UNIVERSITY OF CALIFORNIA
RIVERSIDE

Vector Borne Disease: Viruses and Antiviral Immunity in *Culex* Mosquitoes and New Insights into Gene Regulation in Malaria Parasites

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

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

Steven M. Abel

March 2024

Dissertation Committee:
Dr. Karine G. Le Roch, Chairperson
Dr. Naoki Yamanaka
Dr. Stefano Lonardi

The Dissertation of Steven M. Abel is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I want to thank my major professor, Dr, Karine Le Roch, for giving me an opportunity to do research in her lab and to expand my skills and learn lab skills beyond the computational ones that I was more familiar with. My other dissertation committee members were also very helpful throughout the process. West Valley and Coachella Valley Mosquito & Vector Control Districts collected and provided invaluable samples for my research. Many collaborators provided opportunities to work on important projects, especially Dr. Rita Tewari and her lab at University of Nottingham.

I must especially thank those who helped me find my footing when I first arrived at UCR. Lab manager Jacques and former graduate students Maggie and Gayani, between them, introduced me to culturing and lab bench techniques, put me on the path to learning the correct ways to do bioinformatics, and helped me to learn how to navigate graduate school. I also want to acknowledge the other labs that I rotated in: the Judelson and Murn Labs, some of whom spent much more time teaching me than they needed to. I am very grateful for the help of all of the above.

I certainly want to acknowledge those who made the long haul of doing a PhD an enjoyable experience: Thomas (who I also learned much from), Zeinab, Todd, Desiree, Mohit, Trevor, Tina, Anthony, Loic, and all others of the lab. This group made coming to work every day more fun than I could have expected.

Last but not least, a special acknowledgment for my parents, my brother, my dear friends, and my dog Lucky for helping me through a time that was very enriching professionally but of course presented inevitable difficulties. Thank you!

ABSTRACT OF THE DISSERTATION

Vector Borne Disease: Viruses and Antiviral Immunity in *Culex* Mosquitoes and New Insights into Gene Regulation in Malaria Parasites

by

Steven Matthew Abel

Doctor of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, March 2024
Dr. Karine G. Le Roch, Chairperson

Mosquitoes harbor and transmit a variety of pathogens which are dangerous to humans and incur a constant and significant public health cost. Gaining more detailed knowledge about these pathogens, their interactions with the mosquito, and their molecular biology and genetics will allow for new techniques to be developed to prevent harmful effects on humans. These pathogens vary in complexity from viruses like West Nile virus to eukaryotic apicomplexan parasites like *Plasmodium falciparum*, the human malaria parasite.

Culex mosquitoes routinely transfer viruses like West Nile virus in the United States representing a predominant health threat there. These mosquitoes are also routinely infected by a wide variety of other viruses, which have received very little research attention, yet could affect the transmission of pathogens like West Nile virus and St. Louis encephalitis virus. Therefore, in the first chapter of this dissertation work, we

performed small RNA sequencing to examine the full virome of field-caught *Culex* mosquitoes in multiple geographical regions of southern California. These data also allowed us to analyze the interactions between viruses and identify potential pairs of co-infecting or mutually excluding viruses, as well as closely look at the small RNA immune response of *Culex* mosquitoes against these viruses, expanding the known role of the antiviral piRNA response in *Culex*. We also identified mosquito miRNAs that may be involved in antiviral immunity or virus infection by comparing highly infected mosquito pools against lowly infected ones.

Apart from viruses, single-celled eukaryotic parasites are also transmitted from mosquitoes to humans. One of the deadliest of these pathogens is the malaria parasite, *Plasmodium falciparum*. Important regulators that propel the life cycle of this parasite also represent potential drug targets, that when identified could lead to new treatment options to lessen the impact of malaria in Africa and across the globe. We performed a wide variety of experiments and studies in this vein. In the second chapter, we examine the role of long non-coding RNAs (lncRNAs) on parasite gene regulation and predict lncRNAs across the genome, categorizing their properties and their essentiality for parasite survival. We also use experimental techniques to determine the binding sites of several of these predicted lncRNAs, and look at the effect of one in particular, lncRNA-14, by knocking it out and using transcriptomic and phenotypic analysis. The third chapter, on the other hand, uses *Plasmodium berghei*, a mouse malaria parasite, as a model for human malaria and examines parasites proteins SMC2 and SMC4, the homologs of proteins that make up the condensin complex in model eukaryotes. We

determine that, as in other eukaryotes, *Plasmodium* SMC2 and SMC4 form a condensin complex and are key in cell division, particularly in the mosquito stages of the parasite life cycle. ChIP-seq determined that these proteins bind at the centromeres, and transcriptomic and phenotypic analysis revealed the exact roles of these proteins and their importance.

Finally, the remaining two chapters examine RNA-dependent and RNA-binding proteins in *Plasmodium falciparum*, key elements in post-transcriptional regulation, another important aspect of gene regulation in the life cycle. In the fourth chapter, we performed a screen for RNA-dependent proteins using the R-DeeP protocol, obtaining a list of likely proteins and examining RNA-dependent complexes. We also characterized one of the proteins found as an RNA-binding protein using various techniques including enhanced crosslinking and immunoprecipitation followed by high-throughput sequencing (eCLIP-seq). In the fifth chapter, this technique and others were also used to determine the binding sites and functions of two other predicted RNA-binding proteins with RAP (RNA-binding domain abundant in apicomplexans) domains. These RAP proteins were determined to regulate the parasite mitochondrial rRNAs (mitoribosome).

All in all, this dissertation work reveals new insights into mosquito-associated pathogens, their interactions with mosquitoes and mosquito immunity, and important molecular components of their life cycles which could be targeted to reduce their impact on humans.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Introduction

Mosquito species and associated pathogens

Three main genera of mosquitoes dominate transmission of pathogens to humans, each harboring and passing on different types of pathogens. *Culex* mosquitoes transmit West Nile and St. Louis encephalitis viruses, among others. *Aedes* mosquitoes transmit Zika, Yellow Fever, Dengue, and Chikungunya viruses. Finally, *Anopheles* mosquitoes transmit O'nyong nyong virus and, most prominently, human malaria parasites of the *Plasmodium* genus. The most common type of mosquito present, and thus the associated pathogen and prevalence of diseases caused by mosquitoes, varies by region across the globe. In sub-Saharan Africa and parts of southeast Asia, for example, *Anopheles* mosquitoes abound and transmit *Plasmodium* parasites, making malaria common [1], while in the United States, malaria is nearly nonexistent and the predominant mosquito-associated threat is transmission of viruses, particularly West Nile by *Culex* mosquitoes [2]. It is thought that habitable regions for mosquitoes may change, and in many cases expand, as climate changes, making research into different types of mosquitoes and pathogens relevant and crucial [3,4].

Culex mosquitoes and viruses

Culex mosquitoes are among those that transmit viruses that impact public health across the globe. known to transmit West Nile Virus (WNV) and related viruses like St. Louis encephalitis virus (SLEV), which both cause infections in the United States [5,6], as well as

parasitic nematodes and avian malaria parasites. *Culex* is of more immediate interest in California as it much more widespread in the state than *Aedes*, which was only detected there in 2013 [7]. Accordingly, the most common mosquito-transmitted viral disease in California is West Nile. WNV, considered the most prevalent cause of viral encephalitis worldwide, reached New York City in 1999 and spread to the rest of North America within three years, including California in 2002. The virus is mainly transmitted between mosquitoes and birds but can also be incidentally transmitted to humans as well as horses. Around 80% of human cases are asymptomatic, but severe neuroinvasive disease can occur, especially in older patients and those with chronic medical conditions [6]. *Culex* mosquitoes also harbor viruses that are not arboviruses (i.e. are not transmitted to humans) but do establish persistent infections in mosquitoes, and do evoke small RNA immune responses. These include a diverse group of insect-specific viruses (ISVs). Despite not leaving the mosquito, these ISVs are of interest as they may affect, at least partially, the transmission of arboviruses. It is possible that ISVs may decrease vector competence similarly to what is observed with infection by *Wolbachia*, an endosymbiotic bacterium that has been shown to greatly reduce or block virus transmission [8–10]. Recent studies have also suggested that level of ISV infection can modulate transmission of arboviruses [11–13]. However, preliminary results *in vitro* were more compelling than what has been observed *in vivo*, and some of the initial studies have used *A. albopictus* mosquito cell line C6/36, which does not have a functional RNAi system, making the biological relevance questionable. Although research is at an early stage, there seems to be some promise, especially if ISVs related to arboviruses compete with them and reduce

their transmission. Related arboviruses WNV and St. Louis Encephalitis Virus have been shown to compete with each other in this way [14]. It has also been suggested that ISVs could be used as biological control mechanisms or novel vaccine platforms to exploit the host range of ISVs to safely introduce aspects of dangerous viruses to humans [15]. Finally, dual-host viruses such as arboviruses that infect humans are thought to have evolved from ISVs. Some dual-host viruses may have even lost the ability to infect mosquitoes and stayed within the second host. Thus, mosquitoes may played a central role in viral evolution, and may cause novel viral threats to humans and other organisms [16].

Small RNA immunity in mosquitoes and sequencing approach

Several pathways have been implicated in antiviral immunity in insects. Amongst these, the small RNA interference (RNAi) system has been shown to play a central role. The predominant RNAi pathway involved in response to most viruses is the exogenous small interfering RNA (siRNA) pathway. Replicative intermediates in the form of dsRNA are often generated during viral infection, and these intermediates can be processed into 21-nt long siRNAs by Dicer-2, an RNase III enzyme. The siRNAs are loaded onto the RNA-induced silencing complex (RISC), which results in the loss of one of the two strands. The remaining guide strand will be complementary to further invading viral sequences, and will guide the RISC to the virus genome sequences, which will be degraded by Argonaute-2, a component of the RISC with endonuclease activity [17]. The antiviral siRNA pathway functions against both arboviruses and insect-specific viruses.

Another type of small RNA pathway is the PIWI-interacting RNA (piRNA) pathway that has a well-established role in silencing transposons to maintain germline integrity [18]. However, piRNAs have also been implicated in antiviral activity in mosquitoes, although this activity is not yet well-understood [19]. Interestingly, while many small RNA features are shared among dipterans (two-winged flying insects), this expanded piRNA activity appears to be present in mosquitoes but not *Drosophila* fruit flies. In flies, the Piwi-clade contains only three known proteins and is restricted to the ovaries to perform its role in silencing transposons. Mosquitoes possess a significantly expanded repertoire of Piwi-clade proteins as compared to fruit flies, they are expressed in somatic cells as well as follicular cells, and the purified proteins can be associated with viral-derived sequences [20,21]. Expanded piRNA activity in mosquitoes to combat viruses seems very likely. Lower numbers of viral-derived piRNAs leads to cytopathological changes caused by infection, suggesting that the piRNA pathway is indeed involved in antiviral activity, if auxiliary to the siRNA pathway [22].

Yet another class of small RNAs, miRNAs, are 22-nt long and made from processing of long pri-miRNAs transcribed from the eukaryotic genome by RNA polymerase II. Although miRNAs do not seem to be directly involved in antiviral activity in mosquitoes, they are differentially expressed when viral infection is present. miRNAs have a more clearly defined role in regulation of gene expression by binding to cellular mRNAs and preventing their translation or causing them to decay altogether, depending on the exactness of the base pairing.

Antiviral immunity also includes innate immune pathways such JAK-STAT, Toll, and Imd, which have been most extensively researched in *Drosophila melanogaster* fruit flies. Knockouts of genes in these pathways, particularly JAK-STAT, in fruit flies, has led to increased viral titers, and some of these genes were upregulated upon viral infection [23]. However, this type of research has yet to be widely performed in mosquitoes. RNA sequencing has been widely used to detect viruses in many species including mosquitoes. Some studies use direct enrichment of viral RNA, while others use total small RNA sequencing, which captures a great deal of host RNA but is also highly sensitive for viral detection as it can also detect RNAi involved in the immune response. A viral metagenomics study previously done to study the *Culex* virome in California isolated specifically viral RNA, and found a diverse set of viral sequences with differing degrees of relationship to previously known viruses. Some were already known to infect *Culex* in other parts of the world, and some were from viral families that are only known to infect plants or birds and thus most likely did not represent actual mosquito infections, but rather were present because of feeding [24]. Although this approach allows virus detection and novel virus discovery, no information about the mosquito host is retained. Another metagenomics study, in Western Australia, used total small RNA sequencing to determine the virome of the Australian mosquitoes in a sensitive manner. This study found that *Culex*, unlike *Aedes*, tolerates a high abundance and diversity of viruses, and also found viruses that were highly related to those found in *Culex* in China and Indonesia [25]. Our work adds to the evidence from these prior studies, compares and

contrasts the virome found in *Culex* in the Inland Empire region of California to these, and analyzes mosquito antiviral small RNA immunity in detail as well.

Malaria and Plasmodium falciparum

According to estimates in recent years, around 200 to 250 million cases of malaria infection occur annually, spread over at least 90 countries and resulting in more than 600,000 deaths [1]. Ninety percent of malaria cases and deaths occur in Sub-Saharan Africa, but the disease is also widespread in southeast Asia and South America. More than two-thirds of malaria deaths occurred in children under 5.

Multiple apicomplexan parasites of the genus *Plasmodium* can cause malaria in humans. These include *P. falciparum*, *P. vivax*, *P. malariae*, *P. knowlesi*, and *P.ovale*, which in truth likely represents two distinct species, *P. o. curtisi* and *P. o. wallikeri* [26]. However, because *P. falciparum* is responsible for the vast majority of human fatalities (1), it has received the bulk of research attention. This unicellular eukaryotic parasite traverses a complex life cycle. After transmission from mosquito to human in the form of sporozoites, the parasite invades liver cells and develops for a period of around 10 days, before travelling to the bloodstream and undergoing an asexual cycle of invasion and replication inside red blood cells. With each cycle, some parasites commit to becoming gametocytes, which can be taken up by mosquitoes, where the parasites will undergo sexual reproduction and once again form sporozoites that can infect new human hosts. The genome of *P. falciparum* is composed of 23 million base pairs per haploid genome,

arranged into 14 chromosomes [27]. The most striking trait of the parasite genome is its high AT-content of around 80% and rising to 90-95% in intergenic regions, making this the most AT-rich eukaryotic genome yet sequenced. Changes in the parasite's needs and environment occur as the parasite progresses through its life cycle, necessitating large-scale shifts in gene expression [28,29] While several AP2 transcription factors have been identified as potential master regulators of transcription and stage transitions [30–36], the 27 putative *Plasmodium*-specific TFs remain extremely low in number compared to other eukaryotes to regulate the expression of ~ 5500 parasite protein-coding genes [30,33]. As an example, gene expression in the similarly sized yeast genome has been shown to be regulated by 169 specific TFs [37]. As a result, despite the existence of a few other types of identified DNA-binding factors [38–40], researchers remain perplexed as to how such a small number of TFs can govern a complex gene expression program. A good deal of evidence now points toward additional mechanisms such as epigenetics, post-transcriptional regulation, and lncRNAs for controlling gene expression in the parasite.

Transcriptional activity and chromatin structure in P. falciparum

While the most conserved elements of eukaryotic transcriptional machinery are present in *P. falciparum*, parasite-specific features affect how transcription is carried out and regulated. Similar to other eukaryotes, the parasite possesses the RNA polymerase II complex and associated general TFII transcription factors, including the TATA-binding protein (TBP) that is part of the TFIID subunit [33,41]. However, while TFIID in most eukaryotes possesses TBP-associated factors (TAFs) with histone fold domains, the

relatively few TAFs that have been identified in *P. falciparum* do not contain the histone fold domain [41]. As the histone fold domain is involved in heterodimerization of TAFs, the lack of this domain in parasite TAFs suggests a divergent TFIID complex compared to other eukaryotes. The low number of TAFs may also point to alternative mechanisms being more important for transcriptional regulation in parasite.

At the epigenetic level, chromatin and nucleosome organization in *Plasmodium* show reduced stability as compared to chromatin and nucleosome organization in higher eukaryotes, reflecting an increased accessibility of the parasite genome [42]. One factor that may contribute to the lower stability nucleosomes and overall openness of chromatin structure in *Plasmodium* is the apparent absence of linker histone H1 in Apicomplexa and other single-celled eukaryotes[43,44]. Most importantly, with the exception of the telomere ends and a few internal loci, which are marked by the repressive histone mark H3K9me3 and heterochromatin protein 1 (PfHP1) [45–47], most of the chromatin in the nucleus exists as euchromatin with active histone modification marks such as H3K4me3, H3K9ac, and H4K8ac [45,48,49] observed through the genome. Only a few parasite-specific gene families including gene families coding for clonally variant antigens, proteins involved in erythrocyte invasion, and other key proteins such as the gametocyte-promoting transcription factor, PfAP2-G, are known to be maintained in one or more heterochromatin cluster(s) around the periphery of the parasite nucleus [45,48,50–53]. These particular features are addressed in more detail below.

Genome-wide nucleosome and histone trends in P. falciparum

Nucleosome mapping and other techniques have shown that *P. falciparum* retains some features of typical eukaryotic nucleosome landscape. First, the parasite possesses nucleosome-depleted regions (NDRs) in promoter regions [54,55]. As in other eukaryotes, a more pronounced NDR correlates with a higher level of transcription, where more open chromatin structure in the promoter leads to a higher level of gene expression. Second, and still under debate among researchers, genic regions show higher levels of nucleosome occupancy as compared to intergenic regions. These results arise from initial nucleosome mapping [55,56] and FAIRE-seq [57] studies, and are in line with observations in all other eukaryotic genomes [58–60] including *Tetrahymena thermophila* [61], another organism with an AT-rich genome. Some studies dispute this finding and suggest that the more nucleosome-sparse intergenic regions may be caused by preferential digestion of AT-rich regions during nucleosome mapping [54]. Although still controversial, it is possible that with the exception of the telomere ends, the histone variant H2A.Z found ubiquitously throughout intergenic regions of the *Plasmodium* genome generate a weak interaction with the DNA. It has been demonstrated in humans, mice and plants that nucleosomes containing H2A.Z confer lower nucleosome stability compared with other H2A variants [62–64]. In *Plasmodium*, we can speculate that H2A.Z containing nucleosomes could play a chromatin-destabilizing role, which may be important for transcriptional activation in an organism that seems to lack a large amount of specific transcription factors.

While classical eukaryotic features of nucleosome positioning are clearly conserved in *P. falciparum*, some traits of parasite chromatin are known to be divergent from other eukaryotes while others remain controversial in the field. First, evidence has shown that the strongly positioned +1 nucleosome that is found immediately downstream of the transcription start site (TSS) in other eukaryotes is missing in *P. falciparum* [55,65], with strongly positioned nucleosomes instead observed at the beginnings and ends of coding regions [65]. However, it is important to highlight that more recent studies displayed a conserved +1 nucleosome relative to TSS locations that arose from RNA-seq data [54] or modified CAGE (cap analysis of gene expression) [66]. If present, the +1 nucleosome may be more weakly positioned than in other eukaryotes [54]. In fact, a more recent machine learning algorithm incorporating several published epigenetic data sets demonstrated that epigenetic features and nucleosome positioning at the start codons outperformed TSS for predicting transcription in *P. falciparum* [67]. Second, otherwise conflicting nucleosome studies concur that the arrays of nucleosomes within genes display a less phased and more random distribution than is typically observed in eukaryotes [54,55]. It has been proposed that the high AT-content of the parasite genome may cause nucleosomes to bind at preferential locations rather than at fixed distances from other nucleosomes, and that the especially AT-rich intergenic regions cause strong nucleosome positioning at the beginnings and ends of coding regions by acting as barriers [55]. Third, as stated above, histone variants H2A.Z and H2B.Z are found ubiquitously throughout intergenic regions of the core chromosomes [68,69], rather than simply marking active promoters to poise genes for transcription as in other eukaryotes [70–72]. As their levels positively correlate

with AT content [69], it is possible that these variants may have a specialized function for promoting nucleosome deposition in AT-rich regions. H2A.Z and H2B.Z have also been found to be prominently acetylated during the asexual replication cycle [73]. These acetylations may also be important for nucleosome stability and chromatin organization in the AT-rich intergenic regions of the parasite genome.

Another area of debate pertaining to parasite nucleosome landscape is the variation of nucleosome levels during the progression of the parasite life cycle. Evidence suggests that global nucleosome levels drop during the trophozoite stage to allow the transcription of thousands of genes, then rise again as the life cycle progresses through the schizont stage toward egress of merozoites and invasion of new red blood cells [55,57]. In this model, the weak correlation observed between changes in nucleosome positioning vs. mRNA steady state transcript levels for at least 30% of parasite genes [57,74] could be explained by mechanisms regulating gene expression at the post-transcriptional level [75,76]. This hypothesis is supported by the presence of a large number of mRNA-binding proteins identified and validated in the parasite [77,78] as well as mechanisms of gene regulation identified at the translational level [76]. As opposed to global nucleosome depletion tied to large-scale transcriptional activation, other studies propose that changes in nucleosome positioning at regulatory regions generally correlate with the amount of transcription observed at the mRNA steady state level [54]. In addition, recent nascent transcript capture using 4-thiouracil (4-TU) incorporation via pyrimidine salvage [79] as well as ATAC-seq (Assay for Transposase-Accessible Chromatin coupled to next-generation sequencing) [80] experiments during the intra-erythrocytic cycle have displayed

a dynamic change of ATAC-seq signal that correlates with the cascade of stage-specific expression of the associated genes [81,82]. It is highly possible that discrepancies observed between studies could be explained by cell cycle timing and data normalization. Normalization by parasitemia or number of nuclei largely reconciles the differences in these opposing datasets. A recent adaptation of single-cell RNA-sequencing (scRNA-seq) experiments was able to resolve some of these issues [83]. As scRNA-seq produces transcriptomic profiles for multiple individual cells, the authors were able to observe sharp transcriptional transitions over the asexual life cycle, which was previously thought to be a continuous process described as a "cascade of transcripts" [2879]. These results further confirmed discrete transcriptional signatures observed using nascent RNA sequencing technology [74] that correlate with sharp changes in nucleosome positioning and chromatin structure and suggest that gene expression throughout parasite development is not as continuous as commonly thought. In addition, a wide variety of complementary approaches confirm that nucleosome occupancy changes throughout the parasite life cycle. These include Western blots [75], mass spectrometry [55,73,84], MNase-seq, FAIRE-seq [57], and ChIP-seq [55] experiments. Furthermore, ATAC-seq results have also validated that the highest number of promoter peaks representing accessible chromatin regions are found during the trophozoite stage [81]. Collectively, all these datasets propose a model for gene regulation where nucleosome eviction, open chromatin structure and the limited number of validated TFs drive the active transcriptional state observed at the trophozoite stage, followed by an increase in nucleosome levels and reduction in gene expression during the schizont stage. At this later stage, parasite-specific transcription

factors (AP2), as well as histone PTMs are likely regulating transcription at the initiation level in a more classical manner.

P. falciparum epigenetic regulation

Along with the genome-wide nucleosome landscape, other layers of epigenetic regulation such as histone modifications correlate with gene expression in *P. falciparum*. With only a small number of TFs to regulate the predicted 5472 protein-coding genes in the parasite (genome version: 06-18-2015, http://plasmodb.org/plasmo), changes in histone modifications have been demonstrated to play significant roles in controlling gene expression. This is particularly true for virulence genes involved in immune evasion (*var* genes). *Var* genes encode variants of erythrocyte membrane protein 1 (PfEMP1), a protein exported to the surface of the infected erythrocyte. PfEMP1 plays a key role in cytoadherence of the RBC and immune evasion inside the human host [85]. Approximately 60 *var* genes are present in the parasite genome, but only one is expressed at a given time, and switching expression creates antigenic variation and allows the parasite to evade the host immune system [86–88] The mechanisms regulating *var* gene expression have been thoroughly studied *in vitro*, and results from these studies have unveiled complex epigenetic features.

The 59 silenced *var* genes, and other associated silenced genes like PfAP2-G, the gametocyte stage-promoting transcription factor, are marked by the repressive histone modification H3K9me3 and heterochromatin protein 1 (PfHP1) [45,48,50,51]. PfHP1 binds to H3K9me3 and is essential in maintaining repressive heterochromatin. The absence of

PfHP1 results in simultaneous expression of nearly all *var* genes as well as cell-cycle arrest during the asexual cycle and an abnormally high rate of sexual differentiation due to de-repression of PfAP2-G [89]. *Var* genes have also been shown to cluster together in one or more repressive regions at the nuclear periphery of the parasite [46–48,51,57,90]. Other proteins have also been demonstrated as playing key roles in maintaining the heterochromatin cluster(s), including nuclear class II protein PfHDA2, which has been validated as essential for silencing *var* genes and PfAP2-G, the transcription factor critical for sexual differentiation [91]. Other chromatin-modifying enzymes include histone methyltransferases (HKMTs) such as PfSET2, an enzyme which marks nucleosomes with H3K36me3 and was determined as essential for *var* gene repression [92,93]. Disruption of PfSET2 or its interaction with RNA pol II results in expression of nearly the entire *var* gene family. Finally, the sirtuin proteins PfSIR2A or PfSIR2B have been shown to have a role in chromatin condensation and *var* gene silencing through histone deacetylase activity. Initial studies conducted in the 3D7 strain showed that the loss of *PfSir2a* causes de-silencing of multiple *var* genes [94], while the loss of *PfSir2b* causes a modest loss of silencing [95] [94,95]. However, more recent evidence suggests that monoallelic *var* expression in other lab strains is much less sensitive to the loss of these sirtuin proteins [96]. Complementary studies of the initial PfSIR2A knockout strain demonstrated extensive chromosome rearrangements including large deletions in the genome. This latest result created uncertainty about the exact role of these proteins *in vivo*. It is most likely that other component(s) of the genome depleted in the initial knockout experiment are in fact

responsible for the loss of regulation of *var* expression in 3D7. Further experiments will be required to identify the presence of such additional regulatory elements.

In wild-type parasites, the single active *var* gene is transcribed at the late ring and trophozoite stages and is prominently marked by H3K4me3, H3K9ac, and H3K27ac [48,90,94,97]. The *Plasmodium* histone lysine methyltransferase (HKMT) PfSET10 seems to associate with the active *var* gene and is likely responsible for transcription and epigenetic memory by keeping this particular *var* gene poised for activation in daughter parasites [98]. Finally, predominantly euchromatic marker H4K8ac, found to be one of the most sensitive modifications to HDAC inhibitors in *Plasmodium*, also functions to induce expression of the active *var* gene [49]. The many ties between histone-modifying proteins and the expression and repression of *var* genes illustrate that chromatin structure is the most prominent method for the parasite to control mechanisms involved in immune evasion.

Epigenetics of Sexual Differentiation

With each round of asexual replication, a fraction of parasites commits to sexual differentiation into gametocytes, the stage that is transmitted from human to mosquito. Transcription factor PfAP2-G, located on chromosome 12, is well-established as a master regulator of this pathway, positively regulating a set of genes that promote gametocytogenesis [31,32]. In asexual parasites, this gene is silenced by H3K9me3 and PfHP1 and is found to colocalize with repressed clonally variant gene families such as *var* [48,50]. Experimental PfHP1 deletion is sufficient to activate PfAP2-G and increase the

rate of gametocyte production *in vitro* [89]. Recently, GDV1 (gametocyte development 1) protein was identified as an upstream activator of sexual differentiation [99]. GDV1 functions by evicting PfHP1 from H3K9me3 sites, and is itself repressed during the asexual cycle by an antisense multi-exon long noncoding RNA (lncRNA) transcribed from the *gdv1* locus. Although we know that GDV1 and the gametocytogenesis pathway are activated by environmental conditions [99], it is yet unknown how this lncRNA transduces the signal from these conditions and ceases to repress GDV1.

Post-transcriptional regulation in Plasmodium falciparum

Along with transcriptional control by epigenetic regulation of chromatin structure, the tight control of gene regulation in *Plasmodium* could be explained by the many post-transcriptional regulation mechanisms that have been identified in the parasite. Early studies on protein and mRNA transcript abundance showed translational delays for many transcripts, a result which was later confirmed by polysome and ribosome profiling experiments [75,100]. Nascent RNA sequencing showed that there is a burst of transcription that occurs in the trophozoite stage, which also disagrees with the "just-in-time" cascade of gene expression shown by steady-state RNA-seq, suggesting the importance of post-transcriptional regulation mechanisms that are likely regulated by the many RNA-binding proteins (RBPs) identified in the parasite genome [74]. The roles of some of these parasite RBPs have already been validated, including essential protein PfAlba1, which binds to specific mRNA transcripts to repress their translation until later in the asexual life cycle [101]. Similarly to PfAlba1, CAF1, a component of the CCR4-NOT complex and a key

eukaryotic regulator of mRNA decay, has also been shown to regulate transcript abundance for approximately 20% of the parasite genome [102]. Two other genes, DOZI and CITH, have been demonstrated to be essential in the repression of specific transcripts in the female gametocyte stage until they are needed after fertilization in the developing zygote [103,104]. A more recent study has also identified a large amount of additional mRNA-bound proteins in the parasite genome [78] that will need to be further investigated at the functional level. Chapters 4 and 5 contain additional background.

lncRNAs in Plasmodium falciparum

Interaction between RNA and protein can also go beyond simple repression of mRNA translation. Ribonucleoprotein complexes (RNPs), in which RNAs interact with proteins, are dynamic, with frequently changing components including proteins, mRNAs and long non-coding RNAs (lncRNAs) [105]. lncRNAs are defined as RNAs which are over 200 nucleotides in length and are not translated into protein [106]. They seem to be a diverse group of RNAs that perform many functions, and only the functions of some specific lncRNAs have been well-characterized. This is specifically the case for lncRNA Xist in mammals, which is responsible for X-chromosome inactivation in females. In *Plasmodium,* an antisense lncRNA encoded in the intron of genes involved in antigenic variation (termed *var* genes) has been demonstrated to be essential in the expression of these specific genes [107,108]. Other known lncRNAs in *Plasmodium* include the TARE lncRNAs that are specifically found at the telomeres of the parasite chromosomes and may have a role in telomere maintenance, and a lncRNA that regulates the expression of

GDV1, a key gene involved in sexual commitment [99,109]. While a majority of lncRNAs have been found to have a role in transcriptional regulation, others have been demonstrated to be essential in protein-protein interaction [110]. However, the functions of this class of RNAs are still being investigated, and much is not yet understood, even in model organisms. Knowing the molecular complexes in which these lncRNAs interact may help us better understand their function, and in *Plasmodium* may give clues toward the functions of both lncRNAs and unknown proteins, which still make up a large proportion of the *P. falciparum* genome annotation. Chapter 2 contains additional background.

In this dissertation work, I furthered research into mosquito-associated pathogens through multiple avenues. Given the potential benefits of exploring the full virome of disease-transmitting mosquitoes for both surveillance and biological control, I used small RNA sequencing to take a detailed snapshot of viruses present in *Culex* mosquitoes in southern California. Since a full understanding of how the mosquito responds to present viruses will also be useful for research into controlling mosquito transmission, I also performed a detailed analysis of small RNA immunity in *Culex*, expanding the known use of the antiviral piRNA pathway. However, as the most pressing global public health issue related to mosquitoes is malaria caused by transmission of *Plasmodium* parasites, I put particular effort into understanding the detailed mechanisms of gene regulation and molecular biology that allow the parasite to progress through its life cycle and cause disease. In particular, I explored how the parasite regulates its genome beyond the more well-established aspects of epigenetics and chromatin structure explained in this

introduction, expanding the known repertoire of regulatory factors used by the parasite by focusing on lncRNAs as well as RNA-dependent and RNA-binding proteins that facilitate post-transcriptional regulation. These new data include results of genome- and proteome-wide screens that reveal predicted lncRNAs across the genome and RNA-dependent proteins across the proteome, in both cases providing new candidates for exploration into lncRNA function and post-transcriptional regulation, respectively. They also include specific, fully characterized regulators like lncRNA-14, involved in promoting the sexual stage of the parasite life cycle, and RNA-binding proteins like RAP01, RAP21, and PF3D7_0823200, all of which bind to and regulate specific transcripts that have now been identified. Finally, as molecular processes like cell division are also crucial for the parasite life cycle, I also participated in the characterization of the *Plasmodium* condensin complex, composed of SMC proteins which facilitate the atpical cell division of these parasites. The following chapters detail each of these efforts and the findings which represent knowledge that will, in my hope, be useful toward reducing the burden of mosquito-transmitted disease and malaria in particular.

References

1. WHO: World Malaria Report 2023. *World Health Organization* (2023). Available at: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023.

2. Ronca, S. E., Ruff, J. C. & Murray, K. O. A 20-year historical review of west nile virus since its initial emergence in north america: Has west nile virus become a neglected tropical disease? *PLoS Negl. Trop. Dis.* **15**, 1–20 (2021).

3. Reiter, P. Climate Change and Mosquito-Borne Diseases. *Environ. Health Perspect.* **109**, 141–161 (2001).

4. Colón-González, F. J. *et al.* Projecting the risk of mosquito-borne diseases in a warmer and more populated world: a multi-model, multi-scenario intercomparison modelling study. *Lancet Planet. Heal.* **5**, e404–e414 (2021).

5. Diaz, A., Coffey, L. L., Burkett-Cadena, N. & Day, J. F. Reemergence of St. Louis encephalitis virus in the Americas. *Emerg. Infect. Dis.* **24**, 2150–2157 (2018).

6. Chancey, C., Grinev, A., Volkova, E. & Rios, M. The global ecology and epidemiology of west nile virus. *Biomed Res. Int.* **2015**, (2015).

7. Pless, E. *et al.* Multiple introductions of the dengue vector, Aedes aegypti, into California. *PLoS Negl. Trop. Dis.* **11**, 1–17 (2017).

8. Blagrove, M. S. C., Arias-Goeta, C., Failloux, A. B. & Sinkins, S. P. Wolbachia strain wMel induces cytoplasmic incompatibility and blocks dengue transmission in Aedes albopictus. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 255–260 (2012).

9. Ant, T. H., Herd, C. S., Geoghegan, V., Hoffmann, A. A. & Sinkins, S. P. The Wolbachia strain wAu provides highly efficient virus transmission blocking in Aedes aegypti. *PLoS Pathog.* **14**, 1–19 (2018).

10. Dutra, H. L. C. *et al.* Wolbachia Blocks Currently Circulating Zika Virus Isolates in Brazilian Aedes aegypti Mosquitoes. *Cell Host Microbe* **19**, 771–774 (2016).

11. Hobson-Peters, J. *et al.* A New Insect-Specific Flavivirus from Northern Australia Suppresses Replication of West Nile Virus and Murray Valley Encephalitis Virus in Co-infected Mosquito Cells. *PLoS One* **8**, 1–12 (2013).

12. Hall-Mendelin, S. *et al.* The insect-specific Palm Creek virus modulates West Nile virus infection in and transmission by Australian mosquitoes. *Parasites and Vectors* **9**, 1–10 (2016).

13. Goenaga, S. *et al.* Potential for co-infection of a mosquito-specific flavivirus, nhumirim virus, to block west nile virus transmission in mosquitoes. *Viruses* **7**, 5801–5812 (2015).

14. Pesko, K. & Mores, C. N. Effect of sequential exposure on infection and dissemination rates for west nile and St. Louis encephalitis viruses in culex quinquefasciatus. *Vector-Borne Zoonotic Dis.* **9**, 281–286 (2009).

15. Bolling, B. G., Weaver, S. C., Tesh, R. B. & Vasilakis, N. Insect-specific virus discovery: Significance for the arbovirus community. *Viruses* **7**, 4911–4928 (2015).

16. Öhlund, P., Lundén, H. & Blomström, A. L. Insect-specific virus evolution and potential effects on vector competence. *Virus Genes* **55**, 127–137 (2019).

17. Samuel, G. H., Adelman, Z. N. & Myles, K. M. Antiviral Immunity and Virus-Mediated Antagonism in Disease Vector Mosquitoes. *Trends Microbiol.* **26**, 447–461 (2018).

18. Vagin, V. V *et al.* A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. *Science (80-.).* **313**, 320–324 (2006).

19. Léger, P. *et al.* Dicer-2- and Piwi-Mediated RNA Interference in Rift Valley Fever Virus-Infected Mosquito Cells. *J. Virol.* **87**, 1631–1648 (2013).

20. Lewis, S. H., Salmela, H. & Obbard, D. J. Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of Piwi and Aubergine. *Genome Biol. Evol.* **8**, 507–518 (2016).

21. Miesen, P., Girardi, E. & Van Rij, R. P. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. *Nucleic Acids Res.* **43**, 6545–6556 (2015).

22. Morazzani, E. M., Wiley, M. R., Murreddu, M. G., Adelman, Z. N. & Myles, K. M. Production of Virus-Derived Ping-Pong-Dependent piRNA-like Small RNAs in the Mosquito Soma. **8**, (2012).

23. Dostert, C. *et al.* The Jak-STAT signaling pathway is required but not sufficient for the antiviral response of drosophila. *Nat. Immunol.* **6**, 946–953 (2005).

24. Sadeghi, M. *et al.* Virome of > 12 thousand Culex mosquitoes from throughout California. *Virology* **523**, 74–88 (2018).

25. Shi, M. *et al.* High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J. Virol.* **91**, (2017).

26. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542**, 101–103 (2017).

27. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).

28. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol.* **1**, E5 (2003).

29. Roch, K. G. Le *et al.* Discovery of Gene Function by. *Gene Expr.* **301**, 1503–1508 (2003).

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

31. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–52 (2014).

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

33. Coulson, R. M. R., Hall, N. & Ouzounis, C. A. Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res.* **14**, 1548–54 (2004).

34. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. *PLoS Pathog.* **6**, e1001165 (2010).

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

36. Modrzynska, K. *et al.* A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle. *Cell Host Microbe* **21**, 11–22 (2017).

37. Templeton, T. J. *et al.* Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.* **14**, 1686–95 (2004).

38. Gissot, M., Briquet, S., Refour, P., Boschet, C. & Vaquero, C. PfMyb1, a Plasmodium falciparum transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. *J. Mol. Biol.* **346**, 29–42 (2005).

39. Briquet, S. *et al.* High-mobility-group box nuclear factors of Plasmodium falciparum. *Eukaryot. Cell* **5**, 672–82 (2006).

40. Bertschi, N. L. *et al.* Malaria parasites possess a telomere repeat-binding protein that shares ancestry with transcription factor IIIA. *Nat. Microbiol.* **2**, 17033 (2017).

41. Callebaut, I., Prat, K., Meurice, E., Mornon, J.-P. & 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* **6**, 100 (2005).

42. Li, X. Z., Zhang, L. & Poole, K. Role of the multidrug efflux systems of Pseudomonas aeruginosa in organic solvent tolerance. *J. Bacteriol.* **180**, 2987–91 (1998).

43. Sullivan, W. J., Naguleswaran, A. & Angel, S. O. Histones and histone modifications in protozoan parasites. *Cell. Microbiol.* **8**, 1850–61 (2006).

44. Gill, J. *et al.* Structure, localization and histone binding properties of nuclear-associated nucleosome assembly protein from Plasmodium falciparum. *Malar. J.* **9**, 90 (2010).

45. Salcedo-Amaya, A. M. *et al.* Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9655–60 (2009).

46. Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* **407**, 1018–22 (2000).

47. Ay, F. *et al.* Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.* **24**, 974–88 (2014).

48. Lopez-Rubio, J.-J., Mancio-Silva, L. & Scherf, A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* **5**, 179–90 (2009).

49. Gupta, A. P. *et al.* Histone 4 lysine 8 acetylation regulates proliferation and host-pathogen interaction in Plasmodium falciparum. *Epigenetics Chromatin* **10**, 40 (2017).

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

51. Pérez-Toledo, K. *et al.* 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.* **37**, 2596–606 (2009).

52. Crowley, V. M., Rovira-Graells, N., Ribas de Pouplana, L. & Cortés, A. Heterochromatin formation in bistable chromatin domains controls the epigenetic repression of clonally variant Plasmodium falciparum genes linked to erythrocyte invasion. *Mol. Microbiol.* **80**, 391–406 (2011).

53. Freitas-Junior, L. H. *et al.* Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* **121**, 25–36 (2005).

54. Kensche, P. R. *et al.* The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Res.* **44**, 2110–24 (2016).

55. Bunnik, E. M. *et al.* DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics* **15**, 347 (2014).

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

57. Ponts, N. *et al.* Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res.* **20**, 228–38 (2010).

58. Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–5 (2004).

59. Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453**, 358–62 (2008).

60. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lakc of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).

61. Beh, L. Y., Müller, M. M., Muir, T. W., Kaplan, N. & Landweber, L. F. DNA-guided establishment of nucleosome patterns within coding regions of a eukaryotic genome. *Genome Res.* **25**, 1727–38 (2015).

62. Abbott, D. W., Ivanova, V. S., Wang, X., Bonner, W. M. & Ausió, J. Characterization of the stability and folding of H2A.Z chromatin particles: implications for transcriptional activation. *J. Biol. Chem.* **276**, 41945–9 (2001).

63. Osakabe, A. *et al.* Histone H2A variants confer specific properties to nucleosomes and impact on chromatin accessibility. *Nucleic Acids Res.* **46**, 7675–7685 (2018).

64. Henikoff, S. & Smith, M. M. Histone variants and epigenetics. *Cold Spring Harb. Perspect. Biol.* **7**, a019364 (2015).

65. Ponts, N., Harris, E. Y., Lonardi, S. & Le Roch, K. G. Nucleosome occupancy at transcription start sites in the human malaria parasite: a hard-wired evolution of virulence? *Infect. Genet. Evol.* **11**, 716–24 (2011).

66. Adjalley, S. H., Chabbert, C. D., Klaus, B., Pelechano, V. & Steinmetz, L. M. Landscape and Dynamics of Transcription Initiation in the Malaria Parasite Plasmodium falciparum. *Cell Rep.* **14**, 2463–75 (2016).

67. Read, D. F., Lu, Y. Y., Cook, K., Le Roch, K. & Noble, W. S. Predicting gene expression in the human malaria parasite Plasmodium falciparum. *bioRxiv* 1–16 (2018). doi:10.1101/431049

68. Petter, M. *et al.* 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.* **87**, 1167–82 (2013).

69. Hoeijmakers, W. A. M. *et al.* H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the Plasmodium falciparum genome. *Mol. Microbiol.* **87**, 1061–73 (2013).

70. Lohman, T. G. *et al.* Relationships among fitness, body composition, and physical activity. *Med. Sci. Sports Exerc.* **40**, 1163–70 (2008).

71. Tolstorukov, M. Y., Kharchenko, P. V, Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res.* **19**, 967–77 (2009).

72. Guillemette, B. *et al.* Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.* **3**, e384 (2005).

73. Saraf, A. *et al.* Dynamic and Combinatorial Landscape of Histone Modifications during the Intraerythrocytic Developmental Cycle of the Malaria Parasite. *J. Proteome Res.* **15**, 2787–801 (2016).

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

75. Le Roch, K. G. *et al.* Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res.* **14**, 2308–18 (2004).

76. Vembar, S. S., Droll, D. & Scherf, A. Translational regulation in blood stages of the malaria parasite Plasmodium spp.: systems-wide studies pave the way. *Wiley Interdiscip. Rev. RNA* **7**, 772–792 (2016).

77. Reddy, B. P. N. *et al.* A bioinformatic survey of RNA-binding proteins in Plasmodium. *BMC Genomics* **16**, 890 (2015).

78. Bunnik, E. M. *et al.* The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum. *Genome Biol.* **17**, 147 (2016).

79. Painter, H. J. *et al.* Genome-wide real-time in vivo transcriptional dynamics during Plasmodium falciparum blood-stage development. *Nat. Commun.* **9**, 2656 (2018).

80. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–70 (2015).

81. Ruiz, J. L. *et al.* Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum. *Nucleic Acids Res.* **46**, 9414–9431 (2018).

82. Toenhake, C. G. *et al.* Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. *Cell Host Microbe* **23**, 557–569 (2018).

83. Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, 1–29 (2018).

84. Oehring, S. C. *et al.* Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum. *Genome Biol.* **13**, R108 (2012).

85. Miller, L., Good, M. & Milon, G. Malaria Pathogenesis. *Science (80-.).* **264**, 1878–1883 (1994).

86. Biggs, B. A. *et al.* Antigenic variation in Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 9171–4 (1991).

87. Chen, Q. *et al.* Developmental selection of var gene expression in Plasmodium falciparum. *Nature* **394**, 392–5 (1998).

88. Scherf, A., Lopez-Rubio, J. J. & Riviere, L. Antigenic variation in Plasmodium falciparum. *Annu. Rev. Microbiol.* **62**, 445–70 (2008).

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

90. Lopez-Rubio, J. J. *et al.* 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol. Microbiol.* **66**, 1296–305 (2007).

91. Coleman, B. I. *et al.* A Plasmodium falciparum histone deacetylase regulates antigenic variation and gametocyte conversion. *Cell Host Microbe* **16**, 177–186 (2014).

92. Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* **499**, 223–7 (2013).

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

94. Duraisingh, M. T. *et al.* Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. *Cell* **121**, 13–24 (2005).

95. Tonkin, C. J. *et al.* Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum. *PLoS Biol.* **7**, e84 (2009).

96. Merrick, C. J. *et al.* Functional analysis of sirtuin genes in multiple Plasmodium falciparum strains. *PLoS One* **10**, e0118865 (2015).

97. Watanabe, M., Tokita, Y. & Yata, T. Axonal regeneration of cat retinal ganglion cells is promoted by nipradilol, an anti-glaucoma drug. *Neuroscience* **140**, 517–28 (2006).

98. Volz, J. C. *et al.* PfSET10, a Plasmodium falciparum methyltransferase, maintains the active var gene in a poised state during parasite division. *Cell Host Microbe* **11**, 7–18 (2012).

99. Filarsky, M. *et al.* GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* **359**, 1259–1263 (2018).

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

101. Vembar, S. S., Macpherson, C. R., Sismeiro, O., Coppée, J. & Scherf, A. The PfAlba1 RNA-binding protein is an important regulator of translational timing in Plasmodium falciparum blood stages. 1–18 (2015). doi:10.1186/s13059-015-0771-5

102. Balu, B. *et al.* CCR4-associated factor 1 coordinates the expression of Plasmodium falciparum egress and invasion proteins. *Eukaryot. Cell* **10**, 1257–1263 (2011).

103. Mair, G. R. *et al.* Translation Repression is essential for Plasmodium sexual development and mediated by a DDX6-type RNA helicase. *Science (80-.).* **313**, 667–669 (2006).

104. Mair, G. R. *et al.* Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. *PLoS Pathog.* **6**, (2010).

105. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).

106. Fang, Y. & Fullwood, M. J. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics Bioinforma.* **14**, 42–54 (2016).

107. McHugh, C. A., Chen, C.-K., Chow, A., Surka, C. F. & Tran, C. The Xist lncRNA directly interacts with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).

108. Amit-Avraham, I. *et al.* Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E982-91 (2015).

109. Broadbent, K. M. *et al.* A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol.* **12**, R56 (2011).

110. Peng, W., Koirala, P. & Mo, Y. LncRNA-mediated regulation of cell signaling in cancer. 5661–5667 (2017). doi:10.1038/onc.2017.184

Chapter 1: Small RNA sequencing of field *Culex* mosquitoes identifies patterns of viral infection and the mosquito immune response

Steven M. Abel[1†], Zhenchen Hong[2†], Desiree Williams[1], Sally Ireri[1], Michelle Q. Brown[3], Tianyun Su[3], Kim Y. Hung[4], Jennifer A. Henke[4], John P. Barton[2††], and Karine G. Le Roch[1††*]

1. Department of Molecular, Cell and Systems Biology, Center for Infection Disease and Vector Research, University of California, Riverside, CA 92521.

2. Department of Physics and Astronomy, University of California, Riverside, CA 92521.

3. West Valley Mosquito & Vector Control District, Ontario, CA 91761.

4. Coachella Valley Mosquito & Vector Control District, Indio, CA 92201.

*Corresponding author: karine.leroch@ucr.edu

† †† These authors contributed equally to this work.

A version of this chapter has been published in *Scientific Reports*, 2023.

Preface

Transmission of viruses such as West Nile (WNV) or St. Louis encephalitis virus by *Culex* mosquitoes represent a predominant threat of mosquito-borne disease in the United States. Despite substantial interest in curtailing this public health threat, much remains unknown about the virome of field mosquitoes and their impact on the mosquito's immune response or capacity to transmit viruses to humans, animals, and plants. Increasing knowledge in this area may lead to new methods to reduce vector-borne disease transmission around the world. To this end, we sequenced small RNA samples from over 60 pools of *Culex* mosquitoes (~2500 mosquitoes) from two areas of Southern California over a 3-year period. Using a novel genome assembly and computational approach, we not only detected viruses present in field mosquitoes including WNV but their co-infection patterns and the immune responses they induced in *Culex*. Finally, we characterized detected small RNA using their read length as well as nucleotide and strand bias to examine the immune response pathways (either siRNA and/or piRNA) induced against each virus. Using this method, we confirmed a role of the piRNA pathway against some pathogens. I performed nearly all of the computational data analysis for this project and wrote the manuscript which is presented in this chapter with editing from others. This chapter provides the base of this dissertation by addressing host-pathogen interaction and immunity before presenting new insights into parasites carried by mosquitoes, especially *Plasmodium*.

Abstract

Mosquito-borne disease remains a significant burden on global health. In the United States, the major threat posed by mosquitoes is transmission of arboviruses, including West Nile virus by mosquitoes of the *Culex* genus. Virus metagenomic analysis of mosquito small RNA using deep sequencing and advanced bioinformatic tools enables the rapid detection of viruses and other infecting organisms, both pathogenic and non-pathogenic to humans, without any precedent knowledge. In this vein, we sequenced small RNA samples from over 60 pools of *Culex* mosquitoes from two major areas of Southern California from 2017 to 2019 to elucidate the virome and immune responses of *Culex*. Our results demonstrated that small RNAs not only allowed the detection of viruses but also revealed distinct patterns of viral infection based on location, *Culex* species, and time. We also identified miRNAs that are most likely involved in *Culex* immune responses to viruses and *Wolbachia* bacteria, and show the utility of using small RNA to detect antiviral immune pathways including piRNAs against some pathogens. Collectively, these findings show that deep sequencing of small RNA can be used for virus discovery and surveillance. One could also conceive that such work could be accomplished in various locations across the world and over time to better understand patterns of mosquito infection and immune response to many vector-borne diseases in field samples.

Introduction

Transmission of arboviruses to humans by mosquitoes is a persistent public health threat around the world. In the United States, *Culex* mosquitoes transmit arboviruses that are endemic in several states. These notably include West Nile virus (WNV), which caused ~2,500 human cases of disease annually in the U.S. between 1999 and 2019 in addition to many times more asymptomatic infections, and St. Louis encephalitis virus, which also causes a small number of cases annually, including periodic outbreaks [1–3]. WNV, considered the most prevalent cause of viral encephalitis worldwide, reached New York City in 1999 and spread to the rest of North America within four years, including California in 2003. Human infections can sometimes result in severe neuroinvasive disease, especially in older patients and those with chronic medical conditions [3].

Many viruses that have been detected in mosquitoes do not infect humans but do establish persistent infections in the mosquito, and evoke small RNA immune responses [4,5]. Viruses in this diverse group include insect-specific viruses (ISVs) [6] and those that can be transmitted to non-human organisms. Little is known about many of these viruses, or their effect on transmission of arboviruses by mosquitoes. Recent studies have presented evidence that some ISVs may decrease arbovirus loads and transmission [7–9] similarly to what is observed with infection by the *Wolbachia* bacterium. *Wolbachia* is a genus of intracellular bacteria that has been shown, when introduced into non-native host *Ae. aegypti*, to significantly reduce the mosquito's ability to transmit dengue, Zika, and other RNA viruses to humans [10–12]. The potential mechanisms of *Wolbachia*-mediated

antiviral effects are not completely clear, but data suggest some evidence of competition for resources between the virus, host, and *Wolbachia* [13,14] as well as use of host microRNAs by the bacterium to contribute to virus inhibition [15]. If ISVs have some effects on arbovirus infection and transmission [7–9,16,17], they could be used as biological control mechanisms or novel vaccine platforms by exploiting the limited host range of ISVs to protect against dangerous viruses infecting humans [18]. Furthermore, constant monitoring of viruses may allow us to detect the re-emergence of arboviruses transmissible to humans.

In insects including mosquitoes, the small RNA interference (RNAi) system has been shown to play a central role in defense against viruses, most prominently the exogenous small interfering RNA (siRNA) pathway [19,20]. Replicative intermediates in the form of dsRNA are often generated during viral infection, and these intermediates can be processed into ~21-nt long siRNAs by Dicer-2. The siRNAs are loaded onto the RNA-induced silencing complex (RISC) and guide it to complementary, invading viral sequences, which will then be degraded [21,22]. By contrast, the PIWI-interacting RNA (piRNA) pathway has a well-established role in silencing transposons to maintain germline integrity [23]. However, piRNAs, which are generally ~24-29 nt in length, have also been implicated in antiviral activity in mosquitoes, although this activity is not yet well-understood [24–28]. Interestingly, this expanded piRNA activity does not seem to be present in *Drosophila*, despite mosquitoes and fruit flies being in the same order, Diptera. Some mosquitoes, particularly *Culex* and *Aedes* species, possess an expanded repertoire of Piwi-clade proteins as compared to *Drosophila*. Some of the proteins are expressed in

somatic cells as well as follicular cells, and when purified were found to be associated with virus-derived sequences [26,29]. piRNAs can be produced through the primary Zucchini (Zuc) – mediated biogenesis pathway, which generates antisense piRNAs with a 1U bias [30,31] or through the "ping-pong cycle", where primary piRNAs are used to generate sense piRNAs with a 10A bias and further 1U antisense piRNAs [32]. Virus-derived piRNAs have been shown to have the antisense 1U and/or sense 10A nucleotide biases [24–26,33], although the mechanisms of viral piRNA generation remain unknown. Finally, a separate class of small RNAs, miRNAs, are ~22-nt long, are transcribed from the host genome, and have been demonstrated as critical components of gene regulation by binding to cellular mRNAs to control their translation, stability, or decay [34].

RNA sequencing has been used to detect viruses in many species, including mosquitoes [35–38]. In our study, we aimed to use total small RNA extracted from whole mosquitoes to not only sample the virome of mosquitoes but also to analyze patterns of viral infection and immune signatures in these mosquitoes. We therefore deep sequenced 63 pools of *Culex* mosquitoes, 58 of them field-caught, and showed snapshots of the *Culex* virome over a three-year period in southern California. We also examined the patterns and correlation of viral infection based on location, year, and mosquito species. Furthermore, as our goal was not only to discover viruses, but rather to analyze the abundance of and host response to viruses, we also mapped reads to the *Culex* genome to elucidate miRNA responses to both viruses and *Wolbachia*. Finally, we generated size profiles and virus genome coverage plots from the small RNAs mapping to individual viruses to analyze induction of small RNA pathways such as siRNA and piRNA in response to viral

infection in field mosquitoes. Taken together, the results demonstrate the power of our approach, which could be used not only for virus discovery, surveillance, and epidemiology, but also to improve our understanding of mosquito immune response to many vector-borne diseases around the world.

Results

Detection of viruses in *Culex* samples based on *de novo* assembly of small RNA reads

We sequenced and analyzed small RNA from 58 pools of either *Cx. quinquefasciatus* or *Cx. tarsalis* mosquitoes from the Inland Empire region of southern California, as well as five *Cx. quinquefasciatus* pools originating from laboratory strains. The most common pool size was 50 or near to this, but pool sizes varied (see Supplementary Table S1). Our experimental pipeline is summarized in Figure 1.1a, and the computational pipeline for virus detection using VirusDetect [39] is displayed in Figure 1.1b. Agarose gel pictures showing extracted total RNA from mosquito pools are shown in Supplementary Figure S1.

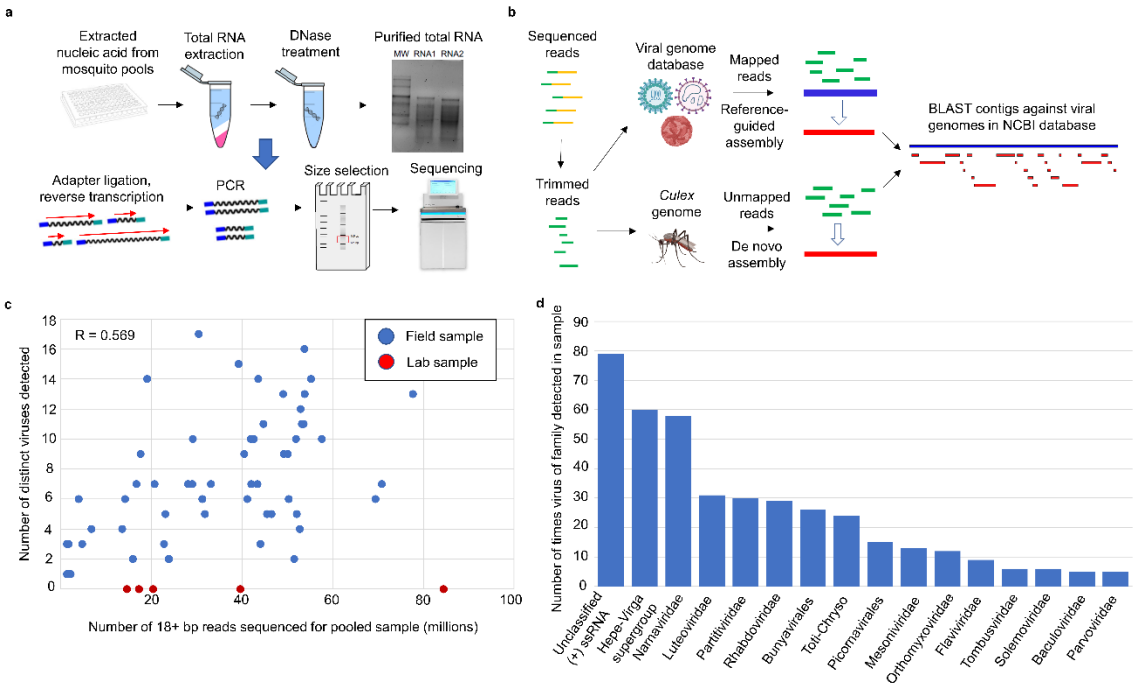

Figure 1.1. Virus detection using small RNA libraries. **(a)** Schematic representation of the experimental protocol. RNA was extracted from mosquito pools, reverse transcribed, PCR amplified, size selected for small RNA, and sequenced. Pool sizes are listed in Supplementary Table S1. **(b)** Schematic representation of virus detection including VirusDetect. Reads were assembled into contigs in two ways and compared to a viral genome database by BLAST. **(c)** Number of known mosquito viruses detected in a sample vs. number of sequenced for that sample (Spearman's R = 0.569 for all samples). **(d)** High nucleotide identity (>90%) virus detections in *Cx. quinquefasciatus* samples by taxonomic group. Hepe-Virga supergroup and Toti-chryso are loose classifications of related virus families.

The pools received an average of 57.9 million reads per sample. The number of sequenced reads directly correlated to the number of distinct viruses that were detected in field samples (Spearman's R coefficient = 0.569). We detected an average of 7 distinct viruses in each sample. Despite high quality sequencing reads for five laboratory samples (average of 72.7 million reads), we did not detect any known mosquito-associated viruses

in these samples (Figure 1.1c), suggesting that lab-grown mosquitoes are not exposed to pathogens as field mosquitoes are. This should be considered when using lab mosquitoes to study viruses. Due to the varying number of sequenced reads between samples, normalization accounting for read number was done whenever samples were compared in downstream analysis.

For most samples, there were several high-identity matches by nucleotide alignment (blastn), which we considered to be high-confidence virus detections as these sequences were closely related to known reference genomes. Many samples also had more distant matches that were detected only through virtual translation of the sequences in the six reading frames (blastx). The list of viruses detected by blastn with 90% nucleotide identity or higher can be found in Supplementary Table S2, while viruses detected only by blastx with 50% amino acid identity or higher are in Supplementary Table S3. The numbers of 90%+ blastn virus detections by virus family/classification in field samples are shown in Figure 1.1d. Virus families detected, separated by mosquito species (*Cx. quinquefasciatus* vs *Cx. tarsalis)* are shown in Supplementary Figure S2. By far the most abundant single virus detected was *Culex* narnavirus 1, which was present in every field sample of both species. Also common were viruses from the Hepe-Virga supergroup [35], a group of (+)ssRNA viruses that has been loosely defined and only recently characterized, reflecting the lack of clear understanding around invertebrate viruses. Supporting this as well is the high prevalence of unclassified (+)ssRNA viruses which could not be placed into any defined families, such as *Bunyavirales*, *Rhabdoviridae*, and *Flaviviridae*, which are well-known to contain ISVs or arboviruses. Interestingly, we were also able to detect

viruses thought to only infect plants (e.g. *Tombusviridae*, *Tymoviridae*, and *Luteoviridae*). For these viruses, as discussed later, we have strong evidence of specific siRNA responses (Supplementary Figure S3, https://github.com/Sabel14/MosquitoSmallRNA_Supplemental_AndCustomScripts), suggesting that they may indeed infect mosquitoes. Many of the viruses detected have widespread geographical range, as some of them were found in other parts of the world including China [35], Mexico [46], and Colombia [47], suggesting many of the same or very similar *Culex* viruses are found throughout the world.

Clustering and patterns of virus-mapped small RNA quantity in mosquito samples

To identify factors affecting viral infection, we used direct mapping of reads to viral genomes (read counts in Supplementary Table S1) and clustered our samples using UMAP [48], a manifold learning technique for dimension reduction (see Methods for details). The resulting numbers of mapped reads represent a combination of viral abundance and intensity of the mosquito immune response, and will be referred to as small RNA quantity. Results show that the most obvious factors determining small RNA quantity in a sample were location and mosquito species (data points for *Cx. tarsalis* cluster apart from those for *Cx. quinquefasciatus*). This was true even for samples collected over multiple years (Figure 1.2a). Year itself as a factor also appears to drive sample clustering but is closely tied to location. As another way to visualize relationships based on small RNA quantity, we generated Pearson correlation matrices [49] between samples and between viruses. The sample correlation matrix (Figure 1.2b) displays which

samples tend to contain the same viruses. Results are similar to those obtained by UMAP. Broadly, blocks of high correlation represented, respectively, Greater LA *Cx. quinquefasciatus* (region a), both locations' *Cx. tarsalis* (region b), Coachella Valley *Cx. quinquefasciatus* (region c), and lab (region d) samples.

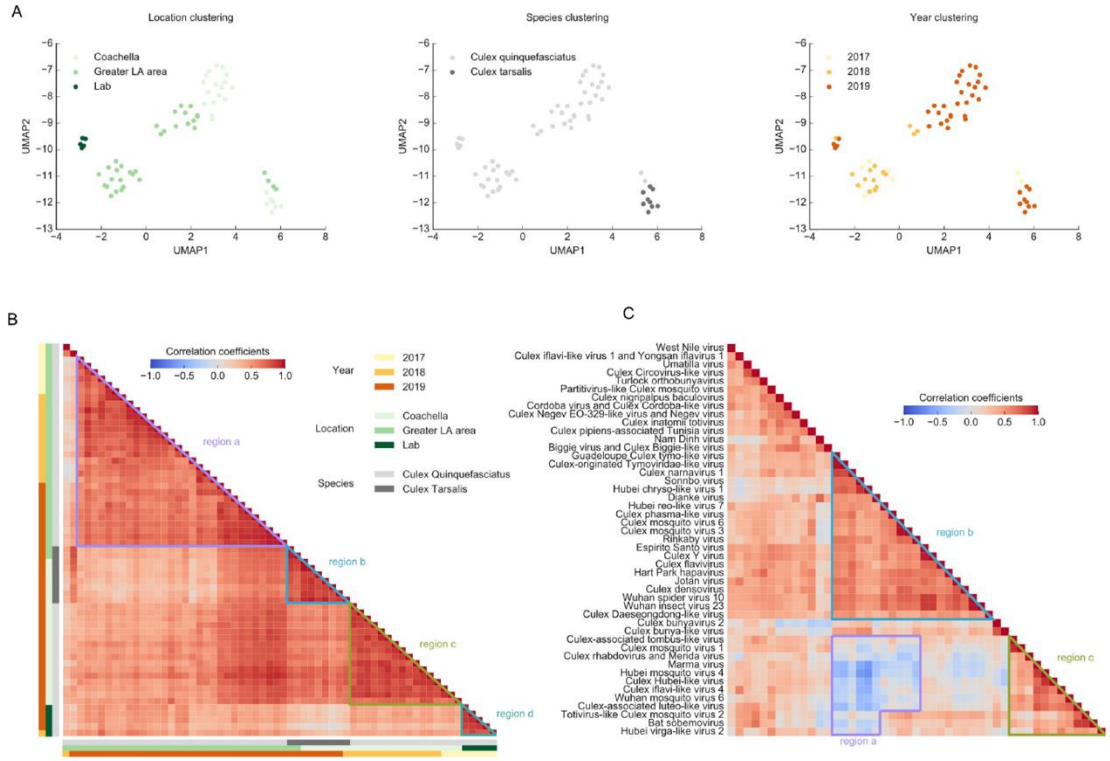

Figure 1.2. Clustering and correlation of mosquito pools by virus small RNA quantities. **(a)** Clustering of mosquito pool samples by virus small RNA quantities using the dimension reduction method UMAP. The three plots differ only by the sample property used to color the data points. **(b)** Pearson correlation matrix of mosquito pool samples by virus small RNA quantities. Sample properties are labeled to the left and below the matrix, and regions of high correlation are denoted. **(c)** Pearson correlation matrix of detected viruses by reads mapped from all samples. Regions of low and high correlation are denoted.

To detect possible virus co-occurrence or suppression in our samples, we generated a virus correlation matrix (Figure 1.2c) using Pearson coefficients for pairs of detected viruses based on their read frequencies across all mosquito pool samples. A positive correlation would mean two viruses tended to infect and generate small RNA in the same samples, while a negative one would mean they are found together in the same sample less often than expected by chance. Negative correlation coefficients were observed between two groups of viruses, as detected in region a. The coefficients in this region range between -0.67 and 0.28, with a median of -0.15. Considering only viruses involved in the most negative correlations, the first group includes Guadeloupe *Culex* tymo-like virus, *Culex*-originated *Tymoviridae*-like virus, Sonnbo virus, Hubei chryso-like virus 1, and Dianke virus, while the second includes Marma virus, Hubei mosquito virus 4, *Culex* Hubei-like virus, *Culex* iflavi-like virus 4, Wuhan mosquito virus 6, and *Culex*-associated luteo-like virus. This suggests that these groups of viruses could exclude each other within the same mosquitoes, potentially providing research direction about virus exclusion, to narrow down the range of possible exclusion candidates. Scatterplots showing frequencies of two viruses for all samples show that the negative correlations are not the result of outlier samples but rather general trends (Supplementary Figure S4). There are also blocks of notably high correlation within two groups of viruses (regions b and c).

Virus correlation matrices including samples collected from the Coachella Valley or *Cx. tarsalis* exhibit patterns that differ from the overall matrix (Supplementary Figure

S5). In *Cx. tarsalis*, WNV is only strongly positively correlated with a select group of other viruses and has a very weak or negative correlation with most. This contrasts with what was observed in *Cx. quinquefasciatus*, where WNV was notably positively correlated with almost all viruses. Thus, it is possible that WNV interacts differently with other viruses depending on the mosquito species, although other possibilities exist such as *Cx. quinquefasciatus* being a more competent vector than *Cx. tarsalis*. These data will need to be further validated as additional samples may provide further insight into viral co-infection patterns.

Small RNA derived from the mosquito genome reveals miRNAs likely to be related to pathogen infection

To investigate the *Culex* response to infection, we explored small RNA reads that mapped to the *Cx. quinquefasciatus* genome (CpipJ2 assembly) - *Cx. tarsalis* samples were not included in this analysis due to the lack of an extensively annotated genome assembly. Approximately 19% of *Culex*-aligned reads mapped to rRNA, tRNA, snRNA, or snoRNA genes, while 32% mapped to pre-miRNA or protein-coding genes, either in coding regions or putative untranslated regions (UTRs) (Figure 1.3a). The remaining 49% mapped to intergenic regions, perhaps representing unannotated transcripts such as novel pre-miRNA genes or lncRNAs. A higher percentage of reads mapped to the antisense of the 3' UTRs as compared to CDSs and 5' UTRs, and to 3' UTRs in general when normalized by total feature length (Supplementary Figure S6), indicating that our

reads are most likely enriched for miRNAs and further validating our methodology as small RNAs, particularly miRNAs, are known to bind to the antisense of the 3' UTRs of targeted genes to regulate transcription at the post-transcriptional level [50].

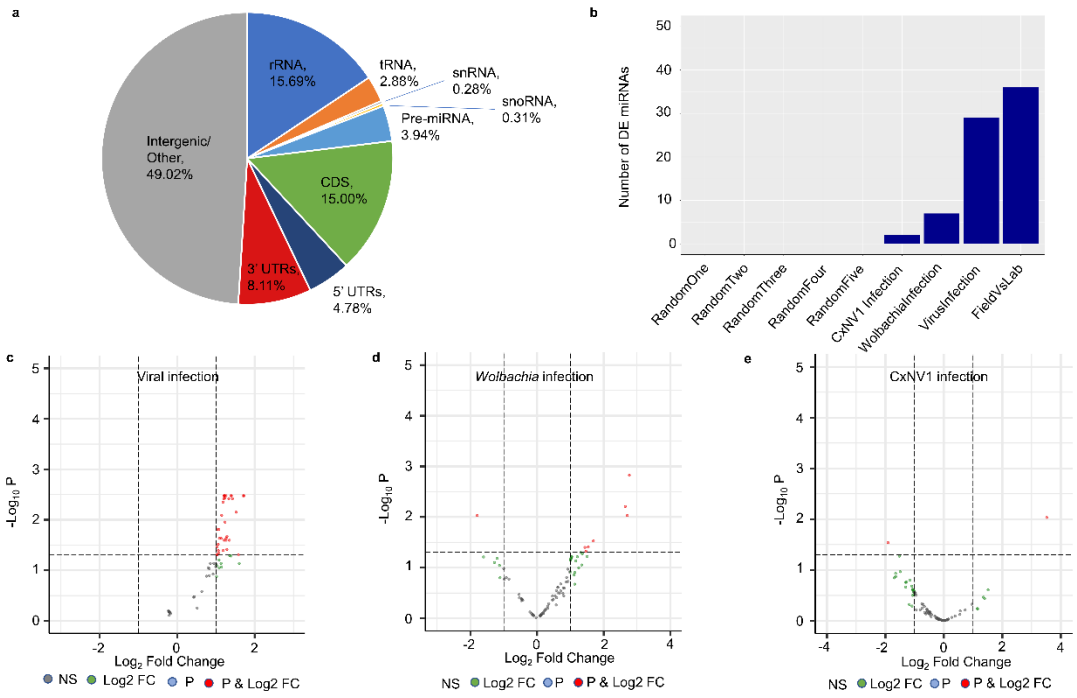

Figure 1.3. Analysis of small RNA derived from the *Culex* genome. **(a)** Percentages of small RNA reads from all *Cx. quinquefasciatus* field samples mapping to each type of genomic feature in the mosquito genome. **(b)** Numbers of miRNA genes determined as differentially expressed for each comparison of field *Cx. quinquefasciatus* samples, with field vs. lab as a point of reference for these comparisons. **(c-e)** Comparison of miRNA expression in samples with higher abundance of **(c)** viruses, **(d)** *Wolbachia*, or **(e)** CxNV1 against those with lower abundance. Volcano significance plots have adjusted P-value cutoff of 0.05 and log2fold change cutoff of 1. NS: not significant. Log2 FC: significant by log2fold change only (threshold +/- 1). P: significant by adjusted P-value only (threshold 0.05). P & Log2 FC: significant by both adjusted P-value and log2fold change.

Next, we performed multiple comparisons in which we segregated all *Cx. quinquefasciatus* samples into two groups based on chosen sample attributes and compared the groups against each other. This was done using DESeq2 [40], a software often used for RNA-seq differential expression analysis, using only sense-mapped reads and restricting the analysis to miRNA genes. DESeq2 normalizes for library size (number of reads) and has been successfully used in various fields for differential expression of small RNAs including miRNAs [51–53]. Our number of samples allowed for a higher number of replicates than typical RNA-seq experiments (48 samples in each field vs. field comparison, 53 samples in a field vs. lab comparison). As a negative control, all 48 *Cx. quinquefasciatus* field samples were randomly assigned into two groups five separate times. Each time, 0 miRNAs were differentially expressed between the two groups (Figure 1.3b). From this, we were confident that any miRNAs that would be called as differentially expressed between selected groups would be due to the chosen factors and not statistical noise.

We first compared samples that were highly infected by viruses against those that were lowly infected, using a threshold of 0.049% of sequenced reads aligning to virus genomes, while controlling for the effects of location and year of collection by including these factors in the DESeq2 generalized linear model (Figure 1.3c). For this analysis, we however excluded *Culex* narnavirus 1, due to its extremely high abundance in all of our samples, and instead analyzed its effect separately (see below). We identified thirty-five pre-miRNA genes that were significantly upregulated in highly infected samples, including twenty-nine unique miRNAs (Figure 1.3c). The full list of upregulated

miRNAs is available in Supplementary Table S4. Interestingly, fourteen of the upregulated miRNAs have already been tied to pathogen infection in previous experiments (see Discussion for details). To assess the putative targets of the top 20 highly expressed of the differentially expressed miRNAs, we used sRNAtoolbox miRNAconsTarget [41], a software that combines four different miRNA target prediction algorithms. The list of putative targets is available in Supplementary Table S5. Gene ontology (GO) enrichment of the targeted genes (Supplementary Table S5), focusing on those agreed upon by at least 2 of the 4 prediction algorithms used, identifies function in translation and cellular respiration as being enriched among the potential identified targets. When we restricted the enrichment to those agreed upon by at least 3 of 4 algorithms, GO enrichment identifies several genes involved in innate immunity, validating further our initial results.

We next examined the effect of *Wolbachia* infection on miRNAs in *Culex* mosquitoes, while controlling for the effects of location and year of collection (Figure 1.3d). High/low infection by *Wolbachia*, as determined by a threshold of 6.34% (the median percentage) of *Culex*-unmapped reads aligned to the *Wolbachia* genome, was associated with a lower number of differentially expressed miRNA genes (8, with 7 of these being unique miRNAs) than infection by viruses (Supplementary Table S6). Two of the seven differentially expressed miRNAs, miR-1889 and miR-12, have been previously associated with *Wolbachia* infection in mosquitoes [54,55] (see Discussion). These results suggest that *Wolbachia* infection induces a more limited but significant miRNA response in the mosquito as compared to viral infection.

As *Culex* narnavirus 1 accounted by itself for 38.6% of virus-mapped reads, we generated a separate analysis between samples with high and low abundance of this virus (Figure 1.3e), determined by a threshold of 0.171% (the median percentage) of sequenced reads aligned to the CxNV1 genome. However, only two miRNAs were detected as differentially expressed, with miR-1889 upregulated and miR-277 downregulated (Supplementary Table S7). Interestingly, miR-1889 was upregulated in both high-*Wolbachia* and high-CxNV1 groups, suggesting a possible general immune function. The effect of CxNV1 infection on miRNAs, while seemingly present to some degree, will need to be explored in future experiments.

Small RNA responses to specific viruses by size profile and genome coverage analysis

Next, we investigated the specific mosquito immune response to individual viruses by examining the size and other properties of reads mapped to each particular virus. The size profiles of the mapped reads, their nucleotide biases, and patterns of sense and antisense genome coverage can be combined to gauge the extent to which siRNA and piRNA response pathways are used in *Culex* against each virus. Because only reads which did not map to the *Culex* genome were used, we can reasonably assume that most observed siRNAs and piRNAs are virus-derived rather than encoded by a viral integration segment in the mosquito genome.

The small RNA size profiles that we detected for each virus are displayed in Figure 1.4a and Supplementary Figure S3. A specific siRNA response was observed for many viruses, with Hubei chryso-like virus 1 being a very clear example in which ~50% of the

total mapped reads were 21 nt in length. For *Culex* bunya-like virus (CbunLV) and *Culex* phasma-like virus (CphasLV), in addition to the 21-nt peak, we detected a clear enrichment for read lengths of 24-29 nt. To validate that the detected read length of 24-29 reads represent a specific piRNA pathway response, we confirmed a sequence bias for an A in the 10th position of the forward reads and for a T in the 1st position of the reverse reads of this size range, indicative of piRNA generation by the ping-pong cycle (Figure 1.4b). We also confirmed that, for CbunLV and CphasLV, there are far more 10-nt overlaps between reads with the ping-pong signature nucleotides than those without, demonstrating additional evidence for this mode of synthesis in the piRNA response pathway (Supplementary Figure S7). Similar profiles were also detected against Turlock orthobunyavirus and Hart Park hapavirus, suggesting the activation of the piRNA ping-pong pathway against these viruses as well (Supplementary Figures S3 and S8). Altogether, clear evidence for the activation of the ping-pong piRNA response pathway was limited to viruses with negative-polarity single-stranded RNA genomes. For *Culex* phasma-like virus segment S, we found that likely piRNAs (24-29 nt, with 1U for antisense reads or 10A for sense reads) target one region directly upstream of the nucleoprotein gene (Figure 1.4c). By contrast, the 21-nt reads that characterize the siRNA response pathway were scattered throughout the virus genome, suggesting that these pathways can specifically target distinct regions in the genome. For this particular segment, a significant number of siRNAs targeted the same site as piRNAs, but this was not the case for all viruses, as mentioned below.

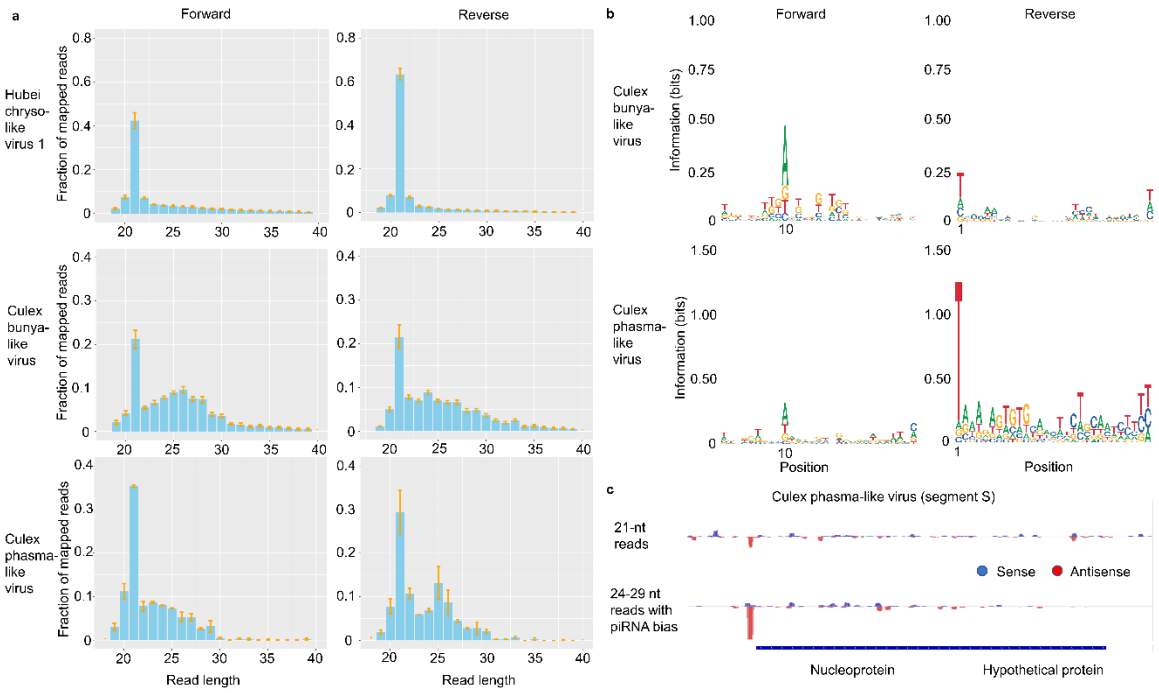

Figure 1.4. Small RNA responses of field *Culex* mosquitoes against specific viruses. **(a)** Examples of small RNA size profiles showing percentages of mapped reads of each size. HCLV1 displays a strong siRNA response (21-nt peaks), while CbunLV and CphasLV display both siRNA and piRNA responses (24-29 nt enrichment). Percent values are averages across all samples in which the virus was detected. Error bars show the average +/- standard error for that small RNA size across all samples. **(b)** Sequence logo plots showing nucleotide bias for 24-29 nt reads mapping to individual viruses. Bias is indicative of piRNA generation by the ping-pong cycle. **(c)** Likely siRNAs and piRNAs mapped to *Culex* phasma-like virus segment S. Only 24-29 nt reads with 1U (for antisense reads) or 10A (for sense reads)are included in the piRNA track.

These results are expanded upon by examining patterns of sense and antisense small RNA coverage for each virus, which support the findings of siRNAs and piRNAs discussed above and also suggest production of piRNAs without ping-pong generation against some viruses. For WNV, a positive-sense RNA virus, antisense 21-nt reads can be found in multiple genomic regions, but there are virtually no antisense 24-29 nt reads

(Figure 1.5a), agreeing with the idea that siRNAs but not piRNAs are generated against WNV as suggested by its size profile and lack of 1U or 10A nucleotide bias (Supplementary Figure S3). This agrees with observations previously made for WNV-infected mosquito cell lines [4]. The coverage plot for a genomic region of *Culex* bunya-like virus confirms the extensive production of siRNAs and piRNAs against it, due to the abundant sense and antisense reads of both size ranges, including many 24-29 nt reads with ping-pong nucleotide bias (Figure 1.5b). Finally, the coverage plot for *Culex* narnavirus 1 reveals new information about this virus (Figure 1.5c), whose small RNA size profile and lack of ping-pong nucleotide bias suggested only siRNA production against it (Supplementary Figure S3). There is a distinct peak of antisense reads upstream of the coding region, a similar pattern observed in *Culex* phasma-like virus (Figure 1.4c), most of which have the piRNA 1U bias. In this case, siRNAs do not target the same site as piRNAs.

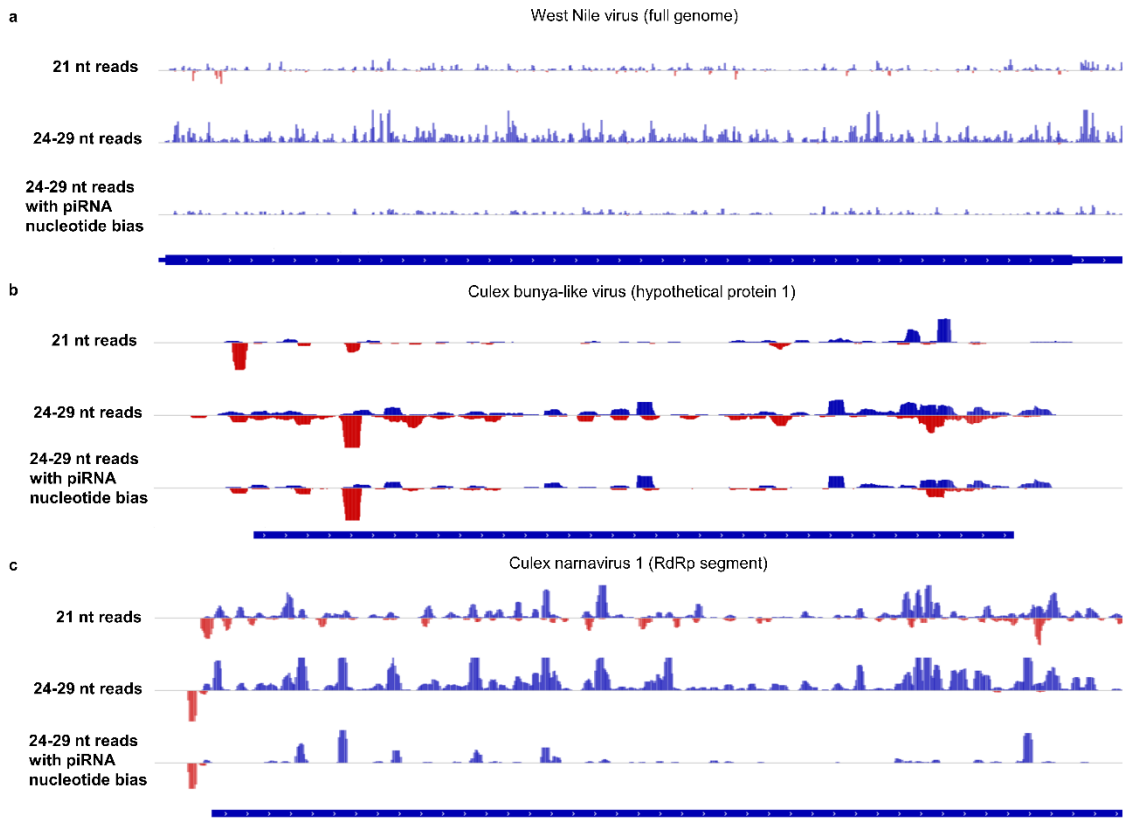

Figure 1.5. Virus coverage plots showing evidence of siRNAs and/or piRNAs in three viruses. **(a)** Sense-mapped reads are shown in the upper half of each track in blue, while antisense-mapped reads are shown in the bottom half in red. Reads shown are from all pools in which the virus was detected by VirusDetect. **(a)** The presence of antisense 21-nt reads but not antisense 24-29 nt reads, as well as lack of nucleotide bias in 24-29 nt reads, suggest siRNA but not piRNA generation against WNV. **(b)** Both sense and antisense reads for both size ranges are generated in abundance against the *Culex* bunya-like virus genomic region shown, and the 24-29 nt reads include a high percentage with the piRNA nucleotide bias. **(c)** Antisense 21-nt reads are abundant against the entire *Culex* narnavirus 1 genome, while antisense 24-29 nt reads are confined to a strong peak upstream of the coding region and many possess the piRNA nucleotide bias. This suggests generation of viral piRNAs without a detectable ping-pong signature.

Coverage plots also suggest possible piRNA production for other viruses without a detectable ping-pong signature or obvious size profile. This can be suggested either by distinct peaks of antisense 24-29 nt reads as for CxNV1 and Hubei mosquito virus 4, or

widespread antisense coverage with 1U bias, as for Wuhan insect virus 23 and Wuhan spider virus 10 (Supplementary Figure S3). Antisense 24-29 nt reads mapped to Wuhan insect virus 23 show a clear preference for 1U (1T in the cDNA sequencing), as do those mapped to Marma virus (Supplementary Figure S8). Although the coverage plot for Marma virus shows that there are few of these antisense 24-29 nt reads against it compared to sense reads (Supplementary Figure S3), the antisense reads may represent a small number of antiviral piRNAs overshadowed by the many reads deriving from the virus genome. Interestingly, several viruses related to families known to only infect plants showed clear 21-nt peaks and antisense 21-nt reads across the genome, suggesting they are generating siRNA responses in the mosquitoes. These viruses include *Culex*-associated luteo-like virus, *Culex*-associated tombus-like virus, *Culex*-originated *Tymoviridae*-like virus, Guadeloupe *Culex* tymo-like virus, and Marma virus (family *Luteoviridae*). Overall, despite clear evidence of piRNA ping-pong generation against four viruses and strong evidence for piRNAs without ping-pong generation for at least four others, the most common feature among the viruses was the detection of the siRNA response pathway. For forty-three out of fifty-four examined viruses (79.6%), 21-nt was the most common mapped small RNA length for sense-mapped reads, antisense-mapped reads, or both, even when accounting for the standard error range for viruses found in multiple samples (Supplementary Figure S3).

Discussion

Mosquitoes are exposed to many pathogens in the field, including many that can be transmitted and are pathogenic to humans, animals, and plants, which they combat in large part by small RNA responses. By using total small RNA sequencing, we detected and characterized patterns of viral infection and improved our understanding of immune response to pathogen infection in the field.

Although our mosquitoes were caught in one general geographic area, the Inland Empire region of southern California, we detected a wide array of viruses in the samples analyzed. While these mosquitoes transmit several human pathogens, many of the detected viruses in this study have yet to be assigned to a family and demonstrate that much of the virosphere in these mosquitoes remains to be fully characterized. Using deep sequencing of viral nucleic acids, Sadeghi et al. explored the virome of over 12 thousand *Culex* mosquitoes in California [36] and detected 56 *Culex*-associated viral strains. While the number of detected viruses between Sadeghi et al and the present studies is different, most likely due to the methodologies and the restricted geographical area use in the studies, both datasets reflect the diversity of viruses present in mosquitoes and a particular abundance of viruses with single-stranded RNA genomes. A previous study done in western Australia demonstrated that *Culex* mosquitoes possess a more diverse range of virus infection than *Aedes* mosquitoes, with 2 to 6 high-abundance viruses found in *Culex* samples and only 0 to 1 in *Aedes* samples [37]. Although we did not test *Aedes* mosquitoes, the high diversity of viruses found in our samples agrees with this claim and

highlights the extended geographic range of our detected viruses, which were also found in Australia, China, and California.

Because gravid females were included in the study, it is also possible that we detected viruses associated with the blood meal rather than the mosquito. However, this is more likely for the viruses detected by blastx or with a weak small RNA signal rather than those with detected strong siRNA signature and other reads mapping to the genome, as discussed below (see Supplementary Table S2 for additional information regarding viruses detected with strong siRNA/piRNA signals). Finally, the lower-identity matches, especially those detected by blastx (Supplementary Table S3), could represent novel viruses or strains that are related to reference genomes present in the databases. The contigs that did not match any sequence could represent novel viruses and will deserve to be further investigated.

Our deep RNA-sequencing strategy not only allowed us to detect virus infection, but also mosquito small RNAs including miRNAs that map in antisense orientation to the 3' UTRs of coding genes and could provide candidate genes that are differentially regulated between highly and lowly infected samples. Several miRNAs that have previously been associated with viral infection in mosquitoes are present in our list of upregulated miRNAs associated with high viral infection. For example, studies on WNV infection in *Culex* mosquitoes have demonstrated that miR-989 and miR-92 both can significantly alter gene expression in WNV-infected mosquitoes, and that miR-989 is downregulated upon infection with WNV while miR-92 is upregulated [56]. In *Aedes*, miR-375 was

described as key to dengue virus replication [57] and miR-252 was shown to target the dengue envelope protein gene to regulate its expression in *Ae. Albopictus* C6/36 cells [58]. Additional experiments demonstrated that introduction of miR-184 and/or miR-275 inhibits dengue virus replication [59], while miR-281 seems to enhance replication [60]. Finally, miR-87 may contribute to the *Aedes* immune response against dengue [61]. Others of these upregulated miRNAs have been associated with non-viral pathogens such as *Wolbachia* in *Aedes* mosquitoes or *Plasmodium* malaria parasites in *Anopheles* mosquitoes. These include bantam, miR-306, miR-305, miR-317, miR-1891, miR-210, and miR-1175 [546263]. The remaining 15 of 29 identified miRNAs will need to be further validated but represent novel candidates for miRNAs with a significant role in controlling viral infection.

To determine specific targets of the differentially expressed miRNAs involved in virus infection or response, we used a combination of four algorithms. While experimental validation will be required to validate some of the potential targets, GO enrichment identified genes involved in translation and innate immunity. These genes are most likely targeted to control infection and stresses induced by the detected virus. As an example, we detected the putative toll protein (CPIJ018343) as a target of miR-989. This gene was predicted by all 4 algorithms and is most likely of particular interest as toll-like receptors are key to innate immunity including antiviral immunity. For mosquito samples that were highly infected with *Wolbachia,* we detected changes in gene expression of seven unique miRNAs. Of these, miR-1889 has been shown to be downregulated in *Wolbachia*-infected *Ae. aegypti* [54], and miR-12 was demonstrated to affect *Wolbachia* density in host

cells by targeting the *MCM6* and *MCT1* genes [55]. miR-309 has not been linked to *Wolbachia* but was shown to be downregulated in *Anopheles stephensi* mosquitoes infected by *Plasmodium* parasites [64]. No miRNAs were differentially expressed due to both viral and *Wolbachia* infection. As *Wolbachia* infection in mosquitoes is currently being used as a biological agent to control the spread of some mosquito-borne disease [65,66], understanding the exact molecular mechanism controlling virus infection in *Wolbachia* infected mosquitoes could help the design of more effective strategies to combat mosquito-borne diseases across the world.

Our designed strategy allowed us to examine small RNA patterns to investigate specific immune responses against viruses using size profiles, nucleotide bias, and/or coverage plots. While this approach has been used previously to investigate the immune response against specific viruses [38,61,67,68], to our knowledge ours is the first study to use a such wide array of viruses in field samples. Our results confirm that the siRNA pathway is the predominant small RNA response used by *Culex* mosquitoes in the field. For some viruses, the 21-nt size profile peak was much more pronounced for antisense reads, while in others, such as Hubei chryso-like virus 1, a strong signal was detected in both sense and antisense (Figure 1.4a). When the 21-nt peak is more pronounced for antisense reads, it is likely that many of the 21-nt sense reads derive from the virus genome rather than siRNA pathways. The fact that we observed clear siRNA responses against viruses that have sequence similarity with plant viruses suggests that these viruses may also replicate in the mosquito. This discovery follows what has recently been shown for narnaviruses in *Culex*. This viral family was previously thought to only infect yeast and oomycetes [69], but

the high coverage of reads and strong siRNA response that we and others [4] detected against *Culex* narnavirus 1 suggest that this virus is replicating in the mosquito.

Interestingly, all four of the viruses with clear evidence of ping-pong piRNA response generation have negative-sense single-stranded RNA genomes, indicating that the (-)ssRNA genome itself may encourage extensive activation of this piRNA pathway in *Culex* mosquitoes. However, virus genome coverage plots suggest piRNAs may be produced against other viruses as well, without use of the ping-pong cycle (Figure 1.5 and Supplementary Figure S3, https://github.com/Sabel14/MosquitoSmallRNA_Supplemental_AndCustomScripts). *Culex* narnavirus 1 represents one example, with an antisense 1U-biased 24-29 nt peak similar to the one for *Culex* phasma-like virus. A recent study done in infected *Aedes albopictus* demonstrated that piRNAs are produced against a specific region of Chikungunya virus while siRNAs target the entire genome [70]. Our data show that Hubei mosquito 4 has a relatively high amount of antisense 24-29 nt reads with 1U bias and 10-nt overlaps with sense reads, in a peak directly outside of the coding region. By contrast, Wuhan insect virus 23 and Wuhan spider virus 10 display multiple regions of antisense 24-29 nt reads which have 1U bias, suggesting a more diffuse pattern of piRNA targeting. This pattern is more similar to those for *Culex* bunya-like virus, Turlock orthobunyavirus, and Hart park hapavirus, which generate widely targeting high-confidence piRNAs with the ping-pong signature. CxNV1, Hubei mosquito virus 4, Wuhan insect virus 23, and Wuhan spider virus 10 all have positive-sense ssRNA genomes, which seem to generate few antisense-mapped reads in general. All together, these data allow us to identify

antisense piRNAs targeting some of these viruses in the absence of a clear ping-pong signature and suggest that different small RNA pathways covering different regions of the genome may be a common pattern across mosquito species against different type of viruses. Although our data suggest that piRNAs may be more common in *Culex* than previously thought, the overall scarcity of evidence for piRNAs in our data does agree with previous observations that piRNA responses occur to a wider array of viruses in *Aedes* compared to *Culex* mosquitoes [4,71]. Overall, the detection of intriguing patterns of viral infection and distinct small RNA immune response demonstrate the need to expand this type of study across different parts of the world, in a wide range of mosquitoes. Such data will allow us to generate an atlas of pathogens and the mosquito immune responses they generate to not only better understand host-pathogen interaction in field samples but to also design novel strategies against many vector-borne diseases.

Materials and Methods

Mosquito collection, pooling, and nucleic acid extraction

For samples collected in both the Ontario and Coachella Valley areas, mosquitoes were amassed using CO_2 traps and gravid traps by the West Valley Mosquito and Vector Control District and Coachella Valley MVCD, respectively. Nucleic acid extraction was performed using the MagMAX Viral RNA Isolation Kit (AMB18365) and samples were deep frozen at -75°C or lower.

RNA extraction and validation

TRIzol was added to nucleic acid extracts for long-term storage, and RNA was extracted from this using chloroform and isopropanol precipitation. Samples were DNase-treated, checked for quality on an agarose gel, purified using Agencourt RNAclean XP beads (Beckman Coulter #A63987), and quantified using a Nanodrop spectrophotometer.

Library preparation and sequencing

Library preparation was performed using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB #E7300S/L), following the provided protocol. Size selection was performed by excising the region corresponding to small RNA on a 6% TBE PAGE gel.

Initial read processing and viral detection

Illumina sequencing results were downloaded in FASTQ form, trimmed of adapter sequence, and, for analysis beyond viral detection, filtered to retain reads of length 18 bp or higher. Trimmed reads were run through VirusDetect, an automated pipeline designed for virus discovery using deep sequencing of small RNAs [39]. We used the default settings for maximum E-value for a hit (1e-5) and minimum percentage identity (25%) for blastn, although our analysis was mostly restricted to matches with at least 90% identity. For blastx hits, we used a cutoff of 50% percentage identity to reduce potentially inaccurate results.

Clustering and prediction based on small RNA quantity

After depleting reads that mapped to the *Cx. quinquefasciatus* (CpipJ2), we mapped reads to a combined file containing all virus genomes that had been detected with high confidence and filtered for uniquely mapped reads. We converted read counts to log-transformed frequencies, used UMAP to generate a lower dimensional visualization for the virus frequency matrix, and generated and inspected Pearson correlation matrices for correlations between samples and between viruses.

Analysis of mosquito-mapped small RNA reads and miRNA analysis

Reads from *Cx. quinquefasciatus* samples were aligned to the *Cx. quinquefasciatus* genome (CpipJ2). DESeq2 [40] was used to find differentially expressed miRNA genes between different groups of samples based on cutoffs of percentages of reads mapping to viruses (0.049% of sequenced reads) or *Wolbachia* (6.34% of Culex-unmapped reads, strain endosymbiont of *Culex quinquefasciatus* Pel strain wPip, NC_010981.1). DESeq2 corrects for differences in number of reads between samples by generation of a size factor for each sample. Putative targets of differentially expressed miRNAs were predicted using sRNAtoolbox miRNAconsTarget [41] with 4 algorithms: Simple seed analysis, TargetSpy [42], Miranda [43], and PITA [44]. GO enrichment was done using Fisher's exact test through VectorBase (https://vectorbase.org).

Analysis of small RNA response to specific viruses

Similarly to clustering analysis, *Culex*-depleted reads were mapped to combined detected virus genomes. Small RNA size profiles and nucleotide bias plots were generated using custom Python and R scripts, and genome-wide coverage plots were made using the Integrative Genomics Viewer (IGV) [45].

References

1. Ronca, S. E., Ruff, J. C. & Murray, K. O. A 20-year historical review of west nile virus since its initial emergence in north america: Has west nile virus become a neglected tropical disease? *PLoS Negl. Trop. Dis.* **15**, 1–20 (2021).

2. Diaz, A., Coffey, L. L., Burkett-Cadena, N. & Day, J. F. Reemergence of St. Louis encephalitis virus in the Americas. *Emerg. Infect. Dis.* **24**, 2150–2157 (2018).

3. Chancey, C., Grinev, A., Volkova, E. & Rios, M. The global ecology and epidemiology of west nile virus. *Biomed Res. Int.* **2015**, (2015).

4. Goertz, G. P. *et al.* Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in Aedes and Culex Mosquito Cells. *Viruses* **11**, 1–18 (2019).

5. Rückert, C. *et al.* Small RNA responses of Culex mosquitoes and cell lines during acute and persistent virus infection. *Insect Biochem. Mol. Biol.* **109**, 13–23 (2019).

6. Agboli, E., Leggewie, M., Altinli, M. & Schnettler, E. Mosquito-specific viruses—transmission and interaction. *Viruses* **11**, 1–26 (2019).

7. Hall-Mendelin, S. *et al.* The insect-specific Palm Creek virus modulates West Nile virus infection in and transmission by Australian mosquitoes. *Parasites and Vectors* **9**, 1–10 (2016).

8. Baidaliuk, A. *et al.* Cell-Fusing Agent Virus Reduces Arbovirus Dissemination in Aedes aegypti Mosquitoes In Vivo. *J. Virol.* **93**, 1–17 (2019).

9. Romo, H., Kenney, J. L., Blitvich, B. J. & Brault, A. C. Restriction of Zika virus infection and transmission in Aedes aegypti mediated by an insect-speci fi c fl avivirus. *Emerg. Microbes Infect.* (2018). doi:10.1038/s41426-018-0180-4

10. Ant, T. H., Herd, C. S., Geoghegan, V., Hoffmann, A. A. & Sinkins, S. P. The Wolbachia strain wAu provides highly efficient virus transmission blocking in Aedes aegypti. *PLoS Pathog.* **14**, 1–19 (2018).

11. Dutra, H. L. C. *et al.* Wolbachia Blocks Currently Circulating Zika Virus Isolates in Brazilian Aedes aegypti Mosquitoes. *Cell Host Microbe* **19**, 771–774 (2016).

12. Chouin-Carneiro, T. *et al.* Wolbachia strain wAlbA blocks Zika virus transmission in Aedes aegypti. *Med. Vet. Entomol.* **34**, 116–119 (2020).

13. Caragata, E. P. *et al.* Dietary Cholesterol Modulates Pathogen Blocking by Wolbachia. *PLoS Pathog.* **9**, (2013).

14. Caragata, E. P., Rancès, E., Neill, S. L. O. & Mcgraw, E. A. Competition for Amino Acids Between Wolbachia and the Mosquito Host , Aedes aegypti. *Microb. Ecol.* **67**, 205–218 (2014).

15. Zhang, G., Hussain, M., Neill, S. L. O. & Asgari, S. Wolbachia uses a host microRNA to regulate transcripts of a methyltransferase , contributing to dengue virus inhibition in Aedes aegypti. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 10276–10281 (2013).

16. Hobson-Peters, J. *et al.* A New Insect-Specific Flavivirus from Northern Australia Suppresses Replication of West Nile Virus and Murray Valley Encephalitis Virus in Co-infected Mosquito Cells. *PLoS One* **8**, 1–12 (2013).

17. Goenaga, S. *et al.* Potential for co-infection of a mosquito-specific flavivirus, nhumirim virus, to block west nile virus transmission in mosquitoes. *Viruses* **7**, 5801–5812 (2015).

18. Bolling, B. G., Weaver, S. C., Tesh, R. B. & Vasilakis, N. Insect-specific virus discovery: Significance for the arbovirus community. *Viruses* **7**, 4911–4928 (2015).

19. Wang, X. *et al.* RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila. *Science (80-.).* **312**, 452–454 (2006).

20. Galiana-arnoux, D., Dostert, C., Schneemann, A., Hoffmann, J. A. & Imler, J. Essential function in vivo for Dicer-2 in host defense against RNA viruses in drosophila. *Nat. Immunol.* **7**, 590–597 (2006).

21. Samuel, G. H., Adelman, Z. N. & Myles, K. M. Antiviral Immunity and Virus-Mediated Antagonism in Disease Vector Mosquitoes. *Trends Microbiol.* **26**, 447–461 (2018).

22. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs : an expanding universe. *Nat. Rev. Genet.* **10**, 94–108 (2009).

23. Vagin, V. V *et al.* A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. *Science (80-.).* **313**, 320–324 (2006).

24. Vodovar, N. *et al.* Arbovirus-Derived piRNAs Exhibit a Ping-Pong Signature in Mosquito Cells. *PLoS One* **7**, (2012).

25. Schnettler, E. *et al.* Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells. *J. Gen. Virol.* **94**, 1680–1689 (2013).

26. Miesen, P., Girardi, E. & Van Rij, R. P. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. *Nucleic Acids Res.* **43**, 6545–6556 (2015).

27. Hess, A. M. *et al.* Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. *BMC Microbiol.* **11**, 24–30 (2011).

28. Petit, M. *et al.* PiRNA pathway is not required for antiviral defense in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4218–E4227 (2016).

29. Lewis, S. H., Salmela, H. & Obbard, D. J. Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of Piwi and Aubergine. *Genome Biol. Evol.* **8**, 507–518 (2016).

30. Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L. & Hannon, G. J. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature* **491**, 279–283 (2012).

31. Mohn, F., Handler, D. & Brennecke, J. piRNA-guided slicing specifies transcripts for Zucchini dependent, phased piRNA biogenesis. *Science (80-.).* **348**, 812–817 (2015).

32. Brennecke, J. *et al.* Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* **128**, 1089–1103 (2007).

33. Léger, P. *et al.* Dicer-2- and Piwi-Mediated RNA Interference in Rift Valley Fever Virus-Infected Mosquito Cells. *J. Virol.* **87**, 1631–1648 (2013).

34. Bartel, D. P. MicroRNAs : Genomics , Biogenesis , Mechanism , and Function. *Cell* **116**, 281–297 (2004).

35. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).

36. Sadeghi, M. *et al.* Virome of > 12 thousand Culex mosquitoes from throughout California. *Virology* **523**, 74–88 (2018).

37. Shi, M. *et al.* High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J. Virol.* **91**, (2017).

38. Wu, Q. *et al.* Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1606–1611 (2010).

39. Zheng, Y. *et al.* VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **500**, 130–138 (2017).

40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

41. Aparicio-Puerta, E. *et al.* sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms. *Nucleic Acids Res.* **50**, W710–W717 (2022).

42. Sturm, M., Hackenberg, M., Langenberger, D. & Frishman, D. TargetSpy: A supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics* **11**, (2010).

43. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, (2010).

44. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).

45. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

46. Charles, J. *et al.* Merida virus , a putative novel rhabdovirus discovered in Culex and Ochlerotatus spp . mosquitoes in the Yucatan Peninsula of Mexico. *J. Gen. Virol.* **97**, 977–987 (2016).

47. Nunes, M. R. T. *et al.* Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus. *Virology* **504**, 152–167 (2017).

48. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

49. Pearson, K. Notes on regression and inheritance in the case of two parents. in *Proceedings of the Royal Society of London* 240–242 (1895).

50. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).

51. Balaskas, P. *et al.* Small non-coding RNAome of ageing chondrocytes. *Int. J. Mol. Sci.* **21**, (2020).

52. Xue, X. *et al.* Dietary Immunostimulant CpG Modulates MicroRNA Biomarkers Associated with Immune Responses in Atlantic Salmon (Salmo salar). *Cells* **8**, 1–22 (2019).

53. Ramachandran, S. R., Mueth, N. A., Zheng, P. & Hulbert, S. H. Analysis of miRNAs in Two Wheat Cultivars Infected With Puccinia striiformis f. sp. tritici. *Front. Plant Sci.* **10**, 1–11 (2020).

54. Mayoral, J. G., Etebari, K., Hussain, M., Khromykh, A. A. & Asgari, S. Wolbachia infection modifies the profile, shuttling and structure of microRNAs in a mosquito cell line. *PLoS One* **9**, (2014).

55. Osei-amo, S., Hussain, M., Neill, S. L. O. & Asgari, S. Wolbachia-Induced aae-miR-12 miRNA Negatively Regulates the Expression of MCT1 and MCM6 Genes in Wolbachia-Infected Mosquito Cell Line. *PLoS One* **7**, (2012).

56. Skalsky, R. L., Vanlandingham, D. L., Scholle, F., Higgs, S. & Cullen, B. R. Identification of microRNAs expressed in two mosquito vectors, Aedes albopictus and Culex quinquefasciatus. *BMC Genomics* **11**, (2010).

57. Hussain, M., Walker, T., O'Neill, S. L. & Asgari, S. Blood meal induced microRNA regulates development and immune associated genes in the Dengue mosquito vector, Aedes aegypti. *Insect Biochem. Mol. Biol.* **43**, 146–152 (2013).

58. Yan, H. *et al.* miR-252 of the Asian Tiger Mosquito Aedes albopictus Regulates Dengue Virus Replication by Suppressing the Expression of the Dengue Virus Envelope Protein. *J. Med. Virol.* **86**, 1428–1436 (2014).

59. Tsetsarkin, K. A. *et al.* Dual miRNA Targeting Restricts Host Range and Attenuates Neurovirulence of Flaviviruses. *PLoS Pathog.* **11**, 1–22 (2015).

60. Zhou, Y. *et al.* MIR-281, an abundant midgut-specific miRNA of the vector mosquito Aedes albopictus enhances dengue virus replication. *Parasites and Vectors* **7**, 1–11 (2014).

61. Aguiar, E. R. G. R. *et al.* Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic Acids Res.* **43**, 6191–6206 (2015).

62. Dennison, N. J., BenMarzouk-Hidalgo, O. J. & Dimopoulos, G. MicroRNA-regulation of Anopheles gambiae immunity to Plasmodium falciparum infection and midgut microbiota. *Dev Comp Immunol.* **49**, 170–178 (2015).

63. Biryukova, I., Ye, T. & Levashina, E. Transcriptome-wide analysis of microRNA expression in the malaria mosquito Anopheles gambiae. *BMC Genomics* **15**, 1–19 (2014).

64. Jain, S. *et al.* Blood feeding and Plasmodium infection alters the miRNome of Anopheles stephensi. *PLoS One* **9**, (2014).

65. O'Neill, S. L. *et al.* Establishment of wMel Wolbachia in Aedes aegypti mosquitoes and reduction of local dengue transmission in Cairns and surrounding locations in northern Queensland, Australia. *Gates Open Res.* **3**, 1–32 (2019).

66. Ahmad, N. A. *et al.* Wolbachia strain wAlbB maintains high density and dengue inhibition following introduction into a field population of Aedes aegypti. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **376**, 20190809 (2021).

67. Webster, C. L. *et al.* The discovery, distribution, and evolution of viruses associated with drosophila melanogaster. *PLoS Biol.* **13**, 1–33 (2015).

68. Belda, E. *et al.* De novo profiling of RNA viruses in Anopheles malaria vector mosquitoes from forest ecological zones in Senegal and Cambodia. *BMC Genomics* **20**, (2019).

69. Hillman, B. I. & Cai, G. The family Narnaviridae: Simplest of RNA viruses. in *Advances in Virus Research* 149–176 (2013).

70. Marconcini, M. *et al.* Profile of small RNAs, vDNA forms and viral integrations in late Chikungunya virus infection of Aedes albopictus mosquitoes. *Viruses* **13**, (2021).

71. Miesen, P., Joosten, J. & van Rij, R. P. PIWIs Go Viral : Arbovirus-Derived piRNAs in Vector Mosquitoes. *PLoS Pathog.* **12**, (2016).

Supplementary Material

Supplementary material for chapter 1 is available with the published paper at

https://www.nature.com/articles/s41598-023-37571-6.

Chapter 2: Novel insights into the role of long non-coding RNA in the human malaria parasite, *Plasmodium falciparum*

Gayani Batugedara[1]*, Xueqing M. Lu[1]*, Borislav Hristov[2], Steven Abel[1], Zeinab Chahine[1], Thomas Hollin[1], Desiree Williams[1], Tina Wang[1], Anthony Cort[1], Todd Lenz[1], Trevor A. Thompson[1], Jacques Prudhomme[1], Abhai K. Tripathi[3], Guoyue Xu[3], Juliana Cudini[4], Sunil Dogga[4], Mara Lawniczak[4], William Stafford Noble[2], Photini Sinnis[3] and Karine G. Le Roch[1✧]

[1]Department of Molecular Cell and Systems Biology, University of California Riverside, Riverside, CA 92521, USA
[2] Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, USA
[3]Department of Molecular Microbiology and Immunology and the Johns Hopkins Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[4]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK

* These authors contributed equally to this work.

A version of this chapter has been published in *Nature Communications*, 2023.

Preface

The deadliest pathogen carried by mosquitoes is the human malaria parasite, *Plasmodium falciparum*. Much research into this parasite aims to discover mechanisms and actors involved in gene regulation, which could be targeted by antimalarial drugs. While most targets identified are proteins, recent evidence suggests that lncRNAs may also be heavily involved in parasite gene regulation. We used a computational approach to predict lncRNAs across the parasite genome, then experimental approaches to validate several of these and characterize one in particular, lncRNA-14. This genome-wide approach of predicting lncRNAs identified many with various properties, including which stage of the parasite life cycle in which they are expressed, whether they are nuclear or cytoplasmic, their size and GC content, and whether they are found in telomeric regions or not. I was involved in various computational analyses for this project. I compared the predicted lncRNA coordinates to previously found piggyBac insertion data from a study which had determined the essentiality of *Plasmodium falciparum* genes. By using the lncRNA coordinates instead of protein-coding genes, I determined which lncRNAs are likely to be essential, meaning the parasite would die if the lncRNA was disrupted. These were more likely to be large lncRNAs in the telomeric (and subtelomeric) regions. I also helped to analyze ChIRP-seq data showing binding locations of several candidate lncRNAs in the genome, and analyzed RNA-seq data showing the transcriptomic effect of knocking out lncRNA-14. This knockout affected parasite at the schizont stage most strongly, interfering with their ability to invade new red blood cells when they normally would at this stage, and also with the ability to

generate gametocytes, the transmission stage in which parasites would be taken up by mosquitoes. This work is representative as one of several I was involved in, analyzing next-generation sequencing data to better understand chromatin-associated gene regulation in *Plasmodium*.

Abstract

The complex life cycle of *Plasmodium falciparum* requires coordinated gene expression regulation to allow host cell invasion, transmission, and immune evasion. Increasing evidence now suggests a major role for epigenetic mechanisms in gene expression in the parasite. In eukaryotes, many lncRNAs have been identified to be pivotal regulators of genome structure and gene expression. To investigate the regulatory roles of lncRNAs in *P. falciparum* we explore the intergenic lncRNA distribution in nuclear and cytoplasmic subcellular locations. Using nascent RNA expression profiles, we identify a total of 1,768 lncRNAs, of which 718 (~41%) are novels in *P. falciparum*. The subcellular localization and stage-specific expression of several putative lncRNAs are validated using RNA-FISH. Additionally, the genome-wide occupancy of several candidate nuclear lncRNAs is explored using ChIRP. The results reveal that lncRNA occupancy sites are focal and sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis and sexual differentiation. Genomic and phenotypic analysis of one specific lncRNA demonstrate its importance in sexual differentiation and reproduction. Our findings bring a new level of insight into the role of lncRNAs in pathogenicity, gene regulation and sexual differentiation, opening new avenues for targeted therapeutic strategies against the deadly malaria parasite.

Introduction

Malaria, a mosquito-borne infectious disease, is caused by protozoan parasites of the genus *Plasmodium*. Among the human-infecting species, *Plasmodium falciparum* is the most prevalent and deadly, with an estimated 627 000 deaths in 2020 [1]. The parasite has a complex life cycle involving multiple biological stages in both human and mosquito hosts. As sporozoites are transmitted from a *Plasmodium*-infected mosquito to the human bloodstream, they migrate to the liver to invade hepatocytes and initiate parasite amplification. After this pre-erythrocytic cycle, which can last 7 to 10 days, tens of thousands of infectious merozoites are released into the bloodstream to invade red blood cells. Within the erythrocyte, the parasite matures from the ring to the trophozoite and to the multinucleated schizont stages. After 48 hours, the newly formed merozoites burst out of the erythrocyte to reinfect new red blood cells. This rupture is usually associated with clinical symptoms. During this intraerythrocytic developmental cycle, a subset of the parasites can differentiate into male and female gametocytes. Once ingested by a female Anopheles mosquito during a blood meal, these gametocytes undergo sexual replication inside the mosquito's gut to form a zygote that can differentiate into a mobile ookinete and an oocyst. The oocyst will grow and produce thousands of new sporozoites that will migrate to the mosquito's salivary glands ready to infect a new human host during a subsequent blood meal. This multi-stage developmental life cycle leads to distinct morphological and physiological changes in response to altered environmental conditions and is tightly regulated by coordinated changes in gene expression.

Gene expression profiling [2,3] including bulk RNA-seq experiments [2,4-7], nascent RNA expression profiles [8], as well as single cell sequencing [9] has revealed that a majority of the genes in the parasite are transcribed in a cascade of gene expression throughout the parasite life cycle but the exact molecular mechanisms regulating these events are largely unknown.

Compared to other eukaryotes with a similar genome size, *P. falciparum* has an extremely AT-rich genome and a relatively low number of sequence-specific transcription factors (TFs), approximately two-thirds of the TFs expected based on the size of the genome. Only 27 apicomplexan apetala2 (ApiAP2) DNA-binding proteins have been identified as specific TFs in the parasite genome. These ApiAP2 are unique to *Apicomplexan* [10] and have been demonstrated to have a major role as activators or repressors of transcription [11]. Our understanding of the regulation of these TFs, and how various TFs could act together to organize transcriptional networks, is still limited but the patterns of gene expression observed are likely the result of a combination of transcriptional [9,12-14] and post-transcriptional regulatory events [15-18]. Additionally, epigenetic studies [19-25] and chromosome conformation capture methods (Hi-C) [26-28] have suggested that the chromatin state and the three-dimensional (3D) genome structure of *P. falciparum* are strongly connected with transcriptional activity of gene families [28]. Machine learning algorithms have also suggested that the ApiAP2 TFs may indeed work in conjunction with epigenetic factors [29]. However, how all the regulators of transcription are recruited to their DNA binding motifs and their chromatin regions remains to be

elucidated. Understanding the exact mechanisms regulating the parasite replication life cycle is essential if we want to identify novel therapeutic targets.

With advances in biotechnology and next generation sequencing technologies, huge strides have been made in genomics studies revealing that the transcriptome of an organism is much larger than expected. In eukaryotes spanning from yeast to human, many non-coding RNAs (ncRNAs) have been detected and linked to diseases ranging from cancers to neurological disorders and are now actively studied for their potential as novel therapeutic and diagnostic agents [30]. Over the past few years, ncRNAs been recognized as key regulators of chromatin states and gene expression [31-33]. One class of ncRNAs, the long noncoding RNAs (lncRNAs), are defined as non-protein coding RNA molecules which are ≥ 200 nucleotides in length. Many lncRNAs share features with mature mRNAs including 5' caps, polyadenylated tails as well as introns [34]. LncRNAs are expressed and functionally associated in a cell-type specific manner. Based on their genomic localization, lncRNAs are categorized as sense, antisense, bidirectional, intronic and intergenic [35]. because lncRNAs can bind DNAs, RNAs and proteins, their functions are diverse [34]. LncRNAs enriched in the nuclear fraction often associate with regulation of transcription [36-39]. By tethering genomic DNA, lncRNAs can control long-range interaction. They can also regulate promoter accessibility by recruiting, guiding or enhancing either TFs or chromatin remodeling enzymes including histone acetyltransferases and methyltransferases. LncRNA have also been shown to interact with spliceosomal factors to affect the frequency and efficiency of mRNA splicing. In the

cytosol, lncRNAs can regulate gene expression by mediating mRNA export, RNA stability and translation.

In mammalian systems, the X inactive specific transcript (Xist) is a well-studied example of a lncRNA mediating X-chromosome inactivation during zygotic development [40]. Deposition of Xist on the X-chromosome recruits histone-modifying enzymes that place repressive histone marks, such as H3K9 and H3K27 methylation, leading to gene silencing and the formation of heterochromatin. Two other lncRNAs, the HOX transcript antisense (HOTAIR) and the antisense lncRNA in the INK4 locus (ANRIL), have also been shown to interact with multiprotein Polycomb PRotein Complexes (PRC1 and PRC2) to catalyze histone marks and silence gene expression [30]. Similarly, long telomeric repeat-containing lncRNAs (TERRA) have been recently identified as a major component of telomeric heterochromatin [41,42].With thousands of lncRNAs transcribed in mammalian cells we are only starting to grasp their role in regulating major biological processes.

Although the role of lncRNA in malaria parasite has only been studied more recently, they are emerging as new players in the development of parasite life cycle stages. To date, several studies have already explored the presence of ncRNAs in *P. falciparum* [43-55]. Technological advances including strand-specific and long read sequencing platforms have identified > 2,500 lncRNA candidates, including 1,300 circular lncRNAs [51-53,56,57] . These initial studies confirmed that parasite lncRNAs are developmentally regulated but

only a few of these annotated ncRNAs have been functionally characterized. Some have been linked to regulation of virulence genes [58-63]**.** It has also been established that GC-rich ncRNAs serve as epigenetic regulatory elements that play a role in activating *var* gene transcription as well as several other clonally variant gene families [64]. In addition, a family of twenty-two lncRNAs transcribed from the telomere-associated repetitive elements (TAREs) has been identified in the parasite [45,47,60]. These TARE-lncRNAs show functional similarities to the eukaryotic family of non-coding RNAs involved in telomere and heterochromatin maintenance [65] and could have a role in regulating virulence factors. More recently, the functional characterization of two lncRNAs, *gdv1*-as-lncRNA and *md1*-lncRNA that were detected during gametocytogenesis, has revealed that sexual differentiation and sex determination in *P. falciparum* is at least partially regulated by lncRNAs[66,67]. While it is becoming evident that lncRNAs serve as an integral part of the mechanisms regulating gene expression in *Plasmodium*, the localization and function of most of the identified lncRNAs remain a mystery.

Here, to investigate the localization and subsequently the potential role of lncRNAs in *P. falciparum,* we explore the intergenic lncRNA distribution separately in nuclear and cytoplasmic subcellular locations. Using deep sequencing and nascent RNA expression profiles [8], we identify a total of 1,768 lncRNAs, of which 41% are novels in *P. falciparum*. We further validate the subcellular localization and stage-specific expression of several putative lncRNAs using RNA fluorescence in situ hybridization (RNA-FISH) and single-cell RNA sequencing (scRNA-seq). Additionally, the genome-wide occupancy

of 7 candidate nuclear lncRNAs is explored using Chromatin Isolation by RNA Purification followed by deep sequencing (ChIRP-seq). Our ChIRP-seq experiments on our candidate lncRNAs reveal that lncRNA occupancy sites within the parasite genome are sequence-specific with a particular enrichment for several parasite-specific gene families, including those involved in pathogenesis, remodeling of the RBC, and regulation of sexual differentiation. We also demonstrate that the presence of some of these lncRNAs correlates with changes in gene expression demonstrating that these lncRNAs can possibly work in cooperation with TFs and epigenetic factors. We further validate the role of a lncRNA identified as enriched in gametocytes. Using the CRISPR-cas9 editing tool, we functionally characterize lncRNA-ch14 and validate its role during sexual differentiation and development, particularly affecting female gametocytes. Transmission studies demonstrate that even partial deletion of this lncRNA significantly affects parasite development throughout all mosquito stages. Collectively, our results provide evidence that in addition to being developmentally regulated, lncRNAs are distributed in distinct cellular compartments in *P. falciparum.* Depending on their nuclear or cytoplasmic localization, they may play important roles in gene regulation at the transcriptional or translational levels respectively, ultimately regulating the malaria parasite life cycle progression.

Results

Identification of lncRNAs.

To comprehensively identify lncRNA populations in *P. falciparum* we extracted total RNA from both nuclear and cytoplasmic fractions using synchronized parasite cultures at early ring, early trophozoite, late schizont, and gametocyte stages (Fig. 2.1a). The samples collected here allow for gene expression profiling during the critical processes of parasite egress, invasion, and sexual differentiation. In brief, extracted parasites were subjected to a modified cell fractionation procedure described in the PARIS kit (ThermoFisher) (see methods). Successful isolation of both subcellular fractions was validated using western blot with an anti-histone H3 antibody as a nuclear marker and an anti-aldolase antibody as a cytoplasmic marker (Fig. 2.1b). After separation of nuclear material from the cytoplasmic material, total RNA and subsequent polyadenylated mRNA was isolated from both fractions. Strand-specific libraries were then prepared and sequenced (see methods for details, Supplementary data 1a). For verification, Spearman correlations in gene expression levels were calculated among nuclear samples and cytoplasmic samples [68] (**Fig. S1**). Once validated, a computational pipeline was implemented for the identification of lncRNAs. Briefly, all nuclear and cytoplasmic RNA libraries were merged, resulting in one nuclear and one cytoplasmic merged file, then assembled into nuclear and cytosol transcriptomes independently using cufflinks. Subsequently, transcripts were filtered based on length, expression level, presence of nascent transcript from previously published GRO-seq dataset [8], and sequence coding potential (Fig. 2.1a). To specifically identify lncRNA candidates within the intergenic

regions and avoid any potential artefacts introduced by PCR amplification of the AT rich genome, we removed any predicted transcripts that have at least 30% overlap with annotated genes. Our goal was to select transcripts that are ≥ 200 bp in length, consistently expressed in both published nascent RNA and steady-state RNA expression profiles, and that are likely to be non-protein-coding genes. As a result, we identified a total of 1,768 intergenic lncRNAs in *P. falciparum* irrespective of the developmental stage (Supplementary data 1b). Nine hundred fifty-one lncRNAs have no overlap with any UTR regions. Overall, 1050 lncRNAs (~59%) overlapped with previously identified intergenic lncRNAs [49,51-53,56,57] and 718 lncRNAs were identified as novel in *P. falciparum* (Fig. 2.1c).

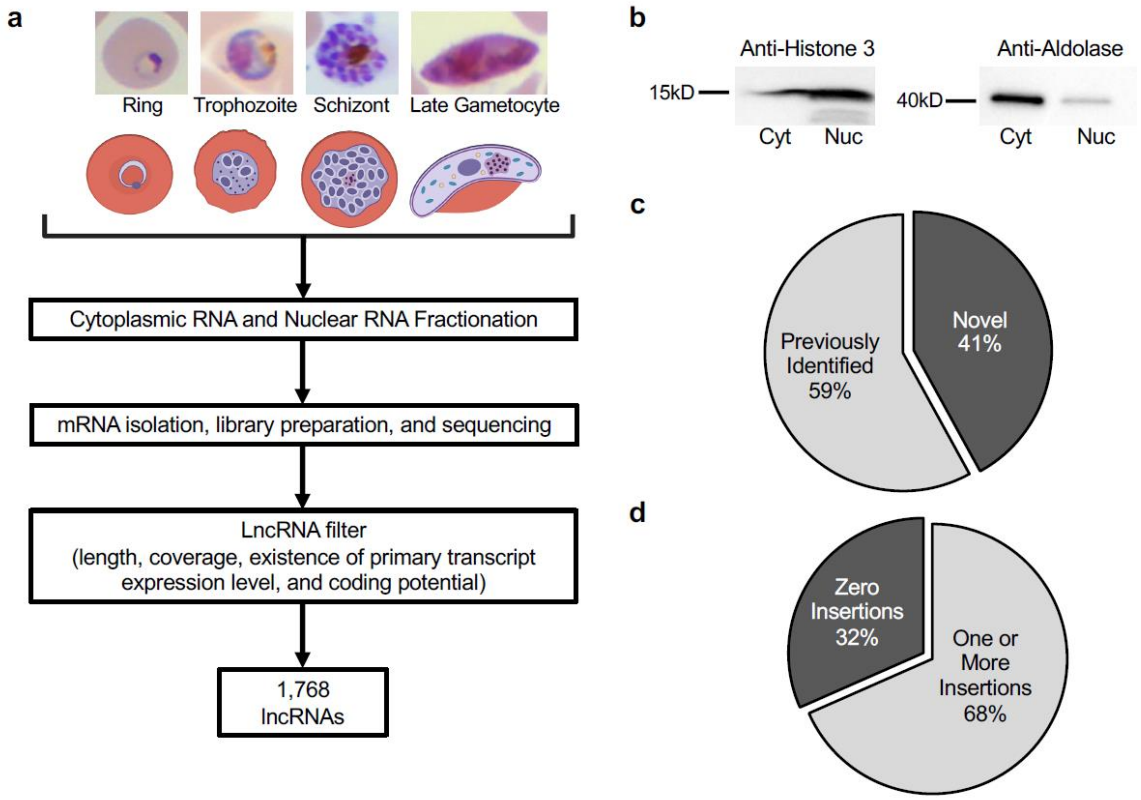

Fig. 2.1: Nuclear and cytoplasmic lncRNA identification. (a) A general overview of the lncRNA identification pipeline. Created with BioRender.com. (b) Validation of cell fractionation efficiency using anti-histone H3 and anti-aldolase as nuclear (Nuc) and cytoplasmic (Cyt) markers. Blot is representative of two independent biological replicates. (c) Comparison of lncRNA candidates with lncRNAs identified from previous publications. (d) Essentiality of lncRNAs using piggyBac insertion (Zhang et al., 2018). LncRNAs that cannot be disrupted are more likely to be essential.

To evaluate the essentiality of the lncRNAs identified in this study, we used piggyBac insertion sites from Zhang and colleagues [69]. In this work, the authors used a high-throughput transposon insertional mutagenesis method to distinguish essential and dispensable genes in the *P. falciparum* genome during the asexual stages of the parasite life cycle. We focused our analysis on the integration of the transposon that occurred within each of the identified lncRNAs. The piggyBac insertion site coordinates were

overlapped with the genomic ranges for all detected lncRNAs. This was performed after accounting for differences in the parasite strains used in the two studies. Overall, we were unable to uncover piggyBac insertion for 558 lncRNAs (31.6%), suggesting that these lncRNAs are either difficult to disrupt or are potentially essential for the parasite asexual development (Fig. 2.1d). Because we observed an insertion in 292 lncRNAs (16.5%) that were specifically detected in gametocyte stage, it will be important to validate at the phenotypic level whether those lncRNAs are essential during sexual differentiation rather than the asexual cycle. Additionally, significantly fewer insertions per possible insertion site (TTAA sequence) were found for telomeric as compared to subtelomeric lncRNAs, and for subtelomeric as compared to other lncRNAs (Fig. S2a). This suggests that piggyBac insertions in telomeric and sub-telomeric lncRNAs are also either difficult to disrupt due to their co-localization with heterochromatin or are more likely to be essential than others. It was also found that the 5' flanking regions of lncRNAs (Fig. S2b) had more insertions per possible site as compared to the rest of the lncRNAs, suggesting that, as a general trend, these regions of the lncRNAs are the most disposable while the gene body and 3' flanking regions may be more important for their function. Further analysis of the 558 lncRNA with zero piggyBac insertions illustrated that 158 of them (28.3%) overlapped by at least 1 bp with a gene, including UTRs, found to be essential in the Zhang study. For the remaining 400 lncRNAs (71.7%), the lack of insertions cannot be tied to a nearby gene rather than the lncRNA itself. In addition, novel lncRNAs were found to have fewer insertions (and thus be more likely essential) than previously known lncRNAs (38.6% of novel lncRNAs had zero insertions, whereas 26.8% of previously

known lncRNAs had zero insertions, and novel lncRNAs had fewer insertions per TTAA site (0.184 vs. 0.204)). Thus, many lncRNAs identified in this study may be essential for parasite survival. Although additional experiments will be needed to validate these results, it is also possible that some of the genes found to be essential in the Zhang study including some of the *rifin*, *stevor*, and pseudogenes, could be due to extensive overlap with the identified lncRNAs.

Length, GC content, and RNA stability of cytoplasmic and nuclear lncRNAs.

LncRNAs exhibit diverse subcellular distribution patterns, ranging from nuclear foci to cytoplasmic localization. Their localization patterns are linked to their distinct regulatory effects at their site of action [70,71]. Therefore, to better understand the potential function of the lncRNA in *Plasmodium*, we categorized the subcellular localization of our candidate lncRNAs into nuclear lncRNAs, cytoplasmic lncRNAs, or indistinguishable lncRNAs that are equally distributed in both fractions. Among the total identified 1,768 lncRNAs, 719 lncRNAs (41%) were enriched in the nuclear fraction, 204 lncRNAs (11%) were enriched in the cytoplasmic fraction, and 845 lncRNAs (48%) showed similar distribution between both subcellular fractions (Fig. 2.2a). Further, we explored the physical properties of these lncRNAs. We observed that lncRNAs are in general shorter in length and less GC-rich as compared to protein-encoding mRNAs (Fig. 2.2b and c). Using total steady-state RNA expression profiles and nascent RNA expression profiles (GRO-seq) (Fig. 2.2d), we then estimated the expression levels and stability of the lncRNAs. RNA stability was calculated as the ratio between steady-state RNA expression levels over

nascent RNA expression levels. We discovered that, although the overall life 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 lncRNAs are particularly lowly expressed and unstable as compared to the other two groups of lncRNAs (Fig. 2.2e). These observations are consistent with previous lncRNA annotation studies in human breast cancer cells [72] and noncoding RNA stability studies in mammalian genomes [73]. Our results suggest that the low expression level and the low stability of these lncRNAs may be the reason why they failed to be detected in previous identification attempts. By taking advantage of primary transcripts detected in our GRO-seq dataset, we significantly improved the sensitivity of lncRNA detection, especially for those localized in the nuclear fraction and expressed at a lower level.

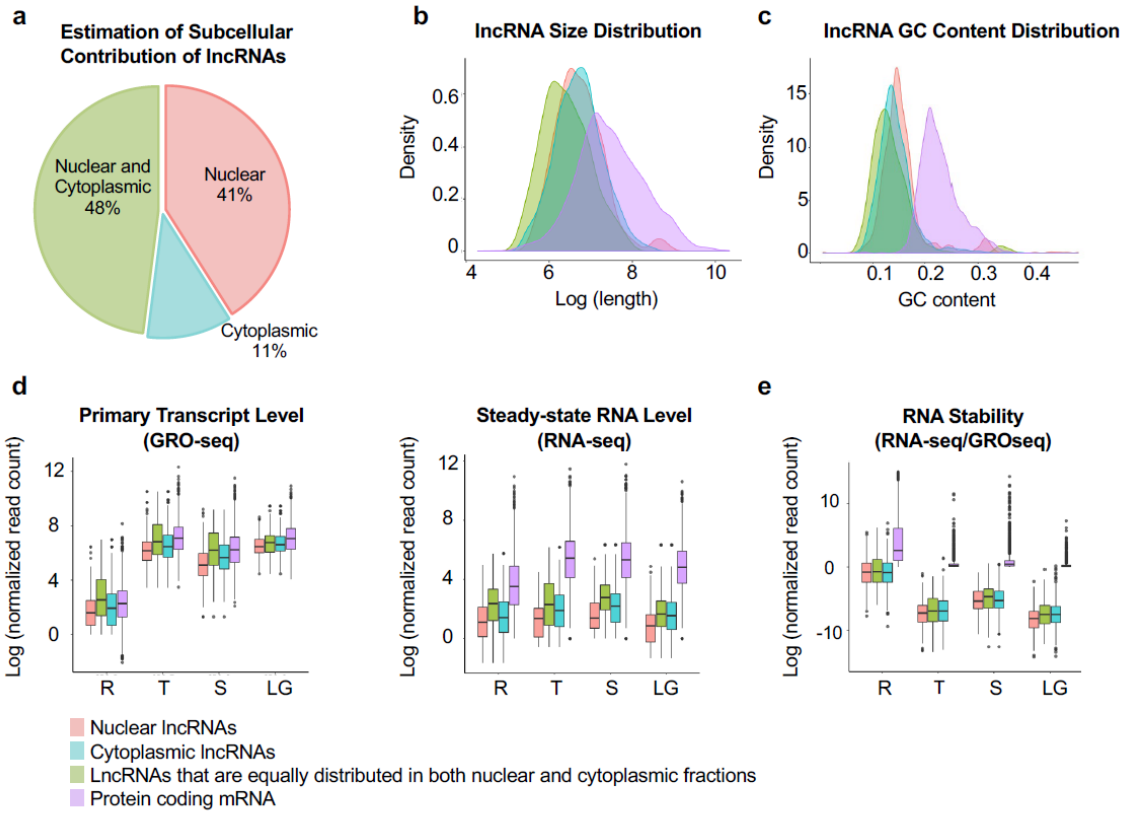

Fig. 2.2: Candidate lncRNA categorization. (a) A total of 1,768 lncRNA candidates were identified, covering 719 nuclear enriched lncRNAs (red), 204 cytoplasmic enriched lncRNAs (blue), and 845 lncRNAs found in both fractions (green). Cellular distribution of predicted lncRNAs is based on $\log_2$ fold change > or < 0.5 of summed nuclear vs cytoplasmic expression level. Density plots of size (b) and GC content (c) of lncRNA candidates and annotated protein encoding mRNAs (purple). (d) Expression levels of primary transcripts (left), steady-state RNA (middle) of lncRNA candidates and annotated protein encoding mRNAs for Ring (R), Trophozoite (T), Schizont (S) and Late Gametocyte (LG) stages. (e) Relative stability of lncRNA candidates and annotated protein encoding mRNAs. The stability is based on the ratio of RNA-seq/GRO-seq transcript level. Each box represents the 25/75 percentiles, the line across the box represents the median, and the whiskers represent maximum and minimum values. Outliers are indicated with dots.

Stage-specific expression of cytosolic and nuclear lncRNAs.

As lncRNAs often exhibit specific expression patterns in a tissue dependent manner, we investigated the stage specificity of identified candidate lncRNAs across the asexual and sexual life cycle stages. Using k-means clustering, we were able to group lncRNAs into 10 distinct clusters (Fig. 2.3a). Generally, nearly all lncRNAs showed a strong coordinated cascade throughout the parasite's life cycle. Similar to what was observed with mRNA, a large fraction of the lncRNAs was highly expressed at mature stages compared to the ring stages (Fig. 2.3b). Cluster 1 contains lncRNAs that are more abundantly expressed in the nuclear fraction of ring stage parasites and are lowly expressed in the nuclear fraction of schizont stage parasites. LncRNAs representative of this cluster are the lncRNA-TAREs. We observed that most lncRNA-TAREs identified in this study (19 out of 21) are clustered into this group with an average expression of 1.18 $\log_2$ fold change of nuclear to cytoplasmic ratio (Fig. 2.3a). The remaining two identified lncRNA-TAREs were found in cluster 6, where transcription peaks at the schizont stage. This finding validates our approach and suggests that lncRNAs in this cluster may contribute to the maintenance and regulation of chromatin structure and telomere ends. Approximately 28% of the identified lncRNAs are more abundantly found in either the nuclear or cytoplasmic fraction at the schizont stage (cluster 6, 7 and 8), after DNA replication and the peak of transcriptional activity observed at the trophozoite stage. We observed a few lncRNAs that are solely expressed during the asexual cycle with distinct changes in heterochromatin marks (Fig. 2.3c). Based on clustering analysis, we also found that 460 of our detected lncRNAs are exclusively expressed at a high level at the

gametocyte stage (cluster 9 and 10). Interestingly, two unique lncRNAs in this cluster, lncRNA-ch9 (Pf3D7_09_v3:1,384,241-1,386,630) and lncRNA-ch14 (Pf3D7_14_v3:3,148,960 - 3,150,115), were identified in a previous study to be located within heterochromatin regions marked by repressive histone marks H3K9me3 at the trophozoite stage [28] (Fig. 2.3d and 2.3e). At the gametocyte stage however, the H3K9me3 was lost. Additionally, both lncRNAs are transcribed from regions adjacent to gametocyte-specific genes. To validate the expression of these gametocyte-specific lncRNAs, we performed RT-PCR (Fig. S3 and Supplementary data 1c) as well as single cell RNA-seq (scRNA-seq) across key-stages of the parasite life cycle. LncRNA expression was visualized on the UMAP embedding generated from coding gene expression (Fig. 2.3f). LncRNA-chr9 and lncRNA-chr14 were expressed in sexual-stage parasites, with lncRNA-ch9 and lncRNA-ch14 were expressed in sexual-stage parasites, with specific enrichment in male and female gametocytes, respectively (Fig. 2.3f, right panel). Collectively, these results emphasize the stage-specific expression of parasite lncRNAs and the potential function of gametocyte-specific lncRNAs in regulating sexual development.

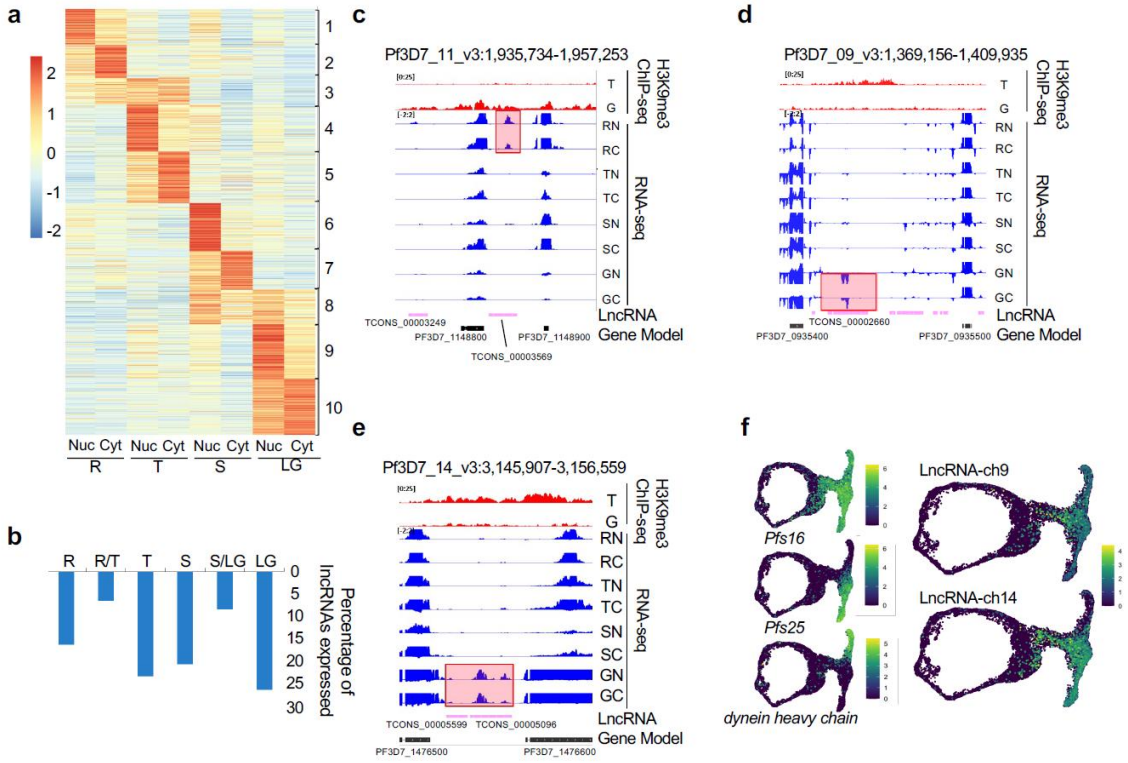

Fig. 2.3. Gene expression pattern of lncRNAs. (a) lncRNAs are grouped into 10 clusters based on their life cycle expression patterns in the Nuclear (Nuc) or Cytoplasmic (Cyt) fraction for the Ring (R), Trophozoite (T), Schizont (S) and Late Gametocyte (LT) stages. (b) Percentage of lncRNAs that are highly expressed at ring, trophozoite, schizont, and late gametocyte stages. Genome browser views of H3K9me3 ChIP-seq and RNA-seq datasets in the region of a representative asexual stage-specific lncRNA on chromosome 11 (c), two gametocyte-specific lncRNAs located at the intergenic regions of chromosome 9, lncRNA-ch9 (d) and 14, lncRNA-ch14 (e). RNA-seq data of Nuclear (N) and Cytoplasmic (C) fraction at Ring (R), Trophozoite (T), Schizont (S) and Gametocyte (G) stages are shown (e.g., RN=nuclear fraction at ring stage). (f) scRNA-seq analysis. 2-dimensional UMAP projection of *P. falciparum* parasites, both asexual (ring) and sexual (T-shape). Each dot represents a single cell. Left panel: Cells colored according to log-normalized gene expression values for gametocytes (*Pfs16* (PF3D7_0406200), top), females (*Pfs25* (PF3D7_1031000), middle), and males (*dynein heavy chain* (PF3D7_0905300), bottom). Right panel: Log-normalized expression of lncRNA-ch9 (top) and lncRNA-ch14 (bottom) across the *P. falciparum* life cycle.

Validation of lncRNA localization and stage-specific expression.

To validate the cellular localization of several candidate lncRNAs, we utilized RNA fluorescence in situ hybridization (RNA-FISH). We selected two candidates that were detected as enriched in the cytoplasmic fraction (lncRNA-267 (Pf3D7_08_v3: 1382128-1382689) and lncRNA-643 (Pf3D7_14_v3:1606672-1607587), four candidates detected as enriched in the nuclear fraction of asexual parasites (lncRNA-13 (Pf3D7_01_v3:491225-494291), lncRNA-178 (Pf3D7_06_v3:53758-54745), lncRNA-271 (Pf3D7_02_v3:590844-592940) and lncRNA-1494 (Pf3D7_06_v3:1311694-1312858) and two candidates detected as enriched in the nuclear fraction of sexually mature gametocytes (lncRNA-ch9 and lncRNA-ch14). Finally, we also selected a lncRNA that had been previously identified and known to be transcribed from the telomere region on chromosome 4, termed lncRNA-TARE4 (Pf3D7_04_v3:1194786-1199684) [47]as our positive control. Briefly, mixed stage parasites were fixed and hybridized to fluorescently labeled ~200-300 nucleotide antisense RNA probes (**Supplementary data 1c-d**). The hybridization images, representative of 15-20 parasites, demonstrate that the nuclear lncRNAs localize to distinct foci within or close to the DAPI-stained nuclei (Fig. 2.4a), while cytoplasmic lncRNAs are localized outside the DAPI-stained genomic DNA (Fig. 2.4b). Additionally, using RNA-FISH, we validated the stage-specific expression of our candidate lncRNAs. Specifically, expression of lncRNA-267 and lncRNA-13 were enriched at the ring and trophozoite stages; lncRNA-178 was expressed at the trophozoite and schizont stages; lncRNA-643 was expressed at the schizont stage only and lncRNA-TARE4 was expressed at all three asexual stages.

LncRNA-ch9 and lncRNA-ch14 were only expressed at the gametocyte stage. These results highlight that, similar to protein-coding transcripts, these candidate lncRNAs are developmentally regulated.

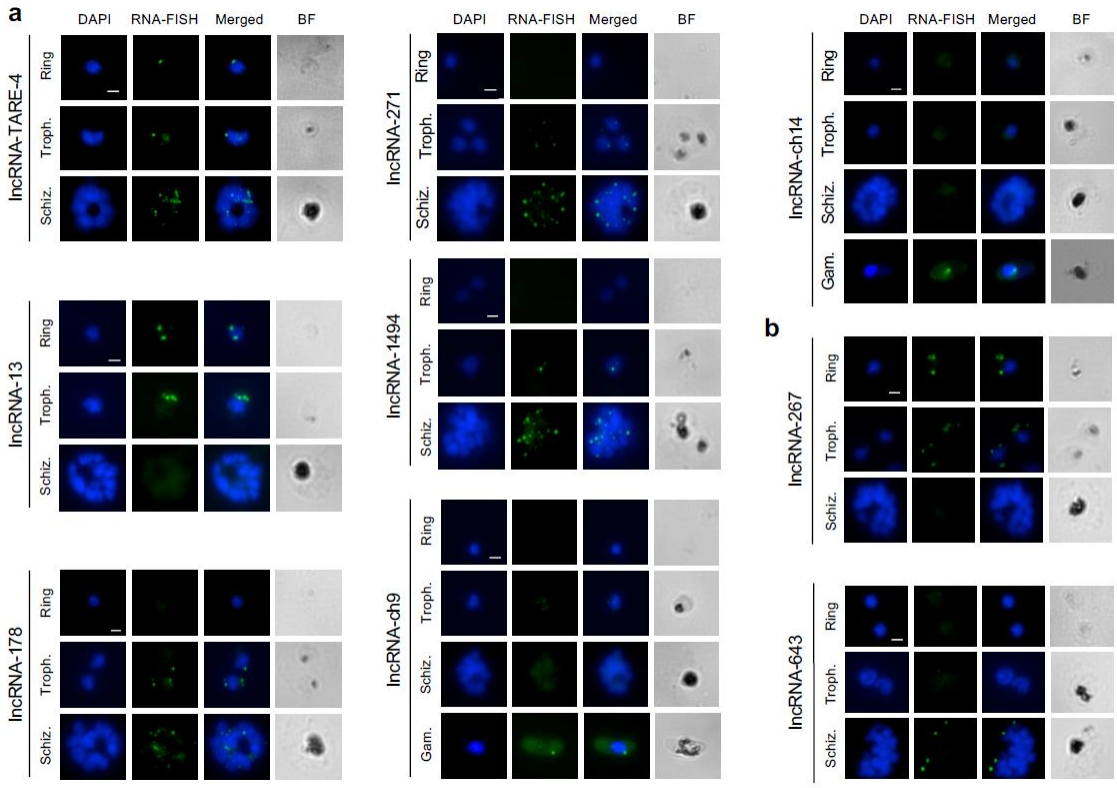

Fig. 2.4: **RNA-FISH experiments to show localization of several candidate lncRNAs**. (a) Nuclear lncRNAs (lncRNA-TARE4, lncRNA-13, lncRNA-178, lncRNA-271, lncRNA-1494, lncRNA-ch9 and lncRNA-ch14) colocalize with nuclei stained with DAPI in Ring, Trophozoite (Troph.), Schizont (Schiz.) and Gametocyte (Gam.) stages. (b) Cytoplasmic lncRNA-267 and lncRNA-643 do not colocalize with the nuclei stained with DAPI. Scale bar indicates 2 μm. Hybridization images are representative of approximately 15 stained parasites from two independent experiments. BF: brightfield.

Genomic maps of RNA-chromatin interactions.

The locations of the binding sites of most lncRNAs remain unknown. Accordingly, the role of lncRNAs in chromatin and gene regulation in eukaryotes, including the malaria parasite, has been mostly deduced from the indirect effects of lncRNA perturbation. To explore the role of lncRNAs in gene expression, we sought to identify occupancy sites of our selected candidate lncRNAs within the parasite genome. For an unbiased high-throughput discovery of RNA-bound DNA in *P. falciparum*, we adapted a method termed Chromatin Isolation by RNA Purification (ChIRP) (Fig. 2.5a)[74,75]. ChIRP-seq is based on affinity capture of target lncRNA:chromatin complex by tiling antisense-biotinylated-oligos to allow the identification of lncRNA-DNA binding sites at single base-pair resolution with high sensitivity and low background [76,77]. Such experiments can identify whether lncRNAs are working in *cis* on neighboring genes or in *trans* to regulate distant genes. ChIRP-seq is applicable to all detected lncRNAs and requires no knowledge of the RNA's structure. This method has recently been used in *Plasmodium* to investigate the role of ncRNA RUF6 on heterochromatin formation [77]. In our experiments, synchronized parasites were extracted and crosslinked. Parasite nuclei were then extracted, and chromatin was solubilized and sonicated. Biotinylated antisense oligonucleotides tiling our candidate lncRNAs (**Supplementary data 1e-f)** were hybridized to target RNAs and isolated using magnetic beads. These candidates correspond to the seven nuclear lncRNAs validated using RNA-FISH: lncRNA-TARE4, lncRNA-13, lncRNA-178, lncRNA-1494 and lncRNA-271 detected in the nuclear fraction of the asexual stages, as well as lncRNA-ch9 and lncRNA-ch14 detected in the nuclear fraction of the sexual

stages. To validate the specificity of the biotinylated oligonucleotides to target our RNA of interest, we performed RT-PCR following our RNA pulldown. RT-PCR results confirmed that lncRNAs and control serine tRNA ligase probes retrieve the selected lncRNA and the serine tRNA ligase RNA (PF3D7_0717700), respectively (Fig. 2.5b). No RNA was retrieved in the negative controls that were incubated with no probes or templates. These results confirm that the biotinylated probes target the RNA of interest with specificity. Purified DNA fragments were then sequenced using next-generation sequencing technology. An input control was used to normalize the signal from ChIRP enrichment.

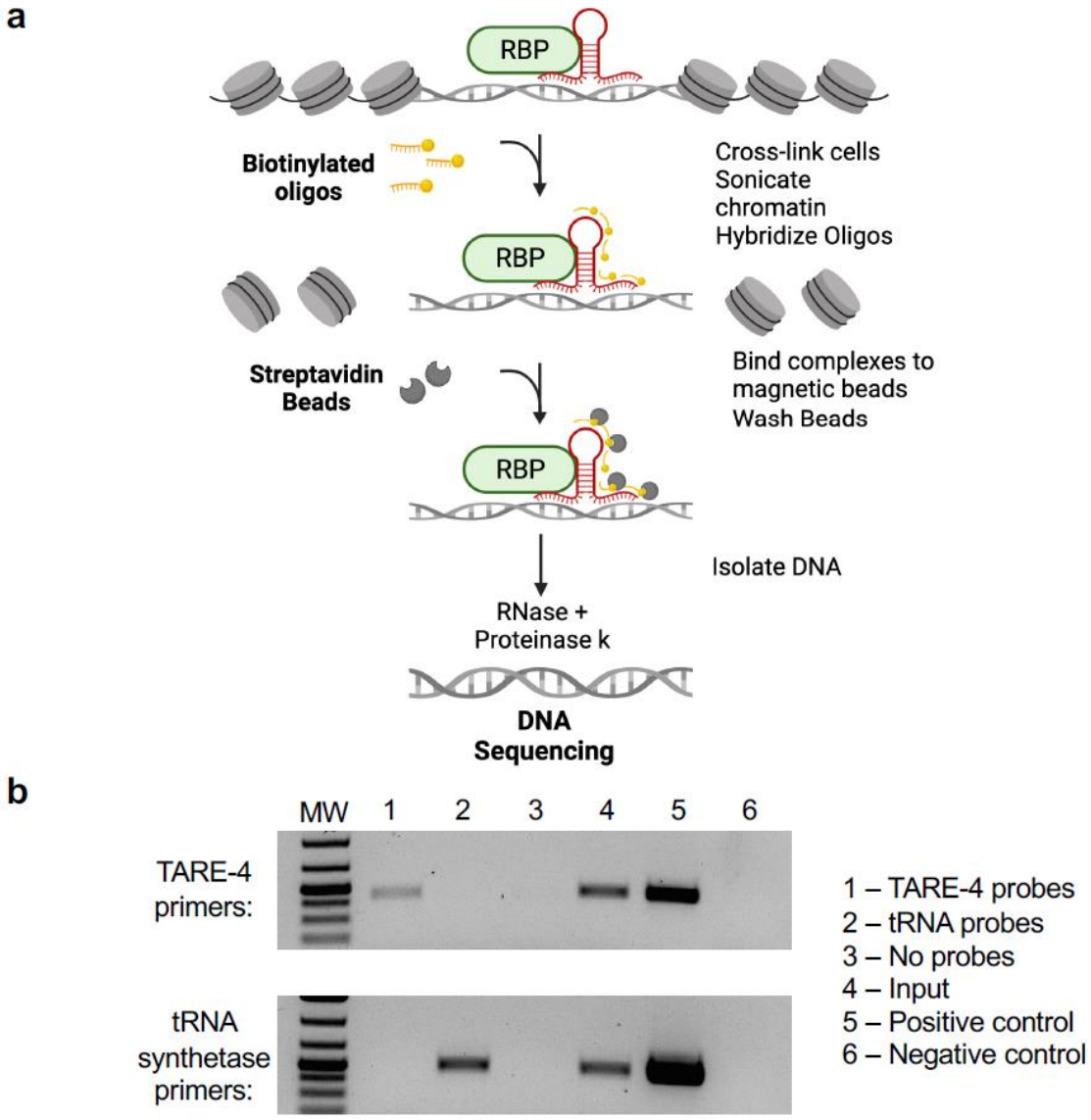

Fig. 2.5: Chromatin Isolation by RNA Purification (ChIRP). (a) Schematic representation of the ChIRP methodology. Created with BioRender.com. RBP: RNA-binding protein. (b) RT-PCR following the ChIRP protocol validates the specificity of the biotinylated antisense probes. Following the ChIRP-pulldown, the RNA fraction was analyzed, and RT-PCR results confirm that lncRNA-TARE-4 probes retrieve the lncRNA-TARE-4 RNA (535 bp PCR product) and the control serine tRNA ligase probes retrieve the serine tRNA ligase RNA (505 bp PCR product), respectively. No RNA was retrieved in the no probe control. RNA from a ChIRP-input sample as well as WT 3D7 parasites (wells 4 and 5, respectively) was used to confirm the lncRNA-TARE-4 and serine tRNA ligase primers. The negative controls (well 6) represent no template controls. RT-PCRs are representative of two independent replicates.

For all seven lncRNAs, ChIRP-seq experiments were performed in duplicate at the stages when the lncRNA was either highly or lowly expressed. We also generated several input controls for each analyzed stage to demonstrate the consistency of our coverage in various heterochromatin and euchromatin regions (Fig. S4). For all lncRNAs investigated here, ChIRP-seq performed at time points where the lncRNAs were least expressed retrieved low to no significant signals (Fig. 2.6a, 2.6b and Fig. S4). The lncRNA-TARE4, transcribed from the telomere region on chromosome 4, and expressed throughout the parasite life cycle stages investigated in this study was found to strongly interact with most telomeres in a very specific manner (Fig. 2.6a). Interestingly, one focus per cell for lncRNA-TARE4 was detected by RNA-FISH experiment at the ring stage. One focus per cell was also observed by IFA using an antibody against histone H3K9me3, known to also localize in the telomeres, as well as the subtelomeric and internal *var* gene cluster of the parasite genome [28,66] (Fig. S5). This data correlates nicely with the Hi-C data published previously [27,28,78] that demonstrate that most telomere ends, including *var* genes, interact with each other in a large heterochromatin cluster before and after DNA replication. LncRNA-13, transcribed on chromosome 1 and highly expressed at the trophozoite stage (Fig. S4), was found to be around surface antigen genes, including PF3D7_0113100 (SURFIN4.1) and PF3D7_1149200 (RESA, ring-infected erythrocyte surface antigen). Both the SURFIN and RESA families of proteins have been implicated in erythrocyte invasion-related processes and are transcribed in mature-stage parasites [79,80]. Given that a trophozoite stage lncRNA was identified adjacent to surface antigen genes which are transcribed at the schizont stage, lncRNA-13 is possibly playing a role in

recruiting chromatin modifying enzymes to edit the epigenetic state of the chromatin and allow the recruitment of transcription factor(s) needed to activate the transcription of these genes. To further investigate the possible link between lncRNAs and their role in epigenetics and regulation of gene expression, we developed a software pipeline in Python to identify all specific binding sites in the genome using Bowtie for mapping and PePr for peak calling [81]. We then aligned lncRNA ChIRP-seq signals across all 5' and 3' UTRs as well as the gene bodies. Similar to what was detected in higher eukaryotes, we discovered that the lncRNA occupancy is enriched either in the gene bodies for lncRNA-13 and lncRNA-178 or near the end of the 5′ UTR of each gene (lncRNA-1494 and lncRNA-271) (Fig. S6). While the data will need to be further validated at the molecular level, this pattern provides support for our candidate lncRNAs to promote either transcriptional elongation or transcriptional initiation, respectively. We next retrieved the genes closest to the identified lncRNA binding sites and calculated the $\log_2$ fold change of their expression from inactive to active stage. We compared the resultant information to the change in the expression profiles for all other genes in the *P. falciparum* genome (Fig. 2.6c). For each of the lncRNAs investigated, we detected a significant increase in the expression of the genes near the ChIRP signals. These results indicate that overall, the presence of lncRNA correlates with a significantly increased gene expression. To further demonstrate that lncRNAs interact specifically with DNA, we looked for motif enrichment (Fig. 2.6d). Motif analysis of ChIRP-seq data revealed one motif for lncRNA-13 (pval=$1.8e^{-3}$) occurring in 131 of the 138 retrieved lncRNA-13 sequences and two motifs for lncRNA-TARE-4 (pval=$3.7e^{-7}$ and pval=$5.1e^{-5}$) occurring in 72% and 61% of

the TARE-4 binding sites. These data demonstrate that our ChIRP-seq experiments were highly sensitive and specific, and that we were able to retrieve biological insights into their function.

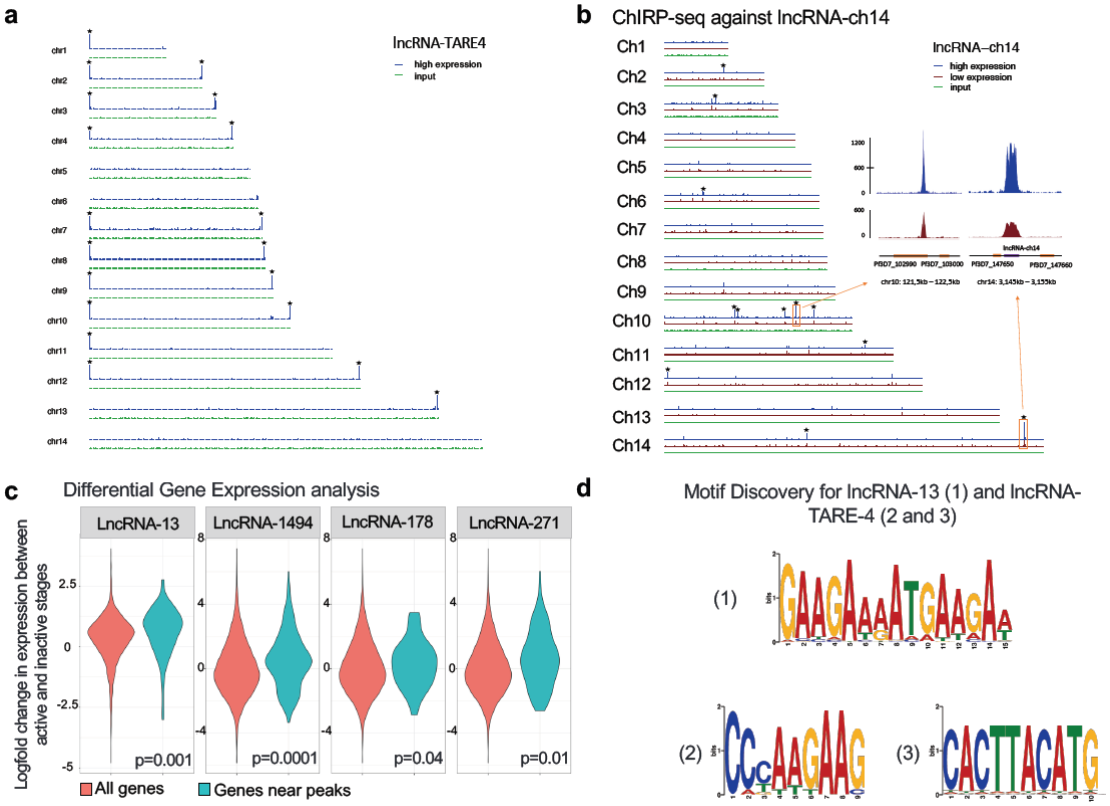

Fig. 2.6: ChIRP-seq reveals candidate lncRNA binding sites. (a) Genome-wide binding sites of lncRNA-TARE-4 and (b) lncRNA-ch14. Mapped, normalized reads from the active stage (top, blue track) and inactive stage (bottom, red track) are shown for each chromosome (Ch). Significant peaks are highlighted with an asterisk. (c) Differential gene expression analysis. The log2-fold change of gene expression was calculated for the genes closest to the lncRNA peaks between the inactive and active stage (right violin). This distribution was compared to the log2-fold change in expression for all other genes in the *Plasmodium* genome (left violin) using a two-sided t-test, and the p-values are reported at the bottom of each panel. (d) Motif identification. 100bp sequences centered at the peaks' summits were extracted, and we used STREME specifying 2nd order Markov model and default for the rest of parameters to search for possible motifs. We identified one motif for lncRNA-13 ($p=1.8e^{-3}$) occurring in 131 of the 138 (95%) lncRNA-13 sequences and two motifs for lncRNA-TARE-4, ($p=3.7e^{-6}$ and $p=5.1e^{-5}$), occurring in 72% and 61%, respectively, of the TARE binding areas.

We then focused our attention on ChIRP-seq data generated using probes against lncRNA-ch9 and lncRNA-ch14, two lncRNAs enriched in gametocytes. ChIRP-seq signals (Fig. 2.6b, Fig. S4, and Supplementary data 1k-l) showed significant enrichment in the genomic regions where the lncRNAs are transcribed. The lncRNA-ch9 lie between genes that have been implicated in gametocyte differentiation. These include PF3D7_0935500 [82], a *Plasmodium* exported protein of unknown function, PF3D7_0935600, a gametocytogenesis-implicated protein, and PF3D7_0935400, Gametocyte development protein 1. These three genes are known to be significantly up regulated in gametocytes and have been demonstrated to be essential to sexual commitment [66]. For lncRNA-ch14, the genes are PF3D7_1476500, a probable protein of unknown function, PF3D7_1476600 [83], a *Plasmodium* exported protein of unknown function and PF3D7_1476700, a lysophospholipase, three genes on chromosome 14 that are only detected either in gametocyte or ookinete stages. When overlaid with previous ChIP-seq data generated against histone H3K9me3 during the asexual and sexual stages of the parasite life cycle, we noticed that the presence of these lncRNAs correlate with a loss of H3K9me3 marks at the gametocyte stage. For lncRNA-ch14, 11 additional peaks were detected as statistically significant in the gametocyte stage (Fig. 2.6b, and Supplementary data 1). Most of these peaks were identified in the promotors of genes that were described as conserved *Plasmodium* protein of unknown function but were also known to be expressed in gametocyte including PF3D7_1145400, a dynamin-like protein overexpressed in female gametocytes [84]**.**While these data will need to be further validated, our results suggest that these lncRNAs may recruit histone demethylase and/or

histone acetyl transferase to change the epigenetic state of the chromatin and activate the expression of these genes during sexual differentiation. Collectively, these experiments propose that lncRNAs in the parasite could be essential to recruit chromatin remodeling and modifying enzymes as well as sequence-specific transcription factors to regulate gene expression.

Role of lncRNA-ch14 in sexual commitment and development.

We next sought to validate the role of one lncRNA in the parasite development. We therefore selected lncRNA-ch14, which was detected as upregulated in female gametocytes. We began by disrupting its full length via the CRISPR-Cas9 editing tool. After four unsuccessful attempts, we concluded that either the system was not able to target this genomic region or this large chromosomal deletion was lethal to the parasite. However, we were able to successfully disrupt the lncRNA-ch14 gene through insertion of a resistance marker spanning the position (Ch14:3,148,960 - 3,150,115) of the gene (Fig. S7a). Parasite lncRNA14 disruptive lines (two clones, named △lncRNA-ch14 B1 and F2) were recovered and validated via PCR and RT-PCR (Fig. S7b). We also confirmed the absence of obvious off-target effects via whole genome sequencing (WGS). We then examined parasite growth using our two selected clones along with wild-type NF54 parasites during the erythrocytic cycle. Growth was monitored in triplicates using Giemsa-stained blood culture smears for two full cycles. No significant difference was observed in the asexual stages of the △lncRNA-ch14 clones compared to

the WT (Fig. S8a). This indicates that partial disruption of lncRNA-ch14 does not have a significant role in the asexual blood stage replication.

We subsequently aimed to analyze the effect of △lncRNA-ch14 in gametogenesis. Relative gametocyte numbers were determined by microscopic examination of Giemsa-stained blood smears prepared from day 16 gametocyte cultures. To consider the impact prolonged culturing times may have on gametogenesis, the assays were conducted between our two △lncRNA-ch14 clones as well as two NF54 strains; a NF54 WT lab strain as well as the NF54 parental line used for the initial transfection. The NF54 parental line was maintained in culture in parallel with our selected △lncRNA-ch14 clones. We detected a significant decrease in mature stage V gametocytes in clones B1 and F2 compared to the WT NF54 line ($n=4$, $p<0.05$). This decrease was however not detected as significant between the parental NF54 line and △lncRNA-ch14 clones (Fig. 2.7a). To better understand discrepancies observed between the NF54 lines, we purified gDNA for WGS. While we confirmed successful disruption of ncRNA-ch14 in our two selected clones, we also identified a nonsense mutation in the gametocyte developmental protein 1 (*gdv1)* gene (PF3D7_0935400), a gene essential in sexual differentiation, in the parental line. The mutation detected indicated a premature stop codon leading to a C-terminal truncation of 39 amino acids (GDV39) similar to what was previously observed by Tibúrcio and colleagues [85,86]. Of the 110 reads covering the mutation site, 43 were shown to retain the reference base while 67 displayed the *gdv1* mutation described. This result suggests the development of a spontaneous mutation after transfection and that the

reduced number of gametocytes observed in the parental line were most likely a result of a significant portion (60%) of parasites with the *gdv1* mutation, not producing mature gametocytes. This mutation was however absent in both of our △lncRNA-ch14 clones as well as our NF54 WT (Supplementary data 2), explaining the discrepancies observed in our gametocyte induction assays with the NF54 parental line. Development of spontaneous mutations in culture attests the need of WGS to validate the phenotypes observed in the parasites for both the WT and genetically modified strains. As the parental line was still capable of generating healthy mature gametocytes, albeit at a lower frequency, and gametocytemia is normalized prior to mosquito feeds, the use of the NF54 parental line as our control to analyze the impacts of lncRNA-ch14 disruption on transmission to the mosquito was not considered to be a major issue because GDV1 is only essential for early sexual commitment. Our reasoning was that the gametocytes produced from the parental line would still closely resemble the △lncRNA-ch14 clones at the genomic level.

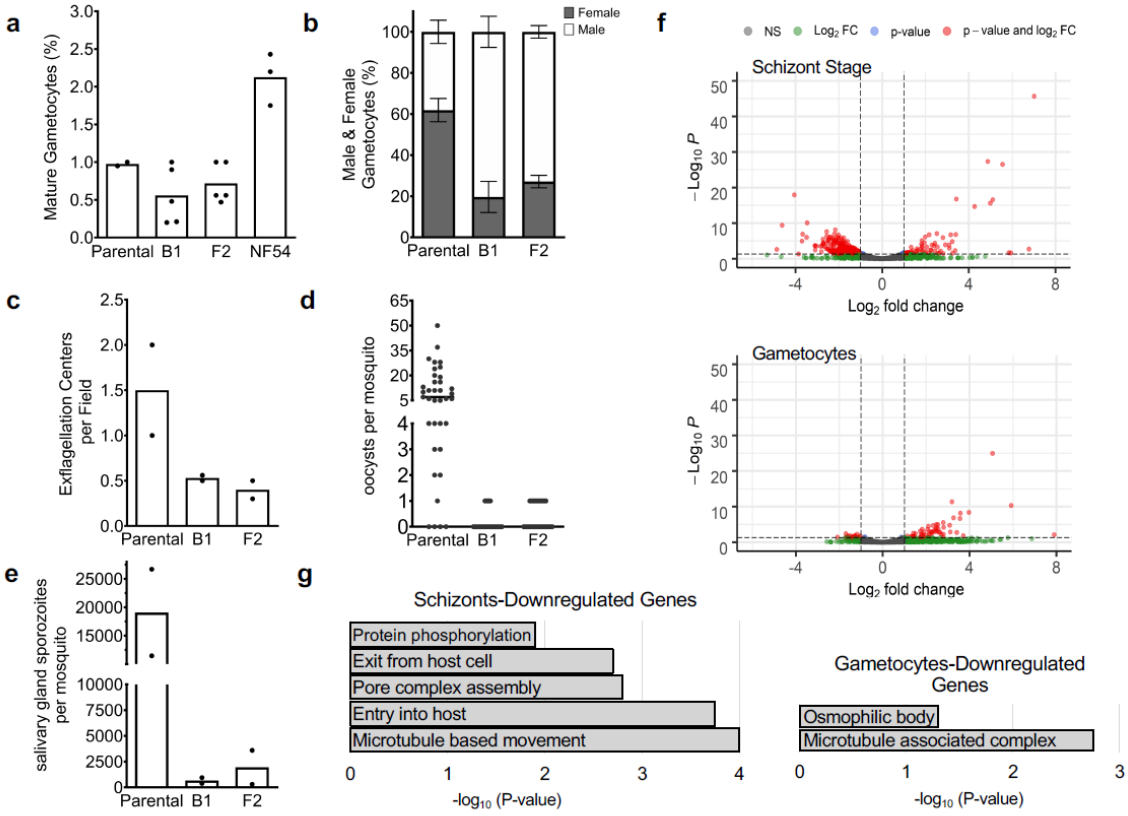
a
Mature Gametocytes (%)
Parental B1 F2 NF54
b
Male & Female Gametocytes (%)
Female
Male
Parental B1 F2
c
Exflagellation Centers per Field
Parental B1 F2
d
oocysts per mosquito
Parental B1 F2
e
salivary gland sporozoites per mosquito
Parental B1 F2
f
NS Log2 FC p-value p – value and log2 FC
Schizont Stage
−Log10 P
Log2 fold change
Gametocytes
−Log10 P
Log2 fold change
g
Schizonts-Downregulated Genes
Protein phosphorylation
Exit from host cell
Pore complex assembly
Entry into host
Microtubule based movement
-log10 (P-value)
Gametocytes-Downregulated Genes
Osmophilic body
Microtubule associated complex
-log10 (P-value)

Fig. 2.7: LncRNA-ch14 disruption design and characterization. (a) Percentage of mature gametocytes. Gametocyte cultures were sampled by Giemsa-stained blood smears to assess the percent of Stage V gametocytes. Assays were performed at least in duplicate and significance of the results was calculated using the one-way ANOVA with Dunnett's multiple comparison test (** $p < 0.01$, *** $p < 0.001$). (b) Percentage of gametocytes identified as male and female. Two independent biological experiments were performed and data are presented as mean values +/- SD (n>500 mature gametocytes). Significance of the results was calculated using the two-way ANOVA with Holm-Šídák correction (** $p < 0.01$, *** $p < 0.001$). (c) Exflagellation assays were performed and the number of exflagellation centers per field (n>10) were counted. Shown are results from two biological replicates. Bars indicate the mean. (d&e) Mosquito passage: Gametocyte cultures were fed to *Anopheles stephensi* mosquitoes. On day 11 post-infection, midguts were removed, and oocysts were counted (d) and on day 17 post-infection salivary glands were harvested and sporozoites were counted (e). Data are pooled from 2 biological replicates. Oocyst counts were performed with 15 to 25 mosquito midguts per experiment. Significance of the results was calculated using the Kruskal-Wallis test with Dunn's post-test (**** $p < 0.0001$). For salivary gland sporozoite counts, salivary glands from 20 mosquitoes were harvested, homogenized and sporozoites counted using a hemocytometer. Shown is the average number of salivary gland sporozoites for each of two experiments. Significance was calculated using the Chi-Squared tests (**** $p < 0.0001$). (f) Volcano plots for gene expression profile between the WT and △lncRNA-ch14 lines by $-Log_{10}$ P (y-axis) and log_2 Fold change (x-axis) in asexual stage parasites (Top) and in mature gametocytes (Bottom). NS: Non-significant; FC: Fold Change. (g). Bar graph representations of selected Gene Ontology (GO) enrichment of downregulated genes between WT and △lncRNA-ch14 lines are presented by Log_{10} P (y-axis) in asexual mature (Left) and mature gametocyte (Right) stages. Exact p-values and raw data are indicated in the Source data file.

We then analyzed the formation and ratio of mature male and female gametocytes between the parental line and △lncRNA-ch14 clones. To discriminate between male and female gametocytes, blood smears prepared from day 16 gametocyte cultures were stained with Giemsa and ≥100 mature stage V gametocytes were counted to determine sex ratio in each line. Male gametocytes can be distinguished from their female counterparts as they are less elongated with rounder ends and their cytoplasm is distinctly pink while the cytoplasm of female gametocytes, with their large stores of RNA, are dark blue. As shown in Fig. 2.7b, the male to female ratio was significantly affected in the △lncRNA-ch14 clones compared to control parasites. In control lines the ratio of female to male gametocytes was approximately 2:1, with the expected larger number of females than males. In contrast, in the lncRNA-ch14 clones it was approximately 2:7, with significantly more males than females ($n=2$, $p<0.05$) (Fig. 2.7b). These data suggest that lncRNA-ch14 has a role in the ratio of male and female gametocytes produced under our culture conditions. Exflagellation assays revealed a dramatic drop in microgametocyte exflagellation with an average of a 65% decrease in exflagellation centers observed in the △lncRNA-ch14 clones compared to the parental lines, indicating a defect in male gametogenesis and microgamete formation in △lncRNA-ch14 parasites (Fig. 2.7c). All together these data demonstrate a role of lncRNA-ch14 in gametogenesis.

We next investigated the transmissibility of △lncRNA-ch14 gametocytes to mosquitoes. Infectious blood meals were prepared with parental and △lncRNA-ch14 stage V gametocytes. Mosquitoes were fed with 0.2% mature stage V gametocyte infected blood

using membrane feeders as described earlier [87]. Mosquito midguts were dissected and analyzed on day 11 post blood feeding. As expected, our mutated parental line was able to produce a high number of oocysts and sporozoites. However, we detected a significant decrease in the number of oocysts per midgut in the △lncRNA-ch14 clones compared to the parental control. Both prevalence and intensity of infection were impacted in the △lncRNA-ch14 clones. While 90% of control infected mosquitoes had oocysts, only 17% and 37% of the △lncRNA-ch14 clones, B1 and F2, respectively were positive for oocysts. Additionally, the mosquitoes infected with the △lncRNA-ch14 clones had only 1 oocyst while the control had a median of 7 oocysts ($n=2$, $p<0.05$) (Fig. 2.7d). As expected, these lower oocyst numbers resulted in significantly lower salivary gland sporozoite numbers. On day 17 salivary glands were dissected from control and △lncRNA-ch14 infected mosquitoes and the average of number of salivary glands sporozoites from 2 biological replicates was 19,000 for the control line and 667 and 1993 for the △lncRNA-ch14 clones B2 and F1, respectively, a decrease of about 96% ($n=2$, $p<0.05$) (Fig. 2.7e). Altogether, though we could only partially disrupt our candidate lncRNA, we clearly demonstrate that lncRNA-ch14 has an important role in gametocyte development and in the infectivity of these gametocytes for mosquitoes.

Transcriptome perturbation in asexual and sexual stages.

Based on our phenotypic assays, we predicted that perturbation of lncRNA-ch14 would affect parasite gene expression. We therefore performed RNA-seq analysis for two biological replicates each of our WT NF54 and △lncRNA-ch14 clones. For each parasite

line, RNA was extracted at both schizont stage, before sexual commitment and at the late gametocyte stages after sexual commitment. For each respective stage, we observed up-regulation of 78 and 57 genes in the △lncRNA-ch14 lines compared to controls as well as downregulation of 383 and 18 genes (Fig. 2.7f and Supplementary data 3). Gene Ontology (GO) enrichment analysis revealed that up-regulated genes were involved in translation and cytoadherence (Pval= $10e^{-14}$) that are known to be expressed in a stage specific manner between the human and vector hosts (Fig. S8b). Downregulated genes were mostly involved in microtubule movement, cell signaling and oxidation-reduction processes that are known to be critical during sexual differentiation (Fig. 2.7g). Five specific genes known to be upregulated in female gametocytes were detected to be significantly downregulated in our △lncRNA-ch14 clones compared to our control line. These genes include PF3D7_1250100, the osmiophilic body protein G377; PF3D7_1407000, LCCL domain-containing protein; PF3D7_0719200, the NIMA related kinase 4; PF3D7_0525900, the NIMA related kinase 2, and PF3D7_1031000, the ookinete surface protein P25. Four male-specific genes were also detected as downregulated in the lncRNA-ch14 clones compared to the control. Those genes included PF3D7_1113900, the mitogen-activated protein kinase 2, PF3D7_1014200, the male gamete fusion factor HAP2; PF3D7_1216700, the perforin-like protein 2 and PF3D7_1465800, the putative dynein beta chain coding gene. All together this data confirms that lncRNA-ch14 controls, at least partially, the regulation of the transcripts known to be critical in gametocyte development including several key kinases involved in cell signaling relevant to gametogenesis.

Discussion

Many lncRNAs are now recognized as essential regulators of chromatin structure and gene expression in eukaryotes. While some of the identified lncRNAs have been shown to work in *cis* on neighboring genes, others seem to work in *trans* to regulate distantly located genes. Specifically, functions of nuclear lncRNAs have been determined as either directly promoting or repressing gene expression activity [88,89], guiding or enhancing the functions of regulatory proteins [37,89-92], or assisting the alteration of chromatin structures by shaping 3D genome organization [38,93,94].

The extent of lncRNA regulation in the human malaria parasite is only now starting to emerge. A few lncRNAs have already been suggested to regulate *var* gene expression [58,64,95] or drive sexual commitment [66,67]confirming that at least some of the identified *P. falciparum* lncRNA candidates may have a functional role in the parasite life cycle progression.

In *P. falciparum*, emerging evidence has shown that chromatin structure and genome organization are of vital importance for the parasite's gene expression and regulation system[28,96]. Depending on their localization and their specific interactions with DNA, RNA and proteins, lncRNAs can modulate chromatin and epigenetics in the nucleus or mRNA stability and translation in the cytosol, ultimately affecting gene expression. Therefore, identification and characterization of nuclear or cytoplasmic enriched lncRNAs may support the discovery of lncRNAs that are either chromatin-associated or

translational-associated regulators of gene expression in the parasite. The dataset generated in this study presents the first global detection of lncRNAs from different subcellular locations throughout several *P. falciparum* life cycle stages. By utilizing published total and nascent RNA expression profiles (GRO-seq [8]), we were able to significantly improve the sensitivity of lncRNA detection, especially for the identification of nuclear lncRNAs. Using both experimental and computational pipelines, we identified 1,768 lncRNAs covering 204 cytoplasmic enriched, 719 nuclear enriched, and 845 lncRNAs that localized to both fractions. Our data suggest that nuclear and cytoplasmic lncRNAs are coordinately expressed but cytoplasmic lncRNAs are less abundant as compared to the number of nuclear lncRNAs in the parasite. In addition, we observed that a small group of cytoplasmic lncRNAs is highly expressed at the trophozoite stage, the stage where a large proportion of genes are transcribed [8]. Though more in-depth studies will be required to confirm the functions of these trophozoite-expressed cytoplasmic lncRNAs, it is possible that some of these lncRNAs are involved in mRNA stability or translational regulation.

In our present work, we also observed that many lncRNAs enriched in the nuclear fraction, including the lncRNA-TAREs, are highly abundant at the ring and schizont stages. This finding suggests that some of these lncRNAs (cluster 1, Fig. 2.3a) are likely to be involved in heterochromatin maintenance or chromatin structure re-organization events, as previous ChIP-seq and Hi-C experiments have shown that epigenetics and chromatin are critical to gene expression at the initiation level [27]. Additionally, ChIRP

experiments mapping genome-wide binding sites of lncRNAs revealed that lncRNA-TARE4 binds to subtelomeric regions on multiple chromosomes as well as regulatory regions around genes involved in pathogenesis and immune evasion. Previous reports showed that subtelomeric regions as well as virulence gene families cluster in perinuclear heterochromatin. Therefore, evidence suggests a role for lncRNA-TARE4 in transcriptional and/or epigenetic regulation of parasite telomeric and subtelomeric regions by interacting with or recruiting histone-modifying complexes to targeted regions to maintain them in a heterochromatin state, much like the case of X chromosome inactivation regulated via lncRNA Xist [97].

Genomic occupancy of other lncRNAs explored here, suggest that these lncRNAs bind around the gene regions. In all cases investigated, a positive correlation was observed between the lncRNA expression and the expression of genes around the lncRNA occupancy sites (Fig. 2.6). Given already existing evidence for lncRNA-associated epigenetic modification and transcriptional regulation in other eukaryotes [98-100], it is likely that the lncRNAs identified in the parasite nucleus are responsible for coordinated recruitment of distinct repressing proteins and/or histone-modifying complexes to target loci. Additionally, we uncovered that the lncRNA binding sites were situated upstream of the start codon of target genes (Fig. S6). This pattern of lncRNA occupancy provides additional support for the idea that the lncRNAs explored here might have a role in recruiting protein complexes to promoter regions of target genes to regulate transcription, either by activating the formation of the pre-initiation complex or recruiting histone

modifiers. However, while additional experiments are needed to confirm the roles of these nuclear lncRNAs in the parasite, using ChIRP-seq, we demonstrate that genome-wide collections of RNA binding sites can be used to discover the DNA sequence motifs enriched by lncRNAs. These findings signify the existence of lncRNA target sites in the genome, an entirely new class of regulatory elements that could be essential for transcriptional regulation in the malaria parasite.

Genetic disruption of lncRNA-ch14, a transcript detected specifically in gametocytes, demonstrates that this lncRNA plays an important role in sexual differentiation and is required for onward transmission to the mosquito (Fig. 2.7a and b). This finding is supported by our transcriptomic analysis where we identified significant downregulation of genes involved in sexual differentiation including NEK and MAP kinases [101,102] ookinete/oocyst development [103] and microtubule function (i.e., dyneins and kinesins) most likely important in reshaping the parasite into sexual stages (Fig. 2.7f and 2.7g, Fig. S8b and Supplementary data 3). Importantly, the skewed sex ratio of the △lncRNA-ch14 parasites does not completely account for the dramatic decrease in the ability of these parasites to be transmitted to mosquitoes. Indeed, our data suggest that the gametocytes of the △lncRNA-ch14 parasites are less infectious, a result that is also supported by the decrease in exflagellation of the male gametocytes (Fig. 2.7c). It is currently difficult to assess infectiousness of female gametocytes, but we would hypothesize that these are also impacted by the disruption of lncRNA-ch14. While further experiments will be needed to further validate the effect of the full deletion or downregulation of the lncRNA-

ch14 transcript in the mosquitoes stage, the results presented here confirm that some of the lncRNAs identified in this study play a role in the parasite's sexual development and onward transmission to the mosquito.

Compared to the progress made in understanding lncRNA biology in higher eukaryotes, the field of lncRNAs in *Plasmodium* is still evolving. Analysis of promoter and gene body regions with available histone modification datasets (H3K9me3, H3K36me3, and H3K9ac) are still needed for further annotation of these candidate lncRNAs. It is clear that lncRNAs represent a new paradigm in chromatin remodeling and genome regulation. Therefore, this newly generated dataset will not only assist future lncRNA studies in the malaria parasite but will also help in identifying parasite-specific gene expression regulators that can ultimately be used as new anti-malarial drug targets.

Materials and Methods

Parasite culture.

P. falciparum 3D7 strain at ~ 8% parasitemia was cultured in human erythrocytes at 5% hematocrit in 25 mL of culture as previously described in [104]. Two synchronization steps were performed with 5% D-sorbitol treatments at ring stage within eight hours. Parasites were collected at early ring, early trophozoite, and late schizont stages. Parasite developmental stages were assessed using Giemsa-stained blood smears.

Nuclear and cytosolic RNA isolation.

Highly synchronized parasites were first extracted using 0.15% saponin solution followed by centrifugation at 1500 x g for 10 mins at 4°C. Parasite pellets were then washed twice with ice cold PBS and re-collected at 1500 x g. Parasite pellets were resuspended in 500 µL ice cold Cell Fractionation Buffer (PARIS kit, ThermoFisher; AM1921) with 10 µL of RNAse Inhibitor (SUPERaseIn 20U/µL, Invitrogen; AM2694) and incubated on ice for 10 minutes. Samples were centrifuged at 500 x g for 5 mins at 4°C. After centrifugation, the supernatant containing the cytoplasmic fraction was collected. Nuclei were resuspended in 500 µL Cell fractionation buffer and 15 µL RNAse Inhibitor as described above. To obtain a more purified nuclear fraction, the pellet was syringed with a 26G inch needle five times. The sample was incubated on ice for 10 mins and centrifuged at 500 x g for 5 mins at 4°C. The nuclear pellet was resuspended in 500 µL of ice-cold Cell Disruption Buffer (PARIS kit, ThermoFisher; AM1921). For both cytoplasmic and nuclear fractions, RNA was isolated by adding 5 volumes of Trizol LS Reagent (Life Technologies, Carlsbad, CA, USA) followed by a 5 min incubation at 37°C. RNA was then isolated according to manufacturer's instructions. DNA-free DNA removal kit (ThermoFisher; AM1906) was used to remove potential genomic DNA contamination according to manufacturer's instruction, and the absence of genomic DNA was confirmed by performing a 40-cycle PCR on the PfAlba3 gene using 200 to 500 ng input RNA.

mRNA isolation and library preparation.

Messenger RNA was purified from total cytoplasmic and nuclear RNA samples using NEBNext Poly(A) nRNA Magnetic Isolation module (NEB; E7490S) with manufacturer's instructions. Once mRNA was isolated, strand-specific RNA-seq libraries were prepared using NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB; E7420S) with library amplification specifically modified to accommodate the high AT content of *P. falciparum* genome: libraries were amplified for a total of 12 PCR cycles (45 s at 98°C followed by 15 cycles of 15 s at 98°C, 30 s at 55°C, 30 s at 62°C], 5 min 62°C). Libraries were then sequenced on Illumina NExtSeq500 generating 75 bp paired-end sequence reads.

Sequence mapping.

After sequencing, the quality of raw reads was analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The first 15 bases and the last base were trimmed. Contaminating adaptor reads, reads that were unpaired, bases below 28 that contained Ns, and reads shorter than 18 bases were also filtered using Sickle (https://github.com/najoshi/sickle) [105]. All trimmed reads were then mapped to the *P. falciparum* genome (v34) using HISAT2 [106] 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 properly paired (samtools v 0.1.19-44428cd [107]) and are likely to be PCR duplicates (Picard tools v1.78,

broadinstitute.github.io/picard/). The final number of working reads for each library is listed in Supplementary data 1. For genome browser tracks, read coverage per nucleotide was first determined using BEDTools [108] and normalized per million mapped reads.

Transcriptome assembly and lncRNA identification.

To identify lncRNAs in the nuclear and cytoplasmic fractions, we first merged all nuclear libraries and cytoplasmic libraries for each replicate, resulting in one pair of nuclear and cytoplasmic dataset per replicates. Next, we assembled the transcriptome (cufflinks v2.1.1) [109] for each of the datasets using the following parameters: -p 8 -b PlasmoDB-34_Pfalciparum3D7_Genome.fasta -M PlasmoDB-34_Pfalciparum3D7.gff --library-type first strand -I 5000. After transcriptome assembly, we filtered out transcripts that are less than 200 basepairs and were predicted to be protein-coding (CPAT, http://lilab.research.bcm.edu). We then merge transcripts in both replicates using cuffmerge and removed any transcripts that are on the same strand and have more than 30% overlap with annotated regions (BEDTools intersect). Lastly, we further selected transcripts that have both primary and steady-state transcriptional evidence. For primary transcription, we used GRO-seq dataset (GSE85478) and removed any transcript that has a read coverage below 15% of the median expression of protein-encoding genes, as well as transcript that has an FPKM count less than 10 at any given stage. The same filtering criteria were also applied to the steady-state RNA-seq expression profiles.

To estimate the cellular location for each predicted lncRNA, we first calculated the summed read count of all nuclear libraries and the summed read count of all cytoplasmic libraries. Then, we measured the log2-fold change (log2 FC) of the summed nuclear signal to the summed cytoplasmic signal. Any transcripts with a log2 FC value above 0.5 were classified as nuclear enriched lncRNAs, and any transcripts with a log2 FC value below -0.5 were classified as cytoplasmic enriched lncRNAs. In addition, lncRNAs with a log2 FC value between the above thresholds were classified as lncRNAs expressed equally in both fractions.

Western blot.

Mixed-stage parasites were collected as described above. Parasite pellets were gently resuspended in 500 µL of ice-cold Cell Fractionation Buffer (PARIS kit, ThermoFisher; AM1921) and 50 µL of 10X EDTA-free Protease inhibitor (cOmplete Tablets, Mini EDTA-free, EASY pack, Roche; 05 892 791 001). Solution was incubated on ice for 10 mins and the sample was centrifuged for 5 mins at 4°C and 500 x g. The supernatant containing cytoplasmic fraction was collected carefully and the nuclear pellet was resuspended in 500 µL Cell Fractionation Buffer followed by needle-lysis 5x using 26 G inch needle. Nuclei were collected again at 4°C and 500 x g. The supernatant was discarded and the nuclei pellet in 500 µL of Cell Disruption Buffer (PARIS kit, ThermoFisher; AM1921) and incubated on ice for 10 minutes. The nuclear fraction was then sonicated 7x with 10 seconds on/30 seconds off using a probe sonicator. Extracted nuclear protein lysates were incubated for 10 mins at room temperature and centrifuged

for 2 mins at 10,000 x g to remove cell debris. Seven micrograms of parasite cytoplasmic and nuclear protein lysates were diluted in a 2X laemmli buffer at a 1:1 ratio followed by heating at 95°C for 10 mins. Protein lysates are then loaded on an Any-KD SDS-PAGE gel (Bio-Rad, 569033) and run for 1 hour at 125 V. Proteins were transferred to a PVDF membrane for 1 hr at 18 V, then stained using commercial antibodies generated against histone H3 (1:3,000 dilution, Abcam; ab1791) and PfAldolase (1:1,000 dilution, Abcam; ab207494), and secondary antibody, Goat Anti-Rabbit IgG HRP Conjugate (1:25,000 dilution, Bio-Rad; 1706515). Membranes were visualized using the Bio-Rad ChemiDoc MP Gel Imager and images were treated using Image Lab software. Uncropped blots are presented in Source data file.

PiggyBac insertion analysis.

To analyze lncRNA essentiality, we used piggyBac insertion sites from [69], who performed saturation mutagenesis to uncover essential genes in *P. falciparum*. Since that study used an NF54 reference genome, we converted the coordinates to be applicable to the 3D7 reference genome (v38, PlasmoDB), using liftOver (Kent tools v427, UCSC Genome Bioinformatics group, https://github.com/ucscGenomeBrowser/kent). A chain file for the two genomes, needed for liftOver, was manually constructed as described here:

http://genomewiki.ucsc.edu/index.php/LiftOver_Howto.

Custom Python scripts were used to overlap insertion site coordinates with lncRNA ranges to count the number of insertions that occurred in each lncRNA, as well as to locate TTAA sites (sites where piggyBac insertions could potentially occur) in the genome and count the number of TTAA sites in each lncRNA. These scripts were also used to determine the normalized location of each TTAA site and insertion site, in one of 50 windows either across the lncRNA range or also including the 5' and 3' flanking regions, which were each given 50% of the length of the lncRNA. The ratio of number of insertion sites to number of TTAA sites within a lncRNA was used as a loose measure of essentiality.

Estimation of transcript stability.

Read coverage values were calculated from total steady-state RNA datasets (SRP026367, SRS417027, SRS417268, SRS417269) using BEDTools v2.25.0. The read counts were then normalized as described in the original publication, and ratios between RNA-seq and GRO-seq coverage values were calculated for each lncRNA and gene. This ratio reflects the relative abundance of the mature RNA transcript over its corresponding primary transcript and is a simple but convenient measurement for transcript stability.

Reverse transcription PCR.

Total RNA was isolated from 10 mL of mixed-stage asexual *P. falciparum* culture and 25 mL of late gametocyte stage culture. Total RNA quality was checked on an agarose gel and genomic DNA contamination was removed using a DNA-free DNA removal kit

(ThermoFisher; AM1906) according to manufacturer's instructions. The absence of genomic DNA was validated using a primer set targeting an intergenic region within PfAlba3 (PF3D7_1006200). Approximately 1 µg of DNase I treated RNA from each sample was used in a 35-cycle PCR reaction to confirm the absence of genomic DNA contamination. DNase-treated total RNA was then mixed with 0.1 µg of random hexamers, 0.6 µg of oligo-dT (20), and 2 µL 10 mM dNTP mix (Life Technologies) in total volume of 10 µL, incubated for 10 minutes at 70°C and then chilled on ice for 5 minutes. This mixture was added to a solution containing 4 µL 10X RT buffer, 8 µL 20 mM MgCl2, 4 µL 0.1 M DTT, 2 µL 20U/µl SuperaseIn and 1 µL 200 U/µL SuperScript III Reverse Transcriptase (Invitrogen, 18080044). 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 (Invitrogen, 10812014), 3 µL 10 mM dNTP mix (Life Technologies), 4 µL 10 U/µl *E. coli* DNA Polymerase (NEB, M0209), 1 µL 10 U/µL *E. coli* DNA ligase (NEB, M0205) and 1 µL 2 U/µL *E. coli* RNase H (Invitrogen, 18021014). Samples were incubated for 2 h at 16°C and double stranded cDNA was purified using AMPure XP beads (Beckman Coulter, A63881). For testing transcription activity of predicted genes, 450 ng of double stranded cDNA was mixed with 10 pmole of both forward and reverse primers. DNA was incubated for 5 minutes at 95°C, then 30s at 98°C, 30s at 55°C, 30s at 62°C for 25 cycles. All primers used for PCR validation are listed in Supplementary data 1.

Single-cell sequencing and data processing.

P. falciparum strain NF54 was cultured in O+ blood in complete RPMI 1640 culture medium at 37°C in a gas mixture of 5% O_2/5% CO_2/90% N_2, as described previously [9,110]. Sexual commitment was induced at 1% parasitemia and 3% hematocrit and culture media were supplemented with 10% human serum. After 4, 6 and 10 days post sexual commitment, samples were taken from the culture for single cell sequencing. Cells from each day were loaded into separate inlets in a 10X chromium controller using the manufacturer's instructions for a 10,000-target cell capture. Libraries for the days 4 and 6 samples were obtained using Chromium 10X version 2 chemistry, whereas libraries for the day 10 sample were obtained using version 3 chemistry. Cells were sequenced on a single lane of a HighSeq4000 using 150-bp paired-end reads. Raw reads were mapped to a custom gtf containing lncRNA coordinates appended to the *P. falciparum* 3D7 V3 reference genome (www.sanger.ac.uk/resources/downloads/protozoa/). Read mapping, deconvolution of cell barcodes and UMIs and the generation of single cell expression matrices were performed using the CellRanger pipeline v 3.0.0. LncRNA regions were labeled as 'protein coding' to be prioritized in STAR mapping in CellRanger. CellRanger was also run separately for each sample using the 3D7 reference genome that did not contain the appended non-coding regions for comparison. Resultant count matrices were loaded into the R package Seurat (v3.2.2) for pre-processing.

Quality control and lncRNA expression.

Single-cell transcriptomes (SCTs) were log-normalized, and expression scaled using Seurat (v.3.2.2). Each cell was assigned a stage by mapping to the Malaria Cell Atlas [110] using scmap-cell (v1.8.0). Cells were assigned the stage of their closest neighbor in the Malaria Cell Atlas if they reached a cosine similarity of > 0.2. Cells identified as an early/late ring or late schizont containing < 50 UMI/cell and < 50 genes/cell were removed due to poor quality. Cells mapped to late stages, or cells not assigned to a stage in the Malaria Cell Atlas were removed if they contained < 100 UMIs/cell or < 80 genes/cell. Data from days 4, 6 and 10 were integrated together using Seurat's IntegrateData function using 2000 integration anchors and 10 significant principal components. A variance stabilizing transformation was performed on the integrated matrix to identify the 750 most highly variable coding genes, and these were used to perform a principal component (PC) analysis. Significant PCs were then used to calculate three-dimensional UMAP embeddings using only coding genes. LncRNA expression was visualized on the UMAP embedding generated from coding gene expression using the package ggplot2 to assign stage-specific expression for the lncRNA.

RNA in situ hybridization (RNA-FISH).

RNA FISH was performed with slight modifications as described by Sierra-Miranda, 2012 [60] on mixed-stage asexual and gametocyte stage parasites. Antisense RNA probes for seven nuclear lncRNAs; -TARE4, -178, -13, -1494, -271, -4076 -ch9, -ch14 and two cytoplasmic lncRNAs; -267, -643, were labeled by in vitro transcription in the presence

of fluorescein. RNA FISH was also performed using sense RNA probes as controls. Briefly, fixed and permeabilized parasites were incubated with RNA probes overnight at 37°C. Parasites were washed with 2x SSC three times for 15 mins each at 45°C followed by one wash with 1x PBS for 5 mins at room temperature. The slides were mounted in a Vectashield mounting medium with DAPI and visualized using the Olympus BX40 epifluorescence microscope. Images were treated with ImageJ. Pictures are representative of 15-20 positive parasites examined.

Immunofluorescence assays.

Parasites were fixed with 4% paraformaldehyde and 0.0075% glutaraldehyde for 15 min at 4˚C, and then sedimented on Poly-D-lysine coated coverslips for 1 h at room temperature. After PBS washes, parasites were permeabilized and saturated with 0.2% Triton X-100, 5% BSA, 0.1% Tween 20 in PBS for 30 min at room temperature. Anti-H3K9me3 mAb (Abcam, ab184677) was diluted at 1:500 in 5% BSA, 0.1% Tween 20 and PBS, and applied for 1 h at room temperature. After PBS washes, Goat anti-Mouse Alexa Fluor 488 (Invitrogen, A11001) was diluted at 1:2000 and applied for 1 h at room temperature. Coverslips were mounted in Vectashield Antifade Mounting Medium with DAPI (Vector Laboratories, H-1200). Images were acquired using a Keyence BZ-X810 fluorescence microscope and treated with ImageJ.

Chromatin isolation by RNA purification (ChIRP).

ChIRP-seq experiments were performed in duplicate for all nuclear lncRNAs, at the time point of highest lncRNA expression and lowest expression. Synchronized parasite cultures were collected and incubated in 0.15% saponin for 10 min on ice to lyse red blood cells. Parasites were centrifuged at 3234 x g for 10 mins at 4°C and subsequently washed three times with PBS by resuspending in cold PBS and centrifuging for 10 mins at 3234 x g at 4°C. Parasites were cross-linked for 15 mins at RT with 1% glutaraldehyde. Cross-linking was quenched by adding glycine to a final concentration of 0.125 M and incubating for 5 mins at 37°C. Parasites were centrifuged at 2500 x g for 5 mins at 4°C, washed three times with cold PBS and stored at -80°C.

To extract nuclei, parasite was first incubated on in nuclear extraction buffer (10 mM HEPES, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 0.5 mM 4-(2-aminoethyl) benzenesulfonyl fluoride hydrochloride (AEBSF), EDTA-free protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Roche)) on ice. After 30 mins, Igepal CA-630 (Sigma-Aldrich, I8896) was added to a final concentration of 0.25% and needle sheared seven times by passing through a 26 G ½ needle. Parasite nuclei were centrifuged at 2500 x g for 20 mins at 4°C and resuspended in shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris–HCl pH 7.5, EDTA-free protease inhibitor cocktail and phosphatase inhibitor cocktail). Chromatin was fragmented using the Covaris UltraSonicator (S220) to obtain 100-500 bp DNA fragments with the following settings:

5% duty cycle, 140 intensity incident power, 200 cycles per burst. Sonicated samples were centrifuged for 10 mins at 17000 x g at 4°C to remove insoluble material.

Fragmented chromatin was precleared using Dynabeads MyOne Streptavidin T1 (Thermo Fisher, 65601) by incubating for 30 mins at 37°C to reduce non-specific background. Per ChIRP sample using 1 mL of lysate, 10 µL each was removed for the RNA input and DNA input, respectively. Each sample was diluted in 2x volume of hybridization buffer (750 mM NaCl, 1% SDS, 50 mM Tris-Cl pH 7.5, 1 mM EDTA, 15% formamide, 0.0005x volume of AEBSF, 0.01x volume of SUPERase-In (Ambion, AM2694) and 0.01x volume of protease inhibitor cocktail). ChIRP probes used for each lncRNA (see Supplementary data 1) were pooled, heated at 85°C for 3 mins and cooled on ice. ChIRP probes were added to each sample (2 µL of 100 µM pooled probes per sample) and incubated at 37°C with end-to-end rotation for 4 hours. Prior to completion of hybridization, Dynabeads MyOne Streptavidin T1 beads were washed three times on a magnet stand using lysis buffer (50mM Tris-Cl pH 7, 10mM EDTA, 1% SDS). After the hybridization, 100 µL of washed T1 beads were added to each tube and incubated for 30 mins at 37°C. Beads were washed with wash buffer (2x SSC, 0.5% SDS, 0.005x volume of AEBSF) and split evenly for isolation of DNA and RNA fractions.

For RNA isolation, the RNA input and chromatin-bound beads were resuspended in RNA elution buffer (100 mM NaCl, 10 mM Tris-HCl pH 7.0, 1 mM EDTA, 0.5% SDS, 1 mg/mL Proteinase K), incubated at 50°C for 45 mins, boiled at 95°C for 15 mins and

subjected to trizol:chloroform extraction. Genomic DNA contamination was removed using a DNA-free DNA removal kit (ThermoFisher, AM1906) according to manufacturer's instructions. The absence of genomic DNA was validated using a primer set targeting an intergenic region within PfAlba3 (PF3D7_1006200) in a 35-cycle PCR reaction. DNase-treated RNA was then mixed with 0.1 µg of random hexamers, 0.6 µg of oligo-dT (20), and 2 µL 10 mM dNTP mix (Life Technologies) in total volume of 10 µL, incubated for 10 minutes at 70°C and then chilled on ice for 5 minutes. This mixture was added to a solution containing 4 µL 10X RT buffer, 8 µL 20 mM MgCl2, 4 µL 0.1 M DTT, 2 µL 20U/µl SUPERase-In and 1 µL 200 U/µL SuperScript III Reverse Transcriptase. 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 followed by a 20 min incubation with 1 µL 2 U/µL *E. coli* RNase H at 37°C. Prepared cDNA was then subjected to quantitative reverse-transcription PCR for the detection of enriched TARE-4 and serine tRNA ligase transcripts with the following program: 5 minutes at 95°C, 30 cycles of 30s at 98°C, 30s at 55°C, 30s at 62°C and a final extension 5 min at 62°C. All primers used for PCR validation are listed in Supplementary data 1.

Libraries from the ChIRP samples were prepared using the KAPA Library Preparation Kit (KAPA Biosystems). Libraries were amplified for a total of 12 PCR cycles (12 cycles of 15 s at 98°C, 30 s at 55°C, 30 s at 62°C]) using the KAPA HiFi HotStart Ready Mix (KAPA Biosystems). Libraries were sequenced with a NextSeq500 DNA sequencer (Illumina). Raw read quality was first analyzed using FastQC

(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the *P. falciparum* genome (v38, PlasmoDB) using Bowtie2 (v2.4.2). Duplicate, unmapped, and low quality (MAPQ < 20) reads were filtered out using Samtools (v1.9), and only uniquely mapped reads were retained. All libraries, including the input, were then normalized by dividing by the number of mapped reads in each of them. For each nucleotide, the signal from the input library was then subtracted from each of the ChIRP-seq libraries, and any negative value was replaced with a zero. Genome tracks were generated by the R package ggplot2.

Peak Calling.

Peaks were called using PePr v1.1. For a given lncRNA of interest, the tool was run in differential binding analysis mode using the filtered ChIRP-seq libraries for when the lncRNA was active versus non-active with the following parameters specified: -peaktype broad -threshold $1e^{-10}$. The top 25% of all reported peaks were selected because they exhibited the strongest signal (see Fig. S4) and used in downstream analyses. For differential gene expression analysis, the closest gene to each peak was selected, and its expression in the two stages (active versus inactive) was obtained from [29,68,111].

LncRNA-ch14 gene disruption.

Gene knockout (KO) for the long non-coding RNA 14 (lncRNA-ch14) spanning position (Ch14:3,148,960 - 3,150,115) on chromosome 14 was performed using a two-plasmid design. The plasmid pDC2-Cas9-sgRNA-hdhfr [112], gifted from Marcus Lee (Wellcome

Sanger Institute) contains the SpCas9, a site to express the sgRNA, and a positive selectable marker human dihydrofolate reductase (*hdhfr*). The sgRNA was selected from the database generated by [113] and cloned into pDC2-Cas9-sgRNA-hdhfr at the BbsI restriction site. The homology directed repair plasmid (modified pDC2-donor-*bsd* without eGFP) was designed to insert a selectable marker, blasticidin S-deaminase *(bsd),* disrupting the lncRNA-ch14 region. The target specifying homology arm sequences were isolated through PCR amplification and gel purification. The right homology regions (RHR), and the left homology regions (LHR) of each gene were then ligated into the linearized vector via Gibson assembly. The final donor vectors were confirmed by restriction digests and Sanger sequencing. All primers are indicated in Supplementary data 1n.

Plasmids were isolated from 250 mL cultures of *Escherichia coli* (XL10-Gold Ultracompetent Cells, Agilent Cat. 200314) and 60 µg of each plasmid was used to transfect ring stage parasites. 24-hrs before transfection, mature parasite cultures (6-8% parasitemia) were magnetically separated using magnetic columns (MACS LD columns, Miltenyi Biotec) and diluted to 1% parasitemia containing 0.5 mL fresh erythrocytes [114]. The next day, ~3% ring stage parasites were pelleted and washed in 4 mL of cytomix [115]. 200 µL of the infected erythrocytes were resuspended with the two plasmids in cytomix to a total volume of 400 µL in a 0.2 cm cuvette. Electroporation was performed with a single pulse at 0.310 kV and 950 µF using the Biorad GenePulser electroporator. Cells were immediately transferred to a flask containing 12 mL media and 400 µL

erythrocytes. The media was exchanged five hours post electroporation with 12 mL of fresh media. The following day, fresh culture media was added and supplemented with 2.5 nM WR99210 and 2.5 μg/mL blasticidin (RPI Corp, B12150-0.1). Media and drug selection were replenished every 48 hours. After 14 days, the culture was split into two flasks and 50 μL of erythrocytes were added every two weeks. Once parasites were detected by microscopy, WR99210 was removed (selection for Cas9). Integration of the *bsd* gene was confirmed by gDNA extraction and PCR.

Isolation of △lncRNA-ch14 clone.

To generate genetically homogenous parasite lines, the transfected parasites were serially diluted to approximately 0.5%, into 96 well plates. 200 μL final volume of cultured parasites were incubated with bsd drug selection for 1 month with weekly erythrocytic and media changes for the first 2 weeks of dilution followed by media changes every 2 days until parasite recovery is observed through Giemsa-stained smears.

Verification of △lncRNA-ch14 line.

Genomic DNA (gDNA) was extracted and purified using DNeasy Blood & Tissue kit (Qiagen, 69504) following instructions from the manufacturer. The genotyping PCR analysis was used to genotype the KO lines using primer indicated in Supplementary data 1n. The PCR amplification was done using 2xKAPA master mix for thirty cycles with an annealing temperature of 50°C and an extension temperature of 62°C. The PCR amplicons were analyzed on a 1% agarose gel electrophoresis.

For whole genome sequencing, genomic DNAs were fragmented using a Covaris S220 ultrasonicator and libraries were generated using KAPA LTP Library Preparation Kit (Roche, KK8230). To verify that the insertion was present in the genome at the correct location in both transfected lines, reads were mapped using Bowtie2 (version 2.4.4) to the *P. falciparum* 3D7 reference genome (v48, PlasmoDB), edited to include the insertion sequence in the intended location. IGV (Broad Institute) was used to verify that reads aligned to the insertion sequence.

△lncRNA-ch14 line genome-wide sequencing and variant analysis.

Libraries were sequenced using a NovaSeq 6000 DNA sequencer (Illumina), producing paired-end 100-bp reads. To verify that the insertion was present in the genome at the correct location in both transfected lines, reads were mapped using Bowtie2 (version 2.4.4) to the *P. falciparum* 3D7 reference genome (v48, PlasmoDB), edited to include the insertion sequence in the intended location. IGV (Broad Institute) was used to verify that reads aligned to the insertion sequence. To call variants (SNPs/indels) in the transfected lines compared to a previously sequenced NF54 control line, genomic DNA reads were first trimmed of adapters and aligned to the *Homo sapiens* genome (assembly GRCh38) to remove human-mapped reads. Remaining reads were aligned to the *P. falciparum* 3D7 genome using bwa (version 0.7.17) and PCR duplicates were removed using PicardTools (Broad Institute). GATK HaplotypeCaller (https://gatk.broadinstitute.org/hc/en-us) was used to call variants between the sample and the 3D7 reference genome for both the transfected lines and the NF54 control. Only variants that were present in both transfected

lines but not the NF54 control line were kept. We examined only coding-region variants and removed those that were synonymous variants or were in *var*, *rifin*, or *stevor* genes. Quality control of variants was done by hard filtering using GATK guidelines.

Assessment of gametocyte development.

Viability of gametocytes was assessed via microscopy in parasite laboratory strains NF54 and two of the △lncRNA-ch14 clones, F2 and B1. The morphology of parasite gametocytes was assessed in a Giemsa-stained thin blood smear. Gametocytes were classified either as viable (normal intact morphology of mature gametocytes) or dead (deformed cells with a decrease in width, a thin needle-like appearance or degraded cytoplasmic content).

Gametocyte cultures and mosquito feeding.

This was performed as outlined previously 50]. Briefly, asexual stage cultures were grown in RPMI-1640 containing 2 mM L-glutamine, 50 mg/L hypoxanthine, 25 mM HEPES, 0.225% $NaHCO_3$, contained 10% *v/v* human serum in 4% human erythrocytes. Five mL of an asexual stage culture at 5% parasitaemia was centrifuged at 500 × *g* for 5 min at room temperature. Gametocyte cultures were initiated at 0.5% asynchronous asexual parasitemia from low passage stock and maintained up to day 18 with daily media changes but without any addition of fresh erythrocytes. The culture medium was changed daily for 15-18 days, by carefully aspirating ~ 70-80% of the supernatant medium to avoid removing cells, and 5 mL of fresh complete culture medium was added

to each well. Giemsa-stained blood smears were made every alternate day to confirm that the parasites remained viable. Instead of a gas incubator, cultures were maintained at 37°C in a candle jar made of glass desiccators. On day 15 to 18, gametocyte culture, containing largely mature gametocytes, were used for mosquito feeds. Cultures were transferred to pre- warmed tubes and centrifuged at 500 x *g* for 5 min. The cells were diluted in a pre-warmed 50:50 mixture of uninfected erythrocytes and normal human serum to achieve a mature gametocytemia of 0.2% and the resulting 'feeding mixture' was placed into a pre-warmed glass feeder. Uninfected *Anopheles stephensi* mosquitoes, starved overnight of sugar water, were allowed to feed on the culture for 30 min. Unfed mosquitoes were removed, and the mosquito cups were placed in a humidified 26ºC incubator, with 10% sugar-soaked cotton pads placed on top of the mosquito cage.

Oocyst and salivary gland sporozoite quantification.

On days 11 and 17 after the infective-blood meal, mosquitoes were dissected and midguts or salivary glands, respectively, were harvested for sporozoite counts. Day 11 midguts were stained with mercurochrome and photographed for oocyst counts by brightfield and phase microscopy using an upright Nikon E600 microscope with a PlanApo 4× objective. On day 17, salivary glands from ~20 mosquitoes were pooled, homogenized, and released sporozoites were counted using a haemocytometer.

Gametocyte quantification, sex determination and exflagellation assay.

Between days 15 to 18, blood smears were prepared from gametocyte cultures, fixed with methanol and stained with Giemsa (Sigma, GS500), prepared as a 1:5 dilution in buffer (pH=7.2) made using Gurr buffered tablets (VWR, 331942F) and filtered. Slides were stained for 20 min, washed with buffer, and allowed to dry before observation using a Nikon E600 microscope with a PlanApo 100× oil objective. For calculation of gametocytemias and male: female ratios, at least 500 mature gametocytes were scored per slide. To count exflagellation centers, 500 µL of mature gametocyte culture was centrifuged at 500g for four min and the resulting pellet was resuspended in equal volume of prewarmed normal human serum. Temperature was dropped to room temperature to activate gametogenesis and after 15 min incubation 10µl of culture was transferred to a glass slide and covered with a cover slip. Exflagellation centers were counted at 40x objective in at-least ten fields for each gametocyte culture. To avoid bias, microscopic examination was performed in a blinded fashion by a trained reader.

△lncRNA-ch14 transcriptome analysis.

Libraries were prepared from the extracted total RNA, first by isolating mRNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490), then using the NEBNext Ultra Directional RNA Library Prep Kit (NEB, E7420). 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 using a NovaSeq 6000 DNA sequencer (Illumina), producing paired-end 100-bp reads.

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), was used to assess raw read quality and characteristics, and based on this information, the first 11 bp of each read and any adapter sequences were removed using Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Bases were trimmed from reads using Sickle with a Phred quality threshold of 20 (https://github.com/najoshi/sickle). These reads were mapped against the *Homo sapiens* genome (assembly GRCh38) using Bowtie2 (version 2.4.4) and mapped reads were removed. The remaining reads were mapped against the *Plasmodium falciparum* 3D7 genome (v48, PlasmoDB) using HISAT2 (version 2.2.1), using default parameters. Uniquely mapped, properly paired reads with mapping quality 40 or higher were retained using SAMtools (http://samtools.sourceforge.net/), and PCR duplicates were removed using PicardTools (Broad Institute). Genome browser tracks were generated and viewed using the Integrative Genomic Viewer (IGV) (Broad Institute).

Raw read counts were determined for each gene in the *P. falciparum* genome using BedTools (https://bedtools.readthedocs.io/en/latest/#).

to intersect the aligned reads with the genome annotation. Differential expression analysis was done by use of R package DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) to call up- and down-regulated genes with an adjusted P-value cutoff of 0.05. Gene ontology enrichment was done using PlasmoDB (https://plasmodb.org/plasmo/app). Volcano plots were generated

using R package EnhancedVolcano (https://bioconductor.org/packages/release/bioc/html/EnhancedVolcano.html).

Statistical analysis

Descriptive statistics were calculated with GraphPad Prism version 9.1.2 (GraphPad Software, San Diego, CA, USA) for determining mean, percentages, standard deviation and plotting of graphs. Excel 2013 and GraphPad Prism 9.1.2 were used for the calculation of gametocytemia of microscopic data.

Data Availability

The GRO-seq data used in this study are available in the Gene Expression Omnibus database under accession code GSE85478 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85478). The steady-state RNA-seq data used in this study are available the NCBI Sequence Read Archive under accession code SRP026367, SRS417027, SRS417268, and SRS417269 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP026367, https://www.ncbi.nlm.nih.gov/sra/?term=SRS417027, https://www.ncbi.nlm.nih.gov/sra/?term=SRS417268, https://www.ncbi.nlm.nih.gov/sra/?term=SRS417269).

WGS, ChIRP-seq and RNA-seq data generated in this study (66 libraries) have been deposited in the NCBI Sequence Read Archive with accession PRJNA869073 and are available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA869073/. All other data

generated in this study are provided in the Supplementary Information, Supplementary data and Source Data file.

References

1 WHO. *World malaria report 2021*, <https://www.who.int/publications/i/item/9789240040496> (2021).

2 Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-1508, doi:10.1126/science.1087025 (2003).

3 Bozdech, Z. *et al.* Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol* **4**, R9, doi:10.1186/gb-2003-4-2-r9 (2003).

4 Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol* **1**, E5, doi:10.1371/journal.pbio.0000005 (2003).

5 Young, J. A. *et al.* The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* **143**, 67-79, doi:10.1016/j.molbiopara.2005.05.007 (2005).

6 Silvestrini, F. *et al.* Genome-wide identification of genes upregulated at the onset of gametocytogenesis in Plasmodium falciparum. *Mol Biochem Parasitol* **143**, 100-110, doi:10.1016/j.molbiopara.2005.04.015 (2005).

7 Lemieux, J. E. *et al.* Statistical estimation of cell-cycle progression and lineage commitment in Plasmodium falciparum reveals a homogeneous pattern of transcription in ex vivo culture. *Proc Natl Acad Sci U S A* **106**, 7559-7564, doi:10.1073/pnas.0811829106 (2009).

8 Lu, X. M. *et al.* Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite Plasmodium falciparum. *Nucleic Acids Res* **45**, 7825-7840, doi:10.1093/nar/gkx464 (2017).

9 Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, doi:10.7554/eLife.33105 (2018).

10 Balaji, S., Babu, M. M., Iyer, L. M. & Aravind, L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res* **33**, 3994-4006, doi:10.1093/nar/gki709 (2005).

11 Shang, X. *et al.* Genome-wide landscape of ApiAP2 transcription factors reveals a heterochromatin-associated regulatory network during Plasmodium falciparum blood-stage development. *Nucleic Acids Res* **50**, 3413-3431, doi:10.1093/nar/gkac176 (2022).

12 De Silva, E. K. *et al.* Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci U S A* **105**, 8393-8398, doi:10.1073/pnas.0801993105 (2008).

13 Gomez-Diaz, E. *et al.* Epigenetic regulation of Plasmodium falciparum clonally variant gene expression during development in Anopheles gambiae. *Sci Rep* **7**, 40655, doi:10.1038/srep40655 (2017).

14 Gupta, A. P. *et al.* Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum. *PLoS Pathog* **9**, e1003170, doi:10.1371/journal.ppat.1003170 (2013).

15 Bunnik, E. M. *et al.* The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum. *Genome Biol* **17**, 147, doi:10.1186/s13059-016-1014-0 (2016).

16 Lacsina, J. R., LaMonte, G., Nicchitta, C. V. & Chi, J. T. Polysome profiling of the malaria parasite Plasmodium falciparum. *Mol Biochem Parasitol* **179**, 42-46, doi:10.1016/j.molbiopara.2011.05.003 (2011).

17 Mair, G. R. *et al.* Regulation of sexual development of Plasmodium by translational repression. *Science* **313**, 667-669, doi:10.1126/science.1125129 (2006).

18 Shock, J. L., Fischer, K. F. & DeRisi, J. L. 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* **8**, R134, doi:10.1186/gb-2007-8-7-r134 (2007).

19 Freitas-Junior, L. H. *et al.* Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* **121**, 25-36, doi:10.1016/j.cell.2005.01.037 (2005).

20 Petter, M. *et al.* 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* **87**, 1167-1182, doi:10.1111/mmi.12154 (2013).

21 Lopez-Rubio, J. J. *et al.* 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol Microbiol* **66**, 1296-1305, doi:10.1111/j.1365-2958.2007.06009.x (2007).

22 Lopez-Rubio, J. J., Mancio-Silva, L. & Scherf, A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* **5**, 179-190, doi:10.1016/j.chom.2008.12.012 (2009).

23 Tonkin, C. J. *et al.* Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum. *PLoS Biol* **7**, e84, doi:10.1371/journal.pbio.1000084 (2009).

24 Ukaegbu, U. E. *et al.* Recruitment of PfSET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in P. falciparum. *PLoS Pathog* **10**, e1003854, doi:10.1371/journal.ppat.1003854 (2014).

25 Saxena, H. & Gupta, A. Plasmodium falciparum PfRUVBL proteins bind at the TARE region and var gene promoter located in the subtelomeric region. *Pathog Dis* **80**, doi:10.1093/femspd/ftac018 (2022).

26 Lemieux, J. E. *et al.* Genome-wide profiling of chromosome interactions in Plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. *Mol Microbiol* **90**, 519-537, doi:10.1111/mmi.12381 (2013).

27 Ay, F. *et al.* Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* **24**, 974-988, doi:10.1101/gr.169417.113 (2014).

28 Bunnik, E. M. *et al.* Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. *Nat Commun* **9**, 1910, doi:10.1038/s41467-018-04295-5 (2018).

29 Read, D. F., Cook, K., Lu, Y. Y., Le Roch, K. G. & Noble, W. S. Predicting gene expression in the human malaria parasite Plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features. *PLoS Comput Biol* **15**, e1007329, doi:10.1371/journal.pcbi.1007329 (2019).

30 Abdi, E. & Latifi-Navid, S. Emerging long noncoding RNA polymorphisms as novel predictors of survival in cancer. *Pathol Res Pract* **239**, 154165, doi:10.1016/j.prp.2022.154165 (2022).

31 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

32 Corcoran, A. E. The epigenetic role of non-coding RNA transcription and nuclear organization in immunoglobulin repertoire generation. *Semin Immunol* **22**, 353-361, doi:10.1016/j.smim.2010.08.001 (2010).

33 Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076, doi:10.1038/nature08975 (2010).

34 Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* **22**, 96-118, doi:10.1038/s41580-020-00315-9 (2021).

35 Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biol* **10**, 925-933, doi:10.4161/rna.24604 (2013).

36 Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol* **19**, 143-157, doi:10.1038/nrm.2017.104 (2018).

37 Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat Rev Mol Cell Biol* **17**, 756-770, doi:10.1038/nrm.2016.126 (2016).

38 Quinodoz, S. & Guttman, M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends Cell Biol* **24**, 651-663, doi:10.1016/j.tcb.2014.08.009 (2014).

39 Nakagawa, S. & Kageyama, Y. Nuclear lncRNAs as epigenetic regulators-beyond skepticism. *Biochim Biophys Acta* **1839**, 215-222, doi:10.1016/j.bbagrm.2013.10.009 (2014).

40 Maclary, E., Hinten, M., Harris, C. & Kalantry, S. Long nonoding RNAs in the X-inactivation center. *Chromosome Res* **21**, 601-614, doi:10.1007/s10577-013-9396-2 (2013).

41 Schoeftner, S. & Blasco, M. A. Chromatin regulation and non-coding RNAs at mammalian telomeres. *Semin Cell Dev Biol* **21**, 186-193, doi:10.1016/j.semcdb.2009.09.015 (2010).

42 Feuerhahn, S., Iglesias, N., Panza, A., Porro, A. & Lingner, J. TERRA biogenesis, turnover and implications for function. *FEBS Lett* **584**, 3812-3818, doi:10.1016/j.febslet.2010.07.032 (2010).

43 Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M. & Wirth, D. F. Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Mol Biol Cell* **12**, 3114-3125, doi:10.1091/mbc.12.10.3114 (2001).

44 Gunasekera, A. M. *et al.* Widespread distribution of antisense transcripts in the Plasmodium falciparum genome. *Mol Biochem Parasitol* **136**, 35-42, doi:10.1016/j.molbiopara.2004.02.007 (2004).

45 Raabe, C. A. *et al.* A global view of the nonprotein-coding transcriptome in Plasmodium falciparum. *Nucleic Acids Res* **38**, 608-617, doi:10.1093/nar/gkp895 (2010).

46 Sorber, K., Dimon, M. T. & DeRisi, J. L. RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* **39**, 3820-3835, doi:10.1093/nar/gkq1223 (2011).

47 Broadbent, K. M. *et al.* A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol* **12**, R56, doi:10.1186/gb-2011-12-6-r56 (2011).

48 Wei, C. *et al.* Deep profiling of the novel intermediate-size noncoding RNAs in intraerythrocytic Plasmodium falciparum. *PLoS One* **9**, e92946, doi:10.1371/journal.pone.0092946 (2014).

49 Liao, Q. *et al.* Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data. *Parasitol Res* **113**, 1269-1281, doi:10.1007/s00436-014-3765-4 (2014).

50 Siegel, T. N. *et al.* Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC Genomics* **15**, 150, doi:10.1186/1471-2164-15-150 (2014).

51 Broadbent, K. M. *et al.* Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* **16**, 454, doi:10.1186/s12864-015-1603-4 (2015).

52 Lee, V. V. *et al.* Direct Nanopore Sequencing of mRNA Reveals Landscape of Transcript Isoforms in Apicomplexan Parasites. *mSystems* **6**, doi:10.1128/mSystems.01081-20 (2021).

53 Chappell, L. *et al.* Refining the transcriptome of the human malaria parasite Plasmodium falciparum using amplification-free RNA-seq. *BMC Genomics* **21**, 395, doi:10.1186/s12864-020-06787-5 (2020).

54 Broadbent, K. M. *et al.* Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* **16**, 454, doi:10.1186/s12864-015-1603-4 (2015).

55 Mourier, T. *et al.* Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. *Genome Res* **18**, 281-292, doi:10.1101/gr.6836108 (2008).

56 Yang, M. *et al.* Full-Length Transcriptome Analysis of Plasmodium falciparum by Single-Molecule Long-Read Sequencing. *Front Cell Infect Microbiol* **11**, 631545, doi:10.3389/fcimb.2021.631545 (2021).

57 Hoshizaki, J. *et al.* Correction: A manually curated annotation characterises genomic features of P. falciparum lncRNAs. *BMC Genomics* **24**, 189, doi:10.1186/s12864-023-09164-0 (2023).

58 Epp, C., Li, F., Howitt, C. A., Chookajorn, T. & Deitsch, K. W. Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. *RNA* **15**, 116-127, doi:10.1261/rna.1080109 (2009).

59 Amit-Avraham, I. *et al.* Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum. *Proc Natl Acad Sci U S A* **112**, E982-991, doi:10.1073/pnas.1420855112 (2015).

60 Sierra-Miranda, M. *et al.* Two long non-coding RNAs generated from subtelomeric regions accumulate in a novel perinuclear compartment in Plasmodium falciparum. *Mol Biochem Parasitol* **185**, 36-47, doi:10.1016/j.molbiopara.2012.06.005 (2012).

61 Simantov, K., Goyal, M. & Dzikowski, R. Emerging biology of noncoding RNAs in malaria parasites. *PLoS Pathog* **18**, e1010600, doi:10.1371/journal.ppat.1010600 (2022).

62 Rovira-Graells, N. *et al.* Deciphering the principles that govern mutually exclusive expression of Plasmodium falciparum clag3 genes. *Nucleic Acids Res* **43**, 8243-8257, doi:10.1093/nar/gkv730 (2015).

63 Jing, Q. *et al.* Plasmodium falciparum var Gene Is Activated by Its Antisense Long Noncoding RNA. *Front Microbiol* **9**, 3117, doi:10.3389/fmicb.2018.03117 (2018).

64 Barcons-Simon, A., Cordon-Obras, C., Guizetti, J., Bryant, J. M. & Scherf, A. CRISPR Interference of a Clonally Variant GC-Rich Noncoding RNA Family Leads to General Repression of var Genes in Plasmodium falciparum. *mBio* **11**, doi:10.1128/mBio.03054-19 (2020).

65 Luke, B. & Lingner, J. TERRA: telomeric repeat-containing RNA. *EMBO J* **28**, 2503-2510, doi:10.1038/emboj.2009.166 (2009).

66 Filarsky, M. *et al.* GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* **359**, 1259-1263, doi:10.1126/science.aan6042 (2018).

67 Gomes, A. R. *et al.* A transcriptional switch controls sex determination in Plasmodium falciparum. *Nature* **612**, 528-533, doi:10.1038/s41586-022-05509-z (2022).

68 Bunnik, E. M. *et al.* DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics* **15**, 347, doi:10.1186/1471-2164-15-347 (2014).

69 Zhang, M. *et al.* Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis. *Science* **360**, doi:10.1126/science.aap7847 (2018).

70 Sun, Q., Hao, Q. & Prasanth, K. V. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends Genet* **34**, 142-157, doi:10.1016/j.tig.2017.11.005 (2018).

71 Noh, J. H., Kim, K. M., McClusky, W. G., Abdelmohsen, K. & Gorospe, M. Cytoplasmic functions of long noncoding RNAs. *Wiley Interdiscip Rev RNA* **9**, e1471, doi:10.1002/wrna.1471 (2018).

72 Sun, M., Gadad, S. S., Kim, D. S. & Kraus, W. L. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol Cell* **59**, 698-711, doi:10.1016/j.molcel.2015.06.023 (2015).

73 Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res* **22**, 885-898, doi:10.1101/gr.131037.111 (2012).

74 Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**, 667-678, doi:10.1016/j.molcel.2011.08.027 (2011).

75 Quinn, J. J. *et al.* Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat Biotechnol* **32**, 933-940, doi:10.1038/nbt.2943 (2014).

76 Chu, C., Quinn, J. & Chang, H. Y. Chromatin isolation by RNA purification (ChIRP). *J Vis Exp*, doi:10.3791/3912 (2012).

77 Fan, Y. *et al.* Rrp6 Regulates Heterochromatic Gene Silencing via ncRNA RUF6 Decay in Malaria Parasites. *mBio* **11**, doi:10.1128/mBio.01110-20 (2020).

78 Bunnik, E. M. *et al.* Comparative 3D genome organization in apicomplexan parasites. *Proc Natl Acad Sci U S A* **116**, 3183-3192, doi:10.1073/pnas.1810815116 (2019).

79 Winter, G. *et al.* SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *J Exp Med* **201**, 1853-1863, doi:10.1084/jem.20041392 (2005).

80 Pei, X. *et al.* The ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum stabilizes spectrin tetramers and suppresses further invasion. *Blood* **110**, 1036-1042, doi:10.1182/blood-2007-02-076919 (2007).

81 Zhang, Y., Lin, Y. H., Johnson, T. D., Rozek, L. S. & Sartor, M. A. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics* **30**, 2568-2575, doi:10.1093/bioinformatics/btu372 (2014).

82 Silvestrini, F. *et al.* Protein export marks the early phase of gametocytogenesis of the human malaria parasite Plasmodium falciparum. *Mol Cell Proteomics* **9**, 1437-1448, doi:10.1074/mcp.M900479-MCP200 (2010).

83 Gardiner, D. L. *et al.* Implication of a Plasmodium falciparum gene in the switch between asexual reproduction and gametocytogenesis. *Mol Biochem Parasitol* **140**, 153-160, doi:10.1016/j.molbiopara.2004.12.010 (2005).

84 Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of P. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res* **44**, 6087-6101, doi:10.1093/nar/gkw536 (2016).

85 Tiburcio, M. *et al.* A 39-Amino-Acid C-Terminal Truncation of GDV1 Disrupts Sexual Commitment in Plasmodium falciparum. *mSphere* **6**, doi:10.1128/mSphere.01093-20 (2021).

86 Usui, M. *et al.* Plasmodium falciparum sexual differentiation in malaria patients is associated with host factors and GDV1-dependent genes. *Nat Commun* **10**, 2140, doi:10.1038/s41467-019-10172-6 (2019).

87 Tripathi, A. K., Mlambo, G., Kanatani, S., Sinnis, P. & Dimopoulos, G. Plasmodium falciparum Gametocyte Culture and Mosquito Infection Through Artificial Membrane Feeding. *J Vis Exp*, doi:10.3791/61426 (2020).

88 Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat Struct Mol Biol* **19**, 1068-1075, doi:10.1038/nsmb.2428 (2012).

89 Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58, doi:10.1016/j.cell.2010.09.001 (2010).

90 Ng, S. Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol Cell* **51**, 349-359, doi:10.1016/j.molcel.2013.07.017 (2013).

91 Prensner, J. R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* **45**, 1392-1398, doi:10.1038/ng.2771 (2013).

92 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693, doi:10.1126/science.1192002 (2010).

93 Mele, M. & Rinn, J. L. "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. *Mol Cell* **62**, 657-664, doi:10.1016/j.molcel.2016.05.011 (2016).

94 Rinn, J. & Guttman, M. RNA Function. RNA and dynamic nuclear organization. *Science* **345**, 1240-1241, doi:10.1126/science.1252966 (2014).

95 Guizetti, J., Barcons-Simon, A. & Scherf, A. Trans-acting GC-rich non-coding RNA at var expression site modulates gene counting in malaria parasite. *Nucleic Acids Res* **44**, 9710-9718, doi:10.1093/nar/gkw664 (2016).

96 Abel, S. & Le Roch, K. G. The role of epigenetics and chromatin structure in transcriptional regulation in malaria parasites. *Brief Funct Genomics* **18**, 302-313, doi:10.1093/bfgp/elz005 (2019).

97 Cerase, A., Pintacuda, G., Tattermusch, A. & Avner, P. Xist localization and function: new insights from multiple levels. *Genome Biol* **16**, 166, doi:10.1186/s13059-015-0733-y (2015).

98 Hanly, D. J., Esteller, M. & Berdasco, M. Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer? *Philos Trans R Soc Lond B Biol Sci* **373**, doi:10.1098/rstb.2017.0074 (2018).

99 Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet* **30**, 348-355, doi:10.1016/j.tig.2014.06.001 (2014).

100 Wierzbicki, A. T. The role of long non-coding RNA in transcriptional gene silencing. *Curr Opin Plant Biol* **15**, 517-522, doi:10.1016/j.pbi.2012.08.008 (2012).

101 Reininger, L. *et al.* A NIMA-related protein kinase is essential for completion of the sexual cycle of malaria parasites. *J Biol Chem* **280**, 31957-31964, doi:10.1074/jbc.M504523200 (2005).

102 Reininger, L. *et al.* An essential role for the Plasmodium Nek-2 Nima-related protein kinase in the sexual development of malaria parasites. *J Biol Chem* **284**, 20858-20868, doi:10.1074/jbc.M109.017988 (2009).

103 Tomas, A. M. *et al.* P25 and P28 proteins of the malaria ookinete surface have multiple and partially redundant functions. *EMBO J* **20**, 3975-3983, doi:10.1093/emboj/20.15.3975 (2001).

104 Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193**, 673-675, doi:10.1126/science.781840 (1976).

105 Joshi NA, F. J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. *[Software].* **(Version 1.33)** (2011).

106 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

107 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

108 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

109 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

110 Howick, V. M. *et al.* The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **365**, doi:10.1126/science.aaw2619 (2019).

111 Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* **499**, 223-227, doi:10.1038/nature12361 (2013).

112 Sonoiki, E. *et al.* A potent antimalarial benzoxaborole targets a Plasmodium falciparum cleavage and polyadenylation specificity factor homologue. *Nat Commun* **8**, 14574, doi:10.1038/ncomms14574 (2017).

113 Ribeiro, J. M. *et al.* Guide RNA selection for CRISPR-Cas9 transfections in Plasmodium falciparum. *Int J Parasitol* **48**, 825-832, doi:10.1016/j.ijpara.2018.03.009 (2018).

114 Ribaut, C. *et al.* Concentration and purification by magnetic separation of the erythrocytic stages of all human Plasmodium species. *Malar J* **7**, 45, doi:10.1186/1475-2875-7-45 (2008).

115 Adjalley, S. H., Lee, M. C. & Fidock, D. A. A method for rapid genetic integration into Plasmodium falciparum utilizing mycobacteriophage Bxb1 integrase. *Methods Mol Biol* **634**, 87-100, doi:10.1007/978-1-60761-652-8_6 (2010).

Supplementary Information

Supplementary material for chapter 2 is available with the published paper at https://www.nature.com/articles/s41467-023-40883-w.

Chapter 3: Plasmodium Condensin Core Subunits SMC2/SMC4 Mediate Atypical Mitosis and Are Essential for Parasite Proliferation and Transmission

Rajan Pandey [1], Steven Abel [2], Matthew Boucher [1], Richard J Wall [3], Mohammad Zeeshan [1], Edward Rea [1], Aline Freville [1], Xueqing Maggie Lu [2], Declan Brady [1], Emilie Daniel [1], Rebecca R Stanway [4], Sally Wheatley [1], Gayani Batugedara [2], Thomas Hollin [2], Andrew R Bottrill [5], Dinesh Gupta [6], Anthony A Holder [7], Karine G Le Roch [8], Rita Tewari [9]

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

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

[3] Wellcome Trust Centre for Anti-Infectives Research, School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK.

[4] Institute of Cell Biology, University of Bern, Bern 3012, Switzerland.

[5] School of Life Sciences, Gibbet Hill Campus, University of Warwick, Coventry CV4 7AL, UK.

[6] Translational Bioinformatics Group, International Center for Genetic Engineering and Biotechnology, New Delhi 110067, India.

[7] Malaria Parasitology Laboratory, The Francis Crick Institute, London NW1 1AT, UK.

[8] Department of Molecular, Cell and Systems Biology, University of California Riverside, 900 University Ave., Riverside, CA 92521, USA. Electronic address: karine.leroch@ucr.edu.

[9] School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK. Electronic address: rita.tewari@nottingham.ac.uk.

A version of this chapter has been published in *Cell Reports*, 2020.

Preface

The search for drug targets against malaria parasites includes not only gene regulators but proteins involved in cellular processes like cell division. Disruption of these processes in the parasite without interfering with human proteins could lead to parasite clearance in sick individuals. *Plasmodium berghei*, a mouse malaria parasite, is often used as a model for *Plasmodium falciparum*. In this study we used *P. berghei* to characterize the importance of condensin complex in *Plasmodium* spp.. In model eukaryotes, the condensin complex composed of SMC2 and SMC4 is instrumental in chromosome condensation and segregation during cell division. Our work showed that the parasite homologs of these proteins play a role in the atypical Plasmodium cell division process. To examine these proteins at the functional level, we used a combination of bioinformatics approaches as well as a variety of experimental techniques including ChIP-seq, live cell imaging, and RNA-seq and phenotypic analysis of protein knockouts. I performed not only all ChIP-seq experiments at the bench but also the computational analysis of ChIP-seq and RNA-seq experiments presented in this project. In analyzing the ChIP-seq data for SMC2, SMC4, and NDC80 (a centromeric marker), I found that SMC2 and SMC4 bind at the centromeres and that the true centromeres for most chromosomes were noticeably shifted from the previously annotated centromeres for *P. berghei*, which have now been corrected. In analyzing the RNA-seq data for the SMC4 knockout, I found that microtubule-related genes were significantly enriched among downregulated genes, pointing to the function of this protein and its partner in chromosome segregation and cell division, and some of this downregulation was confirmed by qPCR. Phenotypic analysis

of SMC2 and SMC4 knockouts confirmed that cell division processes crucial for life cycle progression and proliferation of parasites within mosquitoes are impaired due to the loss of these proteins, demonstrating their importance to parasite biology.

Abstract

Condensin is a multi-subunit protein complex regulating chromosome condensation and segregation during cell division. In Plasmodium spp., the causative agent of malaria, cell division is atypical and the role of condensin is unclear. Here we examine the role of SMC2 and SMC4, the core subunits of condensin, during endomitosis in schizogony and endoreduplication in male gametogenesis. During early schizogony, SMC2/SMC4 localize to a distinct focus, identified as the centromeres by NDC80 fluorescence and chromatin immunoprecipitation sequencing (ChIP-seq) analyses, but do not form condensin I or II complexes. In mature schizonts and during male gametogenesis, there is a diffuse SMC2/SMC4 distribution on chromosomes and in the nucleus, and both condensin I and condensin II complexes form at these stages. Knockdown of smc2 and smc4 gene expression reveals essential roles in parasite proliferation and transmission. The condensin core subunits (SMC2/SMC4) form different complexes and may have distinct functions at various stages of the parasite life cycle.

Introduction

Cellular proliferation in eukaryotes requires chromosome replication and segregation, followed by cell division, to ensure that daughter cells have identical copies of the genome. During classical open mitosis in many eukaryotes, chromosome condensation, centrosome migration, and formation of the mitotic spindle are followed by dissolution of the nuclear envelope (Güttinger et al., 2009). In contrast, in some unicellular organisms such as the budding yeast *Saccharomyces cerevisiae*, mitosis is closed: the nuclear membrane remains intact, and chromosomes are separated by spindles assembled within the nucleus (Sazer et al., 2014). The mechanisms and the various regulatory molecules involved in cell division have been well studied in many eukaryotes. The cell division regulators include cyclins, cyclin-dependent kinases (CDKs), components of the anaphase-promoting complex (APC), and other protein kinases and phosphatases (Chang et al., 2014, Fisher et al., 2012, Harashima et al., 2013).

An essential component of chromosome dynamics is a family of structural maintenance of chromosomes proteins, originally described in budding yeast as stability of minichromosomes (SMC) proteins, which are implicated in chromosome segregation and condensation (Hirano, 2016, Uhlmann, 2016). Most eukaryotes have at least six genes encoding SMC proteins (each 110–170 kDa, with a central hinge region and N- and C-terminal globular domains with Walker A and Walker B motifs forming the ATPase head domain). The six SMCs can be classified as subunits of condensin (SMC2 and SMC4, required for chromosomal condensation), cohesin (SMC1 and SMC3, required for

chromosomal segregation), and the SMC5-SMC6 complex (involved in DNA repair and homologous recombination) (Hirano, 2016, Uhlmann, 2016).

Higher eukaryotic organisms have two condensin complexes, condensin I and condensin II, whereas many single-celled organisms such as yeast have only one condensin complex. SMC2 and SMC4 form the core structure for both condensin I and condensin II in higher eukaryotes (Hirano, 2016) and interact with three additional non-SMC components: one kleisin (Schleiffer et al., 2003) and two Heat protein subunits (Neuwald and Hirano, 2000). Kleisin Iγ (CAP-H), Heat IA (CAP-D2), and Heat IB (CAP-G) form the condensin I complex, whereas Kleisin IIβ (CAP-H2), Heat IIA (CAP-D3), and Heat IIB (CAP-G2) form the condensin II complex (Hirano, 2016, Uhlmann, 2016; Figure 3.1A). Electron microscopy and protein-protein interaction studies have revealed the characteristic architecture and geometry of condensin complexes (Anderson et al., 2002, Onn et al., 2007). Condensin plays a vital role in cell division processes such as chromosomal condensation, correct folding and organization of chromosomes before anaphase, and proper chromosome segregation and separation (Hirano, 2016, Kschonsak et al., 2017, Ono et al., 2013, Rawlings et al., 2011, Uhlmann, 2016). Both SMC and non-SMC components are necessary for full function; for example, chromosomal condensation is not observed in the absence of kleisin, showing its critical role for complex formation and condensation (Cuylen et al., 2011, Rawlings et al., 2011).

Plasmodium, the apicomplexan parasite that causes malaria, undergoes two types of atypical mitotic division during its life cycle: one in the asexual stages (schizogony in the

liver and blood stages within the vertebrate host, and sporogony in the mosquito gut) and the other in male gametogenesis (endoreduplication) during the sexual stage (Arnot et al., 2011, Sinden, 1991b). Division during schizogony/sporogony resembles closed endomitosis with repeated asynchronous nuclear divisions, followed by a final synchronized set of nuclear division forming a multinucleated syncytium before cytokinesis. An intact nuclear envelope is maintained, wherein the microtubule organizing center (MTOC), known as the centriolar plaque or spindle pole body (SPB), is embedded, and rounds of mitosis and nuclear division proceed without chromosome condensation (Arnot et al., 2011, Francia and Striepen, 2014, Gerald et al., 2011, Sinden, 1991a, Sinden, 1991b, Sinden et al., 1976). In male gametogenesis, exposure of the male gametocyte to the mosquito midgut environment leads to activation of mitosis, which results in three rounds of rapid chromosome replication (endoreduplication) within 8–10 min and atypical chromosomal condensation, followed by nuclear and cell division to produce eight motile male gametes (exflagellation) (Guttery et al., 2012b, Sinden, 1991b, Sinden et al., 1976, Sinden et al., 2010). During exflagellation, each condensed haploid nucleus and its associated MTOC, together with a basal body, axoneme, and flagellum, form the microgamete that egresses from the main cellular body (Guttery et al., 2012b, Sinden, 1991b, Sinden et al., 1976, Sinden et al., 2010).

The atypical cell division and proliferation of malaria parasites is controlled by, among others, unique and divergent Apicomplexa-specific CDKs, aurora-like kinases (ARKs), mitotic protein phosphatase 1 (PP1), and only four APC components (Guttery et al., 2014, Roques et al., 2015, Wall et al., 2018, Ward et al., 2004, Wilkes and Doerig, 2008).

However, there are no known classical group 1 cyclins, polo-like kinases (that are major regulators in mitotic entry), or classical mitotic protein phosphatases (CDC14 and CDC25) encoded in the genome (Guttery et al., 2014, Tewari et al., 2010, Ward et al., 2004, Wilkes and Doerig, 2008).

In *Plasmodium*, the role of condensin during cell division and general chromosome dynamics is unknown. Here, we investigated the location and function of the core subunits of condensin (SMC2 and SMC4) during two mitotic division stages in the *Plasmodium* life cycle: during schizogony in the host's blood and during male gametogenesis in the mosquito vector. This study was performed using the rodent malaria model *Plasmodium berghei*. For this analysis, we used a combination of cell biology, proteomics, transcriptomics, chromatin immunoprecipitation, and reverse genetics approaches. Spatiotemporal localization using live cell imaging indicates a dynamic profile for both SMC2 and SMC4, with either discrete foci during early schizogony or more diffuse nuclear localization during late schizogony and male gametogenesis. Genome-wide distribution studies using chromatin immunoprecipitation sequencing (ChIP-seq) experiments suggested that both components (SMC2/SMC4) are located at or near the centromeres during early schizogony, but this strong interaction was not observed during gametogenesis. Interestingly, we identified a differential composition of the condensin complex between the distinct mitotic stages, suggesting divergent mechanisms at the molecular level. Our data demonstrate that the condensin core subunits (SMC2/SMC4) have distinct functions at different stages of the parasite life

cycle. Functional analyses using a conditional gene knockdown approach indicate that condensins are required for parasite proliferation and transmission.

Results

Bioinformatic Analysis Shows SMC2/SMC4, the Core Subunits of Condensin, Are Encoded in the *Plasmodium* Genome

To identify condensin in *Plasmodium*, we screened for the core subunit genes in the *P. berghei* genome using PlasmoDB version 42, revealing both core SMC components of condensin, SMC2 and SMC4 (Bahl et al., 2002). Domain analysis revealed a conserved domain architecture for both SMC2 and SMC4 (Figure 3.1B). A comparative sequence analysis revealed low sequence similarity and identity (~29%–34%), except for the SMC4 homolog in *Arabidopsis thaliana* (65%) (Figure 3.1C), although there was similarity in size and overall domain structure when compared with the proteins in the other studied organisms. We found the *P. berghei* SMC4 N-terminal ATPase domain divided in two by a 44 amino acid insertion; a similar pattern has been observed in other *Plasmodium* species. Subsequently, we generated a 3D model of the *P. berghei* SMC2 and SMC4 ATPase head domains and partial coiled region using homology-based 3D structure modeling (Figure 3.1D). Root-mean-square deviation (RMSD) analysis of the 10 ns molecular dynamics (MD) simulation trajectory of the proteins showed a stable conformation comparable to pre-simulation energy-minimized structures. Radius of gyration analysis also confirmed a stable conformation for the predicted SMC2 and SMC4 domain structures during the 10 ns MD simulation (Figure S1). In this model of

the SMC subunits, the N- and C-terminal ATP-binding cassette (ABC) ATPase head and coiled-coil arms connecting the hinge domain (Figure 3.1D) are present, as in other organisms. It is most likely that the heads of *Plasmodium* SMC2 and SMC4 undergo ATP-dependent engagement and disengagement, and they may perform chromosomal functions similar to those in other eukaryotes (Hirano, 2016).

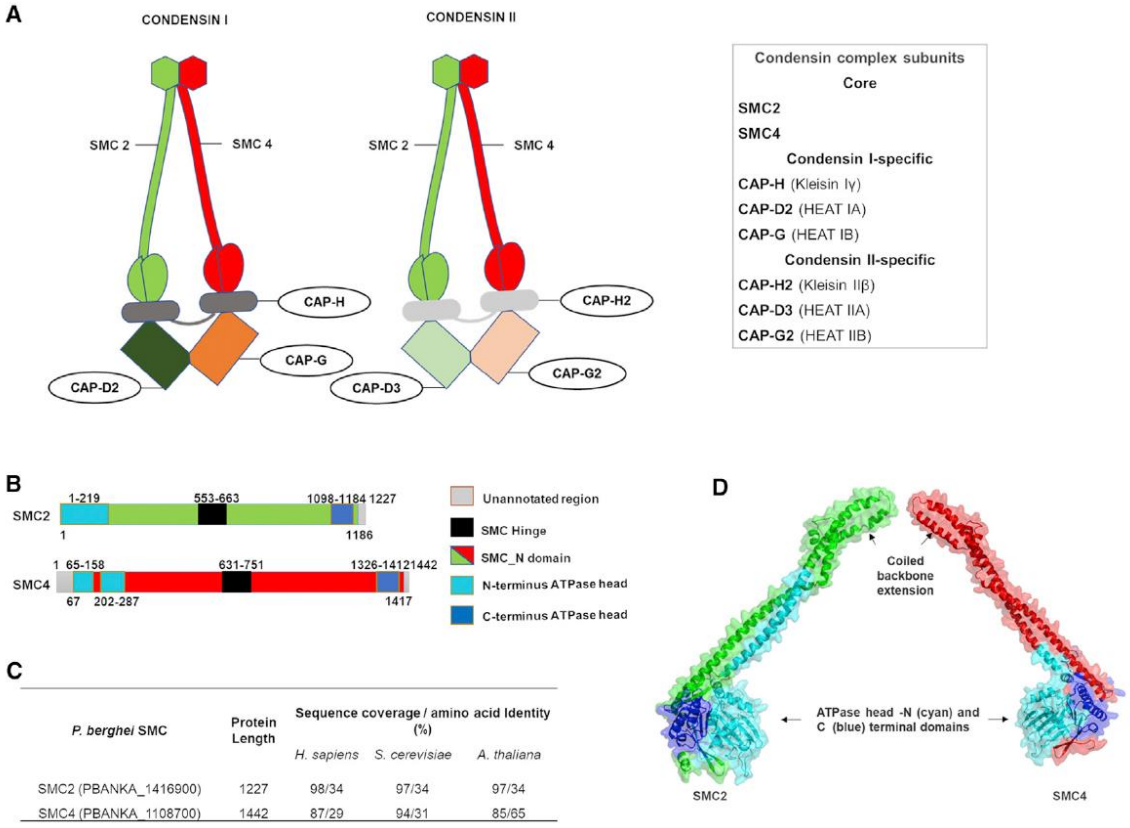

Fig. 3.1 Architecture of Condensin (SMC2/SMC4) in *Plasmodium berghei.* (A) Composition of two conventional condensin complexes (condensin I and condensin II), which are composed of heterodimeric core subunits, SMC2 and SMC4, along with non-SMC regulatory subunits: Kleisin and a pair of Heat subunits specific for either condensin I or II. CAP-H, CAP-D2, and CAP-G form the condensin I complex, whereas CAP-H2, CAP-D3, and CAP-G2 form the condensin II complex (modified from Hirano, 2016, and Uhlmann, 2016). (B) Domain architecture of *P. berghei* SMC2 and SMC4. (C) Sequence coverage and amino acid identity of *P. berghei* SMC2 and SMC4 with *H. sapiens*, *S. cerevisiae*, and *A. thaliana* proteins. (D) Homology-based predicted three-dimensional structures of *P. berghei* SMC2 and SMC4 showing coiled backbone extension without the hinge domain but with ATPase head formation, required for condensin complex.

Condensin Core Subunits Are Expressed at Every Proliferative Stage of the Parasite Life Cycle and Have a Centromeric Location during Early Schizogony

To locate the condensin SMC subunits during two proliferative stages (schizogony and male gametogenesis) of the *Plasmodium* life cycle, transgenic parasite lines were created to express GFP-tagged SMC2 and SMC4 using single-crossover homologous recombination (Figure S2A). Integration PCR and western blot experiments were used to confirm the successful generation of transgenic lines (Figures S2B and S2C). We found that SMC2 and SMC4 are expressed during both schizogony and male gametogenesis. In early schizonts within host red blood cells, we observed discrete foci in the parasite cell adjacent to the nuclear DNA for both SMC2 and SMC4, whereas in mature schizonts, the signal was dispersed throughout the nucleus (Figures 3.2A and 3.2B). During male gametogenesis, the proteins were also dispersed throughout the nucleus (Figures 3.2C and 3.2D). To validate the SMC4 subcellular location, fractionation of cytoplasmic and nuclear extracts derived from purified gametocytes revealed the presence of SMC4 in the nucleus (Figure S2D). In addition, we observed SMC4GFP distributed either as dispersed in the nucleus or at a discrete focus adjacent to the DNA throughout the parasite life cycle, including in female gametocytes, in ookinetes, during oocyst development, and in the liver stages (Figures S3A–S3C), suggesting that condensin core subunits are likely involved at all proliferative stages of the parasite life cycle. To examine whether these foci are centromeric or centrosomal, we used two approaches: we performed a colocalization experiment using parasites expressing SMC4GFP crossed with those expressing NDC80mCherry, a kinetochore/centromeric marker protein (Cheeseman,

2014, McKinley and Cheeseman, 2016, Musacchio and Desai, 2017, Pandey et al., 2019), and we performed immunofluorescence assays using anti-centrin and anti-α-tubulin, together with anti-GFP antibodies. Live imaging using the SMC4GFP and NDC80mCherry genetic cross showed colocalization of SMC4 and NDC80 in early schizonts and discrete foci of NDC80mCherry alone in gametocytes (Figures 3.2E and 3.2F). A similar pattern was observed in ookinetes and oocysts, with centromeric colocalization of SMC4 and NDC80 (Figure S3D). In early schizonts, immunofluorescence assays with anti-centrin antibodies revealed that SMC4 is located between centrin and DAPI-stained nuclear DNA, confirming the non-centrosomal localization of SMC4 (Figure 3.2G). However, partial colocalization was observed with anti-α-tubulin antibodies in schizonts (Figure 3.2G) and during male gametogenesis (Figure 3.2H).

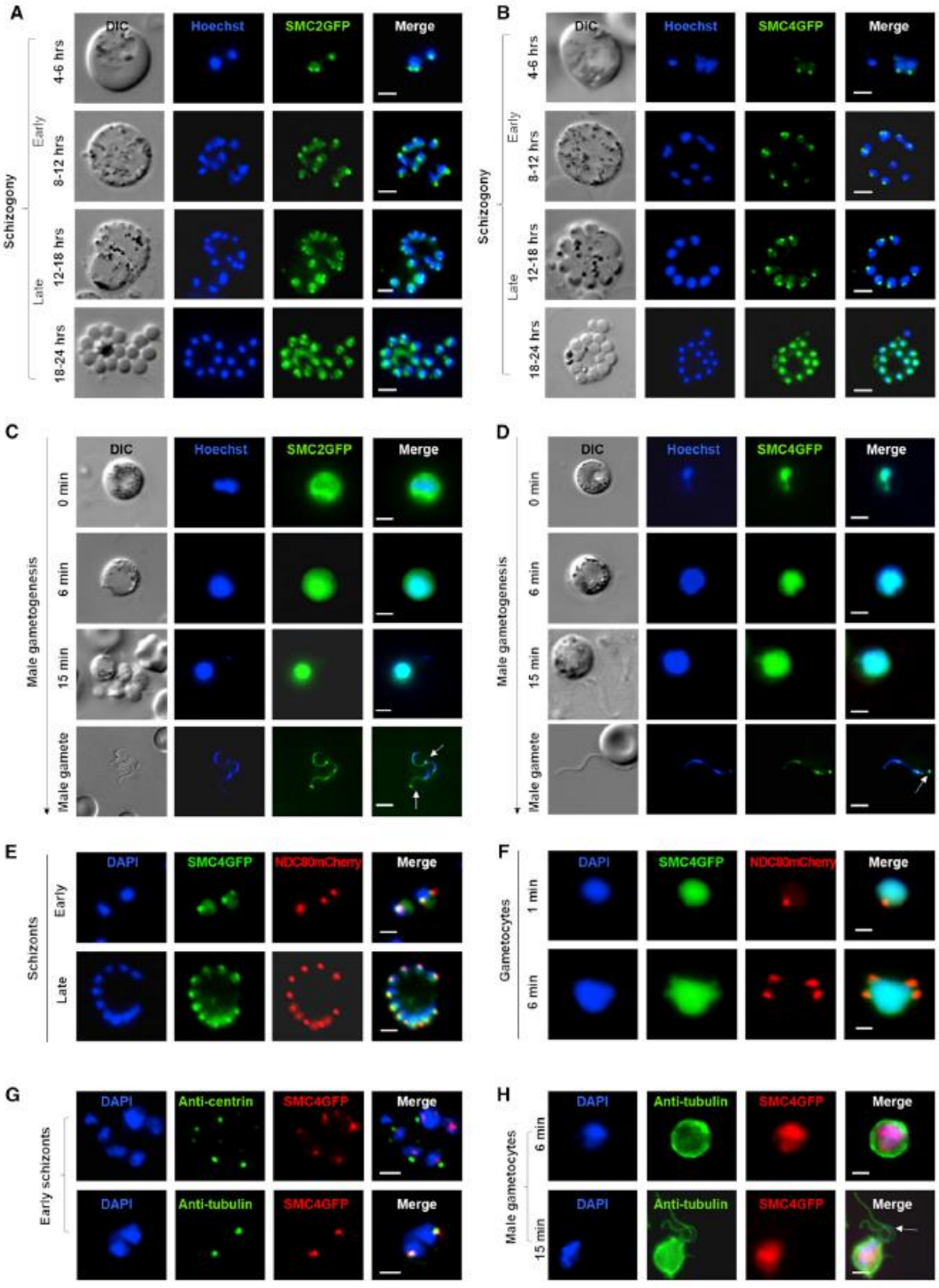
A
DIC
Hoechst
SMC2GFP
Merge
Schizogony
Early
Late
4-6 hrs
8-12 hrs
12-18 hrs
18-24 hrs
B
DIC
Hoechst
SMC4GFP
Merge
Schizogony
Early
Late
4-6 hrs
8-12 hrs
12-18 hrs
18-24 hrs
C
DIC
Hoechst
SMC2GFP
Merge
Male gametogenesis
0 min
6 min
15 min
Male gamete
D
DIC
Hoechst
SMC4GFP
Merge
Male gametogenesis
0 min
6 min
15 min
Male gamete
E
DAPI
SMC4GFP
NDC80mCherry
Merge
Schizonts
Early
Late
F
DAPI
SMC4GFP
NDC80mCherry
Merge
Gametocytes
1 min
6 min
G
DAPI
Anti-centrin
SMC4GFP
Merge
DAPI
Anti-tubulin
SMC4GFP
Merge
Early schizonts
H
DAPI
Anti-tubulin
SMC4GFP
Merge
DAPI
Anti-tubulin
SMC4GFP
Merge
Male gametocytes
6 min
15 min

Fig. 3.2 Temporal Dynamics of Condensin (SMC2 and SMC4) in Two Distinct *Plasmodium* Proliferative Stages (Schizogony and Male Gametogenesis) Undergoing Atypical Mitotic Division (A–D) Live cell imaging of SMC2GFP and SMC4GFP expressed during schizogony (100× magnification) (A and B) and male gametogenesis (63× magnification) (C and D). Time points indicate imaging done for the schizont and gametocyte after the start of respective cultures. The white arrows in (C) and (D) indicate discrete localization in male gametes. DIC, differential interference contrast; merge shows Hoechst and GFP. Scale bar, 2 μm. (E and F) Live cell imaging of SMC4GFP and NDC80mCherry localization during schizogony (E) and male gametogenesis (F). Merge shows Hoechst, GFP, and mCherry. Scale bar, 2 μm. (G and H) Immunofluorescence fixed-cell imaging of SMC4GFP and colocalization with antibodies specific for centrin and α-tubulin in mitotic cells (early schizonts, 100× magnification in G; male gametocytes, 63× magnification in H). The white arrow in (H) indicates an exflagellating male gamete. Scale bar, 2 μm.

To identify the SMC2 and SMC4 DNA binding sites in a genome-wide manner, we performed ChIP-seq experiments for the schizont stage (after 8 h in culture) and gametocyte stage (6 min post-activation) using SMC2GFP- and SMC4GFP-tagged parasites. A wild-type (WT) strain (WTGFP) was used as a negative control. Binding of the SMC2 and SMC4 subunits was restricted to a region close to the previous computationally annotated centromeres (centromere locations of *P. berghei* chromosomes had been predicted using *P. falciparum* as a reference and the conservation of genomic sequences among *Plasmodium* spp.) of all 14 chromosomes at the early schizont stage (Figure 3.3A; Iwanaga et al., 2012). At this stage, we observed significant ChiP-seq peaks with an average of 14.6- and 12.7-fold change (FC) compared with background in all pericentromeric regions for SMC2 and SMC4, respectively. This restriction was not observed during gametogenesis, which instead had a random distribution of condensin core subunits. Although non-significant peaks (less than 1.5 FC) were detected in pericentromeric regions for SMC2, an even smaller increase in ChIP-seq coverage in pericentromeric regions (1.08 FC) observed for SMC4 lead us to believe that the small peak observed during gametogenesis for SMC2 could be explained by a weak interaction of SMC2 or residual asexual signal in gametocyte samples. Identical patterns were obtained between biological replicates for each condition analyzed, confirming the reproducibility of the ChIP-seq experiments and suggesting a distinct function for the core subunits in these two mitotic stages (Figure 3.3A). To confirm the location of the kinetochores/centromeres, we also performed ChIP-seq with activated gametocytes from the NDC80GFP line (Pandey et al., 2019). Strong ChIP-seq

peaks with an average of 74.8 FC were observed at the centromeres of all 14 chromosomes with perfect overlap with SMC2/SMC4 signals. These data clearly confirmed that the SMC2/SMC4 location during early schizogony was the centromeric location of NDC80 (Figure 3.3A).

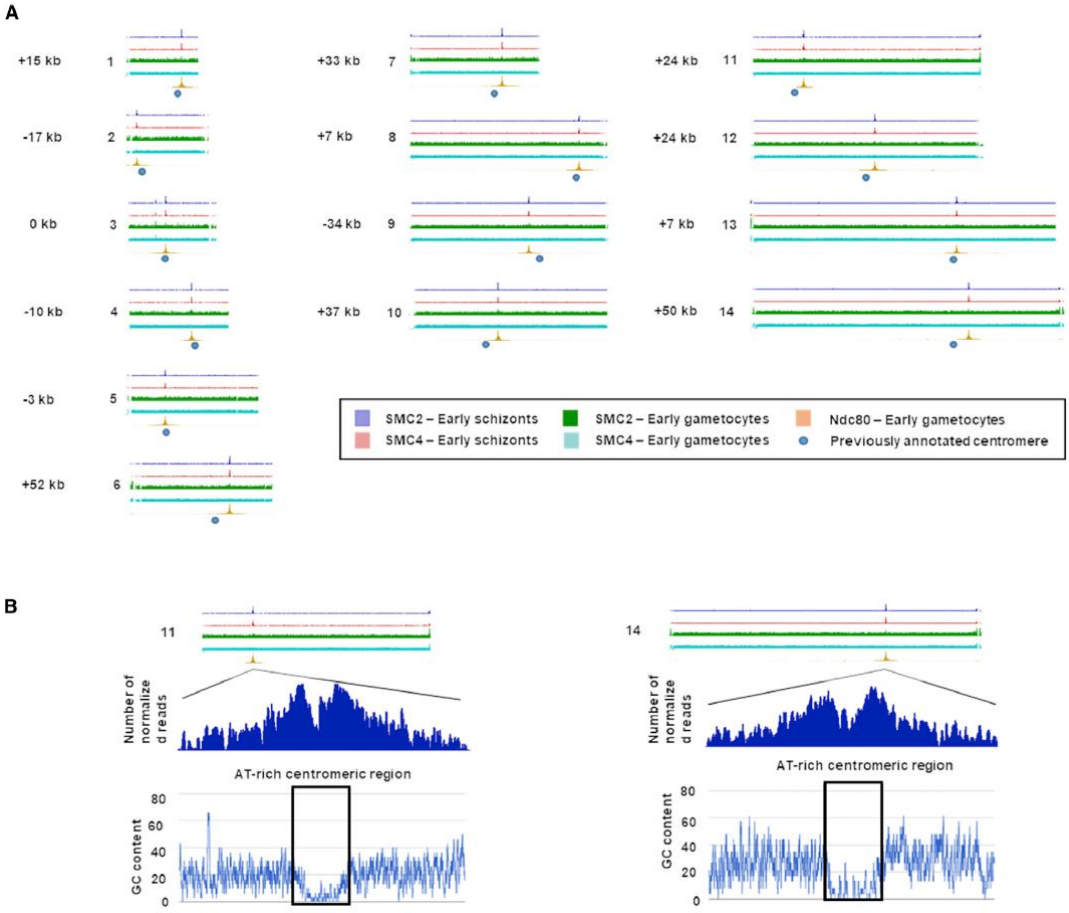

Fig. 3.3 ChIP-Seq Analysis of SMC2GFP, SMC4GFP, and NDC80 Profiles (A) Genome-wide ChIP-seq signal tracks for SMC2GFP and SMC4GFP for all 14 chromosomes in schizont and gametocyte stages. The SMC schizont tracks and the NDC80 track each represent the average of two biological replicates, while the SMC4 gametocyte track represents the average of four biological replicates, one of which had two technical replicates averaged together. The locations of previously annotated centromeres are indicated by blue circles. SMC2 and SMC4 proteins bind near the putative centromere in each of the 14 chromosomes (distance from centromere is shown in ± kilobases). The centromeric location was confirmed by genome-wide ChIP-seq signal tracks for NDC80GFP for all 14 chromosomes in the gametocyte stage. The scale for all SMC tracks is between 0 and 20 normalized read counts, while the NDC80 track is between 0 and 150 normalized read counts because of the considerable FC enrichment observed for the NDC80 ChIP-seq signal. (B) Zoom-in regions associated with the identified ChIP-seq peak. Low GC content at the centers of peaks shown for chromosome 11 and chromosome 14 suggests association of the proteins with these newly defined centromeres in the schizont stage. Signals are plotted on a normalized read per million (RPM) basis.

The chromosome binding sites detected for NDC80 and for SMC2 and SMC4 were slightly offset from the locations previously annotated as centromeres (Iwanaga et al., 2012). However, the ChIP-seq peaks for all 14 chromosomes were centered on distinct regions of very low GC content that are not present in the previously annotated centromeric regions (shown for chromosomes 11 and 14 in Figure 3.3B). AT-rich troughs have been associated with centromeres in yeasts (Lynch et al., 2010). The peaks located within extended intergenic regions indicate the NDC80/SMC binding sites as experimentally validated centromeres for all 14 *P. berghei* chromosomes (Table S1).

The Full Condensin Complex Is Present during Male Gametogenesis and in Mature Schizonts, but Ancillary Proteins Are Absent from Early Schizonts

To examine the colocalization of SMC2 and SMC4 proteins, we generated transgenic parasite lines expressing either SMC2mCherry or SMC4GFP and crossed them genetically. The progeny, expressing both SMC2mCherry and SMC4GFP, showed colocalization of the two proteins during schizogony and gametogenesis (Figure 3.4A) consistent with SMC2 and SMC4 heterodimer complex formation at both stages.

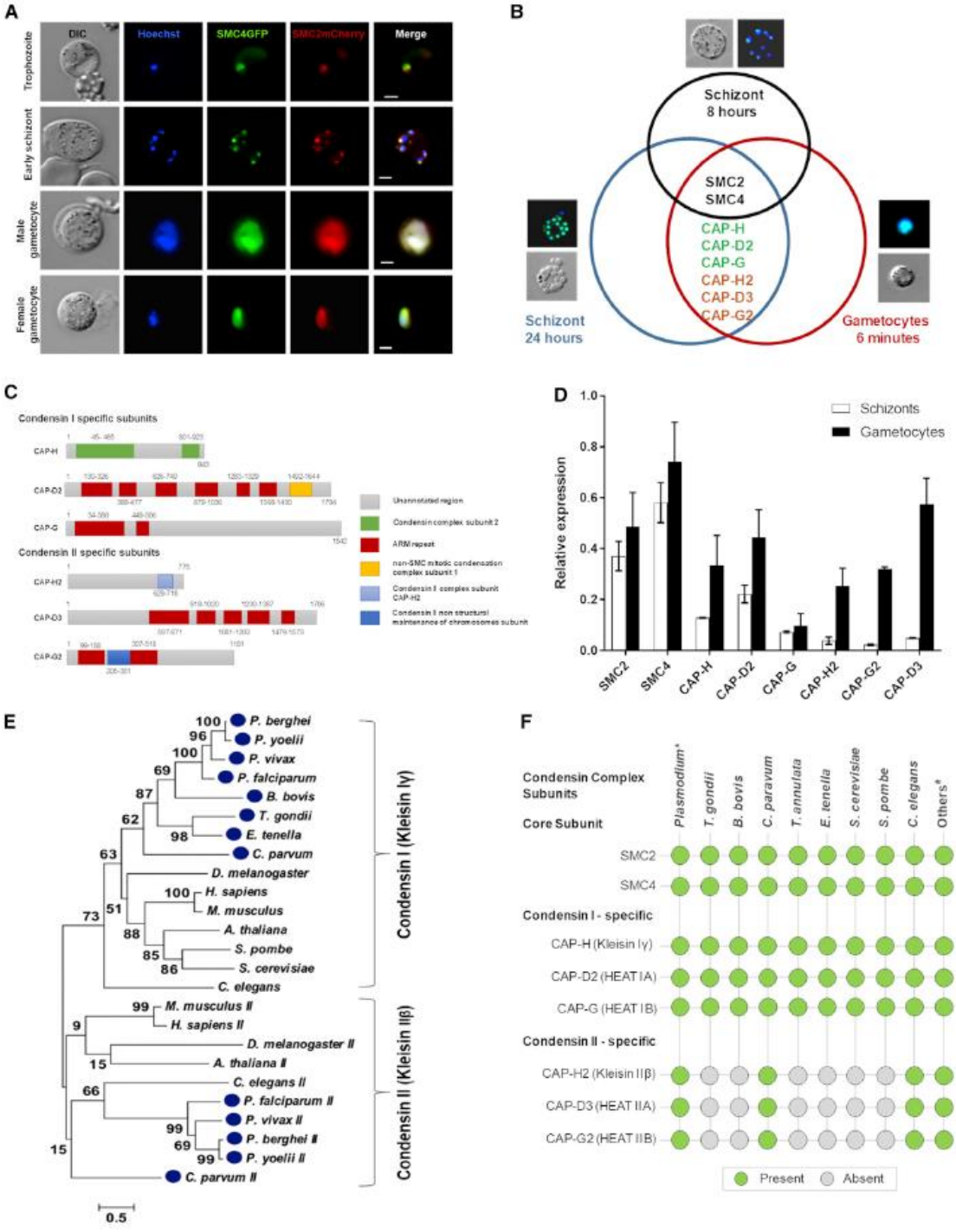
A
DIC
Hoechst
SMC4GFP
SMC2mCherry
Merge
Trophozoite
Early schizont
Male gametocyte
Female gametocyte
B
Schizont
8 hours
SMC2
SMC4
CAP-H
CAP-D2
CAP-G
CAP-H2
CAP-D3
CAP-G2
Schizont
24 hours
Gametocytes
6 minutes
C
Condensin I specific subunits
CAP-H
CAP-D2
CAP-G
Condensin II specific subunits
CAP-H2
CAP-D3
CAP-G2
Unannotated region
Condensin complex subunit 2
ARM repeat
non-SMC mitotic condensation complex subunit 1
Condensin II complex subunit CAP-H2
Condensin II non structural maintenance of chromosomes subunit
D
Schizonts
Gametocytes
Relative expression
1.0
0.8
0.6
0.4
0.2
0.0
SMC2
SMC4
CAP-H
CAP-D2
CAP-G
CAP-H2
CAP-G2
CAP-D3
E
P. berghei
P. yoelii
P. vivax
P. falciparum
B. bovis
T. gondii
E. tenella
C. parvum
D. melanogaster
H. sapiens
M. musculus
A. thaliana
S. pombe
S. cerevisiae
C. elegans
M. musculus II
H. sapiens II
D. melanogaster II
A. thaliana II
C. elegans II
P. falciparum II
P. vivax II
P. berghei II
P. yoelii II
C. parvum II
Condensin I (Kleisin Iγ)
Condensin II (Kleisin IIβ)
0.5
F
Condensin Complex Subunits
Core Subunit
Plasmodium*
T. gondii
B. bovis
C. paravum
T. annulata
E. tenella
S. cerevisiae
S. pombe
C. elegans
Others*
SMC2
SMC4
Condensin I - specific
CAP-H (Kleisin Iγ)
CAP-D2 (HEAT IA)
CAP-G (HEAT IB)
Condensin II - specific
CAP-H2 (Kleisin IIβ)
CAP-D3 (HEAT IIA)
CAP-G2 (HEAT IIB)
Present
Absent

Fig. 3.4 Differential Condensin Complex Formation during Schizogony and Male Gametogenesis, and Phylogenetic Analysis of Kleisin (A) Colocalization of SMC4GFP (green) and SMC2mCherry (red). Merge shows Hoechst, GFP, and mCherry. Scale bar, 2 μm. (B) Venn diagram displays the unique and shared proteins in the condensin complex of schizonts and gametocytes. Analysis of SMC2GFP and SMC4GFP protein complexes was done by tryptic digestion and LC-MS/MS following GFP-specific immunoprecipitation from a lysate of schizonts maintained in culture for 8 h and 24 h, and gametocytes were activated for 6 min. The representative live cell pictures have been taken from Figure 3.2. The list of all identified proteins is provided as Table S2. (C) Different domain architecture for subunits of condensin I and condensin II complexes. The schematic figure displays the domain composition and protein length in the respective complex subunits. (D) qRT-PCR analysis of condensin complex subunit expression in schizont and gametocyte stages of the parasite life cycle. Error bar, ± SD, n = 3. (E) Maximum likelihood phylogeny based on the alignment of kleisin subunits from apicomplexan species (*Plasmodium* spp., *Toxoplasma gondii*, *C. parvum*, *Babesia bovis*, and *Eimeria tenella*) and other selected organisms. Topological support from bootstrapping is shown at the nodes. The protein sequences for selected organisms have been provided in Data S1. (F) Distribution of condensin components across Apicomplexa and other organisms. Presence (green circle) or absence (gray circle) of condensin complex genes in each genome. Asterisk represents 4 *Plasmodium* spp., namely, *P. falciparum*, *P. vivax*, *P. berghei*, and *P. yoelii*; hashmark denotes *H. sapiens*, *A. thaliana*, and *D. melanogaster*.

Next, we directly investigated the interaction between SMC2 and SMC4 and the presence of other potential interacting partner proteins, such as other condensin components (Figure 3.1A). We immunoprecipitated SMC2GFP and SMC4GFP from lysates of cells undergoing asexual endomitotic division at two time points (early schizogony, following 8 h incubation in schizont culture medium *in vitro*, when most parasites are undergoing nuclear division and show discrete SMC2/SMC4 foci, and after 24 h incubation in schizont culture medium, when most parasites are mature schizonts or free merozoites with a dispersed SMC2/SMC4 location). We also immunoprecipitated the proteins from parasites undergoing gametogenesis (at 6 min after activation, when the chromosomes are beginning to condense and cells are in the last phase before cytokinesis). Immunoprecipitated proteins were then digested with trypsin, and the resultant peptides were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS). From all three samples, we recovered peptides from both SMC subunits, confirming SMC2-SMC4 heterodimer formation during both schizogony and male gametogenesis (Figure 3.4B). From early schizonts, only SMC2- and SMC4-derived peptides were recovered, whereas from mature schizonts and gametocytes, we detected kleisins and other components of canonical condensin I and II complexes, together with the SMC subunits except for CAP-G (Figure 3.4B; Table S2). In some early schizont samples, condensin II Heat subunits (CAP-G2 and CAP-D3) were observed; however, kleisin was never recovered in five early schizont experimental replicates; therefore, we assume that formation of the complete condensin II complex does not occur at this stage (Table S2). All conventional condensin I and II complex subunits were identified by *in silico* analysis

of the *Plasmodium* genome; for example, a BLAST search using fission yeast CAP-G revealed *Plasmodium* merozoite organizing protein (MOP) to be *Plasmodium* CAP-G. This is in agreement with the immunoprecipitation data in which we detected CAP-H, CAP-D2, and CAP-G (annotated as MOP) from the condensin I complex, and CAP-H2, CAP-D3, and CAP-G2 from the condensin II complex. The predicted domain architecture of these subunits from *P. berghei* is shown in Figure 3.4C. We also investigated the expression profile of condensin complex subunits in schizont and gametocyte stages of the parasite life cycle. We performed qRT-PCR for all eight components of the condensin complexes. The results revealed comparatively high expression of all condensin subunits in gametocytes compared with schizonts (Figure 3.4D). In addition, levels of non-SMC condensin II components expressed in schizonts were lower than those of non-SMC condensin I components, whereas in gametocytes, comparable expression levels were observed for condensin I and II components except for CAP-G (Figure 3.4D). In ookinetes, the expression of non-SMC condensin components was low, except for CAP-H, which showed a level similar to that of the SMC components (Figure S4). Chromosome condensation in schizonts has not been reported, so why all components of both condensin complexes are present in mature schizonts is unclear. The presence of both condensin I and condensin II in male gametocytes is consistent with the potentially atypical chromosomal condensation that has been previously observed in male gametocytes just before exflagellation (Sinden, 1991b, Sinden et al., 1976, Sinden and Hartley, 1985).

In view of the importance of kleisin to the structure and function of the condensin complexes, we examined the evolutionary relationships of kleisin among some apicomplexans and other organisms, including *S. cerevisiae*, *A. thaliana*, and *Homo sapiens*. The phylogenetic analysis indicated that kleisin is clustered into two groups, which correspond to components of condensin I and condensin II (Figures 3.1A and 3.4E). The presence of both condensin I and condensin II component genes only in *Plasmodium* and *Cryptosporidium* shows that the requirement for both condensin complexes is not a universal feature of Apicomplexa (Figure 3.4F). Other apicomplexans, for example, *Toxoplasma* and *Eimeria*, have only condensin I components, similar to yeast homologs. These data suggest that *Plasmodium* and *Cryptosporidium* possess features of chromosome condensation and segregation that are distinct from those of other members of the phylum.

Knockdown of Condensin (SMC2 and SMC4) Expression Affects Parasite Proliferation and Impairs Parasite Transmission

To examine further the functions of SMC2 and SMC4, we first attempted to delete the two genes. In both cases, we were unable to produce gene knockout (KO) mutants (Figure S5A). Similar results have been reported previously from large-scale genetic screens in *P. berghei* (Bushell et al., 2017, Schwach et al., 2015). Altogether, these data indicate that the condensin subunits SMC2 and SMC4 are likely essential for asexual blood stage development (schizogony). To investigate the function of SMC2 and SMC4 during cell division in male gametogenesis, we used a promoter trap double homologous

recombination (PTD) approach to downregulate gene expression at this stage by placing each of the two genes under the control of the AMA1 promoter (Figure S5B). AMA1 is known to be highly expressed in asexual blood stages, but not during sexual differentiation. This strategy resulted in the successful generation of two transgenic parasite lines: P_{ama1}*smc2* (SMC2PTD) and P_{ama1}*smc4* (SMC4PTD) (Figure S5C).

Because SMC2PTD and SMC4PTD had similar phenotypes, we performed a global transcriptome analysis only on SMC4PTD to identify other affected genes and regulators involved in cell division and proliferation. This analysis of SMC4PTD gametocytes 30 min after activation (when chromosome condensation and exflagellation are complete) confirmed the nearly complete ablation of *smc4* gene expression (Figure 3.5A). For pairs of the two biological samples (WTGFP and SMC4PTD), Spearman correlation coefficients of 0.97 and 0.99 respectively, demonstrate the reproducibility of this experiment (Figure 3.5B). In addition to SMC4, expression of a further 104 genes was significantly downregulated, while expression of only 5 genes was significantly upregulated (Figure 3.5C; Table S3). Gene Ontology (GO) enrichment analysis of the downregulated genes identified several associated with microtubule and cytoskeletal function (Figure 3.5D). The reduced expression levels of 10 of these genes was also examined by qRT-PCR (Figure 3.5E). By this method, there was a statistically significant difference in the level of expression of 9 of these genes when comparing samples from WTGFP and SMC4PTD (Figure 3.5E). Of particular interest are AP2-O2 (an AP2 domain transcription factor) and HMG (putative high-mobility group protein B3), which act as transcription regulators in ookinetes; a putative SET domain protein which is

known to be involved in methyl group transfer from S-adenosyl-L-methionine (AdoMet) to a lysine residue in histones and most likely associated with transcriptional repression; and finally RCC, a protein predicted to be involved in chromosome condensation and chromosomal dynamics (Bahl et al., 2002). Other genes that were significantly downregulated include FRM2, involved in cytoskeleton organization; CCRNOT2 and NOT that form the CCR4-NOT complex, a key regulator of eukaryotic gene expression; and SEC7, involved in regulation of ARF protein signal transduction. Other significantly downregulated genes, include AP2 transcription factor AP2-Sp; molecular motor kinesin-4, a putative regulator of chromosome condensation (PBANKA_0820800); and SMC1, a member of the SMC family, are all known to be involved in either gene expression or chromatid segregation.

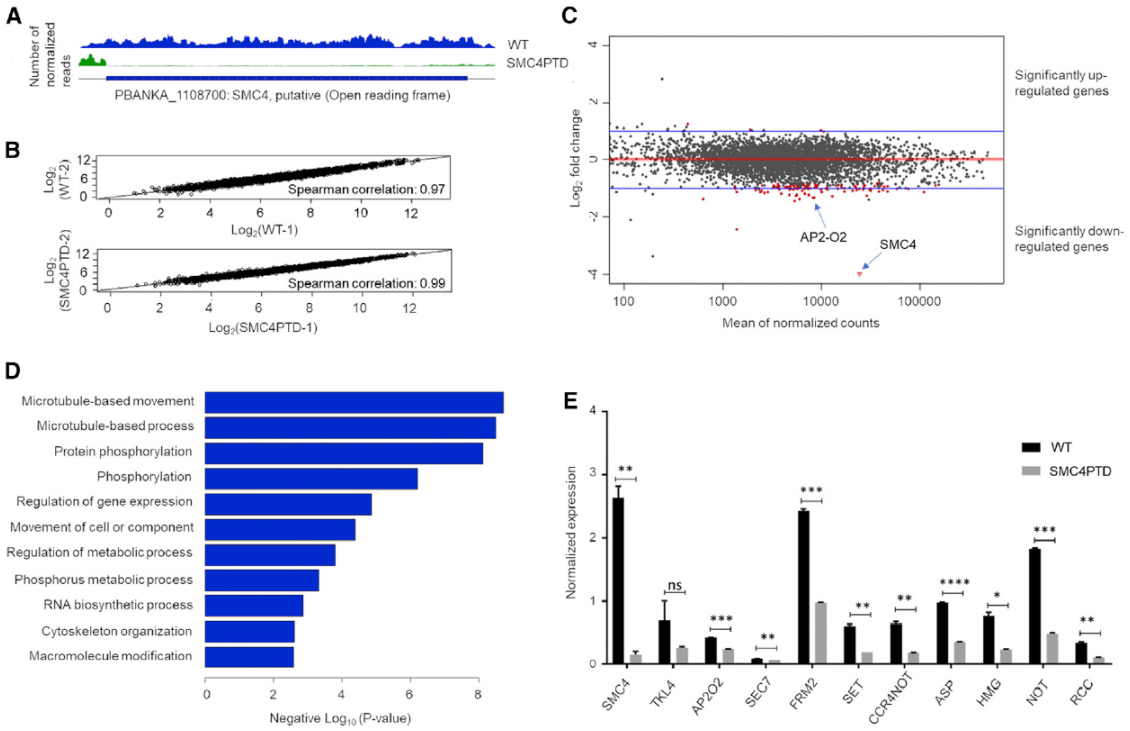

Fig. 3.5 Global Transcriptomic Analysis for SMC4PTD in Activated Gametocytes by RNA-Seq (A) Confirmation of successful depletion of SMC4 transcript in the SMC4PTD line. Tracks shown each represent the average of two biological replicates. (B) Log-normalized scatterplots demonstrating high correlation between replicates for genome-wide expression. The Spearman correlation was calculated using read counts normalized by number of mapped reads per million. (C) MA plot summarizing RNA-seq results. M, log ratio; A, mean average. Every gene is placed according to its log fold expression change in the SMC4PTD line compared with the WT line (y axis), and average expression level across replicates of both lines (x axis). Red indicates statistical significance of differential expression at the false-positive threshold of 0.05. 105 genes are downregulated in the SMC4PTD line, and 5 genes are upregulated. (D) GO enrichment analysis of genes with $\log_{10}$ fold expression change of −0.5 or lower in the SMC4PTD line. (E) qRT-PCR analysis of selected genes identified as downregulated in (C), comparing transcript levels in WT and SMC4PTD samples. Error bar, ± SEM, n = 3. Unpaired t test was performed for statistical analysis: $^{*}p < 0.05$ $^{**}p < 0.001$, $^{***}p < 0.0001$, and $^{****}p < 0.00001$. See also Figure S5 and Tables S3 and S4.

Although we were unable to detect any particular impaired phenotype in the SMC2PTD and SMC4PTD lines at the asexual blood stage (schizogony), and the parasite formed a similar number of schizonts and nuclei compared with the WTGFP line (Figure 3.6A), we observed an ~50% reduction in the number of exflagellation centers during male gametogenesis (Figure 3.6B). Fertilization and zygote formation leading to ookinete conversion was reduced to 10%–15% compared with the WTGFP line (Figure 3.6C). The ookinete motility assay showed normal movement of SMC4PTD ookinetes compared with WT (Videos S1 and S2). In the mosquito gut on 9, 14, and 21 days post-infection, we detected significantly fewer oocysts in the SMC2PTD and SMC4PTD lines (Figures 3.6D and 3.6E). Furthermore, the oocysts were considerably smaller compared with those of WTGFP (Figures 3.6D and 3.6F), with unequal distribution and clusters of DNA in some oocysts at 14 and 21 days post-infection. No sporogony or endomitosis was observed within oocysts (Figure 3.6G). We were also unable to detect sporozoites in the mosquito salivary glands (Figure 3.6H); hence, no parasite transmission from infected mosquitoes to mice was observed for either SMC2PTD or SMC4PTD parasite lines in bite-back experiments (Figure S5D), indicating that condensins are required for parasite transmission.

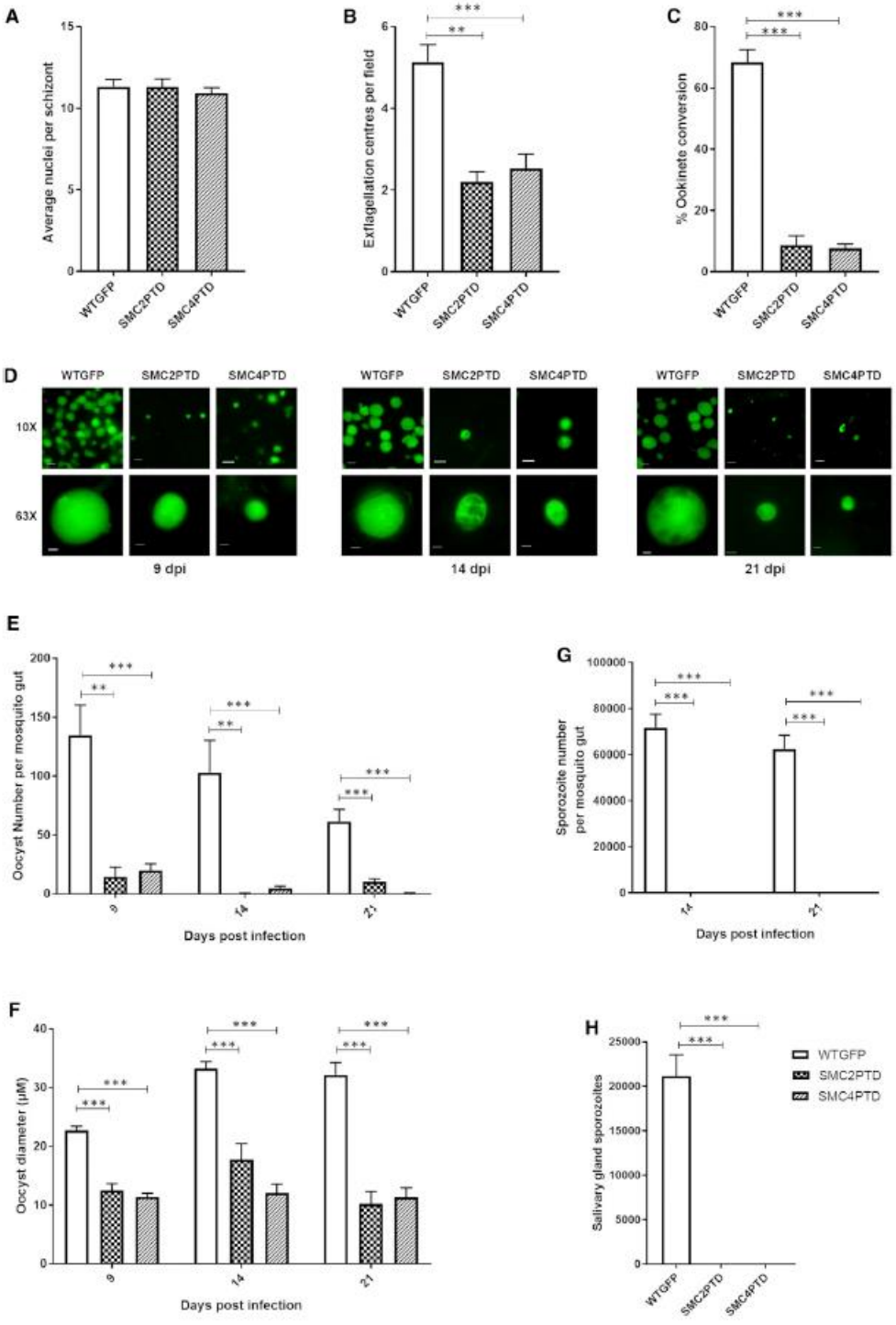
A
Average nuclei per schizont
WTGFP
SMC2PTD
SMC4PTD
B
Exflagellation centres per field
C
% Ookinete conversion
D
10X
63X
9 dpi
14 dpi
21 dpi
E
Oocyst Number per mosquito gut
Days post infection
G
Sporozoite number per mosquito gut
F
Oocyst diameter (µM)
H
Salivary gland sporozoites

Fig. 3.6 Phenotypic Analysis of Conditional Gene Expression Knockdown in SMC2PTD and SMC4PTD Transgenic Lines at Various Proliferative Stages during the Life Cycle (A) Average number of nuclei per schizont (mitotic division within red cell). n = 5 (minimum 100 cells). Error bar, ± SEM. (B) Number of exflagellation centers (mitotic division during male gametogenesis) per field at 15 min post-activation. n = 3 independent experiments (10 fields per experiment). Error bar, ± SEM. (C) Percentage ookinete conversion from zygotes. Minimum of 3 independent experiments (minimum 100 cells). Error bar, ± SEM. (D) Live cell imaging of WTGFP, SMC2PTD, and SMC4PTD oocysts (endomitosis in parasite within mosquito gut) at 9, 14, and 21 days post-infection (dpi), using 10× and 63× magnification to illustrate differences in size and frequency. Scale bar, 5 μm (63×) and 20 μm (10×). (E) Number of oocysts at 9, 14, and 21 dpi. n = 3 independent experiments with a minimum of 5 mosquito guts. Error bar, ± SEM. (F) Oocyst diameter at 9, 14, and 21 dpi. n = 3 independent experiments. Error bar, ± SEM. (G) Number of sporozoites at 14 and 21 dpi in mosquito gut. n = 3 independent experiments with a minimum of 5 mosquito guts. Error bar, ± SEM. (H) Number of sporozoites at 21 dpi in mosquito salivary gland. Minimum of 3 independent experiments. Error bar, ± SEM. Unpaired t test was performed for statistical analysis: $^{*}p < 0.05$ $^{**}p < 0.01$, and $^{***}p < 0.001$.

Discussion

Condensins are multi-subunit complexes that are involved in chromosomal condensation, organization, and segregation and have been widely studied in many eukaryotes (Hirano, 2016, Uhlmann, 2016). The role of condensins in many unicellular protozoans such as *Plasmodium* remained elusive. Here we describe the structure, localization, and functional role of the condensin core subunits (SMC2/SMC4) in the mouse malaria-causing parasite *P. berghei* using MD, live cell imaging, ChIP-seq, protein pull-down, and conditional gene knockdown approaches. *Plasmodium* shows atypical features of closed mitotic division resembling endomitosis during schizogony (with no observed chromosomal condensation and extensive asynchronous nuclear division followed by a final round of synchronous nuclear division before cytokinesis) and endoreduplication in male gametogenesis (with rapid chromosome replication and atypical condensation before nuclear division, cytokinesis, and exflagellation) (Arnot et al., 2011, Sinden, 1991b). Our previous studies have identified an unusual repertoire of proteins involved in the regulation of the parasite cell cycle and cell proliferation: there is no identifiable centrosome, no obvious complement of cell-cycle cyclins, a small subset of APC components, a set of divergent and *Plasmodium*-specific CDKs, and an absence of polo-like kinases and CDC24 and CDC14 phosphatases compared with most organisms that have been studied (Arnot et al., 2011, Francia et al., 2016, Guttery et al., 2012a, Guttery et al., 2014, Roques et al., 2015, Tewari et al., 2010).

Here, using bioinformatics screening, we showed that both condensin I and condensin II complex subunit components are encoded in the *P. berghei* genome, as has been described for *P. falciparum* in PlasmoDB (Bahl et al., 2002). The two core subunits of condensin, SMC2 and SMC4, have low sequence similarity to the proteins in model organisms but a similar protein structure as predicted by molecular modeling (Kelley et al., 2015).

Protein localization studies at different stages of the parasite life cycle using live cell imaging of SMC2GFP and SMC4GFP and immunofluorescence show distinct patterns during the mitotic divisions of early and late schizogony and male gametogenesis. Whereas discrete protein foci were detected during endomitosis in early schizogony, a stage characterized by asynchronous nuclear division, dispersed nuclear localization was observed during late schizogony and male gametogenesis. By immunofluorescence, the discrete foci of SMC2/SMC4GFP in early schizogony were located close to the stained DNA and close to, but not coincident with, centrin, marking the SPB. ChIP-seq analyses suggest that in early schizonts, SMC2 and SMC4 form a complex that binds at or near the centromere of all 14 chromosomes, a result that is substantiated by the dual-labeling and colocalization studies with the kinetochore/centromere marker NDC80. ChIP-seq analysis of NDC80GFP binding during gametogenesis confirms the centromeric location of SMC2/SMC4. These results suggest that the SMC2-SMC4 complex alone is restricted to binding centromeric regions in the highly proliferative early schizont stage, in which it may have a constrained role in sister chromatid cohesion and segregation (Iwasaki and Noma, 2016). Genome-wide studies of condensin distribution in mammalian or yeast

cells have shown that the complex is non-randomly distributed across the chromosomes and often found at the boundaries of topologically associating domains (TADs) within chromosome territories, which supports the proposed role in transcriptional regulation and global chromosomal organization (Kim et al., 2016, Yuen et al., 2017). Because the *Plasmodium* genome lacks classical TADs (Ay et al., 2014, Bunnik et al., 2018, Bunnik et al., 2019), a restricted distribution of the condensin complex on the centromere of all 14 chromosomes suggests a distinct function.

Although we detected only the SMC2-SMC4 heterodimer in early schizogony, located within the nucleus at the centromere and at a discrete focus adjacent to, but distinct from, the SPB, the protein interaction analysis using SMC2/SMC4GFP showed that other subunits of the full condensin I and II complexes were present during late schizogony and male gametogenesis. It is thought that in late schizogony, the last set of divisions is synchronous and followed by cytokinesis to produce mature merozoites and that, at this stage, the dispersed distribution of SMC2/SMC4GFP was observed. A similar dispersed protein pattern in the nucleus was observed during male gametogenesis, which is also associated with the presence of condensin complex I and II proteins preceding exflagellation. No chromosomal condensation has been reported in mature schizonts, although it has been observed in male gametogenesis, as shown by electron microscopy studies (Sinden, 1991b, Sinden et al., 1976, Sinden and Hartley, 1985). The presence of both condensin I and condensin II complexes in late schizogony suggests that the full complexes are only involved in the final synchronous cycle of nuclear division preceding cytokinesis. Previous studies have reported that non-SMC condensin II subunits are

dispensable during schizogony but non-SMC components of condensin I complex are not (Bahl et al., 2002, Schwach et al., 2015). One of the components, CAP-G, also annotated as MOP, has been demonstrated to be essential for cytokinesis in *P. falciparum* asexual blood stages (Absalon et al., 2016). Our bioinformatics analysis shows that *Plasmodium* (two hosts, asynchronous cell division) and *Cryptosporidium parvum* (single host, with long-duration dormant phase outside of the host) are the only two apicomplexan parasites that have components of both condensin I and condensin II complexes encoded in the genome, similar to what is observed in higher eukaryotes. They also have unusual modes of cell division compared with apicomplexans such as *Toxoplasma* and *Babesia*, which display symmetrical modes of division (Francia and Striepen, 2014). Apicomplexan parasites that encode only a single condensin complex show no chromosome condensation. Similarly, other parasites with closed mitosis and no chromosome condensation, for example, *Trypanosoma brucei*, encode only the condensin I complex (Hammarton, 2007), whereas in parasites with chromosome condensation, such as *Giardia intestinalis*, both condensin I and condensin II are present. However, *Giardia* lacks one of the conventional non-SMC Heat subunits (CAP-G and CAP-G2), whereas CAP-D2 and CAP-D3 are present (Tůmová et al., 2015). Another protist, *Tetrahymena thermophila*, shows chromosomal condensation, exhibits noncanonical division between somatic and germline cells, and has an expanded set of condensin I paralogs, with different kleisin components between germline (Cph1 and Cph2) and somatic cells (Cph3, Cph4, and Cph5) (Howard-Till and Loidl, 2018, Howard-Till et al., 2019).

Condensin I and II complexes display distinct localization patterns in various organisms. In the red alga *Cyanidioschyzon merolae* (Fujiwara et al., 2013), condensin II has a centromeric location during metaphase, whereas condensin I distributes more broadly along the chromosome arms. In higher eukaryotes, including *Drosophila melanogaster* (Oliveira et al., 2007), *Caenorhabditis elegans* (Collette et al., 2011), and HeLa cells (Hirota et al., 2004, Ono et al., 2004), condensin I is present in the cytoplasm and has a chromosomal location after the nuclear envelope is dissolved in open mitosis. The nuclear localization of condensin II is observed in interphase, it is stabilized on chromatin during prophase, and the complex remains associated with chromosomes throughout mitosis, at least in HeLa cells. Budding yeast and fission yeast, which undergo closed mitosis like *Plasmodium*, have only a single condensin complex, but there is a differential pattern of subcellular location in each species. In budding yeast, the condensin I complex is located in the nucleus throughout the cell cycle, a pattern observed for condensin II in higher eukaryotes, despite the greater protein sequence similarities of the yeast complex to higher eukaryote condensin I (Thadani et al., 2012). In addition, within the nucleus, the condensin location at the kinetochore is cell cycle dependent (Bachellier-Bassi et al., 2008). In fission yeast, the single condensin complex is predominantly cytoplasmic during interphase and nuclear during mitosis, with the location dependent on CDK phosphorylation at Thr19 of SMC4/Cut3 (Sutani et al., 1999).

The present study shows that the SMC2-SMC4 complex plays an essential role during schizogony, because we, and those who performed previous genome-wide functional screens (Bushell et al., 2017), were unable to disrupt the genes. Our conditional

knockdown using the PTD approach suggests that reduction of both SMC2 and SMC4 expression affects both male gametogenesis and zygote differentiation and causes total impairment of endomitotic cell division in the oocyst, thereby blocking parasite transmission.

The partial defect observed in male gamete formation (exflagellation) in the PTD parasite lines may be because of the necessity of condensin complex formation for proper chromosomal condensation during exflagellation. Transcriptomic analysis of the SMC4PTD line confirmed the reduced expression of the *smc4* gene and identified dysregulated transcripts that are likely critical either for gene expression or for chromosomal segregation and condensation, microtubule assembly, and male gametocyte activation. This further demonstrates that SMC2 and SMC4 complexes are essential for proper chromosome condensation and separation during exflagellation. Among the significantly dysregulated genes, deletion of AP2-O2 has been shown to strongly impair ookinete and oocyst development, leading to an absence of sporozoite formation and blockage of transmission (Modrzynska et al., 2017). The SET protein, which is a post-translational modification protein, is essential for parasite survival (Schwach et al., 2015). RCC is predicted to be a regulator of chromosome condensation, is essential for parasite survival, and acts as anchor for both parasite kinase (CDPK7) and phosphatase (PP1) (Lenne et al., 2018). The phenotype observed in SMC4PTD parasites may therefore reflect contributions from all of these differentially regulated genes.

The reduction in mature ookinete formation in the SMC4PTD line suggests an important role for condensin during meiosis as well. During this stage, chromosomal condensation has been observed (Sinden and Hartley, 1985), and this may be similar to the situation in *Arabidopsis*, where condensin is important in chromosomal condensation during meiosis (Smith et al., 2014). A severe defect in number and size of oocyst formation in the mosquito gut was also observed, and at this stage, multiple rounds of endomitotic division give rise to thousands of sporozoites. The process requires ten or more rounds of DNA replication, segregation, and mitotic division to create a syncytial cell (sporoblast) with thousands of nuclei over several days (Francia and Striepen, 2014, Gerald et al., 2011). The proper segregation of nuclei into individual sporoblasts is organized by putative MTOCs (Roques et al., 2019, Sinden and Strong, 1978). Because condensin has been shown to play an important role in organizing MTOCs (Kim et al., 2014), it may be that in the absence of condensin, the endomitotic division is impaired and no sporozoites are formed. Many mutants reported to cause a defect in oocyst maturation, such as *Pb*MISFIT, *Pb*CYC3, PPM5, kinesin-8X, and G actin-sequestering protein (Bushell et al., 2009, Guttery et al., 2014, Hliscs et al., 2010, Roques et al., 2015, Zeeshan et al., 2019), did not cause a significant change in the *smc4* expression profile, and there was no significant change in the expression of these genes in the present study, suggesting that the SMC4PTD mutant parasite defect in the oocyst is independent of *Pb*MISFIT, *Pb*CYC3, and *Pb*PPM5 function.

In summary, the present study shows that the condensin core subunits SMC2 and SMC4 play crucial roles in the atypical mitosis of the *Plasmodium* life cycle and may perform

distinct functions during different proliferative stages: specifically, during early schizogony, the final chromosome segregation in the last nuclear division during late schizogony, and chromosome condensation before nuclear division and exflagellation during male gametogenesis. Their removal or depletion causes impaired parasite development and blocks transmission. Additional analyses of the non-SMC components of condensin I and II will provide further insight into the function of condensin during *Plasmodium* cell proliferation.

Materials and Methods

Lead Contact and Materials Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rita Tewari (rita.tewari@nottingham.ac.uk). All materials generated in this study will be available from the Lead Contact with a completed Material Transfer Agreement (MTA).

Experimental Model and Subject Details

P. berghei ANKA line 2.34 (for GFP-tagging) or ANKA line 507cl1 (for gene deletion and promoter swap) parasites were used for transgenic line creation as described previously (Wall et al., 2018). All animal work done at the University of Nottingham has passed an ethical review process and has been approved by the United Kingdom Home Office. The work was carried out under UK Home Office Project Licenses (40/3344,30/3248 and PDD2D5182). Six to eight-week-old female Tuck-Ordinary (TO)

(Harlan) or CD1 outbred mice (Charles River) were used for all experiments performed in UK.

Infections of mice in Bern (Switzerland) were performed in accordance with the guidelines of the Swiss Tierschutzgesetz (TSchG; Animal Rights Laws) and approved by the ethical committee of the University of Bern (Permit Number: BE132/16). Female BALB/c mice (6-8 weeks; Janvier laboratories, France) were used to maintain transfected parasites and for feeding of mosquitoes with parasites.

Mice were injected via an intraperitoneal or intravenous route. When parasitemia reached 2%–5%, mice were euthanized in a CO_2 chamber and parasites isolated following exsanguination. For feeding of mosquitoes, upon reaching a parasitemia of 7%–15%, mice were anaesthetized with a terminal dose of ketamine:xylazine and when no longer reacting to touch stimulus were placed on a cage of approximately 150 mosquitoes.

Method Details

Generation of transgenic parasites

GFP-tagged vectors were designed using the p277 plasmid vector and transfected as described previously (Guttery et al., 2014). Targeted gene deletion vectors were designed using the pBS-DHFR plasmid (Tewari et al., 2010). Conditional gene knockdown constructs (SMC2PTD and SMC4PTD) were designed using P_{ama1} (*pSS368*) (Sebastian et al., 2012). *P. berghei* ANKA line 2.34 (for GFP-tagging) or ANKA line 507cl1 (for gene deletion and promoter swap) parasites were transfected by electroporation as

described previously (Wall et al., 2018). Genotypic analysis was performed using diagnostic PCR reaction and western blot. All of the oligonucleotides used to confirm genetically modified tag and mutant parasite lines can be found in Table S4. For western blotting, purified schizonts were lysed using lysis buffer (10 mM TrisHCl pH 7.5, 150 mM NaCl, 0.5 mM EDTA and 1% NP-40). The lysed samples were boiled for 10 min at 95°C after adding Laemmli buffer. The samples were centrifuged at maximum speed (13000 g) for 5 min. The samples were electrophoresed on a 4%–12% SDS-polyacrylamide gel. Subsequently, resolved proteins were transferred to nitrocellulose membrane (Amersham Biosciences). Immunoblotting experiment was performed using the Western Breeze Chemiluminescence Anti-Rabbit kit (Invitrogen) and anti-GFP polyclonal antibody (Invitrogen) at a dilution of 1:1250, according to the manufacturer's instructions.

Phenotypic Analysis and Live Cell Imaging

Phenotypic analyses of the transgenic parasite lines were performed at different points of parasite life cycle as described previously (Guttery et al., 2014). Briefly, infected blood was used to analyze asexual blood stages and gametocytes. Schizont culture was used to analyze different stages of asexual development. *In vitro* cultures were prepared to analyze activated gametocyte, exflagellation, zygote formation and ookinete development. For *in vitro* exflagellation studies, gametocyte-infected blood was obtained from the tails of infected mice using a heparinised pipette tip. Gametocyte activation was performed by mixing 100 µl of ookinete culture medium (RPMI 1640 containing 25 mM

HEPES, 20% fetal bovine serum, 10 mM sodium bicarbonate, 50 µM xanthurenic acid at pH 7.6) with the gametocyte infected blood. Microgametogenesis was monitored at two time points to study mitotic division (6 and 15 min post activation [mpa]). For mosquito transmission and bite back experiments triplicate sets of 40-50 *Anopheles stephensi* mosquitoes were used. The mosquito guts were analyzed on different days post infection (dpi); 9 dpi, 14 dpi and 21 dpi to check oocyst development and sporozoite formation. For live cell imaging, parasites were stained with Hoechst 33342 DNA stain before mounting for fluorescent microscopy. For immunofluorescence assay (IFA), the material was fixed using 2% and 4% paraformaldehyde (PFA) in microtubule stabilizing buffer (MTSB:10 mM MES, 150 mM NaCl, 5 mM EGTA, 5 mM $MgCl_2$, 5 mM glucose) for schizonts and gametocytes, respectively. Immunochemistry was performed using primary antibodies; anti-GFP rabbit antibody (Invitrogen) at 1:250 dilution, anti-alpha-tubulin mouse antibody (Sigma-Aldrich) at 1:1000 dilution, and anti-centrin mouse clone 20h5 antibody (Millipore) at 1:200 dilution. Secondary antibodies were AlexaFluor 568 labeled anti-rabbit (red) and AlexaFluor 488 labeled anti-mouse (green) (Invitrogen) (1:1000 dilution). The slides were mounted in Vectashield with DAPI (Vector Labs) for fluorescent microscopy. Parasites were visualized on a Zeiss AxioImager M2 microscope fitted with an AxioCam ICc1 digital camera (Carl Zeiss, Inc).

For liver stages, 1×10^5 HeLa cells were seeded in glass-bottomed imaging dishes. HeLa cells were grown in MEM (minimum essential medium) with Earle's salts, supplemented with 10% heat inactivated FCS (fetal calf serum), 1% penicillin/streptomycin and 1% l-glutamine (PAA Laboratories) in a humid incubator at 37°C with 5% CO_2. 24 hours after

seeding, sporozoites were isolated from parasite-infected mosquito salivary glands and used to infect seeded HeLa cells. Infected cells were maintained in 5% CO_2 at 37°C. To perform live cell imaging, Hoechst 33342 (Molecular Probes) was added (1 μg/ml) and imaging was done at 48 h and 55 h post-infection using a Leica TCS SP8 confocal microscope with the HC PL APO 63×/1.40 oil objective and the Leica Application Suite X software.

ChIP-seq and global transcriptomic analysis

For the ChIP-seq analysis, libraries were prepared from crosslinked cells (using 1% formaldehyde). The crosslinked parasite pellets were resuspended in 1 mL of nuclear extraction buffer (10 mM HEPES, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 0.5 mM AEBSF, 1X protease inhibitor tablet), post 30 min incubation on ice, 0.25% Igepal-CA-630 was added and homogenized by passing through a 26G x ½ needle. The nuclear pellet extracted through 5000 rpm centrifugation, was resuspended in 130 μl of shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris-HCl pH 7.5, 1X protease inhibitor tablet), and transferred to a 130 μl Covaris sonication microtube. The sample was then sonicated using a Covaris S220 Ultrasonicator for 10 min for schizont samples and 6 min for gametocyte samples (Duty cycle: 5%, Intensity peak power: 140, Cycles per burst: 200, Bath temperature: 6°C). The sample were transferred to ChIP dilution buffer (30 mM Tris-HCl pH 8, 3 mM EDTA, 0.1% SDS, 30 mM NaCl, 1.8% Triton X-100, 1X protease inhibitor tablet, 1X phosphatase inhibitor tablet) and centrifuged for 10 min at 13,000 rpm at 4°C, retaining the supernatant. For each sample,

13 µL of protein A agarose/salmon sperm DNA beads were washed three times with 500 µl ChIP dilution buffer (without inhibitors) by centrifuging for 1 min at 1000 rpm at room temperature, then buffer was removed. For pre-clearing, the diluted chromatin samples were added to the beads and incubated for 1 hour at 4°C with rotation, then pelleted by centrifugation for 1 min at 1000 rpm. Supernatant was removed into a LoBind tube carefully so as not to remove any beads and 2 µg of anti-GFP antibody (ab290, anti-rabbit) were added to the sample and incubated overnight at 4°C with rotation. Per sample, 25 µl of protein A agarose/salmon sperm DNA beads were washed with ChIP dilution buffer (no inhibitors), blocked with 1 mg/mL BSA for 1 hour at 4°C, then washed three more times with buffer. 25 µl of washed and blocked beads were added to the sample and incubated for 1 hour at 4°C with continuous mixing to collect the antibody/protein complex. Beads were pelleted by centrifugation for 1 min at 1000 rpm at 4°C. The bead/antibody/protein complex was then washed with rotation using 1 mL of each buffers twice; low salt immune complex wash buffer (1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 150 mM NaCl), high salt immune complex wash buffer (1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 500 mM NaCl), high salt immune complex wash buffer (1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 500 mM NaCl), TE wash buffer (10 mM Tris-HCl pH 8, 1 mM EDTA) and eluted from antibody by adding 250 µL of freshly prepared elution buffer (1% SDS, 0.1 M sodium bicarbonate). We added 5 M NaCl to the elution and cross-linking was reversed by heating at 45°C overnight followed by addition of 15 µL of 20 mg/mL RNAase A with 30 min incubation at 37°C. After this, 10 µL 0.5 M EDTA,

20 µL 1 M Tris-HCl pH 7.5, and 2 µL 20 mg/mL proteinase K were added to the elution and incubated for 2 hours at 45°C. DNA was recovered by phenol/chloroform extraction and ethanol precipitation, using a phenol/chloroform/isoamyl alcohol (25:24:1) mixture twice and chloroform once, then adding 1/10 volume of 3 M sodium acetate pH 5.2, 2 volumes of 100% ethanol, and 1/1000 volume of 20 mg/mL glycogen. Precipitation was allowed to occur overnight at −20°C. Samples were centrifuged at 13,000 rpm for 30 min at 4°C, then washed with fresh 80% ethanol, and centrifuged again for 15 min with the same settings. Pellet was air-dried and resuspended in 50 µL nuclease-free water. DNA was purified using Agencourt AMPure XP beads.

For the global transcriptome analysis, total RNA was isolated from parasite pellet using RNA extraction Kit and lyophilized. The NEB poly(A) mRNA magnetic isolation module (E7490L) was used to isolate mRNA, while the NEBNext Ultra Directional RNA Library Prep Kit (E7420L) was used to prepare a cDNA library from the isolated mRNA using manufacturer's instructions.

Libraries were prepared using the KAPA Library Preparation Kit (KAPA Biosystems), and were amplified for a total of 12 PCR cycles (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 using a NextSeq500 DNA sequencer (Illumina), producing paired-end 75-bp reads.

Immunoprecipitation and Mass Spectrometry

Schizonts, following 8 hours and 24 hours respectively in *in vitro* culture, and male gametocytes 6 min post activation were used to prepare cell lysates. Purified parasite pellets were crosslinked using formaldehyde (10 min incubation with 1% formaldehyde, followed by 5 min incubation in 0.125M glycine solution and 3 washes with phosphate buffered saline (PBS) pH7.5). Immunoprecipitation was performed using crosslinked protein and a GFP-Trap®_A Kit (Chromotek) following the manufacturer's instructions. Proteins bound to the GFP-Trap®_A beads were digested using trypsin and the peptides were analyzed by LC-MS/MS. Briefly, to prepare samples for LC-MS/MS, wash buffer was removed and ammonium bicarbonate (ABC) was added to beads at room temperature. We added 10 mM TCEP (Tris-(2-carboxyethyl) phosphine hydrochloride) and 40 mM 2-chloroacetamide (CAA) and incubation was performed for 5 min at 70°C. Samples were digested using 1 µg Trypsin per 100 µg protein at room temperature overnight followed by 1% TFA addition to bring the pH into the range of 3-4 before mass spectrometry.

Quantitative RT-PCR

RNA was isolated from different parasite life stages, which include asexual stages, purified schizonts, activated and non-activated gametocytes, ookinetes and sporozoites, using an RNA purification kit (Stratagene). cDNA was prepared using an RNA-to-cDNA kit (Applied Biosystems). Primers for qRT-PCR were designed using Primer3 (Primer-BLAST, NCBI). Gene expression was quantified from 80 ng of total cDNA. qRT-PCR

reactions used SYBR green fast master mix (Applied Biosystems) and were analyzed using an Applied Biosystems 7500 fast machine. Experiments used *hsp70* and *arginine-tRNA synthetase* as reference genes. The primers used for qRT-PCR can be found in Table S4.

Quantification and Statistical Analysis

Bioinformatics analysis

Condensin complex protein sequences were retrieved from PlasmoDB (Bahl et al., 2002), EuPathDB (Aurrecoechea et al., 2010) and from NCBI databases for model organisms (Data S1). An NCBI conserved domain database (CDD) search was used to identify conserved domains. PHYRE2 (Kelley et al., 2015) was used to generate 3D structure models. GROMACS 4.6.3 (Van Der Spoel et al., 2005) with CHARMM27 (Sapay and Tieleman, 2011) force field was used to perform molecular dynamics simulation in an aqueous environment using default parameters. The energy minimization was performed using steepest descent minimization till maximum force reached below 1000 KJ/mol/nm. Temperature (constant temperature) and pressure (constant pressure) equilibrium were done for 1 ns, respectively, before performing the 10 ns production simulation. Pymol (https://pymol.org/2/) was used to visualize 3D protein structure and grace software (https://pkgs.org/download/grace) was used to visualize protein stability. ClustalW was used to generate multiple sequence alignments of the retrieved sequences (Larkin et al., 2007). ClustalW alignment parameters included gap opening penalty (GOP) of 10 and gap extension penalty (GOE) of 0.1 for pairwise sequence alignments, and GOP of 10

and GOE of 0.2 for multiple sequence alignments, gap separation distance cut-off value of 4 and the Gonnet algorithm in protein weight matrix. Other parameters like residue-specific penalty and hydrophobic penalties were "on" whereas end gap separation and use of negative matrix were set to "off." The phylogenetic tree was inferred using the neighbor-joining method, computing the evolutionary distance using the Jones Taylor Thornton (JTT) model for amino acid substitution with the Molecular Evolutionary Genetics Analysis software (MEGA 6.0) (Tamura et al., 2013). Gaps and missing data were treated using a partial deletion method with 95% site-coverage cut-off. We performed 1000 bootstrap replicates to infer the final phylogenetic tree. For ortholog identification, NCBI BLAST and OrthoMCL database search (https://orthomcl.org/orthomcl) were performed. We applied presence of conserved domain or e-value lower than 10^{-5} for protein annotation.

ChIP-seq and global transcriptomic data analysis

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), was used to analyze raw read quality. Any adaptor sequences were removed using Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Bases with Phred quality scores below 25 were trimmed using Sickle (https://github.com/najoshi/sickle). The resulting reads were mapped against the *P. berghei* ANKA genome (v36) using Bowtie2 (version 2.3.4.1) for ChIP-seq and HISAT2 (version 2-2.1.0) for transcriptomic analysis using default parameters. Reads with a mapping quality score of 10 or higher for ChIP-seq and 50 or higher for transcriptomic analysis were retained using Samtools

(http://samtools.sourceforge.net/), and for ChIP-seq, PCR duplicates were removed by PicardTools MarkDuplicates (Broad Institute). For the transcript analysis, raw read counts were determined for each gene in the *P. berghei* genome using BedTools (https://bedtools.readthedocs.io/en/latest/#) to intersect the aligned reads with the genome annotation. BedTools was used for the ChIP-seq to obtain the read coverage per nucleotide. For the transcriptomic analysis, read counts were normalized by dividing by the total number of millions of mapped reads for the library. Genome browser tracks were generated and viewed using the Integrative Genomic Viewer (IGV) (Broad Institute). Proposed centromeric locations were obtained from Iwanaga et al. (2012). GC content was calculated using a sliding window of 30 bp across the peak region as described previously (Lynch et al., 2010). SMC2 gametocyte sample is shown at half height due to higher level of background compared to other samples. Differential expression analysis was done in two ways: (1) the use of R package DESeq2 to call up- and downregulated genes, and (2) manual analysis, in which raw read counts were normalized by library size, and genes above a threshold level of difference in normalized read counts between conditions were called as up- or downregulated. Gene ontology enrichment was done using PlasmoDB (http://plasmodb.org/plasmo/) with repetitive terms removed by REVIGO (http://revigo.irb.hr/).

Mass spectrometry analysis

Mascot (http://www.matrixscience.com/) and MaxQuant (https://www.maxquant.org/) search engines were used for mass spectrometry data analysis. PlasmoDB database was used for protein annotation. Peptide and proteins having minimum threshold of 95% were used for further proteomic analysis.

Statistical analysis of qRT-PCR data

For selected genes identified as downregulated in transcriptomic analysis statistical analysis was performed using Graph Pad Prism 7 software with unpaired t test ($^{*}p < 0.05$ $^{**}p < 0.001$, $^{***}p < 0.0001$ and $^{****}p < 0.00001$) with standard error of the mean (±SEM) deviation.

For condensin complex subunits quantification during *Plasmodium* life cycle, statistical analysis was performed using Graph Pad Prism 7 software with Standard deviation (±SD).

Statistical analysis of phenotypic data

Statistical analysis was performed using Graph Pad Prism 7 software using an unpaired t test to examine significant differences between wild-type and mutant strains for phenotypic analyses; average nuclei per schizont, exflagellation centers per field, percentage ookinete conversion, oocyst number per mosquito gut, oocyst diameter, sporozoite number per mosquito gut and salivary glands sporozoites ($^{*}p < 0.05$, $^{**}p < 0.01$

and $^{***}p < 0.001$). All experiments were performed in three independent biological replicates. Standard error of the mean (±SEM) was applied during phenotypic data analysis.

Data and Code Availability

Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number PRJNA542367. Mass spectrometry proteomic data has been deposited to the PRIDE repository with the dataset identifier PXD016833 and the original data is presented in the excel files in Supplemental Information.

Acknowledgements

We thank Prof. Frank Uhlmann, The Francis Crick Institute, for stimulating discussions and advice on condensins; Dr. Cleidiane Zampronio for assisting in mass spectrometry analysis; and Julie Rodgers for insectary assistance. This project was funded by MRC project grants and MRC Investigators grants awarded to R.T. (G0900109, G0900278, and MR/K011782/1) and BBSRC to R.T. (BB/N017609/1). R.P. was supported by MRC grant MR/K011782/1. A.A.H. was supported by the Francis Crick Institute, UK which receives its core funding from Cancer Research UK (FC001097), the UK Medical Research Council (FC001097), and the Wellcome Trust (FC001097). D.G. was supported by the Department of Biotechnology (DBT), Government of India (BT/BI/25/066/2012). K.G.L.R. was supported by the National Institute of Allergy and Infectious Diseases and

the National Institutes of Health (grants R01 AI06775 and R01 AI136511) and by the University of California, Riverside, USA (NIFA-Hatch-225935).

Author Contributions

R.T., A.A.H., and K.G.L.R. conceived and designed all experiments. R.T., R.P., S.A., M.B., R.J.W., M.Z., E.R., A.F., D.B., E.D., and S.W. performed the GFP tagging and conditional knockdown experiments. R.R.S. performed liver stage imaging. R.P., M.Z., E.D., and R.T. performed protein pull-down experiments. A.R.B. performed mass spectrometry. R.P. and D.G. performed phylogenetic analysis and MD. S.A., R.P., X.M.L., G.B., T.H., K.G.L.R., and R.T. performed RNA sequencing (RNA-seq) and ChIP-seq experiments. R.P., S.A., A.A.H., K.G.L.R., and R.T. analyzed the data. R.P., S.A., A.A.H., K.G.L.R., and R.T. wrote the manuscript, and all others contributed to it.

Declaration of Interests

The authors declare no competing interests.

References

Absalon S., Robbins J.A., Dvorin J.D. An essential malaria protein defines the architecture of blood-stage and transmission-stage parasites. Nat. Commun. 2016;7:11449. [PMC free article] [PubMed] [Google Scholar]

Anderson D.E., Losada A., Erickson H.P., Hirano T. Condensin and cohesin display different arm conformations with characteristic hinge angles. J. Cell Biol. 2002;156:419–424. [PMC free article] [PubMed] [Google Scholar]

Arnot D.E., Ronander E., Bengtsson D.C. The progression of the intra-erythrocytic cell cycle of *Plasmodium falciparum* and the role of the centriolar plaques in asynchronous mitotic division during schizogony. Int. J. Parasitol. 2011;41:71–80. [PubMed] [Google Scholar]

Aurrecoechea C., Brestelli J., Brunk B.P., Fischer S., Gajria B., Gao X., Gingle A., Grant G., Harb O.S., Heiges M. EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res. 2010;38:D415–D419. [PMC free article] [PubMed] [Google Scholar]

Ay F., Bunnik E.M., Varoquaux N., Bol S.M., Prudhomme J., Vert J.P., Noble W.S., Le Roch K.G. 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. [PMC free article] [PubMed] [Google Scholar]

Bachellier-Bassi S., Gadal O., Bourout G., Nehrbass U. Cell cycle-dependent kinetochore localization of condensin complex in *Saccharomyces cerevisiae*. J. Struct. Biol. 2008;162:248–259. [PubMed] [Google Scholar]

Bahl A., Brunk B., Coppel R.L., Crabtree J., Diskin S.J., Fraunholz M.J., Grant G.R., Gupta D., Huestis R.L., Kissinger J.C. PlasmoDB: the *Plasmodium* genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished) Nucleic Acids Res. 2002;30:87–90. [PMC free article] [PubMed] [Google Scholar]

Bunnik E.M., Cook K.B., Varoquaux N., Batugedara G., Prudhomme J., Cort A., Shi L., Andolina C., Ross L.S., Brady D. Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. Nat. Commun. 2018;9:1910. [PMC free article] [PubMed] [Google Scholar]

Bunnik E.M., Venkat A., Shao J., McGovern K.E., Batugedara G., Worth D., Prudhomme J., Lapp S.A., Andolina C., Ross L.S. Comparative 3D genome organization in apicomplexan parasites. Proc. Natl. Acad. Sci. USA. 2019;116:3183–3192. [PMC free article] [PubMed] [Google Scholar]

Bushell E.S., Ecker A., Schlegelmilch T., Goulding D., Dougan G., Sinden R.E., Christophides G.K., Kafatos F.C., Vlachou D. Paternal effect of the nuclear formin-like protein MISFIT on *Plasmodium* development in the mosquito vector. PLoS Pathog. 2009;5:e1000539. [PMC free article] [PubMed] [Google Scholar]

Bushell E., Gomes A.R., Sanderson T., Anar B., Girling G., Herd C., Metcalf T., Modrzynska K., Schwach F., Martin R.E. Functional Profiling of a Plasmodium Genome Reveals an Abundance of Essential Genes. Cell. 2017;170:260–272. [PMC free article] [PubMed] [Google Scholar]

Chang L.F., Zhang Z., Yang J., McLaughlin S.H., Barford D. Molecular architecture and mechanism of the anaphase-promoting complex. Nature. 2014;513:388–393. [PMC free article] [PubMed] [Google Scholar]

Cheeseman I.M. The kinetochore. Cold Spring Harb. Perspect. Biol. 2014;6:a015826. [PMC free article] [PubMed] [Google Scholar]

Collette K.S., Petty E.L., Golenberg N., Bembenek J.N., Csankovszki G. Different roles for Aurora B in condensin targeting during mitosis and meiosis. J. Cell Sci. 2011;124:3684–3694. [PMC free article] [PubMed] [Google Scholar]

Cuylen S., Metz J., Haering C.H. Condensin structures chromosomal DNA through topological links. Nat. Struct. Mol. Biol. 2011;18:894–901. [PubMed] [Google Scholar]

Fisher D., Krasinska L., Coudreuse D., Novák B. Phosphorylation network dynamics in the control of cell cycle transitions. J. Cell Sci. 2012;125:4703–4711. [PubMed] [Google Scholar]

Francia M.E., Striepen B. Cell division in apicomplexan parasites. Nat. Rev. Microbiol. 2014;12:125–136. [PubMed] [Google Scholar]

Francia M.E., Dubremetz J.F., Morrissette N.S. Basal body structure and composition in the apicomplexans *Toxoplasma* and *Plasmodium*. Cilia. 2016;5:3. [PMC free article] [PubMed] [Google Scholar]

Fujiwara T., Tanaka K., Kuroiwa T., Hirano T. Spatiotemporal dynamics of condensins I and II: evolutionary insights from the primitive red alga *Cyanidioschyzon merolae*. Mol. Biol. Cell. 2013;24:2515–2527. [PMC free article] [PubMed] [Google Scholar]

Gerald N., Mahajan B., Kumar S. Mitosis in the human malaria parasite *Plasmodium falciparum*. Eukaryot. Cell. 2011;10:474–482. [PMC free article] [PubMed] [Google Scholar]

Guttery D.S., Ferguson D.J., Poulin B., Xu Z., Straschil U., Klop O., Solyakov L., Sandrini S.M., Brady D., Nieduszynski C.A. A putative homologue of CDC20/CDH1 in the malaria parasite is essential for male gamete development. PLoS Pathog. 2012;8:e1002554. [PMC free article] [PubMed] [Google Scholar]

Guttery D.S., Holder A.A., Tewari R. Sexual development in *Plasmodium*: lessons from functional analyses. PLoS Pathog. 2012;8:e1002404. [PMC free article] [PubMed] [Google Scholar]

Guttery D.S., Poulin B., Ramaprasad A., Wall R.J., Ferguson D.J., Brady D., Patzewitz E.M., Whipple S., Straschil U., Wright M.H. Genome-wide functional analysis of *Plasmodium* protein phosphatases reveals key regulators of parasite development and differentiation. Cell Host Microbe. 2014;16:128–140. [PMC free article] [PubMed] [Google Scholar]

Güttinger S., Laurell E., Kutay U. Orchestrating nuclear envelope disassembly and reassembly during mitosis. Nat. Rev. Mol. Cell Biol. 2009;10:178–191. [PubMed] [Google Scholar]

Hammarton T.C. Cell cycle regulation in *Trypanosoma brucei*. Mol. Biochem. Parasitol. 2007;153:1–8. [PMC free article] [PubMed] [Google Scholar]

Harashima H., Dissmeyer N., Schnittger A. Cell cycle control across the eukaryotic kingdom. Trends Cell Biol. 2013;23:345–356. [PubMed] [Google Scholar]

Hirano T. Condensin-Based Chromosome Organization from Bacteria to Vertebrates. Cell. 2016;164:847–857. [PubMed] [Google Scholar]

Hirota T., Gerlich D., Koch B., Ellenberg J., Peters J.M. Distinct functions of condensin I and II in mitotic chromosome assembly. J. Cell Sci. 2004;117:6435–6445. [PubMed] [Google Scholar]

Hliscs M., Sattler J.M., Tempel W., Artz J.D., Dong A., Hui R., Matuschewski K., Schüler H. Structure and function of a G-actin sequestering protein with a vital role in malaria oocyst development inside the mosquito vector. J. Biol. Chem. 2010;285:11572–11583. [PMC free article] [PubMed] [Google Scholar]

Howard-Till R., Loidl J. Condensins promote chromosome individualization and segregation during mitosis, meiosis, and amitosis in *Tetrahymena thermophila*. Mol. Biol. Cell. 2018;29:466–478. [PMC free article] [PubMed] [Google Scholar]

Howard-Till R., Tian M., Loidl J. A specialized condensin complex participates in somatic nuclear maturation in Tetrahymena thermophila. Mol. Biol. Cell. 2019;30:1326–1338. [PMC free article] [PubMed] [Google Scholar]

Iwanaga S., Kato T., Kaneko I., Yuda M. Centromere plasmid: a new genetic tool for the study of *Plasmodium falciparum*. PLoS ONE. 2012;7:e33326. [PMC free article] [PubMed] [Google Scholar]

Iwasaki O., Noma K.I. Condensin-mediated chromosome organization in fission yeast. Curr. Genet. 2016;62:739–743. [PMC free article] [PubMed] [Google Scholar]

Kelley L.A., Mezulis S., Yates C.M., Wass M.N., Sternberg M.J. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 2015;10:845–858. [PMC free article] [PubMed] [Google Scholar]

Kim J.H., Shim J., Ji M.J., Jung Y., Bong S.M., Jang Y.J., Yoon E.K., Lee S.J., Kim K.G., Kim Y.H. The condensin component NCAPG2 regulates microtubule-kinetochore attachment through recruitment of Polo-like kinase 1 to kinetochores. Nat. Commun. 2014;5:4588. [PubMed] [Google Scholar]

Kim K.D., Tanizawa H., Iwasaki O., Noma K. Transcription factors mediate condensin recruitment and global chromosomal organization in fission yeast. Nat. Genet. 2016;48:1242–1252. [PMC free article] [PubMed] [Google Scholar]

Kschonsak M., Merkel F., Bisht S., Metz J., Rybin V., Hassler M., Haering C.H. Structural Basis for a Safety-Belt Mechanism That Anchors Condensin to Chromosomes. Cell. 2017;171:588–600. [PMC free article] [PubMed] [Google Scholar]

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–2948. [PubMed] [Google Scholar]

Lenne A., De Witte C., Tellier G., Hollin T., Aliouat E.M., Martoriati A., Cailliau K., Saliou J.M., Khalife J., Pierrot C. Characterization of a Protein Phosphatase Type-1 and a Kinase Anchoring Protein in *Plasmodium falciparum*. Front. Microbiol. 2018;9:2617. [PMC free article] [PubMed] [Google Scholar]

Lynch D.B., Logue M.E., Butler G., Wolfe K.H. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol. Evol. 2010;2:572–583. [PMC free article] [PubMed] [Google Scholar]

McKinley K.L., Cheeseman I.M. The molecular basis for centromere identity and function. Nat. Rev. Mol. Cell Biol. 2016;17:16–29. [PMC free article] [PubMed] [Google Scholar]

Modrzynska K., Pfander C., Chappell L., Yu L., Suarez C., Dundas K., Gomes A.R., Goulding D., Rayner J.C., Choudhary J., Billker O. A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the *Plasmodium* Life Cycle. Cell Host Microbe. 2017;21:11–22. [PMC free article] [PubMed] [Google Scholar]

Musacchio A., Desai A. A Molecular View of Kinetochore Assembly and Function. Biology (Basel) 2017;6:E5. [PMC free article] [PubMed] [Google Scholar]

Neuwald A.F., Hirano T. HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. Genome Res. 2000;10:1445–1452. [PMC free article] [PubMed] [Google Scholar]

Oliveira R.A., Heidmann S., Sunkel C.E. Condensin I binds chromatin early in prophase and displays a highly dynamic association with *Drosophila* mitotic chromosomes. Chromosoma. 2007;116:259–274. [PubMed] [Google Scholar]

Onn I., Aono N., Hirano M., Hirano T. Reconstitution and subunit geometry of human condensin complexes. EMBO J. 2007;26:1024–1034. [PMC free article] [PubMed] [Google Scholar]

Ono T., Fang Y., Spector D.L., Hirano T. Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. Mol. Biol. Cell. 2004;15:3296–3308. [PMC free article] [PubMed] [Google Scholar]

Ono T., Yamashita D., Hirano T. Condensin II initiates sister chromatid resolution during S phase. J. Cell Biol. 2013;200:429–441. [PMC free article] [PubMed] [Google Scholar]

Pandey R., Zeeshan M., Ferguson D.J.P., Markus R., Brady D., Daniel E., Stanway R.R., Holder A.A., Guttery D.S., Tewari R. Real-time dynamics of *Plasmodium* NDC80 as a marker for the kinetochore during atypical mitosis and meiosis. bioRxiv. 2019 [Google Scholar]

Rawlings J.S., Gatzka M., Thomas P.G., Ihle J.N. Chromatin condensation via the condensin II complex is required for peripheral T-cell quiescence. EMBO J. 2011;30:263–276. [PMC free article] [PubMed] [Google Scholar]

Roques M., Wall R.J., Douglass A.P., Ramaprasad A., Ferguson D.J., Kaindama M.L., Brusini L., Joshi N., Rchiad Z., Brady D. *Plasmodium* P-Type Cyclin CYC3 Modulates Endomitotic Growth during Oocyst Development in Mosquitoes. PLoS Pathog. 2015;11:e1005273. [PMC free article] [PubMed] [Google Scholar]

Roques M., Stanway R.R., Rea E.I., Markus R., Brady D., Holder A.A., Guttery D.S., Tewari R. *Plasmodium* centrin *Pb*CEN-4 localizes to the putative MTOC and is dispensable for malaria parasite proliferation. Biol. Open. 2019;8:bio036822. [PMC free article] [PubMed] [Google Scholar]

Sapay N., Tieleman D.P. Combination of the CHARMM27 force field with united-atom lipid force fields. J. Comput. Chem. 2011;32:1400–1410. [PubMed] [Google Scholar]

Sazer S., Lynch M., Needleman D. Deciphering the evolutionary history of open and closed mitosis. Curr. Biol. 2014;24:R1099–R1103. [PMC free article] [PubMed] [Google Scholar]

Schleiffer A., Kaitna S., Maurer-Stroh S., Glotzer M., Nasmyth K., Eisenhaber F. Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. Mol. Cell. 2003;11:571–575. [PubMed] [Google Scholar]

Schwach F., Bushell E., Gomes A.R., Anar B., Girling G., Herd C., Rayner J.C., Billker O. PlasmoGEM, a database supporting a community resource for large-scale experimental genetics in malaria parasites. Nucleic Acids Res. 2015;43:D1176–D1182. [PMC free article] [PubMed] [Google Scholar]

Sebastian S., Brochet M., Collins M.O., Schwach F., Jones M.L., Goulding D., Rayner J.C., Choudhary J.S., Billker O. A *Plasmodium* calcium-dependent protein kinase controls zygote development and transmission by translationally activating repressed mRNAs. Cell Host Microbe. 2012;12:9–19. [PMC free article] [PubMed] [Google Scholar]

Sinden R.E. Asexual blood stages of malaria modulate gametocyte infectivity to the mosquito vector—possible implications for control strategies. Parasitology. 1991;103:191–196. [PubMed] [Google Scholar]

Sinden R.E. Mitosis and meiosis in malarial parasites. Acta Leiden. 1991;60:19–27. [PubMed] [Google Scholar]

Sinden R.E., Hartley R.H. Identification of the meiotic division of malarial parasites. J. Protozool. 1985;32:742–744. [PubMed] [Google Scholar]

Sinden R.E., Strong K. An ultrastructural study of the sporogonic development of *Plasmodium falciparum* in *Anopheles gambiae*. Trans. R. Soc. Trop. Med. Hyg. 1978;72:477–491. [PubMed] [Google Scholar]

Sinden R.E., Canning E.U., Spain B. Gametogenesis and fertilization in *Plasmodium yoelii nigeriensis*: a transmission electron microscope study. Proc. R. Soc. Lond. B Biol. Sci. 1976;193:55–76. [PubMed] [Google Scholar]

Sinden R.E., Talman A., Marques S.R., Wass M.N., Sternberg M.J. The flagellum in malarial parasites. Curr. Opin. Microbiol. 2010;13:491–500. [PubMed] [Google Scholar]

Smith S.J., Osman K., Franklin F.C. The condensin complexes play distinct roles to ensure normal chromosome morphogenesis during meiotic division in *Arabidopsis*. Plant J. 2014;80:255–268. [PMC free article] [PubMed] [Google Scholar]

Sutani T., Yuasa T., Tomonaga T., Dohmae N., Takio K., Yanagida M. Fission yeast condensin complex: essential roles of non-SMC subunits for condensation and Cdc2 phosphorylation of Cut3/SMC4. Genes Dev. 1999;13:2271–2283. [PMC free article] [PubMed] [Google Scholar]

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 2013;30:2725–2729. [PMC free article] [PubMed] [Google Scholar]

Tewari R., Straschil U., Bateman A., Böhme U., Cherevach I., Gong P., Pain A., Billker O. The systematic functional analysis of *Plasmodium* protein kinases identifies essential regulators of mosquito transmission. Cell Host Microbe. 2010;8:377–387. [PMC free article] [PubMed] [Google Scholar]

Thadani R., Uhlmann F., Heeger S. Condensin, chromatin crossbarring and chromosome condensation. Curr. Biol. 2012;22:R1012–R1021. [PubMed] [Google Scholar]

Tůmová P., Uzlíková M., Wanner G., Nohýnková E. Structural organization of very small chromosomes: study on a single-celled evolutionary distant eukaryote *Giardia intestinalis*. Chromosoma. 2015;124:81–94. [PubMed] [Google Scholar]

Uhlmann F. SMC complexes: from DNA to chromosomes. Nat. Rev. Mol. Cell Biol. 2016;17:399–412. [PubMed] [Google Scholar]

Van Der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A.E., Berendsen H.J. GROMACS: fast, flexible, and free. J. Comput. Chem. 2005;26:1701–1718. [PubMed] [Google Scholar]

Wall R.J., Ferguson D.J.P., Freville A., Franke-Fayard B., Brady D., Zeeshan M., Bottrill A.R., Wheatley S., Fry A.M., Janse C.J. *Plasmodium* APC3 mediates chromosome condensation and cytokinesis during atypical mitosis in male gametogenesis. Sci. Rep. 2018;8:5610. [PMC free article] [PubMed] [Google Scholar]

Ward P., Equinet L., Packer J., Doerig C. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. BMC Genomics. 2004;5:79. [PMC free article] [PubMed] [Google Scholar]

Wilkes J.M., Doerig C. The protein-phosphatome of the human malaria parasite *Plasmodium falciparum*. BMC Genomics. 2008;9:412. [PMC free article] [PubMed] [Google Scholar]

Yuen K.C., Slaughter B.D., Gerton J.L. Condensin II is anchored by TFIIIC and H3K4me3 in the mammalian genome and supports the expression of active dense gene clusters. Sci. Adv. 2017;3:e1700191. [PMC free article] [PubMed] [Google Scholar]

Zeeshan M., Shilliday F., Liu T., Abel S., Mourier T., Ferguson D.J.P., Rea E., Stanway R.R., Roques M., Williams D. *Plasmodium* kinesin-8X associates with mitotic spindles and is essential for oocyst development during parasite proliferation and transmission. PLoS Pathog. 2019;15:e1008048. [PMC free article] [PubMed] [Google Scholar]

Supplementary Information

Supplementary material for chapter 3 is available with the published paper at

https://www.cell.com/cell-reports/fulltext/S2211-1247(20)30048-6.

Chapter 4: Proteome-Wide Identification of RNA-Dependent Proteins: An Emerging Role for RNAs in *Plasmodium falciparum* Protein Complexes

Thomas Hollin[1], Steven Abel[1], Charles Banks[2], Jacques Prudhomme[1], Laurence Florens[2], William Stafford Noble[3] and Karine G. Le Roch[1]

[1]Department of Molecular, Cell and Systems Biology, University of California Riverside, CA, USA
[2]Stowers Institute for Medical Research, Kansas City, MO, USA
[3]Department of Genome Sciences, University of Washington, Seattle, WA, USA

A version of this chapter has been submitted to *Nature Communications*, 2023.

Preface

Post-transcriptional regulation is another important method of gene regulation in eukaryotes, and another that, along with epigenetics, may be particularly important in *Plasmodium falciparum* due to its paucity of transcription factors. RNA-binding proteins (RBPs) are traditionally thought of as the most important actors in this type of regulation, but increasing evidence suggests that other types of proteins, the RNA-dependent proteins or RDPs that are not directly binding to RNA, may be involved as well. These RDPs may play key function to some critical ribonucleoprotein (RNP) complexes. While most likely important for a number of critical biological functions that include transcription, translation, regulation of gene expression and RNA metabolism, the role of these RDPs have yet to be explored . To start investigating the role of RBPs and RDP in the human malaria parasite, we adapted the R-DeeP protocol, previously done for human cell lines, in *P. falciparum*. R-DeeP, based on sucrose density ultracentrifugation, can screen the entire parasite proteome for RNA-dependent proteins and examine RNA-dependent complexes. We also showed how this type of data can be a starting point for characterizing individual proteins at the functional level. We indeed validated a candidate protein as a true RNA-binding protein and showed exactly which RNAs, and exactly where in the RNA, the RBP binds. I performed the experimental work in this project along with one other colleague, as it is a complex, long, and demanding protocol which depends on extracting total parasite protein and performing all experimental steps before RNA in the Control (no RNase) sample can be degraded. I also assisted in developing the computational approach to analyze the full proteome data, determining whether a “shift”

in sucrose gradient fractions was sufficient enough to call a protein an RNA-dependent protein.

Abstract

Ribonucleoprotein complexes are composed of RNA, RNA-dependent proteins (RDPs) and RNA-binding proteins (RBPs), and play fundamental roles for RNA regulation. However, in the human malaria parasite, *Plasmodium falciparum*, identification and characterization of these proteins are particularly limited. In this study, we use an unbiased proteome-wide approach, R-DeeP, based on sucrose density gradient ultracentrifugation to identify RDPs. Quantitative analysis by mass spectrometry identified 785 RDPs, including 463 proteins newly associated with RNA. This method provides a snapshot of the protein-protein network in presence and absence of RNA. R-DeeP also contributes to reconstruct *Plasmodium* multiprotein complexes based on co-segregation and deciphers their RNA-dependence. Finally, one RDP candidate, PF3D7_0823200, was further characterized and validated as a true RBP. Using enhanced crosslinking and immunoprecipitation followed by high-throughput sequencing (eCLIP-seq), this protein was detected as a regulator of non-coding regions of various plasmodial transcripts including *var* and *ap2* transcription factors.

Introduction

Ribonucleoprotein (RNP) complexes are critical post-transcriptional regulators of gene expression covering all aspects of RNA activity in eukaryotes such as export, splicing, stability, translation and degradation. These RNPs are assemblies of RNA molecules and proteins including RNA-binding proteins (RBPs), and their compositions are highly dynamic to allow adaption to cellular needs and environmental conditions[1]. RNPs contain RNAs ranging from messenger RNAs (mRNAs) to non-coding RNAs (ncRNAs) such as long ncRNAs (lncRNAs), rRNAs and tRNAs. These RNAs can interact with RBPs and act in (post-)transcriptional and (post-)translational regulation, modulate the structures and stability of RNP complexes, or serve as protein decoys. The advent of lncRNA research highlighted the involvement of these transcripts in post-transcriptional control. Recently, lncRNAs have been shown to be preponderant mediators in RNP complexes[2], and elucidating the composition and function of such complexes is a current challenge in RNA biology.

On the other hand, RBPs are also essential components of RNP complexes. They bind RNA through one or multiple RNA-binding domains (RBDs) such as the RNA recognition motif (RRM), K homology (KH), zinc finger, Pumilio (Puf) and DEAD box helicase domains[3]. In humans, 2000-3000 RBPs have been identified through several RNA interactome studies, while around 1000 are annotated in various model organisms such as *Mus musculus* and *Saccharomyces cerevisiae*[4,5]. In *Plasmodium falciparum*, the deadliest human malaria parasite, RBPs also regulate a wide range of essential processes, but our knowledge gaps in identifying and characterizing their role in the parasite represent a

critical roadblock in the fight against malaria. To date, there are only two *in silico* approaches and one mRNA interactome capture dataset available in *P. falciparum*. The first bioinformatics analysis was published in 2015 and reported 189 putative RBPs[6]. These proteins belonged to 13 RBP families, including some of the most prominent, such as RRM, KH and zinc finger domain. The RBP repertoire of *P. falciparum* was then expanded with a hidden Markov model (HMM) search using 793 RNA-related domains[7]. A total of 988 RBPs were identified and corresponded to 18.1 % of the parasite proteome. This study also included an experimental capture of mRNA-binding proteins (mRBPs) using oligo d(T) beads followed by mass spectrometry identification. The authors captured 199 candidate mRBPs, with an enrichment of RRM, DEAD, LSm and Alba domains.

These two proteome-wide studies have provided the most complete core set of RBPs in *P. falciparum*. However, the discoveries of unconventional RBPs, generally lacking canonical RBDs, highlighted the great challenges to comprehensively identify RBPs within eukaryotic proteomes[5]. Development of unbiased and complementary RNA interactome approaches are therefore necessary to elucidate the RBP repertoire and composition of RNP complexes to facilitate our understanding of their biological functions. Recently, a quantitative proteome-wide screen based on density gradient ultracentrifugation (R-DeeP) was developed to identify RBPs as well as RNA-dependent proteins (RDPs) within RNP complexes. RDPs are proteins that do not bind directly to RNA but interact with other RDPs or RBPs and are crucial components of RNP complexes. Applying this approach, the authors identified 1,784 RDPs, 537 of which had never been associated with RNA in human HeLa cells[8]. More recently, a total of 1189 candidates, including 170 unknown

RDPs, were detected in A549 lung adenocarcinoma cells[9]. Here, we applied the R-DeeP method to *P. falciparum* and identified 785 RDP candidates, including novel and uncharacterized proteins. Furthermore, we demonstrated that this approach can be used to interpret *Plasmodium* complexes and protein clusters, and identify RNP complexes. Finally, we experimentally characterized PF3D7_0823200 protein using complementary approaches including high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (eCLIP-seq), and validating this protein as a novel RNA stabilization and/or splicing factor.

Results

Identification of RNA-dependent proteins using R-DeeP

The R-DeeP method is based on the separation of proteins on a sucrose density gradient in the presence (= Control) or absence (= RNase) of RNA (Fig. 4.1a). After ultracentrifugation and fractionation, 25 fractions were collected and processed by mass spectrometry or western blot analysis. The separation of proteins or multiprotein complexes depends on their respective molecular weights (MW), with larger proteins or complexes found in higher density fractions. For RNA-dependent proteins (RDPs), the RNase treatment may impact their interactome and thus result in a shift to lower fractions. In this study, we extracted mid-late trophozoites by saponin lysis and prepared soluble protein extracts in RNase-free conditions (Fig. 4.1a). After DNase I treatment (for Control samples) or DNase I and RNase A/H/I treatment (for RNase samples) (Supplementary Fig.

4.1a), 2-2.5 mg of proteins were loaded onto a 5% to 50% sucrose density gradient. After ultracentrifugation, 25 fractions were collected for each condition and analyzed by mass spectrometry or western blot.

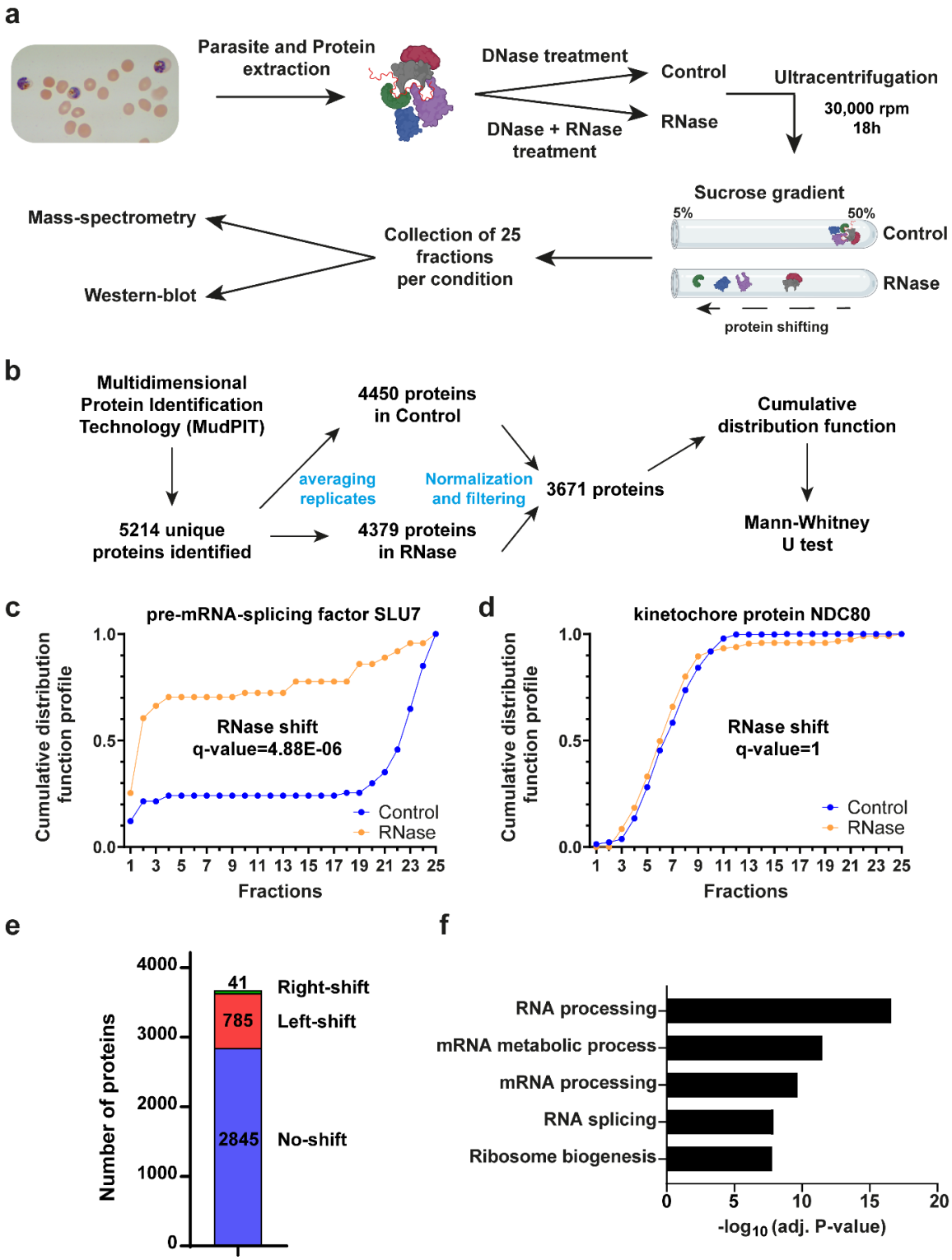
a
Parasite and Protein extraction
DNase treatment
Control
RNase
DNase + RNase treatment
Ultracentrifugation
30,000 rpm
18h
Sucrose gradient
5%
50%
Control
RNase
protein shifting
Collection of 25 fractions per condition
Mass-spectrometry
Western-blot
b
Multidimensional Protein Identification Technology (MudPIT)
5214 unique proteins identified
averaging replicates
4450 proteins in Control
4379 proteins in RNase
Normalization and filtering
3671 proteins
Cumulative distribution function
Mann-Whitney U test
c
pre-mRNA-splicing factor SLU7
Cumulative distribution function profile
RNase shift
q-value=4.88E-06
Control
RNase
Fractions
d
kinetochore protein NDC80
RNase shift
q-value=1
e
Number of proteins
41
785
2845
Right-shift
Left-shift
No-shift
f
RNA processing
mRNA metabolic process
mRNA processing
RNA splicing
Ribosome biogenesis
-log10 (adj. P-value)

Fig. 4.1 R-DeeP approach to identify RNA-dependent proteins in *P. falciparum*. a. Schematic overview of the R-DeeP method. NF54 parasite protein lysates were treated with DNase (Control) or DNase + RNases (RNase) and loaded on a sucrose gradient. After ultracentrifugation, 25 fractions were collected and further processed by mass spectrometry and western blot analysis. b. Bioinformatics workflow for the mass spectrometry data analysis. After multiple filtering (see Methods), a final list of 3671 proteins was obtained and a cumulative distribution function (CDF) was calculated for each protein. CDF profiles of pre-mRNA-splicing factor SLU7 (c) and kinetochore protein NDC80 (d) illustrate an RNase-shifted and non-RNase-shifted protein, respectively. e. The graph shows the number of left-shifted, right-shifted and non-shifted proteins detected in this R-DeeP. f. GO enrichment analysis of the 785 left-shifted proteins. The significance of Biological Process terms is shown by $-\log_{10}$ (adjusted P-value) (Fisher's exact test with Bonferroni adjustment).

First, we generated Control and RNase samples in duplicates and quantified the protein abundance using MudPIT mass spectrometry (Fig. 4.1b). A good coverage of the *Plasmodium* proteome was obtained, with identification of 5214 unique proteins. After filtering and normalization (see Methods), we generated a final list of 3671 proteins, reproducibly detected in each sample (Supplementary Data 4.1). To identify RDPs, we calculated the cumulative distribution function (CDF) for each protein's abundance across the 25 fractions and used a Wilcoxon rank-sum test to detect proteins that exhibit a statistically significant shift. A total of 785 unique proteins were identified as significantly left-shifted, suggesting that their interactions are RNA-dependent (Fig. 4.1c and Supplementary Data 4.1). Additionally, 41 proteins were detected as right-shifted, but no particular pathway seemed to be associated with them (Supplementary Data 4.1). These proteins may have interacted with newly accessible partners in the absence of RNA. For the 785 left-shifted proteins, Gene Ontology (GO) enrichment analysis showed a strong enrichment for diverse RNA pathways such as mRNA processing, mRNA metabolic process, RNA splicing, and ribosome biogenesis (Fig. 4.1d), confirming the robustness of our R-DeeP experiment.

We next generated a list of proteins already defined as RNA-associated proteins, including RNA-binding proteins (RBPs), using PlasmoDB's GO resource (see details in Methods), two *in silico* datasets[6,7] and an mRNA interactome capture experiment[7]. A collection of 1319 unique RNA-associated proteins were obtained (Supplementary Data 4.2) and compared to our R-DeeP list. A total of 41% (322/785) of our shifted proteins were previously associated with RNA, which included 23 ribosomal proteins, 17 RNA helicases

and 7 LSm proteins as well as poorly characterized proteins with 32 conserved proteins and 21 putative RBPs (Fig. 4.2a). For the 463 unshared, R-DeeP only proteins, we identified 172 uncharacterized proteins of which 132 are annotated as *Plasmodium* specific. As the R-DeeP screen does not only identify classical RBPs but also unconventional RBPs and RDPs, we suggest that some of these proteins could be RDPs or unknown RBPs and may require further characterization. This cluster also contains six heptatricopeptide repeat (HPR) proteins which are related to pentatricopeptide repeat (PPR) proteins, a well-known RBP family in land plants. In *P. berghei*, one HPR protein, PbHPR1, bound in vitro to mitochondrial RNAs, suggesting that these proteins are bona fide RBPs[10]. Interestingly, 13 AP2 transcription factors were also significantly shifted, indicating that the stability of their respective complex may depend on transcriptional activity. Among the RNA-associated complexes depleted in our experiment, we noticed an enrichment of 142 ribosomal proteins, 30 (eukaryotic) translation initiation factors, 21 pre-mRNA splicing factors, 14 elongation factors and 9 exosome components (Fig. 4.2a). These data suggest that these particular protein complexes, although involved in RNA processes, are not RNP complexes or RNA-dependent for their formation and/or stability, at least in our experimental conditions.

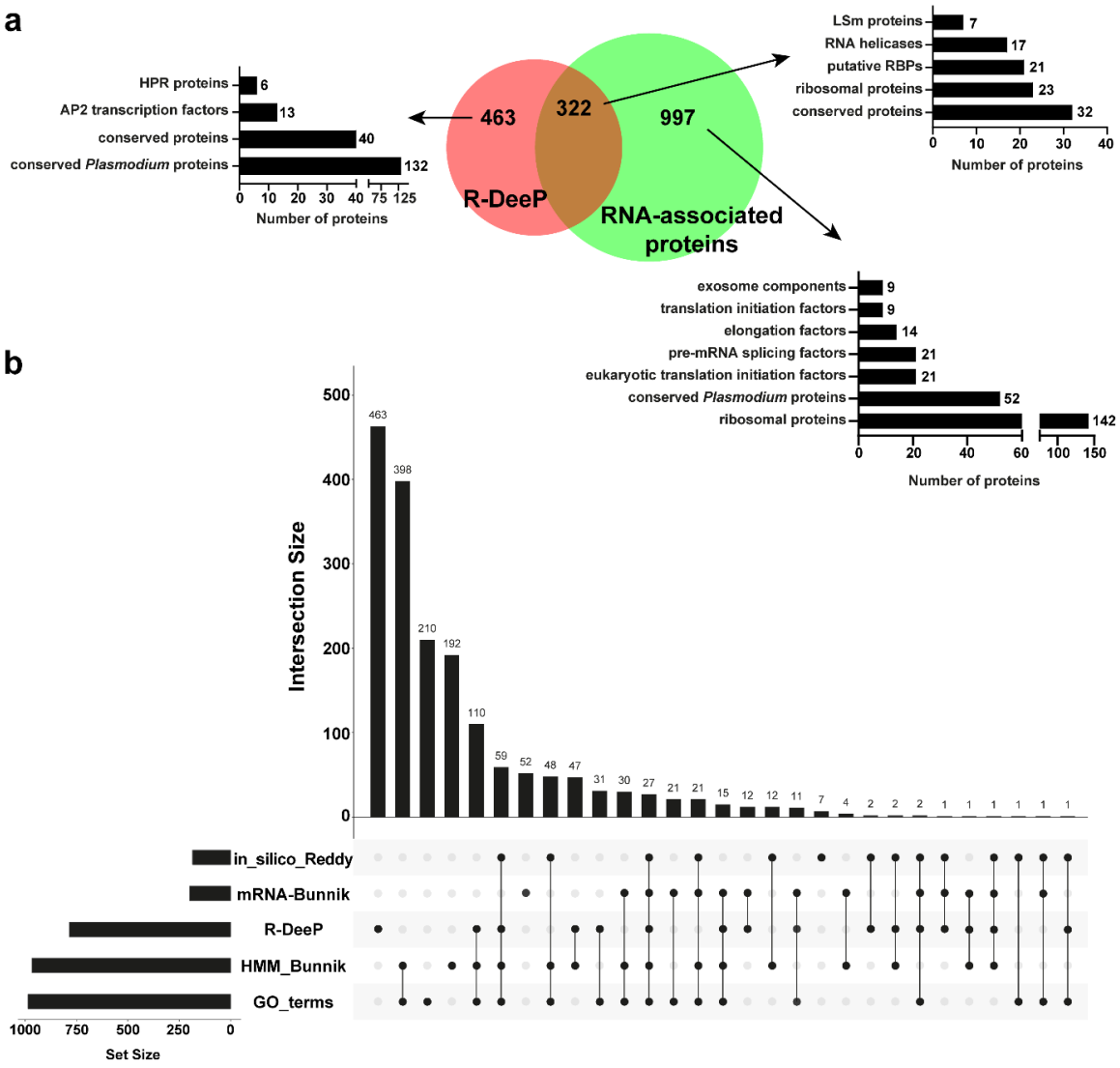

Fig. 4.2 Comparison of the significant left-shifted proteins. a. Venn diagram reporting the number of proteins overlapping with a custom list of known RNA-associated proteins (see Methods). For each cluster, some of the most represented protein families/groups are indicated. b. UpSet plot summarizing the number of unique and shared RNA-dependent proteins between different datasets (in_silico_Reddy[6]; mRNA_Bunnik [7]; R-DeeP from this study; HMM_Bunnik[7]; GO_terms from PlasmoDB).

Detailed comparisons showed that among the 322 RNA-associated proteins shared between R-DeeP and other datasets, 230 proteins (71%) were previously identified in at least two other datasets, confirming the robustness of our R-DeeP experiment (Fig. 4.2b). Additionally, 27 proteins were detected in all datasets and correspond to, among others, well-known RBPs such as RNA helicases DDX6 and PRP22, Alba 2 and 4, CUGBP Elav-like family member 1 and 2, and polyadenylate-binding protein 3.

Validation of the R-DeeP screening

To further validate our R-DeeP results, we analyzed the profiles of some proteins using an independent R-DeeP replicate by western blot analysis. As the availability of commercial antibodies is limited in *P. falciparum*, we produced eight different custom rabbit polyclonal antibodies. We selected peptides targeting six significantly shifted proteins, including two unknown *Plasmodium* proteins (PF3D7_0528600 and PF3D7_1354900), two putative RBPs (PF3D7_1360100 and PF3D7_0823200) and two characterized RBPs (Musashi (PF3D7_0916700) and Alba 4 (PF3D7_1347500)) (Supplementary Data 4.3). Proteasome subunit alpha type-7 (PSA 7, PF3D7_1353900) and cytochrome c oxidase subunit 6A (COX6A, PF3D7_1465000) were chosen as negative controls, as well as fructose-bisphosphate aldolase (PF3D7_1444800), for which a commercial antibody is available. The reactivity of these antibodies was tested on total parasite protein lysates, and only PF3D7_0823200, PF3D7_1347500 and PF3D7_1353900 successfully showed specific recognition (Supplementary Fig. 4.1b). Immunoblots containing all the 50 R-DeeP fractions were carried out using these different antibodies as well as the anti-Aldolase.

Protein signals were normalized for both conditions to obtain relative abundance and then matched to the mass spectrometry distribution profiles. As shown in Fig. 4.3, we validated the shifting of PF3D7_0823200, while the aldolase immunoblots reflected our mass spectrometry results, thus supporting the conclusion that this is an RNA-independent protein. Similarly, immunoblots confirmed Alba 4 and PSA 7 as RNA-binding and RNA-independent proteins, respectively (Supplementary Fig. 4.1c and 4.1d). Interestingly, a difference of migration was observed for PF3D7_0823200 between Control (~ 30 kDa) and RNase (~ 45 kDa) conditions. This discrepancy might be attributable to post-translational modifications only present when the protein is part of its RNP complex, but this requires further investigation.

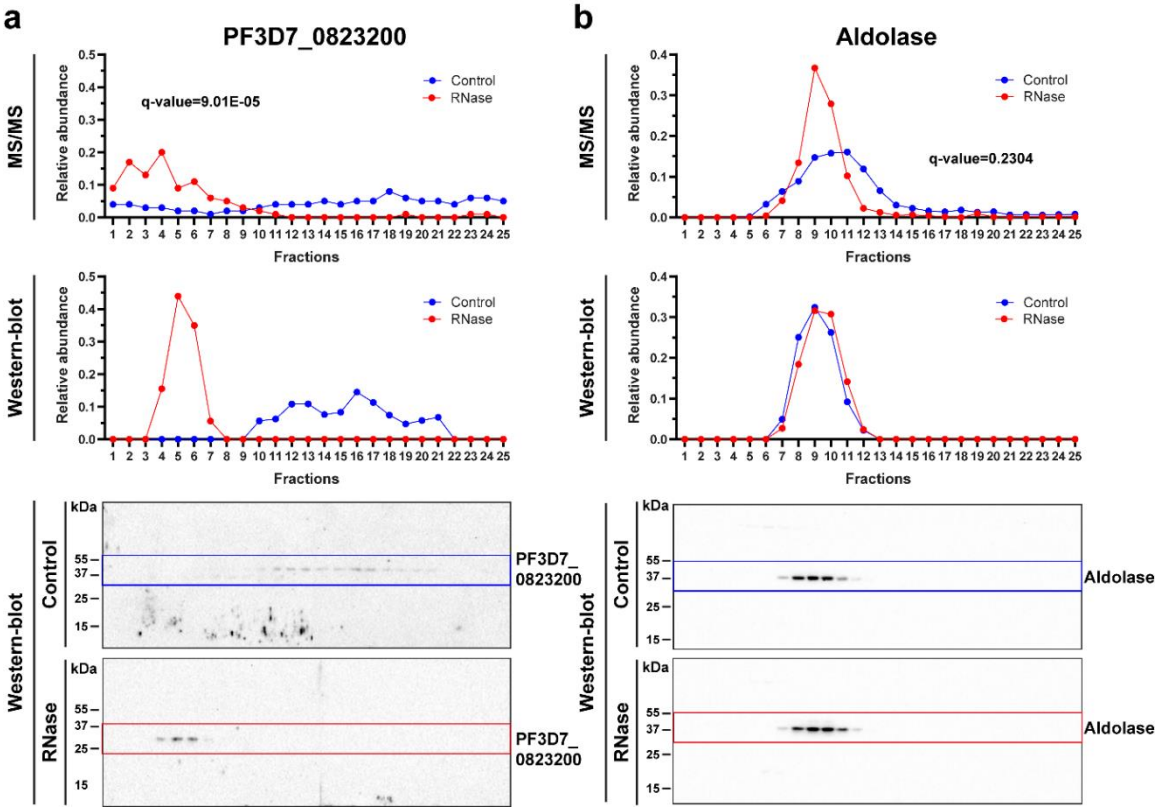

Fig. 4.3 Validation of the R-DeeP protocol by western blot analysis. The mass spectrometry (MS/MS) data (top panel) were compared to the quantitative analyses of immunoblots (center panel) obtained with anti-PF3D7_0823200 (a) and anti-Aldolase (b) (bottom panel).

The fate of protein-protein interaction networks in the presence and absence of RNA

With the R-Deep methodology, the position of each protein in the sucrose gradient is determined by its respective MW, structure and interactome. Based on this information, we determined the network status of each protein in the Control and RNase samples. To do this, we calculated the halfway value indicating at which fraction 50% of the total amount of each protein was detected (CDF = 0.5). Using a previous R-DeeP calibration generated with human reference proteins (RNase A, BSA, Aldolase, Catalase and Ferritin)[8] and these halfway values, we were able to determine an apparent MW for all proteins. Then, the ratio

between apparent and theoretical MW was used to classify the proteins according to their molecular state. Proteins appearing to be smaller, identical, or larger than their theoretical MW are indicated as ‘smaller’, ‘monomeric’ and ‘larger’, respectively.

For the Control condition, we identified 2797 proteins (76.2%) with a higher apparent than theoretical molecular weight (ratio > 2) suggesting that they were in complex (Fig. 4.4a). Only 269 (7.3%) and 602 (16.2%) proteins were detected as smaller and monomeric, respectively. By contrast, in the RNase condition, we observed 634 (17.3%) smaller and 694 (18.9%) monomeric proteins (Fig. 4.4b). A clear shift can be noticed with 2343 (63.8%) larger proteins, instead of 76.2% in Control, a result similar to that obtained with HeLa cells (61%)[8]. GO enrichment analysis confirmed that non-complexed proteins (smaller and monomeric) are mainly involved in nucleic acid binding and mRNA/RNA binding (Fig. 4.4c). On the other hand, proteins categorized as still being in complex were mainly associated with ribosomes and cation transmembrane transporters such as V-type proton ATPases, ATP synthases and cytochromes (Fig. 4.4d).

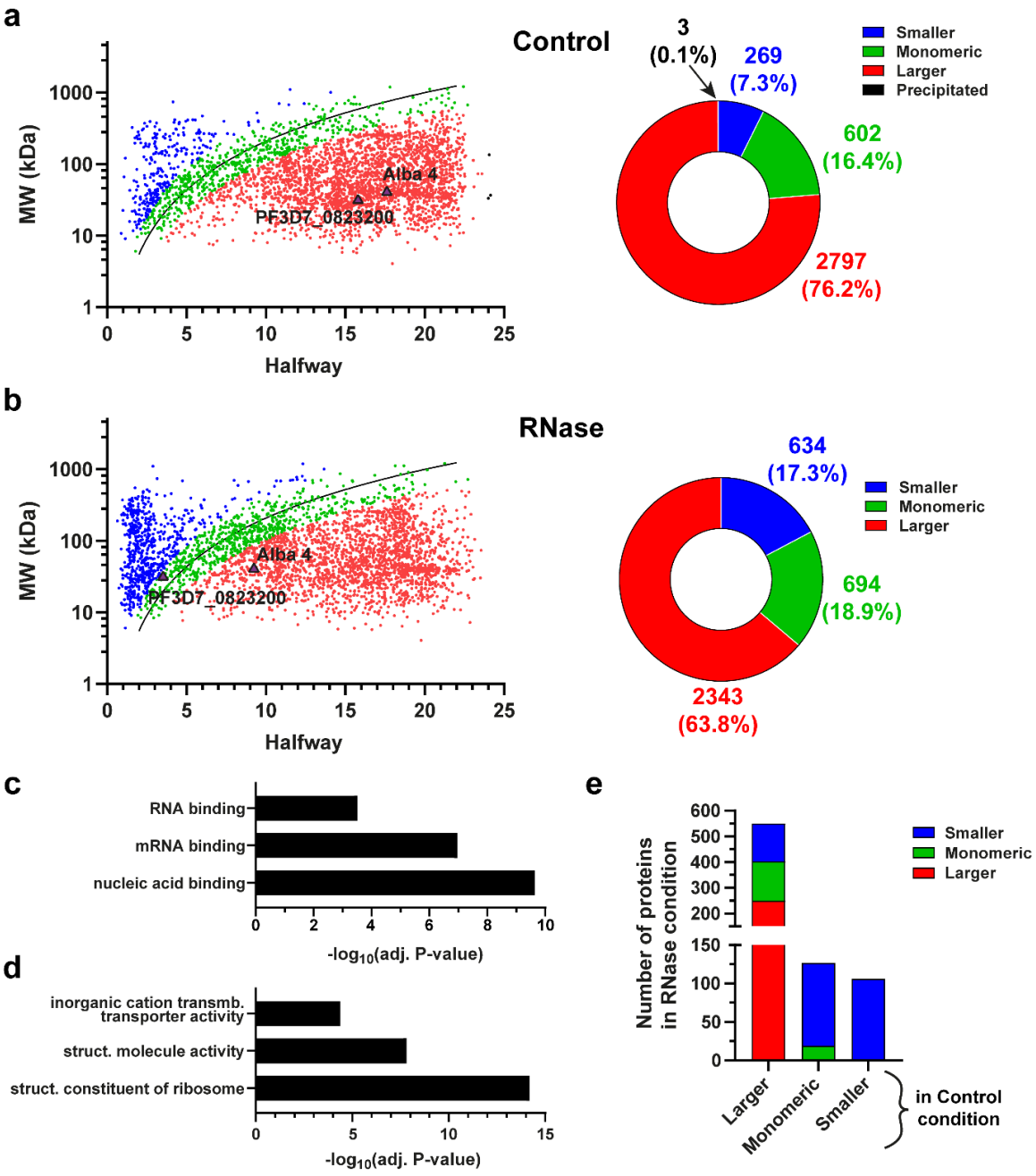
a
Control
MW (kDa)
Halfway
Alba 4
PF3D7_0823200
3
(0.1%)
269
(7.3%)
602
(16.4%)
2797
(76.2%)
Smaller
Monomeric
Larger
Precipitated
b
RNase
MW (kDa)
Halfway
Alba 4
PF3D7_0823200
634
(17.3%)
694
(18.9%)
2343
(63.8%)
Smaller
Monomeric
Larger
c
RNA binding
mRNA binding
nucleic acid binding
-log10(adj. P-value)
d
inorganic cation transmb. transporter activity
struct. molecule activity
struct. constituent of ribosome
-log10(adj. P-value)
e
Number of proteins in RNase condition
Larger
Monomeric
Smaller
in Control condition

Fig. 4.4 Interaction networks prior and after RNase treatment. Plots showing the halfway value and the theoretical molecular weight (MW) for each protein in Control (a) and RNase condition (b) (Left panel). Pie charts indicating the number and percentage of proteins considered smaller (blue), monomeric (green), larger (red) and precipitated (black) (Right panel). The reference extrapolation was used to calibrate the sucrose gradient and classify each protein into the different categories based on their MW ratio. The exact positions of Alba 4 and PF3D7_0823200 are indicated in control and RNase conditions. c and d. GO enrichment analysis of proteins considered smaller and monomeric (c) and larger (d) after RNase treatment. The significance of Molecular function terms is represented as $-\log_{10}$ (adjusted P-value) (Fisher's exact test with Bonferroni adjustment). e. Graph showing the displacement of the 785 significantly shifted proteins. These proteins were separated into three groups based on their classification in the control condition.

Next, we focused on our list of 785 shifted proteins to evaluate their structural fate in the two different experimental conditions. Interestingly, 301 out of 550 proteins considered as larger in the control condition were shifted to monomeric (154) and smaller (147) categories, indicating that the degradation of RNA resulted in loss of complex formation/stability (Fig. 4.4e). Among the shifted proteins remaining in complex in the RNase condition, we identified ribosomal and LSm proteins, suggesting that smaller but still interacting subunits may have formed from the full complexes. Moreover, 127 monomeric proteins were identified as monomeric in the Control, and 108 of them became smaller after RNase treatment (Fig. 4.4e). In total, 409/785 proteins (52.1%) were shifted to a lower category (larger > monomeric > smaller) confirming the deterioration of complex integrity for many of them. For PF3D7_0823200, the protein was shifted from larger to monomeric with an apparent MW 17.98 and 0.61 times that of the theoretical in the Control and RNase condition, respectively (Fig. 4.4a and 4.4b). Although Alba 4 was still considered as larger after RNase treatment, the MW ratio decreased significantly from 17.7 to 4.12, suggesting a partial shifting with the destabilization of the main complex and the formation of smaller subunits. Altogether, these results showed the benefit of R-DeeP profiling to decipher the fate of proteins in various biological conditions, including presence and absence of RNA.

Investigation of multiprotein complexes in *P. falciparum*

Based on the previous observations, we hypothesized that the integrity of protein complexes is preserved in the control condition, except for those that are DNA-dependent. Thus, the interactions among members of a protein complex can be analyzed by co-segregation. To test this idea, we first generated different random complexes ranging from 2 to 15 individual proteins and then assessed the mean correlation of these proteins within the same complex. As expected, the correlation was low (-0.162 to 0.535, average 0.12) for these false protein complexes, especially for those with a large number of partners (Fig. 4.5a, Supplementary Fig. 4.2 and Supplementary Data 4.5). Then we analyzed 45 different complexes identified in *P. falciparum* or conserved in eukaryotes, and which are considered as RNP complex or not (Supplementary Fig. 4.3 and Supplementary Data 4.5). Among them, the complexes involved in splicing such as U2, U6, Prp19-CDC5 showed a high correlation (> 0.8) as well as 20S and 26S proteasomes (Fig. 4.5b and 4.5c)[11,12]. The U6 RNP complex is nuclear and composed of a U6 small nuclear RNA (snRNA) and 7 LSm proteins (LSms 2-8). Conversely, LSm1 is not associated with U6 and is present in the cytoplasm, although it does interact with a low proportion of LSm proteins[13]. Co-segregation analysis of LSm1 confirmed the low correlation between this protein and the U6 complex (Supplementary Fig. 4.3 and Supplementary Data 4.5). Additional complexes also showed a good correlation (> 0.6), such as 60S ribosomal subunits, mitochondrial complexes II and IV, RNA polymerase I, RNA exosome and mediator complex (Fig. 4.5a, Supplementary Fig. 4.3 and Supplementary Data 4.5).

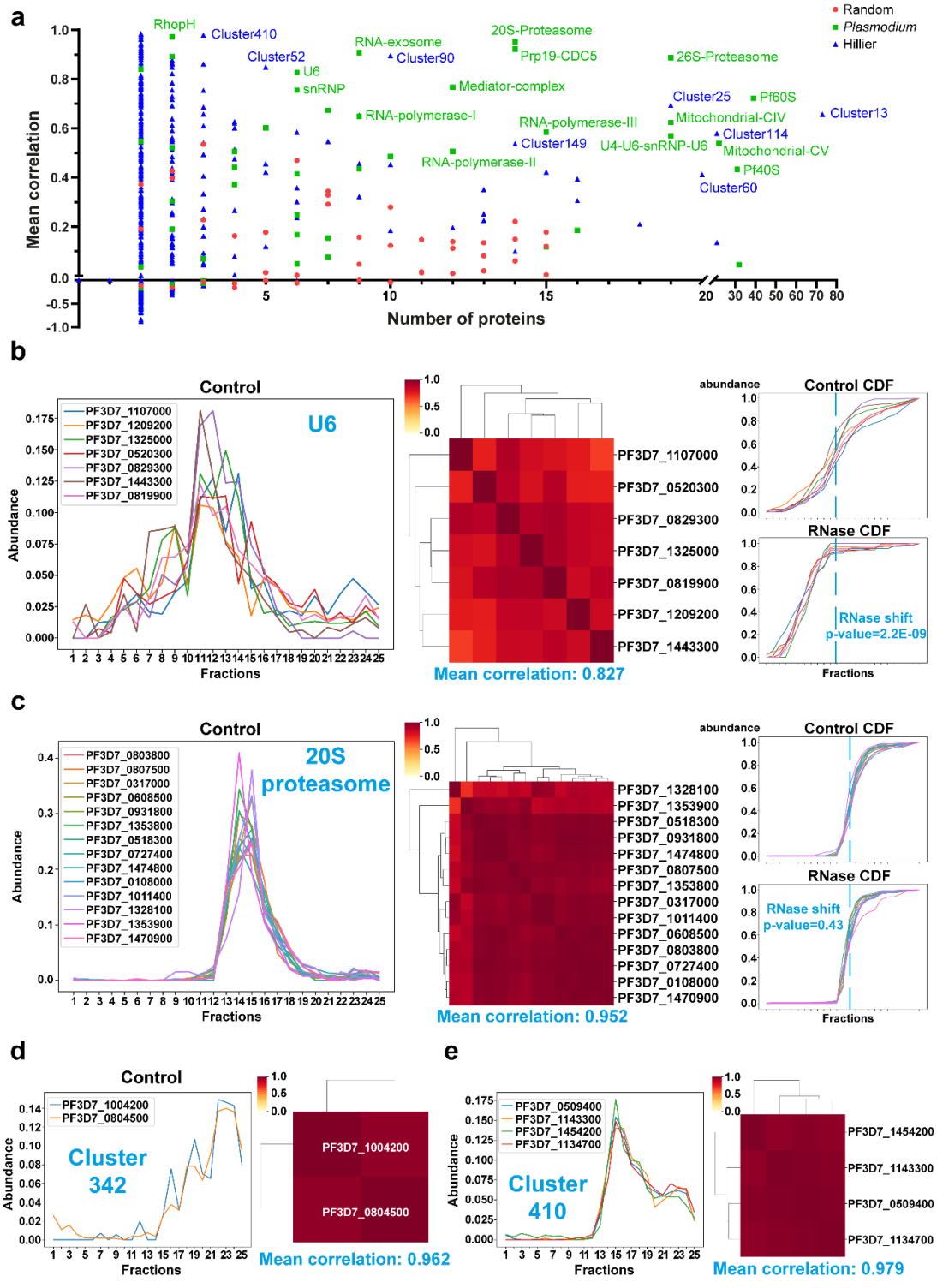
a
Random
Plasmodium
Hillier
RhopH
Cluster410
Cluster52
U6
snRNP
RNA-exosome
Cluster90
20S-Proteasome
Prp19-CDC5
26S-Proteasome
Mediator-complex
RNA-polymerase-I
RNA-polymerase-III
Cluster149
RNA-polymerase-II
U4-U6-snRNP-U6
Cluster25
Pf60S
Mitochondrial-CIV
Cluster13
Cluster114
Mitochondrial-CV
Pf40S
Cluster60
Mean correlation
Number of proteins
b
Control
U6
PF3D7_1107000
PF3D7_1209200
PF3D7_1325000
PF3D7_0520300
PF3D7_0829300
PF3D7_1443300
PF3D7_0819900
Abundance
Fractions
Mean correlation: 0.827
Control CDF
RNase CDF
RNase shift
p-value=2.2E-09
abundance
c
Control
20S proteasome
PF3D7_0803800
PF3D7_0807500
PF3D7_0317000
PF3D7_0608500
PF3D7_0931800
PF3D7_1353800
PF3D7_0518300
PF3D7_0727400
PF3D7_1474800
PF3D7_0108000
PF3D7_1011400
PF3D7_1328100
PF3D7_1353900
PF3D7_1470900
Mean correlation: 0.952
RNase shift
p-value=0.43
d
Control
Cluster 342
PF3D7_1004200
PF3D7_0804500
Mean correlation: 0.962
e
Cluster 410
PF3D7_0509400
PF3D7_1143300
PF3D7_1454200
PF3D7_1134700
Mean correlation: 0.979

Fig. 4.5 Co-segregation of *Plasmodium* protein complexes and their RNA-dependence. a. Mean correlation of protein complexes in *P. falciparum*. Graph showing the mean correlation and number of proteins for all protein complexes investigated. These complexes are grouped in three different categories: random (red), *Plasmodium* (green) and from Hillier publication (blue). b. Co-segregation and RNA-dependence of the U6 complex. Graph showing the mass-spectrometry profiles of all components of the U6 complex under the control condition (left part). Using these different profiles, mean correlations were calculated and represented in the heatmap as well as an overall mean correlation for the U6 complex (center part). Based on the cumulative distribution function (CDF) profiles and associated p-values for all components of the U6 complex, an RNase shift p-value was calculated at the complex level (right part). c. Co-segregation and RNA-dependence of the 20S proteasome. d and e. The control profiles, heatmaps and mean correlations are depicted for cluster 342 (d) and cluster 410 (e).

Taking advantage of this method to reconstruct protein complexes using R-DeeP data, we analyzed a recent publication that claimed to identify over 20,000 putative protein interactions in *Plasmodium* using quantitative mass spectrometry and machine learning[14]. Among a total of 593 clusters, with a mix of known and unknown complexes, we obtained a correlation coefficient for 442 of them containing at least two detected proteins (Supplementary Data 4.5). A subset of 73 clusters (16.5%) obtained a correlation > 0.7, suggesting that they are likely to be real complexes and require further investigation (Supplementary Fig. 4.4). This was the case. For example, with cluster 342 comprising two unknown proteins and displaying a correlation at 0.962, or cluster 410 containing 3 RNA polymerase I components and PF3D7_1454200, an unknown protein which could probably be associated with transcription in view of its partners (Fig. 4.5d and 4.5e). Additional clusters showed lower correlations (Fig. 4.5a), but some are composed of two different subunits like cluster 25 with PA700 and eIF3 complexes, for which the separate correlations are high. Cluster 13 contains a total of 73 proteins, part of the 60S and 40S ribosomal subunits, and had a mean correlation at 0.657. Overall, these results confirm that our R-DeeP data can benefit the malaria community to elucidate known or hypothetical complexes, regardless of their RNA dependence.

RNA-dependence of *Plasmodium* ribonucleoprotein complexes

Reconstruction of different complexes in *P. falciparum* demonstrated that the proteins mostly remained structurally organized in the control condition, indicating that we can study the impact of RNase treatment at the complex level and not just the protein level. We

calculated the significance for the different complexes by multiplying the independent shift p-values of each partner using Fisher's method. Thus, a p-value representative of RNA-dependence was assigned to the previous 45 complexes and 442 clusters (Supplementary Fig. 4.3 and 4.4, and Supplementary Data 4.5). As expected, several complexes associated with splicing were significantly shifted such as U2 (p-value=5.00E-11), U6 (p-value = 2.20E-09) and Prp19-CDC5 (p-value = 0.003) (Fig. 4.5d and Supplementary Data 4.5). The different RNA polymerase I, II and III complexes showed significant p-values at 0.032, 0.041 and 4.30E-12, respectively, confirming their dependence with RNA, as well as the mediator complex (p-value = 0.023), which is also associated with transcription[15]. At the protein level, we noticed that only 23 ribosomal proteins were shifted (Fig. 4.2a). Although the 40S ribosomal subunits showed a p-value of 3.20E-05, Pf60S was slightly above the significance threshold (p-value = 0.056), confirming that this complex is moderately disturbed by the by RNase treatment even though it is an RNP complex (Supplementary Fig. 4.3 and Supplementary Data 4.5). Similarly, the RNA exosome is composed of 9 proteins[16], all absent from our list of shifted proteins (Fig. 4.2a), and our analysis confirmed that they were in complex with a mean correlation at 0.908; however, their interaction did not appear to be RNA-dependent with a p-value at 0.05. For the 20S proteasome, no significant shift was observed (p-value = 0.43), as well as mitochondrial complexes (p-values from 0.89 to 1), confirming that their stabilities are not linked to RNA (Fig. 4.5d and Supplementary Fig. 4.3).

PF3D7_0823200 is well conserved in *Plasmodium*

Given the R-DeeP mass spectrometry analysis, western blot validation, and the specificity of our custom antibody, we decided to further study the PF3D7_0823200 protein. On the PlasmoDB database (v61), this protein is annotated as a putative RBP since two RNA recognition motif (RRM) domains were identified by SMART and ScanProsite. In 2021, PF3D7_0823200 was described as ortholog of UIS2 protein (PBANKA_0506200), a critical RBP for gametocyte development and production of sporozoite in *P. berghei*[17], a malaria rodent model. However, although PF3D7_0823200 is the closest homolog of PBANKA_0506200 in *P. falciparum*, this first protein shares higher identity with PBANKA_0707400 (global identity 88% vs 7%), indicating that UIS2 is most likely not its ortholog. In fact, PF3D7_0823200 is well conserved in *Plasmodium* achieving >81% identity with various *Plasmodium* species, including *P. vivax*, *P. yoelii* and *P. chabaudi* (Supplementary Data 4.6). This homology decreases considerably with other Apicomplexa but still matches uncharacterized RBPs in *Theileria* and *Neospora*. For *Cyclospora cayetanensis* and *Babesia microti*, two apicomplexan parasites, homology was detected (40% and 31% identity, respectively) with CUGBP Elav-like proteins, family implicating in pre-mRNA alternative splicing, mRNA stability and translation[18].

PF3D7_0823200 is a nucleo-cytoplasmic protein interacting with splicing and translational factors

To further characterized PF3D7_0823200, we performed immunofluorescence assays (IFAs) to detect its localization in the parasite. We showed that PF3D7_0823200 was

detected in all asexual stages of *P. falciparum*, and although the protein appeared to be enriched in the parasite nucleus, it was also present in the cytoplasm (Fig. 4.6a). This protein was previously identified in both cellular compartments by MS[19,20].

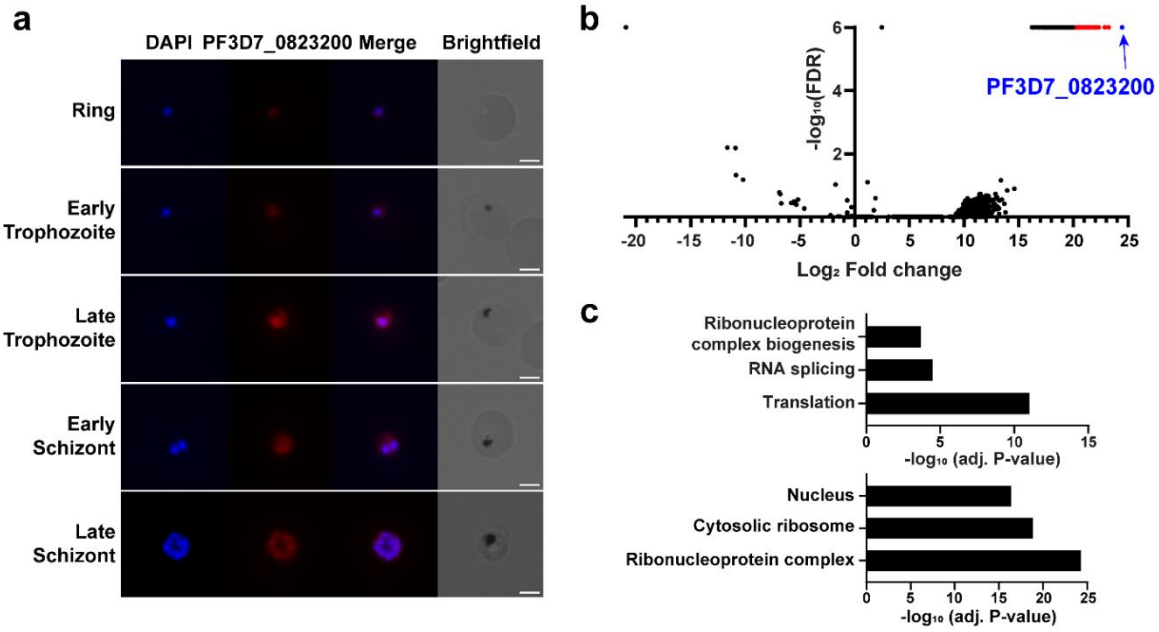

Fig. 4.6 Localization and interactome of PF3D7_0823200. a. Immunofluorescence assay of PF3D7_0823200 on ring, early trophozoite, late trophozoite, early schizont and late schizont. PF3D7_0823200 was labeled using its respective custom antibody and parasite nucleus was stained with DAPI. Merge shows both signals. Scale bar: 3 µm. b. Immunoprecipitation of PF3D7_0823200. Volcano significance plot highlights the 94 proteins significantly enriched (in red) in the three affinity purifications compared to three control purifications using anti-IgG. PF3D7_0823200 is highlighted in blue. c. GO enrichment analysis of the significantly enriched proteins. The top 3 terms of Biological Process (top) and Cellular Component (bottom) are represented as $-\log_{10}$ (adjusted P-value) (Fisher's exact test with Bonferroni adjustment).

Based on the position of PF3D7_0823200 in the sucrose fractions, we established above that the protein was part of a large complex (MW ratio = 17.98) in the control condition. Thus, to determine its potential partners, we performed immunoprecipitation followed by mass spectrometry (IP-MS) using the anti-PF3D7_0823200 antibody on soluble protein extracts of 3D7 parasites. Proteins were filtered with QPROT-calculated Log2 fold change >3, FDR < 0.01 and mean dNSAF ≥ 0.0005 compared to values measured with anti-IgG. PF3D7_0823200 was the most abundant protein detected, and 94 additional proteins were significantly co-purified (Fig. 4.6b and Supplementary Data 4.7). GO enrichment analysis showed that the detected proteins are mainly involved in translation, RNA splicing and ribonucleoprotein complex biogenesis (Fig. 4.6c), confirming a role of PF3D7_0823200 in RNA pathways. These candidates were cytoplasmic as well as nuclear, validating the presence of PF3D7_0823200 in both cellular compartments. Among these potential partners, we detected CUGBP Elav-like family member 1 and 2 (PF3D7_1359400 and PF3D7_1409800, respectively), supporting the hypothesis that PF3D7_0823200 may be associated with the CUGBP Elav-like family.

PF3D7_0823200 is an RBP regulating *var* and *ap2* transcripts

Based on its interactome and predicted function, we sought to identify RNA targeted by PF3D7_0823200 and performed an eCLIP-seq experiment. Briefly, RNA-protein complexes were UV-crosslinked, immunoprecipitated and RNAs were reverse-transcribed for high-throughput sequencing. eCLIP-seq experiments were performed in duplicate and anti-IgG was used as negative control. Using Piranha, a CLIP-seq peak caller[21], and

stringent filters (see Methods), we detected a total of 512 peaks after comparison of PF3D7_0823200 IP and Input samples while only one peak was significantly identified with IgG samples (Supplementary Fig. 4.5 and Data 8). These peaks were distributed across 307 genes and 87.9% of the peaks were found in the same orientation as the associated gene (Fig. 4.7a). Only 17 peaks were detected on gene coding regions indicating a preferential binding of PF3D7_0823200 on untranslated RNA sequences. Indeed, 126 peaks were significantly identified on distal intergenic (24.6%), 155 on 5' UTR (30.3%), 130 on intron (25.4%) and 84 on 3'UTR (16.4%) (Fig. 4.7b) suggesting that this RBP may play a role in RNA stabilization and/or splicing. GO enrichment analysis showed that these RNAs were associated with host cell surface, cell adhesion binding and DNA-binding transcription factor activity (Fig. 4.7c). Closer inspection revealed an important diversity of targets suggesting the binding of PF3D7_0823200 does not appear to be directly linked to the protein families but rather to RNA regions. However, an exception was noticed for *var* genes with 56 significant peaks and 12 AP2 transcription factors, for which 44 peaks were called, including AP2-EXP, AP2-G5 and AP2-L (Supplementary Data 4.8). These two protein families were uniquely responsible for the previous GO enrichment observed. Although not all *var* peaks were called, we noticed a particular pattern with a first peak mapping at the start of the intronic region in the opposite orientation to the gene while a larger second peak was detected at the end of the intron in the gene orientation (Fig. 4.7d). This arrangement is identical to the orientation and position of the well-studied sense and antisense lncRNAs of *var* genes that are transcribed from the bidirectional intron promoter. Although further experiments are required, this result suggests that PF3D7_0823200 could

interact with *var* lncRNAs and participate in the regulation of the *var* gene family and their mutually exclusive expression. For the AP2 transcription factors, the majority of the peaks were identified on the distal intergenic and 5' UTR regions indicating that the protein interacts upstream of the AP2 coding regions. The detection of intergenic sequences could reveal incorrect UTR annotations or the presence of ncRNAs and upstream ORF (uORF). Whatever the explanation, this RBP seems to contribute to the dynamic balance between these master transcription factors. Similarly, we detected five peaks mapping the region of *gdv1*, an activator of sexual commitment, but the opposite orientation of these peaks relative to the gene most likely indicated that PF3D7_0823200 interacted with the antisense lncRNA. This lncRNA is described to inhibit *gdv1* transcription, thereby maintaining *ap2-g* in a repressed state. Otherwise, the 5' and 3' UTR of PF3D7_0823200 transcripts exhibited seven peaks indicating a feedback loop but additional experiments will be necessary to determine if this feedback is positive or negative (Fig. 4.7e).

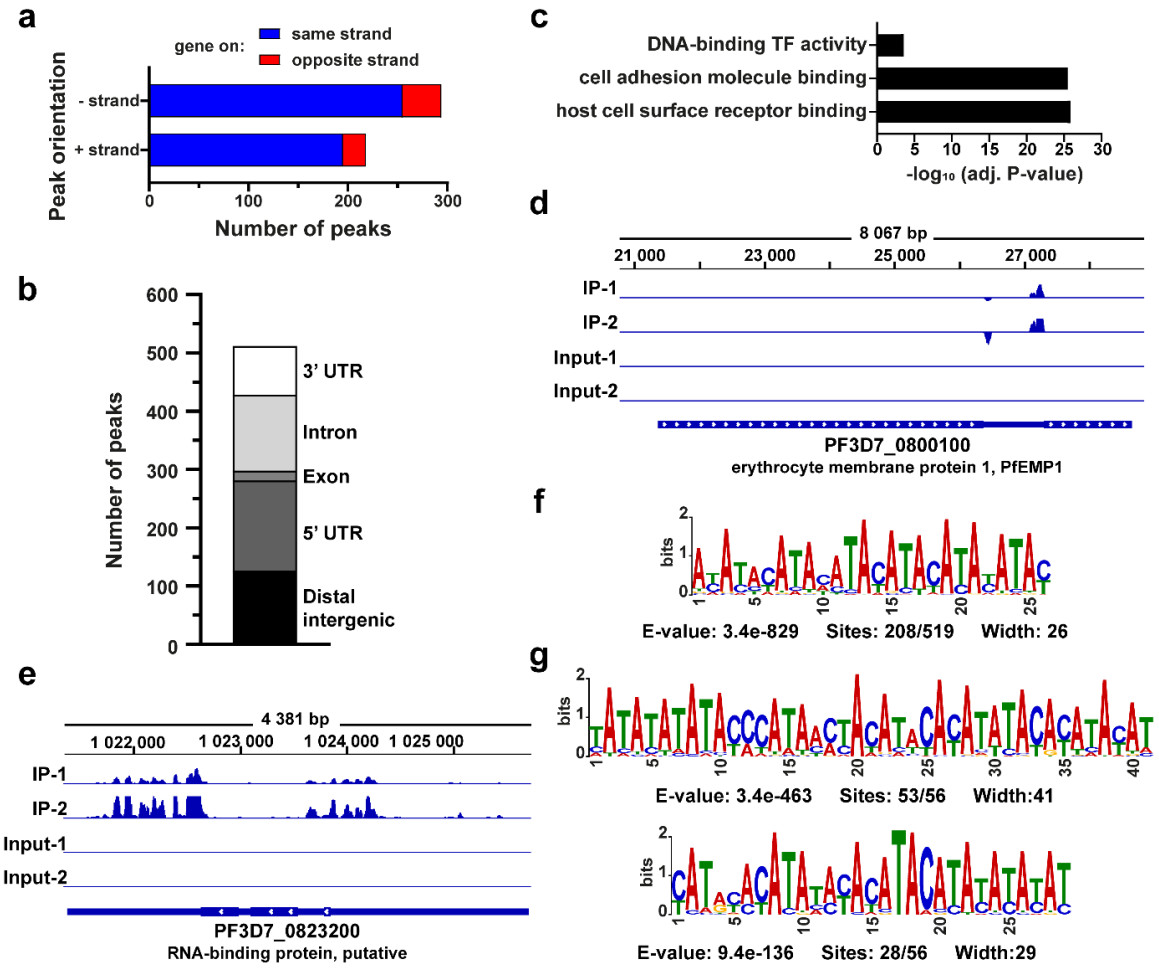

Fig. 4.7 Identification of PF3D7_0823200 targets using eCLIP-seq. a. Number and orientation of the peaks detected using Piranha. b. Distribution of the peaks identified. c. GO enrichment analysis of the significantly enriched transcripts. The top 3 terms of Molecular Function are represented as $-\log_{10}$ (adjusted P-value) (Fisher's exact test with Bonferroni adjustment). d and e. Tracks showing the eCLIP-seq peaks spanning on the region of the *var* gene, PF3D7_0800100 (d), and PF3D7_0823200 (e). The scales are 0-100 and 0-150, respectively. f. Sequence logo of the most significant motif identified by MEME Suite search using the 512 eCLIP-seq peaks. g. Sequence logo of the two most significant motifs using the 56 peaks mapping on *var* genes. All significant motifs identified by MEME Suite search are represented in Supplementary Fig. 4.6.

Finally, the sequences of the 512 peaks were used to search enriched motifs using MEME Suite tool[22]. We found six enriched motifs with E-value < 0.05 (Supplementary Fig. 4.6a) but the motifs 2-4 were not considered biologically meaningful because of their AT-richness, characteristic of non-coding regions in *Plasmodium*. The most significant motif was identified in 208 peaks (40%) and showed an E-value at 3.4e-829 (Fig. 4.7f). The 56 peaks mapping *var* genes were also analyzed separately and four motifs were significantly detected (Fig. 4.7g and Supplementary Fig. 4.6b). One particular motif was identified in 53 out of 56 sequences indicating its high specificity (E-value = 3.4e-463).

Collectively, these eCLIP-seq results confirmed that PF3D7_0823200 is a true RBP interacting with untranslated regions of various transcripts including *var* and *ap2* genes.

Discussion

Our work provides the first proteome-wide screening of RDPs and RNP complexes in *P. falciparum*. Using the R-DeeP methodology, we identified 785 RDPs for which 41% were already associated with RNA. However, the main disadvantage of R-DeeP approach is the limited detection of known RBPs interacting alone or in small complexes with RNA as well as those whose RNA is not critical for stability/formation of the RNP complex(es). For the newly identified RDP candidates, a large proportion of them (172/463) corresponded to uncharacterized conserved or *Plasmodium*-specific proteins. These proteins may not directly interact with RNA, explaining why they were not previously classified as RBPs. Interestingly, 13 AP2 transcription factors were significantly impacted by RNase treatment indicating that they form RNP complexes. Among them, we detected

AP2-G, AP2-G2 and AP2-G5, three major transcriptional regulators in the transition from asexual to sexual programs. Although AP2-G expression is under the control of gdv1-lncRNA, no direct association has been described to our knowledge between the AP2-G transcriptional complex and (lnc)RNAs. Furthermore, AP2-HC is associated with heterochromatin[23], while AP2-SIP2 and AP2Tel bind to subtelomeric regions of *P. falciparum*[24,25]. These regions are known to be enriched in heterochromatin marks and lncRNAs such as lncRNA-TAREs, involved in telomere maintenance[26]. Although AP2-SIP2 was slightly above our statistical threshold (q-value = 0.05169), unlike AP2-HC and AP2Tel, models have advanced potential synergies between AP2-SIP2 and lncRNA-TAREs for transcriptional regulation and recruitment of heterochromatin components[26]. Recruitment, assembly and regulation of these transcription factors may require the participation of lncRNAs to fulfill their respective biological functions.

The R-DeeP method also provided a snapshot of *Plasmodium* protein-protein interactions in presence and absence of RNA. As expected, a large majority of the proteins (76%) were in complex in the control condition. This percentage is most likely higher in cellular condition since all DNA-dependent complexes were affected by our DNase treatment in both experimental conditions. After RNase digestion, 12.4% of the proteins were no longer in complex indicating that the interactions of these proteins were RNA-dependent.

Co-segregation analysis of our R-DeeP date allowed us to reconstruct multiprotein complexes conserved in eukaryotes or specific of *Plasmodium*. A total of 84 protein complexes or clusters had a mean correlation > 0.7 suggesting that they were most likely detected in complex in our R-DeeP. Thus, this approach could be a useful tool for the

malaria community to validate protein complexes in *P. falciparum* identified by various techniques such as *in silico* analysis, immunoprecipitation followed by mass spectrometry and cryo-electron microscopy.

Our analysis also evaluated the RNA-dependence of these multiprotein complexes to discriminate which ones are ribonucleoprotein particles. As expected, the majority of spliceosomal complexes were unstable after RNase treatment as well as RNA-polymerase I-III complexes, while the integrity of 20S and 26S proteasomes were not sensitive to RNase activity. RNA-exosome and snRNP are well conserved complexes, including in *P. falciparum*, and are involved in RNA quality control and splicing, respectively[11,16]. Despite their RNA-associated functions and high correlation (> 0.75), these two complexes were not significantly shifted suggesting that they are not RNA-dependent. The architecture of the exosome core is highly conserved in the tree of life and so far, this complex is not described to require (nc)RNAs to scaffold its structure supporting the lack of shift in this R-DeeP experiment. On the contrary, although the snRNP particle is composed of small nuclear RNAs and proteins, this RNP complex showed a non-significant p-value at 0.58 suggesting that the RNA of some RNP complexes may be protected from RNase activity. Similarly, the 60S ribosomal subunit, assembly of ribosomal RNAs and proteins, also had a p-value above the threshold (0.056) and only 3/39 proteins were significantly shifted. We can hypothesize that these proteins were located on the surface of the complex and only them could have been detached from the rest during the RNase treatment. Thus, for large RNP complexes having RNA embedded inside their structure and not accessible to RNases,

their detection by the R-DeeP technique would then be limited. Complementary approaches are therefore necessary to unveil the complexity of RDP and RNP repertoires.

Identified in our R-DeeP experiment, the significant shifting of PF3D7_0823200 was validated by western blot using a custom antibody. *In silico* analysis suggested that this protein is well conserved in *Plasmodium* and shares similarities with CUGBP Elav-like family members in (which organisms?). This observation was supported by our IP-MS experiment showing a significant enrichment of translational and splicing factors, including the parasite CUGBP Elav-like family member 1 and 2. In Human, six CUGBP Elav-like family members, annotated as CELF1-6, are identified and regulate several steps of RNA processing. In the nucleus, CELF proteins are involved in alternative splicing of pre-mRNA and RNA editing, while in the cytoplasm, they are associated with mature mRNAs and participate in deadenylation, stability and translation of the transcripts[18,27]. Several studies demonstrated that CELFs expression can be dysregulated by miRNAs and lncRNAs and are associated with development of human cancers[27]. Reciprocally, CELF proteins also have cooperative or antagonistic roles on ncRNA expression and function. In view of the nucleo-cytoplasmic localization and eCLIP-seq results, it is tempting to consider that PF3D7_0823200 could also be involved in similar processes. Indeed, the large majority of the eCLIP-seq peaks were detected in UTR and intronic regions suggesting a potential role in mRNA stabilization and/or RNA splicing. Although PF3D7_0823200 appeared to interact with a wide variety of transcripts, an enrichment was observed for *var* genes involved in antigenic switching and immune evasion, and AP2 transcription factors, master regulators of stage conversion.

Interestingly, this protein was also associated with ncRNA transcripts, including the antisense lncRNA of *var* genes and *gdv1*, emphasizing its fundamental role in regulation of RNA metabolism in *P. falciparum*. Recently, two large-scale genetic screening studies using *piggyBac* mutagenesis in *P. falciparum*[28] and barcoded *P. berghei* knockout mutations[29] were performed to identify essential genes in *Plasmodium*. Both studies indicated that disruption of PF3D7_0823200 or PBANKA_0707400 locus impaired development of asexual stages with a mutant fitness score of -2.165 in *P. falciparum* and a growth rate of 0.74 in *P. berghei*. The PBANKA_0707400 knock-out mutant line was also severely impacted during transitions from blood stage to oocyst, oocyst to sporozoite and sporozoite to blood stage (differences in relative abundance between -3.02 to -4.01)[30]. Further functional and phenotypic assays using CRISPR transgenic lines are required to elucidate the fundamental role of PF3D7_0823200 in the parasite and RNA regulation.

Materials and Methods

Parasite Lysate Preparation and RNase Treatment

Cultures of *P. falciparum* NF54 were synchronized by D-sorbitol treatments and 5 x10^{10} late trophozoites were treated with 0.15% saponin. After PBS washing, parasites were evenly separated into 2 samples (Control and RNase) and suspended in lysis buffer (25 mM Tris-HCl pH 7.4, 150 mM KCl, 0.5% (v/v) Igepal CA-630, 2 mM EDTA, 0.5 mM DTT, 1X protease inhibitor cocktail (Roche, 04693159001), 1X phosphatase inhibitor (Roche, 04906837001) and for Control sample only: 1200 units of RiboLock RNase

Inhibitor (Thermo Scientific, EO0381)). After 30 min of incubation on ice with vortexing every 5 min, a freeze-thaw followed by homogenization using a 26 ½ G needle was performed on the parasite extract to improve the lysis efficiency. This step was repeated for a total of 3 freeze-thaw cycles. The soluble protein extract was obtained after centrifugation at 13,000 rpm for 15 min at 4°C. Subsequently, the Control sample was treated with 100 units of DNase I (NEB, M0303) and 1X DNase I reaction buffer for 1 h at room temperature, while the RNase sample was treated with 100 units of DNase I, 500 units of RNase I (Ambion, AM2294), 50 units of RNase H (NEB, M0297), 200 µg of RNase A (Invitrogen, 12091021) and 1X DNase I reaction buffer. During the incubation, protein concentrations were quantified by Bradford assay (Sigma-Aldrich, B6916). The quality of the enzymatic treatments was assessed on 1.2% agarose gel after phenol:chloroform:isoamyl alcohol (25:24:1, v/v) purification (Supplementary Fig. 4.1a).

Sucrose Density Gradient Preparation and Ultracentrifugation

Ten sucrose solutions from 50% to 5% (w/v) sucrose were prepared in 10 mM Tris (pH 7.5), 1 mM EDTA (pH 8) and 100 mM NaCl as previously described [8]. First, 1 mL of the 50% sucrose solution was added to the bottom of the tube (Beckman Coulter, 344059) and flash frozen in liquid nitrogen. Then each sucrose solution was layered on top of the previous solution and frozen prior to addition of the next layer with the 5% sucrose solution on top of the tube. The sucrose density gradients were stored at -20°C and thawed slowly on ice before adding protein lysates.

The Control and RNase samples were carefully overlaid on top of the thawed sucrose gradients avoiding any disturbance. For each condition and replicate, 2 to 2.5 mg of proteins were loaded onto the sucrose gradients. Ultracentrifugation was performed in Beckman L8-70M Ultracentrifuge equipped with a SW 41 Ti Swinging-Bucket Rotor (Beckman Coulter, 331362) at 30,000 rpm for 18 h at 4°C. After centrifugation, 25 fractions (~440 mL each) were carefully transferred by pipetting into fresh 1.5 mL tubes. Fraction 1 corresponded to the top of the tube and fraction 25 to the bottom. The different fractions were stored at -80°C for western blot analysis or precipitated with 20% trichloroacetic acid (TCA) for mass spectrometry.

Preparation of Samples for Mass Spectrometry

TCA precipitated samples were resuspended in 30 µl buffer containing 100 mM Tris-HCl, pH 8.5 and 8 M urea. Disulfide bridges were reduced with tris(2-carboxyethyl)phosphine (5 mM final concentration) for 30 minutes at room temperature. Free SH groups were alkylated with chloroacetamide (CAM, 10 mM final concentration) for 30 minutes at room temperature in the dark. Proteins were first digested with 0.1 µg endoproteinase Lys-C for 6 hours at 37°C. Samples were then diluted with 100 mM Tris-HCl, pH 8.5 to reduce the concentration of urea to 2 M, CaCl2 was added (2 mM final concentration), and digestion was continued with the addition of 0.5 µg trypsin. Samples were incubated at 37°C overnight with shaking and reactions were quenched with the addition of formic acid (5% final concentration).

Mass Spectrometry Analysis

Each sample was loaded onto a split triple-phase fused silica microcapillary column prepared as described previously[31]. Peptides were eluted from the column using a series of 10 ~2-hour MudPIT steps. Mass spectrometry was performed using an Orbitrap Elite Hybrid mass spectrometer in positive ion mode.

Peptide detection and quantification

Mass spectrometry data was generated, in two replicates, from 25 sucrose gradient fractions in both Control and RNase. A total of 1108 runs were carried out, and each raw file was converted to mzML format using the msconvert command in Proteowizard[32] using default parameters.

Proteins were detected and quantified separately in each fraction using Crux version 3.2-46bb0c1[33], in four steps. First, the canonical *Plasmodium* and human reference proteomes were downloaded from PlasmoDB (release 47) and UniProt (UP000005640), respectively, and concatenated into a single FASTA file. A peptide index was created using the tide-index command, requiring fully tryptic peptides, up to two missed cleavages, and up to three methionine oxidations per peptide. These settings yielded a total of 5,669,753 distinct tryptic peptides. For each target peptide, a corresponding shuffled peptide decoy was also stored in the index. Second, the Tide search engine (tide-search command in Crux) was used to search spectra from each fraction against the Tide index. The search employed the exact p-value score function[34], allowing isotope errors of 1 or 2 m/z, using an m/z bin width of 1.0005079 m/z, and using Param-Medic[35] to automatically select an appropriate

precursor window size. Other Tide parameters were left at their default values. Third, all of the resulting peptide-spectrum matches (PSMs) from each fraction were analyzed jointly using Percolator[36], also via Crux, using default parameters. This step yielded, for each fraction, a list of proteins with associated q-values. Fourth, the Crux spectral-counts command was used to compute a normalized spectral abundance factor (NSAF)[37] for each protein in each fraction. In the NSAF calculation, only peptides identified at 1% peptide-level FDR by Percolator were considered. The NSAF values were aggregated into a set of matrices, one per replicate and treatment, in which rows are proteins and columns are sucrose gradient fractions. Proteins with no corresponding PSMs in a given fraction receive an NSAF value of zero. The total number of distinct proteins identified in at least one of the four settings (treatment or control, two replicates) was 5214.

R-DeeP: Statistical analysis

Prior to statistical analysis of the four NSAF matrices, four preprocessing steps were performed. First, protein quantifications were averaged across replicates, yielding one matrix for control and one for RNase treatment. In this step only, values of zero in either matrix were ignored, so that the "average" of a non-zero value x in one matrix with a corresponding zero value in the other matrix was simply x. Second, we required that each protein have at least two consecutive non-zero average NSAF values in both treatment and control. Proteins that failed this criterion for either treatment or control were eliminated from both matrices. This step reduced the number of rows in each matrix from 4146 to 3671. Third, each matrix row was normalized to have a total abundance of 1. Fourth, each

row was converted to a cumulative density, so that the value at row i and column j is the proportion of the abundance associated with protein i that is observed at or before fraction j. Given this preprocessed matrix, we use a Wilcoxon rank-sum test to detect proteins that exhibit a statistically significant shift in NSAF values between treatment and control. The statistic is based on the first 24 entries in each row, since the 25th entry is 1 by definition. Note that this step is equivalent to computing the area under a receiver operating characteristic (ROC) curve between the two distributions. In our setting, we are only interested in proteins for which the RNase peak comes before the control peak; hence, we use a one-tailed test in which values >0.5 correspond to shifts in the desired direction. Finally, we subject the Wilcoxon p-values to FDR control using the Benjamini-Hochberg procedure[38].

Comparison of RNA-associated protein datasets

A list of *Plasmodium* RNA-associated proteins was produced by collecting the datasets of the two *in silico* studies[6,7] and the mRNA proteome capture experiment[7]. In addition, we also integrated proteins annotated on PlasmoDB with GO terms associated with RNA (GO:0016071: mRNA metabolic process; GO:0140098 catalytic activity, acting on RNA; GO:0006396: RNA processing; GO:0003723: RNA binding; GO:0005840: Ribosome). This final collection of 1319 unique RNA-associated proteins is described in Supplementary Data 4.2 and was compared to the 785 significant proteins provided by this R-DeeP experiment. The UpSet plot was generated using UpSetR[39].

Custom antibody production and purification

The different candidates were selected based on their shifting rank, molecular weight and (un)known function. The list included 6 proteins significantly shifted: PF3D7_0528600, PF3D7_1354900, PF3D7_1360100, PF3D7_0823200, PF3D7_0916700 and PF3D7_1347500, and two negative controls: PF3D7_1353900 and PF3D7_1465000. Peptide antigens were designed to target the C-terminal region and are indicated in Supplementary Data 4.3. They were used to immunize two rabbits and antisera from day 72 post-immunization were collected (Thermo Fisher Scientific). Antibody specificity was tested by western blot analysis on total *P. falciparum* protein extract. For each protein, the best antiserum was affinity-purified (Thermo Fisher Scientific) and validated by western blot analysis (Supplementary Fig. 4.1b).

Western blot analysis

For each fraction, 27 µL (6% of total volume) was suspended with 8 µL of 4x Laemmli Sample Buffer (BioRad, 1610747). The samples were boiled at 95°C for 5 min and loaded to 10% polyacrylamide gel. After migration, proteins were transferred onto a PVDF membrane using a Trans-Blot SD Semi-Dry Transfer Cell (BioRad) at 15V for 30 min. Then the different membranes were blocked for 1 h at room temperature in WesternBreeze™ Solution (Invitrogen, WB7050) and incubated with the respective primary antibody (1:50 in WesternBreeze; 1:10,000 for anti-Aldolase) at 4˚C overnight with regular shaking. After 3 washes with WesternBreeze™ Wash Solution (Invitrogen, 46-7005), the blots were probed with HRP-labeled Goat anti-Rabbit IgG (H + L) (1:10,000,

Novex™, A16104). Next, Clarity™ Western ECL Substrate (Bio-Rad, 1705060) was applied to reveal the membranes. For each antibody, all 50 fractions distributed over four membranes (25 fractions per condition) were analyzed simultaneously by a ChemiDoc™ (BioRad).

Molecular weight analysis

For each protein, the halfway was calculated and corresponded to the value for which the CDF reached 0.5, indicating that 50% of the total protein amount was detected. The apparent MW was obtained using the reference extrapolation ($y = 1146.9x^{2.2577}$; $R^2 = 0.9984$) based on position and molecular weight of reference human proteins[8]. Proteins were filtered by setting a cut-off of 0.5 and 2 for the apparent MW/theoretical MW ratio. Each protein was classified as smaller (< 0.5), monomeric ($0.5 < x < 2$), larger (< 2) or precipitated (hallway ≥ 24).

Analysis of random, *Plasmodium* complexes and protein clusters

Proteins were randomly distributed within each random complex. Three complexes were generated for each number of individual proteins (2 to 15). The different *Plasmodium* complexes were selected from various publications: Pf60S and Pf40S[40], U1, U2, U2-related, SF3a, SF3b, U5, U4-U6, snRNP, U6, tri-snRNP, Prp19-CDC5, non-snRNP, NMD and SR-nRNP[11], Invasion-AMA1[41,42], IMC and Glideosome[43], PTEX[44], RAP[45], RNA-exosome[16], Mitochondrial complexes[46], Mediator-complex[15], SIP2[24], PfAP2Tel[25], DOZI-eIF4E[47], PfHSP40 and PfHSP70x[48], Kae1api[49], 20S-Proteasome and 26S-Proteasome[12],

RhopH[50], Basal-complex[51], and CCR4-NOT, AMA1-MSP-RON, mRNA-decapping, RNA-polymerase complexes[42]. The protein clusters were extracted from a large-scale protein interactome study[14]. The protein composition of all complexes and clusters investigated are indicated in Supplementary Data 4.5.

For each pair of proteins within a complex, we computed the correlation of the proteins' normalized quantification profiles. These correlations were averaged across all non-identical pairs within the complex, and the associated p-values were multiplied together and adjusted using Fisher's method (Supplementary Data 4.5).

Immunofluorescence assays

3D7 *P. falciparum* parasites were fixed with 4% paraformaldehyde and 0.0075% glutaraldehyde for 15 min at 4˚C and then sedimented on coverslips coated with Poly-L-ornithine (Sigma-Aldrich, P4957) for 1 h at room temperature. After two PBS washes, fixed parasites were permeabilized and blocked with 0.2% Triton X-100, 5% BSA, 0.1% Tween 20 in PBS for 30 min at room temperature. The custom anti- PF3D7_0823200 was diluted at 1:100 in PBS, 5% BSA, 0.1% Tween 20, and applied for 1 h at room temperature. After washing with PBS, parasites were incubated for 1 h with Donkey anti-Rabbit Alexa Fluor 568 (1:2000, Invitrogen, A10042). Slides were mounted in Vectashield Antifade Mounting Medium with DAPI (Vector Laboratories, H-1200). Images were acquired using a Keyence BZ-X810 fluorescence microscope and treated with ImageJ (n>20 parasites).

Immunoprecipitation followed by MudPIT mass spectrometry

A total of 7.5 x10^9 3D7 parasites enriched in late asexual stages were extracted by saponin lysis and resuspended in 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton X-100, 5 mM EDTA, 1 mM AEBSF and EDTA-free protease inhibitor cocktail (Roche, 11873580001). Soluble proteins were extracted using a 26G needle and treated with 100 units of DNase I (NEB, M0303) for 10 min at room temperature. After centrifugation at 14,000 x *g* for 15 min at 4°C, protein lysates were precleared with Dynabeads™ Protein A (Invitrogen, 10001D) for 1 h at 4°C. Our custom anti-PF3D7_0823200 (1:100) was added in the precleared supernatant and were incubated overnight at 4°C. Purified Rabbit IgG was used in the same condition as negative control (1:100, MP Biomedicals, 0855944). Immunoprecipitations of antibody-protein complexes were performed using Dynabeads™ Protein A for 1 h at 4°C. Subsequently, proteins were washed twice in PBS, 1% Triton X-100, 1 mM EDTA, once in PBS, 1% Triton X-100, 1 mM EDTA and 0.5 M NaCl, then twice in PBS, 1 mM EDTA. Before TCA-precipitation, proteins were eluted in 0.1 M glycine, pH 2.8 and neutralized with 2 M Tris-HCl, pH 8.0.

Samples were processed and analyzed by MudPIT mass spectrometry as described above. Resulting .raw files were processed using the in-house software package RAWDistiller v1.0 to generate .ms2 files. These were searched using the ProLuCID search engine against a database containing 5527 *P. falciparum* protein sequences, 36661 human protein sequences, sequences for 419 common contaminants, and shuffled versions of all the above sequences for estimating false discovery rates. A static modification of +57.02146 Da was

used for cysteine residues (carbamidomethylation) and a variable modification of +15.9949 Da for methionine residues (oxidation). After searching, the resulting .sqt files were processed using DTASelect (v1.9)[52] using our in-house software swallow to select peptide spectrum matches such that false discovery rates at the peptide and protein levels were less than 5%. Peptides and proteins from all samples were compared using Contrast[52], and dNSAF values calculated using our in-house software NSAF7 (v0.0.1). The statistical framework QPROT[53] was used to determine a subset of proteins enriched by the anti-PF3D7_0823200 antibody compared with negative controls (log2 fold change > 3, FDR < 0.01, mean dNSAF ≥ 0.0005).

Enhanced crosslinking and immunoprecipitation followed by high-throughput sequencing (eCLIP-seq)

A total of 7.5 x10^9 3D7 parasites enriched in late asexual stages were extracted by saponin lysis and crosslinked on ice by 254 nm UV light for a total of 1200 mJ/cm² with 2 min breaks using Spectrolinker™ XL-1000. The custom anti-PF3D7_0823200 or purified Rabbit IgG (15 µg) were coupled with Dynabeads™ M-280 Sheep Anti-Rabbit IgG (Thermo Fisher, 11203D) for 1 h at room temperature and then added to the parasite lysate. The following steps were processed using eCLIP Library Prep Kit (Eclipse BioInnovations, ECEK-0001) according to the manufacturer's instructions and as previously described[45]. Based on the molecular weight of PF3D7_0823200 (~32 kDa), regions of the nitrocellulose membrane ranging from 27 to 100 kDa were isolated and digested with proteinase K to release RNA. Libraries were generated by PCR amplifications consisting of 98°C (30 sec)

followed by 6 cycles of (98°C (15 sec), 70°C (30 sec), 72°C (40 sec)), then 11 cycles of (98°C (15 sec), 72°C (45 sec)) and 72°C (1 min). Library fragments of 175 to 350 bp were gel size-selected using MinElute Gel Extraction Kit (Qiagen, 28604) and the quantity and quality of the final libraries were assessed using a Bioanalyzer (Agilent Technology Inc). All samples were multiplexed and sequenced by dual indexed run (PE100) on the Illumina NovaSeq 6000 sequencer at the UC San Diego IGM Genomics Center and on the Illumina NextSeq 2000 Sequencing System at UC Riverside.

Bioinformatic analyzes were performed as previously described[45]. To normalize, all read counts were divided by the number of millions of mapped reads for each particular sample. Peak calling was performed using Piranha with the options -z 50 (bin size 50), -l (convert covariates to log scale), and the default q-value cutoff 0.01. Some peaks were also added manually. R package ChIPseeker was used for peak annotation. Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number **PRJNA690830**.

Data availability

eCLIP-seq datasets generated in this study have been deposited in the NCBI Sequence Read Archive under accession number PRJNA690830. The R-DeeP datasets have been deposited in the ProteomeXChange (PXD023308) via the MassIVE repository (MSV000086636 with [https://doi.org/10.25345/C5R795]). The IP-MS datasets have been deposited via MassIVE repository with identification number MSV000091228. For reviewer access, reviewers can login to MassIVE with username MSV000091228_reviewer and password rdeep. Original data underlying this manuscript generated at the Stowers Institute can be accessed from the Stowers Original Data Repository (**http://www.stowers.org/research/publications/LIBPB-2374**).

Code availability

The custom Python scripts used for eCLIP-seq analysis have been previously published[54]. The entire in-house software suite (Kite) used for the MudPIT mass spectrometry analysis is available in Zenodo **(https://doi.org/10.5281/zenodo.5914885)[55].**

References

1. Zarnack, K. *et al.* Dynamic mRNP Remodeling in Response to Internal and External Stimuli. *Biomolecules* **10**, 1310 (2020).

2. Briata, P. & Gherzi, R. Long Non-Coding RNA-Ribonucleoprotein Networks in the Post-Transcriptional Control of Gene Expression. *Noncoding RNA* **6**, 40 (2020).

3. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol Cell* **78**, 9–29 (2020).

4. Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics 2020 22:3* **22**, 185–198 (2020).

5. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* vol. 19 327–341 Preprint at https://doi.org/10.1038/nrm.2017.130 (2018).

6. Reddy, B. N. *et al.* A bioinformatic survey of RNA-binding proteins in Plasmodium. *BMC Genomics* **16**, (2015).

7. Bunnik, E. M. *et al.* The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum. *Genome Biol* **17**, 147 (2016).

8. Caudron-Herger, M. *et al.* R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation. *Mol Cell* **75**, 184-199.e10 (2019).

9. Rajagopal, V. *et al.* Proteome-Wide Identification of RNA-Dependent Proteins in Lung Cancer Cells. *Cancers (Basel)* **14**, 6109 (2022).

10. Hillebrand, A. *et al.* Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. *Nucleic Acids Res* **46**, 10417–10431 (2018).

11. Sorber, K., Dimon, M. T. & DeRisi, J. L. RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* **39**, 3820–3835 (2011).

12. Aminake, M. N., Arndt, H.-D. & Pradel, G. The proteasome of malaria parasites: A multi-stage drug target for chemotherapeutic intervention? *Int J Parasitol Drugs Drug Resist* **2**, 1–10 (2012).

13. Beggs, J. D. Lsm proteins and RNA processing. *Biochem Soc Trans* **33**, 433–438 (2005).

14. Hillier, C. *et al.* Landscape of the Plasmodium Interactome Reveals Both Conserved and Species-Specific Functionality. *Cell Rep* **28**, 1635-1647.e5 (2019).

15. Iyer, U. B., Park, J. E., Sze, S. K., Bozdech, Z. & Featherstone, M. Mediator Complex of the Malaria Parasite Plasmodium falciparum Associates with Evolutionarily Novel Subunits. *ACS Omega* **7**, 14867–14874 (2022).

16. Droll, D. *et al.* Disruption of the RNA exosome reveals the hidden face of the malaria parasite transcriptome. *RNA Biol* **15**, 1206 (2018).

17. Müller, K., Silvie, O., Mollenkopf, H.-J. & Matuschewski, K. Pleiotropic Roles for the Plasmodium berghei RNA Binding Protein UIS12 in Transmission and Oocyst Maturation. *Front Cell Infect Microbiol* **11**, (2021).

18. Dasgupta, T. & Ladd, A. N. The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins. *Wiley Interdiscip Rev RNA* **3**, 104–121 (2012).

19. Oehring, S. C. *et al.* Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum. *Genome Biol* **13**, R108 (2012).

20. Briquet, S. *et al.* Identification of Plasmodium falciparum nuclear proteins by mass spectrometry and proposed protein annotation. *PLoS One* **13**, e0205596 (2018).

21. Uren, P. J. *et al.* Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* **28**, 3013–3020 (2012).

22. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).

23. Carrington, E. *et al.* The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the malaria parasite Plasmodium falciparum. *iScience* **24**, 102444 (2021).

24. Flueck, C. *et al.* A Major Role for the Plasmodium falciparum ApiAP2 Protein PfSIP2 in Chromosome End Biology. *PLoS Pathog* **6**, e1000784 (2010).

25. Sierra-Miranda, M. *et al.* PfAP2Tel, harbouring a non-canonical DNA-binding AP2 domain, binds to *Plasmodium falciparum* telomeres. *Cell Microbiol* **19**, e12742 (2017).

26. Broadbent, K. M. *et al.* A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol* **12**, R56 (2011).

27. Nasiri-Aghdam, M., Garcia-Garduño, T. & Jave-Suárez, L. CELF Family Proteins in Cancer: Highlights on the RNA-Binding Protein/Noncoding RNA Regulatory Axis. *Int J Mol Sci* **22**, 11056 (2021).

28. Zhang, M. *et al.* Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis. *Science* **360**, eaap7847 (2018).

29. Bushell, E. *et al.* Functional Profiling of a Plasmodium Genome Reveals an Abundance of Essential Genes. *Cell* **170**, 260-272.e8 (2017).

30. Stanway, R. R. *et al.* Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage. *Cell* **179**, 1112-1128.e26 (2019).

31. Swanson, S. K., Florens, L. & Washburn, M. P. Generation and analysis of multidimensional protein identification technology datasets. *Methods Mol Biol* **492**, 1–20 (2009).

32. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–6 (2008).

33. Park, C. Y., Klammer, A. A., Käll, L., MacCoss, M. J. & Noble, W. S. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *J Proteome Res* **7**, 3022–3027 (2008).

34. Howbert, J. J. & Noble, W. S. Computing Exact p-values for a Cross-correlation Shotgun Proteomics Score Function. *Molecular & Cellular Proteomics* **13**, 2467–2479 (2014).

35. May, D. H., Tamura, K. & Noble, W. S. Param-Medic: A Tool for Improving MS/MS Database Search Yield by Optimizing Parameter Settings. *J Proteome Res* **16**, 1817–1824 (2017).

36. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923–925 (2007).

37. Paoletti, A. C. *et al.* Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proceedings of the National Academy of Sciences* **103**, 18928–18933 (2006).

38. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

39. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983–1992 (2014).

40. Wong, W. *et al.* Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife* **3**, (2014).

41. Volz, J. C. *et al.* Essential Role of the PfRh5/PfRipr/CyRPA Complex during Plasmodium falciparum Invasion of Erythrocytes. *Cell Host Microbe* **20**, 60–71 (2016).

42. Ginsburg, H. Malaria Parasite Metabolic Pathways. https://mpmp.huji.ac.il/ (2023).

43. Ferreira, J. L. *et al.* The Dynamic Roles of the Inner Membrane Complex in the Multiple Stages of the Malaria Parasite. *Front Cell Infect Microbiol* **10**, (2021).

44. Elsworth, B. *et al.* Proteomic analysis reveals novel proteins associated with the *Plasmodium* protein exporter PTEX and a loss of complex stability upon truncation of the core PTEX component, PTEX150. *Cell Microbiol* **18**, 1551–1569 (2016).

45. Hollin, T. *et al.* Functional genomics of RAP proteins and their role in mitoribosome regulation in Plasmodium falciparum. *Nat Commun* **13**, (2022).

46. Evers, F. *et al.* Composition and stage dynamics of mitochondrial complexes in Plasmodium falciparum. *Nat Commun* **12**, 3820 (2021).

47. Tarique, M., Ahmad, M., Ansari, A. & Tuteja, R. Plasmodium falciparum DOZI, an RNA helicase interacts with eIF4E. *Gene* **522**, 46–59 (2013).

48. Zhang, Q. *et al.* Proteomic analysis of exported chaperone/co-chaperone complexes of P. falciparum reveals an array of complex protein-protein interactions. *Sci Rep* **7**, 42188 (2017).

49. Mallari, J. P., Oksman, A., Vaupel, B. & Goldberg, D. E. Kinase-associated Endopeptidase 1 (Kae1) Participates in an Atypical Ribosome-associated Complex in the Apicoplast of Plasmodium falciparum. *Journal of Biological Chemistry* **289**, 30025–30039 (2014).

50. Ho, C.-M. *et al.* Native structure of the RhopH complex, a key determinant of malaria parasite nutrient acquisition. *Proceedings of the National Academy of Sciences* **118**, e2100514118 (2021).

51. Morano, A. A. & Dvorin, J. D. The Ringleaders: Understanding the Apicomplexan Basal Complex Through Comparison to Established Contractile Ring Systems. *Front Cell Infect Microbiol* **11**, 269 (2021).

52. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1**, 21–26 (2002).

53. Choi, H., Kim, S., Fermin, D., Tsou, C.-C. & Nesvizhskii, A. I. QPROT: Statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. *J Proteomics* **129**, 121–126 (2015).

54. Hollin, T., Abel, S. & le Roch, K. G. Genome-Wide Analysis of RNA–Protein Interactions in Plasmodium falciparum Using eCLIP-Seq. *Methods in Molecular Biology* **2369**, 139–164 (2021).

55. Wen, Z. kite: a software suite for processing and analysis of tandem mass spectrometry data. *Zenodo* v1.0.0 (2022) doi:10.5281/zenodo.591488

Supplementary Material

Since writing of this dissertation, the final paper has been published. Supplementary material for chapter 4 is available with the published paper at https://www.nature.com/articles/s41467-024-45519-1.

Chapter 5: Functional genomics of RAP proteins and their role in mitoribosome regulation in Plasmodium falciparum

Thomas Hollin[1], Steven Abel[1], Alejandra Falla[2], Charisse Flerida A. Pasaje[2], Anil Bhatia[3], Manhoi Hur[3], Jay S. Kirkwood[3], Anita Saraf[4], Jacques Prudhomme[1], Amancio De Souza[3], Laurence Florens[4], Jacquin C. Niles[2], and Karine G. Le Roch[1]

[1]Department of Molecular, Cell and Systems Biology, University of California Riverside, Riverside, CA, USA.
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.
[3]Metabolomics Core Facility, University of California, Riverside, CA 92521, USA.
[4]Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA

A version of this chapter has been published in *Nature Communications*, 2022.

Preface

In addition to developing a screen to identify RNA-dependent and RNA-binding proteins (R-DeeP, Chapter 4), we performed additional experimentation to characterize two proteins that had been identified in a previous screen as parasite specific mRNA-binding proteins, called RAPs as RNA binding proteins abundant in Apicomplexan parasites. These RAP have been shown to be involved in RNA mitochondrial metabolism in some model organisms. We performed a full range of experiments to characterize two of these RAPs, which we called PfRAP01 and PfRAP21, including the first eCLIP-seq experiment done in *Plasmodium* research. eCLIP-seq is a complex protocol that uses antibodies and crosslinking to reveal what RNA in the transcriptome is bound by a given protein. Most crucially, this technique showed that the two RAP proteins bind to the mitochondrial genome rather than the main nuclear genome, and small RNA sequencing showed that mitochondrial rRNAs are differentially expressed when PfRAP21 is knocked out. I performed much of the computational analysis for this project. I analyzed transcriptomic (RNA-seq) data which showed that the knockout of these proteins (most strikingly, PfRAP21) has a wide-ranging effect on parasite gene regulation, including upregulation of several other known RAP proteins in a possible attempt to compensate for the knockout (Figure 5.4). I also analyzed the eCLIP-seq and small RNA sequencing data mentioned above, which provided the key to understanding the exact functions of these proteins, showing at nucleotide resolution where they act on the mitochondrial genome (Figure 5.6). As important regulators, these proteins could represent drug targets against the malaria parasite.

Abstract

The RAP (RNA-binding domain abundant in Apicomplexans) protein family has been identified in various organisms. Despite expansion of this protein family in apicomplexan parasites, their main biological functions remain unknown. In this study, we use inducible knockdown studies in the human malaria parasite, Plasmodium falciparum, to show that two RAP proteins, PF3D7_0105200 (PfRAP01) and PF3D7_1470600 (PfRAP21), are essential for parasite survival and localize to the mitochondrion. Using transcriptomics, metabolomics, and proteomics profiling experiments, we further demonstrate that these RAP proteins are involved in mitochondrial RNA metabolism. Using high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (eCLIP-seq), we validate that PfRAP01 and PfRAP21 are true RNA-binding proteins and interact specifically with mitochondrial rRNAs. Finally, mitochondrial enrichment experiments followed by deep sequencing of small RNAs demonstrate that PfRAP21 controls mitochondrial rRNA expression. Collectively, our results establish the role of these RAP proteins in mitoribosome activity and contribute to further understanding this protein family in malaria parasites.

Introduction

RNAs associate with RNA-binding proteins (RBPs) to regulate a wide range of essential processes in eukaryotes. Diverse RNA-binding domains (RBDs) have been identified, including the RNA-recognition motif, zinc finger, K Homology domain, and Pumilio homology domain[1,2]. These domains are involved in a variety of cellular processes including splicing, mRNA processing, stability, and translation[3]. While several RBDs are

well conserved in eukaryotes, some domains have species-specific roles. This is particularly true for a domain known as RBD abundant in Apicomplexans or RAP. This domain was originally described as specifically enriched in apicomplexan genomes with 11–15 members in Plasmodium spp. and Toxoplasma gondii, while only 4–6 members (annotated as FASTK) were identified in mammalian genomes[4]. Recent studies reported that the enrichment is even higher than initially thought, with 21–23 RAP proteins in the apicomplexan parasites[5–7]. The exact function of the RAP proteins remains to be determined. In eukaryotes, RAP proteins have been predicted to be involved in RNA binding[4] and seem to be linked to organelle function since most of the characterized proteins have been located in mitochondria[8,9] or chloroplasts[10–12]. Several reports in model organisms have also confirmed their role in RNA metabolism in humans and plants[9,10,12–17]. Structurally, RAP proteins are divided into a variable N-terminal region composed of helical repeat protein motifs and the RAP domain on the C-terminal side[7]. These repeated sequences are stacked together and form a superhelix with an RNA-binding groove[18,19]. They have been identified in some RAP proteins as Heptatricopeptide repeat (HPR) motif[20], structurally and functionally related to Octotricopeptide and Pentatricopeptide repeat (PPR) motifs. Most of these proteins are predicted to target organelles and play different roles in RNA metabolism and translation[12,18–21]. Regarding the RAP domain, the conserved part is composed of ~60 amino acids and has an α/β topology, similar to restriction endonuclease-like folds, with multiple aromatic and charged residues[4,7,14].

Plasmodium falciparum, the deadliest human malaria parasite, possesses 22 potential RAP proteins7, 18 of which are putatively essential for the asexual parasite survival22, indicating the crucial role of this family. Despite the potential of these proteins as novel therapeutic targets, they remain poorly characterized. Experimental capture of RBPs in the blood stages of P. falciparum using oligo d(T) beads captured 199 proteins including two predicted RAP proteins, PF3D7_0105200 (PfRAP01) and PF3D7_1470600 (PfRAP21)5. Here, using transgenic lines in which PfRAP01 and PfRAP21 expression was conditionally regulated, we validated the essentiality of these proteins and showed their localization to the parasite mitochondrion. Furthermore, using transcriptomics, metabolomics, and proteomics profiling experiments, we demonstrated that down-regulation of these RAP proteins affects distinct RNA biology and mitochondrial processes. Finally, using high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (eCLIP-seq) and mitochondrial enrichment followed by sequencing of small RNAs, we confirmed these proteins are true RBPs that bind mitochondrial rRNAs in situ, thus validating their role in parasite mitoribosome metabolism.

Results

Generation of PF3D7_0105200 (PfRAP01) and PF3D7_1470600 (PfRAP21) transgenic lines

Comparative protein sequence analysis revealed that PF3D7_0105200 (PfRAP01) is highly conserved across Plasmodium spp., with 72% and 76% of identity with its homologs in P. vivax and P. berghei, respectively (Supplementary Data 1).

PF3D7_1470600 (PfRAP21) is also conserved in Plasmodium spp., but has lower sequence identity with P. vivax (38%) and P. berghei (38%).

Previous failed attempts to disrupt PfRAP01 and PfRAP21 in P. falciparum22 and their respective homologs in P. berghei23 suggested that both proteins are essential. To gain insights into the function of these proteins, we used an inducible knockdown based on the TetR-DOZI/RNA aptamer system, in which translation levels of the target protein is controlled by anhydrotetracycline (aTc)24–26. The lines were generated using CRISPR/Cas9 and homology directed repair in the NF54::pCRISPRINT line with T7 RNA polymerase and SpCas9 integrated at the cg6 locus25 (Fig. 5.1a, c). PfRAP01 and PfRAP21 were modified to include a C-terminal 3x-HA tag during the genome editing step used to install the regulatory components needed to achieve conditional expression. Clonal lines, PfRAP01 F2 and F4 for PF3D7_0105200, and PfRAP21 G1 and G4 for PF3D7_1470600, were obtained by limiting dilution and used in subsequent studies. The expected integration events in the PfRAP01 and PfRAP21 loci were validated by PCR (Fig. 5.1b, d). We also performed whole-genome sequencing of the parental, PfRAP01, and PfRAP21 clones and confirmed the correct insertion of the inducible system (Supplementary Fig. 1). No major deletions or duplications were detected. Mutations in var genes and a nonsense mutation in ap2-g were observed in the parental, PfRAP01, and PfRAP21 clones, validating the absence of off target editing. The ap2-g mutation is consistent with lack of gametocyte production in all of these parental and transgenic cell lines.

We verified the expression of PfRAP01 and PfRAP21 across the intraerythrocytic development cycle (IDC) by Western blot analysis using the 3x-HA tag (Fig. 5.1e, f). We observed a peak of expression at the mature trophozoite stage for both RAP proteins, in correlation with the RNA expression profiles available[27–29]. Collectively, these data confirmed successful generation of PfRAP01 and PfRAP21 transgenic lines.

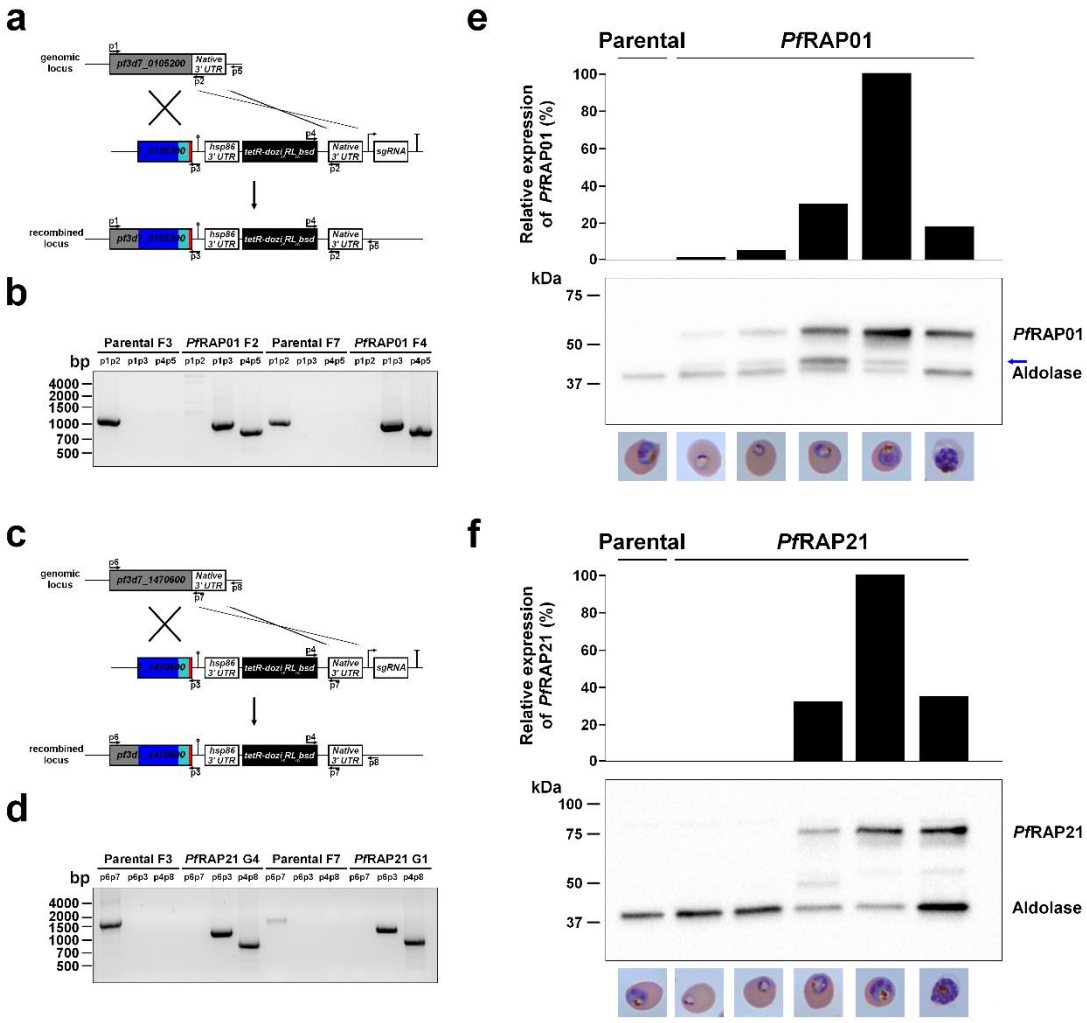

Fig. 5.1 Validation of PfRAP01 and PfRAP21 transgenic lines. Schematic representation of the TetR-aptamer integration strategy into endogenous pf3d7_0105200 (a) and pf3d7_1470600 (c). TetR-dozi-BSD cassette (black), HA-tag (red), PCR primers, homologous regions used (blue) and recodonized sequences (light blue) are indicated. Genotype analyses of parental, PfRAP01 (b) and PfRAP21 (d) clone lines. PCRs were performed on genomic DNA from indicated lines using different primer combinations. The PCRs are representative of three independent experiments. Immunoblot detection of PfRAP01 (e) and PfRAP21 (f) expression across the asexual blood cycle. The membranes were probed with anti-HA and anti-aldolase was used as loading control. The expression of PfRAP01 and PfRAP21 were normalized by aldolase expression. The blue arrow indicates PfRAP01 degradation. Scale bar: 2 μm. The immunoblots shown are representative of two independent experiments.

Mitochondrial localization of PfRAP01 and PfRAP21

In accordance with the predominant localization of RAP proteins to the mitochondria or plastid, PfRAP01 and PfRAP21 have been predicted to localize to the parasite mitochondrion (Supplementary Data 1). To validate this prediction, we performed anti-HA immunofluorescence assays (IFAs) to detect the 3x-HA tagged PfRAP01 and PfRAP21 proteins. We showed that PfRAP01 and PfRAP21 both co-localized with MitoTracker, a red-fluorescent dye that stains mitochondria in live parasites (Fig. 5.2a, b). For PfRAP01, this is consistent with the mitochondrial localization of the P. berghei ortholog, PBANKA_020810020. The lack of PfRAP01 and PfRAP21 detection by IFA in rings and early trophozoites correlates with their respective expression across the asexual cycle (Fig. 5.1e, f). The lack of HA-signal on the parental line confirmed the specificity of these IFAs (Supplementary Fig. 2a).

Using an anti-Cpn60, an apicoplast chaperonin30,31, IFAs revealed a distinct localization between the apicoplast and PfRAP01 or PfRAP21 indicating that they are not directly associated with the plastid (Supplementary Fig. 2b). Pearson's coefficient confirmed this result with a mean score at 0.31 and 0.84 for Cpn60 and MitoTracker signals, respectively (Supplementary Fig. 2c).

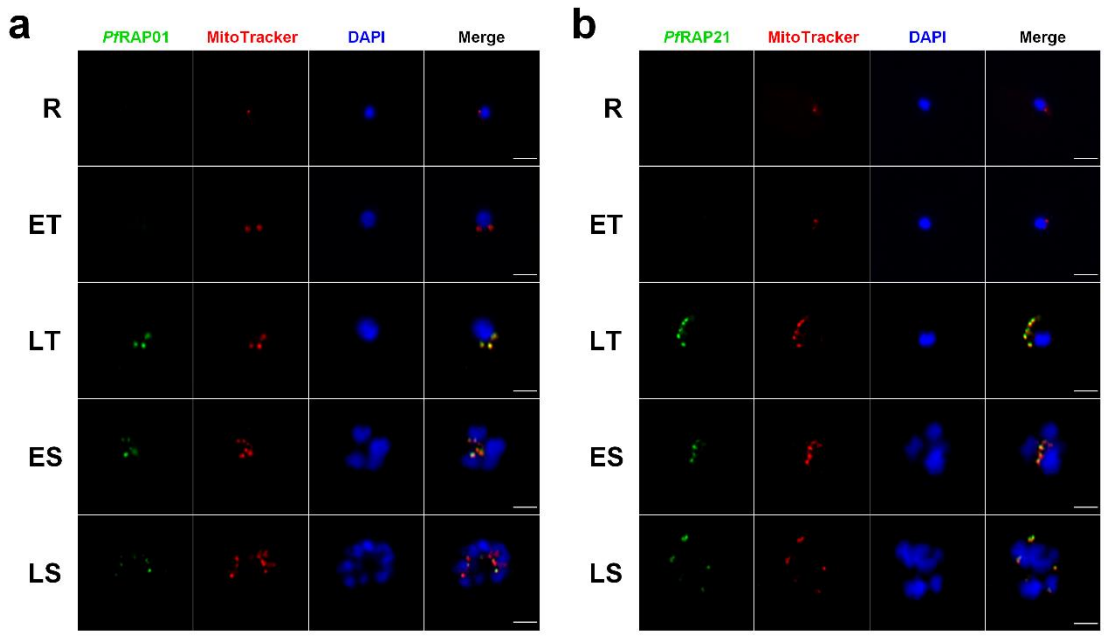

Fig. 5.2 Localization of PfRAP01 and PfRAP21 in asexual blood stages of P. falciparum. Immunofluorescence assays of PfRAP01 (a) and PfRAP21 (b) on ring (R), early trophozoite (ET), late trophozoite (LT), early schizont (ES) and late schizont (LS) stages. Both proteins were labeled with anti-HA and colocalized with parasite mitochondria stained with MitoTracker. Merge shows HA tag, MitoTracker and DAPI signals. Scale bar: 3 μm. The IFAs are representative of three independent experiments.

Essentiality of PfRAP01 and PfRAP21 in erythrocytic cycle

Constitutive knockouts of PfRAP01, PfRAP21, and their respective orthologs in P. berghei have been previously unsuccessfully attempted, suggesting essentiality of these proteins during asexual blood stages22,23. Therefore, we took advantage of our transgenic lines and verified the efficiency of our inducible system by Western blot analysis (Fig. 5.3a, b). Beginning with synchronous ring-stage cultures, we observed that PfRAP01 and PfRAP21 expression remained significantly stable at 24 h post aTc removal (trophozoite from the first IDC) (Fig. 5.3a, b). After 72 h (trophozoite from the

second IDC), which corresponds to the beginning of the transcription (Fig. 5.1e, f), we detected a significant reduction in protein levels for both PfRAP01 and PfRAP21 with a reduction of 77% and 97% respectively when compared to lines cultured with aTc. We then performed IFA studies to simultaneously detect RAP levels (anti-HA) and mitochondrial labeling (MitoTracker) at 72 h post aTc removal. We observed a decrease of 76% and 87% of double positive cells for PfRAP01 and PfRAP21, respectively (Supplementary Fig. 3a–d).

To study further the kinetics of parasite growth upon PfRAP01 and PfRAP21 depletion, tightly synchronized ring-stage parasites (t = 0 h) were maintained either with or without aTc. We observed no change in parasitemia during the first IDC for the PfRAP01 knockdown compared to the parental and control lines (Fig. 5.3c). This result was confirmed by the absence of phenotypic change in the first 56 h and ability of parasites under knockdown conditions to reinvade red blood cells with the same efficiency as the controls (Fig. 5.3e). At 128 h, however, a substantial reduction in parasite proliferation was observed for the RAP-deficient parasites. Microscopic analysis showed that majority of these parasites were blocked in the trophozoite stage during the second IDC (80 h) and began to die while the control parasites continued schizogony and were able to reinvade indicating that PfRAP01 is essential for the IDC. Similar results were obtained upon PfRAP21 knockdown. The parasitemia and developmental stages remained identical throughout the first IDC (Fig. 5.3d, f). Between 72 and 80 h, corresponding to the transcription peak, parasites were mainly blocked at the trophozoite stage and quickly lost viability resulting in a substantial drop in parasitemia. The changes observed with

PfRAP21 appear to be even more drastic than those obtained with PfRAP01, which is consistent with the different knockdown efficiencies detected by Western blot analysis (Fig. 5.3a, b). Altogether, these data confirmed the independent essentiality of PfRAP01 and PfRAP21 for P. falciparum survival. Since clones behaved identically in these studies, the following experiments were performed using only the PfRAP01 F2 and PfRAP21 G4 clonal lines.

To determine if the effects resulting from depletion of RAP proteins during the first IDC can be reversible, PfRAP01 and PfRAP21 parasites were replenished in aTc at 32, 56, or 72 h after initial aTc removal at 0 h (Supplementary Fig. 3e). Quantification of parasitemia using SYBR Green assays at 80 and 128 h showed that the deficiency in PfRAP01 and PfRAP21 can be completely reversed until 56 h (Supplementary Fig. 3f, g). At 128 h, a partial recovery was only obtained for PfRAP01 after a replenishment at 72 h, confirming the more drastic depletion for PfRAP21 as previously shown. This result indicates that no irreversible effects were suffered by the parasites until 56 h leading to complete recovery and pursuit of the IDC.

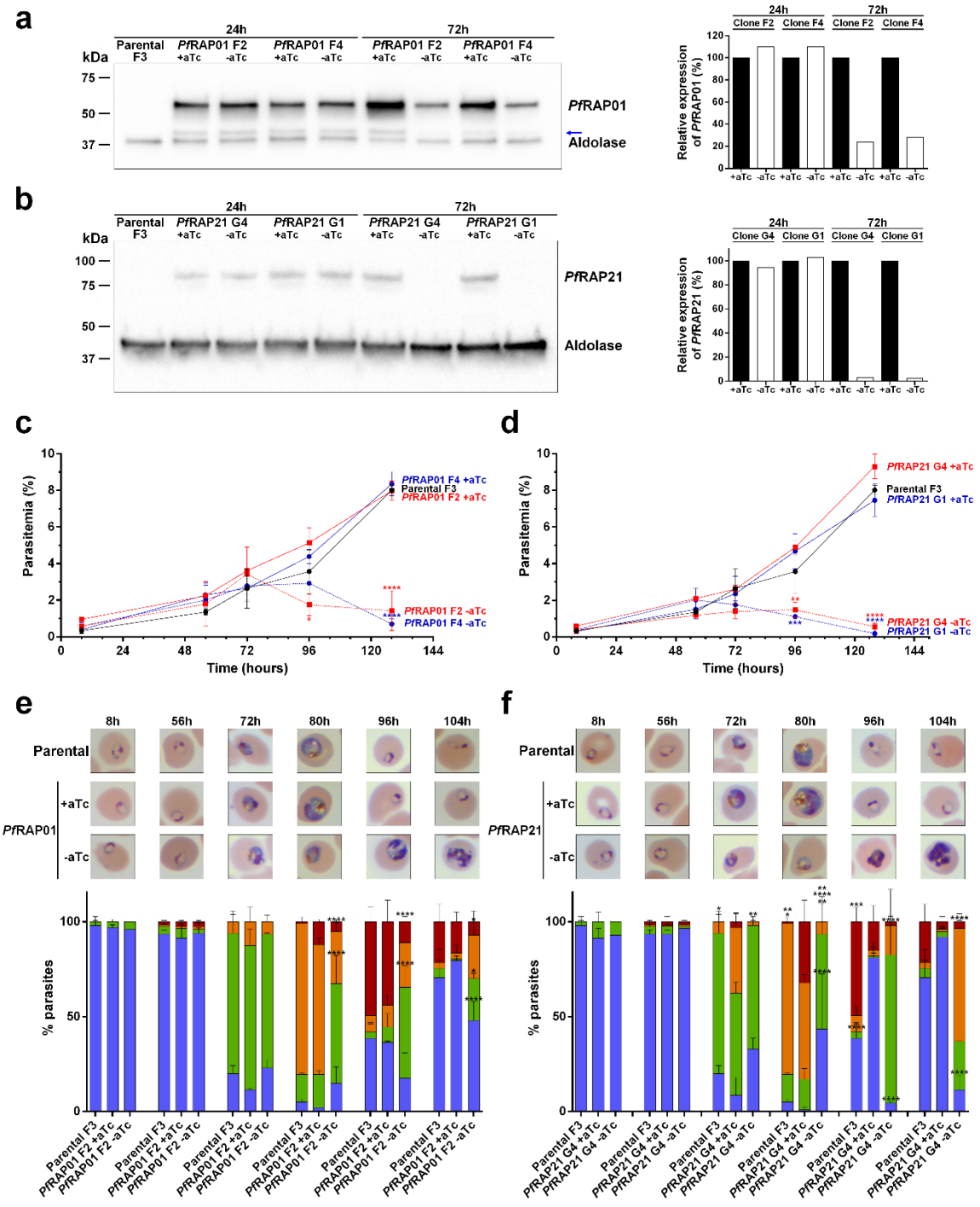
a
24h
72h
Parental F3
PfRAP01 F2
PfRAP01 F4
+aTc
-aTc
kDa
75
50
37
PfRAP01
Aldolase
Clone F2
Clone F4
Relative expression of PfRAP01 (%)
b
Parental F3
PfRAP21 G4
PfRAP21 G1
100
PfRAP21
Clone G4
Clone G1
Relative expression of PfRAP21 (%)
c
Parasitemia (%)
Time (hours)
PfRAP01 F4 +aTc
Parental F3
PfRAP01 F2 +aTc
PfRAP01 F2 -aTc
PfRAP01 F4 -aTc
d
PfRAP21 G4 +aTc
PfRAP21 G1 +aTc
PfRAP21 G4 -aTc
PfRAP21 G1 -aTc
e
8h
56h
72h
80h
96h
104h
Parental
PfRAP01
% parasites
f
PfRAP21

Fig. 5.3 Essentiality of PfRAP01 and PfRAP21 in asexual blood stages. Immunoblot detection of PfRAP01 F2 and PfRAP01 F4 (a), PfRAP21 G1 and PfRAP21 G4 (b) expression in presence or absence of aTc at 24 and 72 h. The membranes were probed with anti-HA and anti-aldolase was used as loading control. The immunoblots are representative of two independent experiments performed. The expression of PfRAP01 and PfRAP21 were normalized by aldolase expression and +aTc condition was considered as 100%. The blue arrow indicates PfRAP01 degradation. Parasitemia of Parental F3, PfRAP01 F2 and PfRAP01 F4 (c), PfRAP21 G1 and PfRAP21 G4 (d) was measured in presence or absence of aTc. Parasitemia was assessed on ten fields and in duplicate. The results of one representative experiment out of two, are shown as the mean parasitemia ± SEM (Two-way ANOVA and Tukey's multiple comparison test, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ and ****$p < 0.0001$ compared to clones +aTc at the same time point). Phenotypic analysis of Parental F3, PfRAP01 F2 (e) and PfRAP21 G4 (f) lines. Percentages of rings (blue), trophozoites (green), late trophozoites (orange), and schizonts (red) are indicated for each time point (n =51–106 parasites counted, mean ± SD). Times indicated correspond to hours post aTc removal. (Two-way ANOVA and Tukey's multiple comparison test, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ and ****$p < 0.0001$ compared to clones + aTc at the same time point). Scale bar: 2 μm.

Transcriptomic analysis of PfRAP01 and PfRAP21 knockdown parasites

To explore the global transcriptomic effects of PfRAP01 and PfRAP21 knockdowns, we performed RNA-seq on the parental line, and the PfRAP01 and PfRAP21 knockdown lines with and without aTc. We selected four time points corresponding to trophozoites from the first IDC (32 h) as well as the rings (56 h), trophozoites (72 h), and trophozoites/schizonts (80 h) from the second IDC (Fig. 5.4a). Two biological samples were generated for each condition. Spearman correlation coefficients ranged between 0.81 to 0.96 for each replicate, demonstrating the reproducibility of our experiment (Supplementary Data 2). Differentially expressed genes were then identified using the DEseq2 software (see "Methods" for further details). Comparative analysis between the parental and PfRAP01 + aTc or PfRAP21 + aTc lines showed that almost no genes were affected (median = 13.5 genes) during the first and second IDCs, confirming that genetic modification of the RAP loci did not grossly alter the parasite's transcriptome (Supplementary Data 2). However, this was not the case at ring stage (56 h) of the second IDC, wherein comparison of the parental and RAP + aTc lines revealed a significantly higher number of differentially expressed genes. Gene Ontology (GO) terms enrichment analysis indicated that most of these genes were involved in invasion, suggesting a potential phase shift in cell cycle between the different lines used.

We then analyzed PfRAP01 and PfRAP21 samples in presence and absence of aTc and grouped the genes based on whether they were significantly differentially expressed at 72 and 80 h (Fig. 5.4b, d). Significantly up-regulated genes were placed in cluster 1 and

downregulated genes were in cluster 2 (Supplementary Data 2). For PfRAP01 knockdown, only eight genes were up-regulated in cluster 1 (Fig. 5.4b). Although GO terms enrichment analysis indicated that glutathione biosynthetic process and response to oxidative stress pathways were significantly up-regulated, the low number of impacted genes was a hindrance to reaching any conclusion (Fig. 5.4c). This low significance likely results from a combination of filtering stringency and incomplete down-regulation (77%) of PfRAP01 (Fig. 5.3a). Cluster 2 contained 23 down-regulated genes, many of which are known to have a role in host-pathogen interactions and invasion. It is highly likely that these genes are identified due to the cell cycle arrest phenotype observed upon PfRAP01 knockdown and are not directly related to the PfRAP01 function. For PfRAP21, 392 genes were described as significantly up-regulated at 72 and 80 h (Fig. 5.4d). These genes are mainly associated with translation and RNA processing, such as the RNA helicases DBP1/7/8, mRNAdecapping enzymes DCP1 and DCP2, and multiple ribosomal proteins (Fig. 5.4e). While some of these retained transcripts could be associated with the cell cycle arrest, our result suggests a potential compensatory mechanism of the parasite in pathways affecting RNA biology and translational regulation. Interestingly, we detected 10 mRNAs for RAP proteins as up-regulated in PfRAP21 knockdown parasites at 72 and 80 h (Fig. 5.4g), including PfRAP01. Such enrichment could confirm a compensatory mechanism, albeit insufficient, of the RAP regulatory network to promote parasite survival upon knockdown of PfRAP21. A similar trend, to a lesser extent, was also observed for the PfRAP01 knockdown line (Fig. 5.4f).

Similar to PfRAP01, cluster 2 of the PfRAP21 samples (528 genes) contains transcripts related to host-pathogen and invasion processes (Fig. 5.4e).

Although not all statistically significant, the three mitochondrial protein-coding genes were detected with a median log2 Fold Change (FC) at −1.7 and −2.5 for PfRAP01 and PfRAP21, respectively (Supplementary Data 2), indicating an important impact on mitochondrial transcriptional regulation. This disturbance seems to have a less significant impact on nuclear genes predicted to be imported into the mitochondria32, since only 40 genes out of 295 were differentially expressed at 72 and 80 h for PfRAP21. As the transcription of these genes is under control of the nucleus, the RAP proteins are probably not directly involved in their regulation, unlike mitochondrial genes. Taken together, these data confirm a significant cell cycle arrest in PfRAP01 and PfRAP21 knockdowns and a possible association with RNA metabolism and translation for PfRAP21.

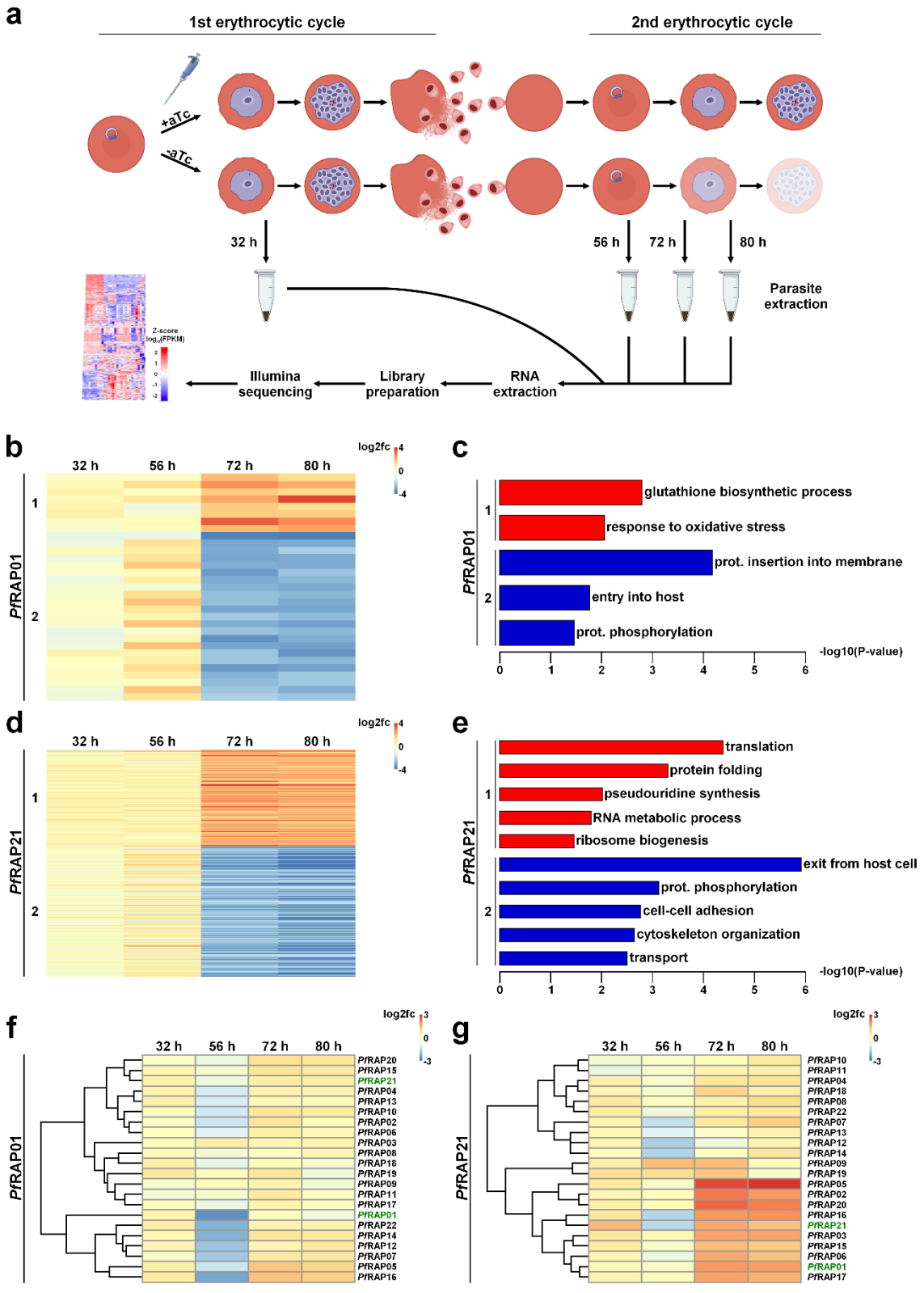
a
1st erythrocytic cycle
2nd erythrocytic cycle
+aTc
-aTc
32 h
56 h
72 h
80 h
Parasite extraction
RNA extraction
Library preparation
Illumina sequencing
Z-score
log10(FPKM)
b
log2fc
32 h
56 h
72 h
80 h
PfRAP01
1
2
c
PfRAP01
glutathione biosynthetic process
response to oxidative stress
prot. insertion into membrane
entry into host
prot. phosphorylation
-log10(P-value)
d
PfRAP21
e
PfRAP21
translation
protein folding
pseudouridine synthesis
RNA metabolic process
ribosome biogenesis
exit from host cell
prot. phosphorylation
cell-cell adhesion
cytoskeleton organization
transport
f
PfRAP01
PfRAP20
PfRAP15
PfRAP21
PfRAP04
PfRAP13
PfRAP10
PfRAP02
PfRAP06
PfRAP03
PfRAP08
PfRAP18
PfRAP19
PfRAP09
PfRAP11
PfRAP17
PfRAP01
PfRAP22
PfRAP14
PfRAP12
PfRAP07
PfRAP05
PfRAP16
g
PfRAP21
PfRAP10
PfRAP11
PfRAP04
PfRAP18
PfRAP08
PfRAP22
PfRAP07
PfRAP13
PfRAP12
PfRAP14
PfRAP09
PfRAP19
PfRAP05
PfRAP02
PfRAP20
PfRAP15
PfRAP21
PfRAP03
PfRAP15
PfRAP06
PfRAP01
PfRAP17

Fig. 5.4 Transcriptome profile of PfRAP01 and PfRAP21 knockdown parasites. a The schematic illustrates the key steps of the RNA-seq protocol. Created with BioRender.com. b–d Clustering of significantly affected genes from PfRAP01 (b) and PfRAP21 (d) knockdown parasites. Clusters 1 and 2 regroup genes significantly up-regulated and down-regulated respectively (two-tailed Wald test with Benjamini–Hochberg adjustment). c–e GO enrichment analysis of genes with −log10 (p value) for clusters 1 (red) and 2 (blue) (weight01 Fisher test). Heatmaps of the log2 FC values of predicted RAP proteins from PfRAP01 (f) and PfRAP21 (g) knockdown samples.

Metabolic perturbation of PfRAP01 and PfRAP21 knockdown parasites

To better understand the metabolic pathways affected by protein knockdown, we performed targeted metabolomics and untargeted lipidomics analyses to assess the effects of PfRAP01 and PfRAP21 loss. Synchronized rings were cultured in the presence or absence of aTc and triplicate samples were collected at 56, 72, and 80 h in the second IDC (Fig. 5.4a), corresponding to just before and during the death phenotype. We successfully detected 120 and 93 polar metabolites, and 558 and 511 lipids in the PfRAP01 and PfRAP21 samples, respectively (Supplementary Data 3). No major changes in ring stage were detected with only one and four metabolites/lipids significantly impacted in PfRAP01 and PfRAP21. For PfRAP21, 32 lipids showed higher abundance levels in deficient parasites at 72 and 80 h. These lipids correspond, among others, to 6 phosphatidylethanolamines (PEs), 6 triglycerides (TGs), 2 cardiolipins, and 3 sterols (Supplementary Data 3). As P. falciparum cannot synthesize sterols and scavenge them from the host, this upregulation is probably unrelated to the function of RAP proteins and due to the growth defect. Conversely, we identified a significant decrease in the relative abundance of 39 polar metabolites and 116 unique lipids, which could arise due to cell cycle arrest upon PfRAP21 knockdown. Mainly, these metabolites were composed of 24 diglycerides, 19 phosphatidylcholines (PCs), 16 TGs, 14 PEs, and 13 phosphatidylglycerol (PGs) (Supplementary Data 3). The lower abundance of CDP-ethanolamine, CDP-choline, and P-choline confirmed this global down-regulation of the phospholipid turnover (Fig. 5.5a and Supplementary Fig. 4b). Although inhibition of the mitochondrial electron transport chain has been shown to affect choline and PCs

abundance in neuroblastoma cells33, it is likely that this effect in our studies is a consequence of parasite death. However, the two most impacted pathways correspond to electron carriers (FAD and NAD) and polar metabolites involved in pyrimidine biosynthesis (Fig. 5.5a and Supplementary Fig. 4b), both of which being required for and/or dependent on mitochondrial functions. Detailed analysis also showed alteration in the abundance of 8 cardiolipins, which are essential constituents of mitochondrial membranes and involved in various mitochondrial processes. Two were significantly more abundant and one was not detected in the PfRAP21 -aTc samples, while the remaining five cardiolipins were only significantly impacted at 80 h (Fig. 5.5b). Altogether, these results suggest that the disruption of PfRAP21 affects overall metabolic activity with a higher impact on mitochondrion-dependent pathways. In the PfRAP01 knockdown parasites, we observed a significant increase in the relative abundance of 8 lipids (2 PEs, 2 PGs, 1 PC, 1 TG), and decrease of 8 polar metabolites and 21 lipids (Supplementary Data 3 and Supplementary Fig. 4a). Interestingly, 33 of these affected metabolites were similarly affected in PfRAP21 (Fig. 5.5c), confirming the same trend despite the lower efficiency of the knockdown system.

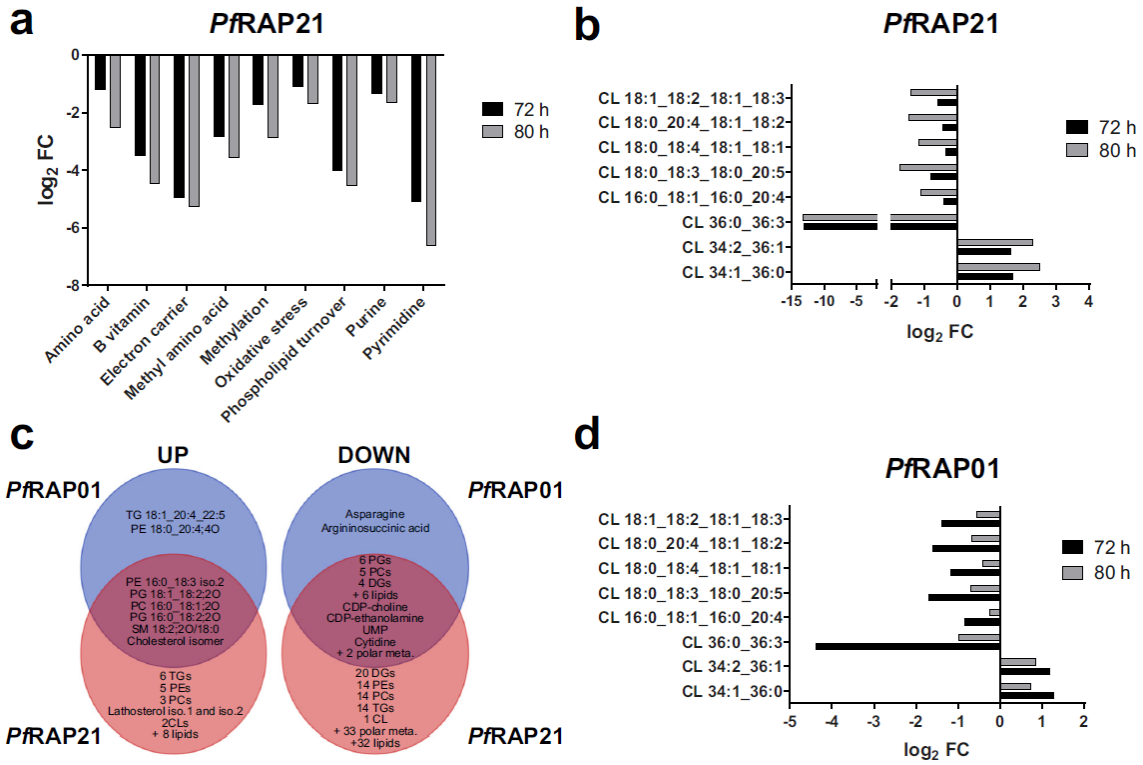

Fig. 5.5 Metabolomics analyses of PfRAP01 and PfRAP21 knockdowns. a Average relative log2 FC (−aTc/+aTc) of the polar metabolites grouped in their respective pathways from PfRAP21 samples at 72 and 80 h. b Relative log2 FC (−aTc/+aTc) of cardiolipins (CL) detected in lipidomics analysis of PfRAP21 lines. c Distribution of metabolites significantly affected in PfRAP01 and PfRAP21 lines. The total number of metabolites in each subgroup is indicated in red. d Relative log2 FC (−aTc/+aTc) of cardiolipins (CL) detected in lipidomics analysis of PfRAP01 lines.

Impact on mitochondrial electron transport in PfRAP01 and PfRAP21 knockdown parasites

To evaluate whether a defect in mitochondrial membrane potential (Δψm) occurred in PfRAP01 and PfRAP21 knockdown parasites, we performed live-cell imaging with MitoTracker staining. Mitochondrial structure was observed for PfRAP01 and PfRAP21 lines □} aTc at 72 and 80 h (Supplementary Fig. 5a–c). A higher number of tubular mitochondria was detected for parasites expressing PfRAP01 or PfRAP21 at both time points. However, this difference is likely due to the cell cycle arrest previously described. Indeed, mitochondrial morphology evolves during parasite maturation and tubular structures are observed in late asexual stages, as confirming by the higher number detected at 80 h compared to 72 h. No diffuse staining was observed as described previously with the knockdown of the mitochondrial ribosomal protein L13 (PfmRPL13)34.

To further explore the impact on mitochondrial metabolism, we carried out a differential sensitivity assay by treating the conditional knockdown lines with atovaquone, an inhibitor of the mitochondrial electron transport required to sustain de novo pyrimidine biosynthesis. We first determined the IC50 values obtained with aTc for PfRAP01 and PfRAP21 lines. For both proteins, we observed a steep dose-response curve suggesting an "all or nothing" action of the aTc with IC50 at 8 and 5 nM for PfRAP01 and PfRAP21, respectively (Supplementary Fig. 5d). We then performed drug inhibition assays beginning with ring-stage parasites with high (500 nM), low (IC values), or in absence of

aTc. Analysis of atovaquone dose-response curves revealed no change in sensitivity of either PfRAP01 or PfRAP21 in high and low aTc conditions (Supplementary Fig. 5e, f), similar to an aptamer-regulated YFP control line (Supplementary Data 4). This absence of hypersensitivity could be due to sufficient expression of each RAP protein at these concentrations of aTc. However, parasites were too affected without aTc to assess the impact of atovaquone. Additionally, with two control compounds that do not directly target mitochondrial functions, quinine and Bafilomycin A1 (a V-type ATPase inhibitor), no change in parasite sensitivity was observed upon PfRAP01 or PfRAP21 knockdown. Thus, conditional depletion of PfRAP01 and PfRAP21 does not appear to increase parasite sensitivity to atovaquone. Knockdown of the mitochondrial PfmtRPS17 protein, with a close phenotype, showed hypersensitivity in a nonspecific manner indicating a minor roles of mitochondrial drugs on severe knockdowns35.

Identification of RAP-protein complexes

To examine further the functions of our RAP proteins, we sought to determine their potential protein and RNA-binding partners. We performed anti- HA immunoprecipitation (IP) and mass spectrometry analysis on soluble protein fractions extracted from the PfRAP01-HA, PfRAP21- HA, and the parental lines (negative control) (Supplementary Data 5). Proteins were filtered with QSPEC-calculated Log FC ≥ 1 and p value ≤ 0.05 compared to values measured from the parental line. With an average dNSAF value of 0.21, PfRAP01 was by far the most abundant protein in its HA-affinity purification (Supplementary Fig. 6a). One other protein, PF3D7_0106300 ATP6, a

calcium transporting ATPase, was detected at a much lower level (about 100 times less). As ATP6 is not predicted to be a mitochondrial protein based on its sequence analysis by MitoProt II (Supplementary Data 5), it is unlikely to be a viable interacting partner. While the PfRAP21 bait was detected at lower levels (its spectral counts contributed <1% of the total spectra in its HA purification), three proteins (PF3D7_0703500, PF3D7_1237100, and PF3D7_0924200) were reproducibly and significantly enriched in the replicate PfRAP21 IP samples (Supplementary Fig. 6b). All three proteins were recovered at abundance levels similar to PfRAP21 and are predicted to localize to the mitochondrion. While annotated as an erythrocyte membrane-associated antigen, PF3D7_0703500 has one N-terminal membrane anchor and a DEAD-like helicase superfamily domain suggesting this protein could be involved in protein transport and/or mRNA-binding. PF3D7_1237100 is also predicted to contain several transmembrane domains. For PF3D7_0924200, the protein exhibits a heptatricopeptide-repeat domain, most likely interacting with RNA20. Although these findings will require further validation, they suggest that PfRAP01 may operate independently while PfRAP21 might be part of a complex to perform their critical mitochondrial functions.

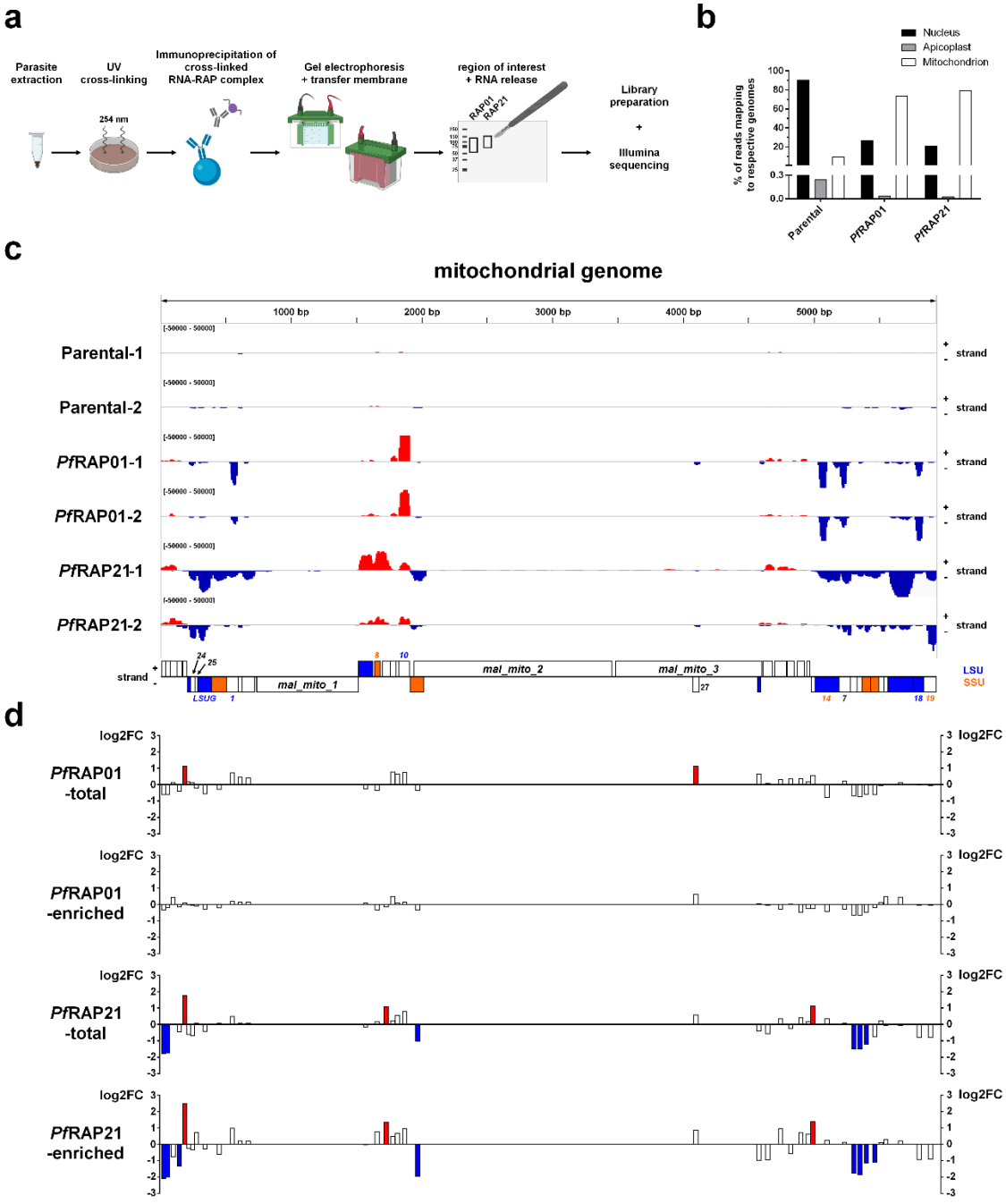
a
Parasite extraction
UV cross-linking
254 nm
Immunoprecipitation of cross-linked RNA-RAP complex
Gel electrophoresis + transfer membrane
region of interest + RNA release
RAP01
RAP21
Library preparation
+
Illumina sequencing
b
Nucleus
Apicoplast
Mitochondrion
% of reads mapping to respective genomes
Parental
PfRAP01
PfRAP21
c
mitochondrial genome
1000 bp
2000 bp
3000 bp
4000 bp
5000 bp
Parental-1
Parental-2
PfRAP01-1
PfRAP01-2
PfRAP21-1
PfRAP21-2
strand
mal_mito_1
mal_mito_2
mal_mito_3
LSU
SSU
d
log2FC
PfRAP01 -total
PfRAP01 -enriched
PfRAP21 -total
PfRAP21 -enriched

Fig. 5.6 PfRAP01 and PfRAP21 participate in mitoribosome regulation. a The schematic illustrates the key steps of the eCLIP-seq protocol. Created with BioRender.com. b Distribution of the reads to the nuclear and mitochondrial genome. c Read density tracks along the mitochondrial genome for parental, PfRAP01, and PfRAP21 samples (n = 2). SSU =small subunit (orange), LSU =large subunit (blue). d Log2 FC values determined for mitochondrial rRNAs from total and enriched RNA of PfRAP01 and PfRAP21 lines. rRNAs identified as differentially up-regulated and down-regulated (log2 FC > 1 or <1) are respectively indicated in red and blue.

Identification of RAP-RNA complexes

To identify potential RNAs directly interacting with PfRAP01 and PfRAP21 proteins, we performed the first eCLIP-seq experiment in P. falciparum. Briefly, RNA-protein complexes from the parental, HA-tagged PfRAP01 and PfRAP21 lines, were UV-crosslinked and immunoprecipitated (Fig. 5.6a). After gel electrophoresis and transfer onto nitrocellulose membrane, RNAs were released by proteinase K treatment for library preparation and high-throughput sequencing on the Illumina NOVAseq platform. eCLIP-seq experiments were performed in duplicate. Approximately 76% of the reads from the PfRAP01 and PfRAP21 lines mapped to the mitochondrial genome, compared to 10% from the parental control, validating significant enrichment of mitochondrial reads in all RAP samples (Fig. 5.6b). Using the MACS2 peak caller, we observed that the non-mitochondrial peaks detected were mainly related to rRNAs, tRNAs, and snRNAs (Supplementary Data 6). However, analysis of normalized read counts indicated that most of these genes were also covered in the parental samples (Supplementary Fig. 7a). MACS2 detected the entire mitochondrial genome with the highest −log10 (q value) in both RAP samples (x2.8 higher than the 2nd highest peak) (Supplementary Data 6). Detailed analysis of normalized read counts showed that although the entire mitochondrial genome was called, PfRAP01 and PfRAP21 interacted mainly and specifically with rRNAs (Fig. 5.6c). PfRAP01 mainly interacted with rRNA1, rRNA3, rRNA7, and large subunit ribosomal RNA (LSU) fragments A and D, while PfRAP21 bound to rRNA3, rRNA7, rRNA8, rRNA24t, rRNA25t, small subunit ribosomal RNA (SSU) fragment E, and LSU fragments F and G. These rRNAs are associated with both

SSU and LSU rRNAs35,36, known to form the mitoribosome, suggesting that binding is not subunitspecific. Likewise, the orientation of the reads showed a nearly perfect correlation with the positioning of the genes on the plus and minus strands (Fig. 5.6c), validating that the RAP proteins interacted directly with the rRNAs and not the intergenic regions or opposite strand of the mitochondrial genome. These rRNAs are completely covered, with the exception of rRNA1, where PfRAP01 reads map mainly to the 5′ side, and rRNA7 and LSU fragments A and D where PfRAP01 reads map to the 3′ end. No significant RNA motifs representing potential PfRAP01 or PfRAP21 binding sites could be identified using the MEME suite37. Loss of PfRAP01 and PfRAP21 proteins could potentially result in mitoribosome destabilization, leading to alteration of mitochondrial functions and initiate a cascade of events lethal to the parasite. Altogether, the eCLIP-seq data confirmed that PfRAP01 and PfRAP21 are true RBPs, and allowed identification of their RNA targets and their importance in mitoribosome function.

Dysregulation of mitoribosome expression in PfRAP01 and PfRAP21 knockdown parasites

To investigate the impact of PfRAP01 and PfRAP21 knockdowns on mitochondrial rRNAs, we performed small RNA library preparation from total parasites and from organelles isolated by N2 cavitation (=enriched RNA). The samples were prepared in duplicate at 72 h after aTc removal and the libraries were size-selected for cDNA fragments between 140 and 350 nucleotides before sequencing (Supplementary Fig. 8a). Approximately 40% of the reads from the enriched RNA mapped to the mitochondrial

genome, compared to 12% from the total RNA, validating significant enrichment of mitochondrial reads (Supplementary Fig. 8b, c). In all samples, high numbers of reads were mapped to rRNAs and tRNAs (32–59% of total reads) (Supplementary Fig. 8d). Among the 45 nuclear tRNAs, 25% were up-regulated (log2 FC > 1) in both RNA preparations in PfRAP21-deficient parasites (Supplementary Data 7), confirming the increase observed for the genes involved in translation in our RNA-seq data. The levels of rRNA expression were similar for the two different RNA isolation methods, validating the results obtained. As expected, mitochondrial reads mapped predominantly to rRNAs. Upon knockdown of PfRAP21, 6 rRNAs were down-regulated (rRNA16, rRNA20, rRNA11, rRNA22, and SSU fragments A and D) while rRNA23t, rRNA2 rRNA13 were up-regulated (Fig. 5.6d). The majority of these rRNAs showed a similar trend in PfRAP01-deficient parasites, although only rRNA23t and rRNA27t showed differential expression in total RNA samples (Fig. 5.6d). This low impact can be explained by the incomplete down-regulation of PfRAP01 as previously noticed (Fig. 5.3a). The results suggest that the level of these ribosomal subunits can be controlled by each RAP protein (Supplementary Fig. 8e). It is however important to note that the most affected rRNAs did not perfectly match the strongest eCLIP-seq binding peaks, which appear to be positioned upstream. This result suggests that the binding of RAP proteins upstream of targeted rRNAs might be necessary to fulfill their function, although this hypothesis will need to be further investigated. Despite the low coverage of nuclear genes, only one rRNA (PF3D7_0112500) was significantly impacted in both conditions in PfRAP21 samples indicating that the RAP protein knockdowns had almost no impact on nuclear

ribosomes (Supplementary Data 7). The knockdowns of PfRAP01 and PfRAP21 thus showed an imbalance in the mitochondrial rRNAs, which could prompt a defect in the mitoribosome function.

Discussion

Mitochondrial genomes in apicomplexan parasites present atypical features, with only three protein-coding genes and multiple highly fragmented rRNAs[36,38,39]. This strong reduction of the genome size requires import of nuclear-encoded proteins to ensure the various essential functions of this organelle[40–42]. In P. falciparum, which has one of the smallest mitochondrial genomes[43,44], hundreds of proteins are imported through a system requiring transit peptides and import machinery involving TOM and TIM proteins[45,46]. Understanding the exact function of parasite specific proteins that are imported to the mitochondria could lead to discovery of new therapeutic strategies. Here, we characterized two RAP proteins encoded by the nuclear genome, but localized to the parasite mitochondrion. Their localization is consistent with the protein family, whose members are associated with mitochondria or plastids in various organisms[7–12]. IFA studies indicated that PfRAP01 and PfRAP21 are expressed in late asexual stages, consistent with Western blot analysis and previous transcriptomic data[27–29]. Using CRISPR/Cas9 genome editing, we generated two inducible knockdown transgenic lines and showed that both RAP proteins are essential for the asexual blood stages, validating previous large-scale screening data in P. falciparum and P. berghei[22,23]. Parasite growth arrest was observed 72–80 h after aTc removal in the trophozoite stage of the subsequent

IDC. The latency period seems to be due to significant loss of RAP proteins, which is only achieved during the second IDC, as confirmed by the growth after replenishment at 56 h. A similar delayed response was also observed with conditional knockdown of PfmtRPS17, a mitochondrial ribosomal protein35. An even higher number of cycles was necessary to observe parasite growth defect after knockdown of two other mitochondrial proteins, PfmRPL1334 and PfmtRPS1235.

Transcriptomics analysis performed to determine the pathways impacted in the knockdown parasites led to an arrest in cell cycle and overexpression of genes implicated in RNA biology, confirming the predicted involvement of the RAP proteins in this particular pathway. Several other RAP proteins were also significantly up-regulated, potentially to partially compensate for lack of PfRAP01 and PfRAP21. Although not statistically significant, the three mitochondrial protein-coding genes were down-regulated, suggesting potential involvement of PfRAP01 and PfRAP21 proteins in the regulation of their expression.

Although it is difficult to distinguish direct and indirect effects of protein knockdown by metabolomics profiling, our results indicate that PfRAP21 is most likely associated with mitochondrial dysfunction since knockdown led to a decrease in the levels of electron carriers and polar metabolites involved in pyrimidine biosynthesis. Although few metabolites were detected as significantly altered by PfRAP01 knockdown, we noticed a trend wherein levels of the same compounds were perturbed in both lines. Metabolic profiling of PfmtRPS17 showed a decrease in abundance of ten metabolites associated

with the pyrimidine de novo synthesis pathway in the absence of aTc35, consistent with our results. Only two intermediates, N-carbamoyl-l-aspartate and dihydroorotate, accumulated in the deficient parasites, but these were not detected in this study.

Although destabilization of mitochondrial electron transport was not detected upon PfRAP01 and PfRAP21 knockdown, an effect on this pathway cannot be excluded. In previous studies, hypersensitivity to the mitochondrial drugs (atovaquone, DSM265, and proguanil) was detected with PfRPS1235 and PfmRPL1334, two mitochondrial proteins whose knockdown caused a medium loss of fitness with effects on parasite growth occurring only 6–8 days after aTc removal. PfmtRPS17, with a rapid onset phenotype similar to PfRAP01 and PfRAP21, was hypersensitive to all inhibitors in a nonspecific manner confirming the difficulty to detect this synergy with severe knockdowns35.

IP-MS experiments identified three candidate partners (PF3D7_0703500, PF3D7_1237100, and PF3D7_0924200) of PfRAP21. Although limited information is available on these proteins, they are predicted to be localized to the mitochondrion and potentially associated with RNA metabolism. Additional experiments will be necessary to validate these partners and exclude the possibility of false-positive interactions or indirect association through RNA bridging. However, it is interesting to note that PF3D7_0924200 is a HPR protein and HPR domains have also been identified in other RAP proteins, including PfRAP01, and could participate in RNA binding20. Although several RAP proteins do not have detectable HPR or PPR motifs, the N-terminal helical structure regions are conserved between all RAP proteins7. Further studies will be

required to assess the contribution of these regions to RNA binding, and to understand what are the critical interactions for regulating proper mitoribosome function.

Adaptation of the eCLIP-seq methodology to Plasmodium allowed us to identify and characterize the RNA targets of PfRAP01 and PfRAP21 proteins in situ. We successfully demonstrated that each protein binds to distinct mitochondrial rRNA transcripts associated with the small subunit (SSU) and large subunit (LSU) rRNAs, which are thought to form the mitoribosome. To further validate our eCLIP-seq results, we sequenced small RNAs from total and mitochondria-enriched RNA samples. While only a few rRNAs were significantly affected by PfRAP01 knockdown, 9 mitochondrial rRNAs were significantly up- and down-regulated for PfRAP21. This difference could arise due to the more efficient depletion of PfRAP21 compared to PfRAP01. The discrepancies observed between the eCLIP-seq and the small RNA-seq results might indicate that the RAP proteins recognize upstream regions of the targeted rRNAs. It is also possible that these proteins interact on polycistronic precursor transcripts prior to maturation events, leading to the high number of binding sites identified by eCLIP-seq. However, a role during the assembly of mitoribosome can also not be excluded. Identifying these binding sites could clarify how and why each protein recognizes and targets these rRNAs. Overall, eCLIP-seq and small RNA-seq validated the involvement of PfRAP01 and PfRAP21 in regulation of the parasite mitoribosome.

In Apicomplexa, the mitochondrial genome exhibits atypically high fragmentation, with 39 rRNAs and only three proteincoding genes[36,38,39]. This contrasts with the human

mitochondrial genome, which has 2 rRNAs and 13 genes encoding the protein subunits of respiratory complexes. The considerable expansion of the number of RAP proteins in apicomplexan parasites may reflect a requirement for supplemental proteins to counteract the rRNA fragmentation and to allow assembly of a functional mitoribosome. Some alveolates such as Tetrahymena thermophila and Symbiodinium microadriaticum, or green alga (e.g., Gonium pectorale or Monoraphidium neglectum) corroborate this interpretation since their mitochondrial genome is fragmented to a lesser degree and have an intermediate number of RAP proteins7. The deficiency in PfRAP01 and PfRAP21 could hamper the stability of fragmented rRNAs and/or mitoribosome assembly leading to down-regulation of the three mitochondrial protein-coding genes as detected in our RNA-seq. This could destabilize mitochondrial respiration and associated metabolic pathways leading ultimately to parasite death. Altogether our results confirm an essential role of PfRAP01 and PfRAP21 proteins in regulation of the parasite's mitoribosome. Although the RAP domain is evolutionarily conserved7, the high variability in the N-terminal regions of these proteins, especially with human FASTKs, could open new opportunities for selectively targeting the Plasmodium proteins for therapeutic purposes.

Materials and Methods

Transfections and cultures of P. falciparum

The P. falciparum lines were grown as previously described47 in 5% of human O + erythrocytes. Parasites were synchronized by two D-sorbitol treatments48 and liberated from red blood cells with 0.15% saponin49.

Using CRISPR/Cas9, we generated parasite lines where we fused the pf3d7_0105200 and pf3d7_1470600 transcripts with a C-terminal 3x-HA epitope tag and RNA aptamers for translational regulation with the TetR-DOZI module25,26. To construct the modification plasmids, we cloned the target specifying sequences, the right homology regions, and the left homology regions (LHR) of each gene into the pSN053 linear vector25 via Gibson assembly. The homology arms were PCR amplified and gBlock synthesized (IDT-DNA) recodonized 3′-end of each target genes were fused to the LHR to prevent cleavage of the modified locus. The target specifying guide RNA sequences were generated by Klenow reaction. In addition to the gene expression regulation module, the pSN053 vector also features the reporter construct Renilla luciferase (Rluc) and the selection marker blasticidin S-deaminase gene. The final donor vectors were confirmed by restriction digests and Sanger sequencing. Primers used in this study are listed in Supplementary Data 8.

Transfections into Cas9- and T7 RNA polymerase-expressing NF54 parasites were carried out by preloading erythrocytes with the donor plasmids as described

previously50. Cultures were maintained in 500 nM anhydrotetracycline (Sigma- Aldrich, 37919) and 2.5 μg/ml of Blasticidin (RPI Corp B12150-0.1). Parasite cell lines stably integrating the donor plasmids were monitored via Giemsa smears and Rluc measurements using the Renilla-Glo® Luciferase Assay System (Promega E2750) and the GloMax® Discover Multimode Microplate Reader (Promega).

Transgenic lines validation

Parental and inducible lines were cloned by limiting dilution. Infected red blood cells were lysed with DNeasy Blood & Tissue kit (Qiagen) to extract genomic DNA. Primers used to genotype pf3d7_0105200-HA and pf3d7_1470600-HA are listed in Supplementary Data 8.

For whole-genome sequencing, genomic DNAs were fragmented using Covaris ultrasonicator and libraries were generated using KAPA LTP Library Preparation Kit (Roche, KK8230). Edited genomes were constructed by adding the desired inserted sequences into the P. falciparum genome, version 43 (http://plasmodb.org). The reads were aligned to these edited genomes using Bowtie2 with default settings51. Reads that aligned with mapping quality of 40 or below were filtered out using Samtools52 (Supplementary Data 9). Alignments were visualized on IGV53. Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number PRJNA690830.

The expression of PfRAP01 and PfRAP21 across the asexual cycle and at trophozoite stages (24 and 72 h after aTc removal) was assessed by Western blot. Blots were probed with anti-HA antibody (1:2500, Abcam, ab9110) and anti- Plasmodium aldolase (1:10000, Abcam, ab207494) followed by HRP-labeled Goat anti-Rabbit IgG (H + L) (1:10,000, NovexTM, A16104). Chemiluminescence detection with ClarityTM Western ECL Substrate (Bio-Rad, 1705060) was applied to reveal the blots. Relative abundance of the RAP proteins was normalized with Pf-aldolase expression using Image Lab software (Bio-Rad).

Immunofluorescence assays

PfRAP01-HA and PfRAP21-HA parasites □} aTc were washed in incomplete medium then incubated with 0.5 µM of MitoTrackerTM Red CM-H2Xros (Invitrogen, M7513) for 30 min at 37 °C. Parasites were washed in incomplete medium and fixed with 4% paraformaldehyde and 0.0075% glutaraldehyde for 15 min at 4 °C, then sedimented on Poly-D-lysine coated coverslips for 1 h at room temperature. After PBS washing, cells were permeabilized and saturated with 0.2% Triton X-100, 5% BSA, 0.1% Tween 20 in PBS for 30 min at room temperature. Anti-HA mAb (Abcam, ab24779) was diluted at 1:500 in 5% BSA, 0.1% Tween 20 and PBS, and applied for 1 h at room temperature, followed by Goat anti-Mouse Alexa Fluor 488 (1:2000, Invitrogen, A11001) for 1 h at room temperature. The coverslips were mounted in Vectashield Antifade Mounting Medium with DAPI (H-1200). The rabbit anti-Cpn60 (kindly provided by Dr. Boris Striepen) was used in the same condition at 1:1000 with Donkey anti-Rabbit Alexa Fluor

568 (1:2000, Invitrogen, A10042). Images were acquired using Zeiss LSM880 microscope with Airyscan (Fig. 5.2) or Leica DMI 6000 (Supplementary Figs. 2 and 3) and treated with ImageJ. Co-localizations of PfRAP01 and PfRAP21- HA signals with MitoTracker or Cpn60 were quantified on trophozoite stage by measuring the Pearson correlation coefficient □} SD using JACoP (n = 10).

Parasitemia and phenotype analyses

To determine the essentiality of the RAP proteins, we synchronized Parental F3, PfRAP01 F2 and F4, and PfRAP21 G4 and G1 lines. Parasitemia and proportion of the different asexual blood stages were determined by counting Giemsa-stained blood smears, under the microscope at the indicated time points (n = 51–106 parasites counted in duplicate). For aTc replenishment, synchronous rings were cultured with or without aTc. At 32, 56, or 72 h, parasites were replenished with 500 nM of aTc. The viability of the cultures was assessed by DNA quantification using SYBR Green (Thermo Fisher, S7523) at 80 and 128 h.

RNA-sequencing

Parasites in ring, trophozoite, and schizont stages (5 × 108 cells) were extracted by saponin treatment before flash freezing. Two independent biological replicates were generated for each time point, culture condition, and line. Total RNA was extracted with TRIzol® LS Reagent (Invitrogen, 10296028) then incubated for 1 h at 37 °C with 4 units of DNase I (NEB, M0303). RNA samples were visualized by RNA electrophoresis and

quantified on SynergyTM HT (Bio- Tek). Then mRNAs were purified using NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB, E7490) according to the manufacturer's instructions. Libraries were prepared using NEBNext® UltraTM Directional RNA Library Prep Kit (NEB, E7420L). Final libraries were amplified by PCR with KAPA HiFi Hot- Start Ready Mix (KAPA Biosystems, KK2602) and the PCR conditions consisted of 15 min at 37 °C followed by 12 cycles of [98 °C (30 s), 55 °C (10 s) and 62 °C (1 min 15)], finished by 5 min at 62 °C. The quantity and quality of the final libraries were assessed using a Bioanalyzer (Agilent Technology Inc). All samples were multiplexed and sequenced on 100 nucleotides paired-end run on the Illumina NovaSeq 6000 sequencer at the UC San Diego IGM Genomics Center to produce at least 10 million of reads per sample (Supplementary Data 9). FastQC54 was used to assess raw read quality. Adapter sequences as well as the first 11 bp of each read were removed using Trimmomatic55. Tails of reads were trimmed using Sickle56 with a Phred base quality threshold of 20, and reads shorter than 18 bp were removed. To remove human RNA contamination, reads were then aligned against the H. sapiens genome (assembly GRCh38) using Bowtie2 (version 2.3.4.1)51, and unmapped reads were retained. These reads were then aligned to the P. falciparum genome (version 43, http://plasmodb.org) using HISAT257. Only properly paired reads with a mapping quality score of 40 or higher were retained, with filtering done using Samtools52. Raw read counts were determined for each gene in the P. falciparum genome using BedTools58 multicov to intersect the aligned reads with the genome annotation. DESeq259 was then used for differential expression analysis. Only the genes with log2 FC > 1 or <1 and adjusted p

value <0.05 at 72 and 80 h were included in heatmaps. R package pheatmap60 was used to generate heatmaps. PlasmoDB was used for GO enrichment analysis. Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number PRJNA690830.

Metabolomics sample preparation

Tightly synchronized parasites (5 × 109 parasites at 56 h and 9 × 108 parasites at 72 and 80 h), were lysed with saponin, flash frozen and stored at −80 °C. Lipids and polar metabolites were extracted from malaria pellets using a biphasic approach. To each sample, 1ml of ice cold 3:2 methyl tert-butyl ether:80% methanol was added. To break up malaria pellets, samples were vortexed 2 min, sonicated for 15 min, vortexed for 2 min, sonicated for 15 min, then vortexed for 30 min at 4 °C. All sonication was performed in an ice bath. In total, 200 µl of water was added to induce phase separation, followed by a 5 min vortex. After centrifugation for 15 min at 4 °C at 16,000 × g, 200 µl of the top, nonpolar layer was transferred to a 2 ml glass vial and the bottom, polar layer was transferred to a new 2ml glass vial then analyzed by LC-MS. The nonpolar fraction was dried under a gentle stream of nitrogen at room temperature then resuspended in 400 µl of 9:1 methanol:toluene and analyzed by LC-MS.

LC-MS lipidomics

LC-MS lipidomics analysis was performed at the UC Riverside Metabolomics Core Facility as described previously61, with minor modifications. Briefly, analysis was

performed on a Waters G2-XS quadrupole time-of-flight mass spectrometer coupled to a Waters Acquity I-class UPLC system. Separations were carried out on a Waters CSH C18 column (2.1 × 100 mm, 1.7 μM). The mobile phases were (A) 60:40 acetonitrile:water with 10mM ammonium formate and 0.1% formic acid and (B) 90:10 isopropanol:acetonitrile with 10mM ammonium formate and 0.1% formic acid. The flow rate was 400 μl/min and the column was held at 65 °C. The injection volume was 2 μl. The gradient was as follows: 0 min, 10% B; 1 min, 10% B; 3 min, 20% B; 5 min, 40% B; 16 min, 80% B; 18 min, 99% B; 20 min 99% B; 20.5 min, 10% B. The MS was scan range was (50–1600 m/z) with a 100 ms scan time. MS/MS was acquired in data dependent fashion. Source and desolvation temperatures were 150 °C and 600 °C, respectively. Desolvation gas was set to 1100 l/h and cone gas to 150 l/h. All gases were nitrogen except the collision gas, which was argon. Capillary voltage was 1 kV in positive ion mode. A quality control sample, generated by pooling equal aliquots of each sample, was analyzed periodically to monitor system stability and performance. Samples were analyzed in random order. Leucine enkephalin was infused and used for mass correction.

LC-MS metabolomics—polar metabolites

Targeted metabolomics of polar, primary metabolites was performed on a TQ-XS triple quadrupole mass spectrometer (Waters) coupled to an I-class UPLC system (Waters). Separations were carried out on a ZIC-pHILIC column (2.1 × 150 mm, 5 μM) (EMD Millipore, 150460). The mobile phases were (A) water with 15mM ammonium

bicarbonate adjusted to pH 9.6 with ammonium hydroxide and (B) acetonitrile. The flow rate was 200 µl/min and the column was held at 50 °C. The injection volume was 2 µl. The gradient was as follows for PfRAP01: 0 min, 90% B; 1.5 min, 90% B; 16 min, 20% B; 18 min, 20% B; 20 min, 90% B; 28 min, 90% B, and as follows for PfRAP21: 0 min, 90% B; 1.5 min, 90% B; 16 min, 10% B; 18 min, 10% B; 20 min, 90% B; 28 min, 90% B.

The MS was operated in selected reaction monitoring mode62. Source and desolvation temperatures were 150 °C and 600 °C respectively. Desolvation gas was set to 1100 l/h and cone gas to 150 l/h. Collision gas was set to 0.15 ml/min. All gases were nitrogen except the collision gas, which was argon. Capillary voltage was 1 kV in positive ion mode and 2 kV in negative ion mode. System stability was monitored by analyzing a quality control sample (generated by pooling together equal volumes of all sample extracts) every 3 injections. Samples were analyzed in random order.

Data processing and analysis

Targeted data processing (manual peak integration) was performed in Skyline software63. Untargeted data processing (peak picking, alignment, deconvolution, integration, normalization, and spectral matching) was performed in Progenesis Qi software (Nonlinear Dynamics). Metabolomics data were normalized by parasite count and lipidomics data were normalized to total ion abundance. Lipidomics features with a CV greater than 30% across QC injections were removed64,65. To aid in the identification of features that belong to the same metabolite, features were assigned a

cluster ID using RAMClust66. An extension of the metabolomics standard initiative guidelines was used to assign annotation level confidence67,68. Annotation level 2a indicates an MS and MS/MS match to an external database. Level 2b indicates an MS and MS/ MS match to the Lipidblast in silico database69 or an MS match and diagnostic evidence. Several mass spectral metabolite libraries were searched including those in Mass Bank of North America, Metlin70, and an in-house library.

Metabolites were described as significantly affected if p value <0.05 and log2 FC < −1 or >1 at 56 h or both at 72 and 80 h.

Live-cell imaging

PfRAP01 and PfRAP21 knockdown lines were washed in incomplete medium then incubated with 0.5 μM of MitoTrackerTM Red CMH2Xros (Invitrogen, M7513) and 8 μM Hoechst 33342 (Invitrogen, H3570) for 30 min at 37 °C. Parasites were washed in incomplete medium and were mounted between slide and coverslip. Images were acquired using Leica DMI 6000 and treated with ImageJ. At least 30 parasites were used for the quantification.

Compound susceptibility assays

Serial dilutions of aTc, atovaquone, quinine, and bafilomycin A were generated to yield final concentrations ranging from 40–0.462 nM, 25–0.048 nM, 432–0.84 nM, and 40–0.08 nM, respectively. Synchronous ring-stage PfRAP01 and PfRAP21 conditional knockdown lines as well as a control cell line expressing an aptamer-regulatable

fluorescent protein were maintained in high aTc (500 nM), low aTc in the case of PfRAP01 (8 nM) and PfRAP21 (5 nM) or no aTc, and were distributed into 384-well assay plates (Corning®, 89176-442). Compounds were transferred to the parasite-containing plates using the Janus® platform (PerkinElmer). DMSO- and dihydroartemisinintreatment (500 nM) served as reference controls. Growth inhibition was analyzed after 72 and 120 h using the Renilla-Glo(R) Luciferase Assay System (Promega, E2750) and the GloMax® Discover Multimode Microplate Reader (Promega). IC50 values were obtained from corrected dose-response curves and plotted using GraphPad Prism 8.

Immunoprecipitation followed by MudPIT mass spectrometry

Purified late asexual parasites from parental, PfRAP01, and PfRAP21 lines (7.5 × 109 to 1.5 × 1010 cells) were suspended in 50mM Tris-HCl pH 7.5, 150mM NaCl, 5mM EDTA, 1% Triton X-100, 1 mM AEBSF and EDTA-free protease inhibitor cocktail (Roche, 11873580001). After lysis by passing 25 times through a 26G needle, the soluble extracts were treated with 100 units of DNase I (NEB, M0303) for 10 min at room temperature and centrifuged at 14,000 × g for 15 min at 4 °C. The lysates were precleared with DynabeadsTM Protein A (Invitrogen, 10001D) for 1 h at 4 °C. The Rb anti-HA antibody (1:100, Abcam, ab9110) was added in each sample and incubated overnight at 4 °C. DynabeadsTM Protein A were used to precipitate antibody-protein complexes and were washed with buffer A (1% Triton X-100, 1 mM EDTA in PBS), buffer B (wash buffer A, 0.5M NaCl) and buffer C (1mM EDTA in PBS). Proteins were eluted using 0.1M

glycine, pH 2.8, and neutralized using 2M Tris-HCl, pH 8.0. Then proteins were precipitated in 20% TCA followed by cold acetone washes.

TCA-precipitated proteins were urea-denatured, reduced, alkylated, and digested with endoproteinase Lys-C (Promega, V1671) followed by modified trypsin (Promega, V5111)71. Peptide mixtures were loaded onto 100 µm fused silica microcapillary columns packed with 5-µm C18 reverse phase (Aqua, Phenomenex), strong cation exchange resin (Luna, Phenomenex), and 5-µm C18 Aqua71. Loaded microcapillary columns were placed in-line with a Quaternary Agilent 1100 series HPLC pump and a LTQ linear ion trap mass spectrometer equipped with a nano- LC electrospray ionization source (Thermo Scientific, San Jose, CA). Fully automated 10-step MudPIT runs were carried out on the electrosprayed peptides, as previously described71. Tandem mass (MS/MS) spectra were interpreted using ProluCID72 v.1.3.3 against a database consisting of 5527 non-redundant (NR) Plasmodium falciparum 3D7 proteins (PlasmoDB-42 release), 36661 NR human proteins (NCBI, 2018-03-30 release), 419 usual contaminants (human keratins, IgGs, and proteolytic enzymes). To estimate false discovery rates (FDRs), the amino acid sequence of each NR protein entry was randomized, which resulted in a total search space of 85246 NR sequences. All cysteines were considered as fully carboxamidomethylated (+57 Da statically added), while methionine oxidation was searched as a differential modification. DTASelect73 v.1.9 and swallow v.0.0.1, an in-house developed software (https://github.com/tzw-wen/kite)74, were used to filter ProLuCID search results at given FDRs at the spectrum, peptide, and protein levels. Here, all controlled FDRs were less than 1.2%. All datasets were

contrasted against their merged data set, respectively, using Contrast v1.9 and in-house developed sandmartin v.0.0.1 (https://github.com/tzw-wen/kite/tree/master/ kitelinux)74. Our in-house developed software, NSAF7 v.0.0.1 (https://github.com/ tzw-wen/kite/tree/master/windowsapp/NSAF7x64)74, was used to generate spectral count-based label free quantitation results75. QSPEC76 was used to calculate log2 FC and p values to statistically compare PfRAP01 and PfRAP21 purifications to negative controls. Proteins with log2 FC ≥ 1 and $p \leq 0.05$ were considered significantly enriched in the RAP purifications (Supplementary Data 5).

eCLIP-seq

Late asexual parasites (7.5 × 109 to 1.5 × 1010 cells) were extracted by saponin lysis and were crosslinked on ice by 254 nm UV light for a total of 1200 mJ/cm2 with 2 min breaks using SpectrolinkerTM XL-1000. Ten µg of rabbit polyclonal HA tag antibody (abcam, ab9110) were coupled with DynabeadsTM M-280 Sheep Anti-Rabbit IgG (Thermo Fisher, 11203D) for 1 h at room temperature. The following steps were processed using eCLIP Library Prep Kit (Eclipse BioInnovations, ECEK-0001) according to the manufacturer's instructions. Briefly, cells were resuspended in 1ml of lysis buffer (50mM Tris-HCl pH 7.4, 100mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate, EDTA-free protease inhibitor cocktail (Roche, 11873580001) and 10 µL of Murine RNase inhibitor (NEB) in nuclease free water) and genomic DNA was sheared by sonication (Covaris ultrasonicator; 5 min, 5% duty cycle, 140 intensity peak incident power, 200 cycles per burst). After RNA fragmentation with 100 units of RNase-I

(Ambion, AM2294) for 5 min at 37 °C, lysates were incubated with antibodycoupled magnetic beads at 4 °C overnight. Two percent of each sample were saved prior to washes and correspond to negative control (Input). Immunoprecipitated (IP) samples were washed with high salt buffer (50mM Tris-HCl pH 7.4, 1M NaCl, 1mM EDTA, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate in nuclease free water) then with wash buffer (20mM Tris-HCl pH 7.4, 10mM MgCl2, 0.2% Tween 20, in nuclease free H2O). 5′ and 3′ RNA ends were repaired with PSP and PNK enzymes, followed by RNA adapter ligation. Protein-RNA complexes from IP and Input samples were eluted in loading buffer, separated by gel electrophoresis and transferred onto a nitrocellulose membrane at 4 °C overnight. The region comprising the protein band of interest up to an additional 75 kDa above (Parental samples: 40–150 kDa; PfRAP21: 40–150 kDa; PfRAP21: 60–175 kDa) were isolated and digested with proteinase K at 37 °C for 20 min then 50 °C for 20 min with interval mixing at 1200 rpm. RNA was cleaned and concentrated using Zymo RNA Clean & Concentrator kit (Zymo Research, R1015) and was reverse transcribed with AffinityScript enzyme (Agilent) at 54 °C for 20 min. After cDNA end repair, a 3’ ssDNA adapter was ligated, and qPCR was performed on Bio-Rad CFX Connect system. Libraries were amplified according to the Ct values obtained. PCR conditions consisted of 98 °C (30 s) followed by 6 cycles of (98 °C (15 s), 70 °C (30 s), 72 °C (40 s)), then (Ct-5) cycles of (98 °C (15 s), 72 °C (45 s)) and 72 °C (1 min). Libraries were loaded into a 3% agarose gel and regions between 175–350 bp were extracted and purified using MinElute Gel Extraction Kit (Qiagen, 28604) (Supplementary Fig. 7b). The quantity and quality of the final libraries were assessed

using a Bioanalyzer (Agilent Technology Inc). All samples were multiplexed and sequenced by dual indexed run (PE100) on the Illumina NovaSeq 6000 sequencer at the UC San Diego IGM Genomics Center to produce 6 million reads per sample (Supplementary Data 9). Only forward reads were used for the computational analysis. FastQC54 was used to assess raw read quality. The 10-bp random-mer sequences at the beginning of the forward reads allowed the use of Clumpify (BBTools)77 to remove PCR duplicates. The random-mer sequences as well as adapter sequences were then removed using Trimmomatic55. Tails of reads were trimmed using Sickle56 with a Phred base quality threshold of 20, and reads shorter than 18 bp were removed. Remaining reads were aligned to the P. falciparum genome using Bowtie2 (version 2.3.4.1)51 with default parameters. Reads with a mapping quality score below 40 were removed using Samtools52 (Supplementary Data 9). The resulting BAM files were converted to BED by bedtools bamtobed, and genome-wide per-nucleotide read counts were obtained using bedtools genomecov with the -d parameter58. To normalize, all read counts were divided by the number of millions of mapped reads for each particular sample. At each nucleotide across the genome, the read counts for each IP sample were then subtracted by the read counts for the corresponding input sample, with negative values being converted to 0. For visualization in IGV53, these final normalized counts were converted to WIG files by a custom Python script, then to TDF files by igvtools totdf. To show strand differences, two BAM files were made from each BAM file after mapping quality filtering, one with only positive-strand reads, and one with only negative-strand reads. This was done using Samtools. Each BAM file was carried through the remaining steps, and the two resulting

TDF files were combined into one IGV track. Peak calling was performed using MACS2 with the options --nomodel --extsize 150 --max-gap 1 -q 0.01. MACS2 was run without modeling due to the lack of a sufficient number of highly significant peaks for the model to be constructed. For each RAP protein, two IP samples were combined as the treatment reads, and two Input samples were combined as the control reads. Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number PRJNA690830.

Mitochondria enrichment

Isolation of mitochondria from PfRAP01 and PfRAP21 transgenic lines was performed as previously described with slight modifications78. Briefly, after 72 h of culture in presence of absence of aTc, ~1010 parasites were extracted by saponin treatment in 120mM KCl, 20mM NaCl, 20mM Glucose, 6mM HEPES, 6mM MOPS, 1mM MgCl2, 0.1mM EGTA, and pH 7. After a final wash, cells were resuspended in 225mM D-Mannitol (Sigma-Aldrich, M4125), 5mM Succinic acid (Sigma-Aldrich, S3674), 5mM L-(−)-Malic acid (Sigma- Aldrich, M1000), 75mM Sucrose, 4.3mM MgCl2, 10mM Tris, 0.25mM EGTA, 15mM HEPES, and pH 7.6. The parasites were disrupted by N2 cavitation after pressurization at 1000 psi for 20 min at 4 °C using a Cell Disruption Vessel (Parr Instrument Company, 4635). Cell debris were removed by centrifugation at 900 × g for 6 min at 4 °C and the supernatants were passed through LS Columns (Miltenyi Biotec, 130-042-401). After centrifugation at 23,000 × g for 20 min at 4 °C, the pellets were resuspended in TRIzol® LS Reagent (Invitrogen, 10296028).

Small RNA-sequencing

Mitochondrial and Total RNAs were purified as described above. RNA was first clean-up with Agencourt RNAclean XP beads (Beckman Coulter, A63987) and 1 µg of RNA was used for library preparation using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB E7300S/L). The following steps were processed according to the manufacturer's instructions. After PCR amplification (94 °C 30 s, followed by 12 cycles of 94 °C 15 s, 62 °C 30 s, 70 °C 15 s, then 70 °C for 5 min), a size selection was performed on a 6% TBE PAGE gel. cDNA fragments between ~145–350 bp were isolated and eluted. After ethanol precipitation, the libraries were assessed on a Bioanalyzer (Agilent Technology Inc). All samples were multiplexed and sequenced on 100 nucleotides paired-end run on the Illumina NovaSeq 6000 sequencer at the UC San Diego IGM Genomics Center to produce at least 10 million of reads per sample (Supplemental Data 9).

FastQC54 was used to assess raw read quality. While sequencing was pairedend, only forward reads were used in this analysis. Adapter sequences were removed using Trimmomatic55. Tails of reads were trimmed using Sickle56 with a Phred base quality threshold of 20, and reads shorter than 15 bp were removed. These reads were then aligned to the P. falciparum genome (version 48, http:// plasmodb.org) using Bowtie251 with seed length 15 (-L 15). Only reads with a mapping quality score of 40 or higher (high-quality unique alignments) were retained, with filtering done using Samtools52. Raw read counts were determined for each gene in the P. falciparum genome using

BedTools58 multicov to intersect the aligned reads with the genome annotation. Read counts were RPM-normalized by dividing by the number of millions of mapped reads for each library. R package pheatmap60 was used to generate heatmaps.

Statistics

Two-tailed Mann–Whitney U test was performed on co-localization quantification. Parasitemia and proportion of asexual stages were analyzed using a two-way ANOVA with Tukey's test for multiple comparisons. Replenishment of aTc was analyzed using a one-way ANOVA with Holm-Šidak correction. For DESeq2, the p values were obtained by two-tailed Wald test are corrected for multiple testing using the Benjamini–Hochberg correction. The GO terms enrichment analysis was determined using weight01 Fisher test. Two-tailed t-test was performed on metabolomics data. For the proteomics, the significance analysis was performed using QSPEC. MACS2 calculates a p value for each peak using Poisson distribution and q values are calculated using the Benjamini–Hochberg correction. Significant differences were indicated as following: * for $p < 0.05$; ** for $p < 0.01$, *** for $p < 0.001$ and **** for $p < 0.0001$. Statistical tests were performed with GraphPad Prism 6.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

WGS, RNA-seq, eCLIP-seq, and small RNA-seq datasets generated in this study have been deposited in the NCBI Sequence Read Archive under accession number PRJNA690830. The MS datasets have been deposited in the ProteomeXChange (PXD023308) via the MassIVE repository (MSV000086636 with [https://doi.org/10.25345/C5R795]), and may also be accessed from the Stowers Original Data Repository (http://www.stowers.org/research/ publications/libpb-1571). The metabolomics data generated in this study have been deposited in the PanoramaWeb [https://panoramaweb.org/Plasmodium_RAPprotein.url]. Source data are provided with this paper.

Code availability

The custom Python scripts used for eCLIP analysis have been previously published79. The entire in-house software suite (Kite) used for the MudPIT mass spectrometry analysis is available in Zenodo (https://doi.org/10.5281/zenodo.5914885)74.

References

1. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. 19, 327–341 (2018).

2. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. 8, 479–490 (2007).

3. Maris, C., Dominguez, C. & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. 272, 2118–2131 (2005).

4. Lee, I. & Hong, W. RAP—a putative RNA-binding domain. Trends Biochemical Sci. 29, 567–570 (2004).

5. Bunnik, E. M. et al. The mRNA-bound proteome of the human malaria parasite Plasmodium falciparum. Genome Biol. 17, 147 (2016).

6. Woo, Y. H. et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. Elife 4, 1–41 (2015).

7. Hollin, T., Jaroszewski, L., Stajich, J. E., Godzik, A. & Le Roch, K. G. Identification and phylogenetic analysis of RNA binding domain abundant in apicomplexans or RAP proteins. Microb. Genomics 7, 000541 (2021).

8. Simarro, M. et al. Fast kinase domain-containing protein 3 is a mitochondrial protein essential for cellular respiration. Biochem. Biophys. Res. Commun. 401, 440–446 (2010).

9. Jourdain, A. A. et al. A mitochondria-specific isoform of FASTK is present in mitochondrial RNA granules and regulates gene expression and function. Cell Rep. 10, 1110–1121 (2015).

10. Rivier, C., Goldschmidt-Clermont, M. & Rochaix, J. D. Identification of an RNA-protein complex involved in chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii. EMBO J. 20, 1765–1773 (2001).

11. Eberhard, S. et al. Dual functions of the nucleus-encoded factor TDA1 in trapping and translation activation of atpA transcripts in Chlamydomonas reinhardtii chloroplasts. Plant J. 67, 1055–1066 (2011).

12. Kleinknecht, L. et al. RAP, the sole octotricopeptide repeat protein in Arabidopsis, is required for chloroplast 16S rRNA maturation. Plant Cell 26, 777–787 (2014).

13. Jourdain, A. A. et al. Survey and summary: the FASTK family of proteins: emerging regulators of mitochondrial RNA biology. Nucleic Acids Res. 45, 10941–10947 (2017).

14. Boehm, E. et al. FASTKD1 and FASTKD4 have opposite effects on expression of specific mitochondrial RNAs, depending upon their endonuclease-like RAP domain. Nucleic Acids Res. 45, 6135–6146 (2017).

15. Boehm, E. et al. Role of FAST kinase domains 3 (FASTKD3) in posttranscriptional regulation of mitochondrial gene expression. J. Biol. Chem. 291, 25877–25887 (2016).

16. Antonicka, H. & Shoubridge, E. A. Mitochondrial RNA granules are centers for posttranscriptional RNA processing and ribosome biogenesis. Cell Rep. 10, 920–932 (2015).

17. Tian, Q., Taupin, J. L., Elledge, S., Kobertson, M. & Anderson, P. Fas-activated serine/threonine kinase (FAST) phosphorylates TIA-1 during fas-mediated apoptmis. J. Exp. Med. 182, 865–874 (1995).

18. Manna, S. An overview of pentatricopeptide repeat proteins and their applications. Biochimie 113, 93–99 (2015).

19. Barkan, A. & Small, I. Pentatricopeptide repeat proteins in plants. Annu. Rev. Plant Biol. 65, 415–442 (2014).

20. Hillebrand, A. et al. Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. Nucleic Acids Res. 46, 10417–10431 (2018).

21. Boulouis, A. et al. Spontaneous dominant mutations in chlamydomonas highlight ongoing evolution by gene diversification. Plant Cell 27, 984–1001 (2015).

22. Zhang, M. et al. Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis. Science 360, eaap7847 (2018).

23. Bushell, E. et al. Functional profiling of a plasmodium genome reveals an abundance of essential genes. Cell 170, 260–272.e8 (2017).

24. Goldfless, S. J., Wagner, J. C. & Niles, J. C. Versatile control of Plasmodium falciparum gene expression with an inducible protein-RNA interaction. Nat. Commun. 5, 5329 (2014).

25. Nasamu, A. S. et al. An integrated platform for genome engineering and gene expression perturbation in Plasmodium falciparum. Sci. Rep. 11, 342 (2021).

26. Ganesan, S. M., Falla, A., Goldfless, S. J., Nasamu, A. S. & Niles, J. C. Synthetic RNA-protein modules integrated with native translation mechanisms to control gene expression in malaria parasites. Nat. Commun. 7, 10727 (2016).

27. Toenhake, C. G. et al. Chromatin accessibility-based characterization of the gene regulatory network underlying plasmodium falciparum blood-stage development. Cell Host Microbe 23, 557–569.e9 (2018).

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

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

30. Florentin, A., Stephens, D. R., Brooks, C. F., Baptista, R. P. & Muralidharan, V. Plastid biogenesis in malaria parasites requires the interactions and catalytic activity of the Clp proteolytic system. Proc. Natl Acad. Sci. U. S. A. 117, 13719–13729 (2020).

31. Agrawal, S., van Dooren, G. G., Beatty, W. L. & Striepen, B. Genetic evidence that an endosymbiont-derived endoplasmic reticulum-associated protein degradation (ERAD) system functions in import of apicoplast proteins. J. Biol. Chem. 284, 33683–33691 (2009).

32. Esveld, S. L. V. et al. A prioritized and validated resource of mitochondrial proteins in Plasmodium identifies leads to unique biology. bioRxiv. https://doi.org/10.1101/2021.01.22.427784 (2021).

33. Baykal, A. T., Jain, M. R. & Li, H. Aberrant regulation of choline metabolism by mitochondrial electron transport system inhibition in neuroblastoma cells. Metabolomics 4, 347–356 (2008).

34. Ke, H., Dass, S., Morrisey, J. M., Mather, M. W. & Vaidya, A. B. The mitochondrial ribosomal protein L13 is critical for the structural and functional integrity of the mitochondrion in Plasmodium falciparum. J. Biol. Chem. 293, 8128–8137 (2018).

35. Ling, L. et al. Genetic ablation of the mitoribosome in the malaria parasite Plasmodium falciparum sensitizes it to antimalarials that target mitochondrial functions. J. Biol. Chem. 295, 7235–7248 (2020).

36. Feagin, J. E. et al. The fragmented mitochondrial ribosomal RNAs of Plasmodium falciparum. PLoS ONE 7, e38320 (2012).

37. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. Nucleic Acids Res. 43, W39–W49 (2015).

38. Hikosaka, K., Kita, K. & Tanabe, K. Diversity of mitochondrial genome structure in the phylum Apicomplexa. Mol. Biochem. Parasitol. 188, 26–33 (2013).

39. Hikosaka, K. et al. Highly conserved gene arrangement of the mitochondrial genomes of 23 Plasmodium species. Parasitol. Int. 60, 175–180 (2011).

40. Van Dooren, G. G., Stimmler, L. M. & McFadden, G. I. Metabolic maps and functions of the Plasmodium mitochondrion. FEMS Microbiol. Rev. 30, 596–630 (2006).

41. Hikosaka, K., Komatsuya, K., Suzuki, S. & Kita, K. Mitochondria of malaria parasites as a drug target. In An Overview of Tropical Diseases (ed Samie, A.) (InTech, 2015).

42. Vaidya, A. B. & Mather, M. W. Mitochondrial evolution and functions in malaria parasites. Annu. Rev. Microbiol. 63, 249–267 (2009).

43. Vaidya, A. B., Akella, R. & Suplick, K. Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite. Mol. Biochem. Parasitol. 35, 97–107 (1989).

44. Suplick, K., Akella, R., Saul, A. & Vaidya, A. B. Molecular cloning and partial sequence of a 5.8 kilobase pair repetitive DNA from Plasmodium falciparum. Mol. Biochem. Parasitol. 30, 289–290 (1988).

45. Deponte, M. Mitochondrial protein import in malaria parasites. In Encyclopedia of Malaria (eds Kremsner, Peter, G. & Krishna, S.) 1–13 (Springer, 2013).

46. Bender, A., Van Dooren, G. G., Ralph, S. A., Mcfadden, G. I. & Schneider, G. Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. Mol. Biochem. Parasitol. 132, 59–66 (2003).

47. Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. J. Parasitol. 91, 484–486 (2005).

48. Lambros, C. & Vanderberg, J. P. Synchronization of plasmodium falciparum erythrocytic stages in culture. J. Parasitol. 65, 418–420 (1979).

49. Umlas, J. & Fallon, J. N. New thick-film technique for malaria diagnosis. Use of saponin stromatolytic solution for lysis. Am. J. Trop. Med. Hyg. 20, 527–529 (1971).

50. Deitsch, K. W., Driskill, C. & Wellems, T. Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. Nucleic Acids Res. 29, 850–853 (2001).

51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

52. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

53. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192 (2013).

54. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

56. Joshi, N. & Fass, J. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle (2011).

57. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360 (2015).

58. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

60. Kolde, R. pheatmap: Pretty Heatmaps. R package version 0.7.7. http://cran.rproject.org/package=pheatmap (2013).

61. Reddam, A. et al. mRNA-sequencing identifies liver as a potential target organ for triphenyl phosphate in embryonic zebrafish. Toxicol. Sci. 172, 51–62 (2019).

62. Helou, D. G. et al. PD-1 pathway regulates ILC2 metabolism and PD-1 agonist treatment ameliorates airway hyperreactivity. Nat. Commun. 11, 3998 (2020).

63. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968 (2010).

64. Dunn, W. B. et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat. Protoc. 6, 1060–1083 (2011).

65. Barupal, D. K. et al. Generation and quality control of lipidomics data for the alzheimer's disease neuroimaging initiative cohort. Sci. Data 5, 1–13 (2018).

66. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A novel feature clustering method enables spectral-matching based annotation for metabolomics data. Anal. Chem. 86, 6812–6817 (2014).

67. Schymanski, E. L. et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ. Sci. Technol. 48, 2097–2098 (2014).

68. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3, 211–221 (2007).

69. Kind, T. et al. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods 10, 755–758 (2013).

70. Smith, C. A. et al. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751 (2005).

71. Florens, L. & Washburn, M. P. Proteomic analysis by multidimensional protein identification technology. Methods Mol. Biol. 328, 159–175 (2006).

72. Xu, T. et al. ProLuCID: an improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J. Proteom. 129, 16–24 (2015).

73. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1, 21–26 (2002).

74. Wen, Z. kite: a software suite for processing and analysis of tandem mass spectrometry data. Zenodo v1.0.0. https://doi.org/10.5281/zenodo.5914885 (2022).

75. Zhang, Y., Wen, Z., Washburn, M. P. & Florens, L. Refinements to label free proteome quantitation: How to deal with peptides shared by multiple proteins. Anal. Chem. 82, 2272–2281 (2010).

76. Choi, H., Fermin, D. & Nesvizhskii, A. I. Significance analysis of spectral count data in label-free shotgun proteomics. Mol. Cell. Proteom. 7, 2373–2385 (2008).

77. Bushnell, B. BBMap short read aligner, and other bioinformatic tools. https://sourceforge.net/projects/bbmap/ (2020).

78. Mather, M. W., Morrisey, J. M. & Vaidya, A. B. Hemozoin-free Plasmodium falciparum mitochondria for physiological and drug susceptibility studies. Mol. Biochem. Parasitol. 174, 150–153 (2010).

79. Hollin, T., Abel, S. & Le Roch, K. G. Genome-wide analysis of RNA–protein interactions in plasmodium falciparum using eCLIP-seq. Methods Mol. Biol. 2369, 139–164 (2021).

Acknowledgements

We thank Tim Wen for his computational support. This work was supported by the National Institutes of Allergy and Infectious Diseases of the National Institutes of Health (grant R01 AI142743 to K.G.L.R.) and the University of California, Riverside (NIFAHatch- 225935 to K.G.L.R.). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929).

Author contributions

T.H. and K.G.L.R. conceived and designed all experiments. A.F., C.F.A.P. and J.C.N. generated the transgenic lines. T.H. validated the transgenic lines and performed microscopy imaging, essentiality experiments, RNA-seq, protein immunoprecipitations, eCLIP-seq, and small RNA-seq. S.A. contributed to the bioinformatics data analyses. J.P. participated in the maintenance of cell cultures and mitochondrial enrichments. A.B., M.H., J.S.K. and A.d.S. performed the metabolomic experiments. A.S. and L.F. performed the mass-spectrometry analysis. C.F.A.P. performed the drug assays. T.H. and K.G.L.R. wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28981-7. **Correspondence** and requests for materials should be addressed to Karine G. Le Roch.

Peer review information Nature Communications thanks Shruthi S Vembar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Material

Supplementary material for chapter 5 is available with the published paper at https://www.nature.com/articles/s41467-022-28981-7.

Conclusion

Mosquitoes-transmitted pathogens represent a significant and potentially increasing threat to humans. This dissertation work has contributed to knowledge in multiple aspects of research into these pathogens, including host-pathogen interaction between viruses and mosquitoes and detailed molecular biology and genetics of the malaria parasites.

The first chapter on mosquitoes and viruses detailed commonly infecting viruses of *Culex* mosquitoes which are under-researched but could impact how well mosquitoes harbor and transmit deadly viruses like West Nile virus, the most significant mosquito-associated pathogen in the United States. Learning more about these co-infecting viruses will hopefully lead to strategies to control the transmissive ability of mosquitoes. This has already been successfully accomplished using the *Wolbachia* bacterium that has been shown to restrict transmission of pathogenic viruses. In addition, detailed information on how *Culex* mosquitoes respond to viruses using small RNA immunity may open new avenues for control of viruses within mosquitoes to prevent transmission.

The remaining four chapters specifically dealt with the details of *Plasmodium* biology with an eye toward discovering important molecular factors that regulate the parasite's gene expression and propel its life cycle progression. These include lncRNAs, which we computationally predicted across the *Plasmodium falciparum* genome and determined which are most likely to be essential to parasite survival. We also did further molecular experiments on a handful of candidate lncRNAs including lncRNA-14, which we showed

to be crucial for sexual differentiation of the parasite. While additional work will be required to reveal the full extent to which lncRNAs are a major force in parasite biology, our work represents a significant advance in the field . Important other parasite factors also include proteins involved in basic cellular processes including those involved in cell division, which are crucial at multiple life cycle stages. We showed that SMC2 and SMC4 bind at the centromeres in *Plasmodium berghei* and are crucial for male gametogenesis in the mosquito stages. These proteins represent potential drug targets against malaria parasites. This is also true for the three RNA-binding proteins characterized in the final two chapters: PF3D7_0823200, which regulates non-coding regions of important transcripts, and PfRAP01 and PfRAP21, which regulate the parasite mitochondrial rRNAs (mitoribosome). These experiments show the effectiveness of the eCLIP-seq technique in determining the RNA binding sites of parasite RNA-binding proteins, and add to the knowledge about post-transcriptional regulation in malaria parasites that is essential in regulating the parasite life cycle progression. The use of eCLIP-seq and other functional genomic approaches we performed for these proteins will continue to elucidate the many mechanisms performing gene regulation at the post-transcriptional level in *Plasmodium*. Finally, the proteome-wide R-DeeP screen that we performed represents a springboard for future research into not only RNA-binding proteins, but all RNA-dependent proteins that do not directly bind RNA but are crucial for RNAs and RNA metabolism.

The work presented here illustrates a step forward toward a fuller understanding of mosquitoes, their interactions with pathogens, and the detailed biology of some of the deadliest pathogens including malaria parasites. With time, our projects together with future work will most likely allow us to curtail the burden of these vector borne diseases on human populations across the globe.