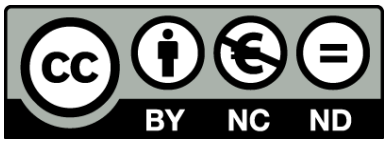
UNIVERSITAT DE BARCELONA

Interplay of genetic, epigenetic and transcription factors in the regulation of transcriptional variation in *Plasmodium falciparum*

Lucas Michel Todó



Interplay of genetic, epigenetic and transcription factors in the regulation of transcriptional variation in *Plasmodium falciparum*

Memòria presentada per Lucas Michel Todó per optar al grau de doctor per la Universitat de Barcelona

Programa de Doctorat en Biomedicina – Bioinformàtica

This thesis was carried out at the Institute of Global Health of Barcelona (ISGlobal), under the supervision of Dr. Alfred Cortés Closas and the tutelage of Dr. Antonio Monleón Getino.

Lucas Michel Todó | Alfred Cortés Closas | Antonio Montleón Getino

UNIVERSITAT DE BARCELONA

ISGlobal Barcelona Institute for Global Health

Per a l'Aba, del teu tesorito

Table of Contents

Abstract

The most severe form of malaria, caused by *Plasmodium falciparum* parasites, still kills over half a million people every year, most of them children under the age of five. Despite huge research efforts, reduction in the global burden of disease has stalled in recent years.

P. falciparum has a very complex life cycle including, among other steps, sexual reproduction in female *Anopheles* mosquitos and an asexual intra-erythoricitic development cycle (IDC) inside the human host, which causes the disease. During the IDC, the parasite needs to continuously adapt to changes in its environment including fluctuations in blood temperature, concentration of nutrients and other metabolites, presence of drugs, and a constant fight against the host's immune system.

In this thesis, we have studied the adaptation mechanisms of *P. falciparum* to this plethora of challenges, with a special focus on clonally variant genes (CVGs). In *P. falciparum*, CVGs are a set of genes, participating in host-parasite interactions, which can be found both in a transcriptionally active state, characterized by euchromatin, or a transcriptionally silenced state, characterized by heterochromatin. The state of CVGs is inherited by the progeny of a parasite, with stochastic switches occurring at a low frequency. Parasites with the most optimal patterns of CVGs expression are continuously selected as the environment changes, leading to adaptation and survival of the infecting population.

In the first paper of this thesis, we have analyzed subcloned parasite populations to characterize, with unprecedented detail, the heterochromatin distribution associated with the active and silenced states of CVGs. This has allowed us to define different kinds of heterochromatin transitions between the active and silenced states of CVGs and has given us new insights on the regulation of *var* genes (one of the main virulence factors for malaria) and into the regulation of sexual conversion, a process crucial for malaria transmission.

Continuing with CVG regulation, in the second paper of the thesis, we have analyzed how patterns of CVG expression are established at the onset of human infections, after passage through

transmission stages. Our results suggest a loss of the epigenetic memory during transmission stages and a reset of the heterochromatin patterns that drive CVG expression. Similar patterns of CVG expression arose in different infected individuals, suggesting that the activation probability of a given CVG is an intrinsic property of the gene.

In the third paper of the thesis, we have further studied the sexual conversion phenomenon. We have generated a conditional over-expression system for *pfap2-g*, the CVG that acts as master regulator of sexual conversion, achieving sexual conversion rates of ~90% after induction. Our results have provided new insights on how heterochromatin at different positions affects expression of pfap2-*g* and have allowed us to characterize the transcriptional profile of the initial stages of sexual commitment with unprecedented sensitivity.

Finally, in the fourth paper of this thesis, we have studied the adaptation of the parasite to heat-shock, which happens in natural infections due to fever episodes. We expected CVGs to participate in this phenomenon, but instead we have identified *pfap2-hs,* a non-clonally variant transcription factor (TF), as the main driver of the heat-shock response in *P. falciparum.* AP2-HS acts as the functional homolog of HSF1 (a TF that drives the heat-shock response from yeast to mammals, but is absent in *P. falciparum*), driving a very tight transcriptional response to heat-shock, characterized by the up-regulation of *hsp70* and *hsp90*. Although the presence of directed responses had previously been demonstrated for other cues, it is the first time that the transcription factor driving such a response is identified in *P. falciparum*.

Taken together, the results of this thesis have broadened our knowledge of the regulation of adaptive mechanisms in *P. falciparum.* Learning about this deadly parasite's defense mechanisms will be instrumental to design better strategies to fight it back in the future.

Introduction

1 Malaria

1.1 A Brief Malaria History: How did we get here?

Malaria has accompanied humanity since its beginnings. Written records referring almost certainly to malaria date back as far as 2700 BC in China, 2000 BC in Mesopotamia, 1570 BC in Egypt or the 6th century BC in India. More reliable ancient sources from early Greeks, from Homer in 850 BC to Hippocrates in 400 BC, among others, already describe the prevalence of fevers, splenomegaly and poor health in people living near marshes and swamps. For over 2500 years malaria was thought to be caused by foul air arising from swamps. In fact, the world malaria is thought to have originated from the Italian expression *"mal'aria"* meaning "bad air". In 1676 Leeuwenhoek discovered microorganisms, and 200 hundred years later, in 1878-79, Pasteur and Koch developed the germ theory of infection, which stated that infectious diseases were caused by microorganisms termed pathogens. The search for the causative agent of many diseases, among them malaria, became a hot topic, and in 1880 Laveran discovered malaria parasites in different stages in blood samples from infected patients [1]. In 1897 MacCullum made the first observation of sexual stages of the parasite (in the closely related organism *Haemoproteus columbae*). In that same year, Ross discovered that malaria parasites (using the avian malaria parasite, *Plasmodium relictum*) were transmitted by the bite of a previously infected mosquito [2], and this was later confirmed in humans by Grassi, Bignami and Bastienelli in 1898 [3]. Pre-erythrocytic liver stages of the parasite were found in 1948 by Shortt and Garnham [4]. In 1976 Trager and Jensen pioneered the techniques to continuously cultivate the malaria parasite *Plasmodium falciparum* intra-erythrocytic stages in vitro [5], a development which hugely facilitated posterior research on the parasite. In 1982 the *Plasmodium vivax* form responsible for the relapses of the disease, the hypnozoite, was discovered by Krotoski [6]. The *var* gene family and its implication in antigenic variation and severity of the disease was first described in 1995 by different teams [7–9] and has remained a hot topic of research until nowadays. The first stable genetic transfection in *P. falciparum* was performed by Crabb and Cowman in 1996 [10], allowing

for the first time to interrogate gene function directly by deletion/addition of genetic material. In 2002 a huge team of scientists determined the whole genome sequence of *Plasmodium falciparum* [11] and in the following years, the whole genomes from many other species were sequenced, enabling malaria research to enter the era of genomics. In recent years, one of the most important discoveries in malaria biology has been the discovery of AP2-G as the master regulator of sexual conversion [12]. AP2-G drives the transition from the asexual replication cycle, responsible for malaria symptoms in humans, to the sexual reproduction cycle, and is thus of central importance for the transmission of the disease.

1.2 Aetiology

There are six different Plasmodium spp. species that can cause malaria in humans: *P. falciparum, P. malariae, P. ovale curtisi, P. ovale wallikeri, P. vivax and P. knowlesi* (a zoonotic parasite that can cause infections in humans). *P. falciparum* is responsible for more than 90% of malaria cases and is responsible for most severe cases and deaths. It is the most abundant form of malaria in sub-Saharan Africa, while *P. vivax* is the most widespread species being the dominant form in South-East Asia and central and South America, accounting for roughly 2% of malaria cases.

1.3 Epidemiology

Malaria has a worldwide distribution, spreading in the tropical and subtropical areas of America, Africa, Asia, and Oceania. Although present all year round in some areas, malaria has a strong seasonal component, with wet seasons carrying a higher burden of disease. According to the WHO reports, there were an estimated 241 million malaria cases worldwide in 2021, with the WHO African Region accounting for 95% of the cases. Nigeria, the Democratic Republic of Congo, Uganda, Mozambique, Angola, and Burkina Faso reported the most cases (accounting for roughly 55% of global cases). The WHO South-East-Asia region accounted for roughly 2% of the cases with

83% of them occurring in India. The rest of the cases where spread across the WHO Eastern Mediterranean, Western Pacific, and the Americas regions [13] (Figure 1).

Roughly following the same distribution, malaria caused approximately 627000 deaths worldwide, of which 77% in children under the under the age of 5. Children and pregnant women continue to be the most vulnerable collectives [13].

Despite huge efforts in malaria treatment, control and elimination, malaria continues to be a devastating disease, especially in low-income settings, highlighting the need for further research that will eventually produce better and cheaper tools against it.

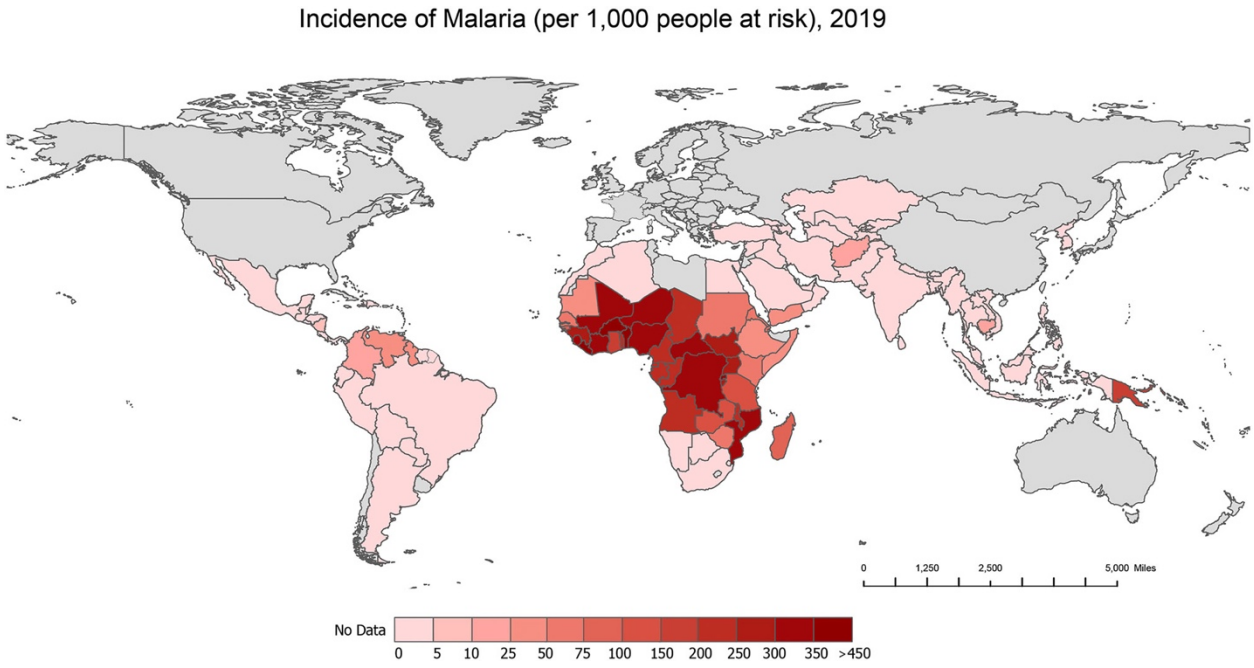


Figure 1. Detected cases of malaria for every 100.000 people at risk per country. Data from WHO 2019. Reproduced from Roberds et. al. [419]

1.4 P. falciparum Life Cycle

Plasmodium falciparum has a very complex life cycle distributed between two hosts: mosquitos of the Anopheles genus, which act as the vector, and humans, which act as the main host. When an infected mosquito bites a human host during a bloodmeal, motile parasite forms residing in the mosquito salivary glands, called sporozoites, are injected into the dermis and penetrate it, reaching the human circulatory system [14,15]. From there, sporozoites migrate into the liver, where they infect hepatocytes [16]. Inside hepatocytes, sporozoites undergo multiple rounds of schizogony (cell division) and convert into multinucleated schizonts which, by budding of parasite-filled vesicles called merosomes, release up to 40.000 merozoites per hepatocyte into the bloodstream [17].

In the bloodstream, merozoites invade red-blood cells (RBCs) and start an asexual intra erythrocytic development cycle (IDC) with 4 differentiated parasite stages: merozoites, rings, trophozoites and schizonts [18]. After invasion, merozoites convert into rings, a relatively low metabolic activity stage in which the parasite cytoplasm adopts a characteristic ring shape (thus the name). Rings develop into trophozoites, a very high cellular activity stage resulting (among other processes) in a significant remodelling of the erythrocyte surface. Finally, the parasites start replicating their genetic material, becoming multinucleated schizonts. Schizonts eventually burst, lysing the RBC and explosively releasing 16-32 new merozoites into the bloodstream. The newly formed merozoites invade other non-infected RBCs and the asexual cycle starts anew. This asexual replicative cycle takes approximately 48h and is responsible for the characteristic periodical fever episodes of malaria as well as most known symptoms of the disease [19,20]

In each round of the IDC, a small proportion of parasites undergo sexual differentiation, exiting the asexual cycle and converting into male and female sexual forms termed gametocytes [21]. Gametocytes sequester in the bone marrow [22] where they undergo a maturation process with 5 distinct morphological stages (I-V) [23]. Stage V gametocytes exit the bone marrow and re-enter the circulation, where they can be ingested by an *Anopheles* mosquito in a subsequent bloodmeal. Inside the mosquito, male and female gametocytes mate resulting in a diploid zygote. The zygote

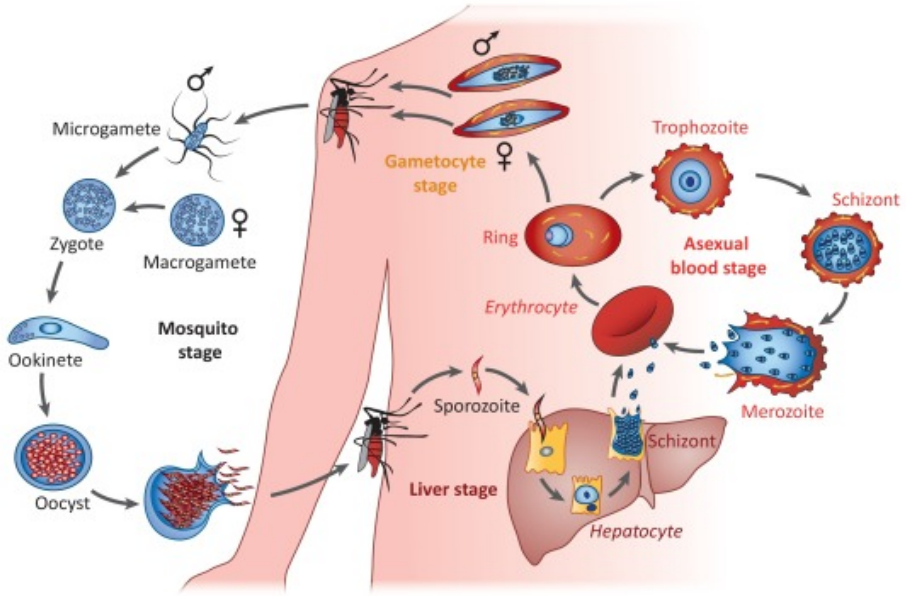


Figure 2. Life cycle of the malaria causing parasite Plasmodium falciparum. Reproduced from Maier et al. 2019. [19]

develops through the ookinete and oocyst stages before replicating asexually again, generating thousands of new sporozoites. Sporozoites migrate to the salivary glands ready to infect a new host, closing the parasite whole life cycle [20] .

1.5 Clinical Manifestations

Malaria curses with a wide variety of largely unspecific symptoms that, in uncomplicated cases, mimic a flu-like syndrome. Symptoms can range from absent or mild to severe, and even result in patient death. All symptoms of the disease are caused by the asexual reproduction of the parasite during the IDC. The severity of an infection may vary according to many factors, including patient age, previous exposure to the parasite, *Plasmodium* species and genotype, host ethnicity, parasite load or previous use of anti-malarial drugs. Malaria is generally further classified into asymptomatic malaria, uncomplicated malaria, and severe malaria.

Asymptomatic malaria is characterized by the absence of symptoms despite the presence of parasites in the blood circulation. It is common in endemic areas with high transmission,

especially in individuals above five years of age. Individuals with repeated exposure to the infection can develop partial immunity and raise tolerance for it, and often do not experience any symptoms. Asymptomatic carriers can be an important reservoir for the disease and contribute significantly to malaria transmission [24].

As we have already mentioned, uncomplicated malaria resembles a typical flu-like syndrome. The most common symptoms are fever, chills, sweats, headache, nauseas, vomits, diarrhoea, abdominal pain, muscular pain, fatigue, and general malaise. The most characteristic trait of symptomatic malaria is periodic fever, which typically occurs every 48h for *P. falciparum,* when schizonts synchronously burst and new merozoites are released into circulation. However, infections are not always synchronous (synchronicity tends to increase as infections persist), and fever episodes can show different time-patterns [25].

Especially if untreated, uncomplicated malaria can sometimes progress into severe malaria, characterized by serious organ failures. Complications of severe malaria are usually related with the pathological binding of iRBCs to endothelial cells and other infected or non-infected RBCs (known as agglutination and rosetting, respectively), which can lead to microvasculature obstruction and inflammation in the affected tissues [26]. Severe malaria can have many different manifestations depending on the affected organs. Symptoms include severe anemia, impaired blood coagulation, hypoglycemia, impairment of consciousness, seizures, coma or other neurologic abnormalities (from cerebral malaria), pulmonary edema, acute respiratory distress syndrome, metabolic acidosis, acute kidney failure, hemoglobinuria, jaundice, hepatic failure and vascular shock caused by cardiovascular collapse. Placental malaria is a type of severe malaria in pregnant women, in which the parasites sequester preferentially in the placenta. It occurs more frequently in primigravid women and can lead to abortions, neonatal death, still births, premature births and low birth weight for the fetus as well as all the symptoms of severe malaria for the mother, leading to a highly increased risk of maternal death.

2 Eukaryotic Gene Expression

2.1 Principles of transcription

To survive and reproduce, all eukaryotic cells rely on the finely regulated expression of their genetic material. The DNA found in the eukaryotic cell nucleus encodes for both protein-coding and non-coding genes which need to be expressed at certain time points and in certain quantities to adequately carry out all cell processes.

Gene expression starts with transcription, which consists in the copying of a DNA sequence into an RNA transcript, which in protein-coding genes is done usually by RNA polymerase II (Pol II). RNA Pol II typically binds to specific regions upstream of genes, termed transcription start sites (TSS). TSSs are themselves located in broader DNA regulatory regions termed promoters. The best studied promoters display a conserved sequence motif known as the TATA box. Promoters act as a binding platform for other proteins necessary for transcription: the TATA-binding proteins (TBPs), the TBP associated factors (TAFs) and the general transcription factors (GTFs), which together with RNA Pol II form a complex known as pre-initiation complex (PIC) [27]. The formation of the PIC at the promoter is sufficient to start transcription, although usually at low basal levels. In most cases, transcription can be enhanced or impaired by more distant regulatory DNA elements, termed enhancers or silencers, respectively, which enable the binding of gene-specific transcription factors (TFs) and cofactors (COFs), although some TF can directly bind to promoters. These factors enable directed responses to specific cues for specific genes [28–30]. There's a third kind of distant regulatory DNA elements known as insulators which prevent genes from being influenced by the transcriptional activity of their neighbouring genes, thus creating "isolated" transcriptional compartments. Insulators can preclude the activity of enhancers or silencers when physically placed between them and the promoter or avoid the expansion of heterochromatin, acting as barrier sequences. The most well know insulator protein in eukaryotes is CTCF, which has been involved both in euchromatin/heterochromatin transitions and in the formation of 3D

chromatin loops [31–33]. Although this seems to be the most general transcription paradigm, recent studies challenge this classical view, and transcription starting in regions outside core promoters (including enhancers) has been demonstrated [34–38].

There are two main types of sequences transcribed into RNA from nuclear DNA: coding and non-coding. Coding sequences will be first processed into messenger RNA (mRNA) and then translated from mRNA into proteins (in a process known as translation), which are the main building blocks of all living cell structures and processes. On the other hand, non-coding sequences (which were for a long time regarded as "genetic garbage") won't be translated into proteins but will generate different kinds on non-coding RNAs (ncRNAs), which have been found to play central roles in gene expression regulation by many different mechanisms (many of which are still incompletely understood) [39].

2.2 Chromatin

In eukaryotes, DNA is wrapped around octameric complexes of proteins known as histones. The unit formed by 147 bp of DNA wrapped around a histone octamer is called a nucleosome [40]. Nucleosomes separated by short sequences of linker DNA form a "string of beads"-like structure which constitutes the first level of chromatin organization (Figure 3). Linker histone H1 organizes chromatin into a 30nm fiber, and supercoiling of this 30nm fiber during mitosis and meiosis produces the metaphase chromosomes, the most condensed form of chromatin [41].

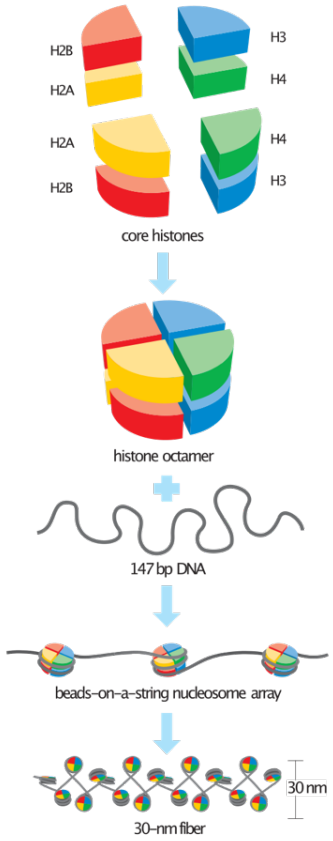


Figure 3. Basic chromatin organization, from histone octamers to string of beads array. (from https://www.wikiwand.com/en/Histone#/Classes_and_variants)

The organization and packaging of chromatin is key in determining the access of the transcriptional machinery to different regions of the genome as well as determining 3D contacts between lineally distant genetic elements, creating partitions in the genome known

as topologically associated domains (TADs) [42–44], and thus represents a major platform for gene expression regulation.

Chromatin can be classified into two broad categories regarding its level of accessibility: a transcription-permissive state termed euchromatin and a more tightly packaged state, refractory to transcription, termed heterochromatin [45,46]. In most eukaryotes, heterochromatin can be further classified into constitutive heterochromatin, which is found in regions of the genome that are always found in a heterochromatic state, such as the peri centromeric regions and the telomeres, and facultative heterochromatin, present in regions of the genome which can be found in both euchromatic and heterochromatic states [47].

2.3 Epigenetic regulation of transcription

Chromatin accessibility and binding capacity with diferent factors are modulated by histone variants, posttranslational modifications (PTMs) of histones, nucleosome remodelling and direct DNA methylation [48]. Some of these modifications carry functional consequences while others are the consequence of previous transcriptional events. In a strict sense, only the modifications with functional consequences which are transmitted to newly formed DNA during cell division can be considered truly epigenetic marks, which carry additional information to the primary DNA sequence. However, all chromatin modifications are usually referred as epigenetic marks in the literature, and we will use this permissive convention in this thesis.

2.3.1 DNA methylation

The fifth carbon of cytosines in DNA molecules can be methylated, originating 5-methylcytosine (5mC), and this modification is widely present in eukaryotic genomes, a phenomenon known as DNA methylation [49]. In eukaryotes, 5mC is found mostly on symmetrical CpG dinucleotides, especially on regions enriched for this dinucleotide known as CpG islands [50,51]. During cell division,

the newly created DNA strain is methylated at the same place where hemi-methylation of CpG in the template strand is found, thus retaining the mark and making it a truly epigenetic heritable mark, which can be transmitted through many generations [52]. DNA methylation has been mostly implicated in transcription repression mechanisms, including silencing of methylated promoters [53–55], silencing of transposable elements [56,57], genetic imprinting [58,59], and the silencing of chromosome X [60,61]. Surprisingly, DNA methylation is also widespread along coding regions of genes, where it positively correlates with transcription [62–64]. Although the mechanism by which DNA methylation enhances transcription in coding sequences in not very clear, it appears to be a highly conserved trait among eukaryotes, highlighting its importance [49,51].

2.3.2 Histone variants

The histone octamer is formed by two copies of each of the core histone proteins H2A, H2B, H3 and H4 [65,66]. This canonical core proteins can be replaced by different histone variants, which can alter the interactions between the DNA and histones and/or interact with different effector proteins [67,68]. Some of the most well-known histone variants are: H3.3 which plays a role in genome integrity during development in mammals [69–72], cenH3 which demarcates centromeres in most eukaryotes [73,74], H2A.X which plays a role in DNA damage repair, chromatin remodeling and X-chromosome inactivation in somatic cells [75] and H2A.Z which can regulate transcription, DNA repair, suppression of antisense RNA and RNA Pol II recruitment [75–85].

2.3.3 Histone Post Translational Modifications

Histone proteins have unstructured amino terminal ends which protrude from the nucleosome, called histone tails. Both histone tails and their globular domains can be subject to numerous covalent post translational modifications (PTMs) such as methylation, acetylation, phosphorylation, ubiquitylation, SUMOylation, ADP-ribosylation and deamination, among others (Figure 4). These modifications can modulate chromatin compaction level and interact with effector proteins, hugely impacting the accessibility of transcriptional machinery to the

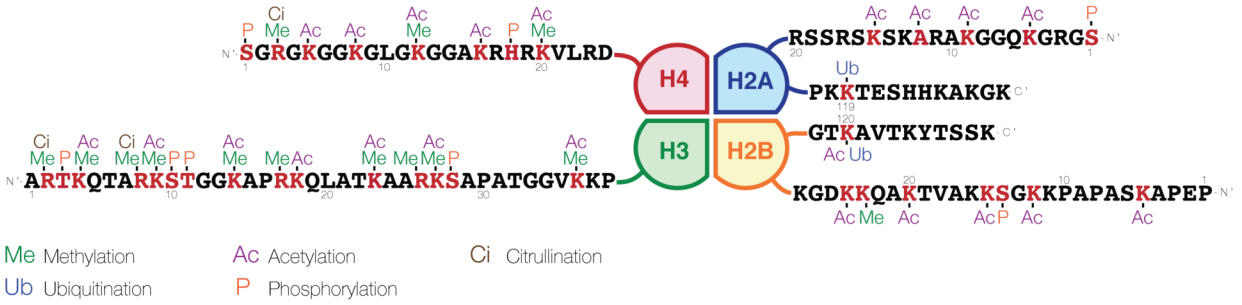


Figure 4. Histone PTMs. The four different core histone proteins (H4, H3, H2A and H2B) and their most common modifications. (from https://www.wikiwand.com/en/Histone#/Classes_and_variants)

underlying DNA sequences and adding a layer of information to the genetic code, which has been hypothesized (together with histone variants) to serve as a true epigenetic code known as the histone code [86,87].

The histone PTMs landscape is shaped by a set of highly specialized proteins, highly interconnected with the rest of the cell metabolism. There are three main kinds of proteins involved in histone PTMs: enzymes which catalyze the formation of specific PTMs, known as writers, proteins which recognize and interact with specific PTMs, known as readers, and enzymes which remove specific PTMs, known as erasers. Almost all specific PTMs have been found to have at least one writer and one eraser, highlighting their dynamic nature, and leading to the hypothesis that the equilibrium between these two ultimately determines histone PTMs spatial and temporal distribution, enabling for a fine regulation. Writers, readers, and erasers can interact with a plethora of cofactors which link them to the different signalling cues they need to respond to [88,89].

2.3.3.1 Histone methylation

Some of the most well studied histone PTMs are lysine acetylation and methylation. Lysines can be mono-, di- or tri-methylated, probably in a sequential manner [90–92]. Methylation is carried out by the writer enzymes histone lysine methyltransferases (HKMTs), which transfer methyl groups

from S-adenosylmethionine (SAM) to the lysine amino group and usually contain the conserved SET domain. The implication of SAM in the methylation process establishes a link between histone PTMs and the general cell metabolism. Demethylation, on the other hand, is carried out by two families of eraser histone demethylases (HDMs): LSD1 and jumonji-C domain containing (JMJD) [93–96]. Methylation marks can be associated with both transcription enhancing and transcription repression, depending on the methylated residue and the context. These marks can be recognized by different readers (containing recognition domains like Chromo, Tudor, PHD, PWWP and MBT) which in turn can recruit other effector proteins to enhance their regulatory function [97–103].

Histone H3 lysine 9 tri-methylation (H3K9me3), one of the best studied histone PTMs, is the hallmark marker for constitutive heterochromatin in eukaryotes. It is enriched in non-transcribed parts of the genome, comprising peri-centromeric regions, telomeric regions, other genomic repetitive sequences, and some silenced genes [104]. H3K9me3 also plays a central in the silencing of transposable elements and it may have evolved precisely for this function [105].

H3K9me3 is deposited by proteins of the SUV39 subfamily (which differ in name in different organisms but are widely conserved), which contain both a chromo reader domain and a SET writer domain. Once deposited, H3K9me3 is bound by heterochromatin protein 1 (HP1) [106,107] through its chromo reader domain. In addition to its reader domain, HP1 also contains a more C-terminal cromoshadow domain (CSD), which dimerizes, forming a binding platform for other proteins [108,109]. SUV39 proteins are one of such proteins, they can bind to HP1, and this coupling generates a positive feedback loop underpinning heterochromatin formation and spreading [110–112]. Other effectors can as well bind HP1, and these complex interactions altogether lead to heterochromatin establishment and allow for the recruitment of writers for other downstream histone methylation marks such as H3K56me3, H3K64me3 and H4K20me3 as well as other effector proteins, which further contribute to transcriptional repression [113–116].

H3K27me3 and monoubiquitylation of histone 2A are produced by the Polycomb repressive complexes PRC2 and PCR1, respectively, and are the hallmarks of facultative heterochromatin.

Genomic regions demarcated by these marks are transcriptionally silenced and most of them are crucial for cell fate determination and development [117–119].

H3K4me3 is enriched in the promoters of most active genes and contributes to the assembly of the transcriptional machinery, although it is not indispensable for transcription [120–123]. The fact that H3K4me3 and other histone PTMs are found in embryonic chromatin before zygote genome activation suggests that they might play an important role in the establishment of active transcriptional states [124–127]. H3K4me3 has also been related with low levels of DNA methylation in CpG islands irrespective of transcription [128].

H3K36me3 is deposited by the SET2 methyltransferase, which is recruited by elongating (active) RNA Pol II, and is thus correlated with active transcription [129,130]. The same mark has also been related with DNA methylation and splicing regulation, although it's role in different organisms is still under investigation [131–133].

2.3.3.2 Histone acetylations

Histone lysines can only be mono acetylated, which neutralizes their positive charge weakening the interaction with the negatively charged DNA strand and fostering a more open chromatin conformation, which facilitates the recruitment of the transcriptional machinery [134]. Additionally, lysine acetylation marks can be recognized by the bromodomain contained in many transcriptional coactivators, enhancing transcription [135]. For these reasons, lysine acetylation is broadly considered to be an activation mark [136–138]. The writers for histone acetylation are histone acetyltransferases (HATs), which use acetyl-CoA to transfer an acetyl group to the amino group of lysines, linking acetylation to general metabolism and availability of nutrients [139]. The erasers of this mark are histone deacetylases (HDACs) which, by restoring the positive charge of lysines, usually act as transcriptional repressors [140].

The most well-known lysine acetylation marks are H3K9ac and H3K27ac. H3K9ac, together with H3K14a, has been found in active promoters, bivalent promoters and active enhancers, and H3K27ac has been found to associate with active promoters (and differentiate them from poised ones). Both H3K9ac and H3K27ac exert their function not only by their positive effects on transcription but also by impeding the methylation of the same residues, since both H3K9me3 and H3K27me3 lead to heterochromatin formation [141–144].

2.3.4 Noncoding RNAs

Some types of housekeeping ncRNAs, essential for translation, have been known since as early as the 60s. Two of the most important ones are transfer RNAs (tRNAs) and ribosomal RNAs (rRNA). rRNAs are the main component of ribosomes, enzymatically active macromolecules that bind a template mRNA and translate it into a protein by "reading" the mRNA and synthetising the corresponding chain of amino acids (forming a peptide). tRNAs are the molecules responsible for the actual translation part, by specifically recognizing triplets of nucleotides (known as codons) and recruiting the corresponding amino acid to the ribosome through the action of aminoacyl-tRNA ligases [145–147].

In the 90s, a new kind of ncRNAs where discovered: small RNA molecules that mediated post-transcriptional silencing of complementary mRNAs [148,149]. Since that seminal discovery, many classes of regulatory ncRNAs have been identified which participate in gene expression regulation by diverse mechanisms (usually involving base complementarity and/or chromatin remodelling). Regulatory ncRNAs have been divided by length into two broad categories: small ncRNAs (<200 bp) and long non-coding RNAs (>200 bp). Noncoding RNAs account for the vast majority of the transcriptional output of most eukaryotes (around 98% in humans), and their amount and variety in a genome correlate better with organism complexity than genome size, hinting their importance in regulating cell optimal transcription [150].

The best studied classes of small ncRNAs are micro RNAs (miRNAs), PIWI associated RNAs (piRNAs) and small interfering RNAs (siRNAs) (Figure 5). miRNAs and siRNAs are processed by the DICER enzyme and they both mediate mRNA targeting and decay by base complementarity. miRNAs exert their function forming a complex containing an ARGONAUTE family protein, known as the RNA interference complex (RISC), where they act as the guide for the complex, targeting specific mRNAs [151]. siRNAs work similarly and they can originate from endogenous RNA or exogenous RNA. Endogenous siRNA have been found to mediate the silencing of transposable elements in the female germ line in mice [152,153], while in plants and some invertebrates, exogenous siRNAs play an important role in anti-viral defense [154,155]. As we have seen, miRNAs and siRNAs have many similarities. Their main differences are their origin, most siRNAs are exogenous while all miRNAs are endogenous, their structure, miRNAs are single-stranded and siRNAs are double-stranded and how they exert their function, miRNAs typically can bind-to and silence many targets while siRNAs are highly specific for a single target. On the other hand, piRNAs are not processed by DICER, they are incorporated into the PIWI proteins, and they mediate the silencing of transposable and repetitive elements in the germ line [156,157].

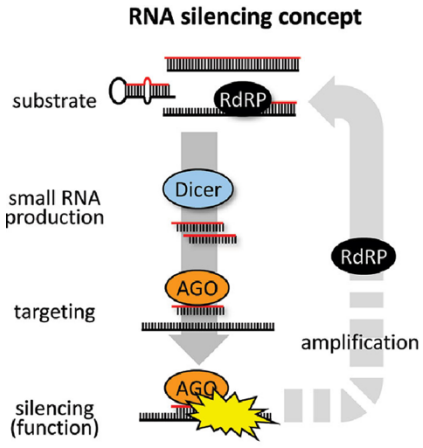


Figure 5. RNA interference scheme. DICER enzyme processes small ncRNAs generated by RNA-dependent RNA-polymerases (RdRP), which are bound by the ARGONAUTE (AGO) complex and guide it to its targets for degradation in a sequence dependent manner. Adapted from Svoboda et al. [420].

Long non-coding RNAs (lncRNAs) are a highly diverse group of RNAs which can originate from sense or antisense transcripts, overlapping genes, intergenic regions or even enhancer and promoter regions, and that constitute the majority of ncRNAs expressed in cells [158]. lncRNAs often contain a poly adenylated (polyA) tail and can be post-transcriptionally spliced, in a similar fashion to mRNAs. They are estimated to be highly abundant in most eukaryotic cells, and although the functionality of many of them is still in debate, they have been implicated in many biological processes such as post-transcriptional control of gene expression [159], regulation of promoter activity [160–162], inactivation of chromosome X [163–165], genetic imprinting [166,167] and nuclear architecture [168–170], among other processes.

3 Transcriptional Regulation in *P. falciparum*

3.1 General aspects of *P. falciparum* transcription

The *P. falciparum* genome contains around 5300 genes, spreading 23.3 million base pairs (bp) organized in 14 chromosomes, with an added 6kb of mitochondrial DNA and 34kb of apicoplast DNA (a plant origin organelle) [171]. It is the most AT-rich of known eukaryote genomes, with a global AT-content of ~80%, rising to ~95% in non-coding regions [172]. Unlike most eukaryotes, most the genome in Plasmodium parasites is found in an euchromatic state at any given time, with almost two thirds of its genes expressed in every life-cycle stage. Intergenic regions are relatively short, making it a very "compact" genome [173]. Most of the main components of the PIC have been identified in *P. falciparum,* including PfTBP, PfTFIIE and RNA pol II, although some subunits of the complex are lacking [174–176]. The typical bipartite structure of cis regulatory enhancer elements and basal promoters, which recruit Pol II to the TSS, exists in *P. falciparum* [177], and the presence of bidirectional promoters, multiple TSS per gene, and sense and antisense ncRNAs has also been documented [178]. Gene-specific TFs, one of the key transcription regulators in most eukaryotes, are remarkably scarce in *P. falciparum,* which has one of the highest ratios of number of proteins to predicted specific transcription factors (a ratio of ~800 in contrast to the ~29 of *S. cerevisiae*) [179] . In addition to that, many of the most common eukaryotic TF families have not been found in *P. falciparum* [175,177]. The main family of TFs in the parasite is the plant derived Apicomplexa specific family ApiAP2. There are 27 members of the ApiAP2 family, most of them displaying sequence-specific binding, and they are thought to be the main regulators of transcription along the parasite's life cycle [180,181]. How such a small number of TFs can drive such a complex life cycle is still a matter of debate, but a combinatorial regulation model has been proposed as the most plausible explanation [182–184]. Interaction with other post transcriptional regulators likely constitutes an added layer of complexity [185].

3.2 Transcription along the life cycle: the ApiAP2 TFs

Progression along the life cycle is, by far, the major source of transcriptional variation in *P. falciparum*. During the parasite's life cycle, each gene is expressed periodically and in a just-in-time manner and transcriptomics analysis have shown a cascade of transcription along the cycle [186–188] . Recent single cell studies have corroborated this view, although transcription flow might be a bit more stepwise than was previously thought [189,190].

As stated previously, the ApiAP2 family is thought to be the main responsible for the regulation of the progression along the life cycle. Systematic screens in *P. berghei* [182] *and P. yoelii* [191] , and more recently in *P. falciparum* [192], have shown that many of its members are essential at particular life cycle stages [177]. Specific binding motifs for most of the members have also been predicted, and some have been found to recognize secondary motifs, favouring the hypothesis of a complex and interconnected regulatory network [193].

Specific functions have been elucidated for many ApiAP2 members. PfAP2-G is the master regulator of sexual conversion, its expression is sufficient for triggering exit from the IDC and progression to gametocytogenesis [12]. Acting downstream of PfAP2-G, PfAP2-G2, PfAP2-G4, and PfAP2-G5 are all implicated in gametocyte maturation [194,195], and AP2-FG(G3) participates in gametocyte maturation only in female gametocytes in *P. berghei* [196]. PfAP2-I is needed for the expression of invasion genes in merozoites [185], but it also targets many other genes related with nucleosome and chromatin remodeling (including seven other ApiAP2 genes), cell-cycle related genes and vesicle transport and host membrane remodeling genes [185,197]. PfAP2-EXP has been shown to be essential during the IDC, and it has been involved in the regulation of multigene families important for host-parasite interactions [198]. PfSIP2 specifically binds to heterochromatic loci in the subtelomeres and telomeric repeats, colocalizing with heterochromatin protein 1 (PfHP1) at chromosome ends (but not in internal heterochromatic islands). Its disruption had no global transcriptional effect suggesting a role in heterochromatin formation and/or genome integrity as opposed to transcriptional control [199]. PfAP2-HC has been found to specifically bind

to heterochromatin in a PfHP1 dependant manner; it does not regulate gene expression or heterochromatin formation, but it appears to be an integral part of heterochromatin in *P. falciparum* [200]. PfAP2-Tel targets telomeric repeats and probably participates in telomere maintenance [201]. PfAP2-LT forms a complex with the histone acetyl transferase (HAT) PfGCN5 and participates in transcriptional activation of late IDC genes [202]. AP2-SP, -SP2 and -SP3 are all required for the formation of mature sporozoites in the rodent malaria parasites *P. berghei and P. yoelii* [182,191,203]. AP2-SP might have a broader regulatory role since it also participates in gene silencing during the IDC in *P. berghei* [182] and its *P. falciparum* homolog is PfAP2-EXP (implicated in expression of multigene families). AP2-L is essential for the development of mature liver stages in *P. berghei* [204]. Finally, AP2-O to AP2-O5 are essential in different parts of the ookinete and oocyst maturation process in rodent malaria parasites [182,191,205].

3.3 Clonally Variant Genes, general aspects

In *P. falciparum*, clonally variant genes are a set of genes, generally involved in host-parasite interactions, which can be found in both an active and a silenced state in genetically identical parasites at the same stage of life cycle progression. The transcriptional state of CVGs is determined at the epigenetic level. Active CVGs are found as euchromatin, while silenced CVGs are found as heterochromatin, and thus transcriptionally repressed. Given its epigenetic nature, the state of CVGs is inherited by the progeny of a cell, with stochastic euchromatin/heterochromatin switches occurring at low frequency. Most CVG gene families have multiple members, which can be switched on and off, generating a huge diversity of possible CVG expression profiles. These profiles confer different functional properties to individual parasites, with new patterns continuously arising from the spontaneous switches in individual CVG states. At the population level, this heritable variability sets the grounds for natural selection of the parasites best suited to the present environment. This process enables adaptation to different challenges, such as the immune system pressure, fluctuation in the availability of nutrients, or the presence of toxic compounds (antimalarials for instance). The dynamic nature of chromatin states, which can be influenced by a plethora of factors, and the reversibility of its modifications

confer this adaptative mechanism different characteristics from natural selection of genetic variants, which occurs at a longer timescale, randomly and in an irreversible manner. This kind of adaptative strategy, in which the adaptations are present in the population prior to any selective pressure, is known as bet hedging [206] and is a prevalent strategy in nature.

There are many families of CVGs: *stevor, rifin, pfmc-2tm, hyp, clag…* But the best studied family is the *var* gene family. The *var* genes encode for ~60 distinct variants of erythrocyte membrane protein 1 (PfEMP1), an erythrocyte surface antigen with strong pathological implications [207]. Of the ~60 members of the family, only one member is expressed in each individual parasite, in a mutually exclusive manner [208,209]. This ensures a limited exposure of the antigen variants to the immune system and allows the parasites to express a different *var* gene (at the population level) when antibodies against one isoform are generated, constituting a major mechanism for immune evasion [7,210] (Figure 6).

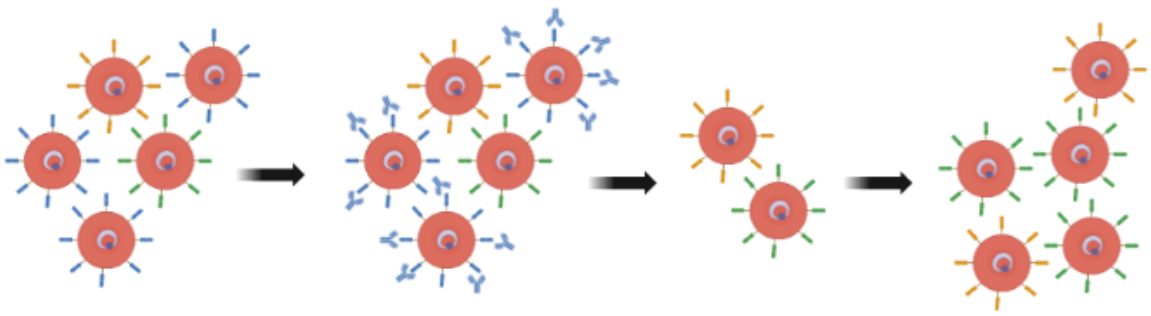


Figure 6. Schematic representation of antigenic variation (linked to the var gene family, among others). Blue, yellow, and green membrane proteins represent different variants of a surface antigen encoded by a CVG (like PfEMP1). When antibodies against one specific variant (blue) are raised, only parasites expressing other variants (green and yellow) will survive. Since CVG expression patterns are inherited by the progeny of a parasite, after some rounds of replication, most of the population will express the surviving parasite's antigen variant (green and yellow).

3.4 Non-coding RNAs in *P. falciparum*

Surprisingly, no RNA interference pathway has been found in *P. falciparum*. The parasite lacks ARGONAUTE and DICER homologs, among other important proteins that from the core of the RNA interference system in other eukaryotes [211–213].

On the other hand, lncRNAs have been found to be abundant and ubiquitously expressed during the whole life cycle of the parasite [214–219]. Most lncRNAs are natural antisense tanscripts (NAT) and they are thought to be important for the regulation of parasite gene expression [218], although their mechanism of action remains unclear, with some studies reporting positive correlation between sense-antisense pairs [215] and others an inverse relation [214].

A few lncRNAs have known or hypothesized functions which are crucial for specific process in parasite biology. The most studied case is the *var* gene family regulation. All *var* genes possess a bidirectional promoter in their intron, from which two lncRNAs are generated, one antisense lncRNA complementary to the first exon, and another sense lncRNA that extends into the second exon. Although their mechanism of action is still not clear, both lncRNAs have been implicated in *var* gene silencing as well as *var* gene activation [220–222]. In another layer of regulation, 136 bp ncRNAs (technically not lncRNA), originated from GC-rich elements, have also been implicated in the regulation of this gene family [223–225].

Another family of lncRNAs, associated with telomere-associated repetitive elements (TAREs), which are coordinately expressed after DNA replication, are thought to play an important role in parasite telomere biology [226].

Finally, lncRNAs have also been implicated in the regulation of sexual conversion. Gametocyte development protein 1 (PfGDV1), an upstream positive regulator of PfAP2-G, the master regulator of sexual conversion, has an antisense lncRNA that negatively regulates its expression thus preventing sexual conversion in the majority of parasites [227].

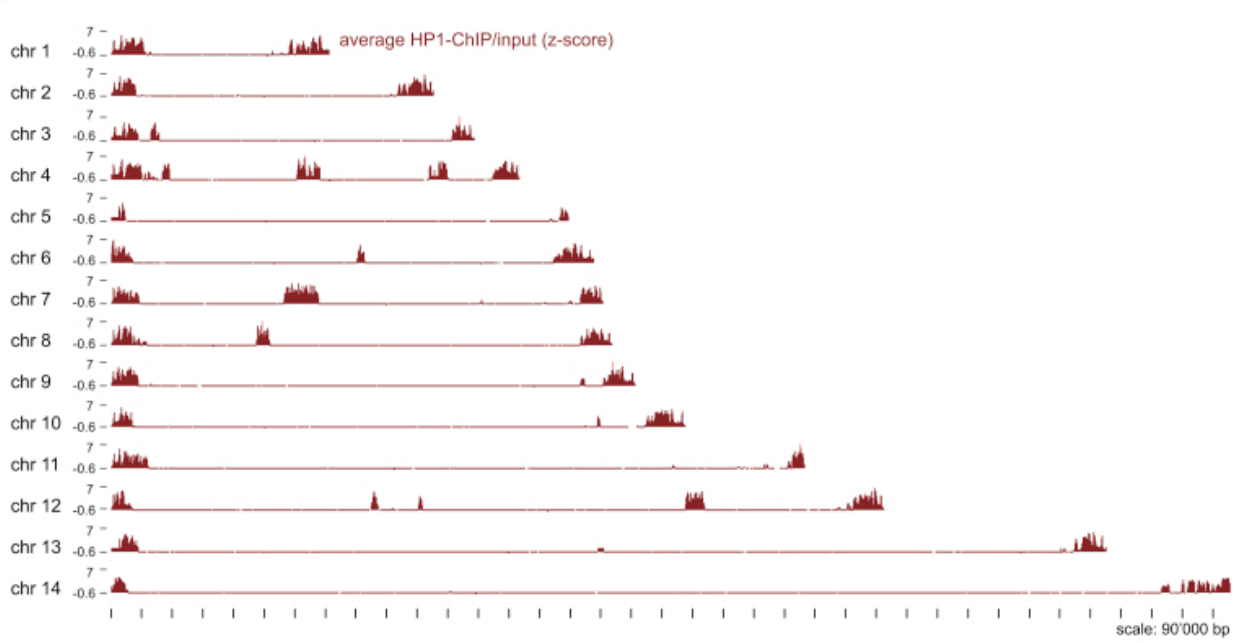


Figure 7. Distribution of HP1 along the P. falciparum genome. Data represent z-score normalized PfHP1-ChIP/input signal. Adapted from Fraschka et al. [273]

3.5 Chromatin in *P. falciparum*

Chromatin architecture in *P. falciparum* has some particularities compared to most other eukaryotes. Most of its genome is found in a euchromatic state [228] and typical chromosome territories in the nucleus are not as well defined as in other eukaryotes [229]. There are some regions in the genome, however, that are found in a heterochromatic state. These regions are limited to the TARES, the subtelomeres and some internal regions of chromosomes [230,231] (Figure 7). The subtelomeric and internal chromosomic regions occupied by heterochromatin comprise the CVGs. Heterochromatin in these regions is not continuous, individual genes can be found in an euchromatic state, suggesting that genes are individually regulated. These clusters of CVGs and their euchromatin-heterochromatin transitions will be the main focus of this thesis.

3.5.1 Histones and Histone Variants

Despite the biased base composition of the *P. falciparum* genome (which a priori could hinder the ability of histones to bind DNA and form nucleosomes), no special histone adaptation to this sequence composition has been observed, and as in most eukaryotes, canonical core histone proteins are conserved in *P. falciparum* [232,233]. Histone variants PfH2A.Z (but not H2A.X), PfH2B.Z (which is apicomplexan specific), PfCenH3 and PfH3.3 have been identified, and seem to play specific roles in gene regulation [234–236].

Histones PfH2A.Z and PfH2B.Z have similar distribution profiles and they are both enriched in intergenic regions of euchromatin, marked by H3K9ac and H3K4me3. Both are also enriched in the promoter of the active *var* gene [234,237]. Coimmunoprecipitation experiments have proven that at least some nucleosomes can display both variants in the same octamer [237,238]. They both remain stable across the IDC and their abundance does not temporally correlate with gene expression, but promoters enriched with these variants display the highest levels of expression along the IDC [234,238]. The *var* genes represent an important exception to this mechanism; their promoters are only enriched in PfH2A.Z during active transcription in ring stage parasites [239].
PfCenH3 (together with PfH2AZ) is found in centromeric DNA, where it demarcates the centromere, that is devoid of heterochromatin in *P. falciparum* [236]. PfH3.3 has not been thoroughly investigated, but it has been found to harbor typical activating PTMs, which may indicate a similar role as in other organisms, where it is enriched in actively transcribed genes. *P. falciparum* (like all apicomplexans) lacks the H1 linker histone, which likely contributes to a looser general 3D genome organization and a lack of typical chromosome territories [207,240].

3.5.2 Nucleosome Landscape

Nucleosome occupancy resembles that of other eukaryotes, with a nucleosome depleted region (NDR) directly upstream the TSS, but unlike model organisms in which strongly positioned nucleosomes (-1, +1) flank the NDR, the -1 nucleosome is lacking, and only a strongly positioned

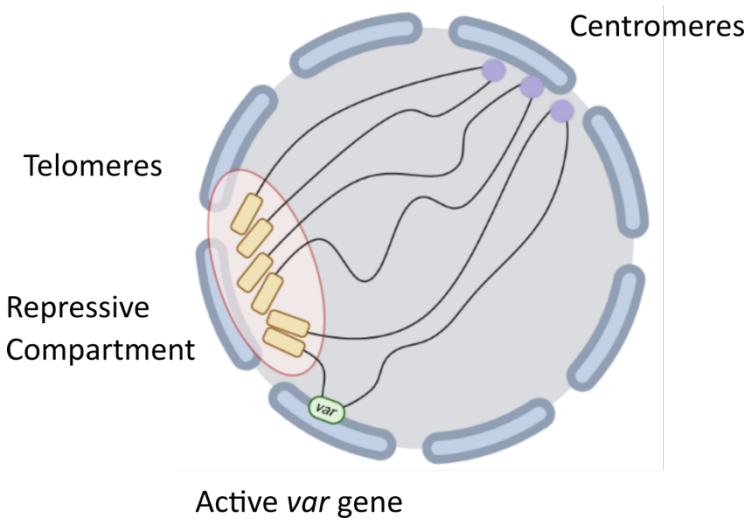


Figure 8. Nuclear organization in P. falciparum. The active member of the var family localizes in a different compartment than the silenced members.

nucleosome downstream of the TSS has been found [241,242]. The lack of nucleosomes at the TSS is more pronounced in highly expressed genes, and dynamic nucleosome depletion correlates with transcription during the IDC [241]. Some studies have suggested that there is a relative depletion of nucleosomes in intergenic regions [243–245], but more recent studies suggest that this could be explained by technical reasons related to the higher AT content in non-coding DNA [241]. Similarly, it had been proposed that the number of nucleosomes was higher at the ring stage, lowered in trophozoites and peaked again during schizogony [246–248], but more recent studies failed to reproduce this result [241] .

3.5.3 3D architecture

As mentioned before, the general 3D structure of the *P. falciparum* nucleus is quite loose, with a lack of typical chromosome compartments [229], but this by no means implies that there is no spatial organization. Chromosomes are folded, with the centromere anchored in one region of the nucleus and telomeres in the opposite side [247]. Heterochromatic foci cluster together in the periphery of the nucleus [229,231,247], while the active *var* gene displays a differential positioning in an actively transcribed compartment (Figure 8). The fact that linearly distant heterochromatic

gene clusters are found together in the nuclear periphery firmly suggests the presence of wide chromatin loops in the parasite chromosomes.

3.5.4 DNA methylation

DNA methylation has been identified in *P. falciparum,* albeit at a very low frequency. Two different cytosine modifications have been found: 5mC and 5hmC-like [249,250]. 5mC was surprisingly found to be present only in the template DNA strand from genes [249]. Also regarding 5mC, core promoters were found to be hypomethylated while transcript levels correlated with methylation at exonic regions of genes [249]. 5hmC is an intermediate of the demethylation from 5mC to C described in other eukaryotes [251,252]. A very similar cytosine modification, termed 5hmC-like, was found to be more predominant in the *P. falciparum* genome than canonical 5mC (5mC 0.01-0.05%, 5hmC-like 0.2-0.4%) [250]. Similarly to 5mC, 5hmC-like was found to be enriched in gene bodies, to be stable across the IDC and its presence positively correlated with transcript levels [250]. Only one single functional DNA methyltransferase, termed PfDNMT, was found in the *P. falciparum* genome, which could mediate the deposition of these marks [249]. Taken together, and contrary to most eukaryotes, DNA methylation does not seem to play a central role in the regulation of transcription in *P. falciparum*.

3.5.5 Histone PTMs in P. falciparum

The combination of mass-spectrometry (MS) and ChIP-Seq-based studies, at different parasite life-cycle stages, is allowing us to unravel the spatial and temporal distribution of histone PTMs and is starting to shed light on the roles different histone PTMs play in transcriptional regulation in *P.falciparum*.

A recent study has identified over 230 histone PTMs in *P. falciparum*, 88 of which had never been identified in other organisms [228,232,253], and 46 of which have also been quantified [254]. Acetylation, methylation, phosphorylation, ubiquitination and SUMOylation have all been found in *P.*

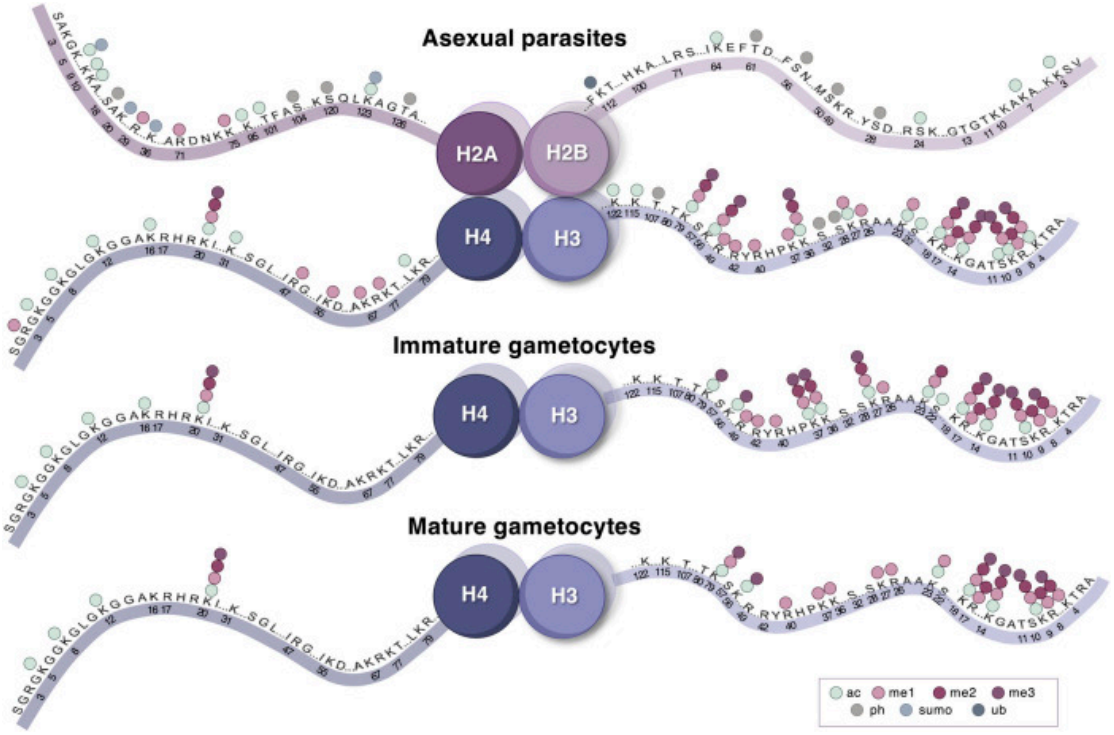


Figure 9. Histone PTMs identified in P. falciparum *by quantitative methods. Reproduced from Connacher et al. 2022* [421].

falciparum histones [228,232,233,253–262], but only acetylation and methylation, the most abundant marks, have been identified along multiple life cycle stages and thoroughly characterized. It has become increasingly clear that not only individual PTMs but also combinations of these PTMs carry specific functional consequences [263], adding a layer of complexity to the histone PTMs landscape. A recent study also demonstrated that histone PTMs presence changes dynamically with life-cycle progression, and that there is a marked difference in the palette of histone PTMs used during asexual proliferation and sexual differentiation [254].

Although some histone PTMs might just be a mere consequence of previous transcriptional events, the ubiquity of histone PTMs, their combinatorial and dynamic nature and the fact different life cycle stages have unique and specific histone PTM signatures [253,254], seem to be clear indicators of the pivotal role of histone PTMs in the regulation of gene expression in *P. falciparum* (Figure 9).

3.5.5.1 Histone Acetylation

Acetylation is the most abundant histone PTM during the IDC and is mainly associated with a transcription-permissive state [253,254]. Acetylation of histones H3K9, H3K14, H3K56 and H4K5, H4K8, H4K12, H4K16 are widespread across the euchromatic regions of the genome [254,255,263–266] and their disruption by HAT/HDAC inhibitors resulted in global transcriptional disruption [267,268], highlighting their implication in active transcription.

Although all these marks are found in euchromatin, only H3K9ac and H3K14ac were found to be enriched in highly transcribed genes when paired with RNA-Seq data [266]. Both marks are more abundant at the 5' end of genes, near the TSS, but they spread along the whole coding region [266]. H3K9ac colocalizes with H3K4me3 and histone variant H2A.Z and its enrichment levels not only correlate with highly expressed genes but also with temporal patterns of gene expression [234].

Some of these marks display specific distributions that vary along the parasite's life cycle. During the IDC, H3K18ac and H3K27ac, together with the histone variant PfH2A.Z, dynamically mark the TSS of active genes and colocalize with PfAP2-I and bromodomain protein 1 (PfBDP1) [265]. H4K8ac is higher in schizont stages, while H3K9ac and H3K14ac are higher in trophozoites [254,255]. H3K4ac is abundant in active promoter regions, and its abundance increases greatly during the schizont stage [254] (a very transcriptionally active stage [187]). H4K16ac has a very similar distribution profile to H3K4ac and correlated with active transcription in a mass spectrometry -based study, but its role is less clear [254].

The fact that histone acetylation is so widely distributed, marking not only actively transcribed genes but also silent genes found in euchromatin, suggests that many of these marks somehow contribute to the maintenance of most of the genome in a poised state, ready to be expressed in a timely manner, when necessary, instead of being actual marks of active transcription.

3.5.5.2 Histone Methylation

Methylation is also widely present along the genome, and unlike acetylation, it might be associated with both transcriptionally active and repressed chromatin.

H3K4me2-3 is a widespread active mark with increased abundance in schizonts. H3K4 methylation also correlated with levels of transcription when paired with RNA-Seq data, although not in a temporal manner [266]. H3K4me3 can recruit the transcriptional activator PfGCN5/ADA2 and its associated reader complexes containing PfBDP1 and PfBDP2 [269,270]. The PfGCN5/ADA2 complex acetylates H3K9 generating H3K9ac, which is a clear example of histone PTMs crosstalk and positive feedback. H3K4me2 has also been implicated in *var* gene regulation, where it decorates the active member of the family, found in a poised state between cell divisions, thus contributing to the active state memory [271,272].

Moving onto repressive marks, H3K9me3, the landmark of constitutive heterochromatin in most eukaryotes, is found in both constitutive and facultative heterochromatin in *P. falciparum.* It spreads from TAREs into the subtelomeric genes, and it is also found in some internal regions in a few chromosomes [231,266,273]. Given its importance as the main marker of heterochromatin in *P. falciparum,* H3K9me3 will be further discussed on a latter chapter.

Other repressive methylation marks have been found in *P. falciparum*, although unlike H3K9me3 they are not restricted to heterochromatin, nor to silenced genes [266]. H3K36me3, which is associated with transcription elongation in other eukaryotes, has a broad distribution along the genome on both active and silent genes in *P. falciparum*, but it's ratio to activating marks has been found to correlate negatively with transcription [266] . In *var* genes, it has been found to be enriched in the TSS of all silent *var* genes as well as the coding sequence of both the silenced and the active members of the family but absent from the TSS of the active *var* gene. Its disruption led to the expression of virtually all members of the family [274] .

The highly conserved heterochromatin mark H3K27me3, which for a long time was thought to be absent in *P. falciparum*, has been found to be present only during gametocytogenesis [254]. In fact, repressive histone PTMs seem to increase considerably during the initial steps of gametocyte maturation [254], when many genes related to the IDC need to be silenced. Also during gametocytogenesis, both H3K27me3 and H3K36me3 were found to have a very high pairwise correlation with the occupancy of the repressive TF PfAP2-G2 [195], suggesting an interesting functional link in gene silencing between these marks and the TF.

H4K20me3, which is associated with heterochromatin in most eukaryotes, has a very unspecific and widespread distribution in *P. falciparum,* without a clear relation with transcriptional status [231,275].

3.5.6 Histone Modifying enzymes in P. falciparum

3.5.6.1 Methyltransferases and Demethylases

There are 10 known histone methyl transferases (HMTs) in *P. falciparum*, all of them from the SET domain containing family [276]. Preferences for certain methylation substrates have been derived in vitro for all of them but two (SET5 and SET9), trough different experiments [276–279]. PfSET1, PfSET4, PfSET6, PfSET7 and PfSET10 all can methylate H3K4 in vitro and PfSET3 and PfSET7 can methylate H3K9 in vitro. Of them, only PfSET3, PfSET6 and PfSET9 (which methylates H3K20) seem to be essential during the IDC [280,281]. PfSET2 mediates methylation at H3K36, and its knock out resulted in the expression of virtually all *var* genes in the parasite [274,282]. PfSET10 has been proposed to mark the active *var* gene with H3K4me2-3 and contribute to the heritability of its state [271], although a latter study did not support these conclusions [283]. PfSET7 has been localized outside the parasite nucleus during the IDC, liver stages and sporozoites indicating a possible role in methylating newly forming histones or non-histone substrates [284].

Regarding demethylases, only 3 Jumonji-C domain-containing histone demethylases (HDMs), with two of them (JMJ1 and JMJ3) preferentially targeting H3K36 in vitro, and the lysine-specific

demethylase LSD1 have been identified [276,285]. None of them seem to be essential during the IDC [280,281], but their inhibition resulted in severe transcriptional dysregulation [285].

3.5.6.2 Acetyltransferases and deacetylases

There are four known histone acetyl transferases (HATs) from the GNAT and MYST families, with specificities characterized for PfGCN5 (H3K9 and H3K14) and PfMYST (H4K5, H4K8, H4K12, H4K16). Disrupting the normal function of these enzymes results in severe transcriptional dysregulation, underlining their importance for optimal gene expression [202,267,286–288]. PfGCN5 has been found to form a complex (highly divergent from the conserved eukaryotic SAGA complexes [289], which play a crucial role as transcriptional cofactors [290,291]) that includes two PHP domain proteins (PfPHP1 and PfPHD2) and an ApiAP2 TF (PfAP2-LT) [202]. The disruption of said complex by domain deletion in either PfGCN5 or PfPHD1 led to a decrease in H3K9ac and H3K4me3, and a phenotype characterized by reduced merozoite invasion and elevated sexual conversion [202].
Regarding histone deacetylases (HDACs), one of class I and two of class II, all of them essential during the IDC [280,281] , and two sirtuins, dispensable for the IDC, have been found. PfHDAC2 disruption led to upregulation of most CVGs including PfAP2-G, which in turn led to higher sexual conversion rates [268]. PfSIR2A and PfSIR2B have been found to play role in silencing CVGs in the subtelomeres [239,264,292–294], although their effect seems to be strain specific [175,295], as well as negatively regulating the expression of ribosomal RNA genes [264].

3.5.7 Readers of histone PTMs

Initial bioinformatic sequence analysis showed that the *P. falciparum* genome encodes for orthologues of histone readers with conserved histone binding domains such as chromo, tudor, PHD or WD40 (which bind to methylated lysines) and bromodomains (which bind to acetylated residues) [175]. To date, six bromodomain, four chromodomain, ten PHD and approximately 90 WD40 motif containing proteins have been identified, although many of them remain to be characterized [296].

Some of the known histone PTM writers possess such domains: PfGCN5 contains a bromodomain, PfSET1 contains a bromodomain and four PHD domains and PfMYST contains a chromodomain [276]. The fact that these histone writers also have reader domains suggests the existence of feedback loops in the establishment of heterochromatin and euchromatin marks, as well as the existence of crosstalk between different PTMs.

PfHP1 contains both a chromo and a chromo shadow domain, which allow it to bind H3K9me3 and exert its function as an integral part of heterochromatin [230,297]. As in most eukaryotes, the cromoshadow domain of PfHP1 has dimerizing capacity, highlighting its potential for nucleosome agregation [297]. PfHP1 is restricted to heterochromatin, and it is not found in centromeres, unlike in other organisms. Functional studies have proven that PfHP1 is essential for the parasite, that genes enriched with PfHP1 are generally in a silenced transcriptional state and that its disruption results in an upregulation of a subset of CVGs, with only minor effects on global transcription [298].

PfBDP1 is a bromodomain containing protein that binds to acetylated H3 residues. It was found to be essential during the IDC and exert its function by binding at the TSS of invasion related genes (and to another bromodomain containing protein, termed PfBDP2), where it is necessary for their expression [299]. Another component of the PfBDP1/PfBDP2 complex was recently characterized, named PfBDP7, which interacts with PfBDP1 at heterochromatic foci and participates in CVG regulation [300].

In other eukaryotes, histone readers usually function as big protein complexes [301]. A recent pull-down experiment in *P. falciparum* identified a total of five different major reader complexes, interacting with different combinations of histone PTMs [269]. These complexes contained known and novel readers as well as TFs, known histone PTM writers and other unknown proteins. As demonstrated for other organisms, many of the complexes showed different "flavours" with different compositions recognizing different histone PTMs or sets of them. For instance, the PfGCN5/ADA2 module could bind through PfPHD1 or PfPHD2, changing its binding substrates. A

high degree of interconnectivity was found between these reader complexes, highlighting a complex regulatory landscape of histone PTMs in *P. falciparum* [269].

3.5.8 Heterochromatin in *P. falciparum*

As previously stated, heterochromatin in *P. falciparum* is characterized by the presence of H3K9me3, and it has some differential characteristics with respect to most eukaryotes. In *P. falciparum*, H3K9me3 is found in both constitutive heterochromatin and facultative heterochromatin. It is worth noting that H3K27me3, the typical mark of facultative heterochromatin in other eukaryotes, has not been detected in *P. falciparum* at IDC stages (it was just recently found to be present at gametocyte stages [254]), and its associated silencing complex, the Polycomb complex, has no known orthologues in *P. falciparum* [175].

Unlike in most eukaryotes, heterochromatin is very restricted in *P. falciparum*. H3K9me3 is only found in telomere-associated repetitive elements (TAREs), demarcating constitutive heterochromatin, and in the subtelomeres and a few internal clusters in specific chromosomes [231,266,273], demarcating facultative heterochromatin, but it is absent from pericentromeric regions. All these H3K9me3-enriched heterochromatic loci are specifically bound by PfHP1, which, as we have seen, constitutes an integral part of heterochromatin and can act as a binding platform for other effector proteins [230].

Virtually all genes marked by H3K9me3 and PfHP1 are CVGs that participate in host-parasite interactions or sexual differentiation, which are rendered transcriptionally silent by these marks [228,230–232,247,302]. H3K9me3 and PfHP1 distribution display relatively sharp edges [230,266,273], and individual genes inside broadly heterochromatic regions can be found in an euchromatic state (the active members of each CVG family), characterized by the absence of H3K9me3 and PfHP1 and the presence of active acetylation marks such as H3K9ac and H3K4me3. How particular genes undergo these heterochromatin-euchromatin transitions is still poorly understood but constitutes the base for the variegated expression of CVGs. The distribution of H3K9me3 and

PfHP1 remains stable along the IDC [303] but an epigenetic reset of CVG expression patterns seems to take place when transitions into transmission stages occur, such as conversion into sexual stages or transitions into mosquito and liver stages, although the exact stage at which this reset occurs remains to be established [304,305].

As we have seen, the regulation of these transitions involves a plethora of histone readers, writers, and erasers which in turn can interact with regulatory ncRNAs, and the presence of other histone PTMs. A recent study, which analysed the binding distribution of 16 PfApiAp2s by ChIP-Seq, identified 8 of these transcription factors, namely PfAP2-HC, PfAP2-G5, PfAP2-O5, PfAP2-12, PfAP2-G2, PfAP2-11A, PfAP2-O2, and PfAP2-EXP, which were preferentially enriched in heterochromatic regions. All these ApiAP2s interacted with heterochromatin in a sequence independent manner, and with a high level of co-occupancy between them [192]. A subset of these factors also showed preferential enrichment in the TAREs, co-localizing with PfAp2-Tel and PfSIP2[192,199,201]. Taken together, these results highlight a complex and multi-layered regulatory network of factors involved in heterochromatin formation in *P. falciparum*, which we are just starting to disentangle.

3.6 Clonally variant genes functions

Around 10% of all *P. falciparum* genes display clonally variant expression and they have been implicated in virtually all host-parasite interaction processes such as antigenic variation, cytoadherence, invasion ligand switching or nutrient acquisition as well as in sexual commitment, highlighting their central role as virulence factors (Figure 10).

As introduced before, the best known and studied family of CVGs is the *var* family, which comprises ~60 *var* genes encoding for the PfEMP1 protein, of which only one member is expressed at a time per parasite, in a mutually exclusive manner. PfEMP1 is exported to the surface of the infected RBCs (iRBCs), where it binds through a transmembrane domain, leaving a hypervariable portion exposed, that mediates cytoadhesion to specific endothelial cell receptors. Cytoadhesion to the endothelium removes iRBCs from the circulation, thus avoiding circulation through the spleen and clearing [208]. Additionally, cythoadhesion to specific tissues such as the placenta or the brain by specific PfEMP1 variants has been linked to the corresponding complications of pregnancy-associated malaria or cerebral malaria, respectively [306,307]. PfEMP1

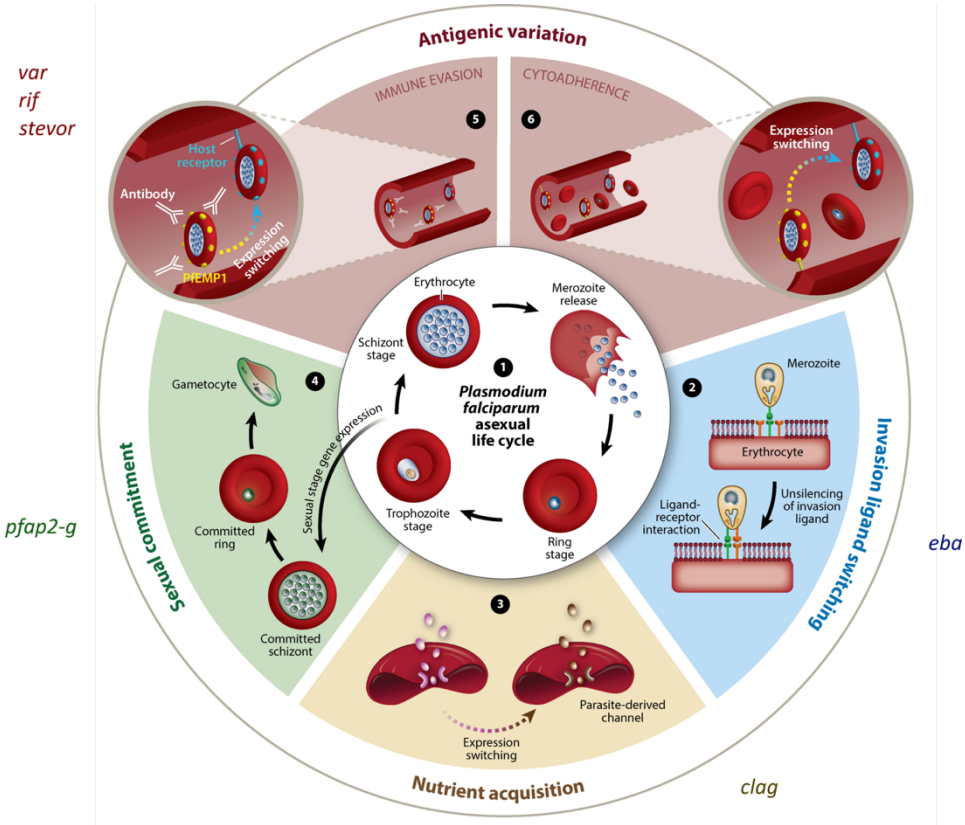


Figure 10. Example of clonally variant gene (CVG) functions in Plasmodium falciparum. CVGs participate in virtually al host-parasite interactions: immune evasion, by switching the antigens presented at the surface of the iRBC, cytoadherence, by expressing different membrane proteins with specific affinity for endothelial cell receptors, invasion ligand switching, by expressing different membrane proteins capable of recognizing different RBC membrane receptors, nutrient acquisition by changing the properties of a parasite derived solute channel and sexual commitment, since the master regulator of sexual conversion is silenced in most parasites in a clonally variant manner. Addapted from Duraisingh et al. 2018 .[400]

can also mediate cythoadherence of iRBCs to non-infected RBCs, a phenomenon known as rosetting, which can cause circulatory problems in the microvasculature [308]. For all the aforementioned reasons, PfEMP1 remains the major malaria virulence factor identified to date.

Like all CVGs, *var* expression is regulated at an epigenetic level. Silenced *var* genes are found in a heterochromatic state, and they localize together in one or a few heterochromatic clusters in the nuclear periphery [309]. Given that there are several *var* gene clusters in very distant chromosome regions, the formation of this repressive compartment probably requires the formation of large chromatin loops. On the other hand, the active member of the family is found in a euchromatic state, localizing in a separate nuclear periphery region than the silenced members [310,311]. As stated before, H3K36me3 also seems to play a role in *var* gene regulation, it is present in the coding regions of all *var* genes, but absent from the TSS of the active one, and its disruption (by PfSET2 inhibition) led to the loss of mutually exclusive expression [274,282].

Histone variants have also been implicated in *var* gene regulation, both H2A.Z and H2B.Z have been found to be enriched in the active *var* promoter, with the presence of H2A.Z temporally limited to the stage of the IDC at which the *var* gene is expressed [237,239].

In another layer of regulation, ncRNAs have also been found to play a role in *var* expression. *Var* genes possess an intron from which an antisense lncRNA is transcribed for the active member of the family, and sense lncRNAs are produced for all members [222,312]. The disruption of the antisense lncRNA led to vast changes in *var* expression [312], reinforcing its implication in regulation. As stated before, 136 bp long ncRNAs originated from GC-rich elements found in the internal chromosomal *var* gene clusters, have also been implicated in the regulation of this gene family, since their overexpression led to the activation of specific subset of *var* genes in individual cells

Family	Num. CVG members	Function	Selected Publications
rifin	184	Exported to iRBC membrane. Some members have been linked to rosetting.	313–316
var	105	Exported to iRBC membrane. Mediate cytoadhesion to different receptors.	208,272,278,317
stevor	42	Exported to the iRBC membrane. Conferes special mechanical properties to the iRBC.	318–320
phist	40	Exported to the iRBC cytoplasm. Participate in iRBC remodelling and protein export.	321–323
hyp	16	Exported proteins with unknown function.	321
pfmc-2tm	13	Exported to the iRBC membrane.	320,324
acs	8	Acetyl-CoA-synthetase. Lypid metabolism.	325
fikk	7	Kinases exported to the iRBC cytoplasm. They participate in iRBC remodelling.	321,326,327
surfin	7	Many localitzations reported including iRBC surface. Unknown function.	328,329
dnaj	6	Chaperons involved in iRBC remodelling	321
clag	4	Formation of the PSAC, transport of solutes through the iRBC membrane.	330–332
gbp	3	Glycophorin binding proteins. Unknown function.	321
6-cys	2	Unknown function, but important for transmission stages.	333
acbp	2	Acetyl-CoA-binding proteins. Lypid methabolism.	325
Other CVGs	130	Other important CVGs are *pfap2-g*, *mspdbl2* or *eba140* (which binds to glycophorin C and participates in RBC invasion). Many others are yet to characterize.	12,206,273,334

Table 1. Adapted from Cortés et al. 2017 [335]

[223]. In addition to this specific mechanisms, *var* genes are regulated by the H3K9me3/ac equilibrium, like the rest of CVGs [272].

Much less is known about the rest of CVGs families. The *rif* (~200 members), *stevor* (~40 members), *pfmc-2tm* (~10 members) and *surfin* (~10 members) families are predicted to be expressed in the surface of iRBCs and participate in antigenic variation [316,324,328]. *Rif, stevor* and *surfin* are expressed also in merozoites, but their role at this stage remains to be elucidated [336]. The *eba* and *Pfrh* families (~5 members) participate in the invasion of RBCs by merozoites. All except *Pfrh5* are non-essential, and they encode for functionally redundant adhesins that mediate merozoite reorientation and apical attachment to RBC through different invasion pathways [337–341]. MSPDBL2 is located at the merozoite surface and appears to be silenced in the vast majority (~99%) of parasites in a population [342]. Its levels correlate with many gametocyte-related genes, and it has been found to be a target of gametocyte development protein 1 (PfGDV1) [227], but it does not correlate with PfAP2-G levels, which has led to the hypothesis that it demarcates a specific subset of parasites that continues asexual growth in gametocyte inducing conditions [334]. Genes *clag3.1*, *clag3.2* and clag2 are members of the wider *clag* family (5 members) that display clonally variant expression. The *clag* family genes participate in the formation of the plasmodial surface anion channel (PSAC), a broad selectivity channel induced by the parasite, through which solutes necessary for normal growth are transported across the iRBC membrane [343,344]. Members *clag3.1* and *clag3.2* display mutually exclusive expression [330], although it is not strict, since parasites expressing both promoters could be found under strong selective pressure in transgenic lines [345]. It has also been observed that under selection with basticidin S, a toxic compound that is transported through the PSAC, parasites with both genes in a silenced state can be selected [346,347]. These findings led to the hypothesis that different *clag3* expression profiles can lead to phenotypical variability in the solute transport characteristics of the PSAC, thus allowing adaptation to variability in the concentration of nutrients, toxic compounds and other solutes in the host blood. Finally, PfAP2G, the master regulator of sexual conversion, is also under epigenetic control and displays clonally variant expression, but it will be addressed in detail in the next section.

3.6.1 Regulation of Sexual Conversion

In each round of the IDC, a few parasites commit to sexual conversion and develop as gametocytes. Since gametocytes are the only parasite forms able to reinvade mosquitos, this fact has profound implications for malaria transmission and has been the object of thorough investigation.

A few years ago, the PfAP2G transcription factor was identified as the master regulator of sexual conversion [348,349]. The *pfap2-g* gene is a CVG found in a heterochromatic state (and thus silent) in the vast majority of asexually reproducing parasites, but in a small proportion of parasites *pfap2-g* is derepressed and transcribed, and sexual conversion is triggered [12]. The basal proportion of parasites that undergo this transition is different among isogenic parasite subclones, and this proportion is maintained stable through multiple generations [12], suggesting some epigenetic heritable mechanism underlying the probability of activation of *pfap2-g*. Once *pfap2-g* is activated, it has a first peak of expression followed by a rapid drop and a new wave of expression, owing to a positive feedback loop, and its activation is sufficient to trigger sexual differentiation [189,350]. If the activation and stabilization of *ap2-g* occurs during the early ring stage, the parasite will directly differentiate into a gametocyte in a process known as same cycle conversion (SCC). If the activation event takes place later, the parasite will undergo another round of asexual

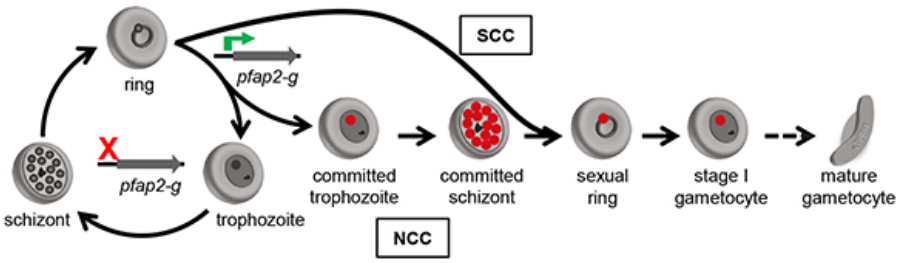


Figure 11. SCC and NCC pathways. Once pfap2-g is activated, parasites can undergo or not an additional round of replication before converting into gametocytes. Adapted from Bancells et al.2019 [351] .

replication as sexually committed forms, going through the committed trophozoite, committed schizont and sexual ring stages, after which all sexual rings (all the progeny of a committed schizont) will differentiate into gametocytes, a process known as next cycle conversion (NCC) [351] (Figure 11).

PfAP2-G targets have been characterized both at the transcriptional level and using ChIP-Seq techniques and include many genes expressed in sexually commited schizonts/rings but also genes expressed during stage I of gametocytogenesis [189,350,352]. Interestingly, PfAP2-G also targets many invasion-related genes in committed schizonts, and it shares many targets with PfAP2-I, suggesting a possible cooperation of these transcription factors in committed schizonts [352].

The mechanisms of *pfap2-g* activation have started to be elucidated recently. PfGDV1 has been found to colocalize with heterochromatin along the genome and its expression led to the removal of PfHP1 from the *ap2-g* and *mspdbl-2* loci specifically, which are then de-repressed and activated [227,353]. How PfGDV1 is able to specifically target these loci for heterochromatin dismantling, despite its genome wide heterochromatin binding, remains to be elucidated [227]. The expression of PfGDV1 itself is regulated by an antisense multi-exon as-lncRNA, that originates downstream of the gene and overlaps its ATG start codon. Disruption of this as-lncRNA resulted in overexpression of PfGDV1 and a marked increase in sexcual conversion, thus suggesting a negative regulation of the sense-antisense pair [227].

In a single-cell RNA-Seq study, the *pfap2-g* neighbor ApiAP2 transcription factor PF3D7_1222400 and the helicases ISWI and SNF2L were upregulated concomitantly with PfAP2-G. Given the role of helicases on chromatin remodelling, this raises the possibility of their implication in making the *pfap2-g* locus accessible to the transcriptional machinery during *pfap2-g* activation [189].

As introduced before, sexual commitment occurs at low frequency at basal level, but it can be increased (a process usually referred as sexual induction) upon certain environmental factors, such as lysophosphatidylcholine restriction [354], high parasitaemia [355] or some antimalarials [356].

Independently of the inducing factor, cultures with higher levels of sexual conversion show higher levels of PfAP2-G [268,349,354,356], but how these factors are sensed by the parasite and the link between them and PfGDV1, PfAP2-G, or any unknown upstream factor, is still an ongoing matter of research.

4 Bioinformatics

Bioinformatics refers to the use of computers and specialized software for the study of biological systems. Bioinformatics can therefore be seen as an instrument in the biologist ever growing set of tools to study living organisms, rather than as an objective by itself. Computers may aid scientists in many ways, but they are particularly suited for intensive calculations and the manipulation of large datasets, which was what they were invented for in the first place. As the amount of biological data available for scientists has increased over the years, so has the need for computational tools able to cope with it, by storing, manipulating, and analysing it in efficient ways. In recent years, the drop in price, and subsequent widespread adoption of next-generation sequencing (NGS) techniques, has led to an exponential growth of the amount of data generated, even by small research groups, and bioinformatics has become an essential part of modern biology.

4.1 A brief history of bioinformatics

The beginnings, Protein Sequence Analysis

Despite its current stellar take-off, bioinformatics did not start with the advent of NGS. In fact, the story of bioinformatics can be traced almost as back as the beginnings of modern computers.
The first applications of bioinformatics were developed in the field of protein sequence analysis. In the late fifties, Margaret Dayhoff and Robert S. Ledley developed a computer software called COMPROTEIN that was able determine a protein primary structure from Edman peptide sequencing data [357] (a very tedious and cumbersome task at the time). COMPROTEIN became the first software specially developed to solve a biological problem, it was the first example of what we would call nowadays a *de novo* sequence assembler. The software was coded in FORTRAN for the IBM 7090 machine (Figure 12) (which weighted almost a ton and costed almost three million dollars at the time) and was entirely written in punch cards. Later, in 1965, Dayhoff

(together with Eck) would also publish the first biological sequence database, called *Atlas of Protein Sequence and Structure* [358], which contained 65 different proteins and their sequences.

Aligning protein sequences became an important and unsolved problem. Sequence similarity was found to be an indicator of phylogenetic proximity and could shed new light into the evolution of proteins, but it was impossible to do it by hand when sequences had differing lengths or where too dissimilar. The first dynamic programming algorithm designed to solve this problem was published by Needleman and Wunsch in 1970 [359], but the first practical multiple sequence alignment (MSA) method, based on Needleman-Wunsch algorithm, wasn't published until 1987, by Da-Fei Feng and Russell F. Doolitle [360]. Only a year later in 1988, the popular MSA software CLUSTAL was developed as a simplification of Feng and Doolitel's algorithm and it is still maintained and widely used today [361].

In 1978 Dayhoff, Schwartz and Orcutt made another major breakthrough in the history of bioinformatics by creating the first probabilistic model for amino acid substitutions. It was based on an experimentally validated set of point accepted mutations (PAM) between closely related

Figure 12. IBM 7090 machine. Computers didn't have a screen at the time, inputs had to be written in punch cards and outputs were printed on paper. From Wikipedia commons.

proteins, and it resulted in the creation of the first PAM substitution matrix, containing the probabilities of each pairwise substitution. PAM matrices and more modern versions of them, based on similar concepts, but making use of the greater amount of sequencing data available nowadays, continue to be used today in MSA problems.

The first genomic sequences

During this same period, many crucial discoveries led to the establishment of DNA as the carrier of heritable information. By 1968, the codon code that translated DNA nucleotide chains into amino acid chains was already deciphered [362], but methods for reading DNA sequences were still lagging. In 1977, Sanger chain termination sequencing [363] (which is still used today) was developed, but the low amounts of DNA present in cells still posed a problem to obtain sequences for study. The next big step in making DNA sequences available was the polymerase chain reaction (PCR), invented by Kjell Kleppe in 1971 [364], and greatly optimized by Kary Mullis in 1987, who developed its current version. With Sanger sequencing and PCR the analysis of genetic sequences, in addition to peptide ones, finally took off. The first software package dedicated to the analysis of Sanger sequencing data was developed by Roger Staden in 1979 (Staden package) [365], and it is still maintained today. In 1984, the University of Wisconsin Genetics Computer Group published the first software suite for the analysis of sequencing data, the GCG, which contained 33 command-line tools to manipulate DNA, RNA and protein sequences.

Free software and Public Databases

In 1985, Richard Stallman published the GNU (GNU is not Unix) manifesto and started the GNU project to create a free to use, edit and distribute Unix-based operating system. The GNU project later evolved in the Free Software Foundation, which proposed that users should be free to run, copy, distribute, study, change and improve the software (https://www.gnu.org/philosophy/free-sw.en.html). This idea has been at the core of many bioinformatics initiatives, such as the European Molecular Biology Open Software Suite (EMBOSS) which started in 1996 as a free

alternative to GCG [366], and open-source and free software has since then been a staple in the bioinformatics field. In the same period, and in a similar spirit, the major public genomic databases (European Molecular Biology Laboratory (EMBL), GenBank and DNA Data Bank of Japan (DDBJ)) united, to standardize data formatting, define minimal information for stored sequences and facilitate data access to the scientific community. This join effort continues today under the International Nucleotide Sequence Database Collaboration (https://www.insdc.org/).

New computers and new scripting languages

During these same years, throughout the 80s, the development of better microprocessors led to new desktop computers, with better hardware capabilities and which run on Unix-like systems. Many new scripting languages emerged during this period, which simplified the coding process by making use of a more natural language-oriented syntax.

One of the first such languages was C, which was built with a focus on speed and code efficiency, which is still widely used for intensive calculations.

Perl was created in 1987 by Larry Wall as an addition to the GNU operating system, with a special focus on the manipulation of text data. This made it ideal for bioinformatics software because biological sequence data can be faithfully represented as text data. The first bioinformatics software written in Perl (GeneQuiz [367]) dates from 1994, and quickly many more followed. Perl became undoubtedly the main scripting language of bioinformatics, which was further favored by the release of BioPerl [368] in 2002, a suite of Perl modules that facilitates common but nontrivial tasks in bioinformatics. Despite its many advantages, Perl's syntax flexibility made reading and maintaining code fairly difficult, which has led to a progressive decline of its use in recent years. Python, another scripting language, with a strong focus on code readability and ease of maintenance (sacrificing some flexibility), was created by Guido van Rossum in 1989 and has slowly but steadily gained support among bioinformaticians (as well as many other programmers). The accessibility of Python syntax makes it an easy scripting language for

```cpp
# include <iostream>
using namespace std;

int main (int argc, in *argv[])
{
    count << "Hello World" << endl;
    return 0;
}
```

```python
print('Hello World')
```

Figure 13. Print statement in C++ (left) and Python (right). C++ is a reimplementation of C, still widely used today, with a focus on speed and efficiency over code readability. The development of new, more user-friendly, scripting languages greatly facilitated the involvement in bioinformatics of people coming from different fields (rather than from computer science only). From https://www.3d-gene.com/en/about/chip/chi_003.html.

beginners, while retaining all the capabilities of any fully grown programing language (Figure 13). In recent years, Python has definitely overthrown Perl as the most used scripting language among the bioinformatics community, and many specialized modules for the analysis of biological data are now publicly available, including a Biopython suite [369].

Another programing language worth mentioning for bioinformatics is R. R was created by statisticians Ross Ihaka and Robert Gentleman in 1991, specifically for statistical analysis of data and graphics generation. Just like Perl and Python, R was released as open-source software (under the GNU license) which has allowed many users to create their own modules and make them available for others, hugely enhancing its ease of use and utility for specific tasks. Despite its more specialized aim, or maybe thanks to it, R has also become a major player in the bioinformatics field, and many biological data analyses and experimental data reports are run on it.

The Genome and the Internet

During the 90s, two main events had a huge impact on bioinformatics: The Human Genome Project (1990-2003) and the appearance of the World Wide Web. Several pioneering sequence assembler software had to be developed (mainly in Perl) to cope with whole-genome scale data that was generated for the first time for the Human Genome Project, fostering a great

advancement in the field. On the other hand, it is impossible to overestimate the impact of the Web in every aspect of human life. The capability of instant communication and sharing of information across the globe has been a huge accelerator for many scientific disciplines. For bioinformatics, it allowed the sharing and rapid dissemination of software, as well as making biological data-bases accessible from virtually anywhere in the world.

The EMBL Nucleotide Sequence Data Library (that includes SWISS-PROT and REBASE, among others) was on the Web since 1993, the NCBI (the United States National Center for Biotechnology Information) launched its website in 1994, which included the tool BLAST (the number one sequence aligner still used today), PubMed started in 1997 and many other important databases, many of which still in use, date from this period.

The Internet also allowed for some bioinformatic tools to be made available online (just like BLAST), usually with user-friendly graphic interfaces. Until then, the use of bioinformatic tools required the knowledge of Unix-like operating systems, terminal command lines and usually some programming language skills, making them restricted to specialized bioinformaticians. Although this is still true for many relatively complex analyses, the appearance of online graphic interfaces allowed at least some common bioinformatic tasks to be performed by general biologists, greatly facilitating day-to-day research.

Structural Bioinformatics

The use of computers to predict the 3D structure of biomolecules, mainly proteins, has been an attractive matter of research since it became a possibility in the 90s [370]. However, the computational power needed for the calculations involved in modeling particle interactions has hindered the predictions, and only proteins with closely related experimentally derived structures could be faithfully modelled. Recently, the great increase in processing power of modern computers, together with advances in machine learning, have led to promising advances in the

field, and new tools like AlphaFold [371] are generating accurate predictions for thousands of proteins.

Next-Generation Sequencing and Big Data

The arrival of second-generation sequencing and massive parallel sequencing technologies has led to a massive drop in the cost of sequencing, and to an exponential increase in the amount of sequencing data generated by research groups around the world. Specialized databases with huge physical infrastructures have had to be deployed in order to cope with this amount of data, such as the Sequence Read Archive (SRA) [372] and the European Nucleotide Archive (ENA) [373]. In the same line, the Genomic Standards Consortium was created in 2005 to standardize data-formats and the information that any record should contain [374]. But not only genomic data has increased in the last decades, other types of quantitative data on different biological molecules (collectively known as omics data from the suffix -omics applied to proteomics, metabolomics, transcriptomics, etc.) have seen a significant increase as well, leading to an unprecedented richness on biological information. Nowadays, even many small research groups generate huge amounts of such data, which makes it impossible to analyse it without the use of specialized software. For this reason, bioinformatics has become an integral part of the research routine, and it has jumped from being used only in very specific/narrow areas of biology to be widespread among the study of life sciences.

This richness of data, paired with modern computation capabilities, has also enabled the use of machine learning techniques, network-based algorithms, and other learning techniques (which may require thousands or even millions of observations and intense calculations to perform well) to be applied to biological questions. This has led to very successful applications in a myriad of challenging areas such as the already mentioned protein structure prediction, the association of complex genomic traits to health disorders, phylogenetic tree reconstruction, the classification of tumours into clinically relevant categories, reconstruction of gene regulatory networks and many others in the present and yet to come.

Although, until recently, most omics datasets have been typically analysed in isolation, there is a growing tendency in taking a more unifying approach into biology. Integrative analysis trying to bring together data from different categories (different omics, phenotypic data, clinical data…) are becoming more and more common. Even a whole new field, known as systems biology, has emerged, which seeks to analyse biological systems as a whole, considering the complex interconnections between different regulatory layers present in them, and producing new biological insights from this holistic view. It goes without saying, that trying to model even the simplest biological system in its entirety requires immense calculation and storage power and modern computers and bioinformatics are playing and will continue to play a crucial role in these efforts.

4.2 Bioinformatic analyses of transcriptomic and epigenomic data

As we have seen, bioinformatics has become a staple in the analysis of all kinds of omics data. In this section, we will refer discuss in more detail the analysis of the type of data generated in this thesis: transcriptional data from gene expression microarrays and H3K9me3/H3K9ac enrichment profiles from ChIP-Seq experiments. We will briefly describe the reasoning behind the approaches commonly used to analyse these types of data and some of the difficulties that arise from *P. falciparum* peculiarities.

4.2.1 Gene Expression Microarrays

Gene expression microarrays are a matrix of small DNA molecules (known as probes) attached to a solid surface. These DNA probes hybridise to known target cDNA (or cRNA) molecules produced from cell transcripts, by sequence specificity. When assessing transcription with microarrays, RNA is extracted from the sample of interest and used as a template for generating cRNA or cDNA copies. These cDNA/RNA molecules are then labelled using a fluorophore or chemiluminescent molecule and laid over the microarray surface. cDNA/RNA molecules then bind to the microarray probes by sequence specificity, and posterior washes ensure non-specific binding is avoided. The readout of the fluorescent dye allows the simultaneous quantification of every DNA target (usually a whole transcriptome) for which a probe is present in the microarray (Figure 14).

Microarrays can be of two main kinds: 1-chanel microarrays and 2-channel microarrays. In 1-channel microarrays, a single dye is used, and they contain many short probes for each gene. In 2-channel microarrays, two dyes are used, meaning that two samples can be hybridized to the same microarray (usually a reference pool and a sample of interest), and they tyìcally contain lesser but longer probes per gene.

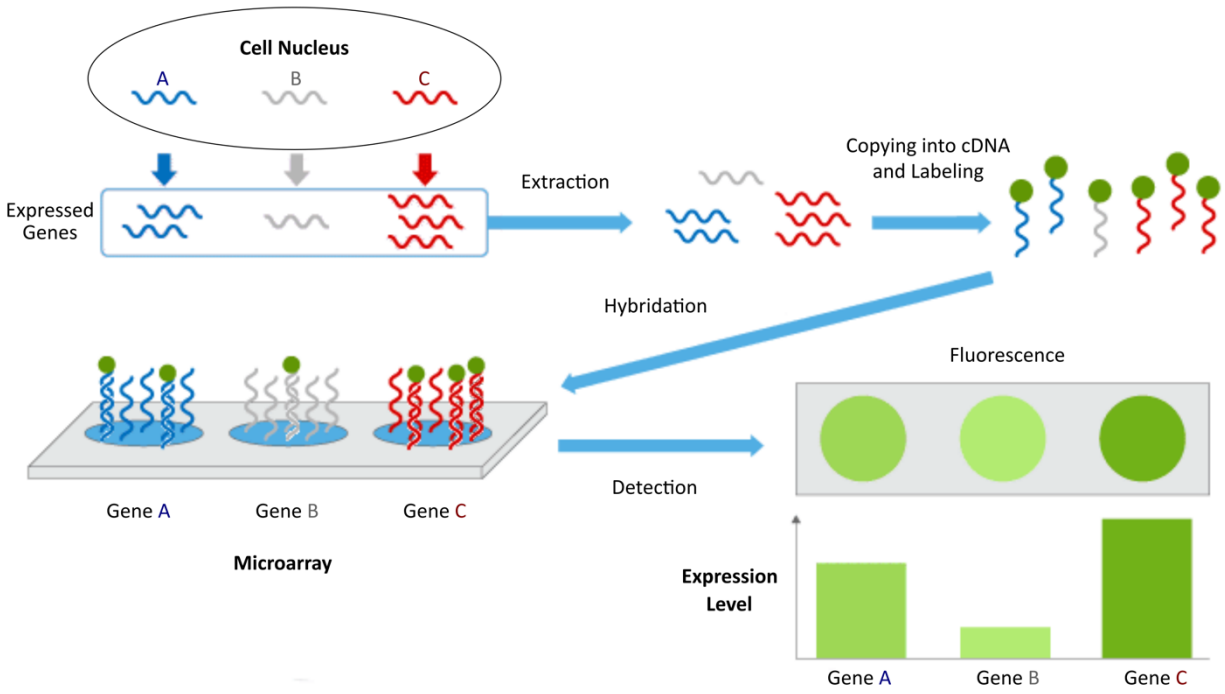


Figure 14. Gene expression microarrays. The fluoroescent label intensity at each spot correlates with the amount of bound DNA/RNA molecules and allows for quantification of the transcript abundance. Reproduced from "https://www.3d-gene.com/en/about/chip/chi_003.html"

Although nowadays many transcriptional studies rely on RNA sequencing to quantify transcription (mainly because it does not require the previous selection of probes), gene expression microarrays remain a robust, and proved by time, method for evaluating the transcriptional output of cells.

4.2.1.1 Analysis of gene expression microarray data

A typical microarray analysis pipeline includes the steps of quality control, normalization, summarization of probes into genes and finally the analysis of the obtained expression values.

The first thing to be done is to check the quality of the primary data. This step involves visually inspecting the distribution of array intensity values to check that no abnormalities are present, which may occur when the distribution of the sample over the array is suboptimal. This can happen for example when the tip of the micropipette used to load the microarray scratches its

surface or when air bubbles form on top of it. Next, a series of diagnostic plots are usually generated to try to identify problematic samples. These may vary in different settings, but usually involve evaluating the intensity of the background signal, the mean intensity of each array, the percentage of probes over the background signal and similar measurements.

Then next step is normalization of data. Normalization aims to compensate for technical variation between arrays while maintaining biologically significant variability. Technical variation can arise from a multitude of sources such as differing starting RNA quantities, differences in the labelling efficiencies of the dyes used, differences in the detection of those labels by the reader system and many others. Diverse methods to normalize microarray data have been proposed over the years, all of which depend on different assumptions on the data.

One of the first and more simple methods is total intensity normalization. Total intensity normalization assumes that, if the amount of total RNA of each sample is equal, total intensity values summed over all elements of the array should be roughly equal for all samples. To normalize the data, we simply calculate the coefficient between the summed intensity values of one sample over another and use this factor to scale the intensity values accordingly. This method relies on the assumption that the expression level of most of the genes does not change substantially between samples. In whole genome microarray experiments this is usually a fair assumption. Other similar approaches consist of deriving such normalization factor using only a subset of the data, such as a known set of housekeeping genes, exogenous RNA controls or rank invariant genes [375] (genes that show a similar rank in both samples), among others. However, none of these methods can account for systematic biases in the data.

One of the most well-known and established systemic biases in microarray data is that the differences between two samples signal intensity systematically depends on the intensity interval where those differences are observed [376,377]. Data points near the background signal (lowest intensities) are especially problematic in this regard. On top of discarding probes with near-background intensities, nowadays, most microarray normalization methods rely on locally

weighted linear regression (lowess) [378] or related methods as a means to remove this intensity-dependent bias. In short, instead of applying the same scaling factor to all the data, lowess methods scale each data point with a factor that gives more weight to data points with a similar intensity value.

A widely used technique for normalization of high-dimensional data (not only from microarray experiments) which does is quantile normalization (QN) [379]. QN consists of the following steps: (1) data for every sample is ranked by magnitude, (2) an average value for each rank (across samples) is calculated, and all values in that rank are set to this average, (3) then data for each sample is reordered back to its original order. This method generates datasets with equal distribution for all samples, which can then be compared between them. There are different ways of performing QN, i.e., splitting data by some class, performing QN on each split then joining the data, or using different aggregate measures instead of the mean, but they all rely on the same principles. The "perfect" distributions of QN normalized data can result very attractive, but the assumption that all samples have underlying similar distributions can lead to the removal of true differences (false negatives) and/or the appearance of false ones (false positives) [380]. Like with all normalization methods, the quality of the results will depend on how well does the data fit the underlying assumptions, something that should always be taken into account.

Once data has been normalized, the different probes that target a same transcript must be summarized into a single magnitude, using some aggregate measure (normally the mean or median value, or some related measure).

Finally, once data is normalized and measurements summarized into transcripts, we can start generating biologically relevant information from the data. This step can involve many different analyses, depending on the underlying hypothesis to be tested. The typical analysis consists in looking for sets of differentially expressed genes between two conditions of interest (treatment vs control, wild-type vs mutant, different cell types, etc.). Once sets of differnentially genes are generated, another typical step is performing a functional enrichment test. There are two main

alternatives to test for functional enrichment, classic Fisher test methods and gene-set enrichment analysis (GSEA) [381,382] methods. Both consists of identifying functional categories (cell processes, cell compartments, metabolic pathways, or any other grouping of genes) that are enriched in DE genes, assigning a functional meaning to the changes in gene expression. One of the most typical gene-sets used in such analysis in the Gene Ontology (GO) [383,384] database, where genes are associated to GO terms describing functions, processes, and locations from curated experimental findings. One important caveat of these methods is that they are reliant on the gene sets used to test for enrichment. This is especially problematic for not so well annotated organisms (such as *P. falciparum*), and it is a known fact that the enrichment results tend to be biased towards the most well annotated gene-sets in the tested suite. Other typical analyses include principal component analysis (PCA), which can be used both as a diagnostic or an exploratory test, different kinds of clustering using gene-expression values (classifying the samples into different phenotypical categories, e.g.), and the generation of custom plots which help to highlight interesting aspects of the data.

4.2.2 ChIP-Seq

ChIP-Seq consists in a series of steps which enable to sequence the fragments of DNA attached to a protein of interest, normally a transcription factor or histones containing a specific PTM. Typically, in the first step of the process, known as crosslinking, formaldehyde is used to create covalent unions between DNA and the proteins with which it is interacting (including histones and TFs). After crosslinking, DNA is fragmented into small size fragments (<500bp), which is usually achieved by sonication. The size of the fragments will be crucial in determining the resolution of the final output. After fragmentation, antibodies targeting the protein of interest are used to immunoprecipitate only those DNA fragments bound to it. Immunoprecipitated fragments are then de-cross-linked, purified and amplified during library preparation and can finally be sequenced generating millions of reads that will constitute the raw output of ChIP-Seq (Figure 15).

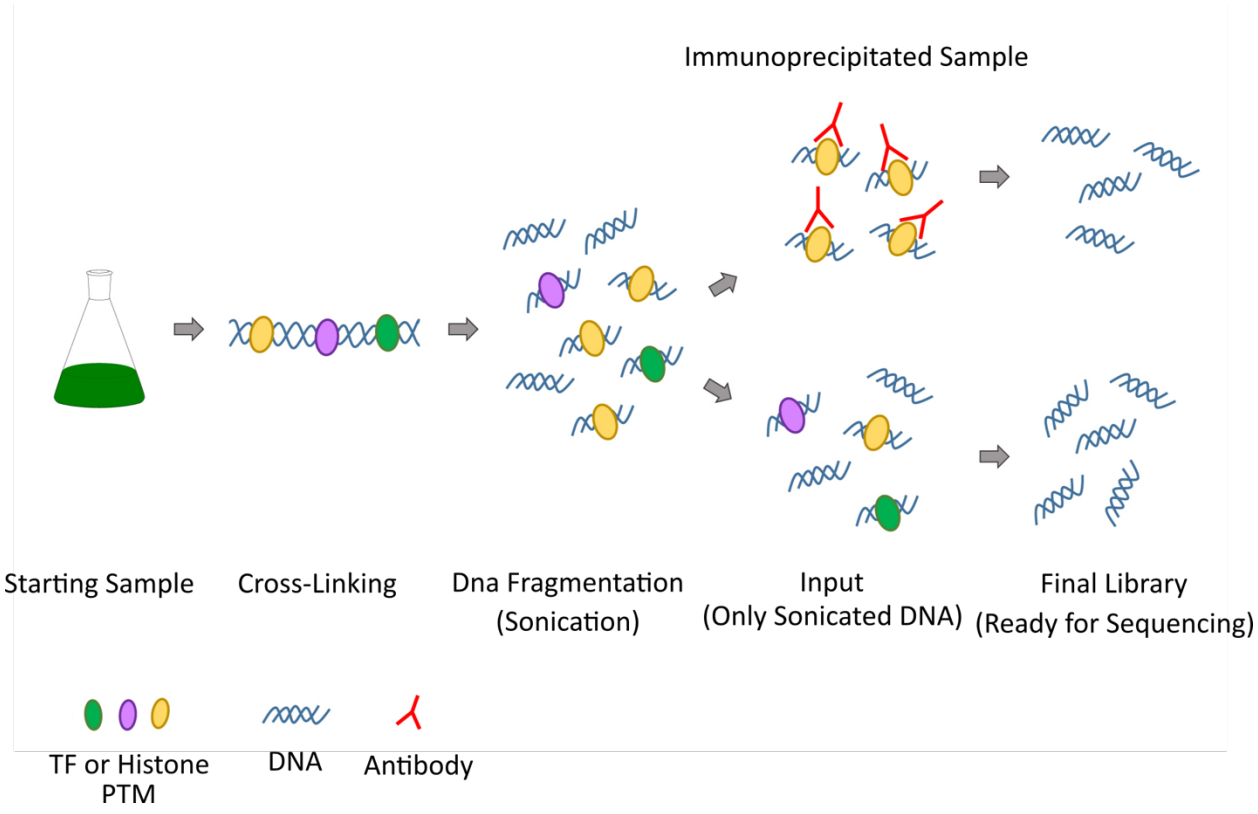


Figure 15. Schematic representation of Chromatin Immuno-Precipitation and Sequencing (ChIP-Seq). Reproduced from Giner-Lamia et al 2018.[422]

4.2.2.1 Analysis of ChIP-Seq data

A typical ChIP-Seq analysis pipeline includes the steps of quality control, "cleaning" of the reads, alignment, quantification and normalization of enrichment into coverage, peak-calling and differential peak-calling, assignment of peaks to genes and finally any downstream analysis to obtain any additional biological insights.

In this section we will describe the steps involved in analysing ChIP-Seq data in particular, but all "-Seq" techniques (RNA-Seq, ATAC-Seq, MNAse-Seq, etc.) start with the same initial analysis steps, they all produce a final set of reads (usually millions of them) and only differ after the aligning step.

As in the case of microarrays, the first step consists in ensuring that the raw data obtained from the sequencer is of optimal quality. The kind of diagnostics may vary from one experimental setting to another, but usually involves checking the number of obtained reads, the mean quality of the base-calls (as reported by the sequencer) along reads (usually the first and last few base pairs of a fragment have a lower quality), the distribution of read lengths, the mean GC content of reads, the duplication level of reads, the presence of adapter sequences and similar metrics. In the specific case of *P. falciparum*, given that the parasite needs to be cultured in blood samples, the presence of human DNA (that may come from any remaining white blood cells) can be checked.

Given that base-calls at the extremes of reads tend to have lower quality and that a certain amount of adapter sequences may still be present in the raw reads, usually some software to prune the reads in both ends and remove those repetitive sequences is applied.

Then comes a crucial step of the process, aligning the reads to a reference genome. Many different aligner tools are publicly available, such as Bowtie2 [385], BWA [386], STAR [387] or HISAT2 [388], among many others. Despite this variety, the use of one aligner or another should not influence the final result of the analysis in a significant manner, although run-time may vary [389]. Aligning the reads to the genome is what allows us to obtain relevant information from the data. After aligning we can finally evaluate which regions of the genome are enriched for our histone PTM or TF of interest. Alignment data contains all the information used by the aligner to allocate each read to a specific location in the genome, plus sequence data for all reads. Since a typical ChIP-Seq experiment generates millions of reads, alignment files require an important amount of computer memory to be stored and manipulated, and more compact format files are usually generated from alignments for downstream analysis, which only contain coverage values as we will explain in short. Once reads are aligned, another common diagnostic test consists of the evaluation of the correlation between the signal in replicate samples.

In a ChIP-Seq experiment we are interested in quantifying the amount of signal over specific regions of the genome. To do that, the genome is binned into intervals (which can be fixed-length intervals or genes or any other genomic feature) and the number of reads overlapping each bin is counted, a value known as coverage. The number of reads over a bin depends (in addition to the biological enrichment in which we are interested) on the total number of reads and the length of the bin. To normalize for these two factors, the number of reads overlapping a bin is divided by the total number of reads of the sample (in millions) and by the length of the bin (in kilo-bases (kb)), obtaining reads per kb per million reads (RPKMs).

Not all regions of the genome are equally amplified during the amplification step of the ChIP-Seq process, owing to differences in the ability of polymerases to amplify certain sequences and other technical factors. Additionally, when aligning to the reference genome, some sequences may be poorly aligned, due to repetitive/unspecific sequences, or sequence discrepancies between the sample and the reference genome. To normalize for this sequence-derived biases, ChIP-Seq samples are normally run with a parallel "input" sample in which the whole protocol is run but no immunoprecipitation step is done. Given that nothing has been enriched by immunoprecipitation, the "input" sample should give a homogeneous coverage over the genome and any enriched/depleted regions are attributable to the underlying sequence effects. This allows the "input" signal to be used to normalize for sequence derived bias, and ChIP-Seq coverage is usually represented as the ratio between the immunoprecipitated sample (IP) coverage and the corresponding "input" coverage, usually in a $\log_2$ scale.

After normalizing for number of reads, bin length, and sequence bias, we finally obtain $\log_2(RPKM_{IP}/RPKM_{input})$ as our coverage value for each bin in the genome. This coverage value for each bin can be represented as a histogram over the whole genome and is the underlying data for the coverage tracks with which ChIP-Seq data is usually visualized.

At this point, coverage tracks are usually visually inspected, using genome browsers such as IGV [390] or the UCSC genome browser [391], to make sure no evident problems are present.

Although in a typical ChIP-Seq experiment enriched regions are immediately distinguishable from non-enriched regions by eye, we need an unbiased approach to demarcate those regions of the genome that constitute an enrichment "peak" at a genome-wide scale. There are many different software tools designed to perform peak-calling on ChIP-Seq samples, some of which can be run directly on the alignment files and others that require a previous quantification step. Currently the most widely used tool is MACS2 [392], although many others exist. Unlike aligning, which produces very similar results irrespective of the aligner and the parameters used to run it, the choice of a peak-calling tool and its parameters have a profound effect on the resulting peaks set obtained. Since every histone PTM and TF has a different distribution profile and different organisms under different conditions may have different distributions of those same marks, it is intrinsically difficult to evaluate what should be a "proper" peak-calling. Although some research papers have addressed this question [393], there is no current consensus on a gold standard peak-caller or set of parameters to run it.

Once peaks in individual samples have been identified, the question remains whether specific regions present differential enrichment between different samples. As with individual peak-calling, many different tools exist to perform differential peak-calling such as MACS2 [392], DESeq2 [394], SICER [395] and MANorm [396], among others. These different tools generate substantially diverging results, and there is no current consensus on which is the most appropriate for a given set of conditions [397].

Unlike in microarray data, were probes target specific transcripts, ChIP-Seq peaks must be assigned to underlying genes based on proximity or some previously defined criteria. This poses additional challenges when the underlying genomes are only partially annotated. Should peaks overlapping non-annotated regions directly upstream of genes, putatively containing regulatory elements, be considered overlapping them? Should peaks that only overlap a small fraction of a gene (possibly at the 3' end) be assigned to that gene? This kind of questions, which usually don't have a correct or incorrect answer, must be addressed when deciding how to annotate the peaks

from a ChIP-Seq experiment, and they can have important implications in the final conclusions extracted from the data.

As was the case for microarray data, the number of downstream analyses that can be run once enrichment over the genome has been calculated, and a set of enriched peaks has been identified, is very variable, and depends on the research questions of interest. Functional enrichment of the genes underlying peaks or differential-peaks is a common possibility. Motif analyses of the sequences that underly enriched regions is another common possibility, especially in the case of ChIP-Seq analysis of TF targets. Enriched sequences can be queried against databases of known sequence motifs or can be tested for “de novo” motifs enriched in the set of sequences. Different kinds of plots can be used to highlight specific aspects of the data, such as the mean distribution of signal over coding genes, the differences between coding and non-coding regions, the differences in coverage in different sets of genes or any other aspect of interest for the researchers.

4.3 Specific difficulties in the analysis of *P. falciparum* transcriptional data

The most important challenge faced for the transcriptional analysis of *P. falciparum* samples is the periodic nature of its transcriptome. When analysing samples from most organisms, one is usually interested in the differences in gene expression between two different conditions. These conditions could be a treatment vs control, different cell types, mutant vs wildtype, cancerous vs normal, and many others, but we will refer to them generally as "conditions". To assess that, the gene expression of samples under both conditions is compared, and differences are attributed to the condition of interest. Replicates might be added to strengthen the results and compensate for background variability. The problem with *P. falciparum* samples is that progression along the IDC is, by far, the main driver of transcriptional change in the parasite. For this reason, minor differences in the age of the parasites between the compared samples can lead to substantial gene expression differences, which might be wrongly attributed to the condition under study. Since the magnitude of age-related transcriptional variation can be significantly higher than the magnitude of transcriptional variation derived from the study condition, it is very difficult to compensate this bias with the use of replicates. Moreover, many study conditions might influence the speed of progression along the IDC, generating an intrinsic age bias which cannot be compensated by replicates. This issue can be partially addressed experimentally by the synchronization of parasite cultures [398], but even if parasites are tightly synchronized to a 1 h window (most studies synchronize parasites to a ~5 h window), after a single round of replication, an age range of 3-5h in progression is expected to arise in the culture and the distribution of parasite ages in the culture might be different in different samples.

Two main strategies have been developed to overcome these limitations. First, the age of parasites in each sample can be estimated from their transcriptomes using the method described by Lemieux et al. [399], instead of relying solely on the experimental culture times. In brief, this method consists in assigning, by maximum-likelihood statistical methods, the sample of interest to the most similar timepoint from a reference transcriptome. By computing the parasite age in

this manner, differences in parasite age distribution in different samples (regardless of them being artefactual or study-condition related) can be taken into account when analysing the data.
The second strategy consists in analysing transcription changes not at singular timepoints, but along pre-defined time intervals [206]. Instead of calculating the gene expression fold-change (FC) comparing two single data-points, FC is averaged over a pre-defined timespan. One limitation of this method is that it requires repeated expression measures along the IDC.

Both strategies can be coupled together by using the statistically estimated "age" of the sample, instead of experimental hours post-invasion (hpi), and averaging FC over a relatively long period of time (e.g., the stage with peak expression of the gene) using these estimated timepoints. Together, these methods ensure that differences attributable to divergence in IDC progression are minimized and allows for detection of differences attributable solely to the condition of interest [206,398] (Figure 16).

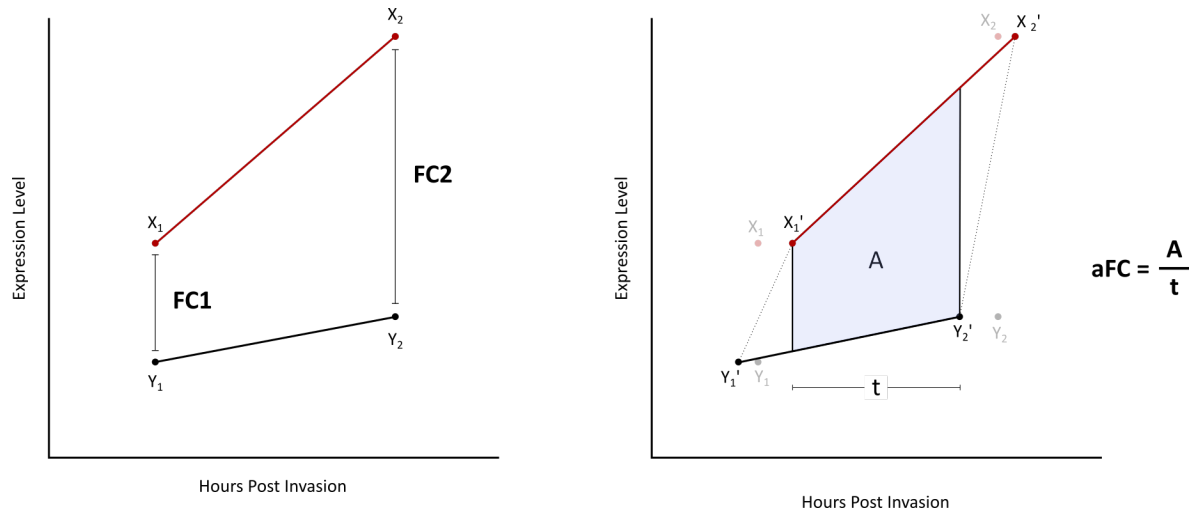


Figure 16. Age imputation and fold-change (FC) calculation on gene expression time-course experiments. X_1 and X_2 represent gene expression values over time for sample X, and Y_1 and Y_2 gene expresion values over time for sample Y. ***Left****: Regular calculation of FC using experimental hours post invasion (hpi).* ***Right****: Estimation of hpi from statistical methods resulting into the X -> X' and Y -> Y' imputations. Calculation of an averaged fold-change (aFC) as the area between the two expression lines on their common time span divided by the same time span*

Objectives

General Objectives

The main aim of this thesis is to better understand the molecular mechanisms that regulate transcriptional variation and adaptation to changes in the host environment in *P. falciparum*. Given the importance of clonally variant genes (CVGs) in the adaptation to many of the challenges during growth in human blood, we have focused our analysis on their regulation and the heterochromatin patterns that drive their transcriptional state. On top of investigating the general regulation of CVGs, we were especially interested in the regulation of sexual conversion, which also involves heterochromatin-based transcriptional control. Finally, we have also aimed to characterize the adaptation of the parasite to heat-shock, which does not involve changes in heterochromatin distribution, as a paradigm of non-CVG driven adaptation.

Specific Objectives

1. Characterize the heterochromatin patterns linked to the active and silenced state of CVGs at a genome-wide level in *P. falciparum*.
2. Gain new insights on the epigenetic regulation of sexual conversion in *P. falciparum*.
3. Characterize the adaptation of *P. falciparum* to heat-shock.
4. Understand how CVG expression patterns are affected by the passage through transmission stages and the establishment of human infections.
5. Optimize existing methods for the analysis of omics data generated by our team, given the peculiarities of the *P. falciparum* genome (cyclic gene expression patterns, extreme AT rich composition and others) and our specific research interests.

Results

Director's Report

Alfred Cortés
ICREA Research Professor

ISGlobal
CEK Building
C./ Rosselló 153, 1st floor
08036 BARCELONA
Catalonia, Spain
Tel.: +34 93 2275400 ext. 4276
alfred.cortes@isglobal.org

16 January 2023

Report of the Thesis supervisor: contribution of the PhD candidate to the scientific articles included in this Thesis

Article 1. "Patterns of Heterochromatin Transitions Linked to Changes in the Expression of *Plasmodium falciparum* Clonally Variant Genes".

Authors: Michel-Todó L, Bancells C, Casas-Vila N, Rovira-Graells N, Hernández-Ferrer C, González JR, Cortés A.

Published in ***Microbiol Spectr.*, 2022**, Dec 14:e0304922. doi: 10.1128/spectrum.03049-22. Online ahead of print. PMID: 36515553

Impact Factor: **9.043** (first quartile, Q1)

Lucas Michel-Todó is first author in this article. His contribution was to perform all the bioinformatic analysis of the data (including the development of new analysis methods), the interpretation of the results, preparing the figures and writing the manuscript, all under my supervision. The main importance of this article relies on data analysis, and the experimental work, performed by other members of the team, was a comparatively smaller part of the total effort involved. This article has not been included in any other PhD thesis.

Article 2. "A heat-shock response regulated by the PfAP2-HS transcription factor protects human malaria parasites from febrile temperatures".

Authors: Tintó-Font E, Michel-Todó L, Russell TJ*, Casas-Vila N, Conway DJ, Bozdech Z, Llinás M, Cortés A.*
** equal contribution*

Published in ***Nat Microbiol.*, 2021,** Sep;6(9):1163-1174. doi: 10.1038/s41564-021-00940-w. Epub 2021 Aug 16. PMID: 34400833

Impact Factor: **30.964** (first quartile, Q1)

Lucas Michel-Todó is second author in this article. He performed the bioinformatic analysis of genome-wide gene expression data (microarrays), which is a key part of the article. This was a complex dataset that required developing specific analysis methods. Many of the results in this article were included in the PhD thesis of Elisabet Tintó-Font, but the thesis (in classic format, not a compendium of articles) was defended long before the article was

published; therefore, the article has not been included in the PhD thesis of Elisabet Tintó-Font (or in any other PhD thesis).

Article 3. "Expression Patterns of *Plasmodium falciparum* Clonally Variant Genes at the Onset of a Blood Infection in Malaria-Naive Humans".

Authors: Pickford AK, Michel-Todó L, Dupuy F, Mayor A, Alonso PL, Lavazec C, Cortés A.

Published in ***mBio*, 2021,** Aug 31;12(4):e0163621. doi: 10.1128/mBio.01636-21. Epub 2021 Aug 3. PMID: 34340541

Impact Factor: **7.786** (first quartile, Q1)

Lucas Michel-Todó is second author in this article. He performed the bioinformatic analysis of genome-wide gene expression data (microarrays) and heterochromatin profiling by ChIP-seq, and the integration of the two datasets. Genome-wide data analysis is a very substantial part of this article. The subject of this article is closely related with the first article listed, and Lucas also did important contributions to the interpretation of the results and developed specific analysis methods. This article was included in the PhD thesis of Anastasia K. Pickford (first author of the article).

Article 4. "Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*".

Authors: Llorà-Batlle O, Michel-Todó L, Witmer K, Toda H, Fernández-Becerra C, Baum J, Cortés A.

Published in ***Sci Adv.*, 2020,** Jun 10;6(24):eaaz5057. doi: 10.1126/sciadv.aaz5057. eCollection 2020 Jun. PMID: 32577509

Impact Factor (year of publication): **14.136** (first quartile, Q1)

Lucas Michel-Todó is second author in this article. He performed the bioinformatic analysis of genome-wide gene expression data (microarrays) and heterochromatin profiling by ChIP-seq. This article was included in the PhD thesis of Oriol Llorà-Batlle (first author of the article).

I confirm that the subject of the four articles is closely related, and all are within the subject of Lucas' Thesis work.

Alfred Cortés Closas
ICREA Research Professor at ISGlobal

1 **Article 1:** Patterns of heterochromatin transitions linked to changes in the expression of *Plasmodium falciparum* clonally variant genes

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Patterns of Heterochromatin Transitions Linked to Changes in the Expression of *Plasmodium falciparum* Clonally Variant Genes

Lucas Michel-Todó,[a] Cristina Bancells,[a] Núria Casas-Vila,[a] Núria Rovira-Graells,[a] Carles Hernández-Ferrer,[b,c] Juan R. González,[b,c] Alfred Cortés[a,d]

[a]ISGlobal, Hospital Clínic, Universitat de Barcelona, Barcelona, Catalonia, Spain
[b]ISGlobal, Barcelona, Catalonia, Spain
[c]Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[d]ICREA, Barcelona, Catalonia, Spain

ABSTRACT The survival of malaria parasites in the changing human blood environment largely depends on their ability to alter gene expression by epigenetic mechanisms. The active state of *Plasmodium falciparum* clonally variant genes (CVGs) is associated with euchromatin characterized by the histone mark H3K9ac, whereas the silenced state is characterized by H3K9me3-based heterochromatin. Expression switches are linked to euchromatin-heterochromatin transitions, but these transitions have not been characterized for the majority of CVGs. To define the heterochromatin distribution patterns associated with the alternative transcriptional states of CVGs, we compared H3K9me3 occupancy at a genome-wide level among several parasite subclones of the same genetic background that differed in the transcriptional state of many CVGs. We found that *de novo* heterochromatin formation or the complete disruption of a heterochromatin domain is a relatively rare event, and for the majority of CVGs, expression switches can be explained by the expansion or retraction of heterochromatin domains. We identified different modalities of heterochromatin changes linked to transcriptional differences, but despite this complexity, heterochromatin distribution patterns generally enable the prediction of the transcriptional state of specific CVGs. We also found that in some subclones, several *var* genes were simultaneously in an active state. Furthermore, the heterochromatin levels in the putative regulatory region of the *gdv1* antisense noncoding RNA, a regulator of sexual commitment, varied between parasite lines with different sexual conversion rates.

IMPORTANCE The malaria parasite *P. falciparum* is responsible for more than half a million deaths every year. *P. falciparum* clonally variant genes (CVGs) mediate fundamental host-parasite interactions and play a key role in parasite adaptation to fluctuations in the conditions of the human host. The expression of CVGs is regulated at the epigenetic level by changes in the distribution of a type of chromatin called heterochromatin. Here, we describe at a genome-wide level the changes in the heterochromatin distribution associated with the different transcriptional states of CVGs. Our results also reveal a likely role for heterochromatin at a particular locus in determining the parasite investment in transmission to mosquitoes. Additionally, this data set will enable the prediction of the transcriptional state of CVGs from epigenomic data, which is important for the study of parasite adaptation to the conditions of the host in natural malaria infections.

KEYWORDS ChIP-seq, H3K9me3, *Plasmodium falciparum*, chromatin, *gdv1*, clonally variant gene expression, epigenetics, heterochromatin, malaria, transcriptomics

The apicomplexan parasite *Plasmodium falciparum*, the causative agent of malaria, has a complex life cycle divided between female *Anopheles* mosquitoes and humans. Human infection starts with the injection of parasites at the sporozoite stage during an

Editor Laura A. Kirkman, Weill Cornell Medicine
Ad Hoc Peer Reviewers Francesca Florini, Weill Cornell Medicine; Jun Miao, University of South Florida

Address correspondence to Alfred Cortés, alfred.cortes@isglobal.org.

The authors declare no conflict of interest.
Received 4 August 2022
Accepted 22 November 2022

Anopheles blood meal. After multiplication in the liver, thousands of parasites at the merozoite stage are released into the bloodstream. Merozoites invade red blood cells (RBCs) and start continuous rounds of the asexual intraerythrocytic development cycle (IDC), in which parasites develop through the intraerythrocytic ring, trophozoite, and schizont stages and then burst and release up to 32 new merozoites. Repeated rounds of the ~48-h IDC are responsible for all of the clinical manifestations of malaria. At each round, some parasites differentiate into sexual forms called gametocytes, which, after developing for ~10 days, can infect an *Anopheles* mosquito during a new blood meal. In the mosquito vector, male and female gametocytes mate and form zygotes that, after meiosis and several asexual replication steps, generate new sporozoites (1).

The complex life cycle of *P. falciparum* requires adaptation to the different conditions of each niche, which is in large part regulated at the transcriptional level. Therefore, the parasite has a different transcriptome at each stage of development (2–4). In addition, the parasite must adapt to fluctuations in the conditions of each niche. In the human blood, where infections can last for several months, parasites are exposed to changes in the availability of nutrients, in temperature (associated with fever episodes), in the presence of toxic compounds such as antimalarial drugs, and in host immune responses, among others (5–8). Rapid adaptation to fluctuations in the conditions of the human blood can be achieved by directed transcriptional responses, which involve sensing followed by changes in gene expression that protect against the challenge. However, malaria parasites can produce directed transcriptional responses to only a limited number of conditions (9–12). An alternative adaptive mechanism to survive under fluctuating conditions is based on epigenetic variation. In *P. falciparum*, epigenetic variation is linked to clonally variant genes (CVGs), which can be found in different transcriptional states (active or silenced) in parasites living in the same environment, with the same genome sequence, and at the same stage of the IDC. The transcriptional state of CVGs is transmitted from one round of the IDC to the next, but low-frequency switches between the active and silenced states occur, ensuring the constant generation of epigenetic diversity within parasite populations (13–17). As a consequence of this expression dynamics, different individual parasites within an isogenic population have different combinations of active and silenced CVGs. Transcriptional heterogeneity in parasite populations results in phenotypic heterogeneity, which provides the grounds for the continuous natural selection of epigenetic variants best adapted to the dynamic conditions of the environment. This constitutes a bet-hedging adaptive strategy (18).

Essentially all CVGs are involved in host-parasite interactions (13–17). The best-characterized family of CVGs is the *var* gene family, consisting of ~60 genes that encode *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), a major virulence factor. PfEMP1 is exposed at the surface of infected RBCs and mediates antigenic variation and adhesion to endothelial cell receptors, which results in parasite sequestration within different host tissues. Individual parasites generally express a single member of the *var* gene family in a mutually exclusive manner, which is important for immune evasion (19–21). Other gene families linked to processes such as antigenic variation (e.g., *rifin* and *stevor*) (21), RBC remodeling (e.g., *pfmc-2tm* and *phist*) (22, 23), RBC invasion (e.g., *eba* and *Rh*) (24), lipid metabolism (e.g., *acs* and *acbp*) (25), or solute transport and nutrient acquisition (*clag*, with *clag3.1* and *clag3.2* displaying mutually exclusive expression) (26–29), among others, also show clonally variant expression (13, 18, 19, 26, 30–33). A peculiar CVG is the *pfap2-g* gene, which encodes a transcription factor of the ApiAP2 family that drives the conversion of asexual forms into nonreplicative sexual gametocytes (34–36). This gene is silenced in asexually growing parasites and, upon activation, triggers a transcriptional cascade that results in sexual conversion (37, 38). The proportion of parasites that activate *pfap2-g* at each round of the IDC determines the proportion of parasites that convert into sexual forms, i.e., the sexual conversion rate.

The epigenetic regulation of the expression of *P. falciparum* CVGs is based on reversible chromatin transitions between euchromatin (active) and facultative heterochromatin (silenced) states (13–17). The heterochromatic state is associated with the

histone posttranslational modification histone H3 trimethylation of lysine 9 (H3K9me3) and the associated heterochromatin protein 1 (PfHP1) (39–42). Furthermore, several ApiAP2 DNA binding proteins such as PfAP2-HC are also associated with heterochromatic regions (43, 44). The active state of CVGs is characterized by the absence of H3K9me3 and PfHP1 as well as the acetylation of H3K9 (H3K9ac) (45–48). The chromatin at CVG loci is considered bistable because, once established, both the euchromatic and the heterochromatic states are maintained across the full IDC and can be transmitted to the next generations. Infrequent transitions between the euchromatic and heterochromatic states at these loci underlie the expression switches of CVGs (15, 16).

In higher eukaryotes, H3K9me3 is typically associated with constitutive heterochromatin localized in subtelomeric and pericentromeric repetitive regions, but recent research has demonstrated that it also participates in cell differentiation processes (49, 50). In *P. falciparum*, H3K9me3 and PfHP1 are absent from pericentromeric regions, but they occupy subtelomeric repeats in addition to CVG loci. CVGs are located mainly in subtelomeric regions and a few chromosome internal clusters, and there are also a few stand-alone CVG loci in internal chromosome regions (40, 42, 44, 51, 52). In total, over 500 genes (~10% of the genome) carry heterochromatin marks and are considered CVGs. In contrast to the restricted distribution of H3K9me3, the histone mark associated with the active state of CVGs, H3K9ac, is widespread across the *P. falciparum* genome and marks constitutively euchromatic genes in addition to active CVGs. In non-CVGs, H3K9ac levels correlate with transcript levels and the temporal patterns of gene expression across the IDC (53, 54).

Previous studies showed that the overall distribution of heterochromatin at CVGs is stable across different stages of the *P. falciparum* IDC (40, 45–47, 52) but changes during transmission stages (52, 55–57). Indeed, in contrast to the relatively stable expression patterns of CVGs during multiple rounds of the IDC, the expression patterns of CVGs are reset during transmission stages (58–60). The overall distribution of heterochromatin is similar across *Plasmodium* spp., suggesting that it plays similar roles across the genus (52). A recent study showed that different *P. falciparum* lines have a similar overall distribution of heterochromatin, with differences at specific loci possibly reflecting different transcriptional states of the genes at these loci among the parasite lines analyzed, although gene expression was not assessed (52). However, the parasite lines used in that study had not been recently subcloned. Consequently, they were likely transcriptionally heterogeneous such that for many CVGs, the population is a mixture of parasites that have the gene in an active state and parasites that have it in a silenced state, as previously demonstrated for other parasite lines under long-term culture (18).

Here, we combined heterochromatin profiling with transcriptomic analysis to investigate in detail the chromatin differences between the active and silenced states of CVGs at a genome-wide level. For this, we compared the distribution of H3K9me3 between subcloned lines of the same genetic background that differ in the expression of dozens of CVGs. The use of subclones ensures that most genes are in a homogeneously active or silenced state within each population. We also compared the genome-wide distribution of heterochromatin between parasite lines that produce different numbers of gametocytes, with the aim of identifying epigenetic differences linked to sexual conversion rates.

RESULTS

Overview of the study. We analyzed the genome-wide distribution of H3K9me3 using chromatin immunoprecipitation followed by sequencing (ChIP-seq) and the transcriptome across the full IDC using two-channel microarrays for five different subclones of the 3D7 genetic background: 1.2B, 10G, A7, E5, and B11 (Fig. 1). The transcriptome analysis of the 1.2B and 10G subclones was previously reported (18). The analysis of 5 subclones enables 10 possible pairwise comparisons to identify genes that are active in one subclone and silenced in another. The selection of the subclones was based on

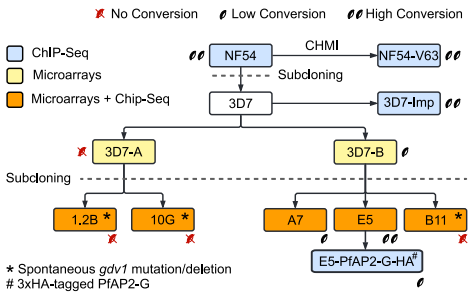


FIG 1 Schematic of the parasite lines used in this study. 3D7-A and 3D7-B are stocks of the clonal 3D7 line that were maintained in different laboratories for some years. Blue indicates parasite lines that were analyzed by ChIP-seq, yellow indicates parasite lines for which new or previously published transcriptomic data (using microarrays) were generated, and orange indicates parasite lines for which both ChIP-seq and transcriptomic data were generated. Parasite lines with a spontaneous deletion or a premature stop codon in *gdv1* (*) or with a genetic edition in the *pfap2-g* locus (#) are indicated. Parasite lines are semiquantitatively classified as high-gametocyte producers (two gametocytes, >10% sexual conversion rate), low-gametocyte producers (one gametocyte, <3% sexual conversion rate), or non-gametocyte producers (red, crossed gametocytes). NF54-V63 was obtained after human infection with the NF54 line in a controlled human malaria infection (CHMI) trial.

maximizing the number of differentially expressed genes among them and covering a broad range of sexual conversion rates. Subclones 1.2B and 10G were derived from the 3D7-A stock of 3D7, whereas A7, E5, and B11 were derived from the 3D7-B stock. The 3D7-A and 3D7-B stocks were maintained in different laboratories for several years and have phenotypic differences in growth rates, sexual conversion rates, density-dependent growth inhibition, the use of invasion pathways, and cytoadherence (26, 34, 61–64). The basal sexual conversion rate of the subclones ranges from zero (10G, 1.2B, and B11) to intermediate (<3%) (A7) or high (~15%) (E5) (34). To preserve transcriptional homogeneity, subclones were maintained in culture for the shortest time needed to obtain sufficient RNA or chromatin for the analyses (~4 to 5 weeks between limiting dilution to obtain the subclones and epigenomic or transcriptomic analysis).

Furthermore, we analyzed H3K9me3 occupancy in additional parasite lines with different basal sexual conversion rates: an E5-derived transgenic parasite line in which PfAP2-G was 3× hemagglutinin (3×HA) tagged (E5-PfAP2-G-HA) and the sexual conversion rate was lower (<3%) than that in the parental E5 line (34, 65); a stock of 3D7 obtained from Imperial College (3D7-Imp) with a high basal sexual conversion rate (66) (conversion rate in Albumax medium, 18.3% [range, 16.6 to 20%] [n = 2]); and the NF54 isolate from which the 3D7 clone was obtained, also a high-gametocyte producer (66) (conversion rate in Albumax medium, 21.8% [range, 18.8 to 24.8%] [n = 2]). In some of the analyses, we also included published ChIP-seq data for the NF54-V63 line that was recovered after infecting a volunteer with the NF54 line in a controlled human malaria infection (CHMI) trial (58). We also analyzed the distribution of H3K9ac in several of the parasite lines.

Transcriptomic analysis of 3D7 subclones. To identify genes that were transcriptionally active in one subclone and silenced in another, we analyzed the transcriptomes of tightly synchronized cultures (5-h age window) of the A7, E5, and B11 lines at six time points along the IDC. The analysis pipeline to identify genes with large expression differences between the subclones, similar to the method used in our previous study including 10G, 1.2B, and 3D7-B (18), was based on the statistically estimated parasite age (67), the average fold change (AFC) in transcript levels (between subclones) at four overlapping time intervals of the IDC, and the maximum (among the time intervals) average fold change (mAFC) parameter (18, 58). After applying several filters (see Materials and Methods), we generated for each pair of subclones a high-confidence list of differentially expressed genes with an mAFC of >4 [$\log_2$(mAFC) of >2].

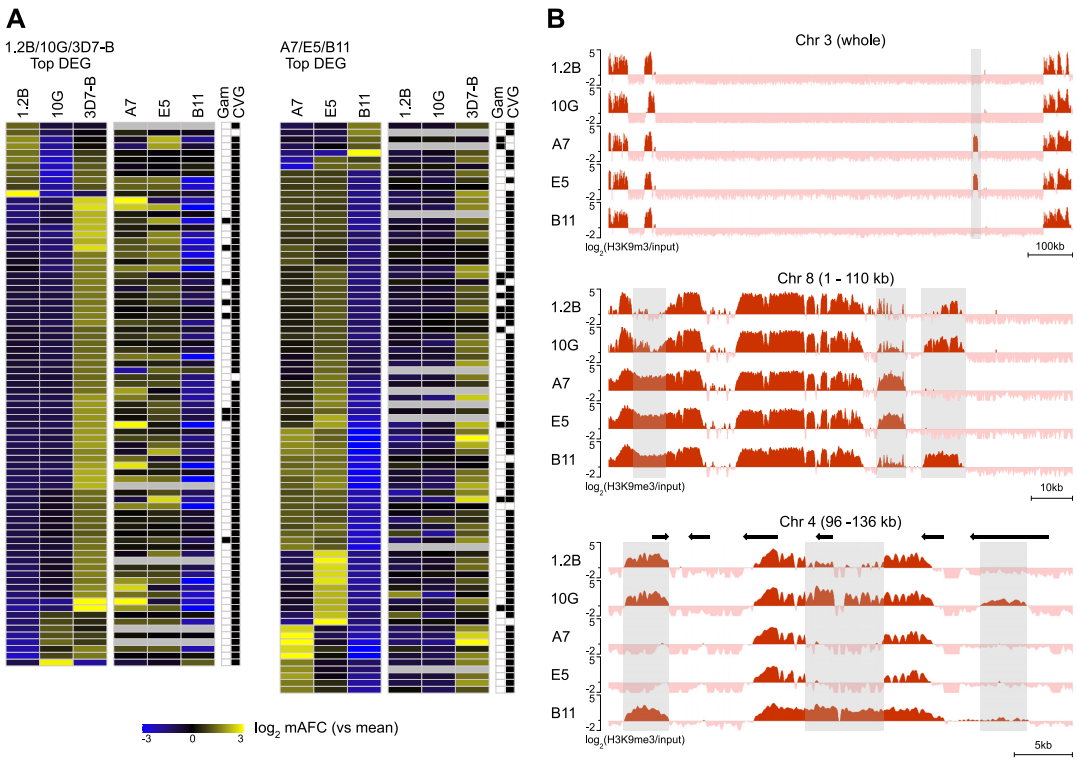


FIG 2 Overview of the transcriptomic and epigenomic analyses of the subclones. (A) Expression patterns of the top differentially expressed genes (DEG). (Left) Genes with differential expression (mAFC of >4) among the 1.2B, 10G, and 3D7-B lines (previously published data [18]). Values for the expression differences in the same genes among the A7, E5, and B11 lines (newly generated data) are shown. All values are the mean-centered $\log_2$ of the mAFC in expression [$\log_2$(mAFC)] among the three parasite lines. (Right) Analogous analysis for the top differentially expressed genes in the comparison of A7, E5, and B11. The last two columns indicate whether a gene is a known gametocyte marker or has been previously classified as a CVG. The values presented in this panel are provided in Data Set S2 in the supplemental material. (B) H3K9me3 distribution in a representative chromosome or chromosome regions in the five subclones at different magnification levels. Gray shades indicate regions with major differences in H3K9me3 coverage among subclones. Values are $\log_2$(H3K9me3/input).

Between 19 and 74 genes were differentially expressed in the pairwise comparisons between 10G and 1.2B or among A7, E5, and B11 (Fig. 2A; see also Data Sets S1 and S2 in the supplemental material). Since 3D7-A subclones (1.2B and 10G) and 3D7-B subclones (A7, E5, and B11) were analyzed using microarrays with slightly different designs (18, 35, 68) and with different preparations of the reference pool, a direct comparison between subclones of the two different groups was not possible. However, we combined data for A7, E5, and B11 with data for the direct comparison of 1.2B and 10G against 3D7-B (18) to identify genes that were likely differentially expressed between 10G or 1.2B and A7, E5, or B11 (see Materials and Methods). In total, 121 genes were differentially expressed with high confidence in at least one of the possible pairwise comparisons, of which 111 were genes previously classified as CVGs based on the presence of heterochromatin marks or variant expression in previous studies (see the list of CVGs in reference 58). A few of the differentially expressed genes are gametocyte markers, and their different transcript levels may reflect the different sexual conversion rates among the subclones (Fig. 2A).

A large fraction of the differentially expressed genes showed lower expression levels in B11 than in A7 and E5. Unexpectedly, many of these genes were also expressed at lower levels in 1.2B and 10G than in 3D7-B, suggesting that the CVG expression patterns in B11 were relatively similar to those in 1.2B and 10G (Fig. 2A). Analysis of

genetic polymorphisms in these parasite lines using the ChIP-seq input data revealed several large deletions in B11, including an ~20-kb deletion affecting the *gdv1* locus (69, 70) (Fig. S1 and Data Set S3). Of note, 10G and 1.2B also have a defect in *gdv1*, in this case a mutation that results in a premature stop codon (Q578X) (Data Set S3) (35).

Heterochromatin distribution in the 3D7 subclones and general association with gene expression. ChIP-seq analysis of H3K9me3 revealed similar global distribution of heterochromatin among the five subclones, although differences were apparent at specific loci (see Fig. 2B for representative examples). The mean pairwise correlation among parasite lines was 0.88 (range, 0.83 to 0.95) using input-normalized coverage values (Fig. S2). We performed individual peak calling using the MACS2 tool (Data Set S4) and differential peak calling for all possible pairwise comparisons using custom scripts. There was an average of 142 (range, 68 to 214) differential peaks in pairwise comparisons, which map to an average of 62 (range, 34 to 99) genes (Data Set S5). Differential peaks involving changes in H3K9me3 coverage at the upstream region (1,000 bp) or the first half of the coding sequence (CDS) of a gene were generally associated with transcriptional differences such that the parasite line with higher coverage had lower transcript levels. However, when H3K9me3 coverage was different in only the second half of the coding sequence or the downstream region of a gene, typically, there were no associated transcriptional differences (see Fig. 3A for the comparison of 10G versus 1.2B and Fig. S3A to C for the other direct pairwise comparisons; Data Set S2). In addition to differences in the heterochromatin occupancy observed at chromosome regions containing genes, there were also differences at the telomere-associated repetitive elements (TAREs) (71, 72) located between the telomeres and subtelomeric genes (Fig. S1).

To further assess the relationship between differences in heterochromatin distribution and transcriptional differences, we first calculated for all differentially expressed genes in any of the pairwise comparisons the level of correlation between the H3K9me3 coverage at different landmark regions of the genes and transcript levels. The H3K9me3 coverage was calculated for fixed-length regions at different positions relative to the start or stop codons and also at the 5′ untranslated region (UTR), the coding sequence, and the 3′ UTR (according to the genome annotation in PlasmoDB v52). The strongest negative correlation between H3K9me3 coverage and transcript levels occurred at the 5′ UTR and in the fixed-length −500 bp to ATG regions, but strong correlations (Pearson's $|r|$ value of >0.6) were observed throughout the upstream and coding sequences (Fig. S4). In genes with large transcriptional differences (mAFC of >4) among the subclones, there was a strong negative association between transcript levels and H3K9me3 coverage at the region from −1000 to +500 bp such that the subclone that expressed a gene at higher levels tended to have less heterochromatin (Fig. 3B, Fig. S3D to F, and Data Set S2).

Changes in heterochromatin distribution at CVG loci between the active and silenced states. To characterize in more detail the changes in the heterochromatin distribution that drive the transition between the silenced and active states of CVGs, we analyzed for each gene the H3K9me3 coverage across 23 bins encompassing the neighboring genes, the upstream region (until the first upstream gene), the CDS, and the downstream region (until the first downstream gene). For the 121 genes in the high-confidence list of differentially expressed genes, we compared the heterochromatin distribution between the two subclones with the maximum expression differences (Fig. 4). The subclone with the highest expression level was used to represent the active state of the gene, and the subclone with the lowest expression level was used to represent the silenced state. For genes in which the largest expression difference occurred between the 1.2B or 10G subclone and 3D7-B, with higher expression levels in 3D7-B than in 10G or 1.2B, we used the 3D7-B subclone (A7, E5, or B11) with the highest expression level as the subclone representative of the active state (and vice versa; i.e., for genes with lower expression levels in 3D7-B than in 10G or 1.2B, the 3D7-B subclone with the lowest expression level was used as the representative of the silenced state). The H3K9me3 coverages in the active and silenced states and the coverage difference

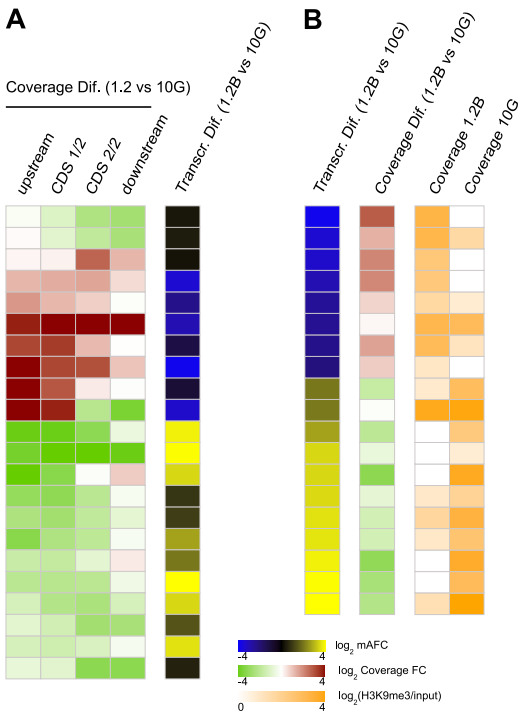


FIG 3 Association between H3K9me3 coverage differences and transcriptional differences. Results for subclones 1.2B and 10G are shown (see Fig. S3 in the supplemental material for the other pairwise comparisons). (A) Transcriptional differences in genes overlapping (from −1000 to +500 bp relative to the ATG start codon) an H3K9me3 differential peak between 1.2B and 10G. The $\log_2$ of the input-normalized coverage fold change (FC) in 1.2B relative to 10G is shown for the upstream region (1,000 bp before the ATG codon), the first or second half of the coding sequence (CDS 1/2 or CDS 2/2, respectively), and the downstream region (1,000 bp after the stop codon). Transcriptional differences are expressed as the $\log_2$ of the mAFC (1.2B versus 10G). (B) Heterochromatin levels in genes differentially expressed between 1.2B and 10G (mAFC of >4). Transcriptional differences are shown as described above for panel A. Heterochromatin coverage ($\log_2$ of the input-normalized coverage) and coverage differences between 10G and 1.2B ($\log_2$ of the input-normalized coverage fold change) are shown for the region from −1000 to +500 bp relative to the ATG codon. The values presented in this figure are provided in Data Set S2.

between the two were used to group the 121 differentially expressed genes into seven distinct clusters using the *k*-means algorithm (Fig. 4 and Data Set S2).

Cluster 1 (22 genes) corresponds to genes that are largely devoid of heterochromatin in the upstream and coding regions in their active state but fully heterochromatic in these regions and also in the downstream region in their silenced state. We called this the "global transition" cluster (Fig. 4 and Fig. 5A to C). Genes in cluster 2 (14 genes) have heterochromatin patterns that are relatively similar to those of the genes in cluster 1, but in the silenced state, the downstream region remains euchromatic ("mainly 5′ transition" cluster) (Fig. 4 and Fig. 5D and E).

Clusters 3, 4, and 5 (20, 13, and 29 genes, respectively) correspond to genes in which the full locus (clusters 3 and 5) or the upstream region and the beginning of the coding region (cluster 4) are marked by H3K9me3 in the silenced state, and in the active state, there is only a localized reduction of the coverage of this mark in the upstream region (Fig. 4 and Fig. 5F and G). We called these the "localized 5′ transition" clusters. In a few genes from these clusters, changes in gene expression were not accompanied by noticeable differences in heterochromatin levels. We cannot exclude

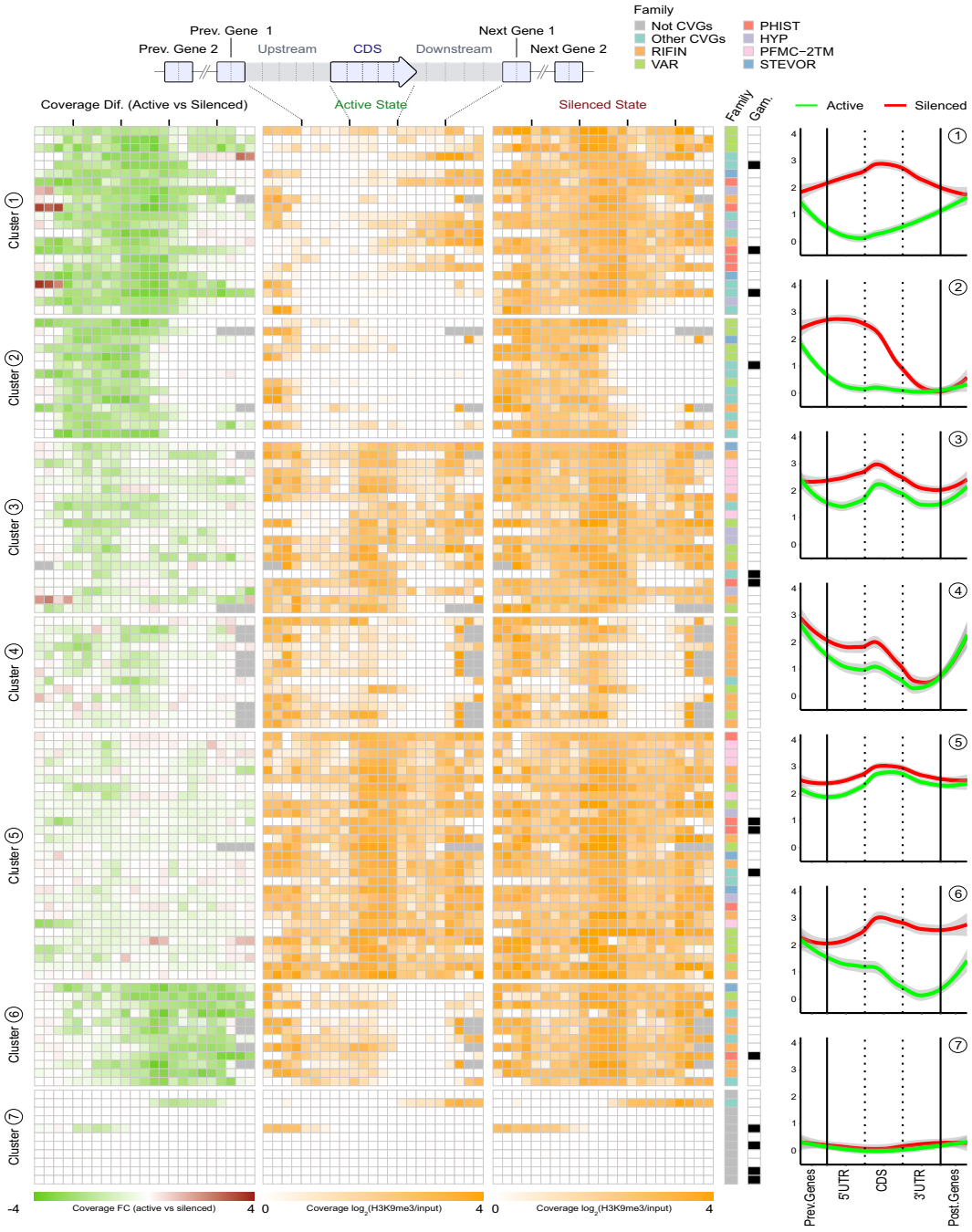


FIG 4 Distribution of heterochromatin in the active and silenced states of CVGs. The H3K9me3 coverage in the top differentially expressed genes among 1.2B, 10G, A7, E5, and B11 in the active compared to the silenced state is shown. For each gene, the subclone with the highest expression level was

(Continued on next page)

the possibility that, even in recently subcloned lines, some of these genes may be expressed in only a small fraction of the parasites in the subclone in which they are tagged as active.

Cluster 6 (12 genes) includes genes in which heterochromatin is present in the upstream, coding, and downstream regions in the silenced state, whereas in the active state, heterochromatin is depleted mainly in the second half of the coding region and the downstream region ("mainly 3′ transition" cluster) (Fig. 4 and Fig. 5H and I). This pattern suggests that heterochromatin in the downstream region of these genes may play a role in their transcriptional regulation, although we cannot exclude the possibility that their expression was actually driven by differences in the upstream region. Indeed, differences in the upstream region between the active and silenced states occurred in the majority of cluster 6 genes, albeit these were less pronounced than the differences in the downstream region.

Finally, cluster 7 (11 genes) corresponds to genes largely devoid of heterochromatin in both the active and silenced states. Therefore, the transcriptional variation of these genes is not controlled by euchromatin-heterochromatin transitions. Of note, the vast majority of the genes in this cluster had not been previously classified as CVGs.

Except for the non-CVGs from cluster 7, in the majority of the differentially expressed genes, there was heterochromatin at some part of the locus or at the neighboring genes, even in the active state there was heterochromatin at some part of the locus or at neighboring genes. Furthermore, the transition from the active to the silenced state was typically associated with an increase in heterochromatin levels in a continuous region spanning from nearby heterochromatic regions to the regulatory regions of the gene. This is consistent with the silencing and activation of CVGs commonly involving the spreading and retraction of heterochromatin, respectively, according to an accordion-like mechanism. In several cases, this involved changes in the position of the outer limit of the heterochromatin domain (Fig. 5A, D, and H to J), whereas in other cases, the transition from the silenced to the active state was associated with reduced heterochromatin levels in a region surrounded by heterochromatin on both sides. The latter is consistent with an opening at an internal point within the heterochromatin domain followed by heterochromatin retraction (Fig. 5B, F, and G). We defined four main transition patterns between the active and silenced states of CVGs (Fig. 5L) and qualitatively assigned each of the 109 classifiable differentially expressed genes in clusters 1 to 6 to the pattern with which they best matched (Fig. 5L and Data Set S5). Only 5 differentially expressed genes showed transitions involving heterochromatin *de novo* formation or complete removal, whereas 65 genes showed transitions consistent with heterochromatin expansion/retraction, either affecting the external limits of a heterochromatin domain ["expansion/retraction (limits)"] (42 genes) or involving an opening within a heterochromatin domain that was continuous in the silenced state ["expansion/retraction (internal)"] (23 genes). In 39 genes, the transition between the active and silenced states involved small changes affecting heterochromatin levels in only a very small region ("localized closing/opening") (Fig. 5L). However, it is important to note that the four general patterns likely represent a mechanistic continuum rather than discrete, fully independent mechanisms, and for some genes, the classification was ambiguous. Furthermore, in some genes, transitions followed a different pattern depending on the state of the neighboring genes.

FIG 4 Legend (Continued)
selected as the representative for the active state, and the subclone with the lowest expression level was selected as the representative for the silenced state. The distribution of heterochromatin in each subclone ($\log_2$ of the input-normalized coverage) and the coverage difference ($\log_2$ of the fold change [FC] in coverage in the active versus the silenced state) are shown for 23 bins spanning the coding sequence (CDS) (divided into 5 bins), the upstream and downstream regions (up to the next gene) (5 bins each), and the coding sequences of the two upstream and downstream neighboring genes (2 bins each). Genes were clustered according to their H3K9me3 coverage and coverage differences (excluding the bins from the neighboring genes) using *k*-means clustering ($k = 7$, guided by silhouette-and-elbow plots). Within each cluster, genes are ordered by fold changes in transcript levels between the active and silenced states. The two columns at the right of the heat map indicate the gene family and whether a gene is a known gametocyte marker. The plots on the right are locally estimated scatterplot smoothing (LOESS) regression plots (shades are 95% confidence intervals) showing the heterochromatin distribution ($\log_2$ of the input-normalized H3K9me3 coverage) in the active and silenced states for each cluster. The details of the genes included in this figure are provided in Data Set S2 in the supplemental material.

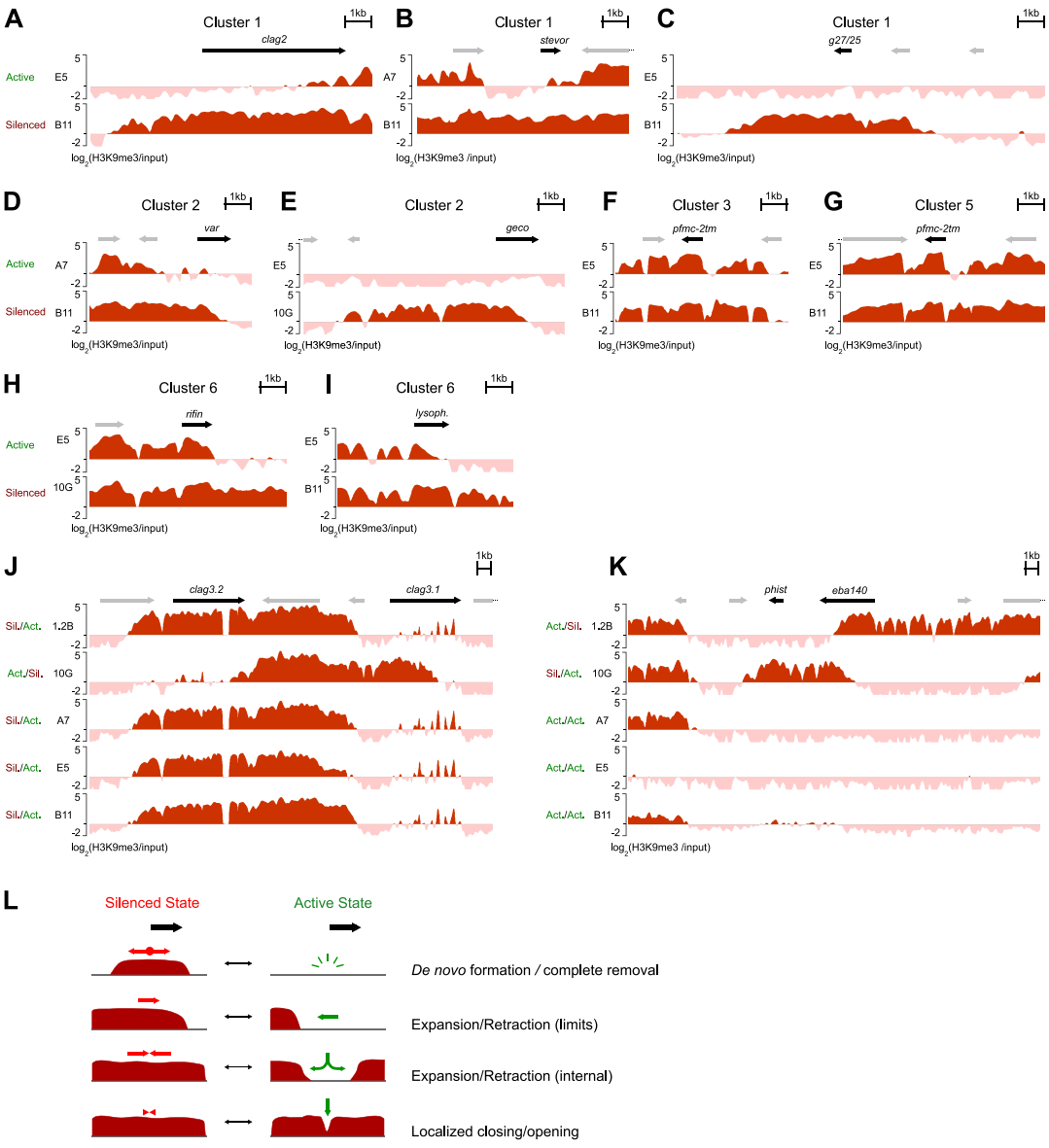


FIG 5 Representative examples of the different modes of heterochromatin transitions between the active and silenced states of CVGs. (A to I) Representative examples of heterochromatin distribution [$\log_2$(H3K9me3/input)] associated with the active or silenced state of CVGs from the different clusters shown in Fig. 4. The genes are representative of "global transition" (cluster 1 [A to C]), "mainly 5′ transition" (cluster 2 [D and E]), "localized 5′ transition" (cluster 3 [F] and cluster 5 [G]), or "mainly 3′ transition" (cluster 6 [H and I]). Panels C and E are representative of transitions that involve the *de novo* formation/complete removal of a heterochromatin domain, whereas the other panels are representative of transitions that can be explained by heterochromatin expansion/retraction. (J and K) Representative examples of heterochromatin distribution in subclones that have a gene in a different transcriptional state and subclones that have it in the same transcriptional state. In all panels, the active or silenced state corresponds to the genes displayed as black arrows (with names). Other annotated genes are shown as gray arrows. The identifiers of the genes displayed as black arrows are PF3D7_0220800 (A), PF3D7_0300900 (B), PF3D7_1302100 (C), PF3D7_0713300 (D), PF3D7_1253000 (E), PF3D7_1039700 (F), PF3D7_0114100 (G), PF3D7_1480000 (H), PF3D7_0936700 (I), PF3D7_0302200 and PF3D7_0302500 (J), and PF3D7_1301500 and PF3D7_1301600 (K). (L) Schematic representation of the four main modalities of heterochromatin transitions between the active and silenced states of CVGs: the *de novo* formation of heterochromatin or the complete disruption of a heterochromatin domain (*De novo* formation/complete removal); the expansion/

(Continued on next page)

While for most genes activation or silencing could be explained by heterochromatin expansion or retraction across a continuous region or small localized transitions, there were 13 H3K9me3 differential peaks in which the *de novo* formation or the complete removal of a heterochromatin domain actually appeared to occur (Fig. 5C and E and Data Set S5). This involved heterochromatin islands isolated from other heterochromatic regions in which, in the active state, the region occupied by these islands and the adjacent regions (at least 5 kb on each side) were fully euchromatic. Unexpectedly, the majority of the 20 genes embedded in these regions were expressed at very low levels both in subclones in which they were euchromatic and in subclones in which they were heterochromatic, and only 5 of them showed differential expression among the 5 subclones (i.e., they were included in the high-confidence list of 121 differentially expressed genes). Indeed, most of these 20 genes were previously reported to be predominantly expressed in gametocytes or mosquito stages rather than during the IDC (Data Set S5). Together, these results suggest that variant gene expression linked to the adaptation of asexual blood-stage parasites to fluctuating conditions during the IDC is regulated mainly by heterochromatin expansion and retraction, whereas the *de novo* formation or the complete removal of heterochromatin is largely restricted to heterochromatin islands containing genes expressed at other stages of development.

The heterochromatin distribution associated with the active or silenced state of a specific CVG is generally conserved. For most CVGs, the analysis included more than one subclone in which the gene was either active or silenced. To compare the distribution of heterochromatin between subclones in which the gene was in the same transcriptional state, we first classified all genes in the five subclones as CVGs or non-CVGs and as active or silenced/inactive based on our transcriptomic data (Cy5 signal and expression differences between subclones) and publicly available transcriptome sequencing (RNA-seq) data sets (see Materials and Methods). Genes for which the active or silenced state could not be unambiguously determined were classified as "undetermined" (Data Set S6). In the majority of CVGs, the distribution of heterochromatin was almost identical among subclones in which the gene was in the same transcriptional state (Fig. 5J, Fig. S5, and Data Set S2). In the few genes in which prominent heterochromatin differences were observed between subclones that had the gene in the same state, differences typically occurred toward the end of the coding sequence or downstream region and could be explained by a different transcriptional state of the neighboring genes. As an example, in the active state, *eba-140* had heterochromatin in the second half of the coding sequence and downstream region only when the downstream *phist* gene was silenced (Fig. 5K).

Heterochromatin differences between the active and silenced states in different families of CVGs. None of the clusters described in Fig. 4 were strictly linked to a specific CVG family, but different families were unevenly represented among the clusters. Different scenarios were observed for different gene families (Fig. S6). In the *rif*, *stevor*, *pfmc-2tm*, and *var* families, the majority of genes that did not show transcriptional differences among the five subclones analyzed were fully heterochromatic, whereas in the *clag*, *acs*, *hyp*, and *phist* families, nondifferentially expressed genes were typically euchromatic. In some gene families, the majority of differentially expressed genes showed similar heterochromatin transition patterns between the active and silenced states (e.g., localized 5′ transition in the *pfmc-2tm* family), but in other families, the transition patterns were rather heterogeneous.

In specific genes for which we had previously characterized the heterochromatin

FIG 5 Legend (Continued)
retraction of heterochromatin at the boundaries of a heterochromatin domain [Expansion/Retraction (limits)]; the opening of heterochromatin within a heterochromatin domain, followed by heterochromatin retraction, or the expansion resulting in a continuous heterochromatin domain from two separate neighboring domains [Expansion/Retraction (internal)]; or the opening/closing within a heterochromatin domain with minimal or no expansion/retraction (Localized closing/opening). Changes associated with silencing and activation are indicated by red and green arrows, respectively.

distribution associated with the active or silenced state in the 10G and 1.2B subclones using chromatin immunoprecipitation-quantitative PCR (ChIP-qPCR) (45), the results were fully consistent. In the mutually exclusively expressed *clag3.1* and *clag3.2* genes (26, 45, 73), a heterochromatin domain centered at the *var* pseudogene located between the two genes expanded toward one gene or the other (Fig. 5J). In *eba-140*, the active state was associated with the absence of heterochromatin in the upstream region and the beginning of the coding sequence, whereas heterochromatin in the downstream region did not influence the expression of the gene and reflected the silencing of the neighboring *phist* gene, as described above (Fig. 5K).

Multiple *var* genes are simultaneously active in some subclones. Genes of the *var* family typically show mutually exclusive expression such that individual parasites express only one gene of the family at a time and keep all of the others repressed (19–21). To obtain a semiquantitative estimate of the expression patterns of *var* genes in the five 3D7 subclones, we used the normalized Cy5 signal (i.e., not relative to the reference pool) (Fig. 6A), as we did in previous studies (18, 58). Subclones 1.2B and B11 showed the expected pattern of mutually exclusive *var* gene expression, with a single *var* gene being predominantly expressed, whereas no expressed *var* gene was identified in 10G. Unexpectedly, subclones A7 and E5 expressed several *var* genes (four and six, respectively) at relatively high levels.

Two different scenarios can explain this observation: (i) authentic simultaneous expression of multiple *var* genes in individual parasites or (ii) transcriptional heterogeneity such that despite the use of recently subcloned lines, different subsets of parasites express different *var* genes. To distinguish between these two possibilities, we used the H3K9me3 ChIP-seq data. If each *var* gene was active in only a subset of the parasites, a partial reduction of heterochromatin levels would be expected, but heterochromatin was almost completely depleted at the regulatory regions of all but one (PF3D7_0413100 in A7) of the highly expressed *var* loci (Fig. 6B). This result suggests that multiple *var* genes were actually simultaneously active in the A7 and E5 subclones. The majority of the *var* genes that were expressed at high levels in one or more of the subclones and silenced in the others showed major differences between the active and silenced states (cluster 1, 2, or 6 in Fig. 4), although two showed more localized differences (cluster 3) (Data Set S2). While an unambiguous demonstration of the simultaneous expression of multiple *var* genes in the same cell requires analyses at the single-cell level, such as RNA fluorescence *in situ* hybridization (RNA-FISH) or single-cell RNA-seq (scRNA-seq), and we did not determine if the multiple *var* genes in an active chromatin state produce full-length mRNA and PfEMP1 protein, our results are consistent with previous reports indicating that the mutually exclusive expression of *var* genes may not be strict (74, 75).

Differences in heterochromatin distribution between parasite lines with different sexual conversion rates. The proportion of parasites that convert into gametocytes at each round of the IDC is called the sexual conversion rate. It is well established that different parasite lines, and even different subclones of the same genetic background, have different rates of basal (noninduced) sexual conversion that are inherited across multiple generations of asexual blood growth (34, 76). Sexual conversion depends on the activation of *pfap2-g*, which is normally silenced by heterochromatin in asexually growing blood-stage parasites. This gene is a peculiar CVG because, unlike other CVGs, only the silenced state can be transmitted from one generation of the IDC to the next: parasites in which the gene is activated convert into nonreplicating gametocytes and abandon the IDC. Therefore, the stable differences in the conversion rates between parasite lines imply the transmission of the probability of activation of *pfap2-g* rather than the transmission of an active or silenced state. To identify potential epigenetic determinants of the probability of activation of *pfap2-g*, we compared the distribution of heterochromatin among the five 3D7 subclones and additional parasite lines with a broad range of sexual conversion rates (0 to >10%) (Fig. 1A). All parasite lines were maintained in standard culture medium containing Albumax II and no human serum, with the exception of NF54-V63 (previously published data), which was analyzed in medium containing human serum (58).

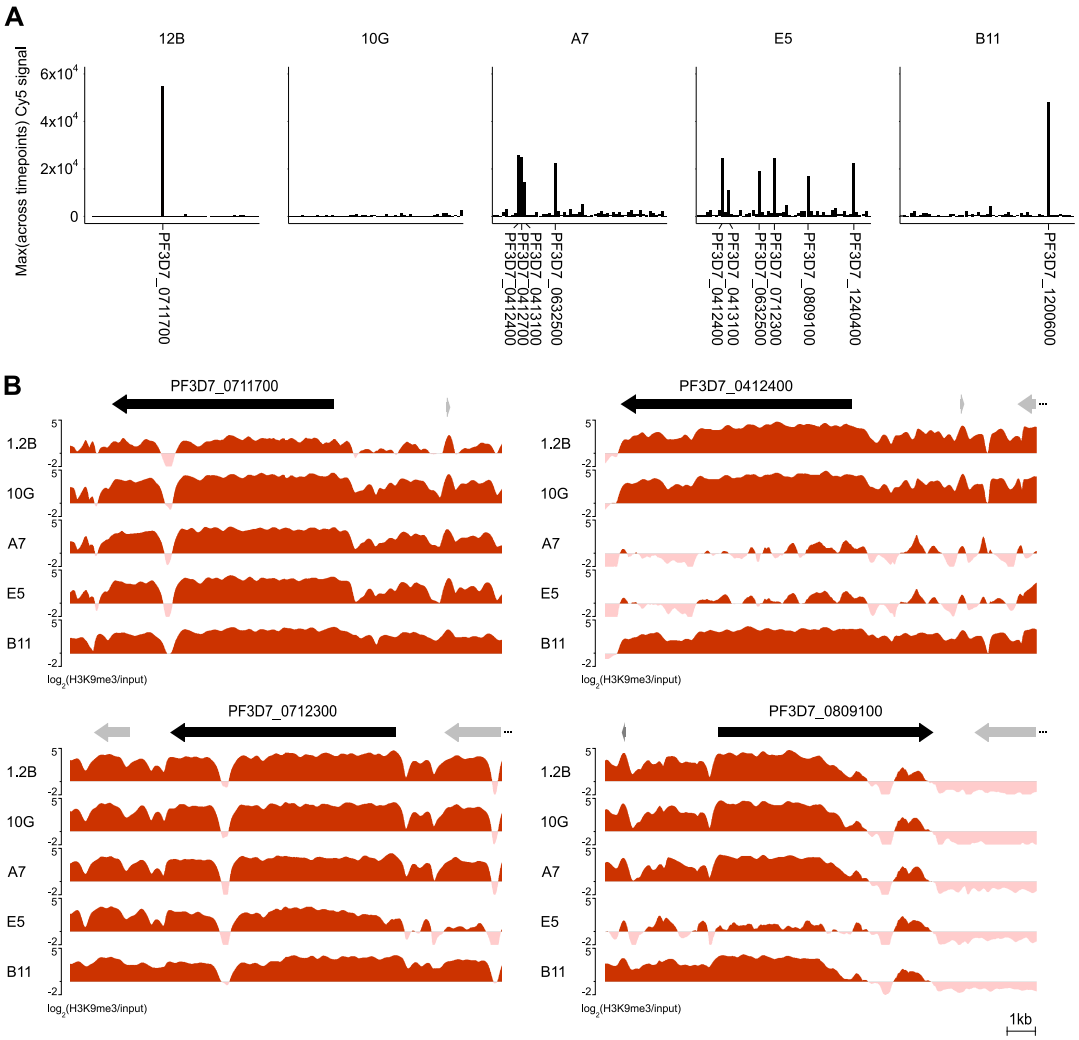


FIG 6 Expression patterns and heterochromatin distribution at *var* gene loci. (A) Expression of *var* genes in the five subclones. Values are the normalized sample signal (Cy5 channel), which provides a semiquantitative estimation of the absolute expression levels. The gene identifier is provided only for *var* genes expressed at high levels (normalized Cy5 signal intensity of >10,000). (B) H3K9me3 distribution at representative *var* genes that are highly expressed in some of the subclones. Black arrows (with identifiers) are the differentially expressed *var* genes; other annotated genes are shown as gray arrows. Values are $\log_2$(H3K9me3/input).

The heterochromatin occupancy at the *pfap2-g* locus was similar among all of the parasite lines, with similar 5′ and 3′ boundaries of the H3K9me3-enriched region (Fig. 7). Therefore, a different extension of the heterochromatin domain at the *pfap2-g* locus does not appear to underlie the different probabilities of *pfap2-g* activation between parasite lines. There were subtle differences between parasite lines in the H3K9me3 coverage at the 5′ and 3′ ends of the H3K9me3 domain, but these may be explained by the different proportions of parasites that were sexually committed (and therefore had *pfap2-g* in an active state) rather than authentic epigenetic differences transmitted across multiple rounds of the IDC.

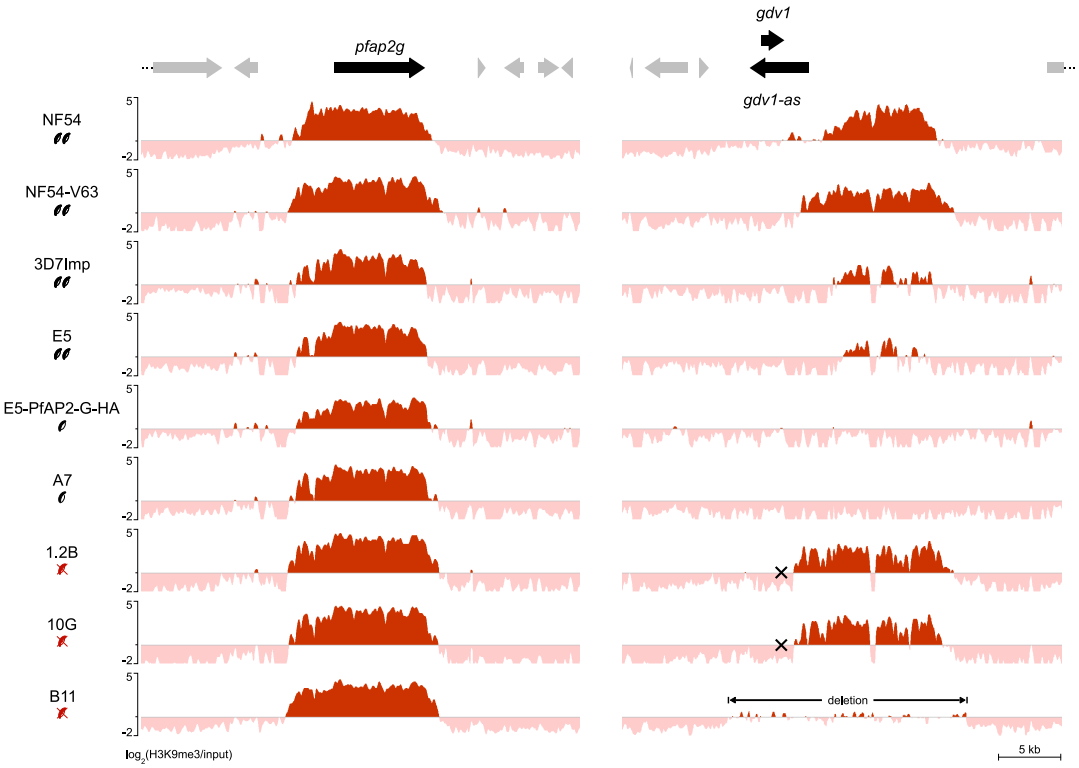


FIG 7 Distribution of heterochromatin at the *pfap2-g* and *gdv1* loci encoding regulators of sexual conversion. The distribution of H3K9me3 at the *ap2-g* and *gdv1* loci in different parasite lines is shown. Values are $\log_2$(H3K9me3/input). Parasite lines are semiquantitatively classified as high-gametocyte producers (two gametocytes, >10% sexual conversion rate), low-gametocyte producers (one gametocyte, <3% sexual conversion rate), or non-gametocyte producers (red, crossed gametocytes), as described in the legend of Fig. 1. Mutations that produce a premature stop codon in *gdv1* (crosses) or a deletion including the full *gdv1* locus are indicated. Other annotated genes are indicated as gray arrows.

In contrast, we observed major differences between the parasite lines in the heterochromatin distribution at the locus encoding GDV1, an upstream positive regulator of *pfap2-g* (69, 70, 77) (Fig. 7). While *gdv1* is typically not considered a CVG or a heterochromatic gene, in some of the parasite lines, we observed a prominent H3K9me3 signal downstream of the coding sequence. This heterochromatin domain overlaps the putative promoter of the long noncoding antisense RNA (*gdv1-as*) that was shown previously to operate as a negative regulator of GDV1 (69) and other previously reported long noncoding RNAs located in this large intergenic region (78). Of note, parasite lines with high (>10%) sexual conversion rates (NF54, NF54-V63, 3D7-Imp, and E5) (34, 58, 66) tended to have prominent heterochromatin at this position, whereas heterochromatin was absent from parasite lines with low (<3%) conversion rates (A7 and E5-PfAP2-G-HA, the latter derived from E5 but with a much lower conversion rate after genome editing and subcloning) (34, 65). The exceptions were the parasite lines 1.2B and 10G, which had prominent heterochromatin at the *gdv1* locus despite not producing gametocytes. However, these subclones have a nonsense mutation that results in a premature stop codon in *gdv1* (Data Set S2) (35) that likely makes the protein nonfunctional and uncouples the regulation of the *gdv1* locus from sexual conversion. Another study also showed that the premature truncation of GDV1 makes the protein nonfunctional (79). The other non-gametocyte-producer line in our study, B11, has a large deletion that spans the *gdv1* coding sequence and the downstream

region (Fig. S1). Together, these results suggest that heterochromatin at the *gdv1* locus plays a role in the regulation of sexual conversion rates, possibly by affecting the expression of *gdv1-as*.

Changes in H3K9ac associated with the transcriptional state of CVGs. We also characterized the distribution of H3K9ac in the 10G, 1.2B, and B11 subclones by ChIP-seq as an exploratory analysis to determine if, in combination with H3K9me3, analysis of this mark may help to identify specific signatures for active and silenced CVGs. H3K9ac is typically associated with the active state of CVGs (45–48) and is also abundant at constitutively euchromatic genes, mainly in intergenic regions (53, 54). There was high overall similarity in the distribution of H3K9ac between subclones (mean pairwise correlation using input-normalized values of 0.92 [range, 0.90 to 0.96]) (Fig. S2A), and the distribution of H3K9ac was largely nonoverlapping with that of H3K9me3, as previously described (51, 54) (Fig. 8A and B). Active CVGs typically had H3K9ac in the region depleted of H3K9me3 (Fig. 8B). We also generated H3K9ac ChIP-seq data for the parasite lines 3D7-Imp, NF54, and E5-PfAP2-G-HA (available in the GEO database), but these data were not further analyzed because transcriptomic analysis was not performed for these parasite lines.

To compare the distribution of H3K9me3 and H3K9ac between the active or silenced CVGs and the active or inactive non-CVGs, we used the classifications of the genes in these different categories for each subclone (Data Set S6). We compared the coverages of H3K9ac and H3K9me3 in the upstream (1,000 bp), coding, and downstream (1,000 bp) regions between genes in the different categories (Fig. 8C). As expected, H3K9me3 was absent from non-CVGs and more abundant in silenced CVGs than in active CVGs. In contrast, the H3K9ac coverage was higher in non-CVGs than in CVGs. Among the CVGs, the H3K9ac coverage was higher in active CVGs than in silenced CVGs, especially in the upstream and coding regions. In these regions, the H3K9ac coverage in active CVGs was almost as high as that in non-CVGs. We also compared the ratios of H3K9ac to H3K9me3, which revealed a pattern similar to that for H3K9ac levels.

To determine if genes in the different categories showed distinct H3K9 signatures, we generated two-dimensional scatterplots with H3K9me3 and H3K9ac coverages at the upstream, coding, and downstream regions (Fig. 8D). Both non-CVGs and silenced CVGs formed compact clusters, whereas the active CVG cluster was more scattered and overlapped the two other clusters. Thus, the global levels of H3K9 epigenetic marks alone cannot be used to fully discriminate between non-CVGs and active CVGs or between active and silenced CVGs. Instead, algorithms that consider the distribution of H3K9 marks associated with the active or silenced state of each specific CVG will be needed to predict the transcriptional state of a CVG from epigenomic data.

DISCUSSION

Changes in the expression of CVGs underlie several important processes in malarial host-parasite interactions, including antigenic variation, alteration of infected RBC cytoadherence tropism and solute permeability, the use of alternative invasion pathways, and sexual conversion (13, 14, 17). Transitions between the heterochromatic and euchromatic states at CVG loci drive expression switches, but the heterochromatin distribution associated with the active and silenced states has been previously characterized for only a few genes (45–48). Here, we systematically characterized the heterochromatin distribution associated with the active and silenced states of over a hundred CVGs, which revealed diverse modalities of heterochromatin changes associated with transcriptional switches. Of note, all of the characterized expression switches and heterochromatin changes occurred during the IDC because all of the subclones were derived from the clonal 3D7 line and had not gone through mosquito passage since the 3D7 line was established. In the majority of the 121 genes for which we compared the distribution of heterochromatin between the active and silenced states, the transitions can be explained by heterochromatin retraction and expansion. Heterochromatin spreading was previously demonstrated to occur in *P. falciparum* (73). Therefore, we propose that CVG expression

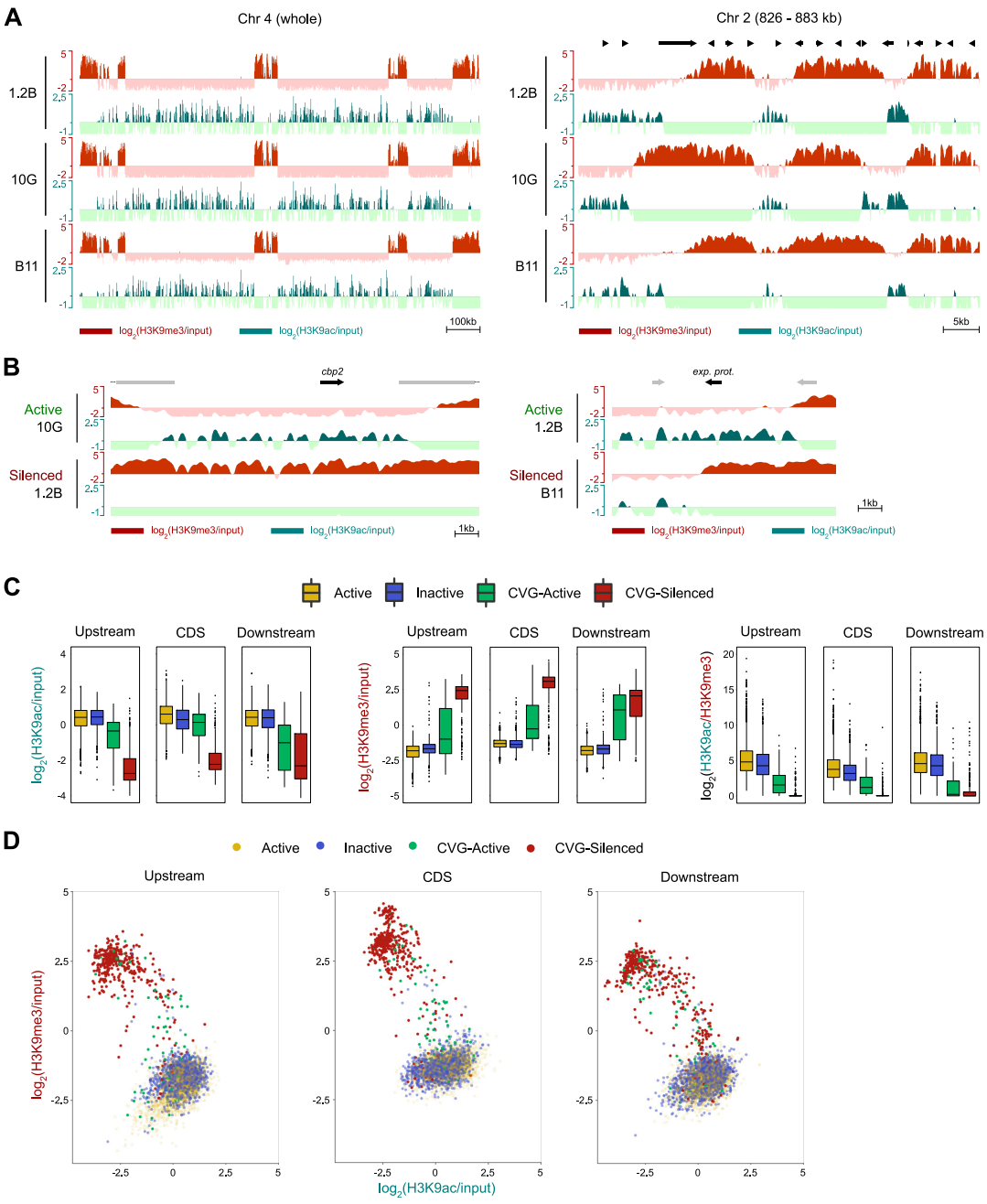


FIG 8 Distribution of H3K9ac and H3K9me3 in CVGs and non-CVGs. (A) General distribution of H3K9ac and comparison with the distribution of H3K9me3 in a representative chromosome or chromosomal region. (B) Representative examples of changes in H3K9me3 and H3K9ac associated with the active or silenced states of CVGs. The identifiers of the genes displayed as black arrows are PF3D7_1301700 (left) and PF3D7_0532600 (right). Other annotated genes are indicated as gray arrows. (C) Mean coverage of H3K9ac and H3K9me3 and H3K9ac/H3K9me3 ratios at the upstream region (1,000 bp), the coding

(Continued on next page)

during the IDC is typically regulated by an accordion-like mechanism rather than by the *de novo* formation or the complete removal of heterochromatin domains (Fig. 5L). The latter mechanism appears to be common only for genes that are expressed at other stages of the life cycle (i.e., transmission stages). The significance of heterochromatin at these genes for expression during transmission stages is not known.

A key aspect of the design of our study was the use of subclones that had been cultured for a limited number of generations after subcloning, which ensures that they were largely homogeneous for the transcriptional state of the majority of the CVGs. In parasite lines that have not been recently subcloned, different individual parasites have different combinations of expressed and silenced CVGs (18), and this heterogeneity precludes the detailed characterization of the active and silenced states of the genes. Other important aspects of our study design were comparing subclones of the same genetic background and combining genome-wide epigenomic analysis at high resolution (using small chromatin fragments) with transcriptomic analysis across the full IDC of the same parasite lines.

Because the *P. falciparum* genome is compact, with relatively small intergenic regions, the distribution of heterochromatin in one CVG locus was often influenced by the state of neighboring CVGs. This likely contributes to the complexity and variety of patterns of heterochromatin differences observed between the active and silenced states. For example, in a gene that has a heterochromatic downstream neighbor, the end of the coding sequence is typically heterochromatic even in the active state. Despite this, CVGs are regulated as individual units such that a CVG can be activated while its neighboring CVGs remain silenced. This suggests the existence of boundary elements, such as barrier insulators (80, 81), that prevent the spreading of heterochromatin beyond certain sequence elements. Based on the observation that there are positions beyond which heterochromatin never spreads, neither in the parasite lines analyzed here nor in parasite lines of different genetic backgrounds (52), we speculate that there may be "strong" barrier elements that strictly prevent spreading into euchromatic areas containing essential non-CVGs. Additionally, there are specific positions where heterochromatin spreading commonly stops in different subclones that have a CVG in the active state but not in subclones in which the gene is silenced (heterochromatin domain limits are generally conserved among subclones with a CVG in the same state). This may reflect the presence of "weak" barrier elements that enable the stable activation of a CVG amid heterochromatic regions but can eventually be overcome, resulting in a switch toward the silenced state.

The phenotype of a parasite depends not only on its genome sequence but also on how it uses its CVGs. In natural human malaria infections, thousands of full parasite genomes have been sequenced (82, 83), but characterizing transcriptomes across the full IDC is far more complex (16). The main difficulty in studying *P. falciparum* gene expression in human infections is that only ring-stage parasites and mature gametocytes are found in the circulation, whereas all other blood stages are sequestered in different organs or tissues and therefore are not available in peripheral blood samples. Consequently, the expression of genes such as *var* or *pfap2-g*, which are active in ring stages, can be directly characterized in human blood samples (20, 84–89), but the expression of genes expressed at other stages of the IDC cannot. In studies using genome-wide transcriptome analysis of patient samples, only the expression of genes expressed in rings is captured (16, 84, 90, 91). Some full-transcriptome studies (58, 92, 93) and studies focused on specific CVGs such as *clag* (60) or invasion genes (31, 94, 95), all expressed in late stages of the IDC, used *ex vivo* parasite cultures to obtain samples for transcriptional analysis at different stages of the IDC. However, this approach is

FIG 8 Legend (Continued)
sequence (CDS), and the downstream region (1,000 bp) of genes classified as being (non-CVG) active, (non-CVG) inactive, CVG active, or CVG silenced. All genes from the 1.2B, 10G, and B11 subclones assigned to one of these categories were included in the analysis. Boxes are interquartile ranges (with medians), and whiskers indicate the 5th and 95th percentiles. (D) Scatterplot of H3K9ac versus H3K9me3 coverage for genes classified as being active, inactive, CVG active, or CVG silenced.

time-consuming and requires culture facilities at the place of sample collection, which compromises its feasibility. Together, these difficulties limit our capacity to investigate CVG expression under different clinical presentations, in hosts of different ages, or under different epidemiological conditions.

The distribution of heterochromatin does not change during the full IDC (45–47, 52), implying that H3K9me3 or PfHP1 ChIP-seq analysis performed at a single time point of the IDC informs about the distribution of heterochromatin for CVGs expressed at any asexual blood stage. Since we found that the distribution of heterochromatin is generally conserved among different parasite lines that have a CVG in the same transcriptional state, it is reasonable to expect that it should be possible to predict the transcriptional state of a CVG across the full IDC from the heterochromatin distribution at a single stage (e.g., the ring stage available in peripheral blood samples). However, the influence of the state of neighboring CVGs needs to be taken into consideration because for some genes, this influence results in more than one possible heterochromatin distribution associated with the same transcriptional state of a gene. We envisage that algorithms using machine learning or other approaches can be developed to accurately predict in human infections the expression state of essentially all CVGs from H3K9me3 ChIP-seq analysis of circulating rings. While developing such classification algorithms is clearly beyond the aim of our study, the data set presented here could be used as a seed for this type of algorithm, which later on would be improved with further data sets as they become available. We are aware that chromatin extraction and ChIP-seq are technically more demanding than RNA extraction and microarray or RNA-seq analysis, but this would clearly be compensated for by avoiding the need for *ex vivo* culture and sampling at multiple times to capture the full IDC. Furthermore, new technical developments such as cleavage under targets and tagmentation (CUT&Tag) (96), which uses streamlined protocols and requires a small amount of sample, may facilitate heterochromatin profiling directly from human infections. In addition to H3K9me3, many other histone posttranslational modifications have been described in *P. falciparum* (54, 97–99), but the majority are a consequence, rather than a cause, of the transcriptional state of a gene, or they fluctuate as the parasite progresses along the IDC. In contrast, H3K9me3-based heterochromatin actually determines the state of CVGs, and its distribution is stably maintained across the full IDC over multiple generations, making it a truly epigenetic mark that carries heritable information (100). Therefore, profiling H3K9me3 (or PfHP1) may be sufficient to predict the transcriptional state of most CVGs. Additionally, profiling H3K9ac may also be informative, especially when a CVG is active in only a small fraction of the parasites.

In addition to the global analysis of heterochromatin changes at CVGs, we paid special attention to specific genes with important known functions in host-parasite interactions. We found that in some subclones, several *var* genes were simultaneously in an active chromatin state (and expressed at high levels) in the majority of parasites in the population. These results suggest that the mutually exclusive expression of *var* genes may not be strict, consistent with previous studies that reported the expression of multiple *var* genes in a single parasite after selection for binding to multiple receptors (74) or the expression of several *var* genes in some recently subcloned lines (75). As an exploratory analysis, we also investigated epigenomic traits that may correlate with sexual conversion rates, a phenotype for which the molecular basis of variation has not been established. This resulted in the identification of a heterochromatin domain at the *gdv1* locus as a candidate regulator of sexual conversion rates. Heterochromatin was present at the putative promoter of the *gdv1-as* long noncoding RNA in parasite lines with high sexual conversion rates or lacking functional GDV1, but not in parasite lines with low sexual conversion rates and an intact *gdv1* gene. This result suggests that heterochromatin at this locus may prevent the expression of *gdv1-as*, which may enable the expression of GDV1 (69) and the subsequent activation of the master regulator of sexual conversion, *pfap2-g*. However, further studies will be needed to test this hypothesis, including analyses under defined conditions that induce or repress sexual conversion (9, 101) and manipulation of the locus using transfection approaches.

MATERIALS AND METHODS

Parasites. The 3D7-A and 3D7-B stocks of the *P. falciparum* clonal line 3D7; the 3D7-A subclones 10G and 1.2B; the 3D7-B subclones A7, E5, and B11; and the transgenic E5-PfAP2-G-HA line (9A subclone) were previously described and characterized (18, 26, 34, 61, 62). The 3D7-Imp line (66, 102) is the stock of 3D7 (103) maintained at Imperial College. The NF54 line (103) was from the stock maintained at Sanaria (104), and the NF54-V63 line was obtained after infecting a volunteer with Sanaria cryopreserved NF54 sporozoites in a CHMI trial (58, 105). Cultures for all of the new experiments presented in this study (and for the previous transcriptomic analysis of 1.2B, 10G, and 3D7-B [18]) were maintained at 3% hematocrit with B^+ RBCs in standard parasite culture medium with Albumax II and no human serum in an atmosphere containing 5% CO_2 and 3% O_2, balanced with N_2. However, the NF54-V63 line (for which ChIP-seq data generated as part of a previous study are included in some analyses) was cultured in medium containing human serum (58). Sexual conversion rates were determined using *N*-acetyl-D-glucosamine treatment and light microscopy, as previously described (65).

Transcriptomic experiments. Samples for transcriptomic analysis of the A7, E5, and B11 lines were obtained from tightly synchronized cultures (5-h age window) obtained by Percoll purification of RBCs infected with mature stages followed by 5% sorbitol lysis 5 h later. RNA was extracted from samples collected at 10 to 15, 20 to 25, 30 to 35, 37 to 42, 40 to 45, and 43 to 48 h postinvasion (hpi) using the TRIzol method, as previously described (18, 106). After reverse transcription and labeling, Cy5-labeled samples were hybridized against a Cy3-labeled reference pool, as previously described (68). The reference pool consisted of a mixture of equal amounts of cDNA from rings, trophozoites, and schizonts from 3D7-A and 3D7-B, similar to the reference pool used for the analysis of 1.2B, 10G, and 3D7-B but from a different biological preparation (18). Samples were hybridized on previously described custom Agilent microarrays (AMADID no. 085763) (35) modified from the AMADID no. 037237 design (68).

ChIP-seq experiments. Samples for ChIP-seq analysis were obtained from sorbitol-synchronized cultures at the late trophozoite/early schizont stage. Since the heterochromatin distribution is stable during the IDC, we did not use cultures tightly synchronized to a 5-h age window for these experiments. Chromatin extraction and immunoprecipitation were performed essentially as previously described (35, 58). In brief, after saponin lysis and formaldehyde cross-linking, chromatin was extracted using the MAGnify chromatin immunoprecipitation system (Life Technologies). Chromatin was sonicated using a Covaris M220 sonicator to obtain fragments of ~150 to 200 bp. For immunoprecipitation, we used 4 μg of chromatin and 8 μg of antibodies against H3K9me3 (catalog no. C15410193; Diagenode) or H3K9ac (catalog no. C15410004; Diagenode). For library preparation, we used 4 ng of input or immunoprecipitated DNA, the NEBNext multiplex oligonucleotides for Illumina (New England BioLabs [NEB]), 8 to 10 cycles of amplification with the Kapa HiFi PCR kit (Kapa Biosystems), and AMPure XP beads for purification steps (35, 107). Sequencing was performed using an Illumina HiSeq 2500 system, obtaining 8 million to 39 million 125-bp paired-end reads per sample.

Microarray data analysis. For the 1.2B, 10G, and 3D7-B lines, we used previously published processed expression values (18). For the A7, E5, and B11 subclones, microarray data were analyzed similarly, but normalized Cy5 and Cy3 values were obtained directly using Agilent Feature Extraction software. The rest of the analysis was performed using mainly the Bioconductor and the Tidyverse suite libraries in an R environment (v.3.6.3), essentially as previously described (58). In brief, for each channel, the background signal intensity was calculated as the median signal of the 100 lowest-signal probes in each array. Probes with a signal for both channels below three times the background were excluded from further analysis. For each probe, the expression value was defined as the $\log_2$(Cy5/Cy3) signal. Probes were collapsed into genes using median polish. For each array, the estimated parasite age in hours postinvasion was calculated using a previously described maximum likelihood method (67). For each gene and subclone, expression plots were generated with gene expression values on the *y* axis and estimated hours postinvasion on the *x* axis. These plots were used to calculate the average fold change (AFC) for each gene over four overlapping time intervals of half the estimated parasite age difference (in hours postinvasion) between the first and last analysis time points. The AFC was calculated for all possible pairwise comparisons among A7, E5, and B11 according to a previously described pipeline (18). The maximum AFC (mAFC) was defined as the AFC at the time interval in which it had the highest absolute value.

From this point, data from 1.2B, 10G, 3D7-B, A7, E5, and B11 were analyzed in parallel, although comparisons were performed separately for 1.2B, 10G, and 3D7-B or for A7, E5, and B11 (because these two groups of samples were analyzed with slightly different microarray platforms and reference pools). Genes with an mAFC of >4 (in at least one of the pairwise comparisons among 1.2B, 10G, and 3D7-B or among A7, E5, and B11) were considered to have markedly different transcript levels. To identify differentially expressed genes with high confidence, we applied several filters, essentially as previously described (58). For each pairwise comparison, only genes that had an mAFC of >4 and passed all of the filters were included in the final list of differentially expressed genes. First, tRNAs and genes lacking a PlasmoDB identification were excluded. Second, genes with very low absolute expression values were excluded. For this, genes with a Cy5 signal within the lowest 15% (at all time points) in the parasite lines with an mAFC of >4 in the pairwise comparison were filtered out. Third, genes that showed large expression differences only at the time periods in which they were expressed at low levels were excluded. For this, genes that at the time interval of maximum expression did not have an AFC of >4 were excluded. Finally, genes (as annotated in PlasmoDB v52, including the 5′ UTR and 3′ UTR when available) that overlapped a duplication or a deletion (see below) in at least one of the parasite lines among which the expression difference was observed were also excluded. To classify genes as previously reported CVGs, we used a previously described list of CVGs (58), whereas for the identification of gametocyte markers (genes with higher expression levels in sexual than in asexual parasites), we used a

list based on previously reported transcriptomic studies of gametocytes (see Data Set S7 in the supplemental material).

For some analyses of differentially expressed genes, we focused on the pairwise comparison with the largest fold difference in expression to identify representative subclones for the silenced and active states. For this, for each gene, we selected the pairwise comparison that passed all filters and had the largest mAFC among the comparisons of 1.2B versus 10G, 1.2B versus 3D7-B, 10G versus 3D7-B, A7 versus E5, A7 versus B11, and E5 versus B11. When the largest mAFC occurred between 3D7-B (which is not a transcriptionally homogeneous subclone and for which we did not generate ChIP-seq data) and 1.2B or 10G, we first determined which 3D7-B subclone (A7, E5, or B11) had the maximum (when the gene was active in 3D7-B and silenced in 10G or 1.2B) or minimum (when the gene was silenced in 3D7-B and active in 10G or 1.2B) expression value and selected this subclone as the representative subclone for the active or silenced state of the gene.

ChIP-seq data analysis. ChIP-seq data analysis, including the generation of plots, was performed using custom Python (v.3.6.9) and R (v.3.6.3) scripts, the latter mainly using Tidyverse suite packages. The initial processing of the data was performed approximately as described previously (35). After a quality check using FastQC (v.0.11.9), reads were trimmed, and repetitive *k*-mers were removed using BBDUK (v.36.99) with the parameters ktrim = r, k = 22, and mink = 6. Reads were aligned to the reference genome (*P. falciparum* 3D7 [PlasmoDB v28]) using Bowtie2 (v.2.3.0), and duplicate reads were excluded using the PICARD suite (v.4.1.9.0) with the RemoveDuplicates command. Peak calling was performed using MACS2 (v.2.2.7.1) with the parameters -f BAMPE -B -g 2.41e7 –keep-dup all –fe-cutoff 1.5 -nomodel –extsize 150. To calculate the normalized coverage of enrichment over the input (log_2 transformed) at a 10-bp resolution, we used the DeepTools (v.3.5.0) BamCompare command with the parameters –normalizeUsing RPKM -bs 10 –smoothLength 200 –pseudocount 10. After setting negative coverage values to zero to prevent the detection of coverage differences in nonheterochromatic regions, the coverage for landmark regions (e.g., upstream, coding, or downstream sequences of genes) was calculated as the mean value for the 10-bp bins within the region using BedTools (v.2.26.0). Custom scripts were used to create BED files demarcating the regions of interest for each gene for each particular analysis.

To identify differential H3K9me3 peaks in all possible pairwise comparisons among subclones 10G, 1.2B, E5, A7, and B11, we used custom scripts. First, the input-normalized coverage was calculated for 100-bp bins in all regions overlapping a MACS2 peak in at least one of the parasite lines of the pairwise comparison. The distribution of coverage intensities was adjusted to a skewed normal distribution, and for each bin, we calculated a cumulative density function (CDF) value for each parasite line. Bins with a CDF difference of >0.3 were considered to have differential coverage. Next, bins with differential coverage separated by <500 bp were merged, and differential peaks of <1,000 bp after merging were filtered out.

Identification of genetic changes in the parasite lines used in this study. To identify genetic polymorphisms, we analyzed the sequences from the input samples of the ChIP-seq experiments. To identify single nucleotide polymorphisms (SNPs) and short indels, we used the variant-calling protocol in the GATK suite according to GATK best practices (Data Set S2). Mutations with a coverage of <20 reads were excluded. Since recently subcloned lines are genetically highly homogeneous, a very high prevalence is expected for real mutations; therefore, we retained only mutations (relative to the reference 3D7 genome in PlasmoDB v52) in which the frequency of the reference allele was ≤0.5. Mutations were annotated using the ENSEMBL variant effect predictor (VEP). To define regions with relatively large (>500-bp) duplications or deletions, we used custom scripts. We first calculated coverage values for the inputs of the ChIP-seq data using the DeepTools (v.3.5.0) BamCoverage command with the parameters –normalizeUsing RPKM -bs 10 –smoothLength 200, similar to the calculation of the coverage for input-normalized ChIP-seq values. For each sample, the mean coverage over the whole genome was calculated. Bins with a coverage value <0.05 times the mean coverage were marked as underrepresented, and bins with a coverage value >1.75 times the mean coverage were marked as overrepresented. Bins marked as underrepresented or overrepresented that were separated by <200 bp were joined, and the resulting fragments were discarded if they were <500 bp. Finally, to avoid marking as deletion regions of low coverage due to the underlying sequence, we removed all regions with putative deletions that were common to all of the subclones analyzed. The remaining fragments were marked as deletions or duplications.

Classification of genes according to their transcriptional state. In order to assign a transcriptional state to each gene in subclones 1.2B, 10G, A7, E5, and B11, we designed a tree-like classification system. The final possible categories were active, inactive, undetermined, CVG active, CVG silenced, and CVG undetermined. The classification system is illustrated in Fig. S7, and the state of each gene is provided in Data Set S6.

The main parameters used for the classification of genes were: (i) previous classification as a CVG on the basis of variant expression or the presence of heterochromatin marks, according to a previously published list of CVGs (58); (ii) differential expression among the five parasite subclones in the data set presented here, with an mAFC of >2 at the time of maximum expression; (iii) Cy5 (sample channel) value percentiles in our data set; and (iv) expression level percentiles in published RNA-seq data sets (53, 107–109). In brief, genes were first classified as CVGs or non-CVGs. Next, regardless of whether a gene was a CVG or not, genes that showed no expression differences among the subclones were classified as active or inactive/silenced according to their maximum expression percentiles, whereas genes with expression differences were classified as active or inactive/silenced according to their relative expression levels. Genes with intermediate expression levels, duplications, or deletions were classified as undetermined or CVG undetermined.

Availability of data. The new microarray and ChIP-seq data described in this article have been deposited in the GEO database with accession no. GSE208561. ChIP-seq data are also available for visualization

at the UCSC genome browser (https://genome.ucsc.edu/s/lucas_michel_todo/3D7A%20and%203D7B%20subclones%20H3K9me3%20and%20H3K9ac and https://genome.ucsc.edu/s/lucas_michel_todo/NF54_3D7imp_E5HA). All of the scripts used for the data analysis are available at GitHub (https://github.com/LucasMichelTodo/Heterochomatin_Transitions.git).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.4 MB.

SUPPLEMENTAL FILE 2, XLSX file, 1.1 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.2 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.1 MB.

SUPPLEMENTAL FILE 5, XLSX file, 0.6 MB.

SUPPLEMENTAL FILE 6, XLSX file, 0.1 MB.

SUPPLEMENTAL FILE 7, XLSX file, 0.4 MB.

SUPPLEMENTAL FILE 8, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We are grateful to Mariona Bustamante (ISGlobal) for help in the acquisition of funding, support in setting up the project, and critical reading of the manuscript; S. Hoffman (Sanaria) for providing the NF54 line; and M. Delves (Imperial College) for providing the 3D7-Imp line.

This work was supported by grants from the Spanish Ministry of Science and Innovation (MCI)/Agencia Estatal de Investigación (AEI) (SAF2016-76190-R and PID2019-107232RBI00/AEI/10.13039/501100011033 to A.C.), cofunded by the European Regional Development Fund (ERDF) (European Union), the la Caixa Banking Foundation (HR18-00267 to A.C.), and the ISGlobal Alliance pilot project program (to C.B., J.R.G., and A.C.). L.M.-T. is supported by a fellowship from the Spanish Ministry of Economy and Competitiveness (BES-2017-081079), cofunded by the European Social Fund (ESF). Our research is part of ISGlobal's Program on the Molecular Mechanisms of Malaria, which is partially supported by the Fundación Ramón Areces. We acknowledge support from the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa 2019–2023 Program (CEX2018-000806-S) and support from the Generalitat de Catalunya through the CERCA Program.

A.C. and C.B. designed the experiments. C.B. optimized and performed ChIP-seq experiments, and N.C.-V. and N.R.-G. performed transcriptomic experiments. L.M.-T. performed the bioinformatic analysis of the data and generated the figures. C.H.-F. and J.R.G. performed preliminary analyses of the initial data. L.M.-T. and A.C. interpreted the results and wrote the manuscript, with input from all authors. A.C. conceived the project. C.B., J.R.G., and A.C. obtained funding.

We declare that we have no competing interests.

REFERENCES

1. Venugopal K, Hentzschel F, Valkiūnas G, Marti M. 2020. *Plasmodium* asexual growth and sexual development in the haematopoietic niche of the host. Nat Rev Microbiol 18:177–189. https://doi.org/10.1038/s41579-019-0306-2.
2. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1:E5. https://doi.org/10.1371/journal.pbio.0000005.
3. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K, Butungi H, Metcalf T, Verzier LH, Rayner JC, Berriman M, Herren JK, Billker O, Hemberg M, Talman AM, Lawniczak MKN. 2019. The Malaria Cell Atlas: single parasite transcriptomes across the complete *Plasmodium* life cycle. Science 365:eaaw2619. https://doi.org/10.1126/science.aaw2619.
4. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301:1503–1508. https://doi.org/10.1126/science.1087025.
5. Cowman AF, Healer J, Marapana D, Marsh K. 2016. Malaria: biology and disease. Cell 167:610–624. https://doi.org/10.1016/j.cell.2016.07.055.
6. Kumar M, Skillman K, Duraisingh MT. 2021. Linking nutrient sensing and gene expression in *Plasmodium falciparum* blood-stage parasites. Mol Microbiol 115:891–900. https://doi.org/10.1111/mmi.14652.
7. Zuzarte-Luis V, Mota MM. 2018. Parasite sensing of host nutrients and environmental cues. Cell Host Microbe 23:749–758. https://doi.org/10.1016/j.chom.2018.05.018.
8. Oakley MS, Gerald N, McCutchan TF, Aravind L, Kumar S. 2011. Clinical and molecular aspects of malaria fever. Trends Parasitol 27:442–449. https://doi.org/10.1016/j.pt.2011.06.004.
9. Brancucci NMB, Gerdt JP, Wang C, De Niz M, Philip N, Adapa SR, Zhang M, Hitz E, Niederwieser I, Boltryk SD, Laffitte MC, Clark MA, Gruring C, Ravel D, Blancke Soares A, Demas A, Bopp S, Rubio-Ruiz B, Conejo-Garcia A, Wirth DF, Gendaszewska-Darmach E, Duraisingh MT, Adams JH, Voss TS, Waters AP, Jiang RHY, Clardy J, Marti M. 2017. Lysophosphatidylcholine regulates sexual stage differentiation in the human malaria parasite *Plasmodium falciparum*. Cell 171:1532–1544.e15. https://doi.org/10.1016/j.cell.2017.10.020.
10. Mancio-Silva L, Slavic K, Grilo Ruivo MT, Grosso AR, Modrzynska KK, Vera IM, Sales-Dias J, Gomes AR, MacPherson CR, Crozet P, Adamo M, Baena-Gonzalez E, Tewari R, Llinas M, Billker O, Mota MM. 2017. Nutrient sensing modulates

malaria parasite virulence. Nature 547:213–216. https://doi.org/10.1038/nature23009.
11. Tintó-Font E, Michel-Todó L, Russell TJ, Casas-Vila N, Conway DJ, Bozdech Z, Llinás M, Cortés A. 2021. A heat-shock response regulated by the PfAP2-HS transcription factor protects human malaria parasites from febrile temperatures. Nat Microbiol 6:1163–1174. https://doi.org/10.1038/s41564-021-00940-w.
12. Tintó-Font E, Cortés A. 2022. Malaria parasites do respond to heat. Trends Parasitol 38:435–449. https://doi.org/10.1016/j.pt.2022.02.009.
13. Cortés A, Deitsch KW. 2017. Malaria epigenetics. Cold Spring Harb Perspect Med 7:a025528. https://doi.org/10.1101/cshperspect.a025528.
14. Duraisingh MT, Skillman KM. 2018. Epigenetic variation and regulation in malaria parasites. Annu Rev Microbiol 72:355–375. https://doi.org/10.1146/annurev-micro-090817-062722.
15. Guizetti J, Scherf A. 2013. Silence, activate, poise and switch! Mechanisms of antigenic variation in *Plasmodium falciparum*. Cell Microbiol 15:718–726. https://doi.org/10.1111/cmi.12115.
16. Llora-Batlle O, Tinto-Font E, Cortes A. 2019. Transcriptional variation in malaria parasites: why and how. Brief Funct Genomics 18:329–341. https://doi.org/10.1093/bfgp/elz009.
17. Voss TS, Bozdech Z, Bartfai R. 2014. Epigenetic memory takes center stage in the survival strategy of malaria parasites. Curr Opin Microbiol 20:88–95. https://doi.org/10.1016/j.mib.2014.05.007.
18. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, Ribas de Pouplana L, Preiser PR, Bozdech Z, Cortés A. 2012. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. Genome Res 22:925–938. https://doi.org/10.1101/gr.129692.111.
19. Scherf A, Lopez-Rubio JJ, Riviere L. 2008. Antigenic variation in *Plasmodium falciparum*. Annu Rev Microbiol 62:445–470. https://doi.org/10.1146/annurev.micro.61.080706.093134.
20. Smith JD, Rowe JA, Higgins MK, Lavstsen T. 2013. Malaria's deadly grip: cytoadhesion of *Plasmodium falciparum*-infected erythrocytes. Cell Microbiol 15:1976–1983. https://doi.org/10.1111/cmi.12183.
21. Wahlgren M, Goel S, Akhouri RR. 2017. Variant surface antigens of *Plasmodium falciparum* and their roles in severe malaria. Nat Rev Microbiol 15:479–491. https://doi.org/10.1038/nrmicro.2017.47.
22. Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, Speed TP, Cowman AF. 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol 7:R12. https://doi.org/10.1186/gb-2006-7-2-r12.
23. Templeton TJ. 2009. The varieties of gene amplification, diversification and hypervariability in the human malaria parasite, *Plasmodium falciparum*. Mol Biochem Parasitol 166:109–116. https://doi.org/10.1016/j.molbiopara.2009.04.003.
24. Wright GJ, Rayner JC. 2014. *Plasmodium falciparum* erythrocyte invasion: combining function with immune evasion. PLoS Pathog 10:e1003943. https://doi.org/10.1371/journal.ppat.1003943.
25. Bethke LL, Zilversmit M, Nielsen K, Daily J, Volkman SK, Ndiaye D, Lozovsky ER, Hartl DL, Wirth DF. 2006. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. Mol Biochem Parasitol 150:10–24. https://doi.org/10.1016/j.molbiopara.2006.06.004.
26. Cortés A, Carret C, Kaneko O, Yim Lim BYS, Ivens A, Holder AA. 2007. Epigenetic silencing of *Plasmodium falciparum* genes linked to erythrocyte invasion. PLoS Pathog 3:e107. https://doi.org/10.1371/journal.ppat.0030107.
27. Nguitragool W, Bokhari AA, Pillai AD, Rayavara K, Sharma P, Turpin B, Aravind L, Desai SA. 2011. Malaria parasite *clag3* genes determine channel-mediated nutrient uptake by infected red blood cells. Cell 145:665–677. https://doi.org/10.1016/j.cell.2011.05.002.
28. Pillai AD, Nguitragool W, Lyko B, Dolinta K, Butler MM, Nguyen ST, Peet NP, Bowlin TL, Desai SA. 2012. Solute restriction reveals an essential role for *clag3*-associated channels in malaria parasite nutrient acquisition. Mol Pharmacol 82:1104–1114. https://doi.org/10.1124/mol.112.081224.
29. Mira-Martínez S, Rovira-Graells N, Crowley VM, Altenhofen LM, Llinás M, Cortés A. 2013. Epigenetic switches in *clag3* genes mediate blasticidin S resistance in malaria parasites. Cell Microbiol 15:1913–1923. https://doi.org/10.1111/cmi.12162.
30. Lavazec C, Sanyal S, Templeton TJ. 2007. Expression switching in the *stevor* and *Pfmc-2TM* superfamilies in *Plasmodium falciparum*. Mol Microbiol 64:1621–1634. https://doi.org/10.1111/j.1365-2958.2007.05767.x.
31. Cortés A. 2008. Switching *Plasmodium falciparum* genes on and off for erythrocyte invasion. Trends Parasitol 24:517–524. https://doi.org/10.1016/j.pt.2008.08.005.
32. Duraisingh MT, Triglia T, Ralph SA, Rayner JC, Barnwell JW, McFadden GI, Cowman AF. 2003. Phenotypic variation of *Plasmodium falciparum* merozoite proteins directs receptor targeting for invasion of human erythrocytes. EMBO J 22:1047–1057. https://doi.org/10.1093/emboj/cdg096.
33. Petter M, Haeggstrom M, Khattab A, Fernandez V, Klinkert MQ, Wahlgren M. 2007. Variant proteins of the *Plasmodium falciparum* RIFIN family show distinct subcellular localization and developmental expression patterns. Mol Biochem Parasitol 156:51–61. https://doi.org/10.1016/j.molbiopara.2007.07.011.
34. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, Williams AE, Drought LG, Kwiatkowski DP, Baker DA, Cortes A, Llinas M. 2014. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507:248–252. https://doi.org/10.1038/nature12920.
35. Llorà-Batlle O, Michel-Todó L, Witmer K, Toda H, Fernández-Becerra C, Baum J, Cortés A. 2020. Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*. Sci Adv 6:eaaz5057. https://doi.org/10.1126/sciadv.aaz5057.
36. Poran A, Notzel C, Aly O, Mencia-Trinchant N, Harris CT, Guzman ML, Hassane DC, Elemento O, Kafsack BFC. 2017. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature 551:95–99. https://doi.org/10.1038/nature24280.
37. Josling GA, Williamson KC, Llinas M. 2018. Regulation of sexual commitment and gametocytogenesis in malaria parasites. Annu Rev Microbiol 72:501–519. https://doi.org/10.1146/annurev-micro-090817-062712.
38. Josling GA, Russell TJ, Venezia J, Orchard L, van Biljon R, Painter HJ, Llinas M. 2020. Dissecting the role of PfAP2-G in malaria gametocytogenesis. Nat Commun 11:1503. https://doi.org/10.1038/s41467-020-15026-0.
39. Brancucci NM, Bertschi NL, Zhu L, Niederwieser I, Chin WH, Wampfler R, Freymond C, Rottmann M, Felger I, Bozdech Z, Voss TS. 2014. Heterochromatin protein 1 secures survival and transmission of malaria parasites. Cell Host Microbe 16:165–176. https://doi.org/10.1016/j.chom.2014.07.004.
40. Flueck C, Bartfai R, Volz J, Niederwieser I, Salcedo-Amaya AM, Alako BT, Ehlgen F, Ralph SA, Cowman AF, Bozdech Z, Stunnenberg HG, Voss TS. 2009. *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog 5:e1000569. https://doi.org/10.1371/journal.ppat.1000569.
41. Perez-Toledo K, Rojas-Meza AP, Mancio-Silva L, Hernandez-Cuevas NA, Delgadillo DM, Vargas M, Martinez-Calvillo S, Scherf A, Hernandez-Rivas R. 2009. *Plasmodium falciparum* heterochromatin protein 1 binds to trimethylated histone 3 lysine 9 and is linked to mutually exclusive expression of *var* genes. Nucleic Acids Res 37:2596–2606. https://doi.org/10.1093/nar/gkp115.
42. Lopez-Rubio JJ, Mancio-Silva L, Scherf A. 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe 5:179–190. https://doi.org/10.1016/j.chom.2008.12.012.
43. Carrington E, Cooijmans RHM, Keller D, Toenhake CG, Bártfai R, Voss TS. 2021. The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the malaria parasite *Plasmodium falciparum*. iScience 24:102444. https://doi.org/10.1016/j.isci.2021.102444.
44. Shang X, Wang C, Fan Y, Guo G, Wang F, Zhao Y, Sheng F, Tang J, He X, Yu X, Zhang M, Zhu G, Yin S, Mu J, Culleton R, Cao J, Jiang M, Zhang Q. 2022. 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. https://doi.org/10.1093/nar/gkac176.
45. Crowley VM, Rovira-Graells N, de Pouplana LR, Cortés A. 2011. 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. https://doi.org/10.1111/j.1365-2958.2011.07574.x.
46. Jiang L, Lopez-Barragan MJ, Jiang H, Mu J, Gaur D, Zhao K, Felsenfeld G, Miller LH. 2010. Epigenetic control of the variable expression of a *Plasmodium falciparum* receptor protein for erythrocyte invasion. Proc Natl Acad Sci U S A 107:2224–2229. https://doi.org/10.1073/pnas.0913396107.
47. Lopez-Rubio JJ, Gontijo AM, Nunes MC, Issar N, Hernandez Rivas R, Scherf A. 2007. 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. https://doi.org/10.1111/j.1365-2958.2007.06009.x.
48. Cabral FJ, Fotoran WL, Wunderlich G. 2012. Dynamic activation and repression of the *Plasmodium falciparum rif* gene family and their relation to

chromatin modification. PLoS One 7:e29881. https://doi.org/10.1371/journal.pone.0029881.
49. Allshire RC, Madhani HD. 2018. Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol 19:229–244. https://doi.org/10.1038/nrm.2017.119.
50. Nicetto D, Zaret KS. 2019. Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. Curr Opin Genet Dev 55:1–10. https://doi.org/10.1016/j.gde.2019.04.013.
51. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AM, Cohen AM, Janssen-Megens EM, van de Vegte-Bolmer M, Selzer RR, Iniguez AL, Green RD, Sauerwein RW, Jensen ON, Stunnenberg HG. 2009. Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. Proc Natl Acad Sci U S A 106:9655–9660. https://doi.org/10.1073/pnas.0902515106.
52. Fraschka SA, Filarsky M, Hoo R, Niederwieser I, Yam XY, Brancucci NMB, Mohring F, Mushunje AT, Huang X, Christensen PR, Nosten F, Bozdech Z, Russell B, Moon RW, Marti M, Preiser PR, Bartfai R, Voss TS. 2018. Comparative heterochromatin profiling reveals conserved and unique epigenome signatures linked to adaptation and development of malaria parasites. Cell Host Microbe 23:407–420.e8. https://doi.org/10.1016/j.chom.2018.01.008.
53. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger TW, Francoijs KJ, Stunnenberg HG. 2010. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. PLoS Pathog 6:e1001223. https://doi.org/10.1371/journal.ppat.1001223.
54. Karmodiya K, Pradhan SJ, Joshi B, Jangid R, Reddy PC, Galande S. 2015. A comprehensive epigenome map of *Plasmodium falciparum* reveals unique mechanisms of transcriptional regulation and identifies H3K36me2 as a global mark of gene suppression. Epigenetics Chromatin 8:32. https://doi.org/10.1186/s13072-015-0029-1.
55. Zanghì G, Vembar SS, Baumgarten S, Ding S, Guizetti J, Bryant JM, Mattei D, Jensen ATR, Rénia L, Goh YS, Sauerwein R, Hermsen CC, Franetich JF, Bordessoulles M, Silvie O, Soulard V, Scatton O, Chen P, Mecheri S, Mazier D, Scherf A. 2018. A specific PfEMP1 is expressed in *P. falciparum* sporozoites and plays a role in hepatocyte infection. Cell Rep 22:2951–2963. https://doi.org/10.1016/j.celrep.2018.02.075.
56. Gomez-Diaz E, Yerbanga RS, Lefevre T, Cohuet A, Rowley MJ, Ouedraogo JB, Corces VG. 2017. Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. Sci Rep 7:40655. https://doi.org/10.1038/srep40655.
57. Bunnik EM, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Cort A, Shi L, Andolina C, Ross LS, Brady D, Fidock DA, Nosten F, Tewari R, Sinnis P, Ay F, Vert J-P, Noble WS, Le Roch KG. 2018. Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. Nat Commun 9:1910. https://doi.org/10.1038/s41467-018-04295-5.
58. Pickford AK, Michel-Todó L, Dupuy F, Mayor A, Alonso PL, Lavazec C, Cortés A. 2021. Expression patterns of *Plasmodium falciparum* clonally variant genes at the onset of a blood infection in malaria-naive humans. mBio 12:e01636-21. https://doi.org/10.1128/mBio.01636-21.
59. Bachmann A, Petter M, Krumkamp R, Esen M, Held J, Scholz JA, Li T, Sim BK, Hoffman SL, Kremsner PG, Mordmuller B, Duffy MF, Tannich E. 2016. Mosquito passage dramatically changes *var* gene expression in controlled human *Plasmodium falciparum* infections. PLoS Pathog 12:e1005538. https://doi.org/10.1371/journal.ppat.1005538.
60. Mira-Martínez S, van Schuppen E, Amambua-Ngwa A, Bottieau E, Affara M, Van Esbroeck M, Vlieghe E, Guetens P, Rovira-Graells N, Gómez-Perez GP, Alonso PL, D'Alessandro U, Rosanas-Urgell A, Cortés A. 2017. Expression of the *Plasmodium falciparum* clonally variant *clag3* genes in human infections. J Infect Dis 215:938–945. https://doi.org/10.1093/infdis/jix053.
61. Cortés A. 2005. A chimeric *Plasmodium falciparum Pfnbp2b/Pfnbp2a* gene originated during asexual growth. Int J Parasitol 35:125–130. https://doi.org/10.1016/j.ijpara.2004.11.004.
62. Cortés A, Benet A, Cooke BM, Barnwell JW, Reeder JC. 2004. Ability of *Plasmodium falciparum* to invade Southeast Asian ovalocytes varies between parasite lines. Blood 104:2961–2966. https://doi.org/10.1182/blood-2004-06-2136.
63. Cortés A, Mellombo M, Mgone CS, Beck HP, Reeder JC, Cooke BM. 2005. Adhesion of Plasmodium falciparum-infected red blood cells to CD36 under flow is enhanced by the cerebral malaria-protective trait South-East Asian ovalocytosis. Mol Biochem Parasitol 142:252–257. https://doi.org/10.1016/j.molbiopara.2005.03.016.
64. Rovira-Graells N, Aguilera-Simon S, Tinto-Font E, Cortes A. 2016. New assays to characterise growth-related phenotypes of *Plasmodium falciparum* reveal variation in density-dependent growth inhibition between parasite lines. PLoS One 11:e0165358. https://doi.org/10.1371/journal.pone.0165358.
65. Bancells C, Llora-Batlle O, Poran A, Notzel C, Rovira-Graells N, Elemento O, Kafsack BFC, Cortes A. 2019. Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. Nat Microbiol 4:144–154. https://doi.org/10.1038/s41564-018-0291-7.
66. Delves MJ, Straschil U, Ruecker A, Miguel-Blanco C, Marques S, Dufour AC, Baum J, Sinden RE. 2016. Routine in vitro culture of *P. falciparum* gametocytes to evaluate novel transmission-blocking interventions. Nat Protoc 11:1668–1680. https://doi.org/10.1038/nprot.2016.096.
67. Lemieux JE, Gomez-Escobar N, Feller A, Carret C, Amambua-Ngwa A, Pinches R, Day F, Kyes SA, Conway DJ, Holmes CC, Newbold CI. 2009. 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. https://doi.org/10.1073/pnas.0811829106.
68. Painter HJ, Altenhofen LM, Kafsack BF, Llinas M. 2013. Whole-genome analysis of *Plasmodium* spp. utilizing a new Agilent Technologies DNA microarray platform. Methods Mol Biol 923:213–219. https://doi.org/10.1007/978-1-62703-026-7_14.
69. Filarsky M, Fraschka SA, Niederwieser I, Brancucci NMB, Carrington E, Carrio E, Moes S, Jenoe P, Bartfai R, Voss TS. 2018. GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. Science 359:1259–1263. https://doi.org/10.1126/science.aan6042.
70. Eksi S, Morahan BJ, Haile Y, Furuya T, Jiang H, Ali O, Xu H, Kiattibutr K, Suri A, Czesny B, Adeyemo A, Myers TG, Sattabongkot J, Su X, Williamson KC. 2012. *Plasmodium falciparum* gametocyte development 1 (*Pfgdv1*) and gametocytogenesis early gene identification and commitment to sexual development. PLoS Pathog 8:e1002964. https://doi.org/10.1371/journal.ppat.1002964.
71. Figueiredo LM, Pirrit LA, Scherf A. 2000. Genomic organisation and chromatin structure of Plasmodium falciparum chromosome ends. Mol Biochem Parasitol 106:169–174. https://doi.org/10.1016/s0166-6851(99)00199-1.
72. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511. https://doi.org/10.1038/nature01097.
73. Rovira-Graells N, Crowley VM, Bancells C, Mira-Martínez S, Ribas de Pouplana L, Cortés A. 2015. Deciphering the principles that govern mutually exclusive expression of *Plasmodium falciparum clag3* genes. Nucleic Acids Res 43:8243–8257. https://doi.org/10.1093/nar/gkv730.
74. Joergensen L, Bengtsson DC, Bengtsson A, Ronander E, Berger SS, Turner L, Dalgaard MB, Cham GK, Victor ME, Lavstsen T, Theander TG, Arnot DE, Jensen AT. 2010. Surface co-expression of two different PfEMP1 antigens on single *Plasmodium falciparum*-infected erythrocytes facilitates binding to ICAM1 and PECAM1. PLoS Pathog 6:e1001083. https://doi.org/10.1371/journal.ppat.1001083.
75. Merrick CJ, Jiang RH, Skillman KM, Samarakoon U, Moore RM, Dzikowski R, Ferdig MT, Duraisingh MT. 2015. Functional analysis of sirtuin genes in multiple *Plasmodium falciparum* strains. PLoS One 10:e0118865. https://doi.org/10.1371/journal.pone.0118865.
76. Stewart LB, Freville A, Voss TS, Baker DA, Awandare GA, Conway DJ. 2022. *Plasmodium falciparum* sexual commitment rate variation among clinical isolates and diverse laboratory-adapted lines. Microbiol Spectr https://doi.org/10.1128/spectrum.02234-22.
77. Boltryk SD, Passecker A, Alder A, Carrington E, van de Vegte-Bolmer M, van Gemert GJ, van der Starre A, Beck HP, Sauerwein RW, Kooij TWA, Brancucci NMB, Proellochs NI, Gilberger TW, Voss TS. 2021. CRISPR/Cas9-engineered inducible gametocyte producer lines as a valuable tool for *Plasmodium falciparum* malaria transmission research. Nat Commun 12:4806. https://doi.org/10.1038/s41467-021-24954-4.
78. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC. 2015. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genomics 16:454. https://doi.org/10.1186/s12864-015-1603-4.
79. Tibúrcio M, Hitz E, Niederwieser I, Kelly G, Davies H, Doerig C, Billker O, Voss TS, Treeck M. 2021. A 39-amino-acid C-terminal truncation of GDV1 disrupts sexual commitment in *Plasmodium falciparum*. mSphere 6:e01093-20. https://doi.org/10.1128/mSphere.01093-20.

80. Gaszner M, Felsenfeld G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7:703–713. https://doi.org/10.1038/nrg1925.
81. Phillips-Cremins JE, Corces VG. 2013. Chromatin insulators: linking genome organization to cellular function. Mol Cell 50:461–474. https://doi.org/10.1016/j.molcel.2013.04.018.
82. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, Ouedraogo J-B, Michon P, Mueller I, Siba P, Nzila A, Borrmann S, Kiara SM, Marsh K, Jiang H, Su X-Z, Amaratunga C, Fairhurst R, Socheat D, Nosten F, Imwong M, White NJ, Sanders M, Anastasi E, Alcock D, Drury E, Oyola S, Quail MA, Turner DJ, Ruano-Rubio V, Jyothi D, Amenga-Etego L, Hubbart C, Jeffreys A, Rowlands K, Sutherland C, Roper C, Mangano V, Modiano D, Tan JC, Ferdig MT, Amambua-Ngwa A, Conway DJ, Takala-Harrison S, Plowe CV, Rayner JC, et al. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature 487:375–379. https://doi.org/10.1038/nature11174.
83. MalariaGEN, Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, Amato R, Amenga-Etego L, Andagalu B, Anderson TJC, Andrianaranjaka V, Apinjoh T, Ariani C, Ashley EA, Auburn S, Awandare GA, Ba H, Baraka V, Barry AE, Bejon P, Bertin GI, Boni MF, Borrmann S, Bousema T, Branch O, Bull PC, Busby GBJ, Chookajorn T, Chotivanich K, Claessens A, Conway D, Craig A, D'Alessandro U, Dama S, Day NPJ, Denis B, Diakite M, Djimdé A, Dolecek C, Dondorp AM, Drakeley C, Drury E, Duffy P, Echeverry DF, Egwang TG, Erko B, Fairhurst RM, Faiz A, Fanello CA, Fukuda MM, Gamboa D, et al. 2021. An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. Wellcome Open Res 6:42. https://doi.org/10.12688/wellcomeopenres.16168.2.
84. Tonkin-Hill GQ, Trianty L, Noviyanti R, Nguyen HHT, Sebayang BF, Lampah DA, Marfurt J, Cobbold SA, Rambhatla JS, McConville MJ, Rogerson SJ, Brown GV, Day KP, Price RN, Anstey NM, Papenfuss AT, Duffy MF. 2018. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen-encoding *var* genes. PLoS Biol 16:e2004328. https://doi.org/10.1371/journal.pbio.2004328.
85. Prajapati SK, Ayanful-Torgby R, Pava Z, Barbeau MC, Acquah FK, Cudjoe E, Kakaney C, Amponsah JA, Obboh E, Ahmed AE, Abuaku BK, McCarthy JS, Amoah LE, Williamson KC. 2020. The transcriptome of circulating sexually committed *Plasmodium falciparum* ring stage parasites forecasts malaria transmission potential. Nat Commun 11:6159. https://doi.org/10.1038/s41467-020-19988-z.
86. Usui M, Prajapati SK, Ayanful-Torgby R, Acquah FK, Cudjoe E, Kakaney C, Amponsah JA, Obboh EK, Reddy DK, Barbeau MC, Simons LM, Czesny B, Raiciulescu S, Olsen C, Abuaku BK, Amoah LE, Williamson KC. 2019. *Plasmodium falciparum* sexual differentiation in malaria patients is associated with host factors and GDV1-dependent genes. Nat Commun 10:2140. https://doi.org/10.1038/s41467-019-10172-6.
87. Kaestli M, Cortes A, Lagog M, Ott M, Beck HP. 2004. Longitudinal assessment of *Plasmodium falciparum var* gene transcription in naturally infected asymptomatic children in Papua New Guinea. J Infect Dis 189:1942–1951. https://doi.org/10.1086/383250.
88. Lavstsen T, Turner L, Saguti F, Magistrado P, Rask TS, Jespersen JS, Wang CW, Berger SS, Baraka V, Marquard AM, Seguin-Orlando A, Willerslev E, Gilbert MT, Lusingu J, Theander TG. 2012. *Plasmodium falciparum* erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. Proc Natl Acad Sci U S A 109:E1791–E1800. https://doi.org/10.1073/pnas.1120455109.
89. Portugaliza HP, Natama HM, Guetens P, Rovira-Vallbona E, Somé AM, Millogo A, Ouédraogo DF, Valéa I, Sorgho H, Tinto H, van Hong N, Sitoe A, Varo R, Bassat Q, Cortés A, Rosanas-Urgell A. 2022. *Plasmodium falciparum* sexual conversion rates can be affected by artemisinin-based treatment in naturally infected malaria patients. EBioMedicine 83:104198. https://doi.org/10.1016/j.ebiom.2022.104198.
90. Mok S, Ashley EA, Ferreira PE, Zhu L, Lin Z, Yeo T, Chotivanich K, Imwong M, Pukrittayakamee S, Dhorda M, Nguon C, Lim P, Amaratunga C, Suon S, Hien TT, Htut Y, Faiz MA, Onyamboko MA, Mayxay M, Newton PN, Tripura R, Woodrow CJ, Miotto O, Kwiatkowski DP, Nosten F, Day NPJ, Preiser PR, White NJ, Dondorp AM, Fairhurst RM, Bozdech Z. 2015. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science 347:431–435. https://doi.org/10.1126/science.1260403.
91. Daily JP, Scanfeld D, Pochet N, Le Roch K, Plouffe D, Kamal M, Sarr O, Mboup S, Ndir O, Wypij D, Levasseur K, Thomas E, Tamayo P, Dong C, Zhou Y, Lander ES, Ndiaye D, Wirth D, Winzeler EA, Mesirov JP, Regev A. 2007. Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients. Nature 450:1091–1095. https://doi.org/10.1038/nature06311.
92. Tarr SJ, Diaz-Ingelmo O, Stewart LB, Hocking SE, Murray L, Duffy CW, Otto TD, Chappell L, Rayner JC, Awandare GA, Conway DJ. 2018. Schizont transcriptome variation among clinical isolates and laboratory-adapted clones of the malaria parasite *Plasmodium falciparum*. BMC Genomics 19:894. https://doi.org/10.1186/s12864-018-5257-x.
93. Mackinnon MJ, Li J, Mok S, Kortok MM, Marsh K, Preiser PR, Bozdech Z. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. PLoS Pathog 5:e1000644. https://doi.org/10.1371/journal.ppat.1000644.
94. Mensah-Brown HE, Amoako N, Abugri J, Stewart LB, Agongo G, Dickson EK, Ofori MF, Stoute JA, Conway DJ, Awandare GA. 2015. Analysis of erythrocyte invasion mechanisms of *Plasmodium falciparum* clinical isolates across 3 malaria-endemic areas in Ghana. J Infect Dis 212:1288–1297. https://doi.org/10.1093/infdis/jiv207.
95. Gomez-Escobar N, Amambua-Ngwa A, Walther M, Okebe J, Ebonyi A, Conway DJ. 2010. Erythrocyte invasion and merozoite ligand gene expression in severe and mild *Plasmodium falciparum* malaria. J Infect Dis 201:444–452. https://doi.org/10.1086/649902.
96. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun 10:1930. https://doi.org/10.1038/s41467-019-09982-5.
97. Gupta AP, Chin WH, Zhu L, Mok S, Luah YH, Lim EH, Bozdech Z. 2013. Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite *Plasmodium falciparum*. PLoS Pathog 9:e1003170. https://doi.org/10.1371/journal.ppat.1003170.
98. Connacher J, von Grüning H, Birkholtz L. 2022. Histone modification landscapes as a roadmap for malaria parasite development. Front Cell Dev Biol 10:848797. https://doi.org/10.3389/fcell.2022.848797.
99. Tang J, Chisholm SA, Yeoh LM, Gilson PR, Papenfuss AT, Day KP, Petter M, Duffy MF. 2020. Histone modifications associated with gene expression and genome accessibility are dynamically enriched at *Plasmodium falciparum* regulatory sequences. Epigenetics Chromatin 13:50. https://doi.org/10.1186/s13072-020-00365-5.
100. Cortés A, Crowley VM, Vaquero A, Voss TS. 2012. A view on the role of epigenetics in the biology of malaria parasites. PLoS Pathog 8:e1002943. https://doi.org/10.1371/journal.ppat.1002943.
101. Portugaliza HP, Miyazaki S, Geurten FJ, Pell C, Rosanas-Urgell A, Janse CJ, Cortés A. 2020. Artemisinin exposure at the ring or trophozoite stage impacts *Plasmodium falciparum* sexual conversion differently. Elife 9:e60058. https://doi.org/10.7554/eLife.60058.
102. Delves MJ, Ruecker A, Straschil U, Lelievre J, Marques S, Lopez-Barragan MJ, Herreros E, Sinden RE. 2013. Male and female *Plasmodium falciparum* mature gametocytes show different responses to antimalarial drugs. Antimicrob Agents Chemother 57:3268–3274. https://doi.org/10.1128/AAC.00325-13.
103. Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, Corcoran LM, Burkot TR, Carter R. 1987. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 236:1661–1666. https://doi.org/10.1126/science.3299700.
104. Roestenberg M, Bijker EM, Sim BKL, Billingsley PF, James ER, Bastiaens GJH, Teirlinck AC, Scholzen A, Teelen K, Arens T, van der Ven A, Gunasekera A, Chakravarty S, Velmurugan S, Hermsen CC, Sauerwein RW, Hoffman SL. 2013. Controlled human malaria infections by intradermal injection of cryopreserved *Plasmodium falciparum* sporozoites. Am J Trop Med Hyg 88:5–13. https://doi.org/10.4269/ajtmh.2012.12-0613.
105. Gomez-Perez GP, Legarda A, Munoz J, Sim BK, Ballester MR, Dobano C, Moncunill G, Campo JJ, Cistero P, Jimenez A, Barrios D, Mordmuller B, Pardos J, Navarro M, Zita CJ, Nhamuave CA, Garcia-Basteiro AL, Sanz A, Aldea M, Manoj A, Gunasekera A, Billingsley PF, Aponte JJ, James ER, Guinovart C, Antonijoan RM, Kremsner PG, Hoffman SL, Alonso PL. 2015. Controlled human malaria infection by intramuscular and direct venous inoculation of cryopreserved *Plasmodium falciparum* sporozoites in malaria-naive volunteers: effect of injection volume and dose on infectivity rates. Malar J 14:306. https://doi.org/10.1186/s12936-015-0817-x.
106. Casas-Vila N, Pickford AK, Portugaliza HP, Tintó-Font E, Cortés A. 2021. Transcriptional analysis of tightly synchronized *Plasmodium falciparum* intraerythrocytic stages by RT-qPCR. Methods Mol Biol 2369:165–185. https://doi.org/10.1007/978-1-0716-1681-9_10.

107. Kensche PR, Hoeijmakers WA, Toenhake CG, Bras M, Chappell L, Berriman M, Bartfai R. 2016. The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res 44:2110–2124. https://doi.org/10.1093/nar/gkv1214.
108. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Berriman M, Newbold C, Llinas M. 2010. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. Mol Microbiol 76:12–24. https://doi.org/10.1111/j.1365-2958.2009.07026.x.
109. Toenhake CG, Fraschka SA, Vijayabaskar MS, Westhead DR, van Heeringen SJ, Bartfai R. 2018. Chromatin accessibility-based characterization of the gene regulatory network underlying *Plasmodium falciparum* blood-stage development. Cell Host Microbe 23:557–569.e9. https://doi.org/10.1016/j.chom.2018.03.007.

SUPPLEMENTARY FIGURES

Patterns of heterochromatin transitions linked to changes in the expression of *Plasmodium falciparum* clonally variant genes

Lucas Michel-Todó, Cristina Bancells, Núria Casas-Vila, Núria Rovira-Graells, Carles Hernández-Ferrer, Juan Ramón González & Alfred Cortés

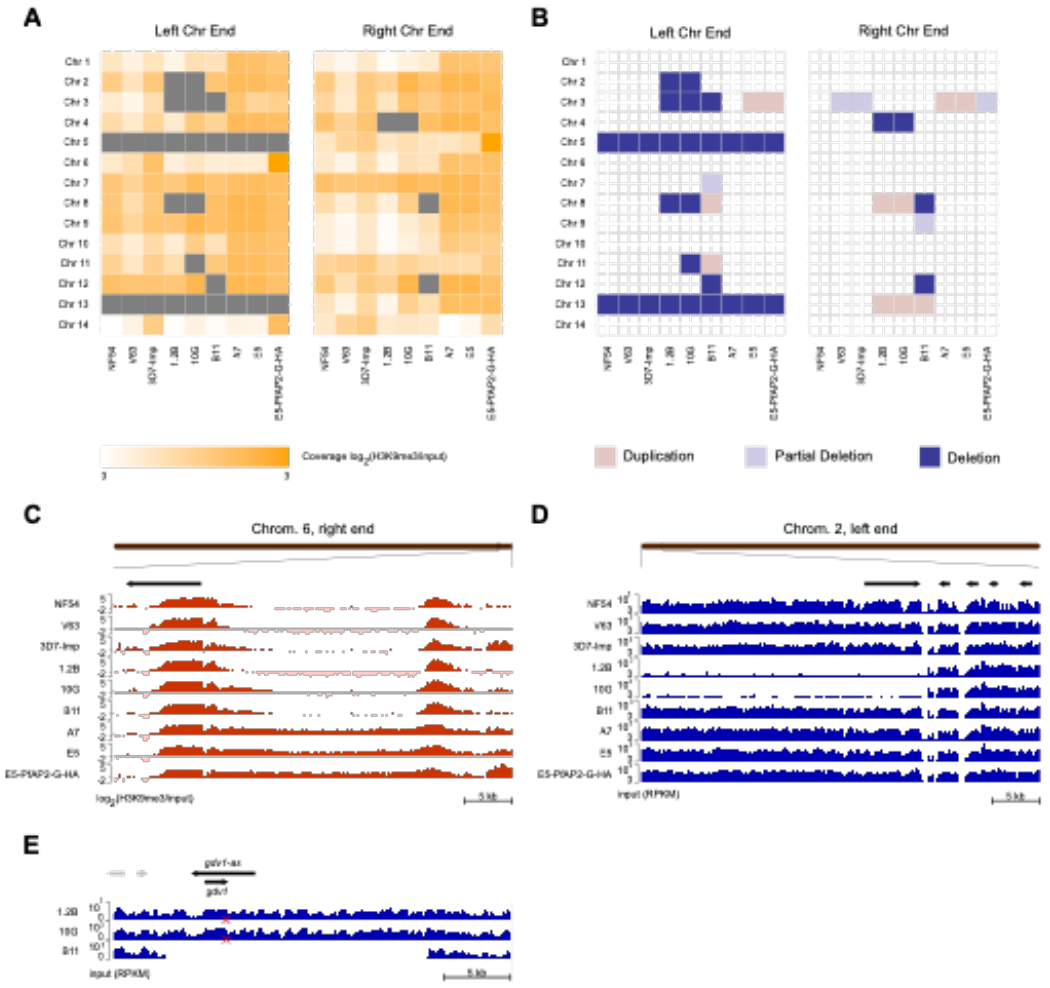


Fig. S1. Characterization of heterochromatin distribution, insertions and deletions at chromosome ends. (A) H3K9me3 coverage at chromosome ends, here defined as the subtelomeric non-coding regions from the most terminal gene to the telomere (right and left defined according to the orientation in PlasmoDB). Grey squares indicate presence of large deletions (>50% of the region analyzed) that preclude measuring heterochromatin levels. **(B)** Large deletions and duplications at chromosome ends (as defined in panel A). Chromosome ends with deletions that encompassed <30% of the region analyzed were marked as "partial deletion". **(C)** Representative example of differences between parasite lines in heterochromatin distribution at chromosome ends. **(D)** Representative example of deletions at chromosome ends. **(E)** Genomic alterations at the *gdv1* locus. A red cross indicates a SNP resulting in a premature STOP codon.

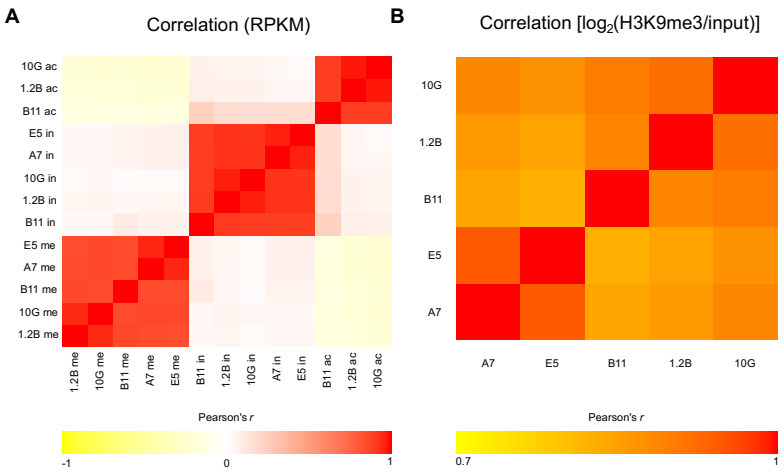


Fig. S2. Correlation between ChIP-Seq experiments. (A) Pearson correlation coefficient (*r*) between H3K9me3 (me), input (in) and H3K9ac (ac) ChIP-Seq signal (normalized coverage in RPKMs) for the five subclones of 3D7 genetic background analyzed. **(B)** Correlation between input-normalized H3K9me3 ChIP-Seq coverage for the five subclones.

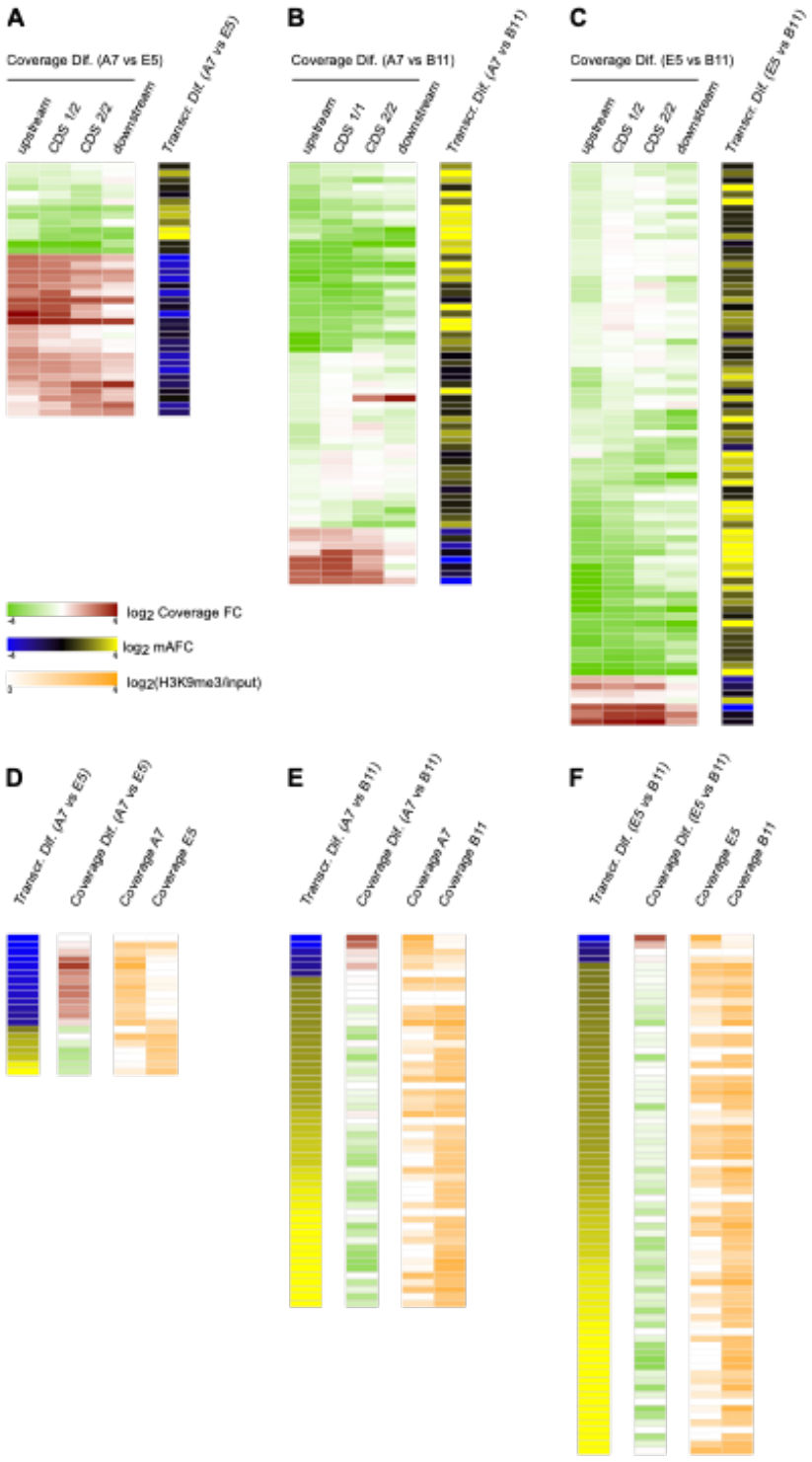

A
Coverage Dif. (A7 vs E5)
upstream
CDS 1/2
CDS 2/2
downstream
Transcr. Dif. (A7 vs E5)
B
Coverage Dif. (A7 vs B11)
upstream
CDS 1/1
CDS 2/2
downstream
Transcr. Dif. (A7 vs B11)
C
Coverage Dif. (E5 vs B11)
upstream
CDS 1/2
CDS 2/2
downstream
Transcr. Dif. (E5 vs B11)
log2 Coverage FC
log2 mAFC
log2(H3K9me3/input)
D
Transcr. Dif. (A7 vs E5)
Coverage Dif. (A7 vs E5)
Coverage A7
Coverage E5
E
Transcr. Dif. (A7 vs B11)
Coverage Dif. (A7 vs B11)
Coverage A7
Coverage B11
F
Transcr. Dif. (E5 vs B11)
Coverage Dif. (E5 vs B11)
Coverage E5
Coverage B11

Fig. S3. Association between H3K9me3 coverage differences and transcriptional differences. Results for subclones A7, E5 and B11 are shown (see main Fig. 3 for 10G and 1.2B). **(A-C)** Transcriptional differences in genes overlapping (-1,000 to +500 bp relative to the ATG) an H3K9me3 differential peak in the pairwise comparisons between A7 and E5 (A), A7 and B11 (B) or E5 and B11 (C). The $\log_2$ of the input-normalized H3K9me3 coverage fold-change (FC) is shown for the upstream region (1,000 bp before ATG), first or second half of the coding sequence (CDS1/2 or CDS2/2, respectively) and downstream region (1,000 bp after the STOP codon). Transcriptional differences are expressed as the $\log_2$ of the mAFC. **(D-F)** Heterochromatin levels in genes differentially expressed in each of the pairwise comparisons (mAFC >4). Transcriptional differences are shown as in panels A-C. H3K9me3 coverage ($\log_2$ of input-normalized H3K9me3 coverage) and $\log_2$ of coverage FC between A7 and E5 (D), A7 and B11 (E) or E5 and B11 (F) are shown for the region from -1,000 to +500 bp relative to the ATG. The values presented in this figure are provided in Data set S2.

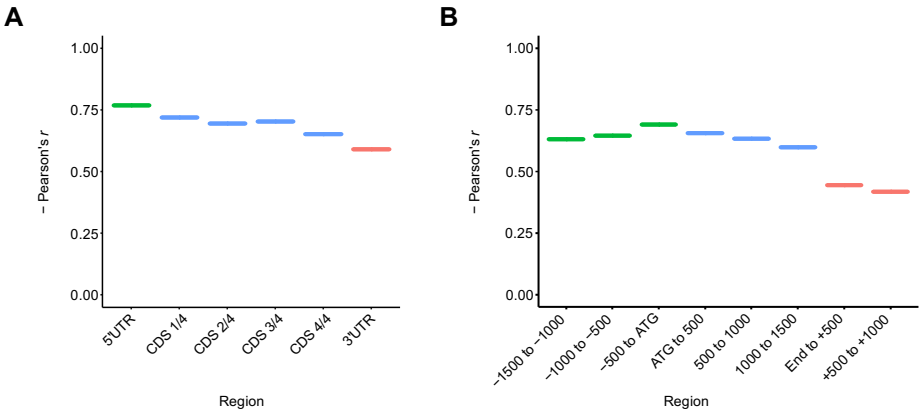


Fig. S4. Correlation between heterochromatin coverage and transcript levels at different positions of clonally variant loci. Correlation (Pearson's correlation coefficient, *r*) between input-normalized H3K9me3 coverage at different positions and transcript levels mAFC for genes with a mAFC >2 in any of the pairwise comparisons between 10G and 1.2B or among A7, E5 and B11. **(A)** H3K9me3 coverage was calculated for intervals of variable length corresponding to the 5'UTR (as annotated in PlasmoDB v52), coding sequence (CDS, divided in four different regions) and 3'UTR (as annotated in PlasmoDB v52). **(B)** H3K9me3 coverage was calculated for fixed length 500 bp regions relative to the start (ATG) or STOP (End) codon.

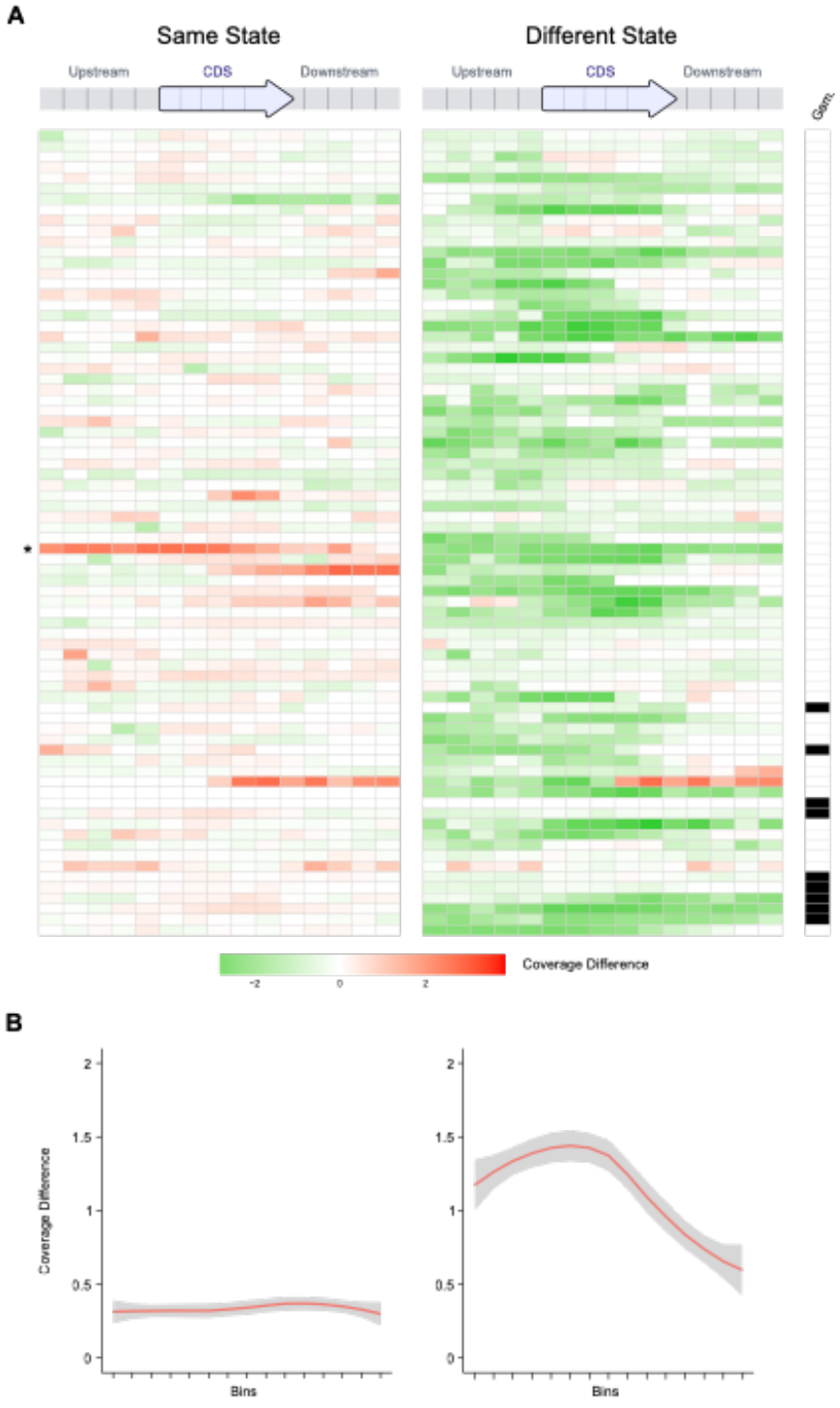
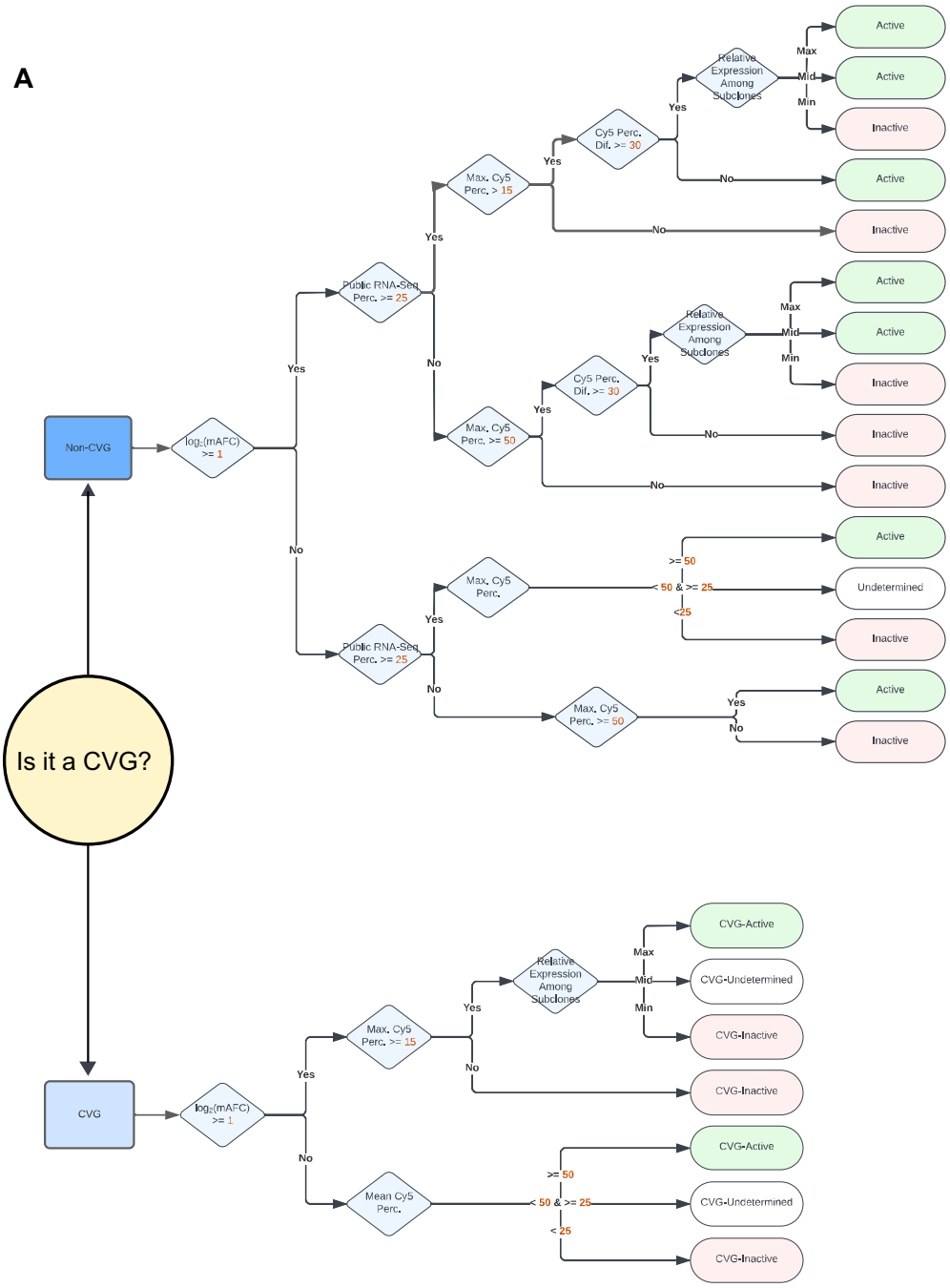

A
Same State
Different State
Upstream
CDS
Downstream
Upstream
CDS
Downstream
Gam
Coverage Difference
-2
0
2
B
Coverage Difference
2
1.5
1
0.5
0
Bins
2
1.5
1
0.5
0
Bins

Fig. S5. Heterochromatin differences at differentially expressed genes between pairs of subclones in which the gene is in the same transcriptional state in the two subclones or in a different state. (A) H3K9me3 coverage differences at genes differentially expressed among the 10G, 1.2B, A7, E5 and B11 subclones. Differentially expressed genes were selected as in main Fig. 4, but only genes that were in the same state (active or silenced) in at least two subclones and in the opposite state in at least one subclone, according to the classification in Data set S6, were included. The coverage difference is shown as in main Fig. 4 but using only the 15 bins corresponding to the upstream sequence, coding sequence (CDS) and downstream sequence. In the left panel, H3K9me3 coverage differences are shown for two randomly-selected subclones that had the gene in the same state (active-active or silenced-silenced). In the right panel, H3K9me3 coverage differences are shown for two randomly-selected subclones that had the gene in a different state (active-silenced, coverage difference in the active vs the silenced state). The gene marked with an asterisk (PF3D7_0800800), which showed large H3K9me3 coverage differences in the upstream and coding regions between two subclones in which the gene was classified as being in the same transcriptional state (silenced), likely corresponds to an incorrect automatic classification (borderline values). Visual inspection revealed that this gene was classified as “silenced CVG” in 1.2B because it did not pass the threshold for differential expression in the classification algorithm [$\log_2$(mAFC): 0.9; threshold: 1], but the expression time course analysis showed consistently higher expression in 1.2B than in 10G, suggesting that actually it is likely an “active CVG” in 1.2B. The values presented in this figure are provided in Data set S2. **(B)** Loess regression plots (shades are 95% confidence intervals) for the absolute value of the data in panel A showing the difference in H3K9me3 coverage between subclones in which a gene is in the same transcriptional state (left) or in a different transcriptional state (right).


Prev. Gene 1
Prev. Gene 2
Upstream
CDS
Downstream
Next Gene 1
Next Gene 2
Coverage Dif. (High. vs Low.)
Highest expression
Lowest Expression
log2(mAFC)
Gam.
ACS
CLAG
HYP
PHIST
STEVOR
VAR
RIFIN
PFMC-2TM
Other CVGs
Coverage FC (active vs silenced)
Coverage log2(H3K9me3/Input)

Fig. S6. Distribution of heterochromatin in different CVG families. Distribution of H3K9me3 coverage at all genes of the main CVG families, regardless of whether or not a gene is differentially expressed among the 10G, 1.2B, A7, E5 and B11 subclones. For each gene, the H3K9me3 coverage for the subclone with highest expression and the subclone with lowest expression (selected as in main Fig. 4) is shown, together with the difference between the two. H3K9me3 coverage is shown for 23 bins spanning the upstream sequence until the previous gene, coding sequence (CDS), downstream sequence until the next gene and the two upstream and downstream neighbor genes, as in Fig. 4. The transcriptional difference [$\log_2$(mAFC)] between the subclone with highest expression and the subclone with lowest expression (or 3D7-B instead of A7, E5 or B11 when the largest difference was between 10G or 1.2B and 3D7-B) is shown. Genes within each family are ordered by transcriptional difference. The column at the right indicates whether a gene is a known gametocyte marker. Information for the genes appearing in this figure is provided in Data set S2.


A
Is it a CVG?
Non-CVG
log2(mAFC) >= 1
Yes
No
Public RNA-Seq Perc. >= 25
Yes
No
Max. Cy5 Perc. > 15
Yes
No
Cy5 Perc. Dif. >= 30
Yes
No
Relative Expression Among Subclones
Max
Mid
Min
Active
Active
Inactive
Active
Inactive
Max. Cy5 Perc. >= 50
Yes
No
Cy5 Perc. Dif. >= 30
Yes
No
Relative Expression Among Subclones
Max
Mid
Min
Active
Active
Inactive
Inactive
Inactive
Public RNA-Seq Perc. >= 25
Yes
No
Max. Cy5 Perc.
>= 50
< 50 & >= 25
<25
Active
Undetermined
Inactive
Max. Cy5 Perc. >= 50
Yes
No
Active
Inactive
CVG
log2(mAFC) >= 1
Yes
No
Max. Cy5 Perc. >= 15
Yes
No
Relative Expression Among Subclones
Max
Mid
Min
CVG-Active
CVG-Undetermined
CVG-Inactive
CVG-Inactive
Mean Cy5 Perc.
>= 50
< 50 & >= 25
< 25
CVG-Active
CVG-Undetermined
CVG-Inactive

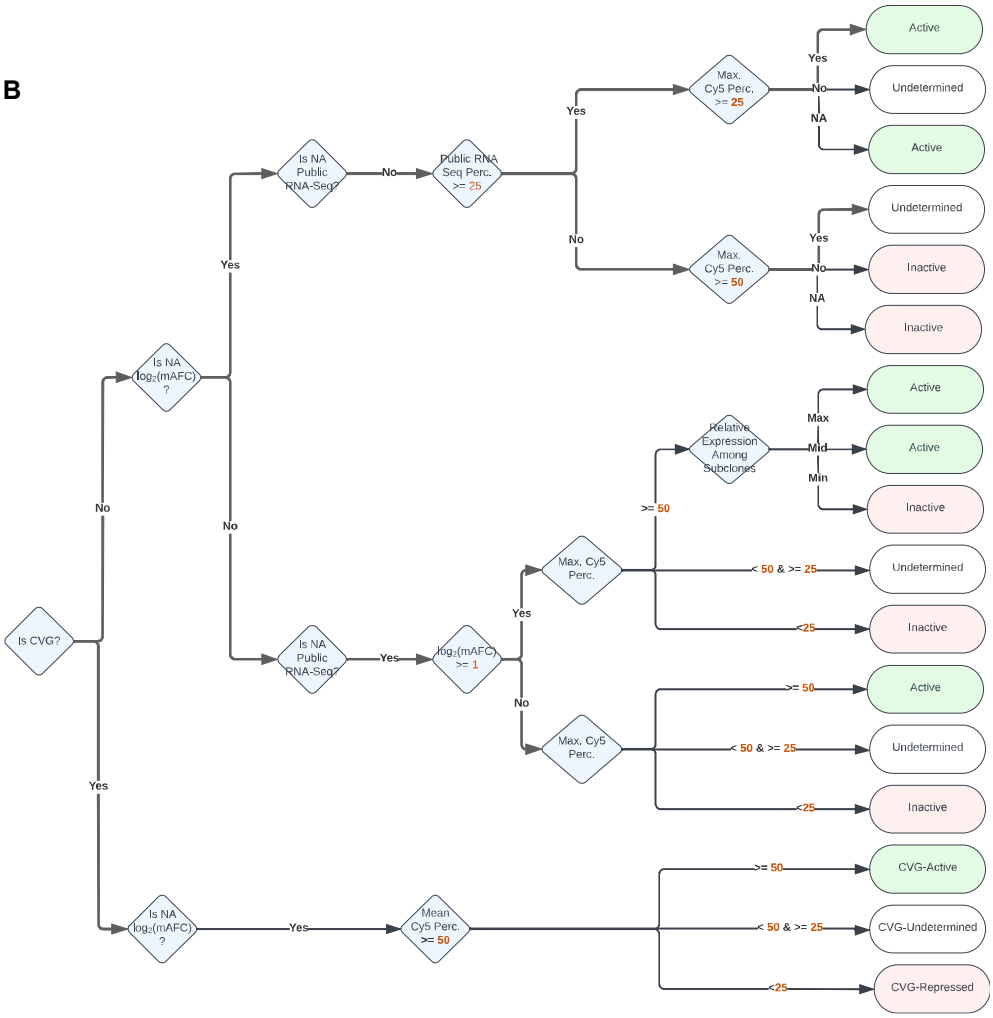


Fig. S7. Algorithm used to classify genes according to their transcriptional state in the subclones analyzed in this study. The algorithm was used to classify genes in the 1.2B, 10G, A7, E5 and B11 subclones. Separate decision trees are shown for genes with values for all classifying variables (A) and for genes with missing values (B). The state of each gene in each of the five subclone is shown in Data set S6. **(A)** Classification tree for genes without missing values. Genes were initially classified as CVGs or non-CVGs based on a previously published list of CVGs (Pickford et al., mBio 2021, PMID: 34340541). Then, 5 different classification criteria where used in the order indicated in the tree: (i) differential expression among the five subclones (log_2(mAFC) >1 in at least one pairwise comparison); (ii) expression level percentile ≥25 in at least one of four selected publicly available RNA-Seq datasets (Public RNA-Seq Perc.) (Bartfai et al. PLoS Pathog. 2010, PMID: 21187892; Otto et al., Mol. Microbiol. 2010, PMID: 20141604; Kensche et al.Nucleic Acids Res. 2016, PMID: 26578577; Toenhake et al., Cell Host Microbe 2018, PMID: 29649445); (iii) expression level in the microarray analysis presented here using the average of the two highest Cy5 percentile values (among all time points), either the maximum among the subclones (Max Cy5 Perc.) or the mean (Mean Cy5 Perc.), with different threshold values; (iv) the maximum difference in Cy5 percentile values between

any two subclones (Cy5 Perc. Dif.); (v) the relative expression between subclones. For this, the subclone with the highest expression value (using average normalized Cy5/Cy3 across the time interval with highest expression) was always classified as “max” and the subclone with lowest expression as “min”. For the remaining subclones, an upper threshold and a lower threshold where defined by dividing the distance between “max” and “min” in three equal intervals. The remaining subclones where classified as “max” if above the upper threshold, as “mid” if between the upper and lower thresholds and “min” if under the lower threshold. Based on these criteria and the decision tree, in each of the five subclones non-CVGs were classified as *Active*, *Inactive* or *Undetermined* and CVGs were classified as *CVG-active*, *CVG-silenced* or *CVG-Undetermined*. **(B)** Classification tree for genes with missing values. The classification criteria were the same as in panel A, but the decision tree structure was modified to account for the missing values.

2 **Article 2:** Expression Patterns of *Plasmodium falciparum* Clonally Variant Genes at the Onset of a Blood Infection in Malaria-Naive Humans

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Expression Patterns of *Plasmodium falciparum* Clonally Variant Genes at the Onset of a Blood Infection in Malaria-Naive Humans

Anastasia K. Pickford,[a] Lucas Michel-Todó,[a] Florian Dupuy,[b] Alfredo Mayor,[a,c,d] Pedro L. Alonso,[a,c*] Catherine Lavazec,[b] Alfred Cortés[a,e]

[a]ISGlobal, Hospital Clínic-Universitat de Barcelona, Barcelona, Catalonia, Spain
[b]INSERM U1016, Centre National de la Recherche Scientifique (CNRS) Unité Mixte de Recherche (UMR) 8104, Université de Paris, Institut Cochin, Paris, France
[c]Centro de Investigação em Saúde de Manhiça, Manhiça, Mozambique
[d]Consorcio de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[e]ICREA, Barcelona, Catalonia, Spain

ABSTRACT Clonally variant genes (CVGs) play fundamental roles in the adaptation of *Plasmodium falciparum* to fluctuating conditions of the human host. However, their expression patterns under the natural conditions of the blood circulation have been characterized in detail for only a few specific gene families. Here, we provide a detailed characterization of the complete *P. falciparum* transcriptome across the full intraerythrocytic development cycle (IDC) at the onset of a blood infection in malaria-naive human volunteers. We found that the vast majority of transcriptional differences between parasites obtained from the volunteers and the parental parasite line maintained in culture occurred in CVGs. In particular, we observed a major increase in the transcript levels of most genes of the *pfmc-2tm* and *gbp* families and of specific genes of other families, such as *phist*, *hyp10*, *rif*, or *stevor*, in addition to previously reported changes in *var* and *clag3* gene expression. Increased transcript levels of individual *pfmc-2tm*, *rif*, and *stevor* genes involved activation in small subsets of parasites. Large transcriptional differences correlated with changes in the distribution of heterochromatin, confirming their epigenetic nature. Furthermore, the similar expression of several CVGs between parasites collected at different time points along the blood infection suggests that the epigenetic memory for multiple CVG families is lost during transmission stages, resulting in a reset of their transcriptional state. Finally, the CVG expression patterns observed in a volunteer likely infected by a single sporozoite suggest that new epigenetic patterns are established during liver stages.

IMPORTANCE The ability of malaria parasites to adapt to changes in the human blood environment, where they produce long-term infection associated with clinical symptoms, is fundamental for their survival. CVGs, regulated at the epigenetic level, play a major role in this adaptive process, as changes in the expression of these genes result in alterations in the antigenic and functional properties of the parasites. However, how these genes are expressed under the natural conditions of the human circulation and how their expression is affected by passage through transmission stages are not well understood. Here, we provide a comprehensive characterization of the expression patterns of these genes at the onset of human blood infections, which reveals major differences with *in vitro*-cultured parasites. We also show that, during transmission stages, the previous expression patterns for many CVG families are lost, and new patterns are established.

KEYWORDS *Plasmodium falciparum*, chromatin, clonally variant genes, controlled human malaria infection, epigenetics, heterochromatin, malaria, transcription

Citation Pickford AK, Michel-Todó L, Dupuy F, Mayor A, Alonso PL, Lavazec C, Cortés A. 2021. Expression patterns of *Plasmodium falciparum* clonally variant genes at the onset of a blood infection in malaria-naive humans. mBio 12:e01636-21. https://doi.org/10.1128/mBio.01636-21.

Editor Thomas E. Wellems, National Institute of Allergy and Infectious Diseases

Address correspondence to Alfred Cortés, alfred.cortes@isglobal.org.

*Present address: Pedro L. Alonso, Global Malaria Program, World Health Organization, Geneva, Switzerland.

Received 8 June 2021
Accepted 2 July 2021
Published 3 August 2021

Plasmodium spp. are responsible for the globally important disease malaria, which causes over 200 million clinical cases and almost half a million deaths per year (1). The complex life cycle of malaria parasites is split between different hosts and cell types. The invertebrate mosquito host injects the infective parasite forms known as sporozoites into the vertebrate host during a blood meal. Sporozoites travel to the liver and multiply asexually within hepatocytes, generating merozoites that are then liberated into the bloodstream, where the intraerythrocytic development cycle (IDC) begins. This cycle consists of erythrocyte invasion followed by asexual multiplication, including the ring, trophozoite, and multinucleated schizont stages, and release of new merozoites. Repeated rounds of the IDC enable parasites to rapidly increase their biomass and establish an enduring infection in the vertebrate host. However, some parasites convert into sexual forms termed gametocytes and abandon the IDC. Mature male and female gametocytes are infective to mosquitoes. Mating occurs within the mosquito midgut, and after several complex development steps, new sporozoites are generated, closing the cycle (2).

Life cycle progression in *Plasmodium* spp. is controlled mainly at the transcriptional level such that each stage is characterized by a specific gene expression program (3–5). This enables parasites to thrive in the multiple different environments they are exposed to, which entail dramatic differences in conditions such as temperature, pH, and nutrient availability. However, in addition to environmental diversity associated with life cycle progression, malaria parasites also have to confront fluctuating conditions within the same niche. These fluctuations, which can occur between individual hosts of the same species or even within the course of a single blood infection, may derive, for instance, from changes in the host's physiological or immunological state or from the effects of antimalarial treatment (6, 7). Adaptation to such diverse conditions occurs through various genetic and nongenetic mechanisms. While in malaria parasites genetic changes play a major role in species evolution and long-term adaptation to new conditions, as in any other organism, rapid adaptation to conditions that fluctuate frequently requires reversible, dynamic mechanisms that provide phenotypic plasticity (i.e., alternative phenotypes from the same genome) (8). One such mechanism is clonally variant gene (CVG) expression, which refers to genes that can be found in a different state (active or silent) in different individual parasites with identical genomes and at the same stage of life cycle progression. While both states are heritable, CVGs undergo low-frequency stochastic switches between the active and silent states, which constantly generates transcriptional heterogeneity within parasite populations (9–11). Changes in the expression of these genes can result in phenotypic variation; therefore, when the conditions of the environment change, natural selection can operate upon this preexisting diversity and eliminate from the population parasites with CVG expression patterns that do not confer sufficient fitness under the new conditions. This is considered a bet-hedging adaptive strategy (12), a type of adaptive strategy commonly observed in many microbial species (13–15).

In *P. falciparum*, the most virulent human malaria parasite species, CVGs include gene families such as *var*, *rif*, *stevor*, *pfmc-2tm*, *hyp1* to *hyp17*, *phist*, and *surfin* linked to pathogenesis, antigenic variation, and host cell remodeling; *mspdbl2*, *eba140*, and *pfrh4* genes linked to erythrocyte invasion; *clag* genes involved in solute transport; *acs* and *acbp* families linked to acyl-CoA metabolism; and *pfap2-g*, the master regulator of sexual conversion (9–12), among others. Specific adaptive roles have been demonstrated for changes in the expression of *var* (16–18) and *clag3* (19–24) genes.

The active or silenced state of CVGs is regulated at the epigenetic level (25). These genes are mainly located in subtelomeric bistable chromatin domains, in which both the active (euchromatin) and the silenced (facultative heterochromatin) states can be stably transmitted for several generations of asexual growth. Spontaneous transitions between the two states underlie the transcriptional switches (8–11). For all CVGs analyzed so far, the silenced state is characterized by the posttranslational histone modification histone H3 lysine 9 trimethylation (H3K9me3) and heterochromatin protein 1 (HP1), whereas the

active state is associated with acetylation of H3K9 (H3K9ac). Transmission of these histone modifications through asexual replication constitutes the epigenetic memory for the transcriptional state of CVGs (8–11, 26–30). In *var* and *clag3* genes, the only *P. falciparum* genes that are known to show mutually exclusive expression (i.e., only one member of the family is active at a time in individual parasites) (31–35), this epigenetic memory is erased during transmission stages (here including gametocyte, mosquito, and liver stages) (23, 36–38).

The expression patterns of CVGs under culture conditions have been characterized for several specific gene families and also at a genome-wide scale (12, 32, 33, 39, 40). However, understanding how CVGs are expressed under the natural conditions of human infection is complicated by multiple factors, including common occurrence of polyclonal infections and genetic diversity among isolates, which mainly affects CVGs. The expression of only some specific CVG families such as *var*, *clag*, and genes involved in erythrocyte invasion has been characterized in some detail in natural infections (8, 16, 23, 41–43). This is an important gap of knowledge because the conditions of the environment influence the expression patterns of CVGs that prevail. Indeed, the limited data available suggest that the transcriptome of malaria parasites grown *in vitro* differs substantially from that observed *in vivo*, including differences in CVG expression (44–47).

Parasites obtained from controlled human malaria infection (CHMI) trials provide many of the advantages of both cultured parasites and parasites from natural human infections because they are exposed to "real" human host conditions, but they have a well-defined genetic background. Furthermore, the conditions of the host and the time of infection are well controlled, which reduces the number of variables and facilitates the interpretation of the results. Therefore, CHMI trials provide a valuable system to study malaria parasite biology, including *in vivo* expression of CVGs. So far, the analysis of gene expression in CHMI samples has focused mainly on the *var* and *clag* families (23, 36–38, 48–50). Two more recent studies analyzed the transcriptome of parasites obtained from CHMI volunteers at a genome-wide level. However, these studies only included ring-stage parasites, precluding the characterization of transcriptional patterns for genes expressed at other stages of the IDC (51, 52).

To provide a complete view of *P. falciparum* CVG expression patterns during the initial phase of a blood infection in malaria-naive humans, we performed a genome-wide transcriptomic comparison across the full IDC between parasites obtained from volunteers participating in a CHMI trial and the parental line maintained in culture. With this controlled approach, we identified transcriptional differences between parasites growing under *in vitro* culture conditions or growing in the human circulation after passage through transmission stages. To confirm the epigenetic nature of the differences observed, we mapped the genome-wide distribution of heterochromatin. We also tested the hypothesis that CVGs other than the mutually exclusively expressed *var* and *clag3* genes undergo an epigenetic reset during transmission stages.

RESULTS

Transcriptomic comparison between parasites obtained from CHMI volunteers and the parental line reveals changes in CVG expression. We performed a time course genome-wide transcriptomic analysis across the full IDC of *P. falciparum* parasites obtained from a CHMI trial in which cryopreserved Sanaria NF54 sporozoites were injected into naive human volunteers (53). Parasites were cryopreserved on day 9 after infection and upon microscopy diagnosis on days 11 to 14. Transcriptomic analysis was performed using parasites collected from four different volunteers (V18, V35, V48, and V63 lines, here collectively termed vNF54) on the day of diagnosis. Parasites were thawed and cultured for the minimum number of cycles needed to obtain sufficient material (4 replication cycles) and then tightly synchronized (involving an additional cycle of replication) before harvesting RNA at defined time points of the IDC (10 to 15, 20 to 25, 30 to 35, and 40 to 45 h postinvasion [hpi]). In parallel, we obtained RNA at the same time points from two independent biological replicates of tightly synchronized cultures of the parental (premosquito) NF54 line (pNF54) (Fig. 1A). Transcript levels were determined

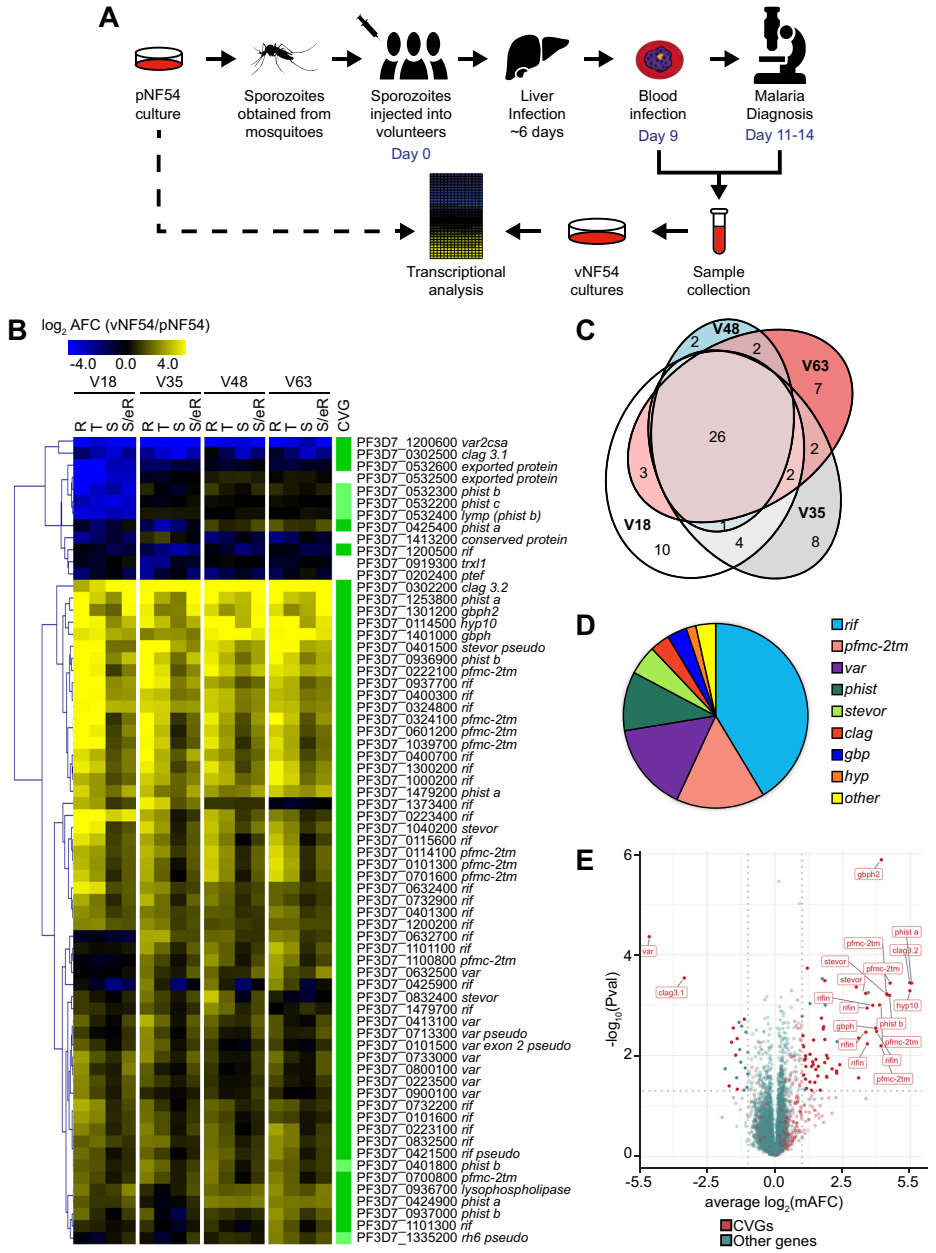


FIG 1 Genes differentially expressed between vNF54 and pNF54 lines. (A) Schematic of the study design comparing transcript levels between vNF54 lines obtained from volunteers participating in a controlled human malaria infection (CHMI) trial and the parental (premosquito) pNF54 line. (B) Heatmap of genes differentially expressed between vNF54 and pNF54 lines, ordered by hierarchical clustering. V18, V35, V48, and V63 are vNF54 lines obtained from different volunteers. Values are the $\log_2$ of the average fold change (AFC) of vNF54 versus pNF54 lines over four overlapping time intervals corresponding to the stages indicated (R, rings; T, trophozoites; S, schizonts; S/eR, schizonts and early rings). Only genes with an absolute value of the $\log_2$ of the maximum AFC (mAFC) of >2 (relative to pNF54) in at least one of the vNF54 lines are shown (see Materials and Methods for exclusion criteria). The

(Continued on next page)

using two-channel, long-oligonucleotide microarrays in which samples were hybridized against a common reference pool to obtain relative expression values (Cy5/Cy3). To quantify transcript-level differences between vNF54 and pNF54 lines, we calculated, for each gene, the maximum average fold change among overlapping time intervals of half the duration of the IDC (mAFC) (12) (see Materials and Methods).

There were 67 genes with a high-confidence mAFC of >4 (see Materials and Methods) in either direction [absolute value of the $\log_2$(mAFC) > 2] between at least 1 of the vNF54 lines and the pNF54 line, 21 of which had a mAFC of >16 (Fig. 1B and Data Set S1 in the supplemental material). The vast majority of differentially expressed genes had been previously identified as CVGs (Fig. 1B), either based on variant expression between isogenic lines (12) or because they carried the H3K9me3 or HP1 heterochromatin marks (54–57) (Data Set S2). Overall, the transcriptional changes (relative to pNF54) observed in the four vNF54 lines were highly similar (Fig. 1B), with the exception of a few genes showing a different pattern only in the V18 line. As a consequence, there was a large overlap in the genes differentially expressed (relative to pNF54) in each of the four vNF54 lines (Fig. 1C). Closer analysis revealed only a small number of genes with large transcript-level differences among the four vNF54 lines, and many of them showed a different pattern only in the V18 line, including a cluster of five downregulated neighbor genes in the distal subtelomeric region of chromosome 5 (Fig. S1).

The majority of genes differentially expressed in all vNF54 lines compared to pNF54 were expressed at higher levels in vNF54 lines, but a small number of genes were expressed at lower levels. The two most downregulated genes (mAFC > 10) were *var2csa* and *clag3.1*. The most upregulated genes in vNF54 lines (mAFC > 32) were a *hyp10* gene encoding an exported protein (PF3D7_0114500) (58), a *phist-a* gene (PF3D7_1253800), *clag3.2* (PF3D7_0302200), and a *pfmc-2tm* gene (PF3D7_0324100). Overall, the majority of genes that showed changes in expression between pNF54 and vNF54 lines belong to the large *pfmc-2tm*, *rif*, *var*, *stevor*, and *phist* CVG families encoding exported proteins (16–18, 58, 59) or to the smaller families *clag*, involved in solute transport (20), and glycophorin binding protein (*gbp*), encoding exported proteins of unknown function (58, 60, 61) (Fig. 1D). Given the overall similarity between the different vNF54 lines, we performed an additional analysis in which the different vNF54 lines were treated as replicates, which enables statistical analysis of the expression differences observed. The genes identified as most differentially expressed in this analysis were roughly the same as in the analysis of each vNF54 line separately, and *gpbh2* (PF3D7_1301200) was the most significantly upregulated gene in vNF54 lines (Fig. 1E and Data Set S1).

Transcriptional changes in specific CVG families. To analyze the expression patterns of CVGs within specific gene families, in addition to the transcript levels relative to a common reference pool (Cy5/Cy3 values), we also used the sample signal (Cy5 channel). The two-channel microarray approach used here is designed for hybridization against a common reference pool, which enables robust comparison of relative transcript levels between samples (3, 12, 62). However, the sample signal alone provides a semiquantitative estimate of the expression intensity of the genes, and it enables the identification of the predominantly expressed members of specific families (12), rather than informing only about the relative levels between samples.

Analysis of the transcriptional differences between pNF54 and vNF54 lines in specific CVG families revealed different scenarios for different families (Fig. 2A and B and

FIG 1 Legend (Continued)
column at the right indicates whether a gene was previously classified as a CVG (dark green; see Data Set S2) or belongs to a gene family in which other genes are CVGs (light green). Twelve genes had values out of the color scale range displayed. (C) Euler diagram showing the overlap between genes with an absolute value of the $\log_2$(mAFC) >2 (relative to pNF54) between the different volunteers. (D) Pie chart showing the distribution of genes with an absolute value of the $\log_2$(mAFC) of >2 between different gene families. (E) Volcano plot representing expression differences between vNF54 and pNF54 lines. Expression fold change values for each gene are the average of the $\log_2$(mAFC) in the four vNF54 lines. *P* values were calculated using an unpaired two-sided *t* test. CVGs are shown in red and other genes in green. Name labels are provided for genes with a mAFC of >10 and a *P* value of <0.01. Vertical dotted lines mark a $\log_2$(mAFC) of 1 or −1, whereas the horizontal dotted line marks a *P* value of 0.05.

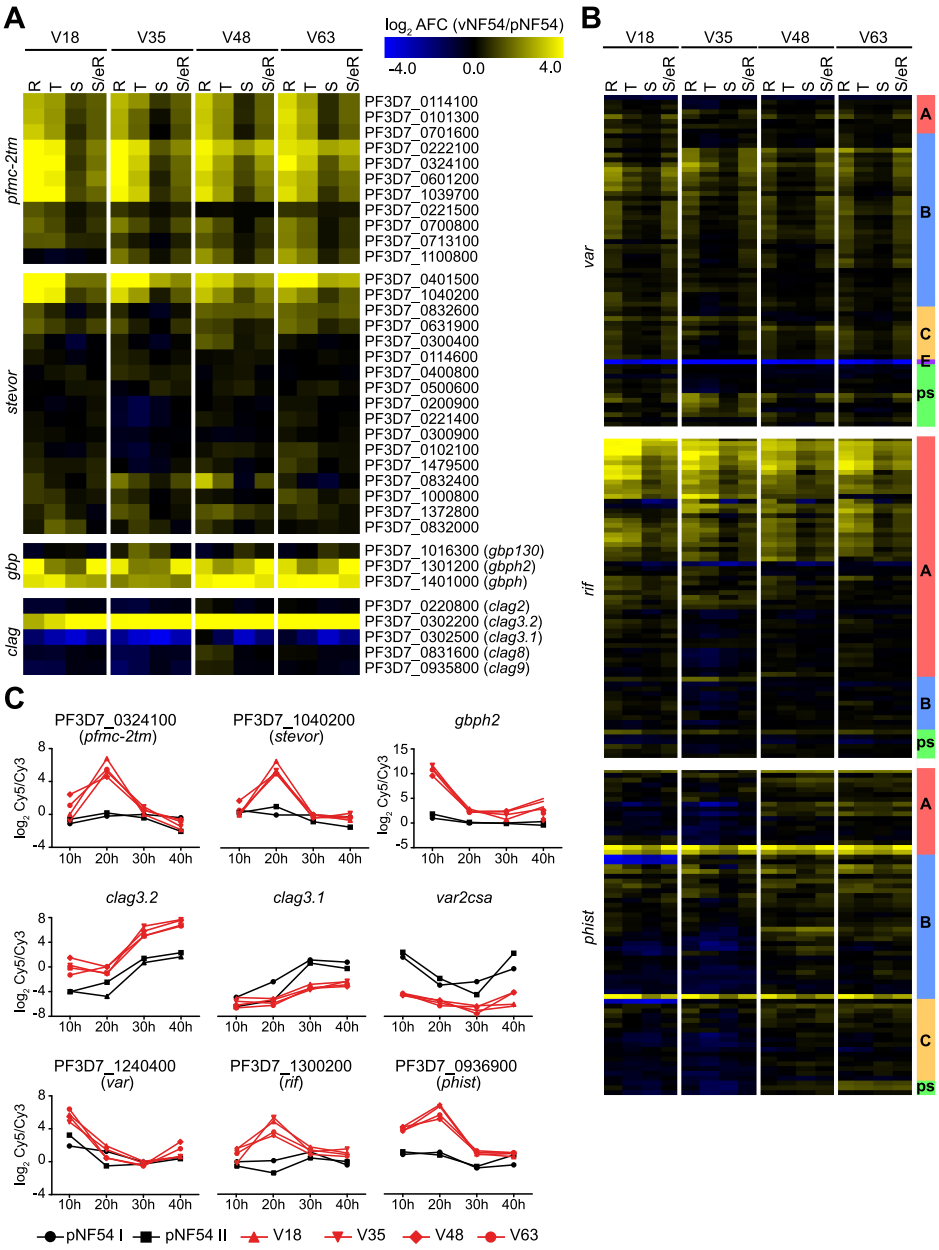


FIG 2 Transcriptional differences in specific CVG families. (A and B) Transcript-level changes in vNF54 lines relative to pNF54, as in Fig. 1B, for the CVG families indicated. See Materials and Methods for exclusion criteria. Eighteen genes (between the two panels) had values out of the color scale range displayed. In panel B, the subfamily of each gene (as annotated in PlasmoDB or in a previous study [12]) and pseudogenes (ps) are indicated in the column at the right. (C) Time course expression plots of selected genes of different CVG families. Values are the $\log_2$ of the sample versus reference pool ratio (Cy5/Cy3) in the vNF54 and pNF54 lines (*y* axis) and time after Percoll-sorbitol synchronization at which the sample was collected (*x* axis).

Fig. S2). Essentially all members of the *pfmc-2tm* family (39, 59, 63) were strongly upregulated in vNF54 lines (mAFC > 8 in all but three genes) (Fig. 2A and C), and this was confirmed by analyzing the Cy5 signal only, which revealed a large increase in the expression of the majority of *pfmc-2tm* genes (Fig. S2A). The analysis of a specific PfMC-2TM protein for which antibodies were available also revealed clearly increased abundance in a vNF54 line compared to pNF54 (Fig. S2B). In contrast, in the *var*, *rif*, *phist*, and *stevor* families, several specific genes were upregulated in vNF54 lines, but many other genes did not change, and, in some families, a few were downregulated. In the case of the mutually exclusively expressed *var* genes, vNF54 lines showed upregulation of many *var* genes, especially of type B, whereas the type E *var2csa* gene was strongly downregulated (Fig. 2B and C). Analysis of the Cy5 values revealed that *var2csa* was the predominantly expressed *var* gene in pNF54, whereas in vNF54 lines, multiple *var* genes were expressed at intermediate levels, similar to findings from previous CHMI studies (36, 38, 48, 52) (Fig. S2C). Considering that individual parasites express a single *var* gene, this result indicates that the parental population was relatively homogeneous, such that the vast majority of individual parasites expressed the same *var* gene (*var2csa*), whereas after passage through transmission stages the population became heterogeneous, with different individual parasites expressing different *var* genes. Some of the most highly expressed *var* genes in vNF54 lines coincided with the most highly expressed genes in a recent CHMI study using the 3D7 line to infect volunteers (52) (Fig. S3A), and preferential upregulation of type B *var* genes in parasites from the volunteers is also consistent with previous reports from other CHMI studies (36, 48, 52).

In the *rif* family (18, 64–67), a large subset of genes encoding type A RIFINS were upregulated in all vNF54 lines, whereas no general differences were observed in genes encoding type B RIFINS. In the *stevor* and *phist* families (18, 58, 68), only a small subset of specific genes showed strong upregulation in vNF54 lines, essentially the same genes in all four lines (Fig. 2A to C). In the *rif*, *stevor*, and *phist* families, which participate in antigenic variation and show expression switches but do not show mutually exclusive expression (39, 67, 69), the analysis of the Cy5 signal revealed that, in spite of changes in the expression of some specific genes, the predominantly expressed genes are similar between pNF54 and vNF54 lines (Fig. S2D to F). None of the genes that showed increased transcript levels in vNF54 lines became the dominantly expressed gene in any of these families. Of note, the most highly expressed *rif* gene in our data set was the same as in a previous CHMI study (52) (Fig. S3A).

Two small CVG families, *gbp* (60) and *clag* (20), showed major changes in the expression of a large proportion of their genes. Two of the three members of the *gbp* family, *gbph* and *gbph2*, were among the most highly upregulated genes (mAFC > 16) in vNF54 lines (Fig. 2A and C and Fig. S2G). In all vNF54 lines, there was increased expression of *clag3.2* and reduced expression of *clag3.1*, consistent with our previous reverse transcriptase quantitative PCR (RT-qPCR) results showing that, in the pNF54 population, essentially all parasites express *clag3.1*, whereas in vNF54 lines, there is a mixture of some parasites expressing *clag3.1* and a majority of parasites expressing *clag3.2* (23). There was no major change in the expression of the other CVG of the *clag* family, *clag2* (32) (Fig. 2A and C and Fig. S2H).

The expression of other CVG families such as acyl-CoA synthetase (*acs*), acyl-CoA binding protein (*acbp*), lysophospholipase, exported protein kinase (*fikk*), *hyp1* to *hyp 17*, *surfin*, and families linked to erythrocyte invasion (*eba*, *pfrh*) was almost identical between pNF54 and vNF54 lines, with the exception of a *hyp10* (PF3D7_0114500) and a lysophospholipase gene (PF3D7_0936700) (Fig. S4A).

Most changes in CVG expression between vNF54 and pNF54 lines are determined at the epigenetic level. The observation that the majority of genes differentially expressed between pNF54 and vNF54 lines are CVGs, which have been previously shown to be regulated by truly epigenetic mechanisms (25), strongly suggests that parasites recovered from the volunteers mainly differ from the parental line in their epigenetic makeup, rather than at the genetic level. To confirm this view, we sequenced the whole genome of the pNF54 and two vNF54 lines (V18 and V63) and

found no genetic differences in either coding or noncoding regions that are likely to explain any of the transcriptional changes common to all vNF54 lines (Data Set S3). However, we observed one duplicated region and several large subtelomeric deletions in V18 that explain the reduced expression of several genes specifically in this parasite line, including the cluster of neighbor downregulated genes in the distal subtelomeric region of chromosome 5 (Data Set S3). Unexpectedly, in the V18 line, there were essentially no reads mapping to the deleted regions, indicating that these large deletions occur with a 100% prevalence. Parasites in the parental pNF54 line (as in any other isolate maintained in culture for a long time) are expected to be a mixture of individual parasites that have accumulated different mutations (70). Therefore, the presence of deletions with a 100% prevalence in the V18 line suggests that this line is genetically homogeneous because it originated from a single sporozoite infecting the liver. Consistent with this view, the majority of differential single nucleotide polymorphisms (SNPs) and small indels identified among the pNF54, V63, and V18 lines were unique to V18 and had a 100% prevalence in this line (Data Set S3). Of note, the volunteer from which the V18 line originated was part of a group of the CHMI trial that was inoculated under a scheme that resulted in blood infection in only one out of six volunteers (group 2, 2,500 *P. falciparum* sporozites [PfSPZ] in 50 μl, intramuscular injection), indicating that very few sporozoites were viable. In contrast, the other parasite lines analyzed (V35, V48, and V63) are from volunteers that received direct venous inoculation resulting in infection of 100% of the volunteers (53). Altogether, these results indicate that the V18 line is genetically homogeneous for the mutations because it likely originated from infection by a single sporozoite.

Next, we performed comparative H3K9me3 chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis to determine the distribution of heterochromatin in one of the vNF54 lines (V63) and pNF54. In the genes showing the largest transcript-level differences between the two lines, higher expression was associated with reduced levels of heterochromatin in the upstream region and beginning of the coding sequence, whereas lower expression was associated with increased levels of heterochromatin (Fig. 3A and C and Data Set S4). This pattern was observed in 7 out of 8 genes with a mAFC of $>$16 between V63 and pNF54, and it occurred in genes from multiple families, including *stevor*, *gbp*, *clag*, *var*, *phist*, *hyp10*, and lysophospholipase. However, there were no apparent differences in heterochromatin prevalence between V63 and pNF54 at the putative regulatory regions of many differentially expressed genes from the large CVG families *pfmc-2tm*, *rif*, *var*, and *stevor* (Fig. 3B and C). The most plausible interpretation for this result is that individual genes of these families were activated in only small subsets of the parasites in V63, accounting for the increased transcript levels observed, but they remained silenced and in a heterochromatic state in the majority of parasites. Indeed, the analysis of a collection of subclones from the vNF54 line V33 (23), which also expressed *pfmc-2tm* genes at much higher levels than pNF54 (Fig. 4A), revealed that individual *pfmc-2tm* genes were silenced in the majority of subclones and expressed at very high levels in others (Fig. 4B). A similar scenario was observed for *rif* and *stevor* genes (Fig. S5). The expression patterns in recent subclones reflect the expression in individual parasites (12). Thus, these results indicate that individual *pfmc-2tm*, *rif*, and *stevor* genes were activated in only small subsets of parasites in V33 and also demonstrate population heterogeneity for the expression of these genes in a vNF54 line. For *var* genes, it has been previously established that in a population with a broad expression pattern, each individual gene is silenced in the majority of individual parasites (31, 36, 37). Together, these results indicate that, in the CVGs for which an association between increased transcript levels and lower levels of heterochromatin was not observed, this was explained by activation of the genes in only a small fraction of the parasites.

Changes in the expression of many CVGs are determined by a reset of epigenetic patterns during transmission stages. For *var* and *clag3* genes, an epigenetic reset during transmission stages results in transcriptional heterogeneity in the parasite population at the onset of a blood infection (23, 36–38, 48, 50). To determine if the epigenetic

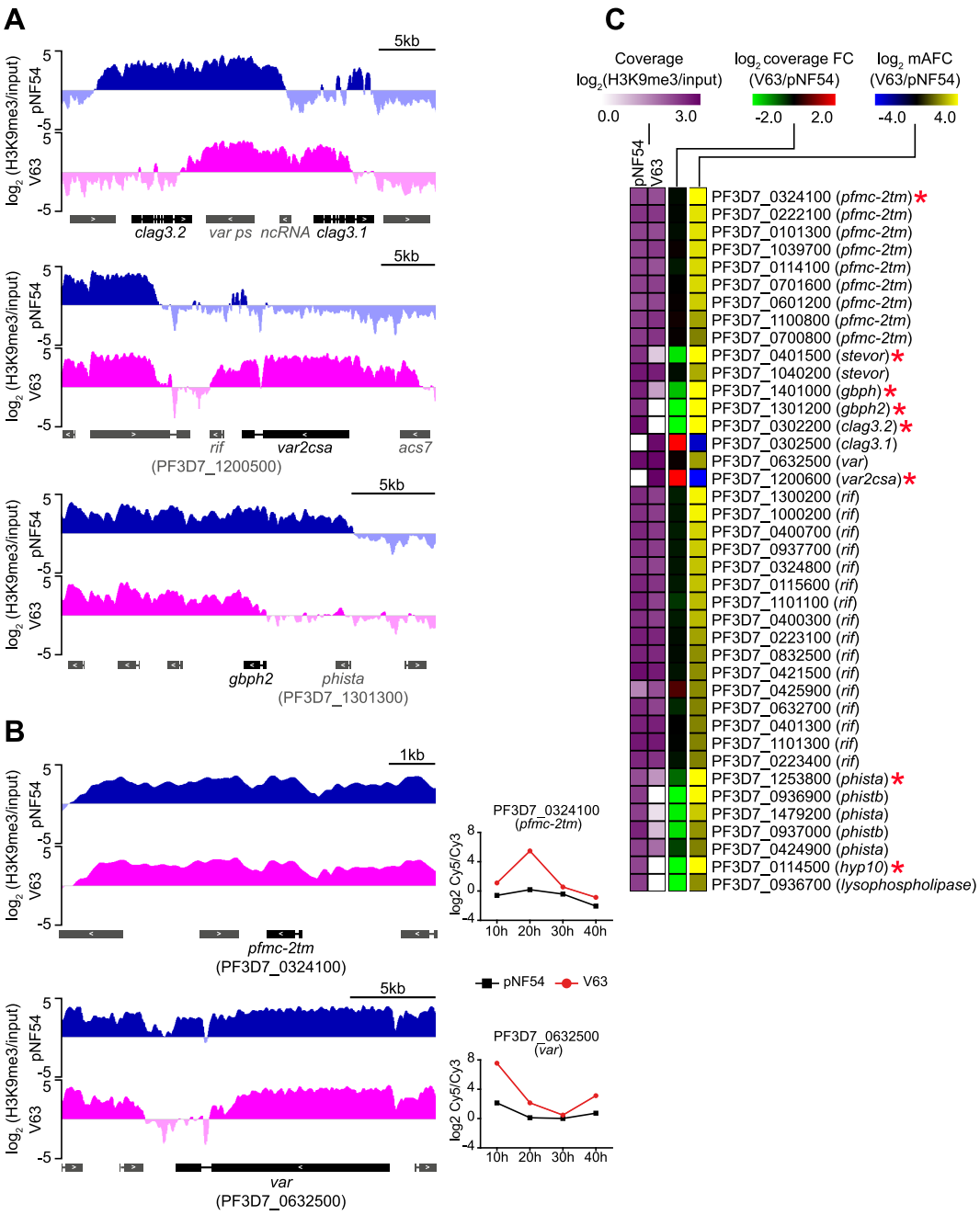


FIG 3 ChIP-seq analysis of vNF54 and pNF54 lines. (A and B) Distribution of normalized H3K9me3 signal relative to input in the pNF54 line and the vNF54 line V63. Representative genes are shown in which transcriptional changes between pNF54 and V63 are (*clag3.1*, *clag3.2*, *var2csa*, and *gbph2*) (A) or are not (*pfmc-2tm* and *var*) (B) accompanied by differences in heterochromatin distribution at their upstream regions. Genes are shown as rectangles (introns displayed as thin lines), with the direction of transcription indicated by arrowheads. The names of neighbor genes (in gray) with different levels

(Continued on next page)

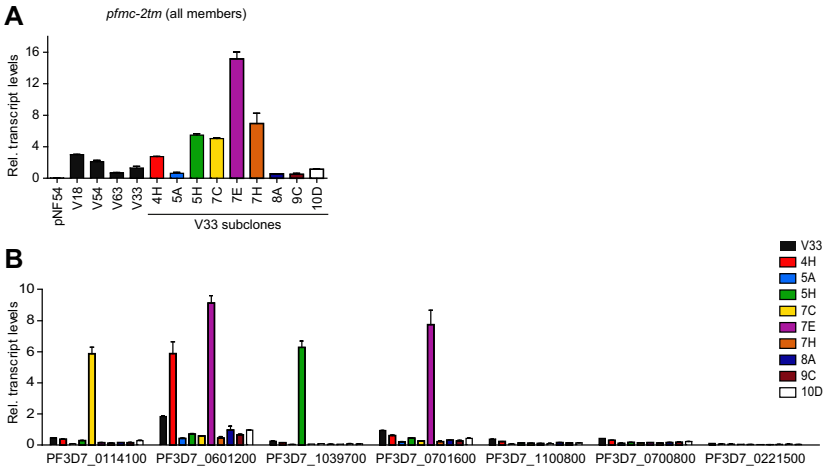


FIG 4 Expression of *pfmc-2tm* genes in subclones of a vNF54 line. (A) Transcript levels of *pfmc-2tm* genes (all members, analyzed with primers that amplify all the genes of the family) in pNF54, different vNF54 lines, and subclones of the V33 line at 20 to 25 h postinvasion (hpi). (B) Transcript levels of selected individual *pfmc-2tm* genes in V33 and V33 subclones. Transcript levels are normalized against serine-tRNA ligase (*serrs*). Data are presented as the average and SEM of two independent biological replicates.

memory for the transcriptional state of other CVGs is also lost during transmission stages, we compared the expression of *gbph2*, *pfmc-2tm*, and *stevor* genes between samples collected from the volunteers on day 9 or the day of microscopy diagnosis (for the samples used, day 11 or 14). Since liver stage development lasts ~6 days (36, 48, 49, 71), day 9 samples correspond to parasites that multiplied in the human circulation for only one round of the IDC after egress from the liver (second-generation blood stages), whereas day 11 or 14 samples are from parasites that replicated for 1 or 3 additional cycles, respectively (third- or fifth-generation blood stages), according to previous estimations (49). We reasoned that if only within-host selection during the IDC was responsible for the differences between pNF54 and vNF54 lines and there was no epigenetic reset during transmission stages, day 9 samples would show expression levels intermediate between pNF54 and day 11 or 14 samples, as parasites with unfavorable expression patterns would be eliminated progressively. In contrast, if an epigenetic reset is the main determinant of the expression patterns observed, similar transcript levels would be expected between day 9 and day 11 or 14 samples (23).

Because the parasitemia of day 9 samples was very low, obtaining sufficient material for RT-qPCR analysis required ~7 cycles of *in vitro* growth, whereas only ~4 cycles were needed for day 11 or 14 samples. Therefore, we first determined whether extended culture affected the expression levels of these genes. After 5 weeks in culture, transcript levels of *gbph2*, one *pfmc-2tm* gene, one *stevor* gene, and total *pfmc-2tm* family transcripts remained stable in both pNF54 and a vNF54 line (V63) (Fig. 5A). This result indicates that growth under *in vitro* conditions does not rapidly alter the expression of these genes and does not represent a confounding factor for the comparison of day 9 with day 11 or 14 samples. These experiments also confirmed the much higher transcript levels for these genes in vNF54 than in pNF54, and they revealed that they are virtually silenced in the latter.

FIG 3 Legend (Continued)
of H3K9me3 between pNF54 and V63 are shown. The time course expression plot for the genes in panel B, not included in Fig. 2C, is shown. These two genes had the largest expression fold increase in V63 relative to pNF54 in the entire *pfmc-2tm* and *var* families. (C) H3K9me3 ChIP-seq coverage (from −1,000 bp or closest upstream gene to +500 bp from the ATG), coverage fold change (FC) between V63 and pNF54, and mAFC between V63 and pNF54 for genes showing an absolute value of the $\log_2$(mAFC) of >2 between V63 and pNF54. Genes with an absolute value of the $\log_2$(mAFC) >4 (values out of the color scale range displayed) are marked with an asterisk.

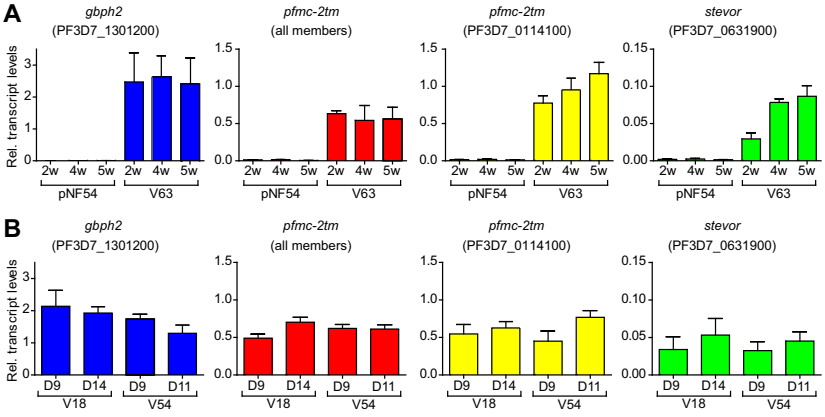


FIG 5 Changes in transcript levels associated with time in culture or time in the human circulation. (A) Effect of duration of growth under culture conditions on the expression of selected CVGs with different transcript levels between V63 (a vNF54 line) and pNF54. Relative transcript levels were determined by RT-qPCR after different times in culture, up to 5 weeks (w), in the pNF54 and V63 lines. RNA for transcriptional analysis was collected at the ring stage (10 to 15 h postinvasion [hpi]) for the *gbph2* gene and at the early trophozoite stage (20 to 25 hpi) for *pfmc-2tm* and *stevor* genes. For "*pfmc-2tm* (all members)", primers that amplify all genes of the family were used. (B) Comparison of relative transcript levels between parasites collected at day 9 postinfection or at days 14 (V18) or 11 (V54) (day of microscopy diagnosis), determined as in the previous panel. Transcript levels are normalized against serine-tRNA ligase (*serrs*). Data are presented as the average and SEM of three independent biological replicates.

Next, we compared the transcript levels of these genes between parasites collected on day 9 or the day of microscopy diagnosis in two vNF54 lines (V18 and V54, with microscopy diagnosis on days 14 and 11, respectively), which revealed no large differences (Fig. 5B). These results indicate that multiplication for a small number of rounds of the IDC in the blood of a naive human host does not have a major impact on the expression of *gbph2*, *pfmc-2tm*, and *stevor* genes. This is consistent with the idea that their increased expression in vNF54 lines compared to pNF54 was established before parasites reached the blood circulation, i.e., during transmission stages, as a consequence of an epigenetic reset. The results from a recent CHMI study in which blood-stage parasites from a volunteer infected via mosquito bite were blood passaged to new volunteers also support this idea (52, 72). Transcriptional comparison of parasites collected before and after blood growth in the new volunteers (three to five cycles), analogous to our day 9 versus day 11 or 14 comparison, revealed no difference in the expression of *var* or *rif* genes (52). We analyzed, in that data set, the expression of other CVGs and also observed no consistent changes (Fig. S3B). The majority of differences between samples collected before and after blood growth in the new volunteers occurred in genes that are not CVGs (Fig. S3C).

Phenotypic comparison of pNF54 and vNF54 lines. We performed exploratory experiments to determine if the transcriptional differences observed between pNF54 and vNF54 lines result in measurable functional differences. We focused on phenotypes that are mainly determined by CVGs and for which assays were available in our laboratories. Since many of the genes showing differential expression are exported to the erythrocyte cytoplasm or membrane (18, 58, 59) and some directly impact the mechanical properties of infected erythrocytes (i.e., *stevor*) (73), we compared the membrane deformability at the trophozoite stage of pNF54- and vNF54 (V63)-infected erythrocytes using a microsphiltration assay. While V63 showed a higher retention rate, indicative of lower deformability, the difference was not significant (Fig. 6A).

Since we observed changes in the expression of *clag3* genes, which determine solute uptake in infected erythrocytes (20), we compared sorbitol permeability between pNF54 and vNF54 lines. Sorbitol uptake resulting in red cell lysis was observed at

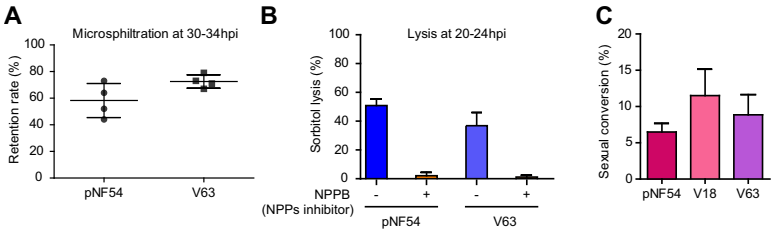


FIG 6 Phenotypic comparison of vNF54 and pNF54 lines. (A) Retention in microsphiltration assays of pNF54 and a vNF54 line (V63). Data are presented as the average, SEM, and individual data points of four independent biological replicates. (B) Sorbitol lysis assays for pNF54 and V63 in the absence or presence of NPPB (an inhibitor of new permeation pathways [NPPs]). Data are presented as the average and SEM of three independent biological replicates. (C) Sexual conversion rate (proportion of parasites that convert into sexual forms) of pNF54 and two vNF54 lines (V18 and V63). Data are presented as the average and SEM of three independent biological replicates. No significant difference ($P < 0.05$) was observed between pNF54 and vNF54 lines in any of the panels, using an unpaired two-sided *t* test.

similar levels in both lines (Fig. 6B) as expected, given that only the simultaneous silencing of the two *clag3* genes, which did not occur in either pNF54 or vNF54, is known to prevent sorbitol uptake (22). Lastly, we compared the sexual conversion rates of pNF54 and vNF54 lines (V18 and V63), which revealed no significant differences (Fig. 6C). This is consistent with the similar transcript levels of *pfap2-g*, the master regulator of sexual conversion (74), and of early gametocyte markers between pNF54 and vNF54 lines (Fig. S4B).

DISCUSSION

Here, we provide the first genome-wide transcriptional characterization across the full IDC of *P. falciparum* parasites obtained during the initial days of a blood infection in malaria-naive humans and compare it with the transcriptome of parasites with the same genome but maintained under *in vitro* culture conditions. We also compared the genome-wide distribution of heterochromatin between parasites obtained from the infected humans or in culture. We found that the largest expression differences occur in CVGs and are associated with epigenetic changes in the distribution of heterochromatin, and we provide an accurate view of how parasites use their CVGs when they establish a new blood infection after egress from the liver. We also show that passage through transmission stages results in a reset of the epigenetic memory involving multiple CVG families.

The majority of changes between premosquito parasites and parasites obtained from the infected volunteers occurred in gene families involved in processes such as antigenic variation, erythrocyte remodeling, or solute transport, whereas other CVG families showed few alterations. Even within the same family, subgroups of genes with different predicted functions showed different expression patterns: type A RIFINS, predicted to encode proteins localized at the infected erythrocyte surface and involved in rosetting (18, 67), were generally upregulated in parasites collected from the volunteers, but not type B RIFINS with different predicted function and localization. The most dramatic changes were observed in the *pfmc-2tm*, *gbp*, *var*, and *clag* families, whereas other families, such as *rif*, *stevor*, or *phist*, showed important transcript-level differences for several specific genes, but the predominantly expressed genes and global expression levels of the family were not altered. Together, these observations suggest that some CVG families have high plasticity in their expression patterns, whereas others have clearly "preferred" patterns.

While the epigenetic state of CVGs is transmitted from one generation of asexual parasites to the next (11, 17, 25), it has been proposed that an "epigenetic reset" may occur during transmission stages such that the epigenetic memory for the transcriptional state of all CVGs is erased and new patterns of expressed and silenced CVGs are established stochastically (75). In the murine malaria parasite *Plasmodium chabaudi*,

mosquito-transmitted parasites were less virulent than those transmitted by blood passage and expressed a broader repertoire of CVGs (75, 76). This led some authors to postulate that repeated cycles of the IDC progressively select for parasites with CVG expression patterns associated with increased fitness, and passage through transmission stages results in a reset of the epigenetic patterns and lower virulence (75), although evidence for this model in human malaria is lacking. In *P. falciparum*, an epigenetic reset during transmission stages was previously demonstrated for *var* and *clag3* genes (23, 36–38). Here, we observed clearly distinct expression patterns between premosquito parasites and parasites obtained from the volunteers for genes from multiple additional CVG families (*gbp*, *pfmc-2tm*, *stevor*, *rif*, *phist*, lysophospholipase, and *hyp10*) and similar expression patterns between day 9 and day 11 or 14 samples in all the CVG families analyzed (*gbp*, *pfmc-2tm*, and *stevor*). Similarly, previous analyses of parasites from CHMI trials showed no consistent CVG expression changes associated with an increasing number of multiplication cycles in the blood of naive humans (36, 52). These results are consistent with a general epigenetic reset of CVGs during transmission stages, although it is also possible that in CVGs for which no transcriptional changes were observed between pNF54 and vNF54 lines, the epigenetic memory is maintained throughout all transmission stages.

We postulate that during transmission stages, the epigenetic memory for multiple CVG families is erased, and later on, new epigenetic patterns are established stochastically such that they differ among individual parasites. However, while the state of a gene in an individual parasite is established stochastically, the probability of the gene acquiring an active or a silenced state may be hardwired and dictated by the underlying DNA sequence. This probabilistic scenario could explain the similar CVG expression patterns observed among the different volunteers and also the relatively constant patterns observed in some gene families. For some CVGs, the active or the silenced state may be strongly favored. Given that in the V18 line, which appears to have originated from a single sporozoite invading the liver, we observed CVG expression patterns similar to the other vNF54 lines, including a broad expression pattern of *var* genes indicative of transcriptional heterogeneity, we postulate that new epigenetic patterns are established during liver development. The stage at which the previous epigenetic patterns are erased is not known, but given the profound heterochromatin remodeling already observed in gametocytes and mosquito stages, it is likely to occur before parasites reach the liver (55, 77, 78).

Of note, in many genes from large multigene families, we observed large transcript-level fold differences that were not accompanied by measurable changes in the levels of H3K9me3 occupancy. We found that in the parasite populations obtained from the volunteers, individual genes of these families were active in only a small subset of the parasites, which still could result in a large fold increase in transcript levels if fewer parasites expressed it in the premosquito population (e.g., a 10-fold increase if a gene was active in 10% of the parasites collected from the volunteers but only in 1% of the premosquito population). This would result in a minor difference in the heterochromatin levels (e.g., 90% versus 99% of the parasites would have the gene from the example above in a heterochromatic state), which is not detectable in a comparative ChIP-seq analysis. Therefore, our results do not indicate that activation of many *var*, *rif*, or *pfmc-2tm* genes occurred independently of heterochromatin changes but, rather, reflect that activation occurred in only a small fraction of the parasites in the population, and reductions in heterochromatin levels were undetectable. Indeed, in the parental line, the *var2csa* gene was in an active state in essentially all the parasites, and this was clearly associated with the absence of heterochromatin at this locus, consistent with previous reports for the epigenetic regulation of *var* genes (28, 30).

Altogether, a model emerges in which a blood infection in a new host starts with a transcriptionally heterogeneous parasite population, consisting of a mixture of individual parasites with different combinations of active and silenced CVGs from multiple gene families. This constitutes the basis of a bet-hedging adaptive strategy because

transcriptional heterogeneity results in phenotypic and antigenic diversity that increases the chances of survival of the parasite population in the unpredictable conditions of the blood of a new human host. Our data provide an accurate picture of how parasites use their CVGs in this initial phase of a blood infection, revealing that a higher proportion of individual parasites (compared to culture conditions) have in an active state the majority of genes of the *pfmc-2tm* and *gbp* families and specific genes of the *rif* (type A), *phist*, *stevor*, and a few other CVG families. Additionally, we demonstrate transcriptional heterogeneity for the expression of *rif*, *stevor*, and *pfmc-2tm* genes in addition to *var* and *clag3* genes (23, 36, 38, 48, 52). While in our study we could not characterize the long-term expression patterns of CVGs because infections were terminated as soon as parasites were detected by light microscopy or symptoms appeared, current knowledge suggests that during chronic blood infection, the expression of CVGs is governed by low-frequency switches and dynamic natural selection of parasites with expression patterns that confer high fitness as the conditions of the host fluctuate (9–11, 50). Future research will be needed to fully characterize the precise phenotypic and antigenic differences that result from specific changes in CVG expression.

MATERIALS AND METHODS

Human samples and parasite culture. The *P. falciparum* NF54 line at Sanaria (pNF54) and lines derived from human volunteers participating in a CHMI study with Sanaria NF54 sporozoites (vNF54 lines) (53) were cultured in B+ erythrocytes at 3% hematocrit in RPMI 1640-based culture medium supplemented with 10% human serum under continuous shaking at 100 rpm and in a 5% CO_2, 3% O_2, and 92% N_2 atmosphere.

Transcriptomic analysis using microarrays. Time course transcriptomic analysis was performed using cultures tightly synchronized to a 5-h age window, which was achieved by Percoll purification of schizonts followed by sorbitol lysis 5 h later (79). Sixty-milliliter cultures at 3% hematocrit and ~10% parasitemia were split into four flasks that were cultured undisturbed for different periods of time (10, 20, 30, or 40 h) before harvesting in TRIzol and freezing at −80°C. Rather than splitting into identical cultures, larger volumes (3- to 4-fold) were used for the early time points because young (ring-stage) parasites contain a smaller amount of RNA per parasite. Cultures for RNA collection at late stages were diluted to ~5% parasitemia with fresh erythrocytes (just after splitting) to prevent culture collapse.

RNA was purified using the TRIzol method, and cDNA was synthesized by reverse transcription (starting with 5 to 10 μg of RNA), purified, and labeled as previously described (62). Samples were analyzed using two-color long oligonucleotide-based custom Agilent microarrays. The microarray design was based on Agilent design AMADID 037237 (62), modified by adding new probes for genes lacking unique probes and for some ncRNAs and reporter genes (new designs, AMADID 084561 and AMADID 085763) and reannotated according to a BLAST analysis (80). All samples (200 to 500 ng), labeled with Cy5, were hybridized against an equal amount of a common reference pool labeled with Cy3, consisting of a mixture of equal amounts of cDNA from rings, trophozoites, and schizonts of pNF54. Microarray hybridization was performed as previously described (62). Images were acquired using a microarray scanner (G2505C; Agilent Technologies) located in a low-ozone hood.

Microarray data analysis. Initial processing of microarray data, including linear and locally weighted scatterplot smoothing (LOWESS) normalization, was performed using the Feature Extraction software (Agilent) with default options. The next steps of the analysis were performed using Bioconductor in an R environment (R version 3.5.3). For each individual sample and channel (Cy3 and Cy5), background signal was calculated as the median of the 100 lowest signal probes. Probes with both Cy3 and Cy5 signals below three times the array background in all samples were excluded from further analysis. Gene-level $\log_2$(Cy5/Cy3) values and statistical estimation of parasite age (81) were computed as previously described (12). For each gene, the $\log_2$(Cy5/Cy3) values were plotted against the statistically estimated culture age (in hpi), and the plots were divided into four overlapping time intervals of identical length that roughly corresponded to the ring, trophozoite, schizont, and late-schizont/early-ring stages (12). For each gene and time interval, the average expression fold change (AFC) between each vNF54 and its control pNF54 line (the replicate of pNF54 analyzed in parallel) was calculated from the difference in the area under the curve in the $\log_2$(Cy5/Cy3) versus estimated age plots. The maximum AFC (mAFC) was the value of the AFC in the time interval at which it had the highest absolute value. All tRNAs were excluded from further analysis because our method is not suitable for the analysis of tRNA expression (it tends to show large technical variability). For the identification of genes differentially expressed among pNF54 and vNF54 lines (i.e., Fig. 1B to D, Fig. 2, Fig. S1 in the supplemental material, and Data Set S1), genes with expression intensity (Cy5 channel) values that were in all samples within the lowest 20th percentile were excluded because expression differences in genes expressed at near-background levels are of low confidence. Based on this criterion, we excluded 17 out of 103 genes with an absolute value of the $\log_2$ mAFC (vNF54/pNF54) of >2 (Fig. 1B to D), 131 out of 889 genes with a value of >1 (Data Set S1), and 4 out of 42 genes with differential expression among vNF54 lines (Fig. S1; see below). Additionally, we excluded transcripts that lack a PlasmoDB ID and genes with apparent artifacts according to visual inspection (i.e., genes with large expression differences observed only at time intervals that do not correspond to their peak expression) from the list of top differentially expressed genes presented

in Fig. 1B to D (6 and 13 genes, respectively) and Fig. S1 (7 and 4 genes, respectively). For the analysis presented in Fig. S1, we compared transcript levels directly between the different vNF54 lines. Since differences observed between lines that were not analyzed in parallel may potentially be attributable to a batch effect, we only included genes with an absolute value of the $\log_2$ mAFC of >2 (in any of the possible pairwise comparisons) among the vNF54 lines compared both directly and after normalizing against the pNF54 line analyzed in parallel. Heatmaps and hierarchical clustering based on Euclidean distance were generated using TMEV 4.9 (82). For the generation of volcano plots, samples from the different volunteers were treated as replicates, and an unpaired two-sided *t* test was performed between volunteer and parental NF54 replicates.

RT-qPCR transcriptional analysis. RNA was purified from parasite samples collected in TRIzol (Invitrogen) using the RNeasy minikit (Qiagen) as previously described (23, 79). Next, purified RNA was reverse transcribed using the reverse transcription system (Promega) alongside parallel reactions without reverse transcriptase to exclude gDNA contamination. Quantitative PCR to analyze cDNAs was performed in triplicate wells using the Power SYBR green master mix (Applied Biosystems) in a StepOnePlus real-time PCR system, essentially as previously described (26, 79). Relative transcript levels were calculated using the standard curve method and the normalizing gene serine-tRNA ligase (*serrs*), which shows stable transcript levels across the IDC. The primers used for qPCR are described in Data Set S5.

Whole-genome sequencing. PCR-free whole-genome Illumina sequencing was used to sequence the whole genome of the parental NF54 line alongside two vNF54 lines (V18 and V63). Genomic DNA was sheared to ~150 to 400 bp using a Covaris S220 ultrasonicator, and the NEBNext Ultra DNA library prep kit for Illumina was used for library preparation with specific paired-end TruSeq Illumina adaptors for each sample. Due to the high AT content of the *P. falciparum* genome, the end repair incubation step at 65°C was omitted. After quality control using the Bioanalyzer DNA high sensitivity kit (Agilent) and quantification using the Kapa library quantification kit (Roche), libraries were sequenced using the Illumina HiSeq2500 system. Over 6 million 125-bp paired-end reads were obtained for each sample.

For the data analysis, read quality was checked (FastQC program), and adaptors were trimmed (Cutadapt program) before mapping sequenced reads to the PlasmodDB *P. falciparum* 3D7 reference genome version 46 (https://plasmodb.org/plasmo/) using the BWA-MEM alignment algorithm. Next, GATK UnifiedGenotyper was used to perform variant calling based on GATK best practices to identify SNPs and small indels. GATK Variant Filtration was used to filter out variants with low calling quality (Phred QUAL $<$ 20) or low read depth (DP $<$ 20). Differences in SNP/indel allelic frequency between the different strains were calculated for each SNP/indel, and those showing a $<$50% difference were filtered out. GenomeBrowse (Golden Helix) was used to visualize alignments and variants, as well as for the detection of large subtelomeric deletions or duplications.

ChIP-seq experiments and analysis. ChIP-seq experiments were performed essentially as previously described (80). In brief, chromatin was extracted from saponin-lysed synchronous parasite cultures at the late-trophozoite/early-schizont stage using the MAGnify chromatin immunoprecipitation system (Life Technologies) (34). After cross-linking and washing, chromatin was sonicated with an M220 sonicator (Covaris) at 10% duty factor, 200 cycles per burst, and 140 W of peak incident power for 10 min. Next, 4 μg of chromatin were immunoprecipitated overnight at 4°C with 8 μg of antibody against H3K9me3 (Diagenode; catalog no. C15410193) previously coupled to protein A/G magnetic beads provided in the kit. Samples were then washed, de-cross-linked, and eluted following the MAGnify ChIP system recommendations but, at all times, avoiding high temperatures that could result in the denaturation of extremely AT-rich intergenic regions. De-cross-linking, proteinase K treatment, and elution were performed at 45°C (for 2 h, overnight, and 1.5 h, respectively).

Using a protocol adapted to a genome with an extremely high AT richness (83), libraries for Illumina sequencing were prepared from 5 ng of immunoprecipitated DNA. After end repair and addition of 3′ A overhangs, NEBNext multiplex oligos for Illumina (NEB; catalog nos. E7335 and E7500) were ligated. Agencourt AMPure XP beads (Beckman Coulter) were used for the purification steps, and the libraries were amplified (9 amplification cycles) with the Kapa HiFi PCR kit (Kapa Biosystems) in Kapa HiFi fidelity buffer (5×). Finally, 0.9× AMPure XP beads were used to purify the amplified libraries and remove adapter dimers. After library size analysis using a 4200 TapeStation system (Agilent Technologies) and quantification using the Kapa library quantification kit (Roche), sequencing was performed with a HiSeq2500 system (Illumina), obtaining 6 to 10 million 125-bp paired-end reads per sample.

Reads were mapped to the 3D7 reference genome using the Bowtie2 local alignment algorithm. Differential peak calling was performed using MACS2 (v2.2.7.1) following the author's recommendations. A first round of peak calling for every sample was performed using the MACS2 callpeak command with parameters as follows: -f BAMPE -g 2.41e7 –fe-cutoff 1.5 –nomodel –extsize 150, and then differential peaks were called using the bdgdiff command with parameters as follows: -g 250 -l 300 –cutoff 5. Results were annotated using custom scripts in Python against the *P. falciparum* 3D7 annotation in PlasmoDB (v46). The Integrative Genomics Viewer (IGV) was used to visualize the data.

H3K9me3 coverage across the genome for each sample was calculated using the DeepTools (v.3.5.0) BamCompare command. After normalizing to reads per kilobase per million (RPKMs), coverage was defined as the $\log_2$(IP/input) and computed for 100-bp intervals. For each gene, we calculated the average coverage for the 1,000 bp upstream plus the first 500 bp of the coding sequence because this is the region in which heterochromatin is typically associated with a silenced state.

Determination of sexual conversion rates. Sexual conversion rates were measured by treating sorbitol synchronized ring-stage cultures with 50 mM *N*-acetyl-D-glucosamine (GlcNAc) (Sigma-Aldrich; catalog no. A32869) for 5 days to eliminate asexual parasites. The sexual conversion rate was calculated as the gametocytemia at day 5 relative to the initial rings parasitemia at day 0 (the day of adding GlcNAc),

as previously described (84). Initial parasitemia was measured by flow cytometry (85) and gametocytemia by light microscopy quantification of Giemsa-stained smears.

Microsphiltration assay. In order to compare the deformability of erythrocytes infected with pNF54 and vNF54 (V63), we used a microsphiltration assay (86). Calibrated metal microspheres (96.50% tin, 3.00% silver, and 0.50% copper; Industrie des Poudres Sphériques) with two different size distributions (5 to 15 μm and 15 to 25 μm in diameter) were used to create a matrix within an inverted 1,000-μl antiaerosol pipette tip (Neptune). For this, 4 g of dry microspheres of each size range were resuspended in 12 ml of parasite culture medium (with 10% human serum), and 400 μl of this microsphere suspension was added into the tip and left to settle and form a 3- to 4-mm-thick layer above the tip filter. Next, 600 μl of tightly synchronized 30- to 34-hpi cultures (1 to 6% parasitemia) were placed on top of the microsphere layer and then perfused through the microsphere matrix at a flow rate of 60 ml/h using an electric pump (Syramed SP6000; Arcomed Ag), followed by a wash with 5 ml of culture medium.

Samples collected after perfusion through the matrix (in triplicate) and the original culture (not passed through the matrix) were analyzed by flow cytometry. For this, 2 μl of erythrocytes pellet were collected in 200 μl of RPMI and washed twice with 200 μl of phosphate-buffered saline (PBS), incubated for 25 min at room temperature with SYBR green (1:2,000 dilution of Sigma S9430 solution), washed twice again with 200 μl of PBS, and finally resuspended in 1.5 ml of PBS. The parasitemia of each sample was then measured with a BD Accuri C6 cytometer.

Sorbitol lysis assay. To test sorbitol sensitivity, tightly synchronized 20- to 24-hpi pNF54 and vNF54 (V63) cultures were treated with a sorbitol-containing isosmotic solution (300 mM sorbitol supplemented with 10 mM HEPES, 5 mM glucose, and adjusted to pH 7.4) or the same isosmotic solution with 100 μM of the general anion channel inhibitor 5-nitro-2-(3-phenylpropylamino) benzoic acid (NPPB, Sigma-Aldrich), which inhibits malarial new permeation pathways (87). Parasitemia was determined before (time zero) and after (time 60) incubation at 37°C for 1 h, using flow cytometry. The lysis percentage was calculated for each condition using the following formula: % lysis = [1 − (parasitemia time 60 min/parasitemia time zero)] × 100.

Western blotting. Synchronous cultures containing mainly ~40-hpi schizonts were purified by magnetic isolation, divided into pellets of approximately 2.5×10^6 schizonts, and stored frozen at −80°C. After thawing, proteins were denaturized in SDS-PAGE protein loading buffer for 5 min at 95°C and resolved by SDS-PAGE on 4 to 12% bis-Tris Criterion XT precast gels (Bio-Rad), transferred to a polyvinylidene difluoride (PVDF) membrane, and blocked for at least 1 h in 1% casein blocking buffer (Sigma). Membranes were incubated overnight at 4°C with the following primary antibodies: purified mouse antiserum against a PfMC-2TM protein (PF3D7_0114100) at 1:200 (kindly provided by Catherine Braun-Breton) (88) and mouse anti-HSP70 antibody (89) at 1:2,000. Membranes were then incubated for 1 h at room temperature with horseradish peroxidase-conjugated anti-mouse IgG secondary antibody (Promega) at 1:10,000, and peroxidase was detected using the Pierce chemiluminescence system (Pierce) following the manufacturer's instructions. To control for equal loading, parts of the membranes corresponding to different molecular weight ranges were separately hybridized with different antibodies. Signal quantification was performed using ImageJ software.

Data availability. The microarray and ChIP-seq data have been deposited in the GEO database with accession numbers GSE166258 and GSE166390, respectively. ChIP-seq data can be visualized at the UCSC genome browser using the following link: http://genome.ucsc.edu/s/apickford/apickford_lmichel. Whole-genome sequence data have been deposited at the Sequence Read Archive (SRA) database with accession number PRJNA699845. The scripts used for microarray data analysis are available in GitHub (https://github.com/CortesMalariaLab/A_Pickford_CHMItranscriptomic).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

DATA SET S1, XLS file, 0.2 MB.
DATA SET S2, XLS file, 0.7 MB.
DATA SET S3, XLS file, 1.6 MB.
DATA SET S4, XLS file, 1.2 MB.
DATA SET S5, XLS file, 0.03 MB.
FIG S1, PDF file, 0.2 MB.
FIG S2, PDF file, 0.3 MB.
FIG S3, PDF file, 0.3 MB.
FIG S4, PDF file, 0.2 MB.
FIG S5, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Catherine Braun-Breton (Université de Montpellier) for providing antibodies against a PfMC-2TM protein; the Genomics Unit at the CRG for assistance with genome sequencing; Gloria P. Gómez-Pérez, José Muñoz (ISGlobal), the volunteers and clinical staff who participated in the CHMI in Barcelona, Sofía Mira-Martínez, and Ariel Magallón-Tejada (ISGlobal) for their contribution to the collection of samples for transcriptional

analysis in the CHMI study; Adam J. Reid (Wellcome Sanger Institute), J. Alexandra Rowe, and Philip J. Spence (University of Edinburgh) for providing processed data from the Milne et al. article; and Stephen L. Hoffman and Kim Lee Sim (Sanaria) for providing cryopreserved *P. falciparum* sporozoites.

This work was supported by grants from the Spanish Ministerio de Ciencia e Innovación (MCI)/Agencia Estatal de Investigación (AEI) (SAF2016-76190-R and PID2019-107232RB-I00 to A.C.), cofunded by the European Regional Development Fund (ERDF, European Union), and grants by the Instituto de Salud Carlos III and Fundación Ramón Areces to P.L.A. for the CHMI trial. F.D. and C.L. were supported by CNRS, INSERM, and the Fondation pour la Recherche Médicale under award number Equipe FRM EQ20170336722D. A.K.P. is supported by a fellowship from the Secretary for Universities and Research, Catalan Government (FI_B 00373), cofunded by the European Social Fund (ESF), European Commission. L.M.-T. is supported by a fellowship from the Spanish Ministry of Economy and Competitiveness (BES-2017-081079), cofunded by the European Social Fund (ESF). Our research is part of ISGlobal's Program on the Molecular Mechanisms of Malaria, which is partially supported by the Fundación Ramón Areces. We acknowledge support from the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa 2019-2023 program (CEX2018-000806-S) and support from the Generalitat de Catalunya through the CERCA Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. World Health Organization. 2020. World malaria report 2020. World Health Organization, Geneva, Switzerland.
2. Venugopal K, Hentzschel F, Valkiūnas G, Marti M. 2020. *Plasmodium* asexual growth and sexual development in the haematopoietic niche of the host. Nat Rev Microbiol 18:177–189. https://doi.org/10.1038/s41579-019-0306-2.
3. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1:E5. https://doi.org/10.1371/journal.pbio.0000005.
4. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301:1503–1508. https://doi.org/10.1126/science.1087025.
5. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K, Butungi H, Metcalf T, Verzier LH, Rayner JC, Berriman M, Herren JK, Billker O, Hemberg M, Talman AM, Lawniczak MKN. 2019. The Malaria Cell Atlas: single parasite transcriptomes across the complete *Plasmodium* life cycle. Science 365:eaaw2619. https://doi.org/10.1126/science.aaw2619.
6. Mackinnon MJ, Marsh K. 2010. The selection landscape of malaria parasites. Science 328:866–871. https://doi.org/10.1126/science.1185410.
7. Zuzarte-Luis V, Mota MM. 2018. Parasite sensing of host nutrients and environmental cues. Cell Host Microbe 23:749–758. https://doi.org/10.1016/j.chom.2018.05.018.
8. Llora-Batlle O, Tinto-Font E, Cortes A. 2019. Transcriptional variation in malaria parasites: why and how. Brief Funct Genomics 18:329–341. https://doi.org/10.1093/bfgp/elz009.
9. Cortés A, Deitsch KW. 2017. Malaria epigenetics. Cold Spring Harb Perspect Med 7:a025528. https://doi.org/10.1101/cshperspect.a025528.
10. Guizetti J, Scherf A. 2013. Silence, activate, poise and switch! Mechanisms of antigenic variation in *Plasmodium falciparum*. Cell Microbiol 15:718–726. https://doi.org/10.1111/cmi.12115.
11. Voss TS, Bozdech Z, Bartfai R. 2014. Epigenetic memory takes center stage in the survival strategy of malaria parasites. Curr Opin Microbiol 20:88–95. https://doi.org/10.1016/j.mib.2014.05.007.
12. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, Ribas de Pouplana L, Preiser PR, Bozdech Z, Cortés A. 2012. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. Genome Res 22:925–938. https://doi.org/10.1101/gr.129692.111.
13. Levy SF, Ziv N, Siegal ML. 2012. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. PLoS Biol 10:e1001325. https://doi.org/10.1371/journal.pbio.1001325.
14. Veening JW, Smits WK, Kuipers OP. 2008. Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol 62:193–210. https://doi.org/10.1146/annurev.micro.62.081307.163002.
15. Starrfelt J, Kokko H. 2012. Bet-hedging–a triple trade-off between means, variances and correlations. Biol Rev Camb Philos Soc 87:742–755. https://doi.org/10.1111/j.1469-185X.2012.00225.x.
16. Smith JD, Rowe JA, Higgins MK, Lavstsen T. 2013. Malaria's deadly grip: cytoadhesion of *Plasmodium falciparum*-infected erythrocytes. Cell Microbiol 15:1976–1983. https://doi.org/10.1111/cmi.12183.
17. Scherf A, Lopez-Rubio JJ, Riviere L. 2008. Antigenic variation in *Plasmodium falciparum*. Annu Rev Microbiol 62:445–470. https://doi.org/10.1146/annurev.micro.61.080706.093134.
18. Wahlgren M, Goel S, Akhouri RR. 2017. Variant surface antigens of *Plasmodium falciparum* and their roles in severe malaria. Nat Rev Microbiol 15:479–491. https://doi.org/10.1038/nrmicro.2017.47.
19. Mira-Martinez S, Pickford AK, Rovira-Graells N, Guetens P, Tinto-Font E, Cortes A, Rosanas-Urgell A. 2019. Identification of antimalarial compounds that require CLAG3 for their uptake by *Plasmodium falciparum*-infected erythrocytes. Antimicrob Agents Chemother 63:e00052-19. https://doi.org/10.1128/AAC.00052-19.
20. Nguitragool W, Bokhari AA, Pillai AD, Rayavara K, Sharma P, Turpin B, Aravind L, Desai SA. 2011. Malaria parasite *clag3* genes determine channel-mediated nutrient uptake by infected red blood cells. Cell 145:665–677. https://doi.org/10.1016/j.cell.2011.05.002.
21. Pillai AD, Nguitragool W, Lyko B, Dolinta K, Butler MM, Nguyen ST, Peet NP, Bowlin TL, Desai SA. 2012. Solute restriction reveals an essential role for *clag3*-associated channels in malaria parasite nutrient acquisition. Mol Pharmacol 82:1104–1114. https://doi.org/10.1124/mol.112.081224.
22. Mira-Martínez S, Rovira-Graells N, Crowley VM, Altenhofen LM, Llinás M, Cortés A. 2013. Epigenetic switches in *clag3* genes mediate blasticidin S resistance in malaria parasites. Cell Microbiol 15:1913–1923. https://doi.org/10.1111/cmi.12162.
23. Mira-Martínez S, van Schuppen E, Amambua-Ngwa A, Bottieau E, Affara M, Van Esbroeck M, Vlieghe E, Guetens P, Rovira-Graells N, Gómez-Perez GP, Alonso PL, D'Alessandro U, Rosanas-Urgell A, Cortés A. 2017. Expression of the *Plasmodium falciparum* clonally variant *clag3* genes in human infections. J Infect Dis 215:938–945. https://doi.org/10.1093/infdis/jix053.
24. Sharma P, Wollenberg K, Sellers M, Zainabadi K, Galinsky K, Moss E, Nguitragool W, Neafsey D, Desai SA. 2013. An epigenetic antimalarial resistance mechanism involving parasite genes linked to nutrient uptake. J Biol Chem 288:19429–19440. https://doi.org/10.1074/jbc.M113.468371.

25. Cortés A, Crowley VM, Vaquero A, Voss TS. 2012. A view on the role of epigenetics in the biology of malaria parasites. PLoS Pathog 8:e1002943. https://doi.org/10.1371/journal.ppat.1002943.
26. Crowley VM, Rovira-Graells N, de Pouplana LR, Cortés A. 2011. 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. https://doi.org/10.1111/j.1365-2958.2011.07574.x.
27. Jiang L, Lopez-Barragan MJ, Jiang H, Mu J, Gaur D, Zhao K, Felsenfeld G, Miller LH. 2010. Epigenetic control of the variable expression of a *Plasmodium falciparum* receptor protein for erythrocyte invasion. Proc Natl Acad Sci U S A 107:2224–2229. https://doi.org/10.1073/pnas.0913396107.
28. Lopez-Rubio JJ, Gontijo AM, Nunes MC, Issar N, Hernandez Rivas R, Scherf A. 2007. 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. https://doi.org/10.1111/j.1365-2958.2007.06009.x.
29. Brancucci NM, Bertschi NL, Zhu L, Niederwieser I, Chin WH, Wampfler R, Freymond C, Rottmann M, Felger I, Bozdech Z, Voss TS. 2014. Heterochromatin protein 1 secures survival and transmission of malaria parasites. Cell Host Microbe 16:165–176. https://doi.org/10.1016/j.chom.2014.07.004.
30. Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ, Hartl DL, Deitsch KW. 2007. Epigenetic memory at malaria virulence genes. Proc Natl Acad Sci U S A 104:899–902. https://doi.org/10.1073/pnas.0609084103.
31. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, Gysin J, Lanzer M. 1998. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of *var* genes during intra-erythrocytic development in *Plasmodium falciparum*. EMBO J 17:5418–5426. https://doi.org/10.1093/emboj/17.18.5418.
32. Cortés A, Carret C, Kaneko O, Yim Lim BY, Ivens A, Holder AA. 2007. Epigenetic silencing of *Plasmodium falciparum* genes linked to erythrocyte invasion. PLoS Pathog 3:e107. https://doi.org/10.1371/journal.ppat.0030107.
33. Dzikowski R, Frank M, Deitsch K. 2006. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. PLoS Pathog 2:e22. https://doi.org/10.1371/journal.ppat.0020022.
34. Rovira-Graells N, Crowley VM, Bancells C, Mira-Martínez S, Ribas de Pouplana L, Cortés A. 2015. Deciphering the principles that govern mutually exclusive expression of *Plasmodium falciparum clag3* genes. Nucleic Acids Res 43:8243–8257. https://doi.org/10.1093/nar/gkv730.
35. Voss TS, Healer J, Marty AJ, Duffy MF, Thompson JK, Beeson JG, Reeder JC, Crabb BS, Cowman AF. 2006. A *var* gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. Nature 439:1004–1008. https://doi.org/10.1038/nature04407.
36. Bachmann A, Petter M, Krumkamp R, Esen M, Held J, Scholz JA, Li T, Sim BK, Hoffman SL, Kremsner PG, Mordmuller B, Duffy MF, Tannich E. 2016. Mosquito passage dramatically changes *var* gene expression in controlled human *Plasmodium falciparum* infections. PLoS Pathog 12:e1005538. https://doi.org/10.1371/journal.ppat.1005538.
37. Peters J, Fowler E, Gatton M, Chen N, Saul A, Cheng Q. 2002. High diversity and rapid changeover of expressed *var* genes during the acute phase of *Plasmodium falciparum* infections in human volunteers. Proc Natl Acad Sci U S A 99:10689–10694. https://doi.org/10.1073/pnas.162349899.
38. Dimonte S, Bruske EI, Hass J, Supan C, Salazar CL, Held J, Tschan S, Esen M, Flotenmeyer M, Koch I, Berger J, Bachmann A, Sim BK, Hoffman SL, Kremsner PG, Mordmuller B, Frank M. 2016. Sporozoite route of infection influences in vitro *var* gene transcription of *Plasmodium falciparum* parasites from controlled human infections. J Infect Dis 214:884–894. https://doi.org/10.1093/infdis/jiw225.
39. Lavazec C, Sanyal S, Templeton TJ. 2007. Expression switching in the *stevor* and *Pfmc-2TM* superfamilies in *Plasmodium falciparum*. Mol Microbiol 64:1621–1634. https://doi.org/10.1111/j.1365-2958.2007.05767.x.
40. Mok BW, Ribacke U, Winter G, Yip BH, Tan CS, Fernandez V, Chen Q, Nilsson P, Wahlgren M. 2007. Comparative transcriptomal analysis of isogenic *Plasmodium falciparum* clones of distinct antigenic and adhesive phenotypes. Mol Biochem Parasitol 151:184–192. https://doi.org/10.1016/j.molbiopara.2006.11.006.
41. Gomez-Escobar N, Amambua-Ngwa A, Walther M, Okebe J, Ebonyi A, Conway DJ. 2010. Erythrocyte invasion and merozoite ligand gene expression in severe and mild *Plasmodium falciparum* malaria. J Infect Dis 201:444–452. https://doi.org/10.1086/649902.
42. Tonkin-Hill GQ, Trianty L, Noviyanti R, Nguyen HHT, Sebayang BF, Lampah DA, Marfurt J, Cobbold SA, Rambhatla JS, McConville MJ, Rogerson SJ, Brown GV, Day KP, Price RN, Anstey NM, Papenfuss AT, Duffy MF. 2018. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen-encoding *var* genes. PLoS Biol 16:e2004328. https://doi.org/10.1371/journal.pbio.2004328.
43. Cortés A. 2008. Switching *Plasmodium falciparum* genes on and off for erythrocyte invasion. Trends Parasitol 24:517–524. https://doi.org/10.1016/j.pt.2008.08.005.
44. Brown AC, Guler JL. 2020. From circulation to cultivation: *Plasmodium* in vivo versus in vitro. Trends Parasitol 36:914–926. https://doi.org/10.1016/j.pt.2020.08.008.
45. Mackinnon MJ, Li J, Mok S, Kortok MM, Marsh K, Preiser PR, Bozdech Z. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. PLoS Pathog 5:e1000644. https://doi.org/10.1371/journal.ppat.1000644.
46. Daily JP, Le Roch KG, Sarr O, Ndiaye D, Lukens A, Zhou Y, Ndir O, Mboup S, Sultan A, Winzeler EA, Wirth DF. 2005. In vivo transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. J Infect Dis 191:1196–1203. https://doi.org/10.1086/428289.
47. Bachmann A, Petter M, Tilly AK, Biller L, Uliczka KA, Duffy MF, Tannich E, Bruchhaus I. 2012. Temporal expression and localization patterns of variant surface antigens in clinical *Plasmodium falciparum* isolates during erythrocyte schizogony. PLoS One 7:e49540. https://doi.org/10.1371/journal.pone.0049540.
48. Wang CW, Hermsen CC, Sauerwein RW, Arnot DE, Theander TG, Lavstsen T. 2009. The *Plasmodium falciparum var* gene transcription strategy at the onset of blood stage infection in a human volunteer. Parasitol Int 58:478–480. https://doi.org/10.1016/j.parint.2009.07.004.
49. Lavstsen T, Magistrado P, Hermsen CC, Salanti A, Jensen AT, Sauerwein R, Hviid L, Theander TG, Staalsoe T. 2005. Expression of *Plasmodium falciparum* erythrocyte membrane protein 1 in experimentally infected humans. Malar J 4:21. https://doi.org/10.1186/1475-2875-4-21.
50. Abdi AI, Hodgson SH, Muthui MK, Kivisi CA, Kamuyu G, Kimani D, Hoffman SL, Juma E, Ogutu B, Draper SJ, Osier F, Bejon P, Marsh K, Bull PC. 2017. *Plasmodium falciparum* malaria parasite *var* gene expression is modified by host antibodies: longitudinal evidence from controlled infections of Kenyan adults with varying natural exposure. BMC Infect Dis 17:585. https://doi.org/10.1186/s12879-017-2686-0.
51. Hoo R, Bruske E, Dimonte S, Zhu L, Mordmüller B, Sim BKL, Kremsner PG, Hoffman SL, Bozdech Z, Frank M, Preiser PR. 2019. Transcriptome profiling reveals functional variation in *Plasmodium falciparum* parasites from controlled human malaria infection studies. EBioMedicine 48:442–452. https://doi.org/10.1016/j.ebiom.2019.09.001.
52. Milne K, Ivens A, Reid AJ, Lotkowska ME, O'Toole A, Sankaranarayanan G, Munoz Sandoval D, Nahrendorf W, Regnault C, Edwards NJ, Silk SE, Payne RO, Minassian AM, Venkatraman N, Sanders MJ, Hill AV, Barrett M, Berriman M, Draper SJ, Rowe JA, Spence PJ. 2021. Mapping immune variation and *var* gene switching in naive hosts infected with *Plasmodium falciparum*. Elife 10:e62800. https://doi.org/10.7554/eLife.62800.
53. Gomez-Perez GP, Legarda A, Munoz J, Sim BK, Ballester MR, Dobano C, Moncunill G, Campo JJ, Cistero P, Jimenez A, Barrios D, Mordmuller B, Pardos J, Navarro M, Zita CJ, Nhamuave CA, Garcia-Basteiro AL, Sanz A, Aldea M, Manoj A, Gunasekera A, Billingsley PF, Aponte JJ, James ER, Guinovart C, Antonijoan RM, Kremsner PG, Hoffman SL, Alonso PL. 2015. Controlled human malaria infection by intramuscular and direct venous inoculation of cryopreserved *Plasmodium falciparum* sporozoites in malaria-naive volunteers: effect of injection volume and dose on infectivity rates. Malar J 14:306. https://doi.org/10.1186/s12936-015-0817-x.
54. Flueck C, Bartfai R, Volz J, Niederwieser I, Salcedo-Amaya AM, Alako BT, Ehlgen F, Ralph SA, Cowman AF, Bozdech Z, Stunnenberg HG, Voss TS. 2009. *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog 5:e1000569. https://doi.org/10.1371/journal.ppat.1000569.
55. Fraschka SA, Filarsky M, Hoo R, Niederwieser I, Yam XY, Brancucci NMB, Mohring F, Mushunje AT, Huang X, Christensen PR, Nosten F, Bozdech Z, Russell B, Moon RW, Marti M, Preiser PR, Bartfai R, Voss TS. 2018. Comparative heterochromatin profiling reveals conserved and unique epigenome signatures linked to adaptation and development of malaria parasites. Cell Host Microbe 23:407–420. https://doi.org/10.1016/j.chom.2018.01.008.
56. Lopez-Rubio JJ, Mancio-Silva L, Scherf A. 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe 5:179–190. https://doi.org/10.1016/j.chom.2008.12.012.
57. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AM, Cohen AM, Janssen-Megens EM, van de Vegte-Bolmer M, Selzer RR,

Iniguez AL, Green RD, Sauerwein RW, Jensen ON, Stunnenberg HG. 2009. Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. Proc Natl Acad Sci U S A 106:9655–9660. https://doi.org/10.1073/pnas.0902515106.
58. Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, Speed TP, Cowman AF. 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol 7:R12. https://doi.org/10.1186/gb-2006-7-2-r12.
59. Templeton TJ. 2009. The varieties of gene amplification, diversification and hypervariability in the human malaria parasite, *Plasmodium falciparum*. Mol Biochem Parasitol 166:109–116. https://doi.org/10.1016/j.molbiopara.2009.04.003.
60. Nolte D, Hundt E, Langsley G, Knapp B. 1991. A *Plasmodium falciparum* blood stage antigen highly homologous to the glycophorin binding protein GBP. Mol Biochem Parasitol 49:253–264. https://doi.org/10.1016/0166-6851(91)90069-i.
61. van Schravendijk MR, Wilson RJ, Newbold CI. 1987. Possible pitfalls in the identification of glycophorin-binding proteins of *Plasmodium falciparum*. J Exp Med 166:376–390. https://doi.org/10.1084/jem.166.2.376.
62. Painter HJ, Altenhofen LM, Kafsack BF, Llinas M. 2013. Whole-genome analysis of *Plasmodium* spp. utilizing a new Agilent Technologies DNA microarray platform. Methods Mol Biol 923:213–219. https://doi.org/10.1007/978-1-62703-026-7_14.
63. Sam-Yellowe TY, Florens L, Johnson JR, Wang T, Drazba JA, Le Roch KG, Zhou Y, Batalov S, Carucci DJ, Winzeler EA, Yates JR, III. 2004. A *Plasmodium* gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. Genome Res 14:1052–1059. https://doi.org/10.1101/gr.2126104.
64. Weber JL. 1988. Interspersed repetitive DNA from *Plasmodium falciparum*. Mol Biochem Parasitol 29:117–124. https://doi.org/10.1016/0166-6851(88)90066-7.
65. Kyes SA, Rowe JA, Kriek N, Newbold CI. 1999. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. Proc Natl Acad Sci U S A 96:9333–9338. https://doi.org/10.1073/pnas.96.16.9333.
66. Fernandez V, Hommel M, Chen Q, Hagblom P, Wahlgren M. 1999. Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*-infected erythrocyte are encoded by the *rif* gene family and are the target of human immune responses. J Exp Med 190:1393–1404. https://doi.org/10.1084/jem.190.10.1393.
67. Petter M, Haeggstrom M, Khattab A, Fernandez V, Klinkert MQ, Wahlgren M. 2007. Variant proteins of the *Plasmodium falciparum* RIFIN family show distinct subcellular localization and developmental expression patterns. Mol Biochem Parasitol 156:51–61. https://doi.org/10.1016/j.molbiopara.2007.07.011.
68. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, Saul A. 1998. *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 97:161–176. https://doi.org/10.1016/s0166-6851(98)00144-3.
69. Witmer K, Schmid CD, Brancucci NM, Luah YH, Preiser PR, Bozdech Z, Voss TS. 2012. Analysis of subtelomeric virulence gene families in *Plasmodium falciparum* by comparative transcriptional profiling. Mol Microbiol 84:243–259. https://doi.org/10.1111/j.1365-2958.2012.08019.x.
70. Frank M, Kirkman L, Costantini D, Sanyal S, Lavazec C, Templeton TJ, Deitsch KW. 2008. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. Int J Parasitol 38:1099–1109. https://doi.org/10.1016/j.ijpara.2008.01.010.
71. Sauerwein RW, Roestenberg M, Moorthy VS. 2011. Experimental human challenge infections can accelerate clinical malaria vaccine development. Nat Rev Immunol 11:57–64. https://doi.org/10.1038/nri2902.
72. McCarthy JS, Sekuloski S, Griffin PM, Elliott S, Douglas N, Peatey C, Rockett R, O'Rourke P, Marquart L, Hermsen C, Duparc S, Möhrle J, Trenholme KR, Humberstone AJ. 2011. A pilot randomised trial of induced blood-stage *Plasmodium falciparum* infections in healthy volunteers for testing efficacy of new antimalarial drugs. PLoS One 6:e21914. https://doi.org/10.1371/journal.pone.0021914.
73. Sanyal S, Egee S, Bouyer G, Perrot S, Safeukui I, Bischoff E, Buffet P, Deitsch KW, Mercereau-Puijalon O, David PH, Templeton TJ, Lavazec C. 2012. *Plasmodium falciparum* STEVOR proteins impact erythrocyte mechanical properties. Blood 119:e1-8–e8. https://doi.org/10.1182/blood-2011-08-370734.
74. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, Williams AE, Drought LG, Kwiatkowski DP, Baker DA, Cortes A, Llinas M. 2014. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507:248–252. https://doi.org/10.1038/nature12920.
75. Spence PJ, Brugat T, Langhorne J. 2015. Mosquitoes reset malaria parasites. PLoS Pathog 11:e1004987. https://doi.org/10.1371/journal.ppat.1004987.
76. Spence PJ, Jarra W, Levy P, Reid AJ, Chappell L, Brugat T, Sanders M, Berriman M, Langhorne J. 2013. Vector transmission regulates immune control of *Plasmodium* virulence. Nature 498:228–231. https://doi.org/10.1038/nature12231.
77. Zanghì G, Vembar SS, Baumgarten S, Ding S, Guizetti J, Bryant JM, Mattei D, Jensen ATR, Rénia L, Goh YS, Sauerwein R, Hermsen CC, Franetich JF, Bordessoulles M, Silvie O, Soulard V, Scatton O, Chen P, Mecheri S, Mazier D, Scherf A. 2018. A specific PfEMP1 is expressed in *P. falciparum* sporozoites and plays a role in hepatocyte infection. Cell Rep 22:2951–2963. https://doi.org/10.1016/j.celrep.2018.02.075.
78. Gomez-Diaz E, Yerbanga RS, Lefevre T, Cohuet A, Rowley MJ, Ouedraogo JB, Corces VG. 2017. Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. Sci Rep 7:40655. https://doi.org/10.1038/srep40655.
79. Casas-Vila N, Pickford AK, Portugaliza HP, Tintó-Font E, Cortés A. Transcriptional analysis of tightly synchronized *Plasmodium falciparum* intraerythrocytic stages by RT-qPCR. Methods Mol Biol, in press. https://doi.org/10.1007/978-1-0716-1681-9_10.
80. Llorà-Batlle O, Michel-Todó L, Witmer K, Toda H, Fernández-Becerra C, Baum J, Cortés A. 2020. Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*. Sci Adv 6:eaaz5057. https://doi.org/10.1126/sciadv.aaz5057.
81. Lemieux JE, Gomez-Escobar N, Feller A, Carret C, Amambua-Ngwa A, Pinches R, Day F, Kyes SA, Conway DJ, Holmes CC, Newbold CI. 2009. 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. https://doi.org/10.1073/pnas.0811829106.
82. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 microarray software suite. Methods Enzymol 411:134–193. https://doi.org/10.1016/S0076-6879(06)11009-5.
83. Kensche PR, Hoeijmakers WA, Toenhake CG, Bras M, Chappell L, Berriman M, Bartfai R. 2016. The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res 44:2110–2124. https://doi.org/10.1093/nar/gkv1214.
84. Bancells C, Llora-Batlle O, Poran A, Notzel C, Rovira-Graells N, Elemento O, Kafsack BFC, Cortes A. 2019. Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. Nat Microbiol 4:144–154. https://doi.org/10.1038/s41564-018-0291-7.
85. Rovira-Graells N, Aguilera-Simon S, Tinto-Font E, Cortes A. 2016. New assays to characterise growth-related phenotypes of *Plasmodium falciparum* reveal variation in density-dependent growth inhibition between parasite lines. PLoS One 11:e0165358. https://doi.org/10.1371/journal.pone.0165358.
86. Lavazec C, Deplaine G, Safeukui I, Perrot S, Milon G, Mercereau-Puijalon O, David PH, Buffet P. 2013. Microsphiltration: a microsphere matrix to explore erythrocyte deformability. Methods Mol Biol 923:291–297. https://doi.org/10.1007/978-1-62703-026-7_20.
87. Kirk K, Horner HA, Elford BC, Ellory JC, Newbold CI. 1994. Transport of diverse substrates into malaria-infected erythrocytes via a pathway showing functional characteristics of a chloride channel. J Biol Chem 269:3339–3347. https://doi.org/10.1016/S0021-9258(17)41868-0.
88. Vincensini L, Richert S, Blisnick T, Van Dorsselaer A, Leize-Wagner E, Rabilloud T, Braun Breton C. 2005. Proteomic analysis identifies novel proteins of the Maurer's clefts, a secretory compartment delivering *Plasmodium falciparum* proteins to the surface of its host cell. Mol Cell Proteomics 4:582–593. https://doi.org/10.1074/mcp.M400176-MCP200.
89. Naissant B, Dupuy F, Duffier Y, Lorthiois A, Duez J, Scholz J, Buffet P, Merckx A, Bachmann A, Lavazec C. 2016. *Plasmodium falciparum* STEVOR phosphorylation regulates host erythrocyte deformability enabling malaria parasite transmission. Blood 127:e42–e53. https://doi.org/10.1182/blood-2016-01-690776.

Supplementary materials for this paper are only available in a screen compatible format.

Find them at: https://journals.asm.org/doi/10.1128/mBio.01636-21?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub_0pubmed - supplementary-materials.

3 **Article 3:** Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*

SCIENCE ADVANCES | RESEARCH ARTICLE

MICROBIOLOGY

Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*

Oriol Llorà-Batlle[1]*, Lucas Michel-Todó[1], Kathrin Witmer[2], Haruka Toda[1], Carmen Fernández-Becerra[1,3], Jake Baum[2], Alfred Cortés[1,4†]

Malaria transmission requires that some asexual parasites convert into sexual forms termed gametocytes. The initial stages of sexual development, including sexually committed schizonts and sexual rings, remain poorly characterized, mainly because they are morphologically identical to their asexual counterparts and only a small subset of parasites undergo sexual development. Here, we describe a system for controlled sexual conversion in the human malaria parasite *Plasmodium falciparum*, based on conditional expression of the PfAP2-G transcription factor. Using this system, ~90 percent of the parasites converted into sexual forms upon induction, enabling the characterization of committed and early sexual stages without further purification. We characterized sexually committed schizonts and sexual rings at the transcriptomic and phenotypic levels, which revealed down-regulation of genes involved in solute transport upon sexual commitment, among other findings. The new inducible lines will facilitate the study of early sexual stages at additional levels, including multiomic characterization and drug susceptibility assays.

INTRODUCTION

Malaria is a vector-borne disease caused by protozoan parasites of the genus *Plasmodium*. Of the five species that infect humans, *Plasmodium falciparum* is the most lethal. During its ~48-hour intraerythrocytic development cycle, the parasite develops through the asexual ring, trophozoite, and multinucleated schizont stages. Upon egress, up to 32 daughter merozoites are released and invade new erythrocytes. The asexual growth of the parasite in the human blood is responsible for all malaria symptoms, but asexual stages cannot infect anopheline mosquito vectors. Human to vector transmission requires that some parasites differentiate into nonreplicating sexual forms termed gametocytes, which develop through the morphologically distinct stages I to V over ~10 days. Immature gametocytes are sequestered in tissues such as the bone marrow until they are released back to the peripheral blood circulation as mature (stage V) gametocytes that are infectious to mosquitoes (*1*, *2*). Since gametocytes are essential for transmission, they are an attractive target for intervention in the context of renewed efforts to eliminate malaria (*3*).

The first step for the production of gametocytes is the commitment of a subset of asexual parasites to sexual development (*1*, *2*, *4*). Commitment, defined as a cell state that irreversibly results in sexual conversion at a later point (*5*), is marked by expression of the master regulator PfAP2-G (*6*–*8*), a transcription factor of the ApiAp2 family (*9*). In asexual parasites, the *pfap2-g* gene is epigenetically silenced by heterochromatin containing the histone mark H3K9me3 and the heterochromatin protein 1 (HP1) (*10*–*12*). Activation of the gene, which triggers the sexual development program, requires eviction of HP1 by the gametocyte development 1 (GDV1) protein (*13*, *14*).

Following commitment, the next step for gametocyte production is sexual conversion, marked by the expression of specific proteins absent from any replicating blood stages (*5*). For many years, the prevailing model was that sexually committed parasites must undergo an additional round of replication before sexual conversion (*4*), but recent research in *P. falciparum* and the murine malaria parasite *Plasmodium berghei* demonstrated that when AP2-G expression starts early enough in the ring stage, conversion can proceed directly without additional replication after commitment (*5*, *15*). The two alternative sexual differentiation pathways are named next cycle conversion (NCC) and same cycle conversion (SCC) (*5*). The first developmental stage upon sexual conversion by either pathway is the sexual ring, also referred to as sexually or gametocyte-committed ring, or gametocyte ring (*1*, *13*, *16*–*18*). Sexual rings develop into stage I gametocytes and then follow sexual development until they reach maturity (stage V).

In *P. falciparum*, the proportion of parasites that abandon asexual growth and start differentiating into gametocytes, referred to as the sexual conversion rate, is typically low (<10%), both under culture conditions and in human infections. The level of investment into production of sexual forms varies between parasite clones and is also affected by the environment: While there is a baseline level of spontaneous *pfap2-g* activation and sexual conversion (*6*), several types of stress, including immune pressure, nutrient limitation, and drug pressure, have been proposed to stimulate conversion rates (*1*, *16*). To date, the best established stimulus is depletion of the serum lipid lysophosphatidylcholine or choline from the culture medium (*19*).

There are no known differences in morphology or physical properties between sexually committed schizonts and their asexual counterparts or between sexual and asexual rings. Together with low levels of sexual conversion, this limits the ability to study these initial stages of sexual development, as it is difficult to obtain sufficient biological material and they cannot be readily separated from asexual parasites (*1*). To overcome this limitation, we generated parasite lines in which expression of *pfap2-g* can be conditionally activated to induce massive synchronous sexual conversion. Using

[1]ISGlobal, Hospital Clinic–Universitat de Barcelona, Barcelona 08036, Catalonia, Spain. [2]Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK. [3]IGTP Institut d'Investigació Germans Trias i Pujol, Badalona 08916, Catalonia, Spain. [4]ICREA, Barcelona 08010, Catalonia, Spain.
*Present address: The Gurdon Institute and Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 1QN, UK.
†Corresponding author. Email: alfred.cortes@isglobal.org

these new parasite lines, we obtained highly pure preparations of sexually committed parasites and sexual rings that enabled the characterization of these largely unexplored parasite stages at multiple levels.

RESULTS

Conditional activation of *pfap2-g* leads to massive sexual conversion

To conditionally activate *pfap2-g* expression and trigger sexual conversion in *P. falciparum*, we designed a conditional activation construct that was inserted immediately upstream of the *pfap2-g* coding sequence [73 base pairs (bp) upstream of the start codon, i.e., between the promoter and the coding sequence] using the CRISPR-Cas9 system (*20*). The construct consisted of a strong constitutive calmodulin promoter (5′*cam*), the *hdhfr* selectable marker (confers resistance to the drug WR99210), and a terminator sequence. The *hdhfr* marker and terminator are flanked by *loxP* sites, such that upon recombination between the two *loxP* sites, the 5′*cam* promoter is adjacent to the *pfap2-g* coding sequence and is expected to control its expression. Recombination is mediated by the inducible Cre recombinase (DiCre), which is active only upon inducing dimerization with rapamycin (Fig. 1A) (*21*).

We integrated the conditional activation construct in the E5 line, a 3D7 subclone with high levels of sexual conversion (*6*). In this parasite line, conversion is stimulated by choline depletion (*22*). In a preliminary set of experiments, the DiCre recombinase was expressed episomally. Treatment with rapamycin resulted in ~30% sexual conversion, whereas gametocytes were never observed in dimethyl sulfoxide (DMSO)–treated control cultures (fig. S1, A and B). To achieve higher conversion rates, we generated a new transgenic line in which the DiCre expression cassette was integrated at the liver specific protein 1 (*lisp1*) locus, which is dispensable during blood stages (Fig. 1, B and C, and fig. S1C) (*23*). The parasite line, termed E5 gametocyte-inducible line (E5ind), was maintained under constant WR99210 pressure to select for parasites that keep the edited *pfap2-g* locus in a transcriptionally active state. Treatment with rapamycin resulted in efficient excision of the floxed region, with undetectable levels of unedited locus by diagnostic polymerase chain reaction (PCR) (Fig. 1C).

To determine the optimal time for induction, we added rapamycin to tightly synchronized E5ind cultures at different stages. Induction at the late trophozoite stage resulted in maximal sexual conversion at the cycle after induction (NCC route of conversion) and maximal total amount of gametocytes (fig. S2). Conversion via the SCC route was observed when cultures were treated at 0 to 5 hours postinvasion (hpi), but it occurred at relatively low levels (<10%) (fig. S2B). The low levels of conversion by the SCC route were expected from the kinetics of DiCre activity, as excision in 50% of the parasites requires >10 hours (*24*). Thus, in the majority of parasites, direct conversion via the SCC route is no longer possible by the time recombination occurs and PfAP2-G is expressed (*5*).

On the basis of these results, we established a straightforward induction protocol in which rapamycin is added to sorbitol-synchronized cultures when the majority of parasites reach the late trophozoite/early schizont stage (~20 hours after synchronization) (Fig. 2A). Using this protocol, we consistently obtained ~90% sexual conversion (e.g., 90% of the rings at the cycle after rapamycin treatment developed as gametocytes), as determined by light microscopy analysis of Giemsa-stained smears (Fig. 2B). By starting with cultures at

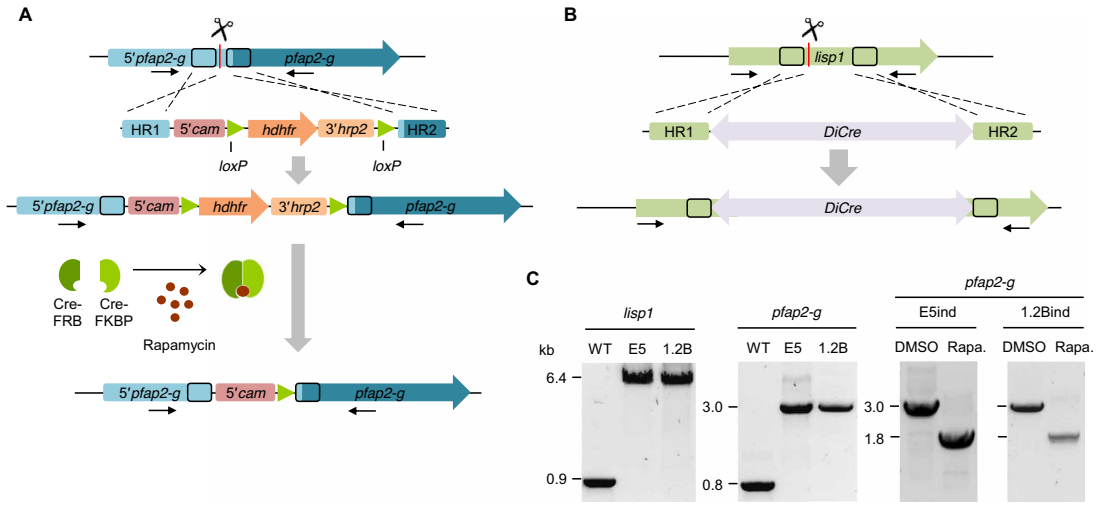


Fig. 1. Generation of transgenic parasite lines to conditionally activate *pfap2-g* expression. (**A**) Overview of the strategy to generate the inducible parasite lines. The conditional activation cassette was integrated using CRISPR-Cas9 technology. Arrows indicate the position of the primers used for diagnostic PCR. Scissors indicate the position targeted by the guide RNA, where Cas9-mediated cleavage is expected. HR refers to homology regions. The part of the *pfap2-g* upstream region (5′*pfap2-g*) presumably including the promoter of the gene remains intact after editing. (**B**) Schematic of the strategy to integrate the DiCre expression cassette in the *lisp1* locus using CRISPR-Cas9 technology. (**C**) Diagnostic PCR analysis to validate the integration of the constructs and to assess recombination in samples collected 24 hours after induction with rapamycin and in DMSO-treated controls. WT, wild type.

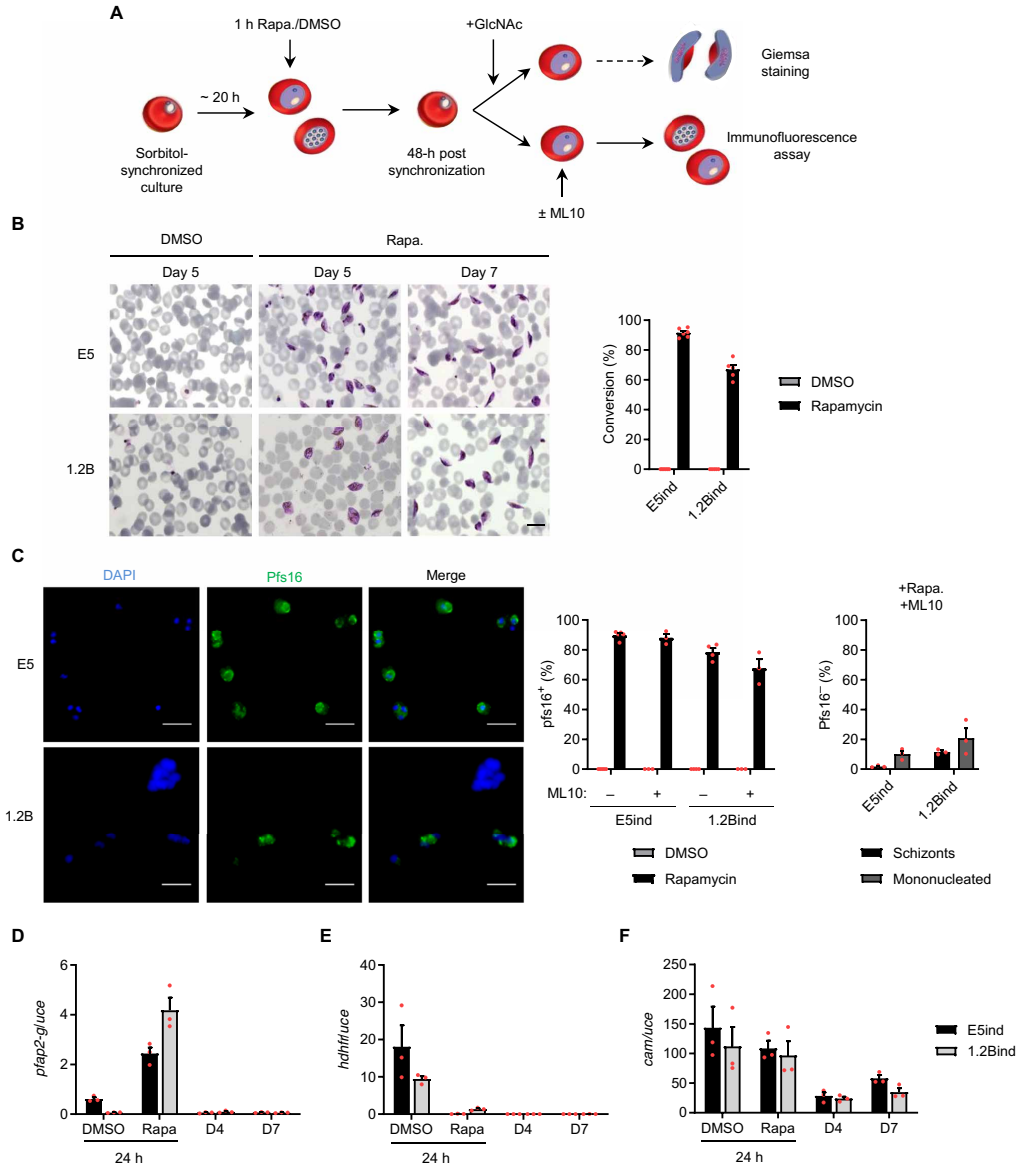


Fig. 2. Sexual conversion upon conditional activation of *pfap2-g*. (**A**) Schematic of the procedure used to induce *pfap2-g* expression and trigger sexual conversion in the inducible lines. GlcNAc, N-acetylglucosamine. (**B**) Representative images of Giemsa-stained smears of the gametocytes obtained after induction and quantification of sexual conversion rates based on analysis of Giemsa-stained smears (day 5 gametocytes). Data are presented as the average and SEM of six (E5ind) or five (1.2Bind) independent experiments. Scale bar, 10 μm. (**C**) Representative Pfs16 IFA images and quantification of the percentage of Pfs16-positive parasites determined 3 days after rapamycin induction, in cultures treated or not with ML10. The bar chart at the right shows the distribution of schizonts and mononucleated cells (according to the number of nuclei) among Pfs16-negative parasites in ML10-treated cultures. Data are presented as the average and SEM of four (left bar chart) or three (right bar chart) independent experiments. In each experiment, >250 parasites were counted. Scale bars, 10 μm. (**D** to **F**) Transcriptional analysis of *pfap2-g*, *hdhfr*, and the endogenous *cam* 24 hours after rapamycin or DMSO (control) treatment and in day 4 or 7 gametocytes. Transcript levels are normalized against ubiquitin-conjugating enzyme (*uce*). Data are presented as the average and SEM of three independent experiments.

high parasitemia, we obtained gametocytemia levels as high as 12%. No gametocytes were observed in DMSO-treated control cultures.

Similar conversion rates were estimated using immunofluorescence assays (IFAs) with antibodies against the early gametocyte marker Pfs16, which is expressed from stage I of gametocyte development onward (*5*). The proportion of Pfs16-positive cells was determined 3 days after induction, when asexual parasites are at the schizont stage and parasites undergoing sexual development are at gametocyte stage I and already express Pfs16. In some experiments, cultures were treated with the cGMP-dependent protein kinase (PKG) inhibitor ML10 to block schizont rupture and reinvasion (*25*), thus preventing amplification of nonconverting parasites. In all cases, ~90% of parasites were Pfs16 positive in rapamycin-induced E5ind cultures, whereas no Pfs16-positive parasites were observed in DMSO-treated control cultures (Fig. 2C). The majority of Pfs16-negative parasites in ML10-treated induced cultures were mononucleated, suggesting that parasites that failed to convert upon induction were predominantly dead or growth arrested. Only ~2% of parasites in these cultures were at the multinucleated schizont stage, the stage expected for healthy replicating parasites 3 days after induction (Fig. 2C).

Conditional activation of *pfap2-g* rescues gametocyte production in a gametocyte nonproducer line

To test the inducible system in a gametocyte nonproducer line, we used the 3D7 subclone 1.2B, derived from the 3D7-A stock. In the 3D7-A line and its subclones, including 1.2B, *pfap2-g* is expressed at very low levels and consequently gametocytes are not produced (*6*, *26*). Next-generation sequencing of the 1.2B genome, conducted as part of an ongoing comparative chromatin immunoprecipitation sequencing (ChIP-seq) study, revealed a nonsense mutation in the *gdv1* gene that introduces a premature STOP codon (Q578*), resulting in a truncated GDV1 protein lacking the last 21 amino acids. The mutation was validated by Sanger sequencing (fig. S3). Given that GDV1 is an upstream regulator of *pfap2-g* activation (*14*), the GDV1 truncation in the 1.2B line likely underlies its inability to form gametocytes.

We edited the 1.2B genome in an analogous way as we did with the E5 line to generate the 1.2Bind line. Induction with rapamycin resulted in ~70% sexual conversion, determined by either light microscopy or Pfs16 IFA, whereas no gametocytes were observed in control cultures (Figs. 1C and 2, A to C). This result shows that the conditional *pfap2-g* activation system can overcome a deficiency in a factor operating upstream of *pfap2-g*, indicating that PfAP2-G activation is sufficient to trigger sexual conversion. It also supports the idea that GDV1 is needed to reverse the silencing of *pfap2-g* (*14*) but is dispensable for normal gametocyte development thereafter.

Transcriptional changes at the *pfap2-g* locus upon induction

Transcriptional analysis of E5ind and 1.2Bind 24 hours after rapamycin treatment revealed activation of *pfap2-g* expression and loss of *hdhfr* expression, as expected (Fig. 2, D and E). However, *pfap2-g* transcript levels were >25-fold lower than the levels of endogenous *cam* transcripts (Fig. 2F) and rather resembled the *pfap2-g* levels in wild-type parasites (*5*). Furthermore, *pfap2-g* transcript levels decreased severely in maturing E5ind gametocytes (day 4 or 7 gametocytes), similar to the *pfap2-g* expression pattern in wild-type parasites (*5*, *27*) but in contrast to endogenous *cam* transcripts (Fig. 2, D and F). Three features of the system can nonexclusively explain these observations: First, there are important differences between the sequence of the 5′*cam* used here (and widely used in malaria research to drive the expression of drug resistance markers) (*28*) and the endogenous *cam* upstream region, which can determine different promoter activity. Second, after rapamycin-induced recombination *pfap2-g* expression is under the control of two promoters in tandem, the intact endogenous *pfap2-g* promoter and the 5′*cam* promoter. Although the 5′*cam* promoter is located in a more proximal position (Fig. 1A), the distal *pfap2-g* promoter and the associated PfAP2-G autoregulatory positive feedback loop (*6*, *7*, *29*) may also contribute to determining *pfap2-g* expression. Transcripts containing sequences of the proximal part of the *pfap2-g* promoter or the 5′*cam* promoter occurred at similarly high levels (comparable to endogenous *cam* transcript levels), which may either reflect transcription from the *pfap2-g* promoter or the bidirectional activity of the 5′*cam* promoter (fig. S4A) (*28*). However, transcripts containing the *pfap2-g* coding sequence were far less abundant than transcripts containing promoter sequences, revealing a complex landscape suggestive of an important role for posttranscriptional mechanisms, or truncation of transcription immediately after the *loxP* site. Third, the *pfap2-g* locus heterochromatin environment may influence the activity of the 5′*cam* promoter. Regardless of the underlying mechanism, the roughly physiological *pfap2-g* transcripts levels and temporal dynamics in the inducible lines are a clear advantage of our system, as it resembles natural sexual conversion more faithfully than if the gene was markedly overexpressed.

To further explore the influence of reversible epigenetic states on *pfap2-g* activation, we maintained E5ind and 1.2Bind cultures without WR99210 pressure for 5 weeks and then induced conversion with rapamycin. In both parasite lines, conversion rates were clearly lower than in cultures under constant drug pressure (fig. S4B). Selecting back the cultures with WR99210 recovered conversion rates (upon induction) to the original levels (fig. S4C), consistent with conversion levels depending on reversible epigenetic states. These results suggest that in the absence of selection, heterochromatin restricts transcription at the *pfap2-g* locus, even after rapamycin-induced recombination. In contrast, drug pressure selects for parasites that have the 5′*cam-hdhfr* chimera in an active state, which determines *pfap2-g* expression upon recombination. Heterochromatin expansion and influence of the epigenetic state of one gene on its neighbors has been previously reported at other *P. falciparum* loci (*30*, *31*).

The few parasites that continue asexual growth after induction have heterochromatin at the endogenous *pfap2-g* promoter

In ~1 week, we could establish stably growing populations from the few parasites that did not convert upon induction and continued proliferating, which we termed E5ind+Rapa_prol and 1.2Bind+Rapa_prol. Diagnostic PCR analysis of genomic DNAs revealed that DiCre-mediated recombination had proceeded correctly in these parasites (Fig. 3A). Unexpectedly, new gametocytes were formed at each growth cycle in E5ind+Rapa_prol but not in 1.2Bind+Rapa_prol, with sexual conversion rates and *pfap2-g* transcript levels that resembled those in the respective wild-type E5 and 1.2B parental lines (Fig. 3B). This observation supports the idea that, despite the constitutive 5′*cam* promoter being immediately upstream of the *pfap2-g* coding sequence, the probability of *pfap2-g* activation is affected by the heterochromatin environment at the *pfap2-g* locus.

To further investigate this possibility, we performed H3K9me3 ChIP-seq analysis to characterize the distribution of heterochromatin

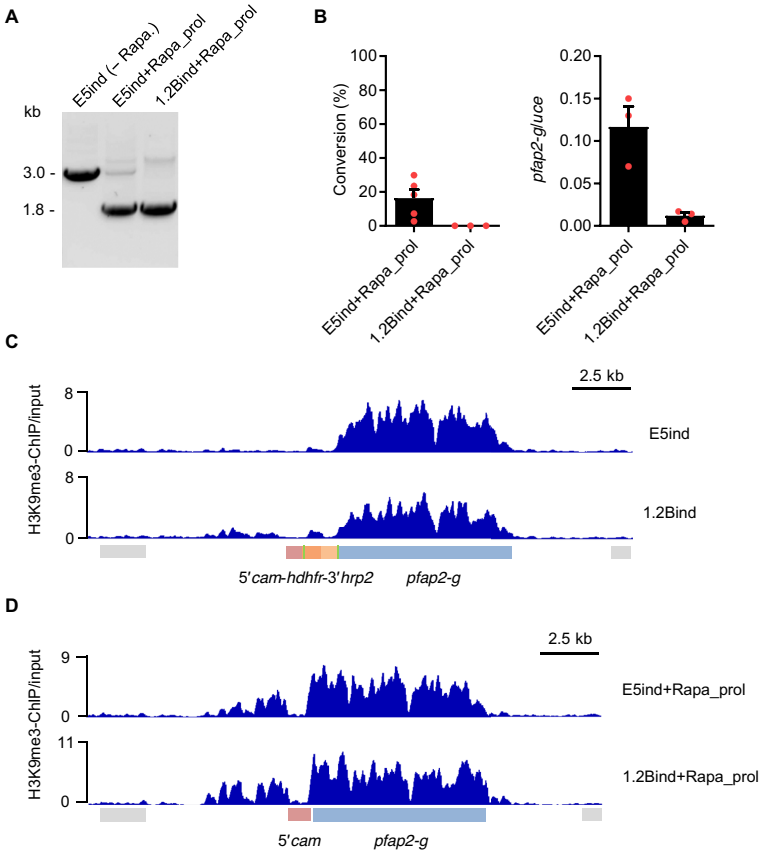


Fig. 3. Characterization of parasites that continue asexual growth after rapamycin-induced recombination. (**A**) Diagnostic PCR to assess recombination at the *pfap2-g* locus in parasites that continue proliferating after rapamycin induction. Genomic DNA was extracted 2 to 3 weeks after induction. (**B**) Sexual conversion rates (left) and analysis of *pfap2-g* relative transcript levels (right) in parasites that continue proliferating after rapamycin induction, determined 2 to 3 weeks after induction. Transcript levels are normalized against ubiquitin-conjugating enzyme (*uce*). Data are presented as the average and SEM of three biological replicates from independent inductions (except for conversion rates in E5ind+Rapa_prol, $N = 5$). (**C** and **D**) ChIP-Seq profiles of normalized H3K9me3 signal relative to input at the *pfap2-g* locus. The E5ind and 1.2Bind lines were analyzed before induction (C) or in parasites that continue proliferating after induction (D).

in the E5ind+Rapa_prol and 1.2Bind+Rapa_prol lines and also in the E5ind and 1.2Bind lines before rapamycin induction. While in wild-type asexual parasites heterochromatin spans the full coding sequence and about 3.5 kb of the upstream region (*32*), in the non-induced E5ind and 1.2Bind lines, H3K9me3 was present at the *pfap2-g* coding sequence but almost absent from the endogenous *pfap2-g* upstream region and the inserted construct (Fig. 3C). However, in parasites that failed to convert into gametocytes upon rapamycin-induced recombination (E5ind+Rapa_prol and 1.2Bind+Rapa_prol lines), both the endogenous *pfap2-g* upstream region and coding sequence were heterochromatic and only the 5′*cam* sequence was devoid of H3K9me3 (Fig. 3D). This result indicates that when E5ind and 1.2Bind cultures are maintained under WR99210 pressure, in the majority of parasites, the endogenous *pfap2-g* promoter remains in a euchromatic state that enables them to readily express *pfap2-g* and convert into gametocytes upon rapamycin-induced recombination. In a small subpopulation of parasites, which was larger in 1.2Bind (Fig. 3C), this promoter is in a heterochromatic conformation that prevents expression of *pfap2-g* even after recombination places the 5′*cam* promoter adjacent to the gene. Since these parasites continue proliferating, they are selected when the culture is maintained for several days after induction and exhibit a sexual conversion rate that likely depends on GDV1-mediated disruption of heterochromatin at the *pfap2-g* promoter (not occurring in 1.2B because GDV1 is truncated).

Characterization of mature-induced gametocytes

Both E5ind and 1.2Bind produced male and female mature gametocytes, with apparently normal morphology and at similar rates to the NF54 control line (fig. S5, A to D). However, because of defective

exflagellation of mature male gametocytes and consequently severely reduced ookinete formation, neither inducible line produced oocysts upon mosquito infection (Fig. 4, A and B, and fig. S5, E and F). The exflagellation defect in E5ind is not attributable to the conditional activation system, as the parental E5 line also failed to exflagellate (fig. S5G). IFA analysis during gametocyte activation suggests that deficient DNA replication underlies the failure of E5ind and 1.2Bind male gametocytes to exflagellate and infect mosquitoes (Fig. 4, C and D, and fig. S5H). The specific molecular defect in the E5 and 1.2B linages has not been identified. While this defect does not prevent gametocyte development or activation (including roundup and egress), it remains formally possible that it results in currently unknown alterations at stages preceding gamete formation.

Transcriptomic profiling of sexually committed schizonts and sexual rings

Our inducible sexual conversion system yields high amounts of synchronous parasites at the initial stages of sexual development, with a level of purity that was not achieved by previous approaches. This provides a unique opportunity to describe the largely uncharacterized sexually committed schizont and sexual ring stages. Here, we characterized the initial transcriptional changes upon *pfap2-g* activation using time course genome-wide transcriptomic analysis.

Overall, we identified 379 genes with altered expression between rapamycin-treated and control E5ind cultures [$\log_2$(fold change) of >1 in two independent biological replicates], of which 77 showed a $\log_2$(fold change) of >3 (Fig. 5, A and B, fig. S6A, and data file S1). Gene families such as *surfin* or *etramp* were enriched in up-regulated genes, whereas large families involved in antigenic variation were generally enriched in down-regulated genes (fig. S6B and data file S2). The majority of transcriptional differences were observed at the latest time point analyzed, which corresponds to the late sexual ring stage (~15 to 20 hpi of the cycle after induction) (Fig. 5C). At the early sexual ring stage (~5 to 10 hpi), the most up-regulated gene was the early sexual marker *gexp02* (Fig. 5, A and B) (*22*). Genes highly up-regulated upon *pfap2-g* activation included well-established early gametocyte markers such as *pfs16*, *pfg27/25*, and *pfg14-744* and many genes up-regulated early during *P. falciparum* sexual development identified by recent genome-wide studies (Fig. 5A) (*6*, *7*, *10*, *14*, *17*, *19*, *33*, *34*). Many of the most up-regulated genes (71% of genes with $\log_2$[fold increase] >3) were bound by PfAP2-G according to a recent ChIP-seq study (Fig. 5A) (*29*), whereas only 7% of the rest of genes in the genome were bound ($P = 2.2 \times 10^{-16}$ using Fisher's exact test). This result indicates that the majority of highly up-regulated genes are direct targets of PfAP2-G. In contrast, very few of the genes down-regulated upon *pfap2-g* activation are bound by PfAP2-G (5% of genes with a $\log_2$[fold-decrease] >3) or had been reported to change during sexual development (Fig. 5A). Down-regulated genes included genes with known functions associated with asexual development, such as *mesa*, *hrpIII*, *pfemp3*, *pf332*, and *kahrp*, and revealed many potential previously unidentified asexual parasite markers. The high level of purity of our sexual and control preparations (~90% sexual and 100% asexual, respectively) enables the identification of genes down-regulated during sexual commitment and development. This was not possible using previous approaches in which many asexual parasites were still present in the preparations enriched in sexual parasites. Only a study using fluorescence-activated cell sorting–sorted early sexual parasites (*17*) identified a substantial

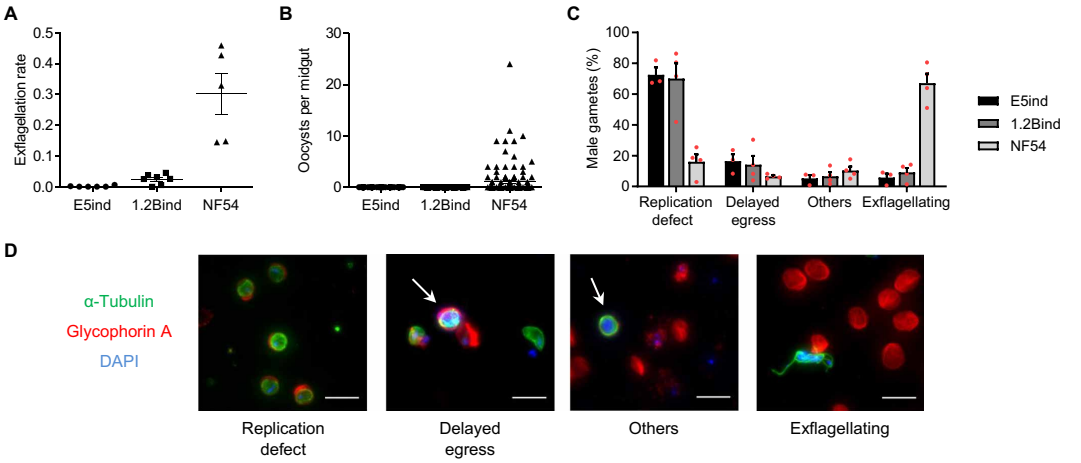


Fig. 4. Characterization of E5ind and 1.2Bind mature gametocytes. (**A**) Exflagellation rate (percentage of exflagellation normalized by gametocytemia) of mature gametocytes of the different parasite lines. Data are presented as the average and SEM of five to seven independent experiments. (**B**) Number of oocysts per mosquito midgut 9 days after blood feeding. Data are presented as average and SEM of the pooled values from three independent feeding experiments (E5, 117 midguts; 1.2B, 155 midguts; NF54, 137 midguts). (**C**) Distribution of male gamete types 25 min after activation. Male gametes were classified into four categories after IFA with anti–α-tubulin (stains microtubules, including the flagella) and anti–glycophorin A (an erythrocyte membrane marker) antibodies and 4′,6-diamidino-2-phenylindole (DAPI) staining of nuclei. Replication defect: Parasites that rounded up but had not replicated the genome (single nucleus). Delayed egress: Parasites that replicated the genome and formed axonemes but were still inside the red blood cells. Others: Other defects such as parasites lacking properly formed axonemes. Exflagellating: Parasites that developed correctly. Data are presented as the average and SEM of three or four independent experiments (E5ind, 389 male gametes; 1.2Bind, 769 male gametes; NF54, 406 male gametes). (**D**) Representative images of the categories described in (C). Scale bars, 10 μm.

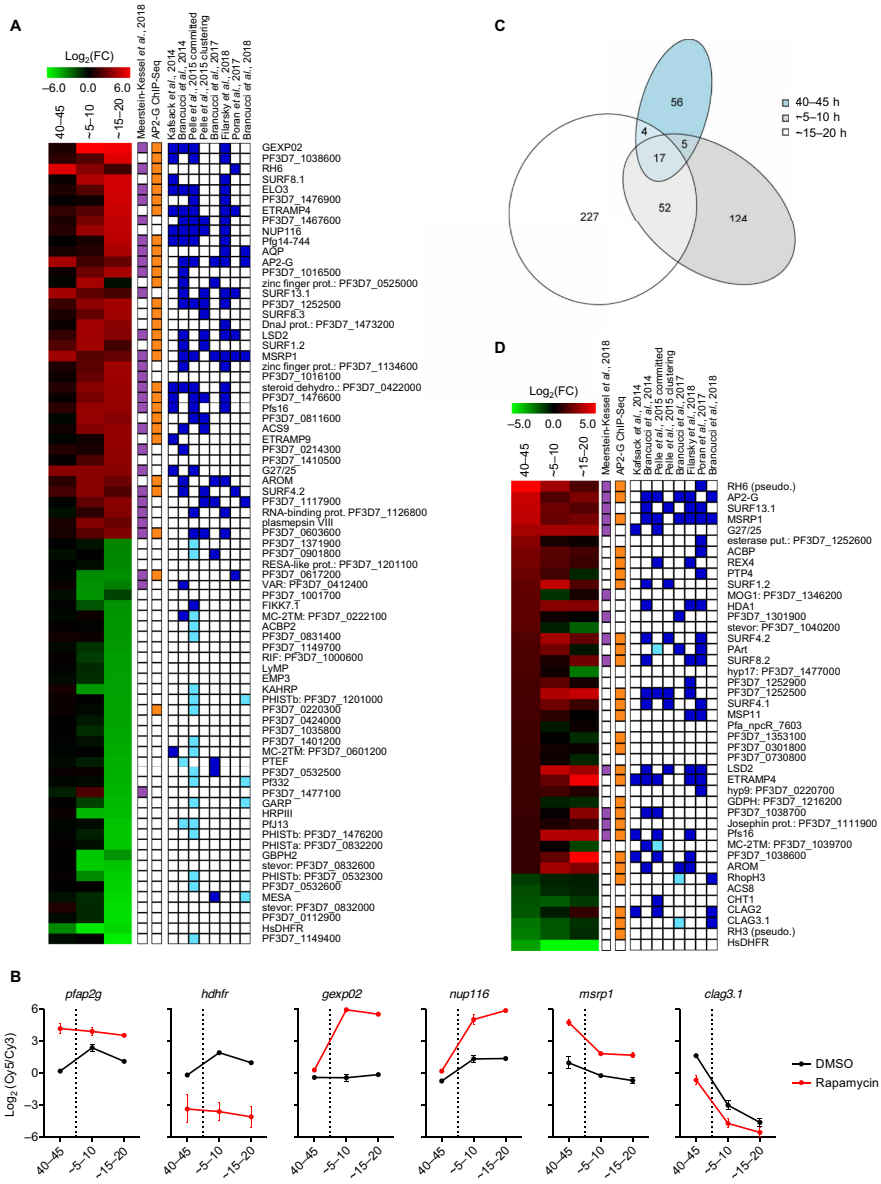


Fig. 5. Transcriptomic analysis of the alterations that ensue upon *pfap2-g* activation. (**A**) Expression profile of genes with a $\log_2$ of the expression fold change (FC) between induced and control cultures of >3 in two independent biological replicates (at any of the time points analyzed). Values are the average of the $\log_2$(FC) in the two replicates. Genes are ordered by FC value. Samples were collected at 40 to 45 hpi of the cycle of induction or ~5 to 10 and ~15 to 20 hpi of the next cycle. PlasmoDB IDs are provided for genes that do not have a unique annotation. Genes that were identified as gametocyte markers in an integrative analysis (purple), direct targets of PfAP2-G as determined by ChIP-seq (orange), and genes up-regulated (dark blue) or down-regulated (light blue) in gametocytes in previous studies are indicated. (**B**) Time course relative expression levels (normalized $\log_2$ ratio of Cy5-labeled samples signal versus Cy3-labeled reference pool signal) of selected genes in DMSO (control) and rapamycin-treated cultures. The dotted lines indicate the time of schizont rupture and reinvasion. (**C**) Venn diagram showing the number of genes with altered expression at each time point [$\log_2$(FC) > 1]. (**D**) Expression FC only for genes showing a $\log_2$(FC) > 1 at 40 to 45 hpi.

fraction of the most down-regulated genes in our analysis, which are genes that show decreased expression at the late sexual ring stage (Fig. 5A).

Next, we focused specifically on genes showing altered expression at the committed schizont stage (40 to 45 hpi of the induction cycle). Thirty-six genes were up-regulated in committed versus non-committed schizonts (log_2 [fold-change] >1 in two biological replicates), including several previously reported commitment markers and PfAP2-G targets, whereas only six genes were down-regulated in addition to the *hdhfr* marker (Fig. 5, B and D). Of note, three of these six genes (*clag3.1*, *clag2*, and *rhoph3*) encode components of the RhopH complex, which participates in erythrocyte invasion and solute transport (*35*–*37*).

A recent report showed that, in sexually committed schizonts, PfAP2-G binds the promoter of many invasion-related genes, which suggests that it may contribute to their regulation (*29*). We found that the expression of the majority of invasion genes bound by PfAP2-G was not altered between induced and control cultures (fig. S6C). We identified a very small number of differentially expressed invasion genes: Excluding pseudogenes, only *msrp1* and, to a lesser extent, *msp11* were up-regulated, whereas only components of the RhopH complex were down-regulated. Both up- and down-regulated invasion genes and pseudogenes are bound by PfAP2-G (fig. S6C), in contrast to genes down-regulated at the sexual ring stage that are generally not bound by PfAP2-G (Fig. 5A). Of note, the up-regulated pseudogene *rh6* is in close proximity to *msrp1*, which is suggestive of regional activation events possibly associated with chromosome topology (*38*).

The RhopH complex and solute transport are down-regulated during sexual development

To determine whether the expression of RhopH complex genes is reduced in wild-type committed cells, we analyzed their transcript levels in the NF54 line under basal conditions (sexual conversion rate, ~10%) and after stimulating conversion by choline depletion (~50% conversion) (*22*). Consistent with the results for the E5ind line, the expression of *clag3.1* and *rhoph3* was lower in stimulated cultures, whereas *rhoph2* and *rap1* transcript levels did not change (Fig. 6A). The modest magnitude of the differences observed in these experiments was expected given that choline-depleted cultures still contain many asexual parasites. Western blot analysis of E5ind schizont pellets and culture supernatants with antibodies against CLAG3 (cytoadherence linked asexual protein 3) revealed a clear reduction in CLAG3 levels in induced compared to control cultures (Fig. 6B), confirming that the lower transcript levels in committed schizonts translate into lower protein levels.

The RhopH complex is initially expressed in schizonts. After reinvasion and maturation to the trophozoite stage, it participates in the formation of the plasmodial surface anion channel (PSAC) (*36*), which determines the permeability of infected erythrocytes to multiple solutes, including sorbitol. To determine whether lower expression of RhopH components in sexually committed schizonts is associated with reduced PSAC activity, we performed sorbitol lysis experiments with E5ind cultures at the cycle following rapamycin induction. A 10-min treatment with 5% sorbitol produced osmotic lysis of all mature parasites (trophozoites and schizonts) in control E5ind cultures, whereas in induced cultures, the majority of parasites (stage I gametocytes) survived. The level of sorbitol survival also correlated with the level of sexual conversion in the independent parasite line E5-PfAP2-G-DD, which has a destabilization domain appended to PfAP2-G to modulate sexual commitment (Fig. 6C) (*6*). These results are consistent with previous reports showing that gametocytes are more resistant to sorbitol lysis than late stage asexual parasites (*39*). We also characterized PSAC function by assessing the uptake of 5-aminolevulinic acid (5-ALA), which requires functional PSAC (*37*). As expected, the proportion of parasites incorporating the compound was much lower in induced than in control cultures (Fig. 6D). Together, these results confirm that the expression of some RhopH complex components is reduced in sexually committed schizonts, resulting in lower PSAC activity in sexual parasites.

Sexually committed merozoites do not have a preference for reticulocyte invasion

The high purity of the E5ind committed schizont preparations enables the phenotypic characterization of parasites at this stage. Differential ligand expression in committed schizonts and merozoites was previously proposed as a mechanism underlying homing of sexual parasites to the bone marrow or enhancing invasion of reticulocytes, which are abundant in this tissue (*29*, *40*). To test whether committed merozoites preferentially invade reticulocytes, we compared the relative invasion of reticulocytes and erythrocytes by Percoll-purified schizonts from induced and control cultures. As a source of reticulocytes and erythrocytes, we used reticulocyte-enriched cord blood. Flow cytometry analysis revealed a similar distribution of new rings in reticulocytes (CD71-positive) and erythrocytes (CD71-negative) between induced and control cultures (Fig. 6, E and F). While we cannot completely exclude the possibility that reticulocytes from other sources may be preferentially invaded by sexually committed merozoites, this result suggests that committed merozoites do not have a specific tropism for reticulocytes.

DISCUSSION

Investigations of sexually committed schizonts and sexual rings have been hampered by the relatively low abundance of these stages in mixed cultures and the difficulty in achieving efficient separation from their asexual counterparts. Here, we describe a conditional activation system for *pfap2-g* that can be used to induce massive synchronous sexual conversion. The high efficiency of the system, with asexually growing parasites almost absent after induction, enables the characterization of committed schizonts and early sexual stages without any further purification step. We demonstrate the utility of the system by providing a detailed characterization of the transcriptome of these stages and showing its suitability for phenotypic characterization. The system can also be used to characterize early sexual parasites at any other level (e.g., drug susceptibility screens, nascent transcript measurements, and epigenomic, metabolomic, or proteomic analysis). We also used the system to show that *pfap2-g* activation is sufficient to trigger productive sexual conversion, even in parasite lines with defects in GDV1. The same approach used here to generate the E5ind and 1.2Bind lines could be applied in the future on the background of a parasite line competent for mosquito infection (e.g., NF54). However, while obtaining parasite preparations highly enriched in the initial stages of sexual development was not possible until now, efficient methods to obtain pure preparations of mature gametocytes in large amounts are already available (*41*, *42*). Thus, we consider that our system would have limited utility for the study of mature gametocytes.

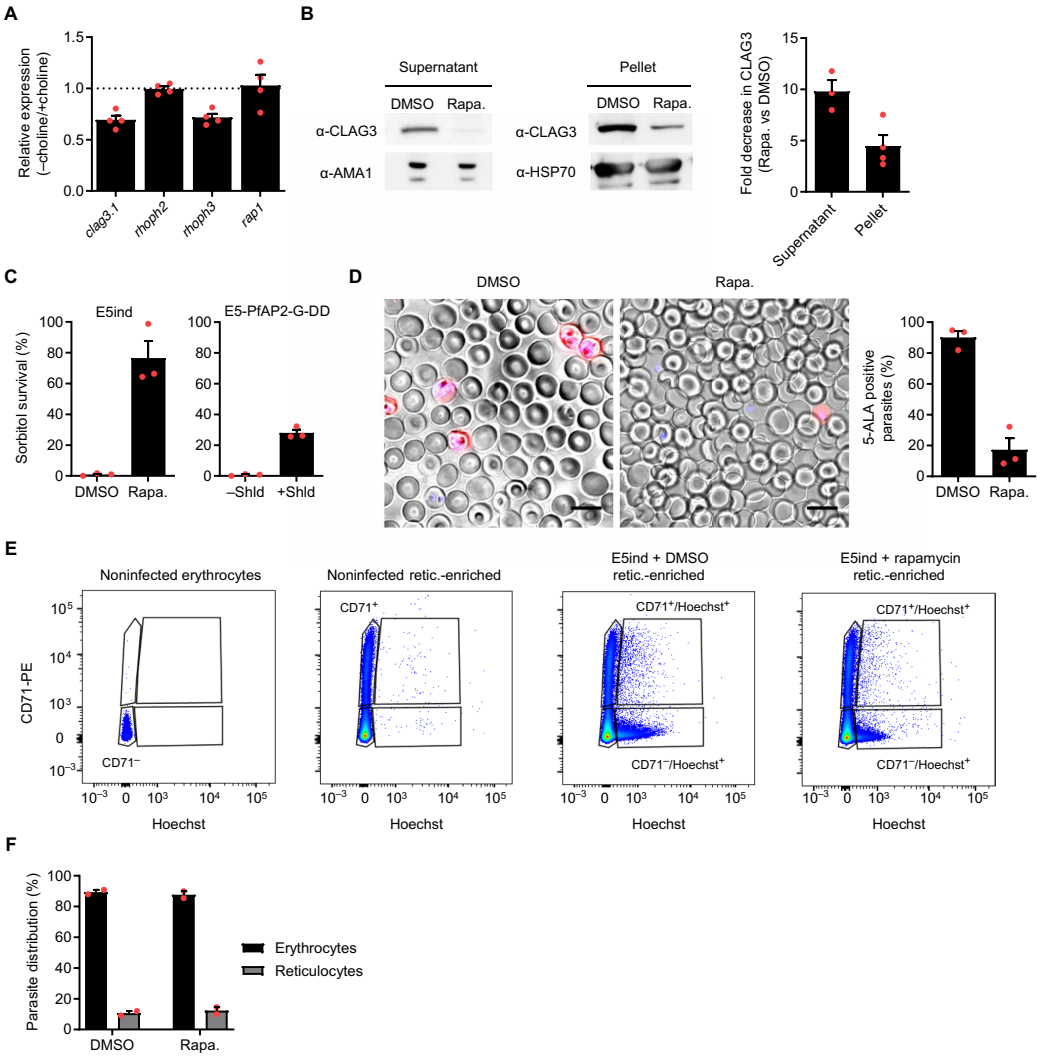


Fig. 6. Functional characterization of sexually committed schizonts and early sexual stages. (**A**) Expression of RhopH complex genes and *rap1* (control) in committed schizonts of the NF54 line. Values are the FC between inducing (−choline) and noninducing (+choline) conditions. Transcript levels were normalized against *rama*. In all panels, data are presented as the average and SEM of independent biological replicates. $N = 4$. (**B**) Western blot analysis of E5ind culture supernatant or mature schizont pellets with antibodies against CLAG3 and the loading controls AMA1 or HSP70. The bar chart shows the fold decrease in normalized band intensity between induced and noninduced cultures. $N = 3$ (supernatant) or 4 (pellet). (**C**) Sorbitol resistance assay performed ~30 to 40 hpi at the cycle after induction. Values are the percentage of pigmented parasites surviving sorbitol treatment. Shield 1 (Shld) stabilizes PfAP2-G in E5-PfAP2-G-DD cultures. $N = 3$. (**D**) Uptake of 5-ALA in induced and non-induced E5ind cultures. Representative fluorescence microscopy images and quantification of the proportion of positive parasites are shown. $N = 3$. Scale bars, 10 μm. (**E**) Flow cytometry analysis of invasion preference by committed and noncommitted merozoites. Reticulocytes were identified by CD71 signal and parasites by Hoechst signal. Representative dot plots are shown. PE, phycoerythrin. (**F**) Distribution of parasites between erythrocytes and reticulocytes after invasion by committed and noncommitted merozoites. $N = 2$.

Previous systems for inducible sexual conversion in *P. falciparum* achieved sexual conversion rates of 30 to 60% (*5*, *6*, *10*, *14*, *19*), as the expression of the master regulator *pfap2-g* was not directly controlled. With a consistent sexual conversion rate of ~90% in the E5ind line and the majority of the remaining ~10% parasites failing to develop rather than growing asexually, the system presented here stands out as the method that yields highest purity *P. falciparum* early sexual parasite preparations. In induced parasites, the 5′*cam* promoter is adjacent to the *pfap2-g* coding sequence, resulting in activation of the gene in the vast majority of parasites. However, the temporal dynamics of expression and transcript levels of *pfap2-g* resemble endogenous *pfap2-g* expression in wild-type parasites rather than expression of the endogenous *cam* gene. While we were unable to fully clarify the underlying mechanism for this unexpected observation, physiological *pfap2-g* expression is a beneficial feature of the system that favors the correct formation of gametocytes and dispensed the need to screen multiple promoters. Conditional activation of sexual conversion has also been described in *P. berghei*. Using an approach similar to the one described here, based on conditional activation of *ap2-g* upon DiCre-mediated promoter flipping, conversion of the majority of parasites was achieved (*15*).

Together with single-cell transcriptomic approaches, purification of sexual parasites, and comparative analysis of gametocyte producing and nonproducing lines (*7*, *17*, *27*, *33*, *43*, *44*), previously reported inducible systems were instrumental to gain insight into the transcriptome of the initial stages of sexual development. However, the high conversion rate obtained with our inducible system provides important advantages for an accurate transcriptomic profiling of these stages. While the majority of genes up-regulated in committed or early sexual parasites had been previously identified by other approaches, the high level of purity of our committed schizont and sexual ring preparations also enabled the identification of down-regulated genes. In previous studies, down-regulation was probably masked by transcripts arising from abundant asexual parasites present in the bulk population. Our approach also provides insight into the temporal order of events in the sexual commitment regulatory cascade: Genes proposed to be altered upstream of PfAP2-G activation, such as the ISWI and SNF2L helicases or GDV1 (*7*, *14*), were not up-regulated upon PfAP2-G activation, supporting the idea that they operate upstream of PfAP2-G. Our results also clearly demonstrate that, while expression of some genes involved in sexual commitment and development may occur independently of PfAP2-G (*18*, *45*), activation of this transcription factor is sufficient to drive the full process. However, characterization of the transcriptional changes occurring immediately upon PfAP2-G activation with high temporal resolution was not possible because of the intrinsic time required for rapamycin-induced recombination with the DiCre system (*24*).

We show that a relatively small number of genes have altered expression at the committed schizont stage, and a broad remodeling of the transcriptome does not occur until the sexual ring stage. Very few genes involved in erythrocyte invasion showed altered expression in sexually committed versus noncommitted schizonts. Apart from pseudogenes, *msrp1* and genes encoding components of the RhopH complex were the only clearly up- and down-regulated invasion-related genes, respectively. Up-regulation of *msrp1* in committed schizonts, of still unknown functional significance, has been consistently observed in several studies (*7*, *10*, *14*, *17*, *19*, *29*, *33*). In contrast, the down-regulation of genes encoding components of the RhopH complex was not previously identified. A previous study reported up-regulation of some components of the complex (*33*). While the reasons for the discrepancy are unclear, our observations are consistent with the known reduction in PSAC function and sorbitol sensitivity of gametocytes (*39*), which we confirmed. Our results suggest that committed schizonts express RhopH components at reduced levels to adjust the permeability of sexually developing parasites upon reinvasion, possibly to adapt to lower nutrient needs or to prevent the entry of toxic compounds (*46*). Of note, genes that are up-regulated and genes that are down-regulated in committed schizonts (but not at other stages) are bound by PfAP2-G. Interactions between PfAP2-G and another ApiAP2 transcription factor, PfAP2-I (*47*), have been proposed to underlie the regulation of invasion genes in committed schizonts (*29*). Complex interactions between the two factors may determine the transcriptional outcome of PfAP2-G binding, ranging from activation for the majority of its sexual ring and gametocyte-stage targets to no transcriptional change for many invasion-related genes, and even reduced expression for genes involved in solute transport such as *rhoph3* and *clag* genes.

The insertion of a constitutive promoter to drive *pfap2-g* expression also provided insight into the regulation of *pfap2-g* itself. In cultures maintained in the absence of drug pressure, sexual conversion rates were lower, and this was reverted upon reselection with the drug. This is suggestive of reversible heterochromatin expansion from neighbor regions influencing the activation of *pfap2-g* in the inducible system. Furthermore, in uninduced E5ind and 1.2Bind cultures under constant WR99210 pressure, heterochromatin is present at the *pfap2-g* coding sequence but not at the promoter, roughly matching the heterochromatin distribution in gametocytes (*32*). In these parasites, the gene is readily activated upon rapamycin-induced recombination, indicating that heterochromatin at the *pfap2-g* coding sequence is compatible with active expression of the gene. In contrast, we found that the few parasites that escaped conversion after rapamycin-induced recombination had the endogenous *pfap2-g* promoter in a heterochromatic state, which was associated with restricted activation of the gene and led to sexual conversion rates that reflected those observed in the respective parental lines. Together, these observations indicate that, in the inducible lines, the constitutive promoter adjacent to the gene, the endogenous promoter and the locus chromatin environment all contribute to *pfap2-g* regulation. We propose a model for endogenous *pfap2-g* regulation whereby the coding sequence is constitutively heterochromatic at all stages of the life cycle, including sexually committed and early sexual stages in which the gene is actively expressed. This may serve as a permanent "reservoir" from which heterochromatin spreads to the promoter region at stages at which the gene needs to be epigenetically silenced, including asexual blood stages and mosquito stages (*32*, *48*). In this scenario, activation of *pfap2-g* expression in committed parasites requires dismantling heterochromatin only at the promoter region.

A possible additional layer of regulation of *pfap2-g* expression in gametocytes of the inducible lines, operating at the posttranscriptional level, is transcript stability. Of note, altering the activity of the exonuclease PfRNase II resulted in up-regulation of PfAP2-G, among other sexual markers (*49*). This raises the intriguing possibility that this exonuclease, involved in *var* gene regulation, also regulates *pfap2-g* expression. In support of this idea, a recent transcriptomic time course analysis of gametocyte development revealed that the expression of the *pfrnase II* gene peaks at stage I

of gametocyte development, coinciding with the drop in *pfap2-g* transcripts (*27*).

In summary, we have developed inducible *P. falciparum* transgenic lines that show the highest level of sexual conversion reported so far. Using these lines, we provide new insight into the regulation of *pfap2-g* and the transcriptomic and functional changes that occur during the initial steps of sexual development. These inducible lines will be instrumental to gain a deeper understanding of the formation of sexual stages, which is expected to facilitate the design of new tools to block malaria transmission.

MATERIALS AND METHODS

Parasite cultures and induction of sexual conversion

The 3D7 subclones E5 (derived from the 3D7-B stock) and 1.2B (derived from the 3D7-A stock), the E5-PfAP2-G-DD line, and the NF54 line have been previously described and characterized (*6*, *26*, *41*). Parasites were cultured in B^+ erythrocytes at 3% hematocrit under standard conditions, with RPMI 1640–based culture medium. For 3D7-derived cultures, we used media supplemented with 0.5% Albumax II (Invitrogen). NF54 cultures were grown in media supplemented with 0.25% Albumax II and 5% human serum (*41*). To produce gametocytes for mosquito infection, a 4% hematocrit was used. Cultures were regularly synchronized by sorbitol lysis to eliminate late asexual stages (trophozoites and schizonts). For some experiments, cultures were tightly synchronized to a 5-hour age window by purification of schizont stages using Percoll gradients (63% Percoll) followed by sorbitol lysis 5 hours later. The PKG inhibitor ML10 was used at a concentration of 80 nM to inhibit merozoite egress (*25*). AquaShield-1 (Cheminpharma) was used at a concentration of 0.5 μM to stabilize PfAP2-G in the E5-PfAP2-G-DD line.

To induce sexual conversion, E5ind or 1.2Bind cultures were sorbitol-synchronized and, after synchronization, maintained without WR99210 (Jacobus Pharmaceutical Co., USA). Typically at ~20 hours after synchronization, cultures were treated with 10 nM rapamycin (Sigma-Aldrich, R0395) or DMSO solvent (control) for 1 hour (*24*), which was followed by one wash with incomplete culture media (culture media without Albumax II) before placing back in culture. After reinvasion, cells were grown in serum-supplemented media, as it improved gametocyte development. Sexual conversion rates were measured by treating cultures at the ring stage after reinvasion (day 0) with 50 mM *N*-acetylglucosamine (Sigma-Aldrich, A3286) for 5 days to eliminate asexual parasites. The sexual conversion rate was calculated as the gametocytemia at day 5 relative to the initial rings parasitemia at day 0 (*5*). Parasitemia and gametocytemia were measured by light microscopy quantification of Giemsa-stained smears.

To stimulate sexual conversion in the NF54 line, we used the choline depletion method (*14*, *19*). Cultures were regularly maintained in RPMI 1640-based culture medium with 0.5% Albumax and 2 mM choline (Sigma-Aldrich, C7527), and choline was removed at the ring stage to stimulate conversion, as previously described (*22*).

Plasmids

The pL7-Ind-ap2g plasmid was derived from the pL6-egfp-yfcu plasmid (*20*). The *yfcu* cassette was removed using restriction sites Not I and Sac II. Next, a *loxP* site was cloned into a BamH I site between the 5′*cam* and the *hdhfr* gene using two annealed and phosphorylated oligonucleotides (p1 and p2) containing the *loxP* sequence. The second *loxP* site was cloned together with the *pfap2-g* homology region 2 (HR2) between the 3′ *hrp2* and HR2. The *loxP* sequence was added to primer p3, which was used together with p4 to amplify the HR2 (positions −73 to +311 bp relative to the *pfap2-g* start codon). The *pfap2-g* HR1 (positions −449 to −160 bp) was amplified with primers p5 and p6. HR1 and HR2 were cloned into Spe I/Afl II and EcoR I/Nco I sites, respectively. Last, to clone the guide, the plasmid was digested with BtgZ I, and two annealed oligonucleotides (p7 and p8) containing the guide sequence (located at −139 to −158 bp) were cloned into this site using the In-Fusion HD Cloning Kit (Clontech).

Plasmid pHH1-cambsd-DiCre was based on plasmids pHH1_SERA5del3DC (*21*) and E140-0 (*50*). The *bsd* coding sequence was amplified with primers p9 and p10 and cloned into E140-0 using BamH I and Xho I restriction sites, replacing the *hdhfr* coding sequence and removing unnecessary plasmid elements. After generating Afl II and Spe I restriction sites upstream of the *bsd* cassette with a PCR-amplified (primers p11 and p12) fragment of the plasmid recloned into Hind III and EcoR I sites, the DiCre cassette obtained by Afl II and Spe I digestion of pHH1SERA5del3DC was cloned into the new Afl II and Spe I sites.

Plasmid pHHI-DiCre-lisp1, which contains the DiCre expression cassette flanked by *lisp1* HRs, was derived from pHH1-cambsd-DiCre. First, the HR2 (positions +5891 to +6235 bp relative to the *lisp1* start codon) was PCR-amplified using primers p13 and p14 and cloned into Spe I and Not I sites in pHHI-cambsd-DiCre. The HR1 (positions +5169 to +5523 bp) was amplified using primers p15 and p16 and cloned into an Afl II site using the In-Fusion kit system. To generate plasmid pDC2-Cas9-hDHFRyFCU-lisp1, a guide covering positions +5526 to +5545 bp from the *lisp1* start codon was prepared by annealing oligonucleotides p17 and p18 and cloning into a Bbs I site of plasmid pDC2-Cas9-hDHFRyFCU (*24*) using the In-Fusion system.

Guides were designed using the EuPaDGT Web-based tool. PCR amplifications from *P. falciparum* genomic DNA were performed using LA Taq DNA Polymerase (Takara). To clone the plasmids, we used *Escherichia coli* DH5α or ultracompetent MAX Efficiency DH5α (Invitrogen) for difficult cloning. Oligonucleotides were from Integrated DNA Technologies. All primers and oligonucleotides are described in table S1.

Generation of transgenic lines

All transfections were performed by electroporation of cultures at the ring stage. For the inducible line with episomal expression of DiCre, E5 cultures were transfected with 60 μg of pUF1-Cas9-ydhodh (*20*) and 12 μg of pL7-Ind-ap2g linearized using a Sca I site (located in the plasmid backbone). Cultures were permanently maintained under 10 nM WR99210 pressure. After confirming that the *pfap2-g* locus was correctly edited, cultures were transfected with 100 μg of the pHHI-cambsd-DiCre plasmid and selected continuously with blasticidin S (2.5 μg/ml) (Thermo Fisher Scientific, R21001). For approaches involving stable integration of the DiCre expression cassette, we used a three-plasmid strategy. Sixty micrograms of pDC2-Cas9-hDHFRyFCU-lisp1, 12 μg of pHHI-DiCre-lisp1, and 12 μg of pL7-Ind-ap2g, the latter two linearized using a Sca I site (located in the plasmid backbone), were cotransfected and selected with 10 nM WR99210 for 4 days as previously described (*24*). Correct edition of the two loci using this strategy was achieved only for the E5 line. For the 1.2B line, we had to use a sequential

editing strategy: Cultures were first transfected with 60 µg of pDC2-Cas9-hDHFRyFCU-lisp1 and 12 µg of linearized pHHI-DiCre-lisp1 plasmids and selected with 10 nM WR99210 for 4 days. After subcloning and treatment with 1 µM 5-fluorocytosine (clinical grade Ancotil, Mylan N.V.) for 2 weeks to eliminate parasites that maintain episomal copies of the pDC2-Cas9-hDHFRyFCU-lisp1 plasmid, a second round of transfection was performed with 60 µg of pUF1-Cas9-ydhodh and 12 µg of linearized pL7-Ind-ap2g plasmids, followed by continuous selection with 10 nM WR99210.

Diagnostic PCR to assess the correct integration of the transgenes was performed using the LA Taq DNA Polymerase (Takara). The primers used for diagnostic PCR are described in table S1, and their relative positions are shown in Fig. 1.

RNA extraction and transcriptional analysis by reverse transcriptase quantitative PCR

For the majority of samples, RNA was extracted using the TRIzol method, deoxyribonuclease-treated and purified with a protocol optimized for low amounts of RNA (*51*), and reverse-transcribed using the AMV Reverse Transcription Kit (Promega) with a mixture of oligo (dT) and random primers. Transcript abundance was measured by real-time quantitative PCR (qPCR) using the standard curve method in a 7900HT Fast Real-Time PCR System and the Power SYBR Green Master Mix (both from Applied Biosystems). For samples from stage V gametocytes, RNA collected in TRIzol was extracted with the Direct-zol RNA MiniPrep Kit (Zymo Research), complementary DNA (cDNA) synthesis performed using the iScript cDNA Synthesis Kit (Bio-Rad), and qPCR analysis of male and female markers (*52*) was performed using the standard curve method in a LightCycler 480 Instrument II (Roche) with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad). For the standard curve, we used genomic DNA from the same parasite line being analyzed. Transcript levels of the ubiquitin-conjugating enzyme (*uce*; ID: PF3D7_0812600), serine-tRNA ligase (*serrs*; ID: PF3D7_0717700), or rhoptry-associated membrane antigen (*rama*; ID: PF3D7_0707300) were used for normalization as indicated. All primers used for qPCR analysis are described in table S1. For the analysis of *pfap2-g* transcript levels in Figs. 2 and 3, we used primers p23 and p24.

IFAs and analysis of 5-ALA uptake

For the analysis of asexual parasites and immature gametocytes, IFA analysis was performed on paraformaldehyde (PFA)–fixed smears essentially as described (*5*). For the analysis of stage V gametocytes or activated gametes, samples were fixed with 4% PFA for 15 min and incubated overnight on top of glass coverslips previously coated with poly-L-lysine (Merck Millipore), followed by permeabilization, blocking, and antibody incubations. To identify ookinetes in mosquito midguts (24 hours after blood feeding), air-dried blood drops from single midguts were PFA-fixed and analyzed by IFA. The primary antibodies used were mouse anti-Pfs16 (1:400 to 1:2000; 32F717:B02, a gift from R. Sauerwein, Radboud University), mouse anti–α-tubulin (1:700; DM1A; Sigma-Aldrich, T6199), rabbit anti–glycophorin A (1:200; EPR8200; Abcam, ab129024), rabbit anti-Pfg377 (1:1000; batch 6809; a gift from L. Ranford-Cartwright, University of Glasgow), and mouse anti-Pfs25 clone 4B7 (1:3000; MRA-315, BEI Resources) conjugated to Cy3 (GE Healthcare) (*41*). The secondary antibodies were goat anti-mouse immunoglobulin G (IgG)–Alexa Fluor 488 (1:500 to 1:1000; Thermo Fisher Scientific, A11029) and donkey anti-rabbit IgG–Alexa Fluor 594 (1:500; Thermo Fisher Scientific, R37119). Preparations were visualized with an Olympus IX51 epifluorescence microscope, and images were acquired with an Olympus DP72 camera using cellSens Standard 1.11 software or, for experiments with gametes, a Nikon Ti-E widefield microscope with ×60 to 100 objective lens and captured in 0.3-µm slice z-stack images, which were converted to maximum intensity projections in NIS Elements v4.20. All further image processing was performed using ImageJ.

5-ALA uptake was determined as previously described (*37*, *46*). Synchronized E5ind cultures were treated with rapamycin or DMSO and, after reinvasion, when cultures were at the ring stage, 200 µM 5-ALA (Sigma-Aldrich, A3785) was added. Inside of infected erythrocytes, 5-ALA is converted to fluorescent protoporphyrin IX. Uptake was determined at ~30 to 40 hpi after staining nuclei with Hoechst (2 µg/ml) for 10 min at 37°C in an eight-well chamber slide for live cell fluorescence microscopy. Images were acquired using an Olympus IX51 epifluorescence microscope and analyzed using ImageJ.

Chromatin immunoprecipitation sequencing

Chromatin extraction from cultures at the late trophozoite/schizont stage was performed as previously described (*30*), with minor modifications. Briefly, after the cross-linking and washing steps, the MAGnify Chromatin Immunoprecipitation System (Life Technologies) was used. Samples were sonicated using an M220 sonicator (Covaris) at 10% duty factor, 200 cycles per burst, 140 W of peak incident power for 10 min. Immunoprecipitations were performed overnight at 4°C with 4 µg of chromatin and 8 µg of antibodies against H3K9me3 (Diagenode, C15410193) previously coupled to protein A/G magnetic beads provided in the kit. Washing, decross-linking, and elution were performed following the MAGnify ChIP System recommendations but avoiding high temperatures that may result in denaturation of extremely AT-rich intergenic regions: De-crosslinking, proteinase K treatment, and elution were performed at 45°C (for 2 hours, overnight, and 1.5 hours, respectively).

Libraries for Illumina sequencing were prepared from 5 ng of immunoprecipitated DNA using a protocol adapted to a genome with an extremely high AT richness (*53*). Briefly, after end repair and addition of 3′ A-overhangs, NEBNext Multiplex Oligos for Illumina (NEB, E7335 and E7500) were ligated. Purification steps were performed with Agencourt AMPure XP beads (Beckman Coulter). Libraries were amplified using the KAPA HiFi PCR Kit (Kapa Biosystems) in KAPA HiFi Fidelity Buffer (5×) with the following conditions: 95°C for 3 min, nine cycles at 98°C for 20 s and 62°C for 2.5 min, and 62°C for 5 min. Amplified libraries were purified using 0.9× AMPure XP beads to remove adapter dimers. The library size was analyzed in a 4200 TapeStation System (Agilent Technologies). We obtained 12 to 26 million 125 bp paired-end reads per sample using a HiSeq2500 System (Illumina).

Transcriptomic analysis using microarrays

Time course transcriptomic analysis was performed using tightly synchronized cultures induced at 25 to 30 hpi. Samples for transcriptomic analysis were collected 15, 28, and 38 hours after induction (corresponding to 40 to 45 hpi schizonts of the first cycle and ~5 to 10 or ~15 to 20 hpi rings of the next cycle, respectively). RNA was purified using the TRIzol method, and cDNA was synthesized, purified, and labeled essentially as described (*26*, *54*). Samples

were analyzed on two-color long oligonucleotide-based custom Agilent microarrays. The microarray design was based on Agilent design AMADID #037237 (*54*), modified by adding new probes for genes lacking unique probes and for some noncoding RNAs and reporter genes (new designs: AMADID-084561 and AMADID-085763; data file S3). All samples, labeled with Cy5, were hybridized against a reference pool labeled with Cy3, consisting of a mixture of cDNA from rings, trophozoites, and schizonts of E5 and E5ind in equal amounts. Microarray hybridization and image acquisition were performed as described using a Microarray Scanner (no. G2505C, Agilent Technologies) located in a low ozone hood (*54*).

Bioinformatic analysis

For the analysis of ChIP-seq data, we first removed repetitive *k*-mers from the reads using BBDuK (v36.99) with parameters *ktrim* = r, *k* = 22, and *mink* = 6. After this initial filtering, reads where aligned to modified 3D7 reference genomes reflecting the alterations introduced by CRISPR-Cas9 genome editing and rapamycin-induced recombination (i.e., integration of the DiCre expression cassette into the *lisp1* locus, integration of the conditional activation cassette upstream of the *pfap2-g* coding sequence, and recombination between *loxP* sites in rapamycin-treated samples). Reads alignment was performed using Bowtie (v1.2) with parameters *--very-sensitive* and *--local*, trimming 4 bp from both the 5′ and 3′ ends and restricting fragment size to 50 to 200 bp. After aligning, duplicate reads where removed using PicardTools MarkDuplicates (v2.9.4). For each sample, per-base coverage was calculated using BEDtools (v2.27.1) and divided by the number of million reads (RPM). For each base, H3K9me3 enrichment was calculated as ChIP (RPM)/input (RPM) with a pseudocount of 0.1 to avoid division by 0. Data were visualized using IGV (Integrative Genomics Viewer; v2.4.10).

Initial processing of microarray data, including normalization, was performed using Feature Extraction software (Agilent) with default options. The next steps of the analysis were performed using Bioconductor in an R environment (R version 3.5.3). For each individual microarray, we calculated Cy3 and Cy5 background signals as the median of the 100 lowest signal probes for each channel. Probes with both Cy3 and Cy5 signals below three times the array background were excluded. Gene level $\log_2$(Cy5/Cy3) values and statistical estimation of parasite age were computed as described (*26*). Genes with missing values or genes that in all samples had expression values within the lowest 20th percentile of intensity (Cy5 channel) were excluded from downstream analysis, including identification of differentially expressed genes, because differences in genes expressed at near-background levels are of low confidence. Heatmaps were generated using TMEV 4.9. Gene set enrichment analysis (GSEA) with gene lists corresponding to gene families (data file S2) was performed using GSEA preranked.

Western blot

Sample preparation for Western blot analysis was performed essentially as previously described (*50*). Briefly, Percoll-purified schizonts were placed in culture at a 0.3% hematocrit, and, after 13 to 20 hours, culture supernatants were harvested and mixed with an equal volume of 2× SDS–polyacrylamide gel electrophoresis (PAGE) protein loading buffer. Schizont extracts were prepared by resuspending Percoll-purified schizonts into 20 pellet volumes of phosphate-buffered saline (PBS) and adding an equal volume of 2× SDS-PAGE protein loading buffer. Samples were heated for 5 min at 95°C before storing at −80°C. For SDS-PAGE, β-mercaptoethanol was added to a 4% final concentration, and, after boiling again for 5 min at 95°C, proteins were resolved in 4 to 8% SDS-PAGE, transferred to nitrocellulose membranes, incubated with antibodies, and washed following standard procedures. The primary antibodies used were mouse anti-CLAG3 polyclonal antibodies (*36*) at 1:2000 (167#2, recognize both CLAG3.1 and CLAG3.2; a gift from S. A. Desai, NIAID-NIH, USA), rabbit anti-3D7 AMA1 polyclonal antibodies at 1:2000 (a gift from R. F. Anders, La Trobe University, Australia), and rabbit anti-PfHSP70 polyclonal antibodies at 1:10,000 (StressMarq Biosciences, SPC-186C, lot no. 1007). As secondary antibody, we used anti-mouse-horseradish peroxidase (HRP) (Sigma-Aldrich, A9044) or anti-rabbit-HRP (Sigma-Aldrich, A6154) at 1:5000. HRP signal was detected using the Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific), and membranes were imaged in a LAS4000 system. Band quantification was performed using ImageJ.

Reticulocyte invasion assay

Human cord blood samples were obtained from the Barcelona Blood and Tissue Bank (www.bancsang.net/), following approval for the protocol and informed consent by the Clinical Research Ethics Committee of Vall d'Hebron University Hospital [PR(CS)236/2017]. Human reticulocyte preparations were performed as previously described (*55*), with some modifications. Briefly, samples were centrifuged at 1000*g* for 15 min (acceleration = 6, deceleration = 2), and plasma was removed. Cell pellets were washed with incomplete culture media at 500*g* for 10 min and resuspended at 50% hematocrit with incomplete culture media. Five milliliters of resuspended blood was carefully loaded on 6 ml of 70% Percoll (GE Healthcare) and centrifuged at 1200*g* for 15 min at 20°C (acceleration = 4, deceleration = 0). Reticulocytes concentrated in the Percoll interface were collected and washed three times with incomplete culture media (centrifugation at 400*g* for 5 min). To deplete leukocytes, reticulocyte-enriched pellets were resuspended to a final volume of 900 μl with incomplete culture media and incubated with 100 μl of CD45 MicroBeads (Miltenyi Biotec, 130-045-801) for 18 min at 4°C, mixing every 3 min. The cells were washed with incomplete culture media and passed through LS Columns (Miltenyi Biotec, 130-042-401) placed on a magnetic stand. Eluents, containing reticulocyte-enriched fractions, were washed twice with incomplete culture media, and reticulocyte enrichment was quantified by incubating for 15 min at room temperature with anti–CD71–phycoerythrin (PE) (Miltenyi Biotec, clone AC102, 130-099-219) at a final dilution of 1:400. Reticulocyte quantification was performed using a BD LSRFortessa flow cytometer.

E5ind schizonts treated with rapamycin or DMSO were purified in 70% Percoll gradients and mixed with reticulocyte-enriched or control regular erythrocyte preparations to achieve a ~1% parasitemia. After establishing 3% hematocrit cultures, parasites were allowed to reinvade overnight and analyzed the following day by flow cytometry. Samples were washed in PBS and incubated for 15 min at room temperature with anti–CD71-PE antibodies. After washing with PBS, parasite nuclei were stained with Hoechst (2 μg/ml) for 20 min. Samples were washed in PBS again before analysis in a BD LSRFortessa Cytometer. A total of 200,000 events were recorded per sample. After gating erythrocytes, forward scatter height (FSC-H) versus forward scatter area (FSC-A) plots were used to define the singlets population. Unstained samples and uninfected erythrocyte controls were used to define the thresholds for positivity.

Gamete activation assays

To evaluate exflagellation, a 30-µl sample from day 12 E5ind and 1.2Bind mature gametocyte cultures was mixed with an equal volume of ookinete medium [100 µM xanthurenic acid, sodium bicarbonate (2 g/liter), and hypoxanthine (50 mg/liter) in RPMI 1640–Hepes (pH 7.4)] with a 1:1500 dilution of anti–Pfs25-Cy3 antibody. As a control, we used NF54 mature gametocytes induced as previously described (*41*). Part of the preparation was immediately transferred to a hemocytometer and incubated at room temperature. After 15 min, exflagellation centers and erythrocyte density were counted using bright-field microscopy (10× magnification) to calculate the percentage of exflagellation (number of exflagellating centers relative to the number of erythrocytes) (*56*). To compare exflagellation between experiments, we determined exflagellation rates by normalizing the percentage of exflagellation by the gametocytemia.

To determine the number of egressing females, the rest of the preparation was incubated overnight at 26°C in the darkness to allow for maximal expression of Pfs25 on the macrogamete surface (*56*). Female numbers were counted on FastRead slides (Immune Systems) and normalized by gametocytemia. For IFA analysis of activated male gametes, 200 µl of stage V gametocyte cultures was mixed with 200 µl of ookinete medium and incubated at room temperature for 25 min before fixing with 4% PFA.

Mosquito infection experiments

E5ind, 1.2Bind, and NF54 mature gametocyte cultures (days 12 to 16) were provided to 3- to 7-day-old *Anopheles stephensi* mosquitoes that were starved the day before the feed. Cultures for mosquito feeding were always manipulated at 37°C to prevent premature activation of the gametocytes. Fifty to 150 µl of prewarmed fresh human erythrocytes was used as a cushion to add 10 to 40 ml of gametocyte culture on top. The mixture was centrifuged at 500*g* for 5 min at 38°C, the supernatant was removed, and the pellet was mixed with an equal volume of prewarmed human serum. Five hundred microliters of the mixture was transferred into three-dimensionally printed feeders attached to a 38°C circulating water bath (Grant Instruments) (*57*), with Parafilm placed at the bottom of the feeders to have a thin membrane through which the mosquitoes can feed. Mosquitoes were allowed to feed for 20 to 30 min and then maintained at 26°C and 80% humidity. Twenty-four hours after feeding, unfed mosquitoes were removed. To determine the presence of ookinetes, mosquitoes were dissected to obtain the midguts (also 24 hours after feeding), the blood bolus was smeared onto a glass slide and fixed with 4% PFA, and ookinetes/females were detected using the Pfs25 antibody. Nine days after the feed, mosquito midguts were stained in 0.1% mercurochrome in PBS for 15 min, fixed in 4% PFA, and stained with 4′,6-diamidino-2-phenylindole (DAPI). Infection intensity was assessed by counting the number of oocysts per midgut with a 20× objective.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/24/eaaz5057/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. G. A. Josling, K. C. Williamson, M. Llinás, Regulation of sexual commitment and gametocytogenesis in malaria parasites. *Annu. Rev. Microbiol.* **72**, 501–519 (2018).
2. E. Meibalan, M. Marti, Biology of Malaria Transmission. *Cold Spring Harb. Perspect. Med.* **7**, a025452 (2017).
3. The malERA Refresh Consultative Panel on Basic Science and Enabling Technologies, malERA: An updated research agenda for basic science and enabling technologies in malaria elimination and eradication. *PLOS Med.* **14**, e1002451 (2017).
4. M. C. Bruce, P. Alano, S. Duthie, R. Carter, Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development. *Parasitology* **100**, 191–200 (1990).
5. C. Bancells, O. Llorà-Batlle, A. Poran, C. Nötzel, N. Rovira-Graells, O. Elemento, B. F. C. Kafsack, A. Cortés, Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. *Nat. Microbiol.* **4**, 144–154 (2019).
6. B. F. Kafsack, N. Rovira-Graells, T. G. Clark, C. Bancells, V. M. Crowley, S. G. Campino, A. E. Williams, L. G. Drought, D. P. Kwiatkowski, D. A. Baker, A. Cortés, M. Llinás, A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–252 (2014).
7. A. Poran, C. Nötzel, O. Aly, N. Mencia-Trinchant, C. T. Harris, M. L. Guzman, D. C. Hassane, O. Elemento, B. F. C. Kafsack, Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **551**, 95–99 (2017).
8. A. Sinha, K. R. Hughes, K. K. Modrzynska, T. D. Otto, C. Pfander, N. J. Dickens, A. A. Religa, E. Bushell, A. L. Graham, R. Cameron, B. F. Kafsack, A. E. Williams, M. Llinás, M. Berriman, O. Billker, A. P. Waters, A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature* **507**, 253–257 (2014).
9. M. D. Jeninga, J. E. Quinn, M. Petter, ApiAP2 transcription factors in apicomplexan parasites. *Pathogens* **8**, E47 (2019).
10. N. M. Brancucci, N. L. Bertschi, L. Zhu, I. Niederwieser, W. H. Chin, R. Wampfler, C. Freymond, M. Rottmann, I. Felger, Z. Bozdech, T. S. Voss, Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host Microbe* **16**, 165–176 (2014).
11. J.-J. Lopez-Rubio, L. Mancio-Silva, A. Scherf, Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* **5**, 179–190 (2009).
12. C. Flueck, R. Bartfai, J. Volz, I. Niederwieser, A. M. Salcedo-Amaya, B. T. Alako, F. Ehlgen, S. A. Ralph, A. F. Cowman, Z. Bozdech, H. G. Stunnenberg, T. S. Voss, *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLOS Pathog.* **5**, e1000569 (2009).
13. S. Eksi, B. J. Morahan, Y. Haile, T. Furuya, H. Jiang, O. Ali, H. Xu, K. Kiattibutr, A. Suri, B. Czesny, A. Adeyemo, T. G. Myers, J. Sattabongkot, X.-z. Su, K. C. Williamson, *Plasmodium falciparum* gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development. *PLOS Pathog.* **8**, e1002964 (2012).
14. M. Filarsky, S. A. Fraschka, I. Niederwieser, N. M. B. Brancucci, E. Carrington, E. Carrió, S. Moes, P. Jenoe, R. Bártfai, T. S. Voss, GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* **359**, 1259–1263 (2018).
15. R. S. Kent, K. K. Modrzynska, R. Cameron, N. Philip, O. Billker, A. P. Waters, Inducible developmental reprogramming redefines commitment to sexual development in the malaria parasite *Plasmodium berghei*. *Nat. Microbiol.* **3**, 1206–1213 (2018).
16. D. A. Baker, Malaria gametocytogenesis. *Mol. Biochem. Parasitol.* **172**, 57–65 (2010).
17. K. G. Pelle, K. Oh, K. Buchholz, V. Narasimhan, R. Joice, D. A. Milner, N. M. Brancucci, S. Ma, T. S. Voss, K. Ketman, K. B. Seydel, T. E. Taylor, N. S. Barteneva, C. Huttenhower, M. Marti, Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. *Genome Med.* **7**, 19 (2015).
18. M. Tibúrcio, M. W. Dixon, O. Looker, S. Y. Younis, L. Tilley, P. Alano, Specific expression and export of the *Plasmodium falciparum* Gametocyte EXported Protein-5 marks the gametocyte ring stage. *Malar. J.* **14**, 334 (2015).
19. N. M. B. Brancucci, J. P. Gerdt, C. Wang, M. De Niz, N. Philip, S. R. Adapa, M. Zhang, E. Hitz, I. Niederwieser, S. D. Boltryk, M. C. Laffitte, M. A. Clark, C. Grüring, D. Ravel, A. Blancke Soares, A. Demas, S. Bopp, B. Rubio-Ruiz, A. Conejo-Garcia, D. F. Wirth, E. Gendaszewska-Darmach, M. T. Duraisingh, J. H. Adams, T. S. Voss, A. P. Waters, R. H. Y. Jiang, J. Clardy, M. Marti, Lysophosphatidylcholine regulates sexual stage differentiation in the human malaria parasite *Plasmodium falciparum*. *Cell* **171**, 1532–1544.e15 (2017).
20. M. Ghorbal, M. Gorman, C. R. Macpherson, R. M. Martins, A. Scherf, J.-J. Lopez-Rubio, Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. *Nat. Biotechnol.* **32**, 819–821 (2014).
21. C. R. Collins, S. Das, E. H. Wong, N. Andenmatten, R. Stallmach, F. Hackett, J.-P. Herman, S. Müller, M. Meissner, M. J. Blackman, Robust inducible Cre recombinase activity in the human malaria parasite *Plasmodium falciparum* enables efficient gene deletion within a single asexual erythrocytic growth cycle. *Mol. Microbiol.* **88**, 687–701 (2013).
22. H. P. Portugaliza, O. Llorà-Batlle, A. Rosanas-Urgell, A. Cortés, Reporter lines based on the *gexp02* promoter enable early quantification of sexual conversion rates in the malaria parasite *Plasmodium falciparum*. *Sci. Rep.* **9**, 14595 (2019).
23. T. Ishino, B. Boisson, Y. Orito, C. Lacroix, E. Bischoff, C. Loussert, C. Janse, R. Ménard, M. Yuda, P. Baldacci, LISP1 is important for the egress of *Plasmodium berghei* parasites from liver cells. *Cell. Microbiol.* **11**, 1329–1339 (2009).

24. E. Knuepfer, M. Napiorkowska, C. van Ooij, A. A. Holder, Generating conditional gene knockouts in *Plasmodium* – a toolkit to produce stable DiCre recombinase-expressing parasite lines using CRISPR/Cas9. *Sci. Rep.* **7**, 3881 (2017).
25. D. A. Baker, L. B. Stewart, J. M. Large, P. W. Bowyer, K. H. Ansell, M. B. Jiménez-Díaz, M. El Bakkouri, K. Birchall, K. J. Dechering, N. S. Bouloc, P. J. Coombs, D. Whalley, D. J. Harding, E. Smiljanic-Hurley, M. C. Wheldon, E. M. Walker, J. T. Dessens, M. J. Lafuente, L. M. Sanz, F.-J. Gamo, S. B. Ferrer, R. Hui, T. Bousema, I. Angulo-Barturén, A. T. Merritt, S. L. Croft, W. E. Gutteridge, C. A. Kettleborough, S. A. Osborne, A potent series targeting the malarial cGMP-dependent protein kinase clears infection and blocks transmission. *Nat. Commun.* **8**, 430 (2017).
26. N. Rovira-Graells, A. P. Gupta, E. Planet, V. M. Crowley, S. Mok, L. Ribas de Pouplana, P. R. Preiser, Z. Bozdech, A. Cortés, Transcriptional variation in the malaria parasite *Plasmodium falciparum*. *Genome Res.* **22**, 925–938 (2012).
27. R. van Biljon, R. van Wyk, H. J. Painter, L. Orchard, J. Reader, J. Niemand, M. Llinás, L.-M. Birkholtz, Hierarchical transcriptional control regulates *Plasmodium falciparum* sexual differentiation. *BMC Genomics* **20**, 920 (2019).
28. B. S. Crabb, T. Triglia, J. G. Waterkeyn, A. F. Cowman, Stable transgene expression in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **90**, 131–144 (1997).
29. G. A. Josling, T. J. Russell, J. Venezia, L. Orchard, R. van Biljon, H. J. Painter, M. Llinás, Dissecting the role of PfAP2-G in malaria gametocytogenesis. *Nat. Commun.* **11**, 1503 (2020).
30. N. Rovira-Graells, V. M. Crowley, C. Bancells, S. Mira-Martínez, L. Ribas de Pouplana, A. Cortés, Deciphering the principles that govern mutually exclusive expression of *Plasmodium falciparum clag3* genes. *Nucleic Acids Res.* **43**, 8243–8257 (2015).
31. T. S. Voss, J. Healer, A. J. Marty, M. F. Duffy, J. K. Thompson, J. G. Beeson, J. C. Reeder, B. S. Crabb, A. F. Cowman, A *var* gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. *Nature* **439**, 1004–1008 (2006).
32. S. A. Fraschka, M. Filarsky, R. Hoo, I. Niederwieser, X. Y. Yam, N. M. B. Brancucci, F. Mohring, A. T. Mushunje, X. Huang, P. R. Christensen, F. Nosten, Z. Bozdech, B. Russell, R. W. Moon, M. Marti, P. R. Preiser, R. Bártfai, T. S. Voss, Comparative heterochromatin profiling reveals conserved and unique epigenome signatures linked to adaptation and development of malaria parasites. *Cell Host Microbe* **23**, 407–420.e8 (2018).
33. N. M. B. Brancucci, M. De Niz, T. J. Straub, D. Ravel, L. Sollelis, B. W. Birren, T. S. Voss, D. E. Neafsey, M. Marti, Probing *Plasmodium falciparum* sexual commitment at the single-cell level. *Wellcome Open Res.* **3**, 70 (2018).
34. L. Meerstein-Kessel, R. van der Lee, W. Stone, K. Lanke, D. A. Baker, P. Alano, F. Silvestrini, C. J. Janse, S. M. Khan, M. van de Vegte-Bolmer, W. Graumans, R. Siebelink-Stoter, T. W. A. Kooij, M. Marti, C. Drakeley, J. J. Campo, T. J. P. van Dam, R. Sauerwein, T. Bousema, M. A. Huynen, Probabilistic data integration identifies reliable gametocyte-specific proteins and transcripts in malaria parasites. *Sci. Rep.* **8**, 410 (2018).
35. D. Ito, M. A. Schureck, S. A. Desai, An essential dual-function complex mediates erythrocyte invasion and channel-mediated nutrient uptake in malaria parasites. *eLife* **6**, e23485 (2017).
36. W. Nguitragool, A. A. Bokhari, A. D. Pillai, K. Rayavara, P. Sharma, B. Turpin, L. Aravind, S. A. Desai, Malaria parasite *clag3* genes determine channel-mediated nutrient uptake by infected red blood cells. *Cell* **145**, 665–677 (2011).
37. E. S. Sherling, E. Knuepfer, J. A. Brzostowski, L. H. Miller, M. J. Blackman, C. van Ooij, The *Plasmodium falciparum* rhoptry protein RhopH3 plays essential roles in host cell invasion and nutrient uptake. *eLife* **6**, e23239 (2017).
38. Q. Szabo, F. Bantignies, G. Cavalli, Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**, eaaw1668 (2019).
39. A. Saul, P. Graves, L. Edser, Refractoriness of erythrocytes infected with *Plasmodium falciparum* gametocytes to lysis by sorbitol. *Int. J. Parasitol.* **20**, 1095–1097 (1990).
40. S. K. Nilsson, L. M. Childs, C. Buckee, M. Marti, Targeting human transmission biology for malaria elimination. *PLOS Pathog.* **11**, e1004871 (2015).
41. M. J. Delves, U. Straschil, A. Ruecker, C. Miguel-Blanco, S. Marques, J. Baum, R. E. Sinden, Routine in vitro culture of *P. falciparum* gametocytes to evaluate novel transmission-blocking interventions. *Nat. Protoc.* **11**, 1668–1680 (2016).
42. S. Duffy, S. Loganathan, J. P. Holleran, V. M. Avery, Large-scale production of *Plasmodium falciparum* gametocytes for malaria drug discovery. *Nat. Protoc.* **11**, 976–992 (2016).
43. F. Silvestrini, Z. Bozdech, A. Lanfrancotti, E. Di Giulio, E. Bultrini, L. Picci, J. L. deRisi, E. Pizzi, P. Alano, Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **143**, 100–110 (2005).
44. F. Silvestrini, E. Lasonder, A. Olivieri, G. Camarda, B. van Schaijk, M. Sanchez, S. Younis Younis, R. Sauerwein, P. Alano, Protein export marks the early phase of gametocytogenesis of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell. Proteomics* **9**, 1437–1448 (2010).
45. H. J. Painter, M. Carrasquilla, M. Llinás, Capturing in vivo RNA transcriptional dynamics from the malaria parasite *Plasmodium falciparum*. *Genome Res.* **27**, 1074–1086 (2017).
46. S. Mira-Martínez, A. K. Pickford, N. Rovira-Graells, P. Guetens, E. Tintó-Font, A. Cortés, A. Rosanas-Urgell, Identification of antimalarial compounds that require CLAG3 for their uptake by *Plasmodium falciparum*-infected erythrocytes. *Antimicrosb. Agents Chemother.* **63**, e00052-19 (2019).
47. J. M. Santos, G. Josling, P. Ross, P. Joshi, L. Orchard, T. Campbell, A. Schieler, I. M. Cristea, M. Llinás, Red blood cell invasion by the malaria parasite is coordinated by the PfAP2-I transcription factor. *Cell Host Microbe* **21**, 731–741.e10 (2017).
48. E. Gómez-Díaz, R. S. Yerbanga, T. Lefèvre, A. Cohuet, M. J. Rowley, J. B. Ouedraogo, V. G. Corces, Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in Anopheles gambiae. *Sci. Rep.* **7**, 40655 (2017).
49. Q. Zhang, T. N. Siegel, R. M. Martins, F. Wang, J. Cao, Q. Gao, X. Cheng, L. Jiang, C.-C. Hon, C. Scheidig-Benatar, H. Sakamoto, L. Turner, A. T. Jensen, A. Claes, J. Guizetti, N. A. Malmquist, A. Scherf, Exonuclease-mediated degradation of nascent RNA silences genes linked to severe malaria. *Nature* **513**, 431–435 (2014).
50. A. Cortés, C. Carret, O. Kaneko, B. Y. Yim Lim, A. Ivens, A. A. Holder, Epigenetic silencing of *Plasmodium falciparum* genes linked to erythrocyte invasion. *PLOS Pathog.* **3**, e107 (2007).
51. S. Mira-Martínez, E. van Schuppen, A. Amambua-Ngwa, E. Bottieau, M. Affara, M. Van Esbroeck, E. Vlieghe, P. Guetens, N. Rovira-Graells, G. P. Gómez-Pérez, P. L. Alonso, U. D'Alessandro, A. Rosanas-Urgell, A. Cortés, Expression of the *Plasmodium falciparum* clonally variant *clag3* genes in human infections. *J. Infect. Dis.* **215**, 938–945 (2017).
52. F. Santolamazza, P. Avellino, G. Siciliano, F. A. Yao, F. Lombardo, J. B. Ouédraogo, D. Modiano, P. Alano, V. D. Mangano, Detection of *Plasmodium falciparum* male and female gametocytes and determination of parasite sex ratio in human endemic populations by novel, cheap and robust RTqPCR assays. *Malar. J.* **16**, 468 (2017).
53. P. R. Kensche, W. A. Hoeijmakers, C. G. Toenhake, M. Bras, L. Chappell, M. Berriman, R. Bártfai, The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Res.* **44**, 2110–2124 (2016).
54. H. J. Painter, L. M. Altenhofen, B. F. Kafsack, M. Llinás, Whole-genome analysis of *Plasmodium* spp. utilizing a new agilent technologies DNA microarray platform. *Methods Mol. Biol.* **923**, 213–219 (2013).
55. M. Díaz-Varela, A. de Menezes-Neto, D. Perez-Zsolt, A. Gámez-Valero, J. Seguí-Barber, N. Izquierdo-Useros, J. Martinez-Picado, C. Fernández-Becerra, H. A. Del Portillo, Proteomics study of human cord blood reticulocyte-derived exosomes. *Sci. Rep.* **8**, 14046 (2018).
56. M. J. Delves, A. Ruecker, U. Straschil, J. Lelièvre, S. Marques, M. J. López-Barragán, E. Herreros, R. E. Sinden, Male and female *Plasmodium falciparum* mature gametocytes show different responses to antimalarial drugs. *Antimicrob. Agents Chemother.* **57**, 3268–3274 (2013).
57. K. Witmer, E. Sherrard-Smith, U. Straschil, M. Tunnicliff, J. Baum, M. Delves, An inexpensive open source 3D-printed membrane feeder for human malaria transmission studies. *Malar. J.* **17**, 282 (2018).

Acknowledgments: We thank H. A. del Portillo (ICREA at ISGlobal) for helpful scientific discussions; G. A. Josling and M. Llinás (Pennsylvania State University) for providing the list of PfAP2-G ChIP-seq targets before publication; M. Llinás for useful comments on the manuscript; R. Bartfai (Radboud University) for technical advice for ChIP-seq; H. Portugaliza and C. Bancells (ISGlobal) for providing NF54 cDNAs and for useful discussion; A. Churchyard, S. Yahiya, and F. Dahalan (Imperial College) for technical assistance with gamete activation and mosquito feeding assays; C. Collins and M. Blackman (The Francis Crick Institute) for providing plasmid pHH1_SERA5del3DC; J. J. López-Rubio (University of Montpellier) for plasmid pL6-egfp-yfcu; E. Knuepfer (The Francis Crick Institute) for plasmid pDC2-Cas9-hDHFRyFCU; S. Osborne (LifeArc) and D. A. Baker (LSHTM) for the compound ML10; R. W. Sauerwein (Radboud University) for the anti-Pfs16 antibody; L. Ranford-Cartwright (University of Glasgow) for the anti-Pfg377 antibody; S. A. Desai (NIAID-NIH) for the anti-CLAG3 antibody; and R. F. Anders (La Trobe University) for the anti-AMA1 antibody.
Funding: This work was supported by a grant from the Spanish Ministry of Economy and Competitiveness (MINECO)/Agencia Estatal de Investigación (AEI) (SAF2016-76190-R to A.C.), cofunded by the European Regional Development Fund (ERDF, European Union), and also received funding from "la Caixa" Banking Foundation under the project code HR18-00267 awarded to A.C. O.L.-B. was supported by a FPU fellowship from the Spanish Ministry of Education, Culture and Sports (FPU014/02456) and an EMBO Short-Term Fellowship (no. 7732). L.M.-T. was supported by a fellowship from the Spanish Ministry of Economy and Competitiveness (BES-2017-081079). H.T. was supported by Secretaria d'Universitats i Recerca del Departament d'Economia i Creixement, Generalitat de Catalunya. C.F.-B. was partially supported by a grant from MINECO (SAF2016-80655-R). Work in the Baum laboratory was supported through funding from Wellcome (Investigator Award 100993/Z/13/Z). This research is part of ISGlobal's Program on the Molecular Mechanisms of Malaria, which is partially supported by the Fundación Ramón Areces. We acknowledge support from the Spanish Ministry of Science and Innovation through the "Centro de Excelencia Severo Ochoa 2019-2023" Program (CEX2018-000806-S) and support from the Generalitat de Catalunya

through the CERCA Program. **Author contributions:** O.L.-B. performed the experiments. L.M.-T. and O.L.-B. performed the bioinformatics analysis. K.W. performed experiments, provided training, and, together with J.B., provided supervision for experiments with mature gametocytes and mosquito infections. H.T. and C.F.-B. provided reticulocyte-enriched cord blood and supervision for reticulocyte invasion experiments. O.L.-B. and A.C. conceived the project, designed the experiments, and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The ChIP-seq and microarray data have been deposited in the GEO database with accession number GSE135006. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Materials and protocols are available from the corresponding author on reasonable request. The scripts used for the analysis of microarray and ChIP-seq data will be made available upon reasonable request, without any restrictions.

Submitted 15 September 2019
Accepted 15 April 2020
Published 10 June 2020
10.1126/sciadv.aaz5057

Citation: O. Llorà-Batlle, L. Michel-Todó, K. Witmer, H. Toda, C. Fernández-Becerra, J. Baum, A. Cortés, Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*. *Sci. Adv.* **6**, eaaz5057 (2020).

Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/24/eaaz5057/DC1

Supplementary Materials for

Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*

Oriol Llorà-Batlle, Lucas Michel-Todó, Kathrin Witmer, Haruka Toda, Carmen Fernández-Becerra, Jake Baum, Alfred Cortés*

*Corresponding author. Email: alfred.cortes@isglobal.org

Published 10 June 2020, *Sci. Adv.* **6**, eaaz5057 (2020)
DOI: 10.1126/sciadv.aaz5057

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/24/eaaz5057/DC1)

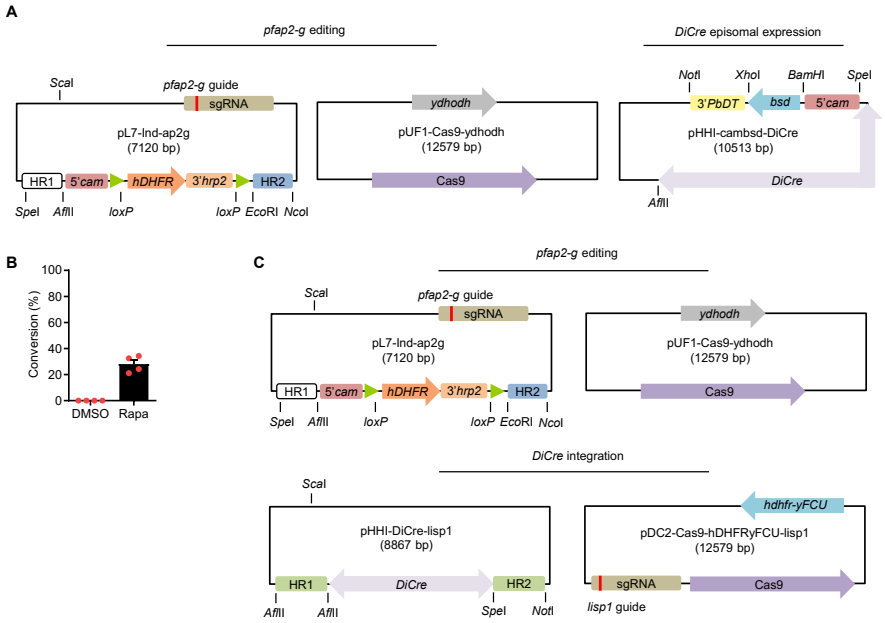


Fig. S1. Plasmids used to generate the transgenic lines and characterization of the line expressing DiCre episomally. (A) Schematic (not to scale) of the plasmids used for the *pfap2-g* conditional activation system with episomally-expressed DiCre. Restriction sites used for the generation of the plasmids are shown. **(B)** Sexual conversion levels in the transgenic line with the *pfap2-g* conditional activation system and episomally-expressed DiCre. Data are presented as the average and s.e.m. of four independent experiments. **(C)** Schematic (not to scale) of the plasmids used for the conditional activation system with the DiCre expression cassette integrated in the genome (at the *lisp1* locus). Plasmid pUF1-Cas9-ydhodh was not used to generate the E5ind line, for which the other three plasmids were simultaneously co-transfected (see Materials and Methods).

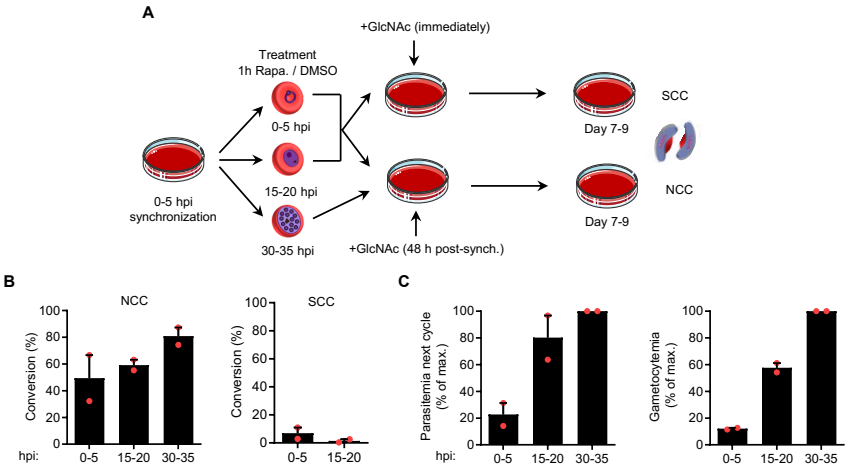


Fig. S2. Optimization of the protocol to induce sexual conversion in the E5ind line. (A) Schematic of the procedure used to determine the optimal time to induce conversion. E5ind cultures synchronized to a 5 h age window were treated with rapamycin at different times (in h post-invasion, hpi). GlcNac was added immediately after rapamycin treatment to measure only sexual conversion within the same cycle of *pfap2-g* activation (SCC route), or after reinvasion to measure conversion at the next cycle (NCC route). **(B)** Sexual conversion rate by the NCC and SCC routes after treating with rapamycin at different times. **(C)** Left, parasitemia at the cycle after rapamycin treatment, in cultures treated at different times, relative to the maximal parasitemia (obtained when treating at 30-35 hpi). Right, gametocytemia under NCC conditions in cultures induced with rapamycin at different times, relative to the maximal gametocytemia (obtained when inducing at 30-35 hpi). Induction at early stages results in reduced growth, which together with the conversion rate explains the lower number of gametocytes obtained. Data in panels **B** and **C** are presented as the average and s.e.m. of two independent experiments.

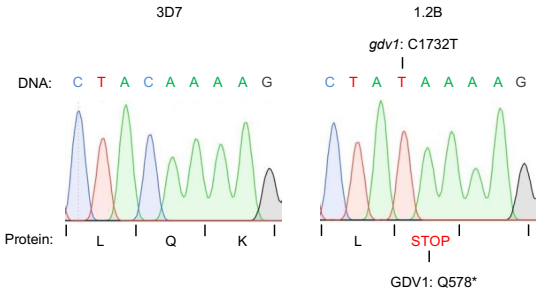


Fig. S3. The 1.2B parasite line has a premature stop codon in *gdv1*. Sanger sequencing validation of the mutation in *gdv1* showing the presence of a C to T mutation at position 1,732 of the coding sequence, resulting in a premature STOP codon (Q578*).

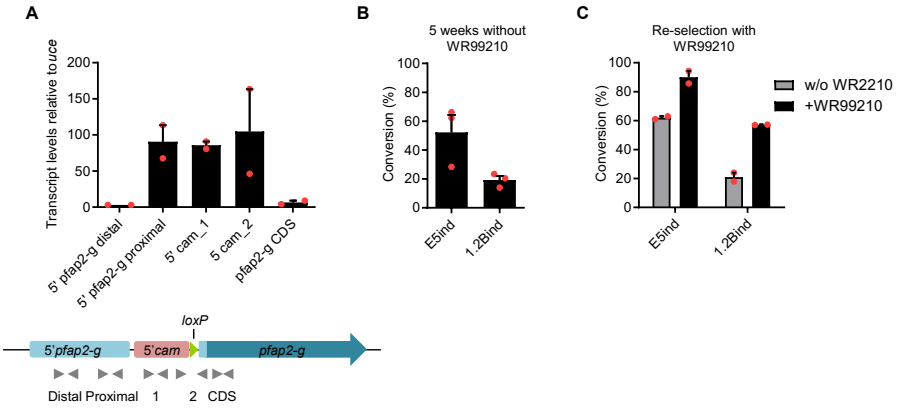


Fig. S4. Transcriptional and epigenetic events at the *pfap2-g* locus in the inducible lines. (A) Reverse transcriptase-quantitative PCR (RT-qPCR) analysis of transcript levels containing different elements of the *pfap2-g* locus in induced schizonts of the E5ind line (40-45 h post-invasion). The primers used for the analysis (arrowheads in the scheme, not to scale) were located at positions -1412 to -1302 (5' *pfap2-g* distal) and -449 to -351 (5' *pfap2-g* proximal) of the endogenous *pfap2-g* upstream region (relative to the position of the *pfap2-g* start codon in wild type parasites), -699 to -581 (5'cam_1) and –201 to -68 (5'cam_2, using a reverse primer located in the region between the *5'cam* and the *pfap2-g* coding sequence) of the 5'*cam* promoter (relative to the position of the *pfap2-g* start codon in the inducible line after recombination), and +259 to +382 of the *pfap2-g* coding sequence (*pfap2-g* CDS). Transcript levels at the coding sequence were similar to levels measured with primers located further downstream within the coding sequence (Fig. 2). Transcript levels are normalized against ubiquitin-conjugating enzyme (*uce*). Data are presented as the average and s.e.m. of two independent experiments. **(B)** Sexual conversion rates in rapamycin-treated E5ind and 1.2Bind cultures maintained for 5 weeks in the absence of WR99210 before induction. Data are presented as the average and s.e.m. of three independent experiments (each experiment corresponds to a separate culture maintained in the absence of WR99210 before induction). **(C)** Sexual conversion rates upon rapamycin induction of E5ind and 1.2Bind cultures reselected with WR99210 for 6-

8 days after 5 weeks without the drug (+WR99210). Conversion rates upon induction were also measured in cultures maintained in parallel without WR99210 (w/o WR99210) during the 6-8 days. Data are presented as the average and s.e.m. of two independent experiments.

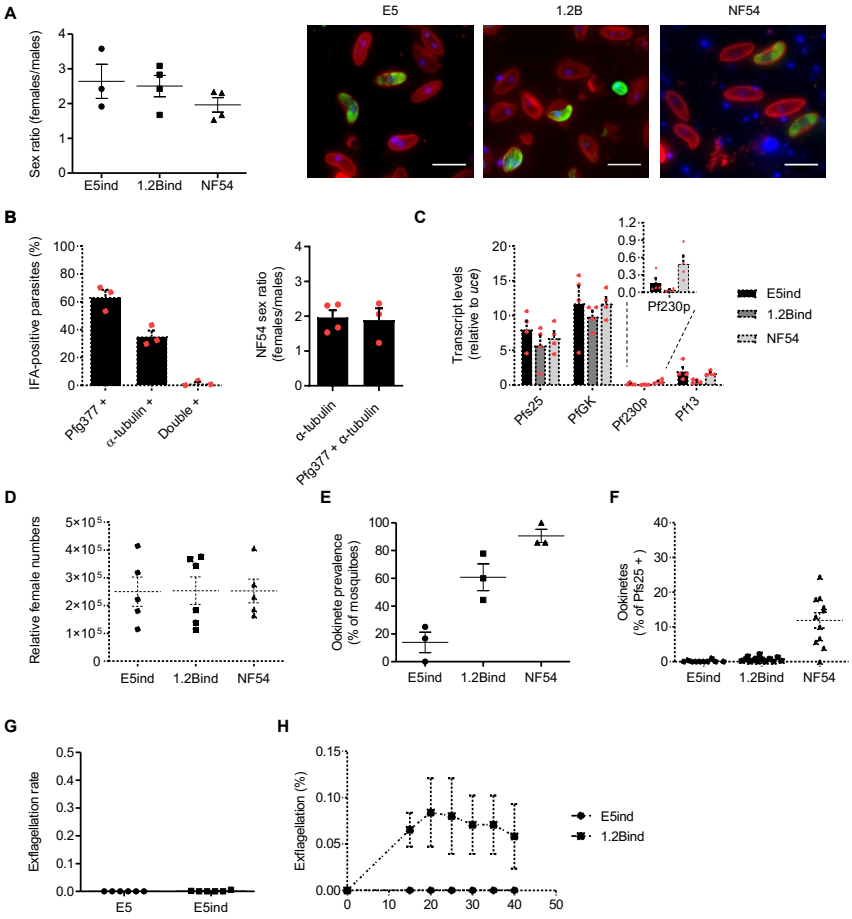


Fig. S5. Characterization of E5ind and 1.2Bind mature gametocytes. (A) Sex ratio determined by IFA of stage V gametocytes using antibodies against α-tubulin (green, male-specific at stage V) and glycophorin A (red, stains the red cell membrane of all gametocytes). DAPI (blue) stains nuclei. Male and female

gametocytes were identified as α-tubulin-positive and -negative stage V gametocytes, respectively (immature gametocytes were excluded). Data are presented as the average and s.e.m. of three (E5ind) or four (1.2Bind and NF54) independent experiments. The total number of parasites scored was >2,000 for each parasite line. Representative IFA images are shown. Scale bar, 10 μm. **(B)** Validation of α-tubulin as a male-specific marker in stage V gametocytes. Left, quantification of IFA co-staining with anti-α-tubulin and anti-Pfg377 (a female-specific marker) antibodies in the NF54 line. Double-positive parasites were virtually not seen. Right, sex ratios determined by only using α-tubulin or both α-tubulin and Pfg377 antibodies. Data are presented as the average and s.e.m. of three or four independent experiments. The total number of parasites scored was >500. **(C)** Transcriptional analysis of female (*pfs25*, *pfGK*) and male (*pf230p*, *pf13*) specific markers (*52*) in day 12-14 gametocyte cultures. Transcript levels are normalized against ubiquitin-conjugating enzyme (*uce*). Data are presented as the average and s.e.m. of four independent experiments. **(D)** Number of Pfs25-positive females/ml (determined using an anti-Pfs25 antibody coupled with Cy3) normalized by gametocytemia, in day 12-14 gametocyte cultures. Data are presented as the average and s.e.m. of five (E5ind and NF54) or six (1.2Bind) independent experiments. **(E)** Ookinete prevalence (percentage of mosquito midguts with at least one ookinete) 24h after feeding. Data are presented as the average and s.e.m. of three independent experiments. The number of mosquitoes analyzed between the three experiments was: E5ind, 20 midguts; 1.2Bind, 23 midguts; NF54: 18 midguts. **(F)** Ookinete intensity (percentage of ookinetes relative to the total Pfs25-positive cells present in each mosquito midgut). Data are presented as the average and s.e.m of 10 midguts (E5), 14 midguts (1.2B), or 11 midguts (NF54), derived from two independent feeding experiments. **(G)** Exflagellation rate (percentage of exflagellation normalized by gametocytemia, see Materials and Methods) of E5ind and parental E5 gametocyte cultures at day 12-14. Data are presented as the average and s.e.m. of six independent experiments. **(H)** Time-course percentage of exflagellation analysis. To assess the possibility that in E5ind and 1.2Bind DNA replication and exflagellation proceed more slowly rather than

failing to occur, exflagellation centers were first scored 15 min after activation and then every 5 min. The decreasing trend in the percentage of exflagellation after 20 min excludes this possibility. Data are presented as the average and s.e.m. of two independent experiments.

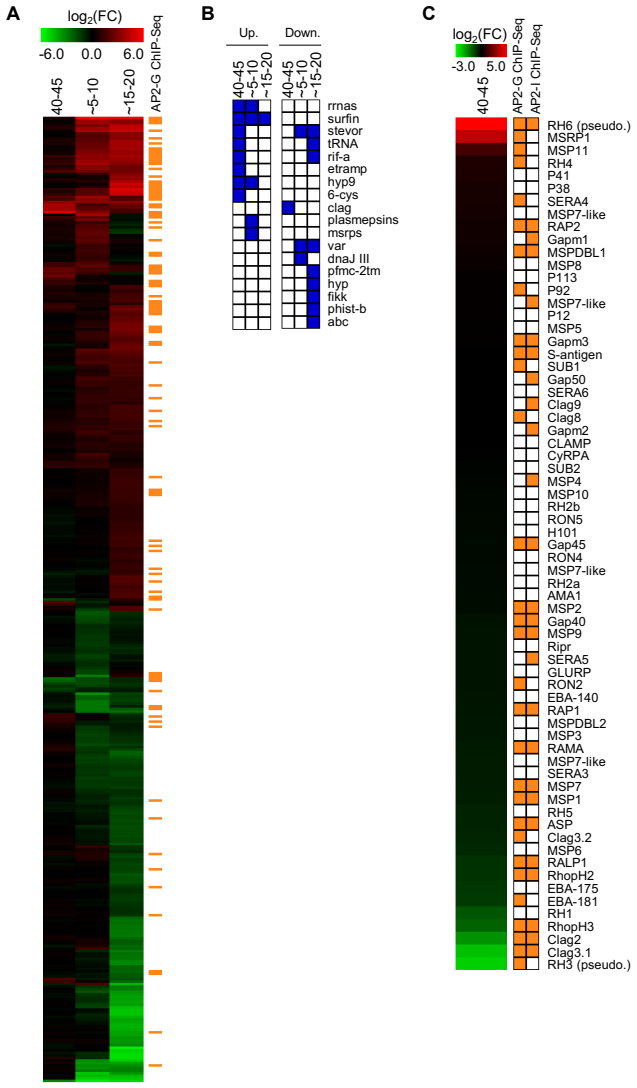


Fig. S6. Transcriptomic alterations upon *pfap2-g* activation. (A) Hierarchical clustering of genes differentially expressed in a time-course analysis of DMSO-treated (uninduced control) and rapamycin-treated (induced) E5ind cultures. Values are the average of the $\log_2$ of the expression fold-change (FC) (induced/non-induced) in the two replicates. Genes with a $\log_2$(FC)>1 in the two independent experiments at any of the time-points analyzed are shown. Samples were collected at 40-45 h post-invasion (hpi) of the cycle of induction or ~5-10 and ~15-20 hpi of the next cycle. Genes bound by PfAP2-G according to ChIP-Seq experiments are indicated in orange (*29*). **(B)** Gene set enrichment analysis (GSEA) of gene families and other gene sets. Only gene sets significantly upregulated (Up.) or downregulated (Down.) in induced vs non-induced cultures are shown (FDR<0.05). **(C)** Expression FC (ordered by FC intensity) of known genes involved in erythrocyte invasion. Genes bound by PfAP2-G and PfAP2-I according to ChIP-Seq data are shown in orange (*29, 47*).

Table S1. Oligonucleotide sequences. List of oligonucleotides used in this study, classified according to their use: cloning procedures, diagnostic PCR or quantitative PCR (qPCR). In lowercase: non-annealing part of the oligonucleotide; in bold: restriction sites included in the oligonucleotide; underlined: loxP sequence; dashed underlined: guide sequences.

Primers used for cloning procedures		
Number	**Name**	**Sequence (5'-3')**
p1	LoxP_site_BamHI_ends_F	**gatcc**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**g**
p2	LoxP_site_BamHI_ends_R	**gatcc**ATAACTTCGTATAGCATACATTATACGAAGTTAT**g**
p3	HR2_ap2g_-73_LoxP_EcoRI_F	tggaga**gaattc**ATAACTTCGTATAATGTATGCTATACGAAGTTATAACAGTTTTATATCGGACTAAC
p4	HR2_ap2g_+311_NcoI_R	tggtgtt**ccatgg**GTTGATAATCGTATCTTCGAG
p5	HR1_ap2g_-449_SpeI_F	tggaga**actagt**ATATGTCCTATAGGTGTCAAAC
p6	HR1_ap2g_-160_AflII_R	tggtgtt**cttaag**GGAGATATTTGAATGTACCTAC
p7	ap2-g_-158_guide_F	TTCTAGCTCTAAAACTTATATTGGCACTAATTTAGAATATTATATACTTA
p8	ap2-g_-139_guide_R	TAAGTATATAATATTCTAAATTAGTGCCAATATAAGTTTTAGAGCTAGAA
p9	pHLHbsdR_bsd_+1_BamHI_F	tgttggt**ggatcc**tATGGCACCTTTGTCTCAAGAAG
p10	pHLHbsdR_PbDT3'_+22_R	CGAACATTAAGCTGCCATATCC
p11	pHH1eba140_7305_AflII_SpeI_F	tggc**aagctt**g**actagt**tgtt**cttaag**GCGATCCATATAATTATTAATAGGT
p12	pHH1eba140_7387_R	GACGGCCAGTGAATTGTAATA
p13	LISP1_HR2_+5891_F_SpeI	tgttgg**actagt**TACCTATAGAGGATAAGGAGAA
p14	LISP1_HR2_+6235_R_NotI	tggttg**gcggccgc**GTAAGTGTTGTGGGTATGCTT
p15	LISP1_HR1_+5169_F_Infusion	tatggatcgccttaaAAGACAATGGGAAATGGTGTTA
p16	LISP1_HR1_+5523_R_Infusion	tttttattaagttccTTCTGGGACGACATTTATTGTT
p17	LISP1_Guide_+5526_F	TAAGTATATAATATTGAGGAACTGGGAACATGTAGGTTTTAGAGCTAGAA
p18	LISP1_Guide_+5545_R	TTCTAGCTCTAAAACCTACATGTTCCCAGTTCCTCAATATTATATACTTA

Primers for diagnostic PCR		
Number	**Name**	**Sequence (5'-3')**
p19	LISP1_+5088_F	TATGAAGAATATATTGAACGAATC
p20	LISP1_+6006_R	AGTATACCCAGGAGTGGATAA
p21	Pfap2-g_5'UTR_-460_F	GCTTCTTTAATGTTGTATGTATG
p22	Pfap2-g_CDS_+345_R	GATACATTCTCGTTACTCTGC

Primers for qPCR		
Number	**Name**	**Sequence (5'-3')**
p23	PF3D7_1222600_F1 (pfap2-g)	AACAACGTTCATTCAATAAATAAGG
p24	PF3D7_1222600_R1 (pfap2-g)	ATGTTAATGTTCCCAAACAACCG
p25	hdhfr_qPCR_F	AGTAGAAGGTAAACAGAATCTG
p26	hdhfr_qPCR_R	GGCATCATCTAGACTTCTGG
p27	Calmodulin_qPCR_F	GATGGAACTATAACAACTAAGG
p28	Calmodulin_qPCR_R	CATTAAGGTTAGAAATTCGGGA
p29	Seryl_qPCR_F (serrs)	AAGTAGCAGGTCATCGTGGTT
p30	Seryl_qPCR_R (serrs)	TTCGGCACATTCTTCCATAA
p31	PfUCE_F	GGTGTTAGTGGCTCACCAATAGGA
p32	PfUCE_R	GTACCACCTTCCCATGGAGTA
p33	5'pfap2-g_distal_F	TTAATAATACGTATGCTTGTGGA
p34	5'pfap2-g_distal_R	AAGAAAGAACTACAACATTCTG
p35	5'pfap2-g_proximal_F	ATATGTCCTATAGGTGTCAAAC
p36	5'pfap2-g_proximal_R	ACAAATACTTATTATATAAATTGTAC
p37	5'cam_1_F	ATTTTTAAATGCTTACTTAATTATCT
p38	5'cam_1_R	TACATTGTTTAATACTACTACATGT
p39	5'cam_2_F	TTTTAAACTAGAAAAGGAATAACTA
p40	5'pfap2-g_-45_R	TATGTTAGTCCGATATAAAACTG
p41	Pfap2-g_CDS_+259_F	GAAGAGAGCATGCAATGAAGT
p42	Pfap2-g_CDS_+382_R	TTGTCCATGCAACTATTCGATA
p43	Pfs25_F1	AATGCGAAAGTTACCGTGGA
p44	Pfs25_R1	TACATTCCAAATGACCACTC
p45	PfGK_F	GTCAGAACATGTTCCGGACT
p46	PfGK_R	TAGCTTGACCAATGCATGCA
p47	Pf230p_F1	CCCAACTAATCGAAGGGATGAA
p48	Pf230p_R2	TGTTGTTCGATTCCAGTTGGT
p49	Pf13_F	ACGAATATGCTCGAGAACGA
p50	Pf13_R	GCCTTTTCATCTGACACGT
p51	RAP1_+479_F	TCAGCTAGTCCACATGGTGAA
p52	RAP1_+619_R	TTTTTGGTGCAGGAGGTGCTT
p53	RAMA_+1735_F	GACCTTCTAGATGTCCAACAA
p54	RAMA_+1855_R	ATGCATGTACGTCGCTAACAT
p55	RhopH2_2F	TGTTGCTGTCCATATTTAGTTTT
p56	RhopH2_2R	AATATATCGCTACATAACTTCGT
p57	RhopH3 _+1289_F	TATCTGTTCAATGCCCAACTTA
p58	RhopH3_+1430_R	CTGCAGAAGGGTGTTTACTTT
p59	clag3.1_clag3.2_6F	TAGTAATGAGAATTAGTTGGACA
p60	clag3.1_6R	ATAAATATTTGGATGCTTCAGCA

4 **Article 4:** A heat-shock response regulated by the PfAP2-HS transcription factor protects human malaria parasites from febrile temperatures

nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-021-00940-w

Check for updates

A heat-shock response regulated by the PfAP2-HS transcription factor protects human malaria parasites from febrile temperatures

Elisabet Tintó-Font[1], Lucas Michel-Todó[1,7], Timothy J. Russell[2,7], Núria Casas-Vila[1], David J. Conway[3], Zbynek Bozdech[4], Manuel Llinás[2,5] and Alfred Cortés[1,6]✉

Periodic fever is a characteristic clinical feature of human malaria, but how parasites survive febrile episodes is not known. Although the genomes of *Plasmodium* species encode a full set of chaperones, they lack the conserved eukaryotic transcription factor HSF1, which activates the expression of chaperones following heat shock. Here, we show that PfAP2-HS, a transcription factor in the ApiAP2 family, regulates the protective heat-shock response in *Plasmodium falciparum*. PfAP2-HS activates the transcription of *hsp70-1* and *hsp90* at elevated temperatures. The main binding site of PfAP2-HS in the entire genome coincides with a tandem G-box DNA motif in the *hsp70-1* promoter. Engineered parasites lacking PfAP2-HS have reduced heat-shock survival and severe growth defects at 37 °C but not at 35 °C. Parasites lacking PfAP2-HS also have increased sensitivity to imbalances in protein homeostasis (proteostasis) produced by artemisinin, the frontline antimalarial drug, or the proteasome inhibitor epoxomicin. We propose that PfAP2-HS contributes to the maintenance of proteostasis under basal conditions and upregulates specific chaperone-encoding genes at febrile temperatures to protect the parasite against protein damage.

A temperature increase of only a few degrees Celsius above the optimal growth temperature of any organism causes aberrant protein folding and aggregation, which contributes to an imbalance in protein homeostasis (proteostasis) that can lead to cell-cycle arrest or cell death[1]. To counteract the effect of high temperatures and other proteotoxic conditions, cells have a well-characterized heat-shock response that induces the expression of molecular chaperones that aid protein refolding and prevent non-specific protein aggregation[1,2]. In most eukaryotes, the immediate upregulation of chaperone-encoding genes during heat shock depends on the conserved transcription factor HSF1, whereas other transcriptional changes during thermal stress are driven by different transcription factors[3–6].

The human response to blood-stage infection with malaria parasites involves periodic fever episodes, which are the hallmark of clinical malaria[7–9]. Fever is an important part of the human innate immune response and may contribute towards reducing the total parasite burden[8–10]. Infection with *Plasmodium falciparum*, which causes the most severe forms of human malaria, results in fevers that typically occur on alternate days (tertian fever). Tertian fever reflects the approximately 48 h duration of the asexual intra-erythrocytic development cycle (IDC), during which parasites progress through the ring, trophozoite and multinucleated schizont stages. Fever episodes are triggered by schizont rupture and the release of invasive merozoites[8,9]. In vitro, febrile temperatures inhibit parasite growth, with maximal effect on trophozoites and schizonts[9,11,12], and induce conversion of asexual parasites into sexual forms that mediate transmission to mosquitoes[13].

Despite the importance of the heat-shock response for the survival of human malaria parasites during the host fever episodes, the regulation of the heat-shock response has not been characterized in these organisms. The genomes of the *Plasmodium* spp. lack an orthologue of HSF1 but they encode the main chaperone families described in other organisms[14], and several specific *P. falciparum* chaperones have been shown to be essential for heat-shock survival[15–19]. Phosphatidylinositol 3-phosphate and apicoplast-targeted pathways are also essential for parasite survival under thermal stress[18,19]. The transcript levels of over 300 genes are altered at febrile temperatures[20] but how these transcriptional changes are regulated is not known. Here we set out to characterize the regulation of the protective response of *P. falciparum* to increased temperature.

Results

A nonsense mutation in *pfap2-hs* is associated with low survival from heat shock. To understand the molecular basis of heat-shock resistance in *P. falciparum*, we first analysed parasite lines that had been previously selected with periodic heat shock (3 h at 41.5 °C) for five consecutive rounds of the IDC[21] (Fig. 1a). Although the parental parasite line (3D7-A) seemed to have lost the ability to withstand heat shock (approximately 30% heat-shock survival) during growth in vitro, it re-adapted to heat-shock pressure (>75% heat-shock survival) in only three generations[21], suggesting that this line contained a selectable subpopulation of parasites resistant to heat shock. To evaluate whether the heat-shock-resistance phenotype of 3D7-A had a genetic or an epigenetic basis, we first analysed the transcriptome across the full IDC under basal conditions (no heat shock) of two independently selected lines (3D7-A-HS r1 and r2) and non-selected cultures maintained in parallel (3D7-A r1 and r2). This analysis failed to identify any differences in the basal

[1]ISGlobal, Hospital Clínic—Universitat de Barcelona, Barcelona, Spain. [2]Department of Biochemistry and Molecular Biology and Huck Center for Malaria Research, Pennsylvania State University, University Park, PA, USA. [3]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK. [4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [5]Department of Chemistry, Pennsylvania State University, University Park, PA, USA. [6]ICREA, Barcelona, Spain. [7]These authors contributed equally: Lucas Michel-Todó, Timothy J. Russell. ✉e-mail: alfred.cortes@isglobal.org

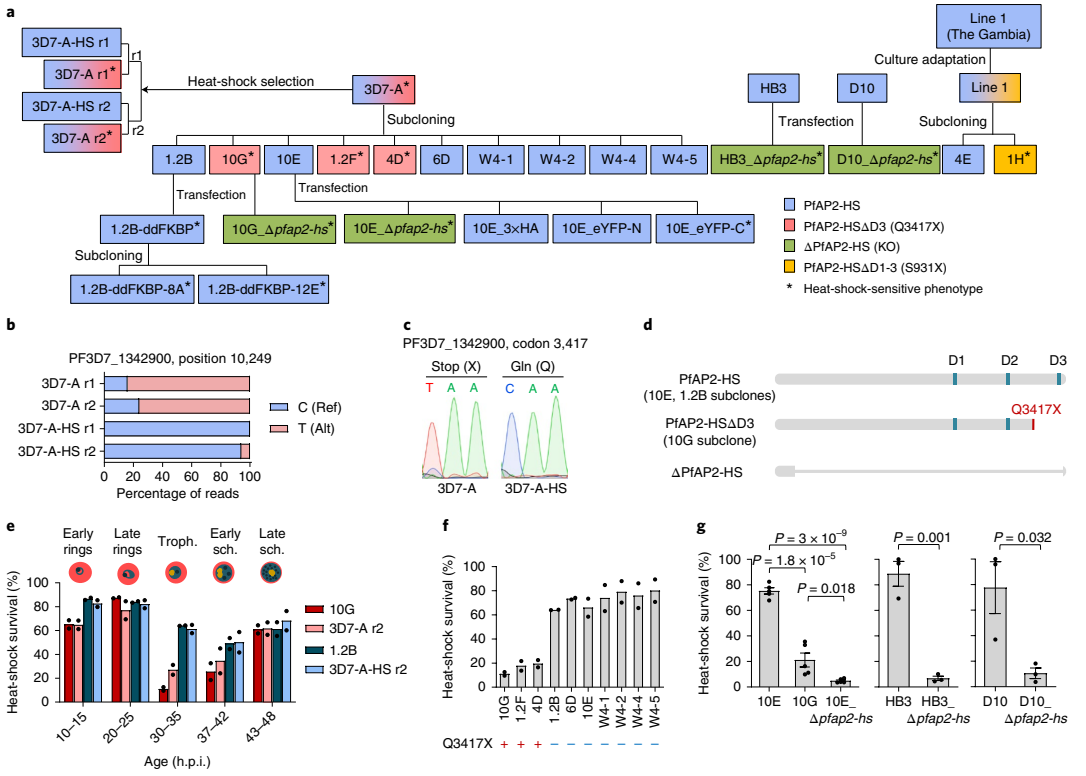
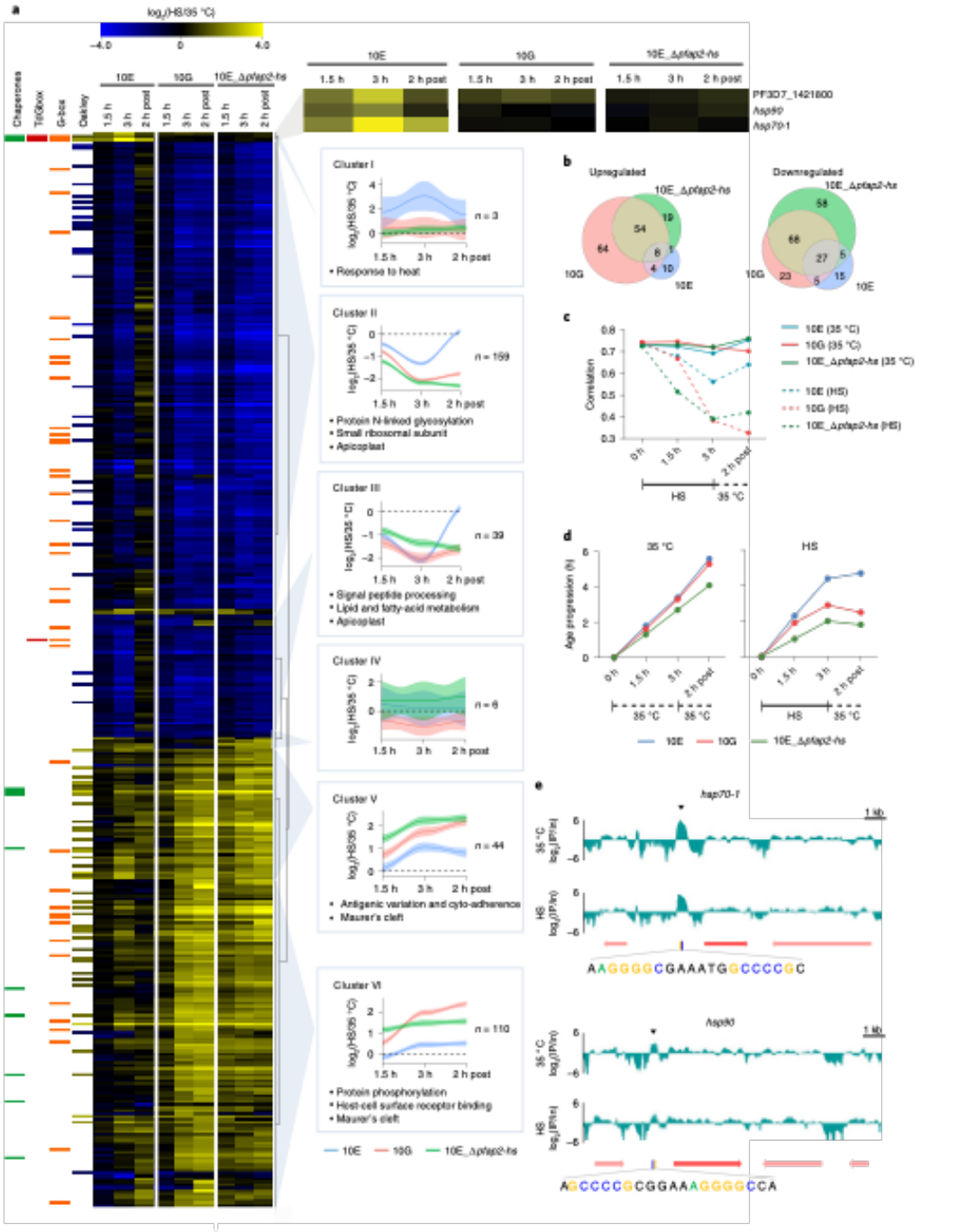


Fig. 1 | Mutations in PfAP2-HS and sensitivity to heat shock. a, Schematic of the parasite lines used in this study. Wild-type PfAP2-HS and truncated forms lacking AP2 D3 (ΔD3), the three AP2 domains (ΔD1–3) or virtually the full protein (KO) are indicated in different colours. The parasite lines shown with a colour gradient consist of a mixture of individual parasites expressing different versions of the protein. Asterisks indicate a heat-shock-sensitive phenotype, and r1 and r2 are independent replicates of the selection of 3D7-A with periodic heat-shock (3D7-A-HS r1 and r2 are the selected lines, whereas 3D7-A r1 and r2 are controls maintained in parallel at 37 °C). **b**, Proportion of Illumina reads with (Alt) or without (Ref) a nonsense mutation in *pfap2-hs* in two independently selected heat-shock-adapted cultures (3D7-A-HS r1 and r2) and their controls (3D7-A r1 and r2). **c**, Sanger-sequencing confirmation of the mutation (in the r1 replicate; representative of r1 and r2). **d**, Schematic of wild-type PfAP2-HS, PfAP2-HSΔD3 and ΔPfAP2-HS. The position of the AP2 domains is indicated (D1–3). **e**, Survival of tightly synchronized cultures exposed to heat shock at different ages (in hours post invasion, h.p.i.) for two heat-shock-sensitive (3D7-A r2 and 10G) and two heat-shock-resistant (3D7-A-HS r2 and 1.2B) lines (mean of $n = 2$ independent biological replicates). Troph., trophozoites; and sch., schizonts. **f**, Heat-shock survival at the trophozoite stage of 3D7-A subclones with or without the Q3417X mutation (mean of $n = 2$ independent biological replicates). **g**, Heat-shock survival of tightly synchronized cultures of parasite lines expressing wild-type or mutated PfAP2-HS. Values are the mean ± s.e.m. of $n = 5$ (lines of 3D7 origin; left) or 3 (HB3 and D10 lines; middle and right, respectively) independent biological replicates. *P* values were calculated using a two-sided unpaired Student's *t*-test. **e**–**g**, Values are survival relative to control cultures maintained at 37 °C (**e**,**f**) or 35 °C (**g**).

transcriptomes that could explain the heat-shock resistance phenotypes (Extended Data Fig. 1). Therefore, we sequenced the genomes of these lines, which revealed a novel single nucleotide polymorphism (SNP) that was predominant in non-selected cultures but virtually absent after heat-shock selection (Fig. 1b,c and Supplementary Table 1). This mutation was also absent from two other 3D7 stocks in our laboratory and the 3D7 reference genome, indicating that it arose spontaneously in the 3D7-A stock during culture. This SNP results in a premature stop codon (Q3417X) in the gene *PF3D7_1342900*, which encodes a putative transcription factor of the ApiAP2 family[22–24] that we termed PfAP2-HS. PfAP2-HS has three AP2 domains (D1–3), and the Q3417X mutation results in a truncated protein that lacks D3 (PfAP2-HSΔD3; Fig. 1d). This result indicates that the adaptation of 3D7-A to heat shock involved the selection of parasites expressing full-length PfAP2-HS, consistent with a role for this protein in the heat-shock response. In support of this idea, the first AP2 domain (D1) of PfAP2-HS was previously reported to recognize in vitro a DNA motif termed G-box[23], which is enriched in the upstream region of some heat-shock protein (HSP) chaperone genes[25].

To test the involvement of PfAP2-HS in heat-shock resistance, we used a heat-shock survival assay consisting of heat-shock at 41.5 °C for 3 h (ref. [21]) at the mature trophozoite stage, because maximal survival differences between heat-shock-sensitive and -resistant parasite lines were observed when the parasites were exposed at this stage (Fig. 1e). The analysis of a collection of 3D7-A subclones revealed that all subclones with the Q3417X mutation (for example, the 10G subclone) have a heat-shock-sensitive phenotype, whereas subclones with the wild-type allele (for example, the 10E subclone) have a heat-shock-resistant phenotype (Fig. 1a,f).

NATURE MICROBIOLOGY ARTICLES



Deletion of PfAP2-HS reduces survival from heat shock. To further characterize PfAP2-HS, we sought to disrupt the entire gene using clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR associated protein 9 (Cas9) technology. After several unsuccessful attempts with different 3D7 subclones at 37 °C (the physiological temperature for *P. falciparum*), we

Fig. 2 | Global transcriptional alterations in parasites exposed to heat shock. a, Hierarchical clustering of genes with altered transcript levels (≥fourfold change at any of the time points analysed) during (1.5 and 3 h) or 2 h after finishing (2 h post) heat shock (HS). The log_2-tranformed values of the fold change in expression in the HS versus control (35 °C) cultures are shown. Thirteen genes had values out of the range displayed (actual range: −4.78 to +4.93). For each cluster, mean values (lines) with the 95% confidence interval (shading) for the genes in the cluster and representative enriched Gene Ontology terms are shown. The columns on the left indicate annotation as chaperone[14], presence of the G-box[23] or tandem G-box (TdGbox) in the upstream region, and log_2-transformed fold-change values after HS from a previous study[20] (Oakley). **b**, Overlap in the genes that were altered (upregulated, left; and downregulated, right) in the three parasite lines following HS. **c**, Pearson's correlation of the genome-wide transcript levels of each culture versus the most-similar time point of a high-density time-course reference transcriptome[28]. **d**, Age progression during the HS assay, statistically estimated[61] from the transcriptomic data, under HS (right) or control (35 ºC, left) conditions. **e**, ChIP–seq analysis of HA-tagged PfAP2-HS, representative of $n=5$ and 3 independent biological replicates for 35 °C and HS, respectively. The log_2-transformed ChIP/input ratio (IP/in) at the *hsp70-1* (top) and *hsp90* (bottom) loci is shown. The position of the tandem G-box is indicated.

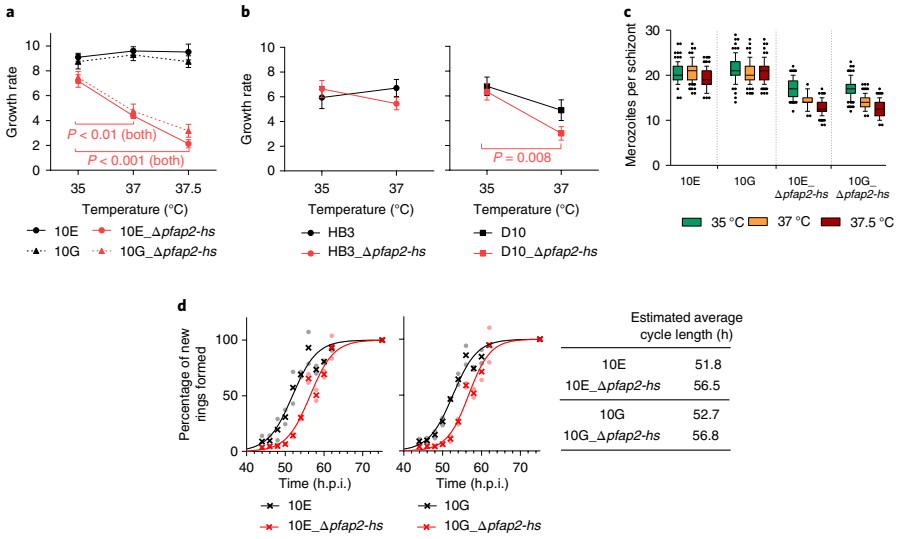


Fig. 3 | Phenotypes of parasite lines lacking PfAP2-HS. a, Growth rate of Δ*pfap2-hs* and parental lines with a 3D7 genetic background at different temperatures (mean ± s.e.m. of $n=4$ independent biological replicates). 10E_Δ*pfap2-hs*: 37 versus 35 °C, $P=2.3\times10^{-3}$; 37.5 versus 35 °C, $P=1.7\times10^{-4}$; and 10G_Δ*pfap2-hs*: 37 versus 35 °C, $P=0.011$; 37.5 versus 35 °C, $P=0.001$. **b**, Same as in **a** for parasite lines with a HB3 and D10 genetic background (mean ± s.e.m. of $n=4$ independent biological replicates). **a**,**b**, The *P* values were calculated using a two-sided unpaired Student's *t*-test. Only significant *P* values ($P<0.05$) are shown. **c**, Number of merozoites per schizont (horizontal line, median; box, quartiles; and whiskers, 10–90 percentile). Values were obtained from 100 schizonts for each parasite line and condition. **d**, Duration of the asexual blood cycle. The cumulative percentage of new rings formed at each time point (left; mean of $n=2$ independent biological replicates) and the estimated average length of the asexual blood cycle (right) are shown.

reasoned that PfAP2-HS may play a role in regulating the expression of chaperones under basal conditions in addition to being necessary for heat-shock survival. Therefore, we attempted to knock out the gene in cultures maintained at 35 °C, as mild hypothermia is expected to reduce protein unfolding and favour proteome integrity[26,27]. Knockout of *pfap2-hs* was indeed readily achieved at 35 °C in both the heat-shock-resistant 10E and the heat-shock-sensitive 10G subclones of 3D7-A (10E_Δ*pfap2-hs* and 10G_Δ*pfap2-hs* lines, respectively; Fig. 1a,d and Extended Data Fig. 2a). Deletion of *pfap2-hs* resulted in a substantial increase in the sensitivity to heat shock, with a level of heat-shock survival below that of parasites expressing PfAP2-HSΔD3. Deletion of the gene in two additional parasite lines with an unrelated genetic background, HB3 and D10, also resulted in a major reduction in heat-shock survival (Fig. 1g).

The PfAP2-HS-driven transcriptional response to heat shock is extremely compact. To define the PfAP2-HS-dependent and -independent heat-shock response, we carried out a time-course transcriptome analysis of the 10E (wild-type PfAP2-HS), 10G (PfAP2-HSΔD3) and 10E_Δ*pfap2-hs* lines during and after heat shock (Fig. 2a, Extended Data Fig. 3a and Supplementary Table 2). Hierarchical clustering based on changes in cultures exposed to heat shock compared with control cultures maintained in parallel without heat shock revealed one cluster of transcripts (cluster I) that are rapidly increased during heat shock in 10E but not in 10G or 10E_Δ*pfap2-hs*. Cluster I comprises only three genes: a gene of unknown function (*PF3D7_1421800*), the cytoplasmic *hsp70* (*hsp70-1*; *PF3D7_0818900*) and *hsp90* (*PF3D7_0708400*; Fig. 2a). The regulatory regions of these two chaperone-encoding genes contain the best two matches in the full genome for a tandem G-box[23,25] (Extended Data Fig. 3b). Although hundreds of genes in the *P. falciparum* genome contain a single G-box, only *hsp70-1* and *hsp90* showed PfAP2-HS-dependent activation during heat shock, suggesting that the tandem G-box arrangement may be needed for activation by PfAP2-HS. The strongest transcriptional response to heat

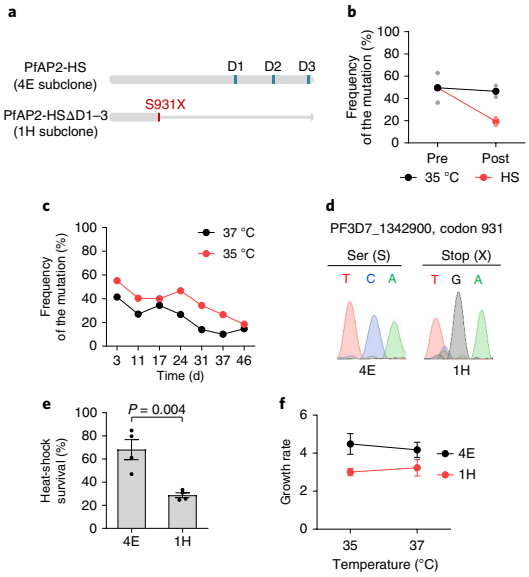


Fig. 4 | Characterization of a culture-adapted field isolate with mutations in *pfap2-hs*. **a**, Schematic of wild-type PfAP2-HS and PfAP2-HSΔD1–3 occurring in Line 1 from The Gambia after culture adaptation (C>G mutation at codon 931, S931X). The position of the AP2 domains is indicated (D1–3). **b**, Frequency of the mutation (as determined by Sanger sequencing) in culture-adapted Line 1 before (Pre) and after (Post) performing heat shock (HS) at the trophozoite stage and culturing for an additional cycle (mean of $n=2$ independent biological replicates). **c**, Frequency of the mutation during culture at different temperatures. The frozen stock from The Gambia (culture-adapted for 91 days) was placed back in culture on day 0. **d**, Sanger-sequencing determination of the presence or absence of the mutation at codon 931 in the Line 1 subclones 4E and 1H. **e**, Heat-shock survival of tightly synchronized 4E and 1H cultures relative to the control cultures maintained at 35 °C (mean ± s.e.m. of $n=4$ independent biological replicates). The *P* value was calculated using a two-sided unpaired Student's *t*-test. **f**, Growth rate of 4E and 1H at different temperatures (mean ± s.e.m. of $n=5$ independent biological replicates). No significant difference ($P<0.05$) was observed between growth at 35 and 37 °C (two-sided unpaired Student's *t*-test).

shock was observed for *hsp70-1* (approximately 16-fold increase versus about fourfold for *hsp90*).

To validate the observation that rapid activation of the cluster I genes following heat shock depends on PfAP2-HS and requires its D3, we analysed the heat-shock response of *pfap2-hs*-knockout parasite lines with different genetic backgrounds and several 3D7-A mutant subclones expressing PfAP2-HSΔD3 (Extended Data Fig. 4). The *hsp70-1* response to heat shock was delayed and of reduced magnitude in all of the knockout and mutant lines examined. These experiments also confirmed that the *hsp90* response is weaker than the *hsp70-1* response and is delayed in PfAP2-HS mutants.

PfAP2-HS-independent transcriptome alterations induced by heat shock. Genes in the other clusters (II–VI) of our transcriptomic analyses showed changes in expression during heat shock that were independent of PfAP2-HS; these changes were more pronounced in the mutant than wild-type lines (Fig. 2a). More genes with altered transcript levels were identified in the heat-shock-sensitive 10G and 10E_Δ*pfap2-hs* lines than in 10E (Fig. 2b). Furthermore, the alterations in the transcripts from clusters II–VI persisted 2 h after heat shock in both heat-shock-sensitive lines, whereas the majority of transcripts returned to basal levels in 10E (Fig. 2a). This suggests that many of these altered transcripts reflect unresolved cell damage or death. In 10E, the rapid PfAP2-HS-dependent response may protect cells from damage and enable rapid recovery from heat shock, thus limiting (in magnitude and duration) the changes in the expression of genes from clusters II–VI that reflect cell damage. After heat shock the transcriptome of the 10G and 10E_Δ*pfap2-hs* lines indeed showed a more pronounced deviation from a reference transcriptome[28] than 10E (Fig. 2c). Global transcriptional analysis also revealed that heat shock resulted in delayed IDC progression, which was again more pronounced in 10G and 10E_Δ*pfap2-hs* than 10E (Fig. 2d).

In addition to genes reflecting cell damage, clusters II–VI probably include some genes that participate in the PfAP2-HS-independent heat-shock response. In particular, clusters V and VI include several chaperone-encoding genes that were upregulated during heat shock, although at a later time point than the cluster I genes (Fig. 2a). However, the expression of the majority of known *P. falciparum* chaperones[14] was not altered by heat shock and, with the exception of cluster I genes, the alterations occurred mainly in the mutant lines (Extended Data Fig. 5). To provide a clearer view of the wild-type heat-shock response, we analysed heat-shock-induced changes in the 10E line alone. Overall, there was generally good concordance with the genes and the processes that are altered by heat shock described in a previous study using non-synchronized cultures[20] (Extended Data Fig. 6 and Supplementary Table 3). Altogether, we conclude that a number of genes are up- or downregulated during heat shock and some may contribute to heat-shock protection through PfAP2-HS-independent responses but in the absence of the rapid PfAP2-HS-dependent activation of cluster I genes, these responses are insufficient to ensure parasite survival at febrile temperatures.

Genome-wide analysis of PfAP2-HS binding sites. To determine the genome-wide occupancy of PfAP2-HS, we analysed a parasite line expressing haemagglutinin (HA)-tagged endogenous PfAP2-HS (Extended Data Fig. 2b) using chromatin immunoprecipitation (ChIP), followed by sequencing (ChIP–seq). The main binding site of PfAP2-HS coincides with the position of the tandem G-box in the upstream region of *hsp70-1* (Fig. 2e, Extended Data Fig. 7 and Supplementary Table 4). This is the only binding site with a median fold-enrichment value above ten (ChIP versus input) that was consistently detected in five independent ChIP–seq biological replicates, revealing an extremely restricted distribution of PfAP2-HS binding. Similar enrichment was observed between control and heat-shock conditions using ChIP–seq and ChIP with quantitative PCR (ChIP–qPCR; Extended Data Fig. 7), which suggests that PfAP2-HS binds constitutively to this site and is activated in situ by heat shock. This is reminiscent of yeast HSF1, which binds the promoter of *hsp70* and most of its other target promoters under both basal and heat-shock conditions[5]. Enrichment for PfAP2-HS at the *hsp90* promoter also coincided with the position of the tandem G-box, but was weaker, and a significant peak at this position was called in only one of the replicates (Fig. 2e, Extended Data Fig. 7 and Supplementary Table 4). No enrichment was observed at the promoter of the cluster I gene *PF3D7_1421800*, which lacks a G-box motif. The only other site consistently enriched for PfAP2-HS binding, albeit at much lower levels than *hsp70-1*, was the small nucleolar RNA *snoR04* (*PF3D7_0510900*) locus (Extended Data Fig. 7), which also lacks a G-box and was not upregulated during heat shock.

Growth defects under basal conditions in parasite lines mutated for PfAP2-HS. Both knockout lines of 3D7 origin (10E_Δ*pfap2-hs* and 10G_Δ*pfap2-hs*) showed severe temperature-dependent

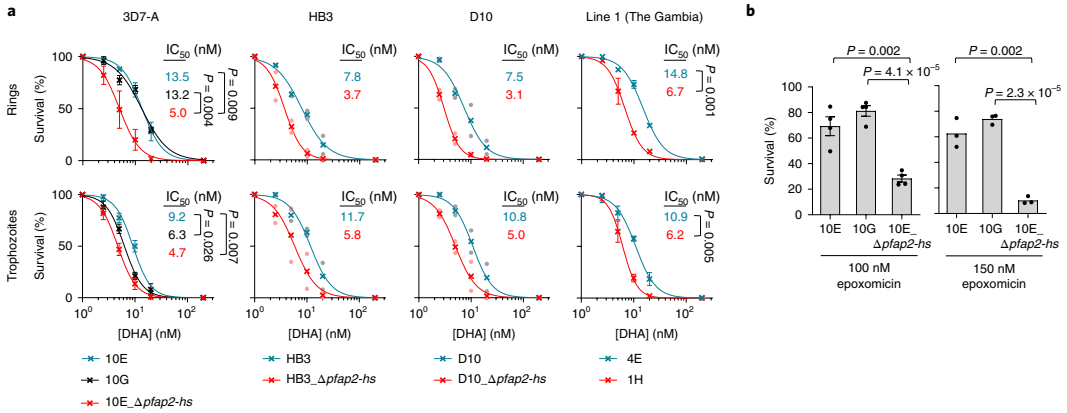


Fig. 5 | Sensitivity of parasites lacking PfAP2-HS to proteotoxic conditions. a, Survival, relative to the untreated control, after a 3 h DHA pulse at the ring (top) or trophozoite (bottom) stage. Values are the mean ± s.e.m. of $n = 3$ (3D7-A and Line 1 genetic backgrounds) or mean of $n = 2$ (HB3 and D10 genetic backgrounds) independent biological replicates. The mean half maximal inhibitory concentration (IC_{50}) for each line is shown (same colour code as the plots). **b**, Survival, relative to the untreated control, after a 3 h epoxomicin pulse at the trophozoite stage. Values are the mean ± s.e.m. of $n = 4$ (100 nM) and 3 (150 nM) independent biological replicates. **a**,**b**, The *P* values were calculated using a two-sided unpaired Student's *t*-test (only for experiments with $n \geq 3$). Only significant *P* values ($P < 0.05$) are shown.

growth defects in the absence of heat shock. They grew at similar rates to the parental lines at 35 °C but their growth was markedly reduced at 37 and 37.5 °C (Fig. 3a). The D10_Δ*pfap2-hs* line also had clearly reduced growth at 37 °C compared with 35 °C, whereas the HB3_Δ*pfap2-hs* line did not (Fig. 3b). Both 3D7 Δ*pfap2-hs* lines also showed a reduced number of merozoites per schizont, especially at 37 and 37.5 °C (Fig. 3c), which partly explains the reduced growth rate. In addition, even at 35 °C, the duration of the IDC was approximately 4 h longer in both knockout lines (Fig. 3d), which is reminiscent of the slower life-cycle progression observed in parasites under proteotoxic stress[29]. In contrast, no differences in growth rate or life-cycle duration were observed between the 10G (PfAP2-HSΔD3) and 10E (wild-type PfAP2-HS) lines, indicating that D3 is necessary for heat-shock survival but not for growth under non-stress (37 °C) conditions (Fig. 3a,c,d). Normal growth at 37 °C, but low heat-shock survival, was also observed in transgenic lines in which bulky carboxy (C)-terminal tags were added to the C terminus of endogenous PfAP2-HS, suggesting interference of the tag with the function of D3, which is located only 18 amino acids from the end of the protein (Fig. 1a and Extended Data Fig. 2).

Genome-wide sequence analysis has previously found that nonsense mutations arise in *pfap2-hs* during adaptation to culture conditions[30,31] but these are virtually absent from clinical isolates (in the www.malariagen.net/data/pf3k-5 dataset[32], only one of >2,500 isolates carries a high-confidence SNP resulting in a premature stop codon). The lack of mutations observed in clinical isolates suggests that there is a selection against loss-of-function mutations in PfAP2-HS during human infections, where parasites are frequently exposed to febrile conditions. We exposed a culture-adapted isolate in which approximately 50% of the parasites carried a mutation that results in the truncation of PfAP2-HS before its first AP2 domain[30] (monoclonal Gambian Line 1, PfAP2-HSΔD1–3) to one round of heat shock (41.5 °C, 3 h) and found that only about 20% of the parasites in the next generation carried the mutation. The frequency of the mutation remained stable in the control cultures maintained in parallel without heat shock (Fig. 4a,b). This result indicates strong selection by heat shock against parasites carrying the PfAP2-HS truncation. In contrast, there was relatively weak selection against mutants during culturing at either 35 or 37 °C, as the prevalence of the mutation only decreased from about 50% to approximately 20% after culturing for 23 generations at either temperature (Fig. 4c). Consistent with these results, a Gambian Line 1 subclone carrying the mutation (1H) was more sensitive to heat shock than a wild-type subclone (4E), but both showed no measurable difference between their growth at 35 and 37 °C (Fig. 4d–f). Together, these results indicate that PfAP2-HS is essential for heat-shock survival in all of the tested genetic backgrounds. However, PfAP2-HS is necessary for normal progression through the IDC at 37 °C in only specific genetic backgrounds (that is, 3D7 and D10), whereas in others (HB3 and the Gambian isolate) it is not essential.

Transcriptional alterations in parasites lacking PfAP2-HS under basal conditions. To gain insight on the molecular basis of the growth defects of some of the knockout lines, we compared the trophozoite transcriptome of 10E_Δ*pfap2-hs* with that of the parental 10E under basal (no heat shock) conditions. This revealed only a small set of genes with a fold decrease ≥ 2 in transcript levels, which included *hsp70-1*, the direct PfAP2-HS target *snoR04* RNA and several genes mainly involved in ribosome formation. The transcript levels for *hsp90* were also reduced (<twofold decrease) in the knockout line (Extended Data Fig. 8a,b and Supplementary Table 2). Reduced *hsp70-1* and *hsp90* transcript levels under basal conditions in 10E_Δ*pfap2-hs* mature trophozoites were independently confirmed by reverse transcription–qPCR and were also observed at the late ring stage and in the knockout lines with D10 and HB3 genetic backgrounds (Extended Data Fig. 8c). These results indicate that PfAP2-HS contributes to the regulation of the basal expression of the same chaperone-encoding genes that it activates following heat shock, among a few other genes. Together with the observation that the growth defect of the *pfap2-hs*-knockout lines is attenuated at 35 °C, this suggests that the knockout parasites have a reduced capacity for proteostasis, such that at 37 °C they are at the edge of proteostasis collapse. Parasite lines that can grow normally at 37 °C in spite of PfAP2-HS deletions including the three AP2 domains must therefore have alternative pathways active to ensure basal proteostasis. We hypothesise that mutant parasites expressing

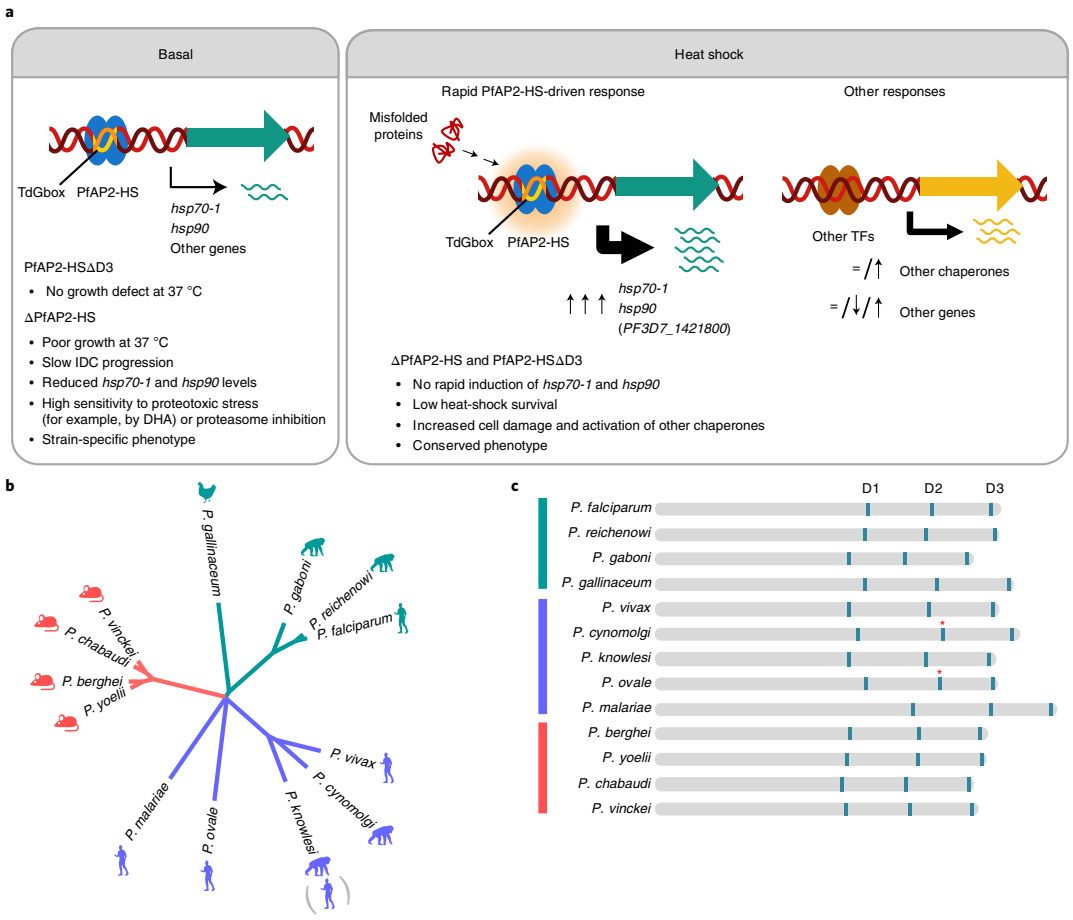


Fig. 6 | Model of the *P. falciparum* heat-shock response and phylogenetic analysis of AP2-HS. a, The *P. falciparum* heat-shock response involves rapid upregulation of the expression of a very restricted set of chaperones by PfAP2-HS. The *PF3D7_1421800* gene (in brackets) shows PfAP2-HS-dependent increased transcript levels following heat shock, but PfAP2-HS binding was not detected in its promoter and it lacks a G-box. The main defects associated with PfAP2-HS deletion or truncation, under heat shock (right) or basal (left) conditions, are listed. TFs, transcription factors. **b**, Phylogenetic analysis of the protein sequences of AP2-HS orthologues in *Plasmodium* spp. **c**, Schematic of the domain structure of the AP2-HS orthologues in *Plasmodium* spp. The position of the AP2 domains (D1–3) is based on domains identified in PlasmoDB release 50, except for those marked with an asterisk, which were annotated manually based on sequence alignments.

truncated PfAP2-HS are frequently selected under culture conditions because the truncations do not pose a fitness cost at 37 °C in the lines in which they appear and they prevent unnecessary activation of the heat-shock response, which can be detrimental[33], by unintended mild stress that may occur during culture.

PfAP2-HS-deficient parasites show increased sensitivity to artemisinin. Artemisinins are potent antimalarial drugs that kill parasites by causing general protein damage[34,35]. Resistance to artemisinin is associated with mutations in the Kelch13 protein[36,37] and involves cellular stress-response pathways such as the ubiquitin–proteasome system and the endoplasmic reticulum-based unfolded protein response (UPR)[34,35,38,39]. Given that PfAP2-HS regulates the expression of key chaperones, we tested the sensitivity of PfAP2-HS-deficient lines to dihydroartemisinin (DHA), the active metabolite of artemisinins. In all four different genetic backgrounds (3D7, D10, HB3 and Gambian Line 1), the knockout of *pfap2-hs* (or truncation before D1) resulted in a higher sensitivity to a pulse of DHA than in the lines with full PfAP2-HS (both at the ring or the trophozoite stage), whereas 10G showed increased sensitivity only when exposed at the trophozoite stage (Fig. 5a). These results indicate that deletion of PfAP2-HS renders parasites more sensitive to chemical proteotoxic stress, in addition to heat shock, probably as a consequence of basal defects in cellular proteostasis. We reasoned that if parasites lacking the PfAP2-HS protein bear constitutive proteome defects, they should have a low tolerance

to the disruption of other factors involved in proteostasis maintenance. The 10E_Δ*pfap2-hs* line was indeed more sensitive to the proteasome inhibitor epoxomicin than the parental 10E line or the 10G line (Fig. 5b). Furthermore, after heat shock there was greater accumulation of polyubiquitinated proteins in the knockout line than in 10E or 10G, reflecting higher levels of unresolved protein damage (Extended Data Fig. 9a). We also assessed the links between the PfAP2-HS-driven heat-shock response and the other main cell stress-response pathway—that is, the UPR. Using phosphorylation of eIF2α as a UPR marker, we found that the UPR does not depend on PfAP2-HS and is not directly activated by heat shock, because the marker was significantly elevated after heat shock only in the knockout line (Extended Data Fig. 9b).

Discussion

Our results show that the PfAP2-HS transcription factor is bound to the tandem G-box DNA motif in the promoter of the chaperone-encoding gene *hsp70-1* and rapidly upregulates the expression of this gene and, to a lesser extent, *hsp90* in response to febrile temperatures (Fig. 6a). Binding of PfAP2-HS to the G-box is mediated by D1 (ref. [23]) but rapid activation of *hsp70-1* and *hsp90* during heat shock requires D3, which is incapable of binding DNA in vitro[23] and in the cell is likely to participate in protein–protein interactions or dimerization[40]. Other components of the protein-folding machinery necessary for heat-shock survival[15–17,19] are either constitutively expressed or induced later, but the rapid PfAP2-HS-driven response is essential to avoid irreversible damage. Importantly, parasites lacking either the entire PfAP2-HS or its D3 cannot survive heat shock.

Although the sequence and domain organization of PfAP2-HS does not show any similarity with HSF1—the conserved master regulator of the heat-shock response in most eukaryotes, from yeast to mammals[3,6]—it serves an analogous role. HSF1 regulates a compact transcriptional programme that includes the *hsp70* and *hsp90* genes[4,5]. In yeast, the only essential role of this transcription factor is the activation of *hsp70* and *hsp90* (ref. [5]), the same chaperone-encoding genes activated by PfAP2-HS during heat shock. In addition to its role in the protective heat-shock response, PfAP2-HS is essential for growth at 37 °C in some *P. falciparum* genetic backgrounds. The function of PfAP2-HS under basal conditions is independent of its D3 domain. Several lines of evidence suggest that the role for PfAP2-HS under basal conditions involves proteostasis maintenance (Fig. 6a), similar to yeast HSF1 (ref. [5]). In other organisms, the heat-shock response mediates protection against different types of proteotoxic stress, in addition to high temperature[1,3]. Here we report that parasites lacking PfAP2-HS have increased sensitivity to artemisinin, and future research will be needed to establish the precise role of the *P. falciparum* heat-shock response in protection against different types of stress. We note that orthologues of *pfap2-hs* are present in all of the analysed *Plasmodium* spp. (Fig. 6b,c and Extended Data Fig. 10), including murine *Plasmodium* species that do not induce host fever. This observation suggests that, at least in these species, the heat-shock response regulated by AP2-HS may play a role in protection against different conditions.

Finally, while several ApiAP2 transcription factors regulate life-cycle transitions in malaria parasites[24,41–43], PfAP2-HS controls a protective response to a within-host environmental challenge. Our findings that the PfAP2-HS transcription factor regulates the activation of a protective heat-shock response settles the long-standing question of whether malaria parasites can respond to changes in within-host environmental conditions with specific transcriptional responses[44].

Methods

Parasite cultures. The 3D7-A stock of the clonal *P. falciparum* line 3D7 (ref. [45]), the 3D7-A subclones 10G, 1.2B, 10E, 4D, 6D, 1.2F, W4-1, W4-2, W4-4 and W4-5 (refs. [46,47]), the HB3B[48] (mosquito and chimpanzee-passaged HB3, provided by O. Kaneko, Ehime University, Japan) and D10 (ref. [49]; provided by R. F. Anders, La Trobe University, Australia) clonal parasite lines, and the culture-adapted Line 1 from The Gambia[30] have been previously described. The heat-shock-selected lines 3D7-A-HS r1 and r2 were derived from 3D7-A by exposing the cultures to 3 h of heat shock (41.5 °C) at the trophozoite stage for five consecutive cycles (each replicate, r1 and r2, is a fully independent selection from the 3D7-A stock), and the 3D7-A r1 and r2 lines are cultures that were maintained in parallel at 37 °C without heat shock[21]. Parasites were cultured in B^+ erythrocytes at a 3% haematocrit under standard culture conditions in RPMI-based media containing Albumax II (without human serum), in an atmosphere of 5% CO_2, 3% O_2 and the balance N_2 (except for cultures for ChIP–seq experiments, in which O^+ erythrocytes were used). Regular synchronization was performed using 5% sorbitol lysis, whereas tight synchronization (1, 2 or 5 h age window) was achieved by Percoll purification followed by sorbitol treatment 1, 2 or 5 h later. All cultures were regularly maintained at 37 °C, with the exception of the *pfap2-hs*-knockout lines that were maintained at 35 °C. For experiments performed in parallel with the knockout lines and other parasite lines, all cultures were maintained at 35 °C for at least one cycle before the experiment.

Generation of transgenic parasite lines. We used two single guide RNAs (hereafter referred to as sgRNA or guide) to knock out *pfap2-hs* (11,577 bp) using the CRISPR–Cas9 system[50] (Extended Data Fig. 2a and Supplementary Table 5). One guide targets a sequence near the 5′ end of the gene (positions 866–885 from the start codon), whereas the other recognizes a sequence near the 3′ end (positions 11,486–11,505). The 5′ guide was cloned into a modified pL6-*egfp* donor plasmid[50] in which the *yfcu* cassette had been removed by digestion with NotI and SacII, end blunting and re-ligation. The 5′ and 3′ homology regions (HR1, positions −2 to 808 of the gene; and HR2, from position +11520 of the gene to 490 bp after the stop codon, respectively) were also cloned in this plasmid, flanking the *hdhfr* expression cassette, to generate the plasmid pL7-pfap2hs_KO_sgRNA5′. The 3′ guide was cloned into a modified version of the pDC2-Cas9-U6-hdhfr[51] plasmid in which we previously removed the *hdhfr* expression cassette by digesting with NcoI and SacII, end blunting and re-ligation, and replaced the BbsI guide cloning site with a BtgZI site. The resulting plasmid was named pDC2_wo/hdhfr_pfap2hs_sgRNA3′. All guides were cloned using the In-Fusion system (Takara) as described[50], whereas homology regions were PCR-amplified from genomic DNA and cloned by ligation using the restriction sites SpeI and AflII (HR1), and EcoRI and NcoI (HR2).

For the constructs aimed at C-terminal tagging of *pfap2-hs* using CRISPR–Cas9 (10E_*pfap2-hs*_eYFP-Cterm and 10E_*pfap2-hs*_3×HA-Cterm lines), we used a guide corresponding to positions 11508–11527 of the gene (Extended Data Fig. 2b,c and Supplementary Table 5). The guide was cloned into the pDC2-Cas9-U6-hdhfr[51] plasmid to obtain pDC2_pfap2hs_sgRNA-C. The donor plasmid for tagging with enhanced yellow fluorescent protein (eYFP; pHR-C_pfap2hs_eYFP) was based on the plasmid pHRap2g-eYFP[52], with the *pfap2-g* homology regions and the 3′ sequence of *hsp90* replaced with *pfap2-hs* homology regions. The HR1 region was generated with a PCR-amplified fragment spanning from nucleotide 10964 to the sequence of the guide (recodonized) and a 47-bp fragment (generated by annealing two complementary oligonucleotides) consisting of a recodonized version of the remaining nucleotides to the end of the gene. The two fragments were cloned simultaneously into the SpeI–BglII sites using the In-Fusion system. The HR2 region was a PCR fragment spanning positions +1 to +590 after the *pfap2-hs* stop codon. It was cloned into the XhoI–AatII restriction sites. The donor plasmid for 3×HA C-terminal tagging (pHR-C_pfap2hs_3×HA_hsp90-3′) was also based on the plasmid pHRap2g-eYFP[52], with the eYFP coding sequence replaced by the 3×HA sequence (amplified from the plasmid pHH1inv-pfap2-g-HA×3; ref. [41]) and the same homology regions as in the plasmid pHR-C_pfap2hs_eYFP (but HR2 was cloned, using the In-Fusion system, into the EcoRI–AatII sites because in this construct the 3′ region of *hsp90* in pHRap2g-eYFP was maintained).

For amino (N)-terminal tagging (10E_*pfap2-hs*_eYFP-Nterm line), we cloned a guide targeting the *pfap2-hs* positions 73–92 in the pDC2-Cas9-U6-hDHFRyFCU[53] plasmid to obtain the plasmid pDC2_pfap2hs_sgRNA-N (Extended Data Fig. 2d and Supplementary Table 5). The donor plasmid (pfap2hs_HR-N_eYFP) consisted of a HR1 region including positions −366 to −1 relative to the *pfap2-hs* start codon, the *eYFP* gene and an in-frame HR2 spanning positions 4–756 of the gene (excluding the ATG) in which the nucleotides up to the position of the guide were recodonized. The HR1 and HR2 regions were cloned into the SacII–NcoI and SpeI–EcoRI sites, respectively. HR2 was amplified in two steps using a nested PCR approach to add the recodonized sequences. The eYFP fragment (PCR-amplified from plasmid pHR-C_pfap2hs_eYFP) was cloned using the In-Fusion system into the SpeI–NcoI sites.

To tag PfAP2-HS with a 2×HA-ddFKBP domain tag (1.2B_ *pfap2-hs*_ddFKBP line), we used a single homologous recombination approach (Extended Data Fig. 2e). To generate the pfap2hs_HA-ddFKBP plasmid, we replaced the *pfap2-g* homology region in the plasmid PfAP2-G-ddFKBP[41] with a PCR-amplified fragment including positions 9551–11574 of *pfap2-hs* in frame with the tag. The fragment was cloned using the restriction sites NotI and XhoI. All of

the oligonucleotides that were used to generate the plasmids are described in Supplementary Table 5. The relevant parts of all plasmids (that is, the new sequences incorporated) were sequenced before transfection.

The transfections were performed by electroporation of ring-stage cultures with 100 µg plasmid (HA-ddFKBP tagging) or a mixture of 12 µg linearized donor plasmid and 60 µg of circular Cas9 plasmid (CRISPR–Cas9 system). Linearization was achieved by digestion with the PvuI restriction enzyme (cleaving the *amp* resistance gene of the donor plasmid). Transfected cultures were selected with 10 nM WR99210 for four days as previously described[53] (transfections using the CRISPR–Cas9 system) or with continuous WR99210 pressure until parasites were observed, followed by three off–on drug cycles and subcloning by limiting dilution (transfections with the pfap2hs_HA-ddFKBP plasmid). In all cases, we used analytical PCR of genomic DNA (Extended Data Fig. 2) with specific primers (Supplementary Table 5) to assess correct integration.

Heat-shock resistance assay. Heat shock was always performed on cultures at the mature trophozoite stage unless otherwise indicated. To measure heat-shock survival, the cultures were tightly synchronized to a defined age window, diluted to 1% parasitaemia, split between two identical petri dishes (heat shock and control) maintained in independent air-tight incubation chambers and exposed to heat shock when the majority of parasites were at the mature trophozoite stage (typically about 30–35 h.p.i.; the Δ*pfap2-hs* lines were tightly synchronized 3 h earlier than the other lines but exposed to heat shock in parallel to account for their slower IDC progression). The exception was experiments to screen many subclones (that is, Fig. 1f) or to characterize transgenic parasite lines (that is, Extended Data Fig. 2) in which cultures were only sorbitol-synchronized and heat shock was performed approximately 20–25 h after sorbitol lysis (mature trophozoite stage). For heat shock, the full incubation chamber was transferred to an incubator at 41.5 °C for 3 h and then returned to 37 or 35 °C (the latter temperature was used for all lines in experiments including the *pfap2-hs*-knockout lines). The chamber with the control cultures was always maintained at 37 or 35 °C. After reinvasion (typically approximately 60–70 h after synchronization to ensure that all parasites had completed the cycle, including parasites subjected to heat shock that demonstrated delayed progression through the IDC), parasitaemia of the control and heat-shock-exposed cultures was measured by flow cytometry using a FACSCalibur flow cytometer (Becton Dickinson) and SYTO 11 to stain the nucleic acids (Supplementary Fig. 1), as previously described[54].

Phenotypic characterization. To determine the growth rate (increase in parasitaemia between consecutive cycles) at different temperatures, the parasitaemia of sorbitol-synchronized cultures was adjusted to 1% and then accurately determined by flow cytometry. The cultures were then split between two or three dishes and maintained in parallel in incubators at the different temperatures tested. Parasitaemia was again determined by flow cytometry at the next cycle to determine the growth rate. To measure the duration of the IDC (at 35 °C) in the different parasite lines, we used a recently described method based on synchronization to a 1 h age window achieved by Percoll purification of schizonts, followed by sorbitol lysis 1 h later[54]. The determination of the number of merozoites per fully mature schizont was based on light-microscopy analysis of Giemsa-stained smears from Percoll-purified schizonts[54]. Sensitivity to DHA (Sigma, cat. no. D7439) or epoxomicin (Selleckchem, cat. no. S7038) was measured after exposing tightly synchronized cultures (1% parasitaemia) at the ring (10–15 h.p.i.; DHA only) or trophozoite (30–35 h.p.i.; DHA or epoxomicin) stage to a 3 h pulse of the compounds at different concentrations (DHA, 2.5, 5, 10, 20 or 200 nM, which is lower than the approximately 700 nM plasma concentration after patient treatment that kills the vast majority of sensitive parasites[35]; epoxomicin, 100 or 150 nM, which is higher than the reported 7.7 nM IC_{50} after exposing parasites for 50 h (ref. [55]) and similar to the concentration used in previous studies with a 3 h pulse[29,38]). Parasitaemia was measured by light-microscopy analysis of Giemsa-stained smears at the next cycle (typically 70–80 h after Percoll–sorbitol synchronization). For these experiments, the Δ*pfap2-hs* lines were tightly synchronized 3 h earlier than the other lines but exposed to DHA or epoxomicin in parallel (13–18 or 33–38 h.p.i.) to account for their slower IDC progression. For the DHA experiments, the drug concentrations were log-transformed and the per cent survival data were fit to sigmoidal dose-response curves to calculate the IC_{50} values using GraphPad Prism.

Transcriptional analysis by reverse transcription–qPCR. RNA from tightly synchronized cultures exposed to heat shock and their controls was obtained using the TRIzol method, DNAse treated and purified essentially as described. Reverse transcription and qPCR analysis of the complementary DNA were also performed as described before[56,57]. Briefly, a mixture of random primers and oligo (dT) were used for reverse transcription, and for qPCR, we used the PowerSYBR Green master mix (Applied Biosystems) and the standard-curve method (each plate included a standard curve for each primer pair). All of the primers used are listed in Supplementary Table 5. Unless otherwise indicated, the transcript levels were normalized to *serine–tRNA ligase* (*PF3D7_0717700*), which shows relatively stable expression throughout the IDC.

Transcriptomic analysis using microarrays. To compare the transcriptome of control and heat-shock-adapted 3D7-A parasite lines across the IDC, we used previously described two-colour long-oligonucleotide-based glass microarrays[21]. RNA was obtained from tightly synchronized cultures (5 h age window) at 8–13, 16–21, 24–29, 32–37 and 40–45 h.p.i. All samples (Cy5-labelled) were hybridized together with a reference pool (Cy3-labelled) consisting of a mixture of equal amounts of cDNA from rings, trophozoites and schizonts from control and heat-shock-adapted lines. Comparative genome hybridization was used to determine which differences in transcript levels were potentially attributable to genetic deletions or duplications. Useful data were obtained from 5,142 genes. Sample preparation, microarray hybridization and data analysis were performed essentially as described previously[21].

To analyse the transcriptome of the 10E, 10G and 10E_Δ*pfap2-hs* parasite lines under control and heat-shock conditions, we used two-colour long-oligonucleotide-based custom Agilent microarrays[58]. The microarray design was based on Agilent design AMADID no. 037237 (refs. [58,59]), but we modified it as previously described (new design AMADID no. 084561)[60]. RNA was obtained from cultures synchronized to a 5 h age window at about 2.5% parasitaemia. Given the slower IDC progression of 10E_Δ*pfap2-hs*, cultures of this parasite line were synchronized to 0–5 h.p.i. 3 h earlier than the 10E and 10G cultures, such that all cultures were approximately at the same stage of IDC progression when heat shock was initiated (in parallel for all lines). Heat shock was started at 30–35 h.p.i. (33–38 h.p.i. for the 10E_Δ*pfap2-hs* line) and samples were collected before, during and after heat shock, as indicated. RNA was prepared using the TRIzol method. Sample preparation and microarray hybridization were performed essentially as described[59]. All samples (Cy5-labelled) were hybridized together with a reference pool (Cy3-labelled) consisting of a mixture of equal amounts of cDNA from rings, trophozoites and schizonts from 3D7-A. Microarray images were obtained using a DNA microarray scanner (Agilent Technologies, cat. no. G2505C) located in a low-ozone area and initial data processing was performed using the GE2 _1105_ Oct12 extraction protocol in the Agilent Feature Extraction 11.5.1.1 software.

Agilent microarray data were analysed using Bioconductor in an R environment (R version 3.5.3). For each microarray, we calculated the Cy3 and Cy5 background signal as the median of the 100-lowest signal probes for each channel, and probes with both Cy3 and Cy5 signals below three times the array background were excluded. Gene-level $\log_2$(Cy5/Cy3) values, statistical estimation of parasite age[61] and estimation of the average fold differences in expression across a time interval (for the comparison between parasite lines in the absence of heat shock) were performed as described[21]. The $\log_2$-transformed value of the fold change in expression following heat shock was calculated, for each gene and time point, as the $\log_2$(Cy5/Cy3) of the heat-shock-exposed sample minus the $\log_2$(Cy5/Cy3) of the control sample at the same parasite age, calculated using linear interpolation in the $\log_2$(Cy5/Cy3) versus estimated age plot. Visual inspection was used to exclude genes with apparent artefacts from further analyses. Genes that were missing data for ≥2 time points (or ≥1 for the comparison between parasite lines in the absence of heat shock across a time interval) or with values within the lowest 15th percentile of expression intensity (Cy5 sample channel) in all samples were also excluded from further analyses. Useful data were obtained from 4,964 genes.

To assess the level of similarity between the transcriptome of our samples and a reference non-stressed transcriptome with high temporal resolution (HB3 line)[28], we calculated the Pearson's correlation between each sample and the time point of the reference transcriptome with which it had a higher similarity. Heatmaps and hierarchical clustering based on Spearman's (Fig. 2) or Pearson's (Extended Data Fig. 6) correlation were generated using Multiple Experiment Viewer (MeV) 4.9 (ref. [62]). Expression trend plots for each cluster were generated using ggplot2, with LOESS smoothing, and Venn diagrams were generated using the eulerr package (both in an R environment). Motif analysis (5–8 bp) was performed using the MEME 5.0.3 software. Functional enrichment analysis using Gene Ontology terms annotated in PlasmoDB release 43 was performed using the Ontologizer 2.1 software[63] with the topology-elim method[64]. Gene set enrichment analysis was performed using GSEA v3.0 Preranked[65].

Whole-genome sequencing analysis, analysis of publicly available genome sequences from field isolates and phylogenetic analysis. To sequence the full genome of control and heat-shock-adapted 3D7-A lines (two biological replicates), we used PCR-free whole-genome Illumina sequencing. Briefly, genomic DNA was sheared to fragments of approximately 150–400 bp in length using a Covaris S220 ultrasonicator and analysed using an Agilent 2100 Bioanalyzer. For library preparation, we used the NEBNext DNA library prep master mix set for Illumina (cat. no. E6040S) using specific paired-end TruSeq Illumina adaptors for each sample. After a quality check by qPCR, we obtained more than six million 150-bp paired reads for each sample using an Illumina MiSeq sequencing system. After checking the quality of the reads (FastQC algorithm) and trimming adaptors (Cutadapt algorithm), the sequence reads were mapped to the PlasmoDB *P. falciparum* 3D7 reference genome release 24 (https://plasmodb.org/plasmo/) using the Bowtie2 local alignment algorithm and the alignment quality was assessed using the QualiMap platform. Average genome coverage was 76- to 98-fold. To identify SNPs and small insertion–deletion mutations (indels), we followed the Genome Analysis Toolkit (GATK)

best-practices workflow, using SAMtools, PicardTools and GATK algorithms. Variant calling was performed using GATK-UnifiedGenotyper. Variants with low calling quality (Phred QUAL < 20) and low read depth (DP < 10) were filtered out using GATK-VariantFiltration, and only variants present in both biological replicates were considered. Differences in the frequency of the SNPs or indels between the control and heat-shock-adapted lines were calculated for each SNP and indel using Microsoft Excel, and those showing <25% difference in any of the two replicates were filtered out. Genome Browse (Golden Helix) was used to visualize the alignments and variants.

For the analysis of publicly available genome sequences, we used the Pf3K project (2016) pilot data release 5 (www.malariagen.net/data/pf3k-5) containing the sequence of >2,500 field isolates. Only SNPs that passed all quality filters and did not fall within a region with multiple large insertions and deletions were considered to be high-confidence. Using these criteria, a single high-confidence polymorphism (occurring in a single isolate) was identified at the *pfap2-hs* gene (producing the C3168X mutation that results in a truncated PfAP2-HS protein that lacks D3).

For sequence alignment and construction of the phylogenetic tree (using the neighbour-joining method), we used Clustal Omega[66] with the default parameters. The cladogram was generated (from the tree without distance corrections) using FigTree 1.4.4.

ChIP experiments and data analysis. For the ChIP experiments, synchronous 50 ml cultures at 2.5–5% parasitemia were harvested at the mid-trophozoite stage. For replicates in which ChIP was performed in parallel under heat-shock and control conditions, the cultures were split off from a single parent flask at the mid-trophozoite stage. Control flasks were immediately returned to 37 °C, whereas the heat-shock flasks were maintained at 41.5 °C for 3 h before harvesting for ChIP analysis.

ChIP, followed by qPCR or Illumina sequencing, was performed as described[67] using the 3F10 rat anti-HA antibody (1:500; Roche, cat. no. 11867423001) to immunoprecipitate the HA-tagged AP2-HS, with the following modification: total chromatin was diluted fivefold in dilution buffer following sonication. The Illumina HiSeq system was used to obtain 125-bp paired-end (replicates 1–3) or 150-bp single-end (replicates 4–5) reads.

Analysis of the ChIP–seq data was performed essentially as described[67]. Briefly, after trimming, quality control, mapping the remaining reads to the *P. falciparum* genome (PlasmoDB release 28) using BWA-MEM and filtering duplicated reads, peak calling was performed using MACS2 (ref. [68]) with a *q*-value cutoff of 0.01. Conversion to $\log_2$-transformed coverage of immunoprecipitate/input was performed using DeepTools BamCompare, selecting the paired-end parameter for all tools when analysing experiments including control and heat-shock conditions. Overlapping intervals within the called peaks for each dataset were determined using Bedtools MultiIntersect. The closest annotated gene coding sequence for each called peak was determined using Bedtools ClosestBed. To visualize the aligned data, we used the Integrative Genomics Viewer (IGV).

The ChIP samples were analysed by qPCR in triplicate wells using the primers described in Supplementary Table 5. All primer pairs were confirmed to have between 80 and 110% efficiency using sheared genomic DNA as a template control. The per cent input was calculated using the formula $100 \times 2^{(Ct\ \text{adjusted input} - Ct\ IP)}$, where C_t adjusted input is the cycle threshold of the adjusted input and C_t IP is the cycle threshold of the immunoprecipitate.

Western blotting. Synchronized cultures at the mature trophozoite stage were exposed to a regular 3 h heat shock or a 1.5 h DHA pulse (10 or 100 nM, used as a positive control for a condition known to produce proteotoxic stress and induce the UPR)[29,38]. Parasites were obtained using saponin lysis (0.15% wt/vol saponin); the pellets were solubilized in 1×SDS–PAGE loading buffer with 4% β-mercaptoethanol and boiled at 95 °C for 5 min. Proteins were resolved by SDS–PAGE on 4–20% TGX Mini-PROTEAN gels (Bio-Rad) and transferred to nitrocellulose membranes (Bio-Rad). After blocking with 5% (wt/vol) bovine serum albumin (Biowest) in TBS-T (0.1% Tween 20 in Tris-buffered saline), the membranes were incubated at 4 °C overnight with the following primary antibodies: rabbit anti-ubiquitin (1:1,000; Cell Signaling Technology, cat. no. 3933), rabbit anti-phospho-eIF2α (1:1,000; Cell Signaling Technology, cat. no. 3398) and rabbit anti-histone H3 (1:1,000; Cell Signaling Technology, cat. no. 9715). After incubation with a goat anti-rabbit IgG–peroxidase (1:5,000; Millipore, cat. no. AP307P) secondary antibody, peroxidase was detected using the Pierce ECL western blotting substrate (Thermo Fisher Scientific) in an ImageQuant LAS 4000 imaging system. To control for equal loading, parts of the membranes corresponding to different molecular-weight ranges were separately hybridized with different antibodies. Signal quantification was performed using ImageJ.

Statistical analysis. Statistical analysis was performed using Microsoft Excel and GraphPad Prism. *P* values were calculated using a two-tailed Student's *t*-test (equal variance). No adjustment for multiple comparisons was made. Only significant *P* values ($P < 0.05$) are shown in the figures. No statistical analysis was performed for experiments involving only two replicates. In all cases, *n* indicates independent biological replicates (that is, samples were obtained from independent cultures).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The microarray data presented in Fig. 2 and Extended Data Figs. 1,5,6,8 has been deposited to the Gene Expression Omnibus (GEO) database under the accession code GSE149394. The genome sequencing and ChIP–seq data presented in Figs. 1b, 2e and Extended Data Fig. 7 have been deposited to the Sequence Read Archive (SRA) database with the accession codes PRJNA626524 and PRJNA670721, respectively. The authors declare that all other relevant data generated or analysed during this study are included in the Article, the Extended Data or the Supplementary Information files. We used data from the Pf3k pilot data release 5 (www.malariagen.net/pf3k) and different releases of PlasmoDB (www.plasmodb.org) databases. The materials described in this article, including the *P. falciparum* transgenic lines, are available from the corresponding author on reasonable request. Source data are provided with this paper.

Code availability

The scripts used for the analysis of microarray and next generation sequencing data are available at github (https://github.com/CortesMalariaLab/PfAP2-HS_Tinto_etal_NatMicrobiol_2021, with https://doi.org/10.5281/zenodo.4775988).

Received: 12 March 2021; Accepted: 22 June 2021;
Published online: 16 August 2021

References

1. Richter, K., Haslbeck, M. & Buchner, J. The heat shock response: life on the verge of death. *Mol. Cell* **40**, 253–266 (2010).
2. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).
3. Anckar, J. & Sistonen, L. Regulation of HSF1 function in the heat stress response: implications in aging and disease. *Annu. Rev. Biochem* **80**, 1089–1115 (2011).
4. Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Mol. Cell* **62**, 63–78 (2016).
5. Solis, E. J. et al. Defining the essential function of yeast Hsf1 reveals a compact transcriptional program for maintaining eukaryotic proteostasis. *Mol. Cell* **63**, 60–71 (2016).
6. Gomez-Pastor, R., Burchfiel, E. T. & Thiele, D. J. Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 4–19 (2018).
7. Milner, D. A. Jr. Malaria Pathogenesis. *Cold Spring Harb. Perspect. Med.* **8**, a025569 (2018).
8. Kwiatkowski, D. Malarial toxins and the regulation of parasite density. *Parasitol. Today* **11**, 206–212 (1995).
9. Oakley, M. S., Gerald, N., McCutchan, T. F., Aravind, L. & Kumar, S. Clinical and molecular aspects of malaria fever. *Trends Parasitol.* **27**, 442–449 (2011).
10. Gravenor, M. B. & Kwiatkowski, D. An analysis of the temperature effects of fever on the intra-host population dynamics of *Plasmodium falciparum*. *Parasitology* **117**, 97–105 (1998).
11. Kwiatkowski, D. Febrile temperatures can synchronize the growth of *Plasmodium falciparum* in vitro. *J. Exp. Med* **169**, 357–361 (1989).
12. Long, H. Y., Lell, B., Dietz, K. & Kremsner, P. G. *Plasmodium falciparum*: in vitro growth inhibition by febrile temperatures. *Parasitol. Res.* **87**, 553–555 (2001).
13. Portugaliza, H. P. et al. Artemisinin exposure at the ring or trophozoite stage impacts *Plasmodium falciparum* sexual conversion differently. *eLife* **9**, e60058 (2020).
14. Pavithra, S. R., Kumar, R. & Tatu, U. Systems analysis of chaperone networks in the malarial parasite *Plasmodium falciparum*. *PLoS Comput Biol.* **3**, 1701–1715 (2007).
15. Muralidharan, V., Oksman, A., Pal, P., Lindquist, S. & Goldberg, D. E. *Plasmodium falciparum* heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. *Nat. Commun.* **3**, 1310 (2012).
16. Silva, M. D. et al. A role for the *Plasmodium falciparum* RESA protein in resistance against heat shock demonstrated using gene disruption. *Mol. Microbiol* **56**, 990–1003 (2005).
17. Kudyba, H. M. et al. The endoplasmic reticulum chaperone PfGRP170 is essential for asexual development and is linked to stress response in malaria parasites. *Cell. Microbiol.* **21**, e13042 (2019).
18. Lu, K. Y. et al. Phosphatidylinositol 3-phosphate and Hsp70 protect *Plasmodium falciparum* from heat-induced cell death. *eLife* **9**, e56773 (2020).
19. Zhang, M. et al. The apicoplast link to fever-survival and artemisinin-resistance in the malaria parasite. Preprint at *BioRxiv* https://doi.org/10.1101/2020.12.10.419788 (2021).

20. Oakley, M. S. et al. Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic *Plasmodium falciparum* parasites. *Infect. Immun.* **75**, 2012–2025 (2007).
21. Rovira-Graells, N. et al. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. *Genome Res.* **22**, 925–938 (2012).
22. 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).
23. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinas, M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6**, e1001165 (2010).
24. Jeninga, M. D., Quinn, J. E. & Petter, M. ApiAP2 transcription factors in apicomplexan parasites. *Pathogens* **8**, 47 (2019).
25. Militello, K. T., Dodge, M., Bethke, L. & Wirth, D. F. Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* **134**, 75–88 (2004).
26. Dobson, C. M., Sali, A. & Karplus, M. Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.* **37**, 868–893 (1998).
27. Masterton, R. J., Roobol, A., Al-Fageeh, M. B., Carden, M. J. & Smales, C. M. Post-translational events of a model reporter protein proceed with higher fidelity and accuracy upon mild hypothermic culturing of Chinese hamster ovary cells. *Biotechnol. Bioeng.* **105**, 215–220 (2010).
28. Bozdech, Z. et al. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, E5 (2003).
29. Dogovski, C. et al. Targeting the cell stress response of *Plasmodium falciparum* to overcome artemisinin resistance. *PLoS Biol.* **13**, e1002132 (2015).
30. Claessens, A., Affara, M., Assefa, S. A., Kwiatkowski, D. P. & Conway, D. J. Culture adaptation of malaria parasites selects for convergent loss-of-function mutants. *Sci. Rep.* **7**, 41303 (2017).
31. Stewart, L. B. et al. Intrinsic multiplication rate variation and plasticity of human blood stage malaria parasites. *Commun. Biol.* **3**, 624 (2020).
32. Manske, M. et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
33. Lamech, L. T. & Haynes, C. M. The unpredictability of prolonged activation of stress response pathways. *J. Cell Biol.* **209**, 781–787 (2015).
34. Blasco, B., Leroy, D. & Fidock, D. A. Antimalarial drug resistance: linking *Plasmodium falciparum* parasite biology to the clinic. *Nat. Med.* **23**, 917–928 (2017).
35. Haldar, K., Bhattacharjee, S. & Safeukui, I. Drug resistance in *Plasmodium*. *Nat. Rev. Microbiol.* **16**, 156–170 (2018).
36. Ariey, F. et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* **505**, 50–55 (2014).
37. Birnbaum, J. et al. A Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. *Science* **367**, 51–59 (2020).
38. Bridgford, J. L. et al. Artemisinin kills malaria parasites by damaging proteins and inhibiting the proteasome. *Nat. Commun.* **9**, 3801 (2018).
39. Mok, S. et al. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. *Science* **347**, 431–435 (2015).
40. Lindner, S. E., De Silva, E. K., Keck, J. L. & Llinas, M. Structural determinants of DNA binding by a *P. falciparum* ApiAP2 transcriptional regulator. *J. Mol. Biol.* **395**, 558–567 (2010).
41. Kafsack, B. F. et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–252 (2014).
42. 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).
43. Zhang, C. et al. Systematic CRISPR–Cas9-mediated modifications of *Plasmodium yoelii* ApiAP2 genes reveal functional insights into parasite development. *mBio* **8**, e01986-17 (2017).
44. Llora-Batlle, O., Tinto-Font, E. & Cortes, A. Transcriptional variation in malaria parasites: why and how. *Brief. Funct. Genomics* **18**, 329–341 (2019).
45. Cortés, A., Benet, A., Cooke, B. M., Barnwell, J. W. & Reeder, J. C. Ability of *Plasmodium falciparum* to invade Southeast Asian ovalocytes varies between parasite lines. *Blood* **104**, 2961–2966 (2004).
46. Cortés, A. A chimeric *Plasmodium falciparum Pfnbp2b/Pfnbp2a* gene originated during asexual growth. *Int. J. Parasitol.* **35**, 125–130 (2005).
47. Cortés, A. et al. Epigenetic silencing of *Plasmodium falciparum* genes linked to erythrocyte invasion. *PLoS Pathog.* **3**, e107 (2007).
48. Walliker, D. et al. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**, 1661–1666 (1987).
49. Anders, R. F., Brown, G. V. & Edwards, A. Characterization of an S antigen synthesized by several isolates of *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **80**, 6652–6656 (1983).
50. Ghorbal, M. et al. Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR–Cas9 system. *Nat. Biotechnol.* **32**, 819–821 (2014).
51. Lim, M. Y. et al. UDP-galactose and acetyl-CoA transporters as *Plasmodium* multidrug resistance genes. *Nat. Microbiol* **1**, 16166 (2016).
52. Bancells, C. et al. Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. *Nat. Microbiol.* **4**, 144–154 (2019).
53. Knuepfer, E., Napiorkowska, M., van Ooij, C. & Holder, A. A. Generating conditional gene knockouts in *Plasmodium*—a toolkit to produce stable DiCre recombinase-expressing parasite lines using CRISPR/Cas9. *Sci. Rep.* **7**, 3881 (2017).
54. Rovira-Graells, N., Aguilera-Simon, S., Tinto-Font, E. & Cortes, A. New assays to characterise growth-related phenotypes of *Plasmodium falciparum* reveal variation in density-dependent growth inhibition between parasite lines. *PLoS ONE* **11**, e0165358 (2016).
55. Prasad, R. et al. Blocking *Plasmodium falciparum* development via dual inhibition of hemoglobin degradation and the ubiquitin proteasome system by MG132. *PLoS ONE* **8**, e73530 (2013).
56. Crowley, V. M., Rovira-Graells, N., de Pouplana, L. R. & 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).
57. Casas-Vila, N., Pickford, A. K., Portugaliza, H. P., Tintó-Font, E. & Cortés, A. Transcriptional analysis of tightly synchronized *Plasmodium falciparum* intraerythrocytic stages by RT–qPCR. *Methods Mol. Biol.* https://doi.org/10.1007/978-1-0716-1681-9_10 (2021).
58. Kafsack, B. F., Painter, H. J. & Llinas, M. New Agilent platform DNA microarrays for transcriptome analysis of *Plasmodium falciparum* and *Plasmodium berghei* for the malaria research community. *Malar. J.* **11**, 187 (2012).
59. Painter, H. J., Altenhofen, L. M., Kafsack, B. F. & Llinas, M. Whole-genome analysis of *Plasmodium* spp. utilizing a new agilent technologies DNA microarray platform. *Methods Mol. Biol.* **923**, 213–219 (2013).
60. Llorà-Batlle, O. et al. Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*. *Sci. Adv.* **6**, eaaz5057 (2020).
61. 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. USA* **106**, 7559–7564 (2009).
62. Saeed, A. I. et al. TM4 microarray software suite. *Methods Enzymol.* **411**, 134–193 (2006).
63. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**, 1650–1651 (2008).
64. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
65. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
66. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
67. Josling, G. A. et al. Dissecting the role of PfAP2-G in malaria gametocytogenesis. *Nat. Commun.* **11**, 1503 (2020).
68. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

Acknowledgements

We thank J. J. López-Rubio (University of Montpellier) for the plasmid pL6-*egfp*, M. Lee (Wellcome Sanger Institute) for the plasmid pDC2-Cas9-U6-hdhfr and E. Knuepfer (The Francis Crick Institute) for the plasmid pDC2-Cas9-U6-hDHFRyFCU. We also thank O. Llorà-Batlle and C. Sànchez-Guirado (ISGlobal) for their assistance with the generation of the plasmids used in this study, N. Rovira-Graells (ISGlobal) and A. Gupta. (Nanyang Technological University) for assistance with the 3D7-A and 3D7-A-HS microarray experiments, O. Billker (Wellcome Sanger Institute) for the experiments attempted in *P. berghei* and H. Ginsburg (The Hebrew University of Jerusalem) for providing data from the Malaria Parasite Metabolic Pathways. This publication uses data generated by the Pf3k project (www.malariagen.net/pf3k). This work was supported by grants from the Spanish Ministry of Economy and Competitiveness (MINECO)/Agencia Estatal de Investigación (AEI) (grant nos SAF2013-43601-R, SAF2016-76190-R and PID2019-107232RB-I00 to A.C.), co-funded by the European Regional Development Fund (ERDF, European Union), and from the NIH/NIAID (grant no. 1R01 AI125565 to M.L.). E.T.-F. and L.M.-T. were supported by fellowships from the Spanish Ministry of Economy and Competitiveness (grant nos BES-2014-067901 and BES-2017-081079, respectively), co-funded by the European Social Fund (ESF). T.J.R. was supported by a training grant from the NIH/NIGMS (grant no. T32 GM125592-01). This research is part of ISGlobal's Program on the Molecular Mechanisms of Malaria, which is partially supported by the Fundación Ramón Areces. We acknowledge support from the Spanish Ministry of Science and Innovation through the 'Centro de Excelencia Severo Ochoa 2019–2023' Program (grant no. CEX2018-000806-S) and support from the Generalitat de Catalunya through the CERCA Program.

Author contributions

E.T.-F. performed all of the experiments, except for those presented in Extended Data Fig. 1 and the western blot and ChIP–seq experiments. L.M.-T., E.T.-F., T.J.R. and A.C. performed the bioinformatics analyses. N.C.-V. performed the western blot experiments. T.J.R. performed, and M.L. supervised, the ChIP–seq experiments. Z.B. provided microarray hybridizations for the experiments presented in Extended Data Fig. 1. D.J.C. provided advice on clinical isolates and provided Line 1 from The Gambia. E.T.-F. and A.C. conceived the project, designed and interpreted the experiments and wrote the manuscript (with input from all authors and major input from M.L. and D.J.C.).

Competing interests

The authors declare no competing interests.

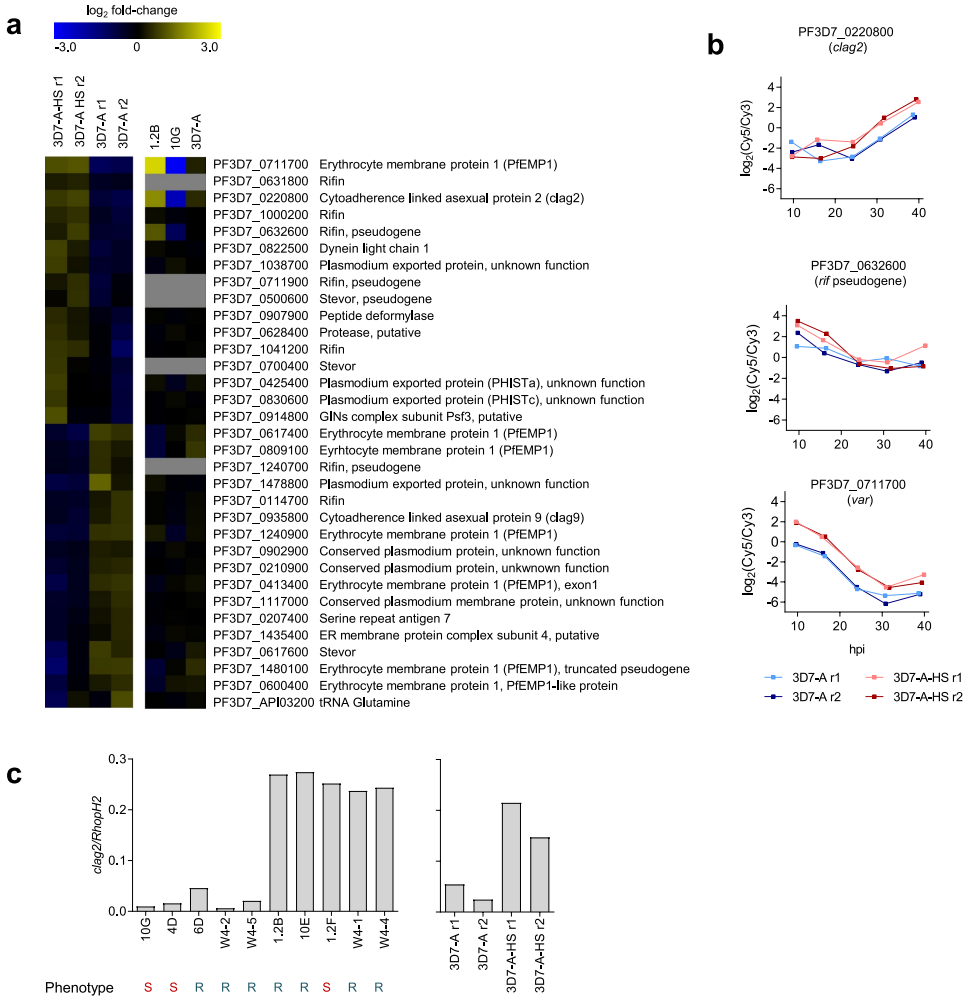
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-021-00940-w.

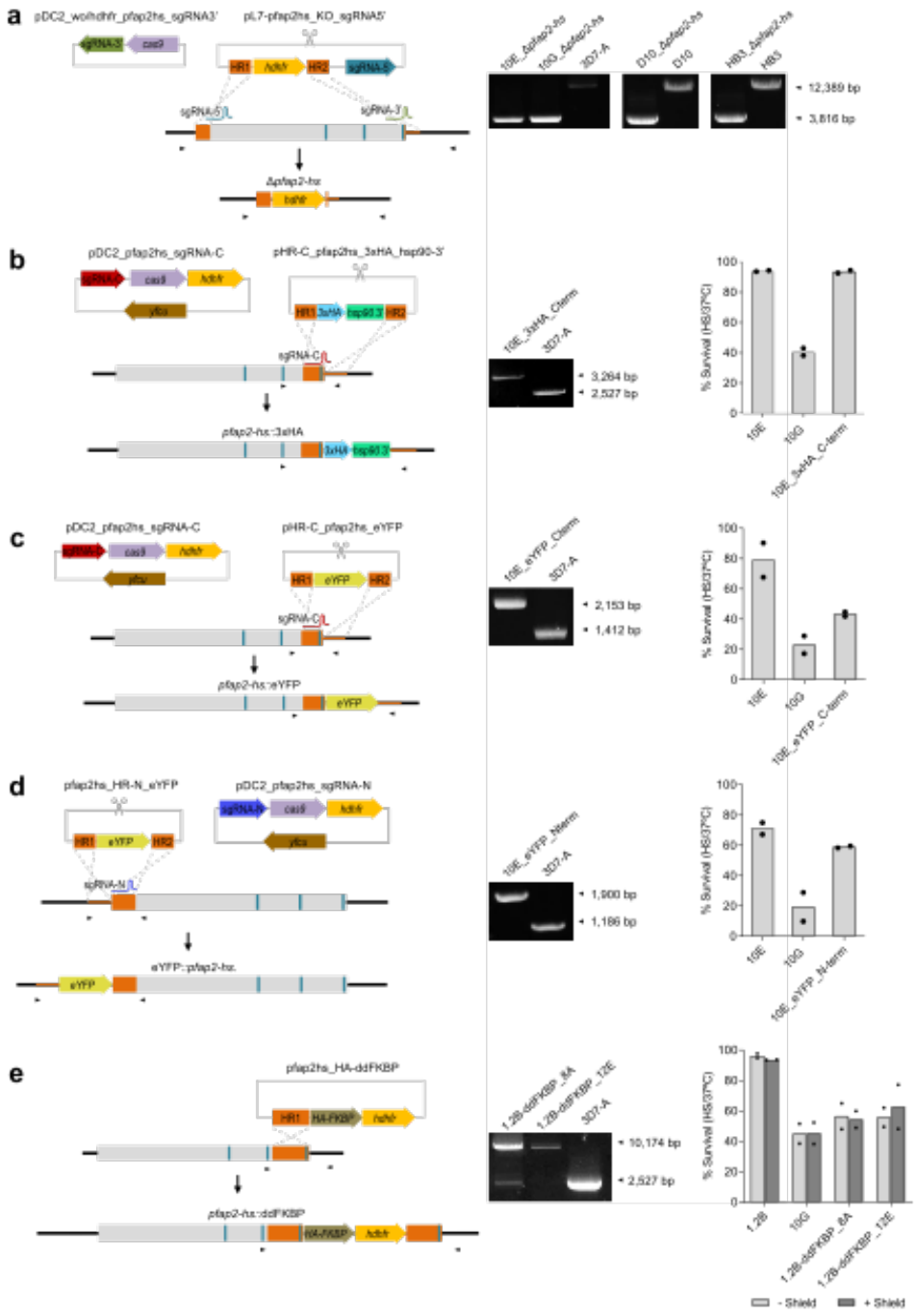
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-00940-w.

Correspondence and requests for materials should be addressed to A.C.

Peer review information *Nature Microbiology* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

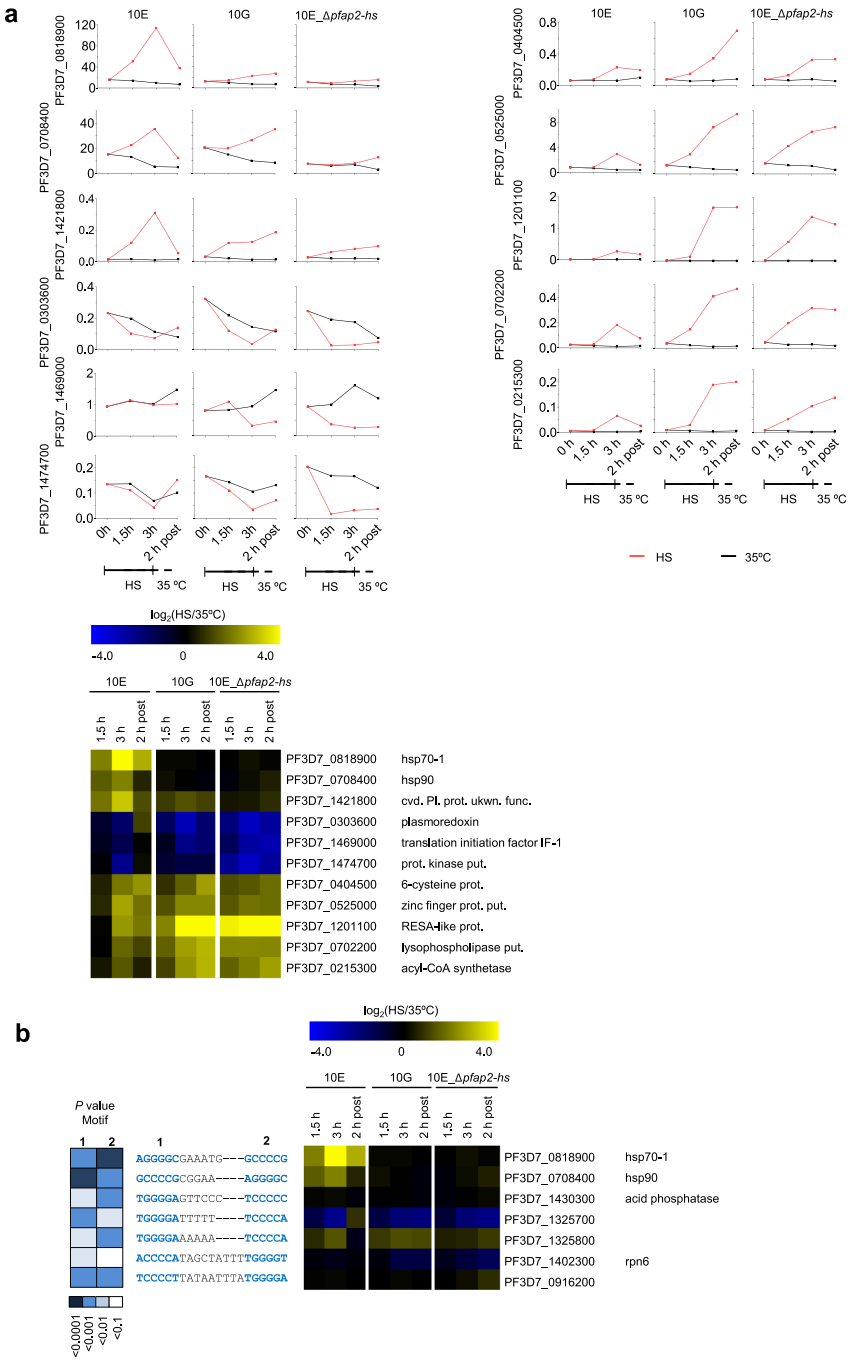


Extended Data Fig. 1 | Transcriptional analysis of heat-shock-adapted and control lines. a, Microarray-based transcriptomic comparison across the asexual blood cycle of 3D7-A cultures adapted to heat shock after five cycles of selection with a 3 h heat shock at 41.5 °C at the trophozoite stage (3D7-A-HS r1 and r2) and control parasite lines maintained in parallel without heat shock (3D7-A r1 and r2). Values are the $\log_2$ of the maximum expression fold-change (from the average of all lines compared) across a time interval corresponding to half the length of the asexual cycle, calculated using the aMAFC score as previously described[21]. Genes with a > 1.5-fold-change in expression in the two independent 3D7-A heat-shock-adapted lines (3D7-A-HS r1 and r2) relative to their respective controls (3D7-A r1 and r2) are shown. Data for parasite lines 10 G (heat-shock-sensitive subclone), 1.2B (heat-shock-resistant subclone) and 3D7-A (right panel) is from Rovira-Graells et al[21]. **b**, Time-course expression of genes in panel a that showed a concordant change in expression between heat-shock-adapted and control cultures, and between the heat-shock-resistant subclone 1.2B and the heat-shock-sensitive subclone 10 G. Based on the predicted function of the three genes, *clag2* was considered the most plausible candidate to play a role in heat-shock resistance. **c**, Expression of *clag2* is neither necessary nor sufficient for heat-shock resistance. RT-qPCR analysis of *clag2* transcript levels (normalized against *rhoph2*) in schizonts of heat-shock sensitive (S) and heat-shock resistant (R) 3D7-A subclones (see Fig. 1f), and of the heat-shock-adapted and control lines ($n = 1$ biological replicates).



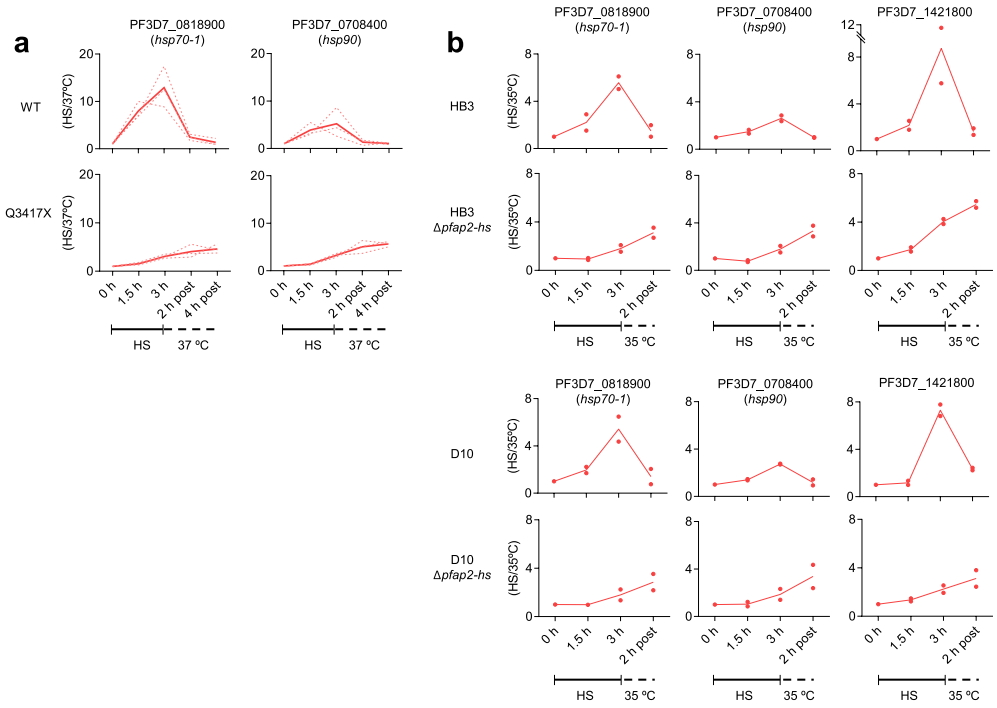
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Generation and characterization of transgenic parasite lines edited at the *pfap2-hs* locus. a, Schematic of the CRISPR–Cas9 strategy used to knockout *pfap2-hs*, using two guide RNAs. **b**–**d**, Tagging of endogenous PfAP2-HS using CRISPR–Cas9 technology. The tags used were a C-terminal 3xHA (**b**), a C-terminal eYFP (**c**) and an N-terminal eYFP (**d**). **e**, C-terminal tagging of endogenous PfAP2-HS by single homologous recombination with a tag consisting of a 2xHA epitope and an FKBP destabilization domain (DD domain). In all panels, the position of the primers used for analytical PCR (arrowheads), guide RNA and AP2 domains (blue vertical bars) is indicated. The electrophoresis images at the right are the analytical PCR validation of the genetic edition (single genomic DNA extraction and PCR analysis), showing correct edition and absence of wild-type locus in all cases except for the 8 A subclone of 1.2B-ddFKBP (8 A and 12E are subclones obtained after drug cycling). The bar charts at the right show the level of survival (mean of $n = 2$ independent biological replicates) of sorbitol-synchronized cultures of the transgenic lines upon heat shock (HS) exposure at the trophozoite stage, with heat-shock-resistant (10E) and heat-shock-sensitive (10 G, expressing PfAP2-HSΔD3) subclones as controls. Addition of a C-terminal eYFP or HA-FKBP tag did not affect growth at 37 °C but resulted in high heat-shock sensitivity, similar to the 10 G line. In contrast, C-terminal addition of the smaller 3xHA tag or addition of an N-terminal eYFP did not affect growth at 37 °C or heat-shock sensitivity. In all cases, tagged PfAP2-HS was not detectable by immunofluorescence or Western blot analysis, probably as a consequence of the very low abundance of this transcription factor (see proteomic data in www.PlasmoDB.org).

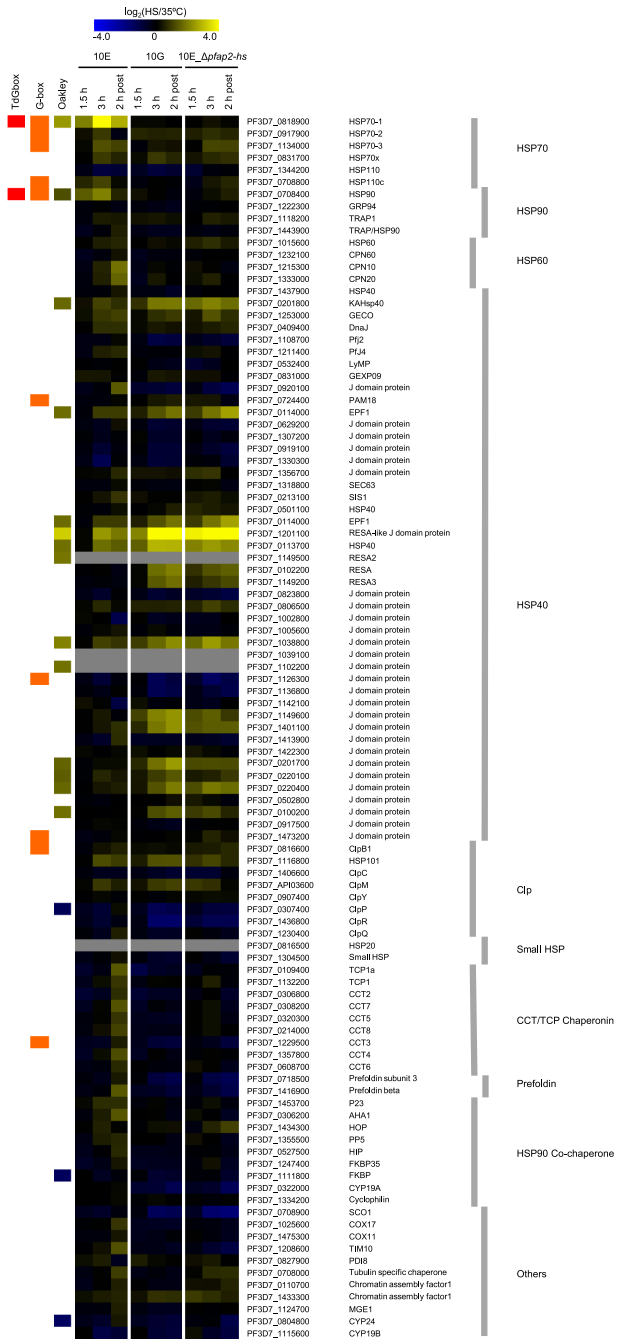


Extended Data Fig. 3 | See next page for caption.

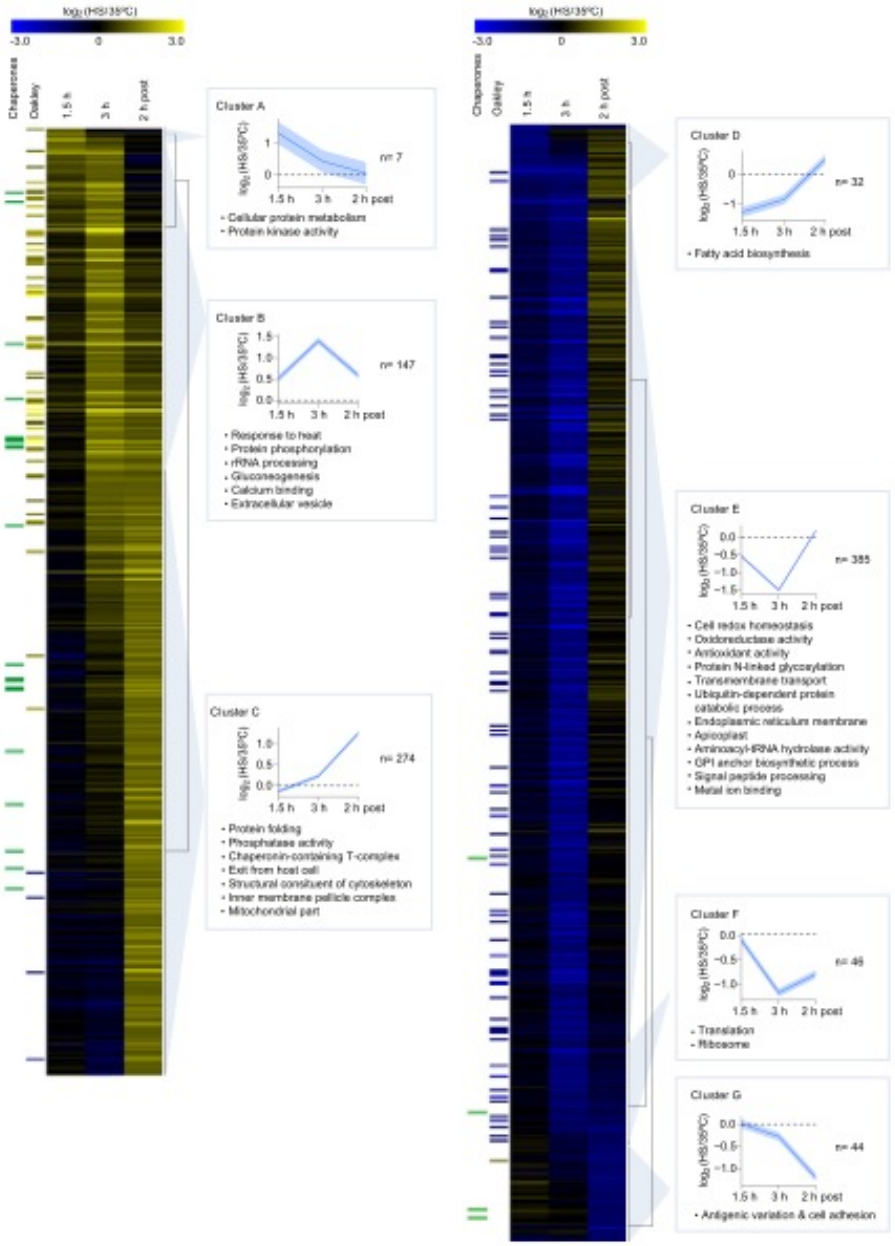
Extended Data Fig. 3 | Validation of the transcriptomic changes upon heat shock and distribution of the tandem G-box motif. a, RT-qPCR analysis of transcript levels (normalized against *serine–tRNA ligase*) of the genes selected for validation, using biological samples independent from the samples used for microarray analysis. Values are the average of triplicate reactions. The $\log_2$ expression fold-change [heat shock (HS) relative to control (35 °C) conditions] for these genes in the microarray analysis (Fig. 2a) is shown in a heatmap to facilitate comparison. **b**, Genes in the *P. falciparum* genome containing tandem arrangements (maximum distance between the two: 9 nucleotides) of the G-box [(A/G)NGGGG(C/A)] motif in their regulatory regions (defined as the 2 kb upstream of the start codon or until the neighbour gene, when it is closer). The sequence of the G-box in each gene is shown in blue, and the level of concordance with the consensus G-box motif is expressed as the *P* value of the match (determined using the FIMO v5.0.5 function in the MEME suite). Expression changes upon heat shock for these genes are shown as in panel **a**.



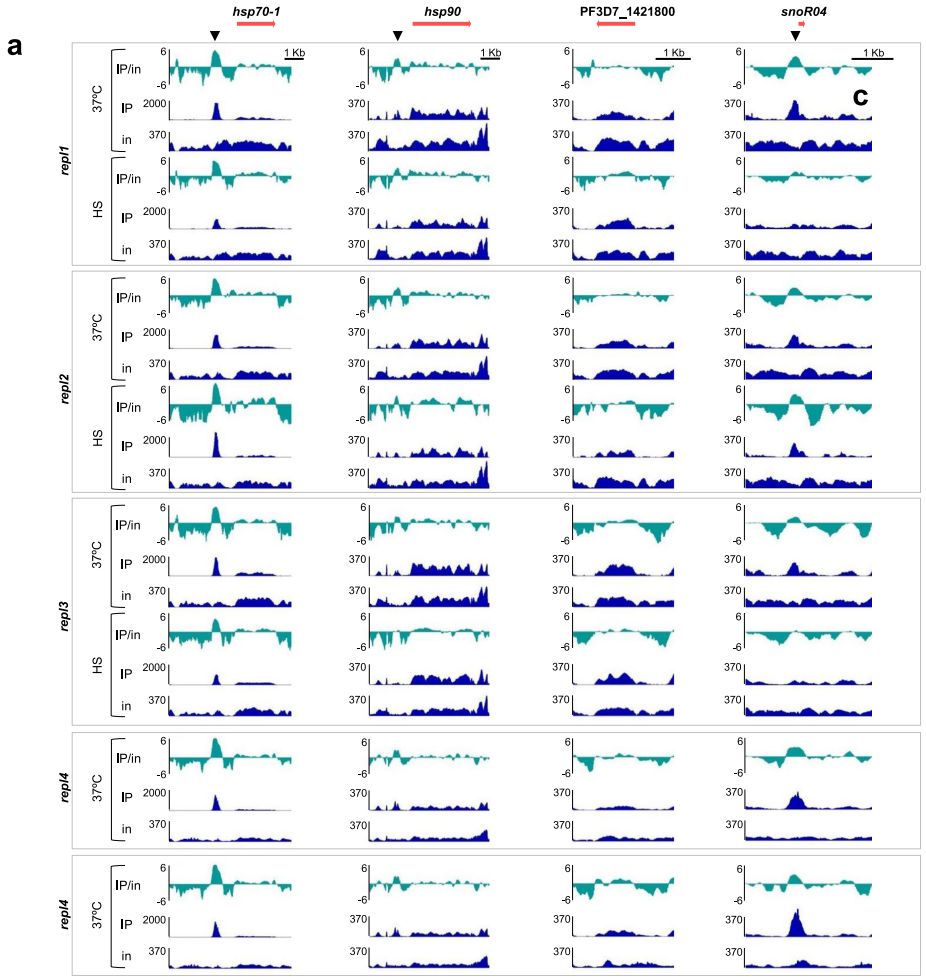
Extended Data Fig. 4 | Changes in *hsp70-1, hsp90* and PF3D7_1421800 transcript levels in parasites lacking the entire PfAP2-HS or D3. Fold increase in transcript levels (determined by RT-qPCR, normalized against *serine–tRNA ligase*) during and after heat shock (HS) starting at 33-35 (**a**) or 30-35 (**b**) h.p.i., relative to cultures maintained in parallel without heat shock (37 or 35 °C). In panel **a**, values for three individual 3D7-A subclones carrying or not the Q3417X mutation are shown as dotted lines, whereas the average of the three subclones is shown as a continuous line. In panel **b**, the mean of $n = 2$ independent biological replicates is shown.



Extended Data Fig. 5 | Transcript level changes upon heat shock in chaperone-encoding genes. Log_2 expression fold-change [heat shock (HS) relative to control (35 °C) conditions, as in Fig. 2a] for all chaperone-encoding genes described by Pavithra and colleagues[14]. Columns at the left indicate presence of the G-box[23] or tandem G-box (TdGbox) in the upstream region, and log_2 fold-change during heat shock in a previous study[20] (Oakley).

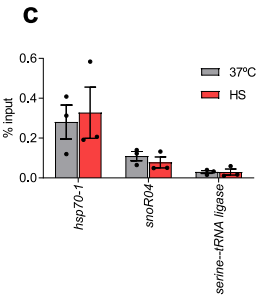


Extended Data Fig. 6 | Transcriptomic characterization of the heat-shock response in parasites expressing complete PfAP2-HS (10E line). Log_2 expression fold-change [heat shock (HS) relative to control (35 °C) conditions, as in Fig. 2a] in the wild-type 10E line determined by microarray analysis. Genes with a fold-change ≥2 at any of the time points analysed are shown. The mean log_2 expression fold-change (with 95% confidence interval) and representative enriched GO terms are shown for each cluster. Columns at the left indicate fold-change during heat shock in a previous study[20] (Oakley), and annotation as chaperone[14]. Ten genes had values out of the range displayed (actual range: -3.89 to +4.03).



b

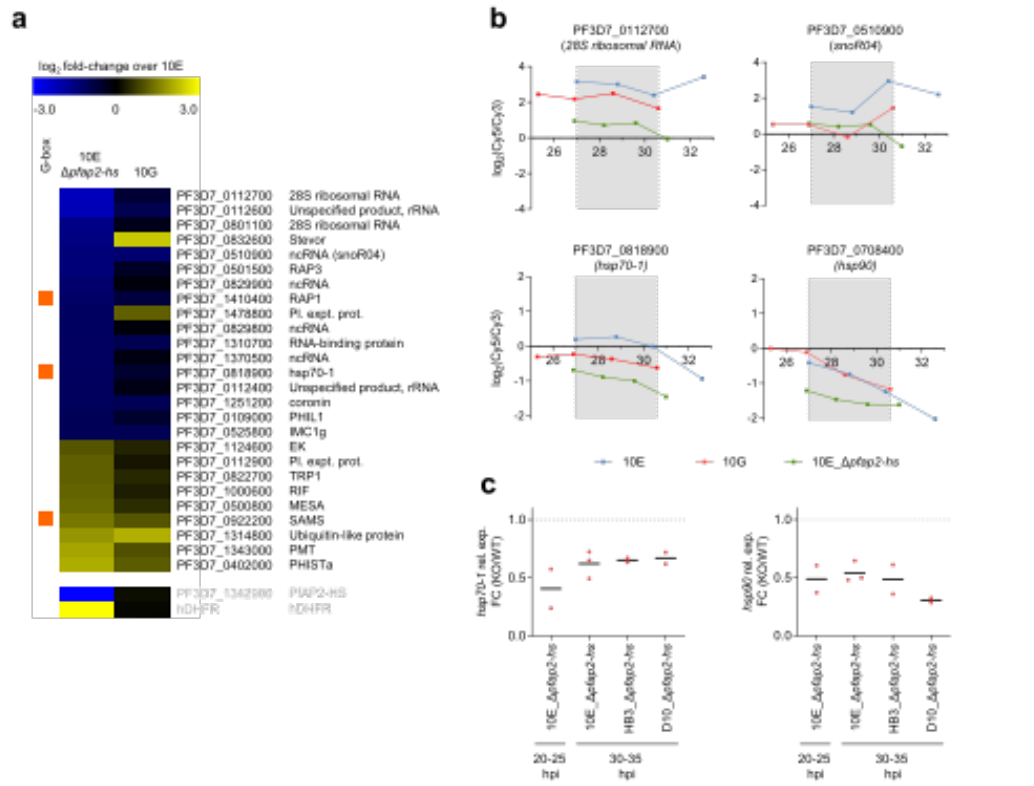
	Chrom.	Start	End	Median MACS2 Score	Median Fold Enrich.	Closest Gene	Gene Name	Location
Control	8	859109	859513	3155	14.6	PF3D7_0818900	*hsp70-1*	5'
	5	465548	465850	429	4.2	PF3D7_0510900	*snoR04*	5' & gene body
	13	2895119	2897998	150	2.3	Telomere	-	--
HS	8	859123	859514	3410	21.8	PF3D7_0818900	*hsp70-1*	5'
	10	1437164	1437430	226	3.5	PF3D7_1036400	*lsa1*	Gene body
	13	2909970	2912982	181	2.9	Telomere	-	-



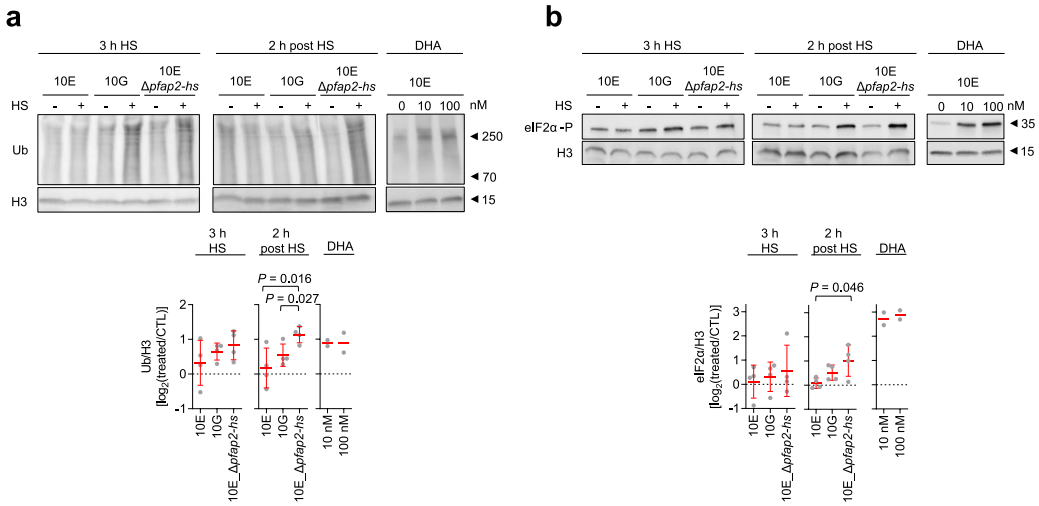
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | ChIP analysis of the chromosomal distribution of PfAP2-HS. a, ChIP–seq analysis of HA-tagged PfAP2-HS. Number of reads of ChIP (IP) and input (in) tracks, and $\log_2$-transformed ChIP/input ratio tracks (IP/in) for five independent biological replicates [three including heat shock (HS) and 37 °C conditions, two including only the 37 °C condition]. Snapshots are shown for the three genes in cluster I (Fig. 2a) and *snoR04*. Binding at the *hsp70-1* and *hsp90* promoters coincides with the position of a tandem G-box motif, whereas PF3D7_1421800 and *snoR04* lack a G-box. The positions of the relevant peaks are indicated by an arrowhead. **b**, Peaks present in ≥3 out of 5 replicate ChIP–seq experiments (37 °C) or ≥2 out of 3 replicate experiments (heat shock) and with a MACS score >100 in each positive replicate. **c**, ChIP–qPCR analysis of HA-tagged PfAP2-HS binding at selected loci, in cultures exposed to heat shock (HS) or control (37 °C) conditions (mean and s.e.m. of % input in $n = 3$ independent biological replicates). No significant difference ($P < 0.05$) was observed between 37 °C and heat shock using a two-sided unpaired Student's *t*-test.

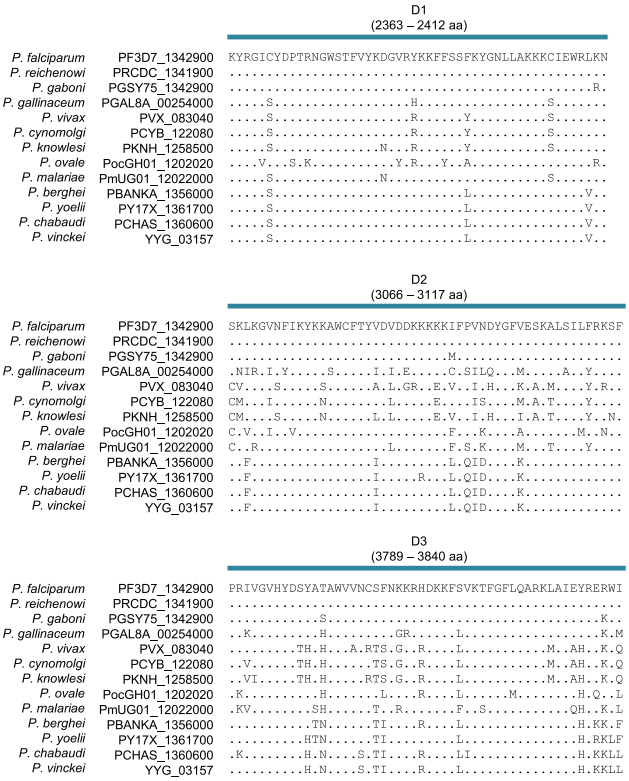
NATURE MICROBIOLOGY ARTICLES



Extended Data Fig. 8 | Transcriptional changes associated with PfAP2-HS deletion under basal (no heat shock) conditions. a, Changes in transcript levels in the absence of heat shock for genes with an average expression fold-change >2 between 10E_Δ*pfap2-hs* and 10E. Values are the $\log_2$ of the average expression fold-change relative to 10E across the time period compared (~27–30.5 hpi). Genes artificially modified or introduced in the knockout line, which serve as controls, are shown at the bottom (their values are out of the range displayed). The column at the left indicates the presence of the G-box[23]. **b**, Expression plots for selected genes under basal conditions. Expression values are plotted against statistically estimated parasite age, expressed in h post-invasion (h.p.i.). Grey shading marks the interval used to calculate the average expression fold-change. **c**, RT-qPCR analysis of *hsp70-1* and *hsp90* transcript levels in *pfap2-hs* knockout (KO) lines compared to their wild-type (WT) controls (the parental line for each knockout line) under basal conditions. Expression values are normalized against *serine-tRNA ligase*, and expressed as the fold-change (FC) in the knockout versus control lines. The mean of $n = 3$ (10E, 30-35 hpi) or $n = 2$ (others) independent biological replicates is shown.



Extended Data Fig. 9 | Analysis of proteome stress and UPR markers in *pfap2-hs* mutants. a, **b**, Western blot analysis (representative of $n=4$) of polyubiquitinated proteins (Ub) (**a**) or phosphorylated eIF2α (eIF2α-P) (**b**) immediately after a 3 h heat shock (3 h HS) and 2 h later (2 h post HS). Histone H3 is a loading control. DHA was used as a positive control, as it is a known inducer of the UPR[29,38]. The Log_2 of histone H3-normalized signal in heat shock or DHA-treated cultures versus control cultures is shown at the bottom (mean and s.e.m. of $n=4$, except for the DHA control mean of $n=2$ independent biological replicates). *P* values were calculated using a two-sided unpaired Student's *t*-test. Only significant *P* values ($P<0.05$) are shown. The position of molecular weight markers is shown (in kDa).



Extended Data Fig. 10 | Sequence alignment of the three AP2 domains (D1–3) present in AP2-HS orthologues in *Plasmodium* spp. Dots indicate identity with the amino acid in the first sequence.

5 Results summary

In this section, we will briefly report the main results of all the papers included in this thesis. All papers deal with transcriptional variation and adaptation in *P. falciparum*, with a special emphasis on epigenetic mechanisms. The only paper in which epigenetics does not play a central role is the one on PfAP2-HS and adaptation to heat-shock. In fact, we expected adaptation to febrile temperatures to be epigenetically regulated, but it turned out to be an exception that proved that directed transcriptional responses, mediated by TFs, also play a role in *P. falciparum* adaptative mechanisms.

6 Summary of Article 1: Patterns of heterochromatin transitions linked to changes in the expression of *Plasmodium falciparum* clonally variant genes

For this study, we generated transcriptomic and heterochromatin profiling data for 5 different recent subclones of the 3D7 genetic background. The study of these homogeneous populations allowed us to characterize in detail the active and silenced states of CVGs at a genome-wide level. As expected, over 90% of differentially expressed genes between subclones where CVGs, confirming that transcriptional differences between subclones were mostly restricted to these kinds of genes. Our data confirmed the negative correlation between the H3K9me3 enrichment in upstream and coding regions of genes and their transcriptional status.

The homogeneity in transcriptional patterns and heterochromatin distribution in the subclones we analyzed allowed us to characterize, with unprecedented detail, the H3K9me3 distribution differences that drive the active and silenced states of CVGs. We identified 121 genes with high transcriptional differences (fold-change > 4) between subclones. For each gene, we selected the subclone with highest expression as the representative of the active state and the subclone with

lowest expression as the representative of the silenced state. We then quantified, again for each gene, the heterochromatin distribution on the upstream, coding, and downstream regions, as well as the neighboring genes in its active and silenced states. From the 121 genes, only 10 were non-CVGs, which indeed showed no H3K9me3 enrichment in the active or silenced states, and whose transcriptional differences might be explained by different mechanisms. This analysis allowed us to define different patterns of active/silence heterochromatin transitions. The first kind, which we named expansion/retraction transitions involved the spreading or retraction of heterochromatin from neighboring genes (65/111). In the silenced state, heterochromatin occupied the upstream and coding sequence of genes. The coding sequence could become completely heterochromatic or only partially. Among genes showing this kind of transition, there was a subset for which differences in H3K9me3 enrichment occurred mainly on the downstream region. However, we could not rule out the possibility that transcriptional differences in this subset were consequence of smaller H3K9me3 occupancy differences on the upstream regions, instead of the observed major changes in the downstream region. The second kind of transitions, which we named, localized opening transitions, involved only small differences in H3K9me3 enrichment in a very small, localized region, typically on the upstream region of genes (39/111). Finally, only a very small subset of genes showed changes in heterochromatin distribution that were suggestive of *de novo* formation of heterochromatin or the complete removal of a whole previously heterochromatic domain (5/111). A vast majority of genes showing this kind of transitions were expressed predominantly in non-IDC stages (gametocyte or mosquito stages). This suggests that virtually all adaptative CVG expression changes during the IDC are driven by the expansion/retraction of heterochromatin whereas the formation/removal of new heterochromatin domains is restricted to genes expressed at other life cycle stages. Most CVG families showed no clear preference for any type of heterochromatin transitions, with the notable exception of the *pfmc-2tm* family, in which almost all differentially expressed members displayed localized transitions in their upstream region. Also of note, we found that heterochromatin distribution over a given CVG was almost identical in different subclones that had it in the same transcriptional state.

We analyzed *var* gene expression in our subclones and found that two of them had expression patterns compatible with the simultaneous expression of more than one single *var* gene. However, bulk transcriptional data cannot distinguish between the possibility that these results arise from heterogeneity in *var* gene expression (which should have originated in the ~4 weeks the subclones where cultured after limiting dilution), or from expression of more than one *var* gene per parasite. ChIP-Seq data, on the other hand, confirmed that all the expressed *var* genes in each subclone were clearly euchromatic in their upstream and sometimes coding region (except for one). Our ChIP-Seq results clearly show that multiple *var* genes are euchromatic in the majority of parasites in the culture, strongly suggesting simultaneous *var* expression in at least some parasites.

Our subclones had a wide range of basal sexual conversion rates. We found only minor heterochromatin differences at the *pfap2-g* locus between subclones, but we found significant differences at the *gdv1* locus, downstream of the gene, in a region that broadly overlaps the putative regulatory region of the *gdv1* antisense lncRNA. Differences at this locus showed a good correlation with basal sexual conversion rates, such that subclones with higher levels of H3K9me3 enrichment showed higher sexual conversion rates.

For three subclones, we also analyzed the distribution of H3K9ac in addition to H3K9me3. As reported previously, H3K9ac and H3K9me3 distribution were mutually exclusive along the genome. Using our microarray data together with publicly available RNA-Seq data, we classified genes in our subclones into five different categories, namely: *non-CVG active*, *non-CVG silenced*, *CVG active*, *CVG silenced* and *Undetermined*. Analysis of H3K9me3 and H3K9ac enrichment profiles in those five categories showed that mean enrichment levels of these marks are different between them. H3K9me3 enrichment was clearly higher in CVGs than in non-CVGs (both active and silenced) and higher in silenced CVGs than active ones. On the other hand, H3K9ac had similar levels in active and silenced non-CVGs, intermediate levels in active CVGs and lowest levels in silenced-CVGs.

7 Summary of Article 2: Expression Patterns of *Plasmodium falciparum* Clonally Variant Genes at the Onset of a Blood Infection in Malaria-Naive Humans

In this study we analyzed the expression patterns of CVGs in parasites cultured before and after the passage through transmission stages and infection of human volunteers, in order to confirm the possible reset of heterochromatin patterns during this process.

We started by analyzing by 2-channel gene expression microarrays the whole transcriptome of parasites obtained from a controlled human malaria infection (CHMI) trial and comparing them with their parental line, wich was kept in contiguous culture in the lab. The samples included 4 different volunteers (V18, V35, V48, and V63 lines, to which we will collectively refer as vNF54) and the parental NF54 (obtained "pre-mosquito", to which we will refer as pNF54) used for these trials. Transcriptional analysis showed that the vast majority of differences occurred in genes previously characterized as CVGs, as expected. In general, transcriptional changes between vNF54 and pNF54 where very similar for all volunteers. Most CVGs where upregulated in vNF54 with respect to pNF54, with a few exceptions. The two most downregulated genes where *var2csa* and *clag3.1*.

Transcriptional changes where different in different CVG families. To perform our analysis, we took advantage of the fact that 2-channel gene expression microarrays can provide both relative expression data (comparing samples) and absolute expression data (from a single channel). Virtually all members of the *pfmc-2tm* family where highly upregulated in vNF54, which was confirmed analyzing normalized absolute expression values. For the *var, rif, phist,* and *stevor* families only a small subset of genes showed expression differences, but the subset was almost identical for all volunteers. Analysis of the normalized absolute expression data for *var* genes revealed that *var2csa* was expressed in the vast majority of pNF54 parasites, while for vNF54 different *var* genes, mostly from the type-B subtype, showed an intermediate level of expression indicating heterogeneity of the population. The *rif* family showed upregulation of mostly type-A

members, while type-B members remained unchanged. Analysis of normalized absolute expression data for *rif, stevor* and *phist* families revealed that, despite the differences in relative expression in some members between vNF54 and pNF54, the most abundantly expressed members of each family remained the same in samples from both origins. The *gbp* and *clag* families showed a huge relative expression difference for most members of the family. Genes from other CVG families did not display differential expression between vNF54 and pNF54, with the exception of *hyp10 and* the phospholipase PF3D7_0936700, which showed high relative differences.

Whole genome sequencing (WGS) of parasite lines obtained from 2 volunteers and pNF54 did not identify any genetic mutations/insertions/deletions that could explain the majority of the observed transcriptional differences, reinforcing the idea that the observed variability derives from changes in the heterochromatin distribution, rather than genetic changes. Indeed, ChIP-Seq profiling of H3K9me3 distribution, on V63 and pNF54, showed a clear depletion of heterochromatin in the upstream region and start of coding sequence of 7/8 genes with FC > 16 expression differences. However, other genes displaying differential expression (FC >4) did not display apparent heterochromatin differences between the two samples. We reasoned that this was probably due to transcriptional differences arising from the expression of some CVG variants in only small subsets of the parasite population. For example, if a given gene is expressed in 10% of parasites in one sample but only in 1% of parasites in another, the expression FC could amount up to 10-fold, but the gene would still be heterochromatic in 90% and 99% of parasites, respectively, rendering differences in ChIP-Seq signal between samples undetectable. To test this hypothesis, we analyzed the expression of all *pfmc-2tm* members and the upregulated *rif* and *stevor* genes in subclones derived from V33. We reasoned that if some of these genes where only expressed in some parasites, only some subclones would express them. Our analysis confirmed that only a few subclones expressed the upregulated genes of each family.

We compared the patterns of expression of *gbph2, pfmc-2tm,* and *stevor* genes between days 9, 11 and 14 after injection on the volunteers, on two vNF54 lines (V18 and V54). Day 9 represents parasites that have undergone only one round of replication after egress from the liver. We

reasoned that if the variability in CVGs expression relative to pNF54 arose from adaptation to host-blood conditions, patterns of expression should be increasingly different between volunteers and pNF54 as days passed. On the other hand, if variability between vNF54 and pNF54 parasites arose from an epigenetic "reset" during transmission stages it should be present in the parasites already at day 9. The analysis showed no significant changes in gene expression differences between days 9, 11 and 14, suggesting that differential CVG patterns occurred from the onset of the infection, and supporting the view that expression patterns of CVGs are reset and reestablished at some point during transmission stages.

We also analyzed possible phenotypic differences between V63 and pNF54 using microsphiltration experiments (which inform about membrane deformability), sorbitol lysis (which informs about solute transport through the iRBC membrane) and sexual conversion assays. None of the analyzed phenotypes was significantly different between V63 and pNF54.

8 Summary of Article 3: Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*

In this study, we produced a conditional expression system for *pfap2-g,* generating a parasite line that reached ~90% sexual conversion upon induction. A strong constitutive calmodulin promoter (5' *cam*) was inserted upstream of the *pfap2-g* coding sequence separated from the coding sequence by a *hdhfr* selectable marker, which confers resistance to the WR99210 drug. This *hdhfr* marker was flanked by *loxP* sites which, upon recombination by the inducible Cre recombinase (DiCre, which was integrated elsewhere in the genome) placed the 5' cam promoter directly upstream of the *pfap2-g* coding sequence, driving its expression. The DiCre recombinase is only active upon dimerization induced by rapamycin, which we can add to the culture at the desired time-point, underlying the conditional expression system. We integrated the constructs in the E5

genetic background (a 3D7 subclone with high basal conversion rate), and we named the resulting parasite line E5ind.

Induction of the E5ind line, maintained under constant selection by WR99210, by addition of rapamycin at the late trophozoite stage, resulted in sexual conversion of ~90% of the parasites. The same inducible system was reproduced in the 1.2B parasite line, which does not produce gametocytes due to a genetic defect in *gdv1*. Induction with rapamycin resulted in sexual conversion of ~70% of the parasites in the 1.2Bind line, which recovered the capacity to produce gametocytes.

Despite the artificial system, expression patterns of *pfap2-g* in the inducible lines resembled those of the endogenous *pfap2-g* in wild-type parasites. In induced lines, *pfap2-g* expression was 25-fold lower than that of endogenous *cam* and its levels were reduced during gametocyte maturation (days 4 to 7). The reasons for this were unclear, but possibly reflect the strong regulatory mechanisms that affect the *pfap2-g* locus, including the effect of heterochromatin in the locus.

Induction of both E5ind and 1.2Bind lines without continuous selection by WR99210 resulted in markedly lower levels of sexual conversion, whereas reselection with the drug recovered the ~90/~70% conversion rates. A plausible explanation is that drug pressure selects for parasites in which the *5'cam-hdhfr* insertion is in an euchromatic state, whereas the lack of selection allows for parasites with heterochromatin at the locus to survive in the culture. Parasites with a heterochromatic locus could fail to convert, thus reducing the level of sexual conversion upon induction.

Continuous culture after induction allowed, in ~1 week, to obtain parasite populations from the few parasites that did not convert into gametocytes and continued asexual replication, generating two parasite populations termed E5ind+Rapa_prol and 1.2Bind+Rapa_prol. In these lines, gametocyte production returned to their parental line levels: ~10% for E5ind+Rapa_prol and no

gametocyte production for 1.2Bind+Rapa_prol. ChIP-Seq analysis of E5ind, 1,2ind, E5ind+Rapa_prol and 1.2Bind+Rapa_prol without induction showed that heterochromatin was virtually absent from the endogenous promoter of *pfap2-g* in E5ind and 1.2Bind (maintained under drug pressure) but was present in the endogenous promoter of *pfap2-g* in both E5ind+Rapa_prol and 1.2Bind+Rapa_prol. Taken together, these results suggest that the presence of heterochromatin at the *pfap2-g* upstream putative regulatory region prevents its expression, and that GDV1 is necessary for its basal activation. In contrast, heterochromatin at the coding region of *pfap2-g* was present in all analyzed lines (even in those reaching ~90% sexual conversion upon induction), indicating that it does not preclude the activation of the gene.

Mature gametocytes from both E5ind and 1.2Ind lines were tested for mosquito infectivity. Both were able to round up and egress, but both failed to exflagellate, and therefore to produce oocysts upon mosquito infection.

The high purity of the obtained sexually committed schizont and sexual ring cultures allowed for the transcriptional analysis of these stages with unprecedented sensitivity. In early sexually committed rings, *gexp02* was one of the first genes to be induced. Known gametocyte markers such as *pfs16, pfg27/25* and *pf14-744* were among the most upregulated genes. 77% of the genes that showed a >3 FC increase were reported as PfAP2-G targets in a previous ChIP-Seq study for this TF, against only a 3% of the rest of the genome. In contrast, only 3% of genes with a >3 FC decrease in expression where among the targets identified by ChIP-Seq. Many of the downregulated genes where among genes known to be necessary for asexual proliferation, such as *mesa, hrpIII, pfemp3, pf332* and *kharp.* In sexually committed schizonts, changes in transcription where less abundant and smaller, and the most remarkable result were the up regulation of the invasion related gene *msrp1* and that, from six downregulated genes in total (FC < 0.5), three (*clag3.1*, *clag2*, and *rhoph3*) where components of the RhopH complex. RhopH complex proteins participate in the formation of the PSAC, which is required for the permeability of iRBC to multiple solutes, including sorbitol. E5ind induced parasites showed an elevated resistance to sorbitol-induced lysis, to which gametocytes are known to be more resistant too.

Similar results were obtained the by assessing the uptake of 5-aminolevulinic acid (5-ALA). These results confirm that the downregulation of RhopH proteins in sexually committed schizonts lowers PSAC activity in sexual parasites.

9 Summary of Article 4: A heat-shock response regulated by the PfAP2-HS transcription factor protects human malaria parasites from febrile temperatures

P. falciparum parasites lack a homolog of heat-shock factor 1 (HSF1), which regulates the response to heat-shock (a rapid increase in environmental temperature) in most eukaryotes, from yeast to mammals. In this work, we identified PfAP2-HS as the functional homolog of HSF1. PfAP2-HS is an apiAP2 transcription factor with three AP2 domains (D1-3), the first one recognizing a tandem G-box motif present in some heat-shock protein (hsp) chaperone genes. We started by observing that non-sense mutations in the domain 3 (D3) of *pfap2-hs* rendered parasites more sensitive to heat-shock at 41.5°C. We generated knock-out lines for *pfap2-hs* and found that mutants were even more sensitive to heat-shock and showed diminished growth at 37°C, but not at 35°C. The analysis of other *pfap2-hs* knock-out parasite lines, of different genetic backgrounds, in comparison with their parental wild-type lines, revealed that the growth defect at basal conditions (37°C) was only observed in some strains, but not in others, whereas the function of *pfap2-hs* in heat-shock resistance was conserved in all strains. Taken together, these results suggest a role for PfAP2-HS in proteostasis at basal conditions, although different parasite lines might have different sensitivity to its disruption.

Transcriptional analysis comparing whole genome expression during heat-shock in a wild-type line, the line with a non-sense mutation in D3 and the mutant line with a complete *pfap2-hs* knock-out revealed many dysregulated genes upon heat-shock, most of which probably related with protein damage caused by heat-shock and the resulting cell-death processes. Only a very small set of genes, comprising *hsp70-1*, *hsp90* and PF3D7_1421800 (a gene of unknown function),

where rapidly up-regulated upon heat-shock only in the wild-type line. Of note, in yeast, the only two targets of HSF1 that are essential for survival under basal conditions are *hsp70* and *hsp90*.

ChIP-Seq analysis of *pfap2-hs* binding revealed, again, a very restricted distribution. The tandem G-box in the upstream region of *hsp70-1* showed a high enrichment in all replicates, and *hsp90* showed a moderate enrichment, also coinciding with its tandem G-box. A small nucleolar RNA (*snoR04*, PF3D7_0510900) locus was the only other enriched locus in the whole genome, albeit at much lower levels than *hsp70-1*.

Artemisinin kills parasites by producing general protein damage. Given that PfAP2-HS participates in protein homeostasis, we hypothesized that PfAP2-HS defective lines would be more sensitive to the toxic effect of artemisinin and its derivatives. We tested whether PfAP2-HS defective parasites were more susceptible to dihydroartemisinin (DHA) treatment, and indeed all tested parasite strains with defects in *pfap2-hs* showed a higher sensitivity to the drug than its wild-type counterparts.

Taken together, our results indicate that PfAP2-HS participates in proteostasis maintenance under basal conditions and drives a protective response to heat-shock through the chaperones *hsp70-1* and *hsp90*, representing the first case of a TF that drives a directed transcriptional response in *P. falciparum*.

Discussion

General Discussion

Clonally variant genes are central to the adaptation of malaria parasites to many of the challenges encountered during the IDC in the human blood circulation. This adaptation allows the survival of the parasite and enables the chronic infections that underly the malaria disease. Reversible heterochromatin-based silencing is the central mechanism that governs the expression of CVGs [335,400]. Changes in the heterochromatin distribution along CVGs clusters result in transcriptional diversity in a population, which in turn results in phenotypic diversity. This diversity, coupled with the heritability of heterochromatin patterns, sets the grounds for bet-hedging adaptative strategies. Parasites with the set of expressed and silenced CVGs conferring the highest fitness in each environment are positively selected and become progressively dominant among an infecting population [206,335,400]. However, bed-hedging adaptative mechanisms are relatively slow and might not be well suited to cope with some of the most rapidly changing challenges the parasites have to face. Directed transcriptional responses, on the other hand, are rapid and transient, and are needed to adapt to such rapidly fluctuating conditions. In this thesis work, we have addressed fundamental questions regarding adaptation both through CVGs and directed transcriptional responses.

Regarding CVGs, we have contributed to a better understanding of many aspects of their heterochromatin-based regulation. We have accurately described the previously uncharacterized heterochromatin distributions associated with the active and silenced states of over a hundred CVGs (only a few had been previously characterized) (Article 1). We have also described how these heterochromatin patterns are reset after passage through transmission stages and characterized them at the onset of a human blood infection (Article 2). One of the most important processes regulated by CVGs is sexual conversion, a crucial step for malaria transmission [12]. In this thesis, we have contributed to a better understanding of this phenomenon by describing the transcriptional and heterochromatin patterns that underly it, both at the *ap2-g* locus (Article 3),

and at the newly described heterochromatic locus at the upstream region of the *gdv1* antisense lncRNA (Article 1).

Regarding directed transcriptional responses, we have characterized the first TF that drives a directed transcriptional response in *P. falciparum*. We have described how PfAP2-HS drives a protective transcriptional response to heat-shock, a common environmental stress faced by the parasite due to the characteristic fever episodes of malaria (Article 4). Indeed, adaptation in *P. falciparum* is not limited to CVGs, and at least for some stimuli, directed transcriptional responses mediated by specific TFs can mount a rapid and transient protective response.

In this discussion we will focus on the aspects of this work that have been least developed in the included papers, on the implications of our results and on future perspectives that arise from our work. We will start by discussing the different kinds of adaptation mechanisms displayed by *P. falciparum* and then focus on CVG-based adaptation. We will discuss the different kinds of heterochromatin transitions that underly the switch between the active and silenced states of CVGs and how the limits of their heterochromatin domains are established. Finally, we will focus on some technical aspects around the proper identification of active and silenced genes and the identification of differential heterochromatin peaks, and how the data generated in our analyses could be used to predict the transcriptional state of CVGs during the whole IDC.

1 Directed Responses vs Bet hedging

Transcriptional changes leading to phenotypic variation lay at the heart of adaptation to changing conditions, which is a vital process for any living organism. Even in the relatively stable conditions of a human blood infection, *P. falciparum* populations need to adapt to frequent changes in the availability of nutrients, significant variations in temperature (from febrile episodes) and the constant pressure of the immune system, among other challenges. Therefore, the capacity of the parasite to adapt its transcriptome to this ever-changing scenario is crucial for its ability to maintain long-lasting infections in the human host.

In *P. falciparum,* transcriptional differences between parasites at the same life cycle stage, can originate mainly from three different sources: genetic changes in the primary DNA sequence of genes or their regulatory regions, epigenetic changes that lead to differential expression of the affected genes, or directed responses, which activate different transcriptional programs in response to specific stimuli. Both genetic and epigenetic derived adaptation are heritable and contribute to adaptation at the population level (meaning that survival of the population is maximized, even if some parasites die). Genetic mutations or epigenetic traits are initially different between individual parasites and confer them different fitness in different environmental conditions. The heterogeneity of the population ensures that upon any challenge, even if many parasites are not well adapted and die, at least some parasites will have stochastically acquired patterns conferring high fitness and will be well adapted. These well adapted parasites will become selected, expand, and ensure the survival of the population. On the other hand, directed responses are quick and transient, and they contribute to adaptation at the individual parasite level (meaning that survival is maximized for each individual parasite). Upon a specific stimulus, all individual parasites will respond activating the same protective response, ensuring survival of most of them and thus ensuring survival of the population.

1.1 Adaptation through genetic changes

Genetic variation is the main means of natural selection and has, undoubtedly, had the largest impact in the evolution of *P. falciparum*. However, genetic changes in eukaryotes occur at a relatively low rate, and this kind of adaptation operates at a very long timescale and is largely irreversible. Genetic variation-based adaptation is too slow to cope with the rapidly changing challenges faced by the parasite during natural infections and, for this reason, the parasite needs other means of adapting to fluctuating conditions.

1.2 Adaptation through epigenetic regulation of clonally variant genes

Adaptation through epigenetic regulation of CVGs stands somewhat between genetic adaptation and directed responses, in many aspects. Like genetic mutations, epigenetic changes are heritable (at least some of them), which allows for natural selection to operate on them, leading to a population-wise adaptation. Also, similarly to genetic changes, the switch between different epigenetic states in CVGs operates stochastically, generating population diversity that allows for bet-hedging adaptative strategies. Unlike genetic mutations though, which are virtually irreversible, epigenetic changes are reversible and allow for the return to a prior state (and associated phenotype). Despite this reversibility, the stochastic switches between epigenetic states in CVGs are relatively rare. For this reason, epigenetic adaptation is more flexible tan genetic adaptation but way more stable than the transient transcriptional changes induced by directed responses.

We have highlighted throughout this thesis the importance that CVG-based epigenetic adaptation has for the pathology of malaria infections. Around 10% of the parasite genome is regulated by facultative heterochromatin, which includes genes implicated in many different aspects of host-parasite interactions and the switch into sexual development [206]. In Article 1 we have characterized the heterochromatin transitions that underly adaptation mediated by CVGs. Our results highlight the dynamic nature of H3K9me3-based silencing in *P. falciparum*, with small

changes and shifts in the heterochromatin distribution, rather than *de novo* formation or complete removal of large heterochromatin domains, governing the CVGs expression landscape. We have described how transitions between the active and the silence state of CVGs can occur at different genomic regions and involve both wide expansions/retractions and narrow localized changes. Regarding sexual commitment, it was already known that the expression of *pfap2-g* was regulated by heterochromatin transitions[12,350], but we have shown that its upstream regulator *gdv1* is also regulated by the presence of heterochromatin in its antisense lncRNA upstream region (Article 1). How CVG expression shifts are regulated, if they respond to stochastic fluctuations on the local concentration of writers/readers/erasers or whether they might be influenced by external factors, remains a question for further research.

1.3 Adaptation through directed transcriptional responses

Directed transcriptional responses are very different to both genetic and CVG-based epigenetic adaptation. They typically involve the sensing of an external stimulus, a signaling cascade that carries the information to the cell nucleus and the activation of a TF (or a set of them) that will drive the transcriptional response to said stimulus. Directed transcriptional responses are usually very quick and transient and most eukaryotes extensively use directed responses to respond to environmental cues.

There has been a longstanding debate in the malaria field around the question of whether *P. falciparum* is capable of mounting directed responses to external stimuli. The general scarcity of specific transcription factors, and the apparent lack of a transcriptional response to some challenges, has led many to propose that the parasite has a somewhat hardwired transcriptional program [401–403] and that other types of adaptation, especially epigenetic-based adaptation, are the parasite's main means to adapt to changing conditions. Recent work, including the one in this thesis, has proven that at least for some specific stimuli, *P. falciparum* can sense external cues and mount specific transcriptional responses to them. This was first demonstrated for changes in the serum component lysophosphatidylcholine (lysoPC). Changes in the concentration of this

critical nutrient resulted in extensive changes in the expression of genes involved in parasite metabolism and an elevated sexual conversion rate [354]. It has also been demonstrated that *P. berghei* (and probably also *P. falciparum*) can sense nutrient restriction, signal it through the KIN kinase, and activate a transcriptional response that results in reduced multiplication rate [404]. In Article 4, we have demonstrated that *P. falciparum* responds to febrile temperatures by activating a very compact transcriptional response through PfAP2-HS, providing the first example of a TF driving a directed transcriptional response in *P. falciparum*. Given these findings, it seems probable that further research will unravel other cues for which the parasite has evolved directed responses. Having proven that *P. falciparum* can mount transcriptional responses, the question remains of the relative importance they bear in the general biology of the parasite and which specific TFs may drive them.

1.4 A mixed mechanism of adaptation

The mechanism that regulates sexual commitment may represent a bridge between epigenetic adaptation and directed transcriptional responses. The master regulator of sexual conversion, *pfap2*-g, is found in a heterochromatic state in most parasites, and this silenced state is inherited by daughter cells. Unlike other CVGs though, its active state is never transmitted through the IDC because parasites in which the gene is active convert into gametocytes and exit the IDC. Despite this, subclones maintain a constant sexual conversion rate for many generations, implying that, somehow, the probability of activation of *pfap2-g* is transmitted from one generation to the next. Our data from Article 1 has unraveled a region upstream of the *gdv1* as-lncRNA in which the enrichment of H3K9me3 correlates with the basal sexual conversion rates of different subclones. This raises the possibility that this region determines the probability of activation of *gdv1* which in turn determines the probability of activation of *pfap2-g*. Additionally, *gdv1* is known to be upregulated in sexual conversion-inducing conditions, such as lysoPC depletion [227,354] or treatment with some antimalarials [356]. How the parasite senses these conditions and how it increases the probability of activation of *gdv1* in response to them remains to be elucidated. As we have seen, current evidence indicates that sexual conversion is determined at the epigenetic

level, and thus is a heritable characteristic driven by epigenetic adaptation, but it is also able to respond to external cues, activating directed responses. Whether only basal sexual conversion rate is determined by the heterochromatin domain of *gdv1* as-lncRNA and external cues act directly on *gdv1* expression or whether these cues can lead to an expansion/retraction of the heterochromatin in the same *gdv1* as-lncRNA domain should be the matter of further research. In any case, this represents an interesting scenario where epigenetic mechanisms could be coupled to environmental sensing and transcriptional responses. Although current evidence only suggests the presence of such epigenetic-coupled sensing for the sexual conversion switch, we cannot rule out the possibility that similar mechanisms may affect other clonally variant genes.

We envision a probabilistic model in which CVGs are switched on and off stochastically, with the probability of activation being a conserved characteristic of each specific member of a CVG family, and with different effectors being able to alter these activation/silencing probabilities, at least for a few CVGs, in response to certain stimuli. This model would account for the stochastic generation of variability, which enables bet-hedging adaptative strategies, but would also accommodate a more targeted adaptative response to environmental cues for some specific cases, like sexual conversion.

Our results from Article 2 are consistent with the probability of activation of each CVG being an intrinsic property of the gene. After the resetting of CVG expression patterns during transmission stages, new patterns are established in our volunteer samples. If each member of each CVG family had a similar probability of activation, for each family of CVGs we would expect all members to be expressed in the population with a similar frequency (in the case of mutually exclusive expressed CVGs, each one in a proportion of 1/n parasites, where “n” would be the number of family members). Instead, very similar global patterns of CVG expression between volunteers arise, with a few members of each CVG family being expressed in all volunteers and only minor differences in the expressed CVGs between volunteers. Under our model, the commonly expressed members of each CVG family would represent the ones with a higher activation probability, while minor differences between volunteers could be attributed to the probabilistic nature of the model. Regarding targeted responses to environmental cues, as we have seen,

previous studies have shown how certain environmental factors such as lysoPC depletion or the presence of antimalarial drugs can influence the expression of *ap2-g,* which is under epigenetic control. Therefore, there must be at least some factors that can alter the probability of activation of either *gdv1* or *ap2-g*, acting on their heterochromatin regulation. The fact that many known ApiAP2 TFs, as well as other chromatin readers/writers/erasers, have been shown to interact with heterochromatin [192,194,202,269,271,274,282,283] raises the possibility that some of them can also modulate CVG activation/silencing probabilities, perhaps in response to certain stimuli. A recent paper has hypothesized that lysoPC depletion can result in a scarcity of methyl donors that ultimately leads to decrease in H3K9me3 levels and *pfap2-g* de-repression [405], a mechanism that would directly couple metabolite concentration sensing and epigenetic regulation. Which set of stimuli can the epigenetic machinery of the parasite respond to, and which different sensing mechanisms does the parasite employ remain questions for further research.

2 Heterochromatin distribution in "pure" active/silenced CVGs

We have seen the importance of CVGs in the adaptation of the parasite to many challenges during human infections. In order to gain a deeper understanding of the heterochromatin transitions that govern CVG switching, we wanted to characterize the H3K9me3 enrichment patterns that define the active and silenced state of CVGs, and then compare them.

Previous studies had demonstrated that the presence/absence of heterochromatin was associated with the silenced/active state of CVGs [230,297,311,331], but the heterochromatin distribution underlying these switches had only been described for a few genes [272,331,406,407], mostly using ChIP-qPCR (which has a very low resolution, compared to ChIP-Seq). The active and silenced state of CVGs usually occur in a subset of the parasites in a population. When trying to analyze the heterochromatin distribution associated with such transcriptional states, mixed results arise from heterogeneity in the population (Figure 17). Single-cell techniques would be an appropiate way to interrogate individual active/silenced states, both at the transcriptional and

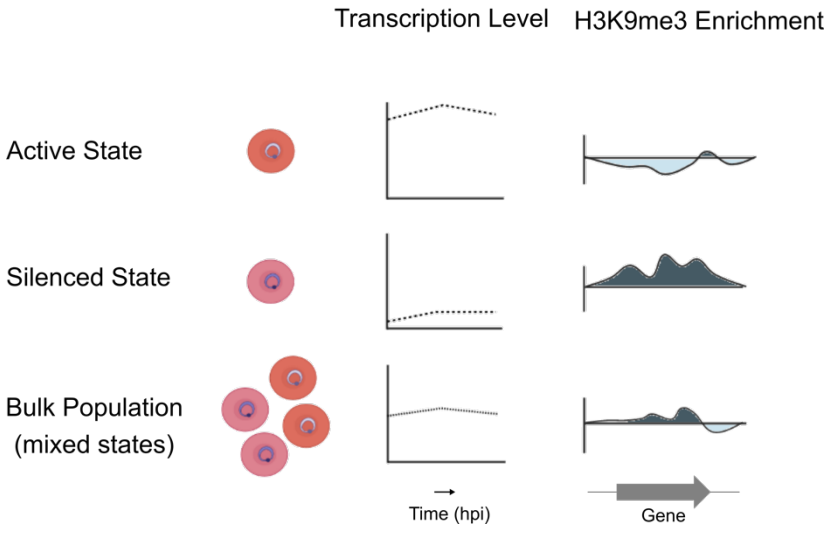


Figure 17. Mixed patterns of expression and heterochromatin in bulk populations. Transcription column represents expression of a specific CVG over time (as we would represent it in a microarray time-course experiment) and H3K9me3 enrichment column represents the data for the same gene as we would represent it on a ChIP-Seq experiment. The bulk population represents a typical population where the gene is active in some parasites and silenced in others.

epigenetic level, but three main limitations hinder their application. First, single-cell ChIP-Seq is still under active development and has never been applied to *P. falciparum*; second, current single-cell RNA-Seq methods capture only a fraction of the transcriptome, the most highly expressed genes, and many CVGs are expressed at low levels; and third, the obtention of both ChIP-Seq and RNA-Seq data from a same single cell is not feasible, due to the differences in extraction protocols. Instead of single cell approaches, we made use of subcloning to overcome the heterogeneity problem. Since the active/silenced state of CVGs is inherited by the progeny of a cell, and clonal populations arise from a single parasite, they should be homogeneous for their heterochromatin distribution at CVGs. Having homogeneous populations allowed us to use bulk techniques (two-channel gene expression microarrays and ChIP-Seq, in our case) and yet obtain considerably "pure" active or silenced states for most CVGs. To obtain sufficient material for microarray and ChIP-Seq analysis, the subclones had to be cultured for ~4-5 weeks after limiting dilution, which could still give rise to some stochastic variation in heterochromatin distribution, but given the low switching rates of CVGs, heterogeneity should be minimal. Indeed, for some genes, we observe transcriptional differences among subclones in CVGs that do not display H3K9me3 enrichment differences in the ChIP-Seq data, probably owing to their activation in only a minority of the parasites in the culture. Despite this limitation, this experimental setting has allowed us to characterize the active and silenced states of over a hundred CVGs, with unprecedented detail.

3 Types of heterochromatin distribution patterns

Our homogeneous parasite populations allowed us to compare, for more than a hundred individual CVGs, the heterochromatin distribution between the active and silenced states of those genes. To characterize the heterochromatin transitions linked to the shift between the active/silenced states of CVGs, we quantified H3K9me3 enrichment in the neighboring, upstream, coding, and downstream regions of highly differentially expressed CVGs (Article 1). Three main types of heterochromatin transitions arose from our analysis.

3.1 Expansion/Retraction transitions

In the first type, which we named expansion/retraction transitions, silencing occurs by the spreading of heterochromatin from neighboring genes into the upstream putative regulatory and coding regions of a CVG. In the silenced state, sometimes heterochromatin spreads over only a fraction of the promoter and coding sequence, other times it can cover the promoter and coding sequence completely or even reach the neighboring genes. Since the common denominator in the silenced state was the presence of heterochromatin on the putative regulatory upstream region, this suggests that, for these CVGs, H3K9me3 enrichment in the putative upstream regulatory region is the determinant of transcriptional state, rather than the heterochromatin presence in the coding sequence.

An especially relevant example of this are the heterochromatin patterns that we observed in the *ap2-g* locus. Our analysis, before induction, of parasite cultures that would display a 90% conversion rate after induction (Article 3) revealed a pattern of H3K9me3 distribution on *pgap2-g* consistent with the importance of heterochromatin at the regulatory upstream region. Even in those parasites that readily converted upon rapamycin induction, *pfap2-g* remained fully heterochromatic in its coding sequence, and only the upstream region of the gene was depleted

of heterochromatin. On the other hand, in parasites which failed to convert, heterochromatin was present in the upstream region of the gene.

Expansion/retraction transitions could occur from the boundary of heterochromatin domains or in the middle of these domains. In the second case, an euchromatic region was observed in the active state inside a region that was continuously heterochromatic in the silenced state. We hypothesize that in these cases, when a CVG becomes active, heterochromatin initially dismantles in a particular region and then it retracts from there, forming the relatively wide euchromatic region we observe in the active state. Heterochromatin dismantling from the boundaries of heterochromatic domains was as frequent as dismantling from the middle of such domains, which could be explained by two non-excluding possibilities. On one hand, the “opening” of a previously heterochromatic region, and posterior retraction of heterochromatin, could be as common as the retraction of heterochromatin from its limits, meaning that “opening” is as frequent as “retraction”. On the other hand, these initial “opening sites” could actually correspond to regions of a continuous heterochromatic domain which are placed on the boundaries of the heterochromatic compartment in the nucleus in its three-dimensional conformation. These regions in the interphase between the euchromatic and heterochromatic compartments could be more prone to transition between the two states. In this second scenario, most transitions would take place in the boundaries of heterochromatic domains, although in a three-dimensional definition of “boundary”.

Some expansion/retraction transitions showed maximal heterochromatin differences at the downstream region of affected genes. Many genes that showed extensive heterochromatin differences at the downstream region did also display minor differences at their upstream region. Therefore, these heterochromatin patterns might reflect true regulatory events at the downstream region of genes, or downstream differences might just be related to the transcriptional status of neighboring genes, while true regulation occurs at the upstream region.

3.2 Localized opening transitions

A second type of transition, which we termed localized opening transitions, was described in genes which were highly heterochromatic in their upstream, coding, and downstream regions both in the active and silenced states. In these genes, only a narrow opening in the upstream heterochromatin domain seemed to drive the transcriptional switch. We hypothesize that this represents a different mode of active/silence heterochromatin transitions. Instead of the spreading/retraction of relatively large heterochromatin domains seen in the first type of transitions, in these genes the targeted removal of heterochromatin in a localized narrow region, again in the putative upstream regulatory region, would be sufficient to drive transcriptional activation. This mechanism seems to be characteristic, but not restricted to, of the members of the *pfmc-2tm* family, which were highly enriched in this category.

3.3 *De novo* formation/complete removal transitions

A third kind of transitions, which we termed *de novo* formation/complete removal transitions, consisted in the formation/removal of heterochromatin domains largely isolated from other heterochromatic regions (more than 5 kb). This kind of transitions occurred very scarcely, and the vast majority of genes that showed them (19/20) were mainly expressed in transmission stages (not during the IDC). Although it might be tempting to relate this *de novo* formation of heterochromatin with the extensive remodeling of the parasite's epigenetic landscape that occurs during transmission stages [254,273], the subclones analyzed in our work (Article 1) where obtained from a 3D7 parental clonal line that has not gone through transmission stages and has been cultured in the lab (in a "perpetual IDC") since its obtention. Therefore, these changes must have happened during the IDC, and why they happen preferentially in genes expressed in other life cycle stages remains something to be further studied. In any case, our results indicate that dynamic adaptation to environmental fluctuations during the IDC does not commonly rely on *de novo* formation or complete removal of heterochromatin domains.

3.4 Differentially expressed clonally variant genes without H3K9me3 enrichment differences

Finally, there was a group of CVGs genes, with high H3K9me3 enrichment along their sequence, and virtually no differences in heterochromatin distribution despite their transcriptional differences. In these genes, the lack of heterochromatin distribution differences might be caused by different factors. One possibility is that these genes are expressed in only a small subset of parasites, owing to stochastic switches in their epigenetic state during culturing. As discussed in Article 2, this could result in big transcriptional changes but undetectable heterochromatin changes. To check whether genes without observable heterochromatin differences had low absolute expression values, we used the Cy5 channel of our gene-expression microarrays to check the expression levels in the active subclone in each comparison. Absolute expression values of genes from cluster 5 from Article 1, which roughly correspond to genes without heterochromatin differences, tended to be lower than absolute expression values from clusters with bigger heterochromatin differences (clusters 1 and 2), but the dispersion of the data was too big to draw firm conclusions (Figure 18). Another option would be that these genes are expressed in gametocytes. Given the differences in sexual conversion rates, the cultures are expected to have

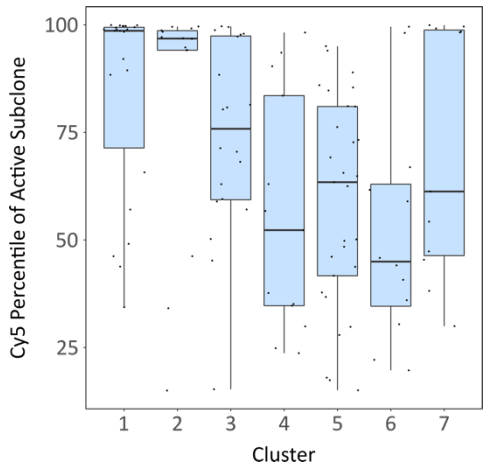


Figure 18. Cy5 signal percentiles for the active subclone in the different clusters from Article 1 Fig. 4.

different amounts of gametocytes, which would justify the transcript level differences. However, very few of these genes were annotated as gametocyte related genes or were highly expressed in gametocytes, when compared to asexual parasites. Yet another possibility would be that transcriptional differences arise from genetic differences that we failed to identify in our genetic variants analysis. In any case, these genes probably represent undetectable heterochromatin differences (probably due to being active in only a small subset of the population) or transcriptional differences not attributable to epigenetic changes, instead of a distinct kind of heterochromatin transition.

3.5 Implications of different heterochromatin transition types

Further research will be needed to assess whether this different heterochromatin transition types represent truly different activating/silencing mechanisms. An exciting possibility would be that expansion/retraction transitions represent stochastic expression switches, while localized transitions represented targeted events. Under this hypothesis, expansion and retraction of heterochromatin would underly the generation of variability in a population and enable bet-hedging adaptative strategies by ensuring the continuous generation of new heterochromatin patterns. On the other hand, localized opening transitions would correspond to the targeted action of some heterochromatin interacting factors, that could be activated in response to different stimuli. However, experimental support for this hypothetical model is lacking.

Using subcloned populations, the application of gene editing technologies to perturb known heterochromatin interacting factors (such as histone PTM writer/eraser/readers or heterochromatin interacting ApiAP2 TFs), coupled with the analysis of the resulting heterochromatin distribution patterns, could help characterize their role in heterochromatin regulation. These kinds of analyses could suggest whether different effectors have different impacts on the different heterochromatin transition types described in this thesis. Coupling genetic editing with changes in the experimental conditions in different samples could also shed

some light into the possible coupling of the sensing of those conditions and heterochromatin regulation at CVGs.

3.6 Heterochromatin transitions as a continuous process

It is important noting that we cannot rule out the possibility that these different heterochromatin transitions are in fact a continuum, rather than discrete, independent mechanisms. Localized openings could represent the first step of CVG activation, from which heterochromatin could then retract, resulting in expansion/retraction transitions. Isolated heterochromatin domains, which we have classified as *de novo* formed heterochromatin, could instead form by the expansion of a previously existing heterochromatin domain followed by an opening and retraction in the middle of it, leaving a portion of the heterochromatin isolated (Figure 19). Should this be the case, all transitions could in fact be governed by a similar opening and expansion/retraction mechanism.

3.7 Experimental alteration of heterochromatin patterns

Technologies that enable the targeted deposition/removal of histone PTMs would be an excellent tool to test some of the hypothesis generated by our data. Recently, the fusing of histone writers/erasers to a catalytically dead Cas9 (dCas9), which can bind to DNA but does not cut it,

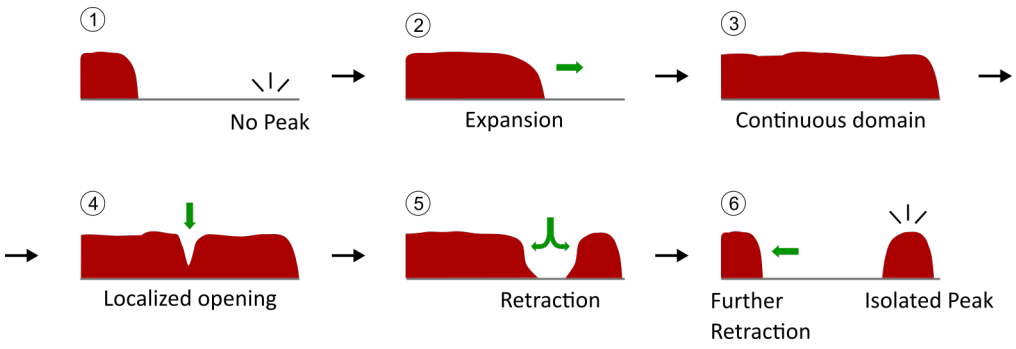


Figure 19. Continuous heterochromatin transitions. The different types of heterochromatin transitions discussed in this section could be different steps of a continuous process rather than fully independent processe . Internal retraction of heterochromatin could be initiated by a localized opening followed by further retraction (steps 4-5). The formation of de novo peaks could start with heterochromatin expansion followed by retraction in an internal region (steps 1-6).

has been proposed as a way to achieve this. Two recent studies, by Xiao *et al.* [408] and by Liang *et al*. [409], have successfully fused the catalytically active domain of the histone deacetylase Sir2a and the active domain of the acetyltransferase GCN5 to dCas9, the latter under the regulation of a DiCre inducible system. The targeting of dCas9-GCN5/Sir2a to specific genes resulted in their up or down-regulation, respectively, proving the functionality of the system [408,409]. The results obtained with this system suggest that the deposition or removal of H3K9ac might be sufficient to trigger the activation or silencing of CVGs. Based on current evidence, we hypothesize that the active/silenced state of CVGs is governed by a dynamic equilibrium between the factors that deposit H3K9me3 and H3K9ac, in such a way that removal of H3K9ac might indirectly result in methylation of the site by alteration of this equilibrium, and deposition of H3K9ac can in turn lead to a removal of H3K9me3. Such system could be used to confirm that localized H3K9me3 depletion at the putative regulatory upstream region is sufficient for the activation of CVGs in which we observe this kind of transition. To confirm this, we should design a guide to specifically target the system to the same localized opening we observe in our active subclones and apply the system to our silenced subclones, with appropriate controls. We should then confirm that a similar limited dismantling of heterochromatin occurs, and that this results in the activation of the targeted CVG. A similar approach could be used to confirm or refute that heterochromatin at the downstream region of some CVGs regulates their expression. Again, we could target a guide to the downstream region which is depleted of heterochromatin in our active subclones and try to reproduce this depletion in our silenced subclones.

4 Presence of “weak” and “strong” insulator elements

Our data set of homogeneous parasite populations, regarding their active and silenced CVG expression patterns, and their associated heterochromatin distribution, results in an enhanced resolution on heterochromatin/euchromatin limits and offers a good opportunity to study the boundaries of heterochromatic domains in *P. falciparum*.

4.1 Insulators, what we already know

Two main kinds of insulator elements govern the three-dimensional organization and the spreading of heterochromatin in most eukaryotes. On one hand, enhancer blocking insulator sequences act by recruiting insulator proteins (like CTCF and its interactor cohesin) which can form loops and insulate TADs, governing the three-dimensional organization of chromosome territories [410]. It is important to note that the disruption of these elements doesn’t result in the loss of the compartmentalization of the nucleus in a euchromatic and a heterochromatic compartment (referred to as A and B compartments, respectively) but rather enhances it, suggesting that the formation of TADs results in the interaction between active and silenced regions of the genome that would otherwise be in separate compartments [411–413]. On the other hand, a second type of insulators exist, usually referