted from Cohesion ChIP-seq library. Genome browser tracks were generated and viewed using Integrative Genomic Viewer (IGV) by Broad institute. Centromere locations were obtained from [81].

Multidimensional Protein Identification Technology (MudPIT)

Proteins were precipitated with 20% trichloroacetate 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 was 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. All samples were run in low resolution mode. A fully automated 10-step chromatography run (for a total of 20 h) was carried out, as described in [82]. 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 SEQUEST [83] against a database consisting of 5538 *P. falciparum* non-redundant proteins (downloaded from PlasmoDB on 24 March 2016), 34,521 *Homo sapiens* non-redundant proteins (downloaded from NCBI on 24 March 2016), 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 [84]. 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 [85].

MudPIT data analysis

A total of two biological replicates, and two technical replicates for each biological replicate was performed for ChEP and cytoplasmic control samples at 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 more than two-fold higher normalized abundance (dNSAF) as compared to the control cytoplasmic sample. List of all proteins that were detected in our samples and individual peptide/spectral counts are provided in Supplemental File 3.2.

Nuclear proteome and chromatin-bound proteome comparison

P. falciparum nuclear proteome dataset from [69] was used for the comparison. For each asexual stage (ring, trophozoite, schizont), 8 replicates from the Oehring dataset and 4 replicates from our MudPIT analysis were merged and averaged. Enrichment for nuclear proteins or chromatin-associated proteins in each experiment was defined as detection of two or more uniquely detected peptides of that protein in the nuclear or ChEP sample and a more than two-fold higher fold change as compared to the control cytoplasmic sample. Enriched GO terms for uniquely detected proteins in each dataset were determined using the PlasmoDB GO enrichment tool.

Discussion and Conclusion

Increasing evidence points towards genome architecture and chromatin structure regulation playing an important role in gene expression throughout the life cycle of *P. falciparum* [37, 86-89]. To better understand how the three-dimensional structure of the genome is being maintained, it is vital to identify which proteins and protein-complexes associating with chromatin throughout parasite development. Although snapshots of the parasite proteome have been generated [26, 69], no such complementary approaches have been performed to generate an accurate view of chromatin-associated proteins throughout the parasite life cycle. The dataset presented here gives the most complete overview of the chromatin-bound proteome in *P. falciparum* to date.

By searching the parasite proteome using a large collection of Pfam (HMM) and NCBI (RPS-BLAST) chromatin-associated domains, we have attempted to identify all CAPs throughout the parasite life cycle. Since *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 current genome annotation to filter our initial broad search to proteins that specifically interact with chromatin. The overall gene expression profiles of the identified candidate CAPs in many stages of the parasite's life cycle show similar expression levels to transcription factors, which indicates the importance of the candidate CAPs in gene regulation (Figure 3.3).

In an unbiased comparison with other eukaryotic organisms, we observed that *P. falciparum* encodes a similar number of CAPs as compared with other apicomplexan parasites but a relatively lower number than the similarly sized budding yeast (Figure 3.4). However, more CAPs are found in *P. falciparum* compared to kinetoplastids *T. brucei, T. cruzii* and *L. major.* Transcriptional regulation in these euglenid organisms is unusual as it is polycistronic, and these organisms regulate their gene expression mostly at the post-transcriptional level [90]. In contrast, accumulating evidence suggests that *P. falciparum* uses chromatin structure as a basal control for gene regulation [37, 86-89, 91], which reflects the importance of CAPs in parasite biology.

Cluster analysis based on the expression of the CADs in a variety of eukaryotic organisms revealed twelve distinct clusters (Figure 3.5). Cluster 1, highly abundant in apicomplexan parasites and plants harbors proteins containing AP2 domains. The enrichment of these parasite and plant-specific domains in our expression analysis further validates our classification. More importantly, these protein domains are underrepresented in mammalian organisms, which points towards the existence of parasite and plant specific chromatin regulation pathways that could be ideal drug targets. Cluster 4, enriches for proteins containing SMC domains and is expressed across many eukaryotes including apicomplexans, kinetoplastids, and yeast. SMC domain-containing proteins are a large family of ATPases that play a role in many aspects of chromosome organization [56]. Cohesin protein complexes containing core subunits SMC1 and SMC3 regulate the separation of sister chromatids during cell division [55], while the condensin complex containing core subunits SMC2 and SMC4 regulate chromosome assembly and segregation during mitosis and meiosis [57]. Additional clusters revealed CADs that are relatively depleted in *Plasmodium* species as compared to one or multiple other organisms. Interestingly, kinetoplastids (cluster 5), yeast (cluster 8), human (cluster 9) and plant (cluster 12) also show enrichment for several CADs as compared other organisms, suggesting that these organisms have also developed particular species-specific chromatin-related mechanisms.

Out of the in silico identified 1,190 *P. falciparum* proteins that contain a chromatin-associated domain, 397 proteins (40%) 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 tried 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 after cell lysis. To validate the efficiency of the ChEP protocol in enriching for chromatin-bound proteins, we compared our analysis to a previously published *P. falciparum* nuclear proteome dataset [69]. Overall, 30% of enriched proteins were shared between the nuclear proteome analysis and ChEP sample. However, many proteins with non-chromatin related functions such as transporter activity were enriched in the nuclear proteome and were not identified using our ChEP methodology and many DNA- and RNA-binding proteins were enriched in our ChEP sample and were not enriched in the nuclear proteome analysis. Taken together, these results demonstrate the use of this novel methodology to enrich for chromatin-bound components in the parasite in an unbiased manner.

As cohesin and condensin protein complexes are conserved from bacteria to human they are not among the most effective drug targets. 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 observe a single SMC3 focus at the trophozoite and schizont stages (Figure 3.8B). 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. According to previously published *P. falciparum* nuclear architecture data, the centromeres of all chromosomes cluster together in one region 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. 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 among the Nuclear Matrix Constituent Protein (NMCP) family of proteins and was originally identified as a protein residing on the nuclear periphery in carrots [92]. Previous studies have demonstrated the importance of CRWN proteins in plant viability as evident by the inability to recover mutants with disruptions in the *CRWN* genes [71]. 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 suggests that these NMCP-related proteins might be functionally analogous to components of the animal nuclear lamina [45]. Despite the critical role of lamina proteins 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 [93], lamina-like proteins have not been detected in *Plasmodium* [34]. 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. The CRWN-like protein localizes to a single focus inside the nucleus at ring and schizont stages and it is possible that this protein regulates heterochromatin regions in the nucleus, much like what has been observed in plant species [71]. 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, the identification of novel plant-related proteins that play an important role in parasite nuclear organization can serve as ideal as 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 *P. falciparum* 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 parasite 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 the malaria parasite.

References

1. WHO: **World Malaria Report. 2016.** http://www.who.int/malaria/publications/world-malaria-report-2016/report/en/. 2016.

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

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

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

16. 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.

17. 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.

18. 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.

19. 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.

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

21. 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.

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

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

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

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

30. 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.

31. 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.

32. 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.

33. 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.

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

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

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

42. 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.

43. 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.

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

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

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

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

54. 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.

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

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

57. 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.

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

59. 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.

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

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

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

63. 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.

64. 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.

65. 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.

66. Iborra FJ, Jackson DA, Cook PR: **The case for nuclear translation.** *J Cell Sci* 2004, **117:**5713-5720.

67. Iborra FJ, Jackson DA, Cook PR: **Coupled transcription and translation within nuclei of mammalian cells.** *Science* 2001, **293:**1139-1142.

68. Fromont-Racine M, Senger B, Saveanu C, Fasiolo F: **Ribosome assembly in eukaryotes.** *Gene* 2003, **313:**17-42.

69. 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.

70. Soding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33:**W244-248.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. Trager W, Jensen JB: **Human malaria parasites in continuous culture. 1976.** *J Parasitol* 2005, **91:**484-486.

78. Lambros C, Vanderberg JP: **Synchronization of Plasmodium falciparum erythrocytic stages in culture.** *J Parasitol* 1979, **65:**418-420.

79. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

80. 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.

81. 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.

82. Florens L, Washburn MP: **Proteomic analysis by multidimensional protein identification technology.** *Methods Mol Biol* 2006, **328:**159-175.

83. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *J Am Soc Mass Spectrom* 1994, **5:**976-989.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. De Gaudenzi JG, Noe G, Campo VA, Frasch AC, Cassola A: **Gene expression regulation in trypanosomatids.** *Essays Biochem* 2011, **51:**31-46.

91. 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.

92. 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.

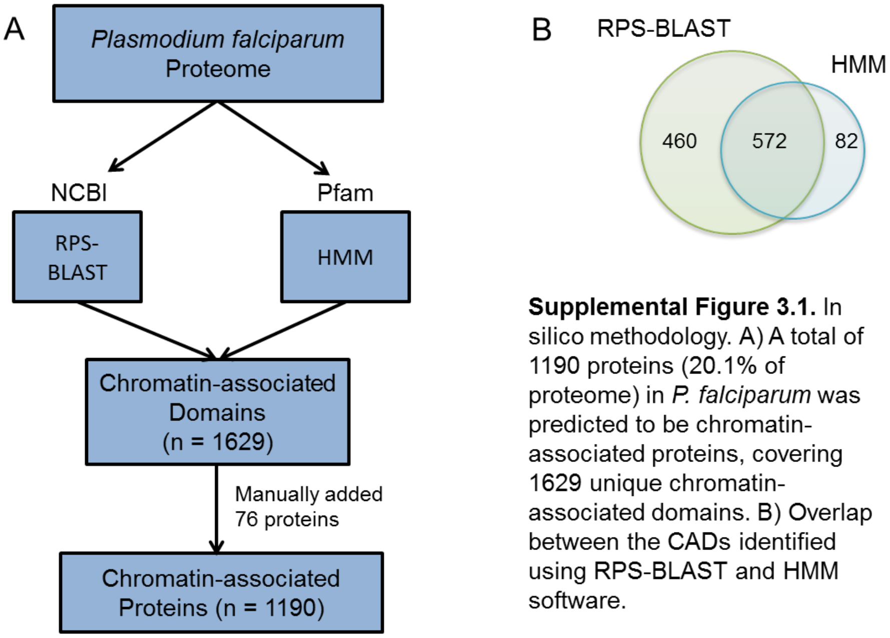
93. DuBois KN, Alsford S, Holden JM, Buisson J, Swiderski M, Bart JM, Ratushny AV, Wan Y, Bastin P, Barry JD, et al: **NUP-1 Is a large coiled-coil nucleoskeletal protein in trypanosomes with lamin-like functions.** *PLoS Biol* 2012, **10:**e1001287.

Supplemental Information

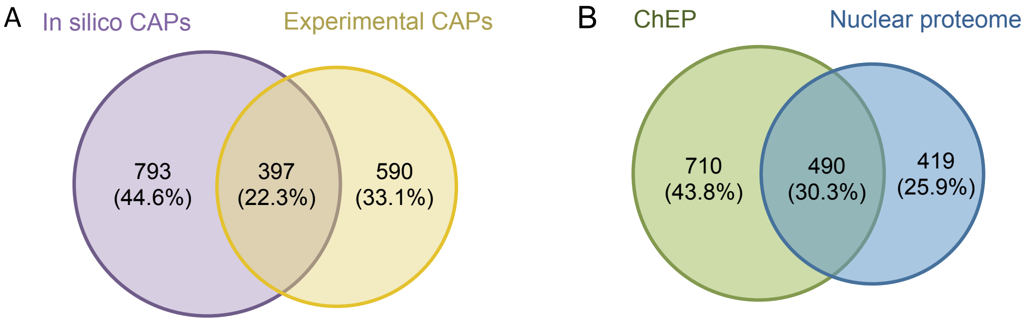
Author's contributions

The co-authors Gayani Batugedara and Xueqing Maggie Lu are equal contributors to this manuscript. Gayani Batugedara performed all wet lab experiments, participated in design of the study, and drafted the manuscript. Xueqing Maggie Lu performed all computational analysis and participated in study design and drafting of the manuscript. Karine Le Roch directed and supervised the research, which forms the basis for this publication.

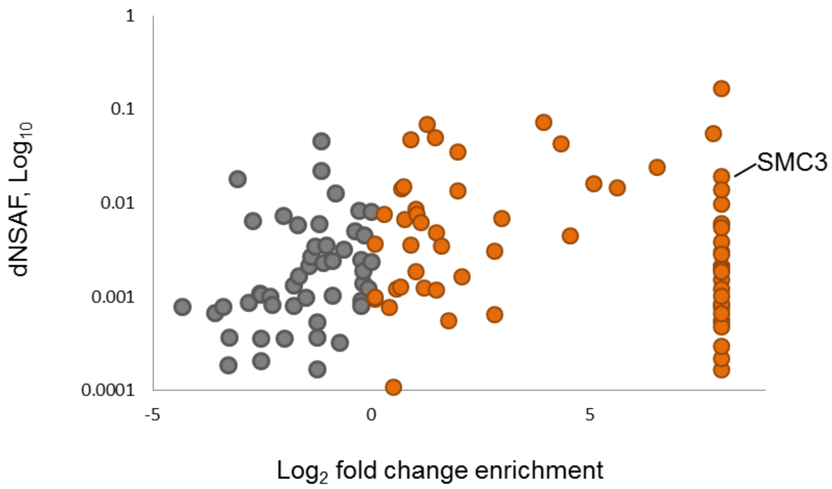
Supplemental Figures


A
Plasmodium falciparum Proteome
NCBI
Pfam
RPS-BLAST
HMM
Chromatin-associated Domains (n = 1629)
Manually added 76 proteins
Chromatin-associated Proteins (n = 1190)
B
RPS-BLAST
HMM
460
572
82

Supplemental Figure 3.1. In silico methodology. A) A total of 1190 proteins (20.1% of proteome) in *P. falciparum* was predicted to be chromatin-associated proteins, covering 1629 unique chromatin-associated domains. B) Overlap between the CADs identified using RPS-BLAST and HMM software.



Supplemental Figure 3.2. Comparison amongin silico CAPs and ChEP and nuclear proteome. A) Comparison of computationally identified CAPs with the ChEP enriched CAPs. B) Comparison of ChEP enriched CAPs with the parasite nuclear proteome enriched proteins based on the number of uniquely detected peptides.



Supplemental Figure 3.3. Fold enrichment of putative proteins interacting with SMC3 in the parasite. Proteins enriched in the immunoprecipitated sample with the SMC3 antibody are indicated in *orange* and proteins enriched in the negative control are indicated in *gray*.

Supplemental Files

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: Enriched chromatin-associated proteins in the ChEP sample, P. falciparum nuclear proteome and the in sillico analysis. (XLSX)

Supplemental File 3.5: Proteins associated with Structural Maintenance of Chromosomes Protein 3 (SMC3) during the IDC. (XLSX)

Chapter 4: Nuclear and Cytoplasmic Long Intergenic Non-coding RNAs in Plasmodium falciparum

Abstract

The most lethal malaria causing parasite, *Plasmodium falciparum*, has a complex life cycle that involves multiple developmental stages and is likely regulated by the coordinate changes in gene expressions. A limited number of transcription factors were identified in the *Plasmodium* genome and increasing evidence has shown that epigenetic regulation and post-transcriptional mechanisms may play essential roles in the parasite's gene regulation system. In all eukaryotes, many long non-coding RNAs have been identified and been shown to be pivotal regulators of genome structure and gene expression. In this chapter, we explore the intergenic long non-coding RNAs distributed in nuclear and cytoplasmic subcellular locations. With the assistance of our recently generated nascent RNA expression profiles, we identified a total of 1,094 lncRNAs, including 574 nuclear enriched lncRNAs, 290 cytoplasmic enriched lncRNAs, and 230 lncRNAs found in both nuclear and cytoplasmic fractions. We observed that these detected lncRNAs are differentially expressed across the parasite's life cycle. The difference in subcellular location and stage specific expression suggests that the lncRNAs are subjected to different regulation processes and have different targets in the genome. It is likely that many of the nuclear enriched lncRNAs are critical for regulation of chromatin structure, while the cytoplasmic enriched lncRNA may be involved in post-transcriptional and translational regulations. Some of these lncRNAs may also have a role in protein complex formation. Further in-depth study and experimental confirmation of these lncRNAs would provide significant insights into the gene expression and regulation systems of this lethal parasite.

Introduction

Malaria is a mosquito-borne infectious disease caused by the protozoan parasite of the genus *Plasmodium*. Among the human-infection *plasmodium* species, *Plasmodium falciparum* is responsible for the most severe forms of malaria and the death of nearly half a million people each year [1]. The parasite has a complex life cycle involving multiple biological stages in both human and mosquito hosts. During the erythrocytic stage of the asexual phase in a human, the parasite replicates in red blood cells and goes through ring, trophozoite and schizont stages. Each 48-hour cell division cycle will produce 16-32 daughter cells or merozoites that are released into the blood stream for new invasions. Large accumulation of infected red blood cells will result in a blockage of the vascular system and cause the host to suffer from recurring fevers, resulting in death of the host in severe cases. In addition, a subset of parasites will sexually develop into gametocytes circulating in the blood stream that are ready to be picked up by a mosquito during a blood meal. The parasite will then develop through asexual cycle in the mosquito and be ready for new malarial infection within the next person. This multiple-stage processing is tightly regulated, but the mechanisms regulating these events are still not well understood.

Compared to other eukaryotes with similar genome size, *P. falciparum* has an extremely AT-rich genome and a relatively low number of detected transcription factors. Thought it is still unclear how this parasite manages its gene expression to coordinate its complex cell cycle, evidence has increasingly shown that epigenetic regulation plays a major role in initiating and directing transcriptional process in *P. falciparum*. In addition, nascent RNA expression profiles revealed that a majority of the genes are transcribed during the trophozoite stage and that the cascade of gene expression observed using messenger RNA (mRNA) is likely the result of a combination of transcriptional [2-6] and post-transcription regulation events [7-11]. Taken together, these data highlight the importance of mechanisms involved in epigenetics and post-transcriptional regulations.

In the past decade, advances in biotechnologies and next generation sequencing technologies have led to huge progress in genomic study and revealed that transcriptome is much larger than what we once expected. Recent evidence suggests that approximately 80 to 90% of the human genome might be transcribed in at least one cell type [12-14]. Though some scientists have argued that many of these transcripts are a result of transcriptional noise, many believe that these transcripts are functional elements with biological roles that are yet to be discovered. It is now becoming clear that what made up this large number of transcripts are non-coding RNAs (ncRNA) with diverse regulatory roles in an orgasm's biological system.

One class of ncRNAs is the long noncoding RNAs (lncRNA). LncRNAs are defined as none protein encoding RNA molecules with length of 200 nucleotides or longer. Many lncRNAs share features with the 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; lncRNA enriched in the nuclear fraction often associated with regulation of epigenetic and transcription regulation [15-18], while lncRNA enriched in the cytoplasm are associated with mRNA processing, post-transcriptional regulation, translational regulation, and cellular signaling process [15, 19-21]. In *Plasmodium falciparum,* though many lncRNAs have been identified, the biological significance of these lncRNAs remains elusive. Majority of the well-studied lncRNAs have been nuclear lncRNA and are thought to be involved in different aspects of chromatin biology. Some of the best-characterized lncRNAs are the telomere-associated repetitive element transcripts (lncRNA-TAREs) [22] and telomeric repeat containing lncRNAs (TERRA) [23]. Both type lncRNAs showed stage-specific expression preferences and are suggested to be associated with heterochromatin environment maintenance and *var* gene regulation. To better understand the biological significance of lncRNAs, we further explore the transcriptome of *P. falciparum* in both nuclear and cytoplasmic fractions. We then attempt to characterize lncRNAs base on their cellular location, peak of expression, physical properties, and epigenetic regulatory marks.

Result

Identification of lncRNAs

To further explore the lncRNA populations in *P. falciparum we extracted total RNA from both* nuclear and cytoplasmic fractions from synchronized parasite cultures at ring (0 hpi), trophozoite (18hpi), schizont (36hpi), and gametocyte stage. Blood smears were used to assess the development of parasite progression (Figure 4.1A). In brief, synchronized parasites are collected from cell culture followed by a modified cell fractionation procedure described in PARIS kit (ThermoFisher). See Methods for detailed cell fractionation procedure. Successful isolation of both subcellular fractions was validated using western blot with the Anti-histone H3 antibody as a nuclear specific marker, and an anti-aldolase antibody as a marker specific to the cytoplasmic fraction (Figure 4.1B). After separation of nuclear material from the cytoplasmic material, total RNAs are extracted using Trizol LS Reagent for both fractions and polyadenylated mRNA are isolated from total RNA. Strand-specific libraries were then prepared and sequenced on the Illumina Next500 sequencing platform. Reads were trimmed and mapped to *P.falicpuarm* genome (v34) using HiSAT2 and only uniquely mapped reads were used for downstream applications. Details about read processing and mapping can be found in the Methods section. As verification, we calculated the spearman correlation in gene expression levels among nuclear samples, cytoplasmic samples, and a previous published steady-state total mRNA dataset generated in our lab [24]. Spearman correlation coefficients are reported in supplemental figure 4.2. Once validated, a computational pipeline was implemented for the identification of lncRNAs. In brief, we first merged all nuclear and cytoplasmic libraries into one dataset, resulting in one single file. We then assembled the nuclear and cytosol transcriptome independently using cufflinks. After transcriptome assembly, we filtered transcripts based on their length, expression level, presence of primary transcript from our GRO-seq dataset, and sequence coding potential (Figure 4.1A). We removed any predicted transcripts that overlap with annotated genes and focused on lncRNA candidates within the intergenic regions. Our goal was to select transcripts that are reasonably long, consistently expressed in both published nascent RNA and steady-state mRNA expression profiles, and that are likely to be non protein-encoding genes. As a result, we identify a total of 1,094 lncRNAs. Three hundred ninety-five lncRNAs (36%) overlapped with previously identified intergenic lncRNA in [25, 26], and 699 lncRNAs were identified as novel in *P. falciparum* (Figure 4.1C).

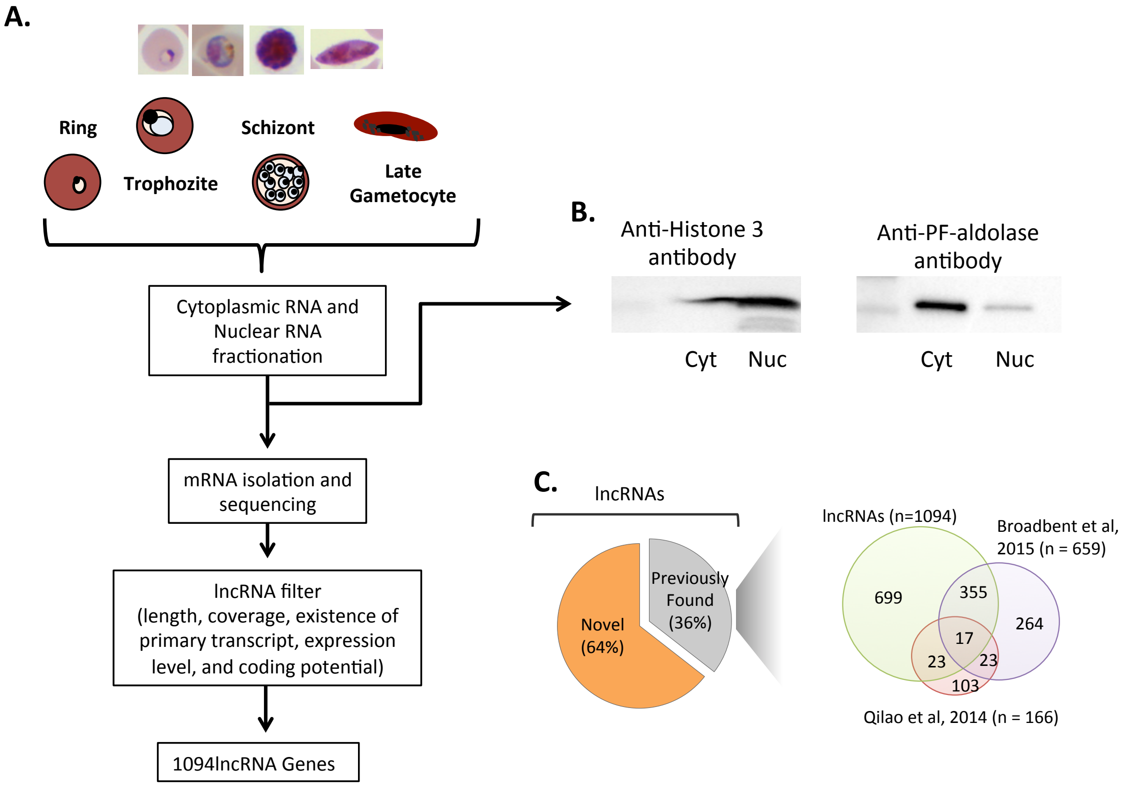

A.
Ring
Trophozite
Schizont
Late Gametocyte
Cytoplasmic RNA and Nuclear RNA fractionation
mRNA isolation and sequencing
lncRNA filter (length, coverage, existence of primary transcript, expression level, and coding potential)
1094lncRNA Genes
B.
Anti-Histone 3 antibody
Cyt
Nuc
Anti-PF-aldolase antibody
Cyt
Nuc
C.
lncRNAs
Novel (64%)
Previously Found (36%)
lncRNAs (n=1094)
Broadbent et al, 2015 (n = 659)
Qilao et al, 2014 (n = 166)
699
355
264
17
23
23
103

Figure 4.1. Nuclear and cytoplasmic lncRNA identification. (A) An general overview of the lncRNA identification pipeline. (B) Cell fractionation efficiency valdiation using anti-histone H3 and anti-aldolases 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 indistinguished lncRNAs that are equally distributed in both fractions. Among the 1,094 lncRNAs, 574 lncRNAs (52%) showed enrichment in the nuclear fraction, 290 lncRNAs (27%) showed enrichment in the cytoplasmic fraction, and 230 indistinguished lncRNAs (21%) showed similar distribution between both subcellular fractions (Figure 4.2A). We then 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). We then estimated the

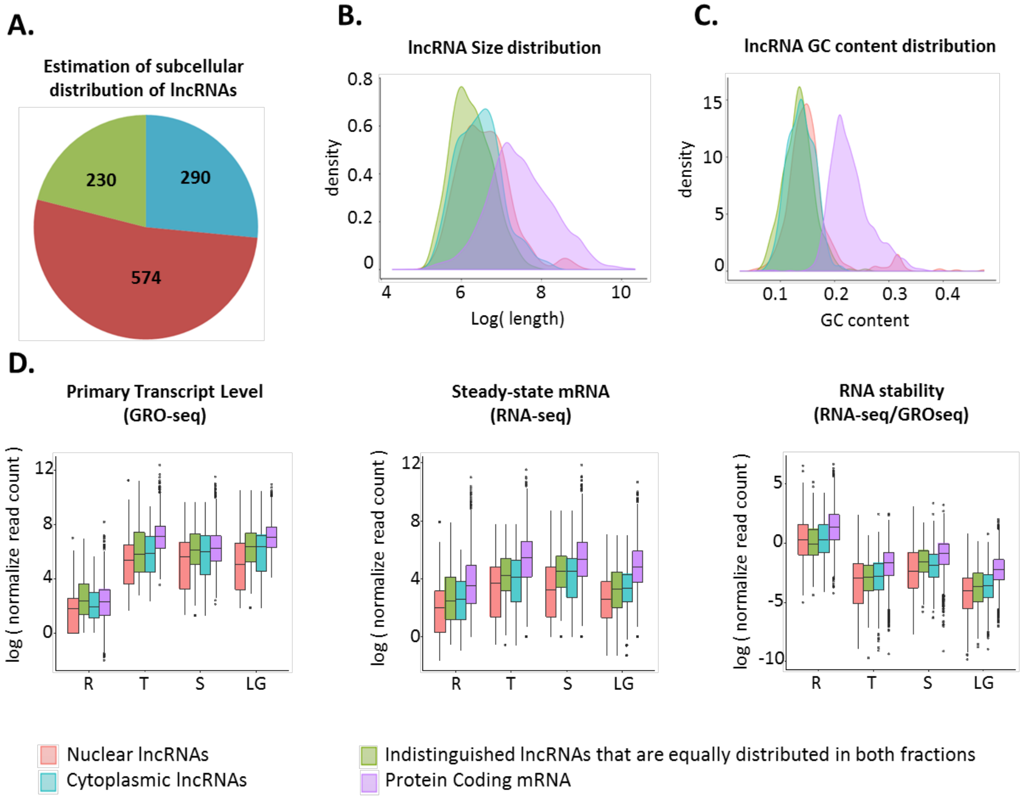


Figure 4.2. (A) A total of 1,094 lncRNA candidate was identified, covering 574 nuclear enriched lncRNA, 290 cytoplasmic enriched lncRNAs, and 230 lncRNAs found in both fractions. Density plot of the size (B) and GC contents (C) of lncRNA candidates and annotated protein encoding mRNAs. Expression level of primary transcripts (left), steady-state mRNA (middle), and relative stability (right) of lncRNA candidates and annotated protein encoding mRNAs.

expression levels and stability of the lncRNAs by using total steady-state mRNA expression profile and nascent RNA expression profile. RNA stability was calculated as the ratio between steady-state mRNA expression levels over nascent RNA expression levels. We found 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 in consistent with previous lncRNA annotation studies in human breast cancer cells [27] and noncoding RNA stability studies in mammalian genomes [28]. 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 the 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.

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 these candidate lncRNAs across cell cycle. Using k-mean clustering, we were able to group these lncRNAs into 7 distinct clusters (Figure 4.3A). Generally, nearly all lncRNAS showed strong coordinated cascade throughout the parasite’s cell cycle. A larger fraction of the lncRNAs are 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 and are also moderately expressed in the nuclear fraction of schizont stages. An example of lncRNAs for this cluster is the lncRNA-TAREs. We observed that all identified lncRNA-TAREs are clustered into this group with an average 2.01 log two fold changes 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 the 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 the DNA replication and the peak of transcriptional activity. Base 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 expressed in gametocyte are surrounded by gametocyte specific genes and are located within heterochromatin regions marked by H3K9me3 (unpublished data) between the asexual and sexual stages (Figure 4.3E and Figure 4.3F). In contrast, we observed few lncRNAs that are solely expressed during the asexual cycle with distinct changes in heterochromatin marks (Figure 4.3D). The presences of some of these lncRNAs are confirmed using reverse transcription polymerase chain reaction (RT-PCR) as demonstrated in Supplemental Figure 4.1.

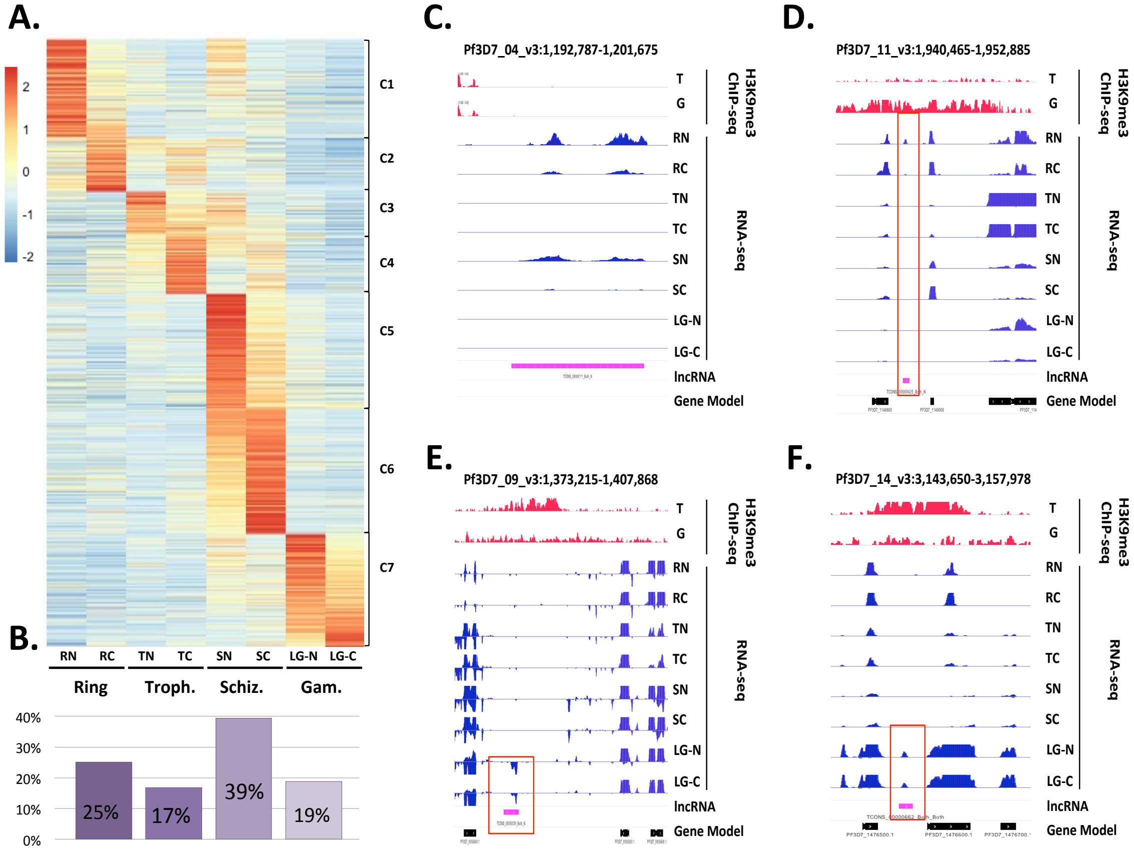


Figure 4.3. Gene expression pattern of lncRNAs. (A) lncRNAs are grouped into 7 clusters based on their cell cycle expression patterns. (B) Percentage of lncRNAs that are highly expressed in each subcellular fractions at ring, trophozoite, schizont, and late gametocyte stage. Genome browser view of H3K9me3 ChIP-seq and RNA-seq datasets on one identified lncRNA-TARE located at the right arm of chromosome 4 (C), one asexual-specific lncRNA (D), and two gametocyte-specific lncRNAs located at the intergenic regions of chromosome 9 (E) and 14 (E).

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 values as previously described in [29]. Two synchronization steps were performed with 5% D-sorbitol treatments at ring stage with eight hours apart. Parasites were collected every 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* strain NF54 strain and were harvested 15 days (stage IV to V) after the induction procedure as previously described in [30].

Nuclear and cytosolic RNA isolation

Highly synchronized parasites were first extracted using 0.15% saponin solution followed by centrifuge at 4500 rpm for 10 minute on ice. Parasite pellets are then washed twice with ice cold PBS and re-collected at 5000 rpm at 4 ^{o}C. Parasite pellets are resuspended in 500 uL ice cold Cell Fractionation Buffer (PARIS kit, ThermoFisher; AM1921) with 10 uL of RNAase Inhibitor (SUPERaseIn 20U/uL, Invitrogen; AM2694). Gently resuspend the cells by pipetting and incubate on ice for 10 minutes. Centrifuge samples for 5 minutes at 4 ^{o}C and 500 xg. After centrifuge, carefully collect the supernatant containing cytoplasmic fraction with micropipetor. Resuspend the nuclear fraction in 500 uL Cell fractionation buffer and 15uL RNAse Inhibitor as described above. To obtain a more purified nuclear fraction, syringe the pellet with 26G ½ inch needle for five times. Incubate on ice for 10 minutes and centrifuge sample again for 5 minutes at 4 ^{o}C and 500 xg. Discard the supernatant and resuspend the nuclear pellet with 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 37 ^{o}C pre-warmed Trizol LS Reagent (Life Technologies, Carlsbad, CA, USA) followed by a 5 minute incubation at 37 ^{o}C. RNAs are 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 PCRs on *Pf*-Alba 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 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 in the following for 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 75bp 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 were 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) [31]. All trimmed reads were then mapped to *P.falciparum* genome (v34) using HISAT2 with the following parameters: –t, -- downstream-transcriptome-assembly, --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 propery paired (samtools v 0.1.19-44428cd [32]), and are likely to be PCR duplicates (Picard tools v1.78[33]). 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 lncRNA in the nuclear and cytoplasmic fraction, we first merged all nuclear libraries and cytoplasmic liberties into two sets: one nuclear library set and one cytoplasmic library set.

Thenk cufflinks (v2.1.1 [34]) was used for transcriptome assembly with the following parameters: -p 8 -b PlasmoDB-34_Pfalciparum3D7_Genome.fasta -M PlasmoDB-34_Pfalciparum3D7.gff --library-type fr-firststrand -I 5000. After obtained the assembled transripts, 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 transcripts with a size shorter than 200bp were also excluded from down stream analysis. Next, for each remaining transcripts, we calculated its primary transcription level using GRO-seq dataset (GSE85478) from [11] and removed any transcripts that has a read coverage fall below the 15% of median of expression levels of all protein encoding genes. After filter 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 remained, we then calculated its protein potential using Coding Potential Calculator (http://cpc.cbi.pku.edu.cn/). Any lncRNA that were predicted to be coding or weak non-coding RNA or with a coding protentail score above -1 were 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 lncRNA showed 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 [35] 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) from [24] using BEDTools v2.25.0 [35]. The read counts are then normalized as described in the original publication, and ratios between RNA-seq and GRO-seq coverage values are 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

Eight 25ml of mixed population of parasite were collected as described before. Gently resuspend parasite pellets in 500ul 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). Incubate solution on ice for 10 minutes and pellet sample for 5 minutes at 4 ^{o}C and 500 xg. Carefully collect the supernatant containing cytoplasmic fraction and resuspend the nuclear fraction in 500 uL Cell Fractionation Buffer followed by 5 times of syringe using 26 ½ inch needle. Pellet nuclei again at 4 ^{o}C and 500 xg. Discard the supernatant and resuspend the nuclei pellet in 500 uL of Cell Disruption Buffer (PARIS kit, ThermoFisher; AM1921) and incubate on ice for 10 minutes. The nuclear fraction is then sonicated seven times 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 13,000 rpm to remove cell debris. Seven micrograms parasite cytoplasmic and nuclear protein lysates were diluted in 2X laemmli buffer at a 1:1 ratio followed by heatings 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 and run for 1 hour at 18 V, then stained using comercial antibodies generate against histone H3 (1: 3,000 dilution, abcam; ab1791) and *plasmdium* aldolase (1:1,000 dilution, abcam; ab207494), and secondary antibody, Goat Anti-Rabbit IgG HRP Conjugate (1:25,000 dilution, Bio-Rad; 1706515). Finally, membrane were visualized using the Bio-Rad ChemidDoc MP Gel Imager.

lncRNA validation using reverse transcription polymerase chain reaction (RT-PCR)

Total RNA was isolated from 10 ml of non-synchronous erythrocytic stage *P. falciparum* culture and 25 ml of late gametocyte stage culture. Total RNA quality was checked on agarose gel and genomic DNA contamination were removed using DNA-free DNA removal kit (ThermoFisher; AM1906) according to manufacturer's instruction. The absence of genomic DNA contamination was validated using a primer set targeting an intergenic region and a primer set targeting PfAlba3 (PF3D7_1006200) from inside exon 1 to within exon 2. Amplification of genomic DNA should give a product with a size of 429 bp including the intronic sequence, whereas amplification of cDNA should result in a fragment with a size of 164 bp. Approximately 1.1 μg of DNase I treated RNA from each sample with 35 PCR cylcle were used to confirm the absent of genomic DNA contamination. In addition, PCR sample with no DNA template was used as negative control (Supplemental Figure 4.1). 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 $MgCl_2$, 4 μl 0.1 M DTT, 2 μl 20 U/μl SuperaseIn and 1 μl 200 U/μl SuperScript III Reverse Transcriptase (all from 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 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). Mixture are 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. Five µl of each PCR sample was used for agarose gel electrophoresis. All primer used for PCR validation are listed in Supplemental File 4.1.

Discussion and Conclusion

The preliminary work described in this chapter is the first genome-wide global and cell cycle detection of lncRNAs from different subcellular locations in *P. falciparum*. Using both cell fraction experimental and computational pipelines, we identified 1,094 lncRNAs covering 574 nuclear enriched, 290 cytoplasmic enriched, and 230 indistinguished lncRNAs that are localized in both fractions. By utilizing nascent RNA expression profiles (GRO-seq dataset), we were able to significantly improve the sensitivity of lncRNA detection, especially for the identification of nuclear lncRNAs. This study revealed 699 lncRNAs that had not been described previously. 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 interfering and regulating gene expression activity [36, 37], guiding or enhancing the functions of regulatory proteins [16, 38-41], or assisting the alteration of chromatin structures by shaping three-dimensional (3D) genome organization[17, 42-44]. Some of the well-characterized nuclear lncRNAs, such as XIST[45], FIRRE[46], and NEAT [47], were shown to be particularly important for nuclear organizing and chromatin conformation change. In *P. falciparum,* emerging evidence has shown that chromatin structure and chromatin organization are of vital importance for the parasite's gene expression and regulation system. Therefore, identification of nuclear enriched lncRNAs may help us to discover chromatin-associated regulators in this parasite. In our present work, we observed that a large number of lncRNAs, including the lncRNA-TAREs, are very 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 environment inducement or chromatin structure re-organization events. This is in line with previous publications that showed condensed chromatin structure at the ring and schizont stages. In addition, we observed that some of the lncRNAs are neighbored with stage-specific genes (i.e., Gametocyte-specific genes or erythrocyte exported genes). This finding implies that lncRNAs found in these phenomena may be involved in local gene regulation and affect the expression level of stage-specific genes.

Compared to nuclear lncRNAs, progress in functional analysis of lncRNA in the *Plasmodium* cytoplasmic faction is significantly lacking. In the last decade, many lncRNAs have been discovered with diverse cellular functions outside of the nucleus. This type of lncRNA has been reported to interact with ribosome[20] and is often associated with post-transcriptional and translational controls [19]. Some cytoplasmic lncRNAs, such as half-STAU1-binding site RNAs (1/2-sbsRNAs) [48, 49] and growth arrested DNA-damage inducible gene 7 (gadd7) [50] , are shown to be able to alter the stability of mRNA, while some cytoplasmic lncRNAs including lncRNA-p21 [51] and AS UCHl1 [52] are shown to be modulating either the repression or promotion of translational process. As of today, some studies of *P. falciparum* have mentioned or used cytoplasmic RNA populations as a comparison control for nuclear RNAs [53], but no one has specifically investigated the function of any of these lncRNAs. The dataset generated from this study provided the first good global view of cytoplasmic lncRNAs expressing 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 detected. In addition, we observed that a small group of cytoplasmic lncRNAs is highly expressed at the trophozoite stage, a stage where massive transcription activity was observed in previous studies [11]. Though more in-depth studies will be required to confirm the functions of these trophozoite-specific cytoplasmic lncRNAs, it is possible that some of these lncRNAs are involved in mRNA stability, alternative splicing, or translational regulation of the transcribed coding mRNAs at the trophozoite stage.

Compared to lncRNA studies in other eukaryotes, the field of lncRNA in Plasmodium is still young, yet full of potential. First, analysis of promoter and gene body regions with available histone modifications datasets (H3K9me3, H3K36me3, H3K9ac) are still required for further annotation of these candidate lncRNAs. In addition, understanding of how these lncRNAs may contribute to the sexual differentiation or promotion of cell progression is still a work in progress. Experimental knockdown or conditional knockout of these lncRNAs will provide new insights into chromatin biology as well as transcriptional and translational regulation processes of this parasite. We are hoping that this newly generated dataset will not only assist future lncRNA studies in this parasite, but also help to identify parasite-specific gene expression regulators that can ultimately be used as new anti-malarial drug targets.

Reference

1. WHO: **World Malaria Report. 2017.** 2017.

2. 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.

3. Duffy MF, Selvarajah SA, Josling GA, Petter M: **Epigenetic regulation of the Plasmodium falciparum genome.** *Brief Funct Genomics* 2014, **13:**203-216.

4. 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.

5. 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.

6. Rai P, Sharma D, Soni R, Khatoon N, Sharma B, Bhatt TK: **Plasmodium falciparum apicoplast and its transcriptional regulation through calcium signaling.** *J Microbiol* 2017, **55:**231-236.

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. 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.

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. Lacsina JR, LaMonte G, Nicchitta CV, Chi JT: **Polysome profiling of the malaria parasite Plasmodium falciparum.** *Mol Biochem Parasitol* 2011, **179:**42-46.

11. 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.

12. Pertea M: **The human transcriptome: an unfinished story.** *Genes (Basel)* 2012, **3:**344-360.

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

14. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

15. Ransohoff JD, Wei Y, Khavari PA: **The functions and unique features of long intergenic non-coding RNA.** *Nat Rev Mol Cell Biol* 2017.

16. 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.

17. Quinodoz S, Guttman M: **Long noncoding RNAs: an emerging link between gene regulation and nuclear organization.** *Trends Cell Biol* 2014, **24:**651-663.

18. Nakagawa S, Kageyama Y: **Nuclear lncRNAs as epigenetic regulators-beyond skepticism.** *Biochim Biophys Acta* 2014, **1839:**215-222.

19. Rashid F, Shah A, Shan G: **Long Non-coding RNAs in the Cytoplasm.** *Genomics Proteomics Bioinformatics* 2016, **14:**73-80.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. Trager W, Jensen JB: **Human malaria parasites in continuous culture.** *Science* 1976, **193:**673-675.

30. Ifediba T, Vanderberg JP: **Complete in vitro maturation of Plasmodium falciparum gametocytes.** *Nature* 1981, **294:**364-366.

31. Joshi NA FJ: **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files** pp. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files 2011:Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

33. Picard.

34. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7:**562-578.

35. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

36. Guil S, Esteller M: **Cis-acting noncoding RNAs: friends and foes.** *Nat Struct Mol Biol* 2012, **19:**1068-1075.

37. 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.

38. 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.

39. 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.

40. 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.

41. Detmer DE, Tyron TJ: **Delivery of surgical care: inferences based on hospital discharge abstract data.** *Surg Forum* 1976, **27:**460-463.

42. Mele M, Rinn JL: **"Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription.** *Mol Cell* 2016, **62:**657-664.

43. Rinn J, Guttman M: **RNA Function. RNA and dynamic nuclear organization.** *Science* 2014, **345:**1240-1241.

44. Caudron-Herger M, Rippe K: **Nuclear architecture by RNA.** *Curr Opin Genet Dev* 2012, **22:**179-187.

45. Cerase A, Pintacuda G, Tattermusch A, Avner P: **Xist localization and function: new insights from multiple levels.** *Genome Biol* 2015, **16:**166.

46. 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.

47. 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.

48. Gong C, Maquat LE: **lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.** *Nature* 2011, **470:**284-288.

49. 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.

50. 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.

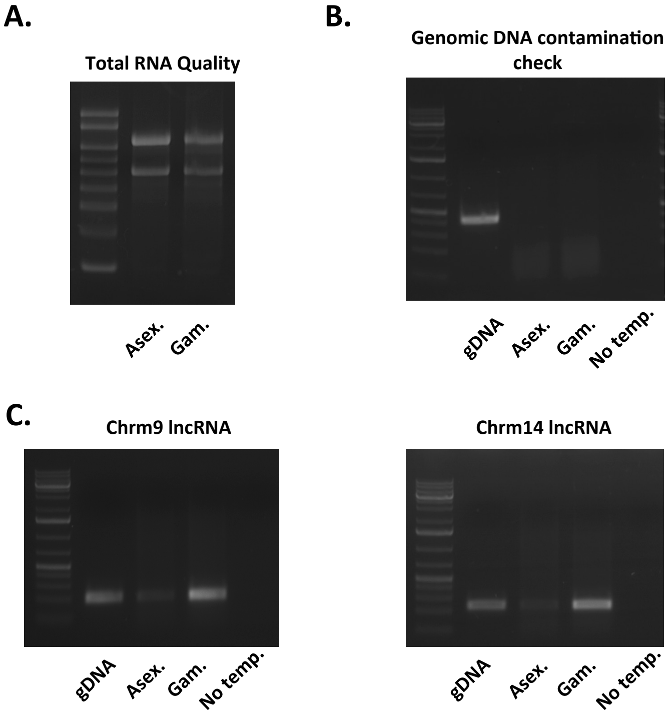
51. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M: **LincRNA-p21 suppresses target mRNA translation.** *Mol Cell* 2012, **47:**648-655.

52. 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.

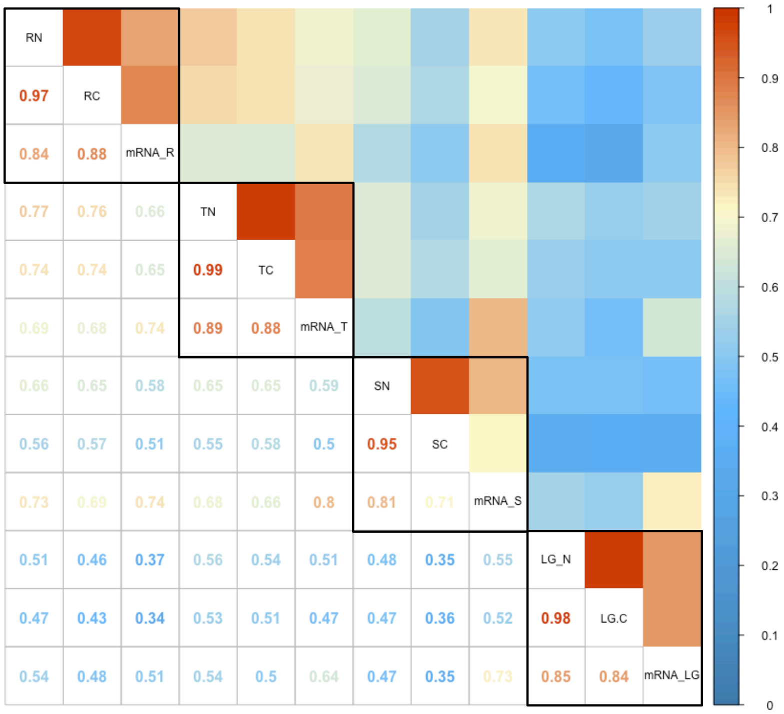
53. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, Martins RM, Sismeiro O, Coppee JY, Scherf A: **Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum.** *BMC Genomics* 2014, **15:**150.

Supplemental Information

Supplemental Figures



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 on agarose gel. (B) Genomic DNA was removed and verified using reverse transcription polymerase chain reaction (RT-PCR) with primers designed to amplify a fragment of 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 stage with high level of H3K9me3 mark.



Supplemental Figure 4.2. Spearman correlations in gene expression levels among nuclear fraction, cytoplasmic fraction, and steady-state mRNA across *P. falciparum* cell cycle.

Supplemental Files

Supplemental File 4.1: identified lncRNAs. (XLSX)

Conclusion

Concluding remarks

Malaria has been one of the most ancient and lethal human infectious diseases known to mankind. For many centuries, humans have been battling and attempting to strike down this malicious disease. However, due to the lack of efficient vaccines, limited drug access, and the rapid development of anti-malaria drug resistant parasite strains, malaria still remains a big health burden in many developing countries. Although malaria is curable, resistant strains have been reported for all of the popular anti-malarial drugs. Therefore, the identification of new drug targets and anti-malaria compounds is urgently needed.

Of the five *Plasmodium* species that cause human infection, *Plasmodium falciparum* is responsible for the most severe and lethal cases of malaria. *P. falciparum* has a complex life cycle with multiple developmental stages and hosts. This life cycle is tightly regulated possibly by the orderly changes in gene expressions. As of today, a limited number of transcriptional factors have been identified in this parasite. Accumulating evidence suggests that *P. falciparum* may use chromatin structure and post-transcriptional elements as alternative mechanisms for its tight transcription regulation. To gain a better understanding of the gene expression and regulation system of the parasite, we explored the transcriptome, epigenome, and proteome of *P. falciparum* in this dissertation work.

First, using nucleosome-positioning landscape, we were able to identify 231 novel putative genes. We demonstrated that nucleosome positioning can be used for gene identification, especially for organisms with high nucleotide bias genomes. Secondly, by generating and analyzing the nascent RNA expression profile, we showed that a majority of the genes are actively transcribed at the trophozoite stage. Data from this study suggested that chromatin structure may provide a basal control for transcription activity. Additionally, the cascade of gene expression observed in steady-state mRNA is likely contributed by various post-transcriptional regulation processes. These findings explain why a weak correlation between the “open-and-closed” chromatin re-organization event and gene expression, measured by steady-state mRNA, was observed in previous studies. As chromatin structure is particularly important for the global transcriptional activity in this parasite, chromatin-associated regulators may be highly effective targets for anti-malaria therapy. Therefore, in the later chapters of this work, we focused on the identifications of potential chromatin-associated regulators such as proteins and lncRNAs. By carefully surveying the proteome, we were able to generate the first and the most up-to-date comprehensive overview of the *plasmodium* chromatin-associated proteome. We also experimentally validated and annotated a few of these important chromatin structural proteins. Since there has been limited knowledge of *Plasmodium* chromatin-associated proteins, further investigation and functional study of these chromatin-associated proteins would significantly advance our knowledge on *P. falciparum* chromatin biology, thus bringing helpful insights to better understand the gene expression systems in this parasite. Beside chromatin-associated proteins, long non-coding RNAs have also been shown to have important roles in both chromatin and post-transcriptional regulation. Therefore, in the last chapter of this dissertation, we performed genome-wide identifications of both nuclear lncRNAs and cytoplasmic lncRNAs within the *P. falciparum* genome. As a result, we identified 1,094 lncRNAs that are differentially expressed not only between subcellular locations but also across the parasite’s life cycle. It is likely that some of the nuclear expressed lncRNAs are critical for chromatin structure regulation, while the cytoplasmic expressed lncRNAs are important for post-transcriptional or translational regulations.

Taken together, the work presented in this dissertation confirmed the strong regulatory roles of chromatin structure in initiating global transcriptional activity; clarified the time of transcription for a majority of the *P. falciparum* genes; and prioritized many proteins and lncRNAs, which are likely to be associated with chromatin regulation, for future malaria studies. The end goal of this dissertation work is to provide new insights and generate meaningful data that will become stepping-stones to further assist the identification of new anti-malaria tools and therapeutic strategies.

Future directions and some after thoughts

Though substantial progress has been made in understanding the chromatin biology, gene expression, and regulation process in *P. falciparum*, many questions are still waiting to be answered. Given the importance of the epigenome for the parasite's gene expression, pinpointing new potential drug targets is still a work in progress. In the past decade, several histone deacetylase (HDAC) inhibitors have been the focus of research. HDAC inhibitors, such as apicidin, trichostatin A(TSA), suberoylanilide hydroxamic acid (SAHA), and 2-aminosuberic acid derivative (2-ASA-9), have all been previously shown to have a profound transcriptional effect on the cascade of gene expression[1, 2]. Together with other HDACs inhibition studies in apicomplexa parasites[3], HDACs are now promising antimalarial drug targets. However, as many of the histone-modifying enzymes are conserved among eukaryotic organisms, HDACs targeted therapies may be toxic to the human host. Therefore, further identification of parasite-specific elements that may interrupt the chromatin structure or the epigenetic control of gene expression is still needed. A number of potential chromatin associated proteins and lncRNAs have been identified in this dissertation. While extensive experimental and functional studies are needed to validate some of these key regulators, they provide possible new targets for novel therapeutic strategies. For example, it is well known that the function of histone tail modification is to provide recognition signals to facilitate the recruitment or stabilization of chromatin-related protein complexes in Eukaryotes. In P. falciparum, though histone readers, bromodomain proteins, PHD fingers, and proteins containing the royal family (i.e. Tudor, Chromo, and MBT domains) have been identified [4], their exact role, binding sites, and associated-pathways remain unclear. Another area of Plasmodium study that is waiting to be expended is the study of molecular component controlling nuclear architecture. Recent studies have shown that the Plasmodium telomere ends are clustered together and that this telomeric heterochromatin environment is important for expression and regulation of genes involved in virulence factors [5-7]. Further mapping and study of the nuclear compartments may provide new insights for our understanding of how nuclear localization of a gene influences its chromatin state or vice versa. While we conducted a large number of genome-wide studies and identified many molecular components that still need to be experimentally validated, we drew a more comprehensive map of gene expression system in this parasite. Most importantly, the work presented in this dissertation will provide significant insights and guidance for future malaria studies.

Reference

1. Andrews KT, Gupta AP, Tran TN, Fairlie DP, Gobert GN, Bozdech Z: **Comparative gene expression profiling of P. falciparum malaria parasites exposed to three different histone deacetylase inhibitors.** *PLoS One* 2012, **7:**e31847.

2. Chaal BK, Gupta AP, Wastuwidyaningtyas BD, Luah YH, Bozdech Z: **Histone deacetylases play a major role in the transcriptional regulation of the Plasmodium falciparum life cycle.** *PLoS Pathog* 2010, **6:**e1000737.

3. Bougdour A, Maubon D, Baldacci P, Ortet P, Bastien O, Bouillon A, Barale JC, Pelloux H, Menard R, Hakimi MA: **Drug inhibition of HDAC3 and epigenetic control of differentiation in Apicomplexa parasites.** *J Exp Med* 2009, **206:**953-966.

4. Cui L, Miao J: **Chromatin-mediated epigenetic regulation in the malaria parasite Plasmodium falciparum.** *Eukaryot Cell* 2010, **9:**1138-1149.

5. 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.

6. 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.

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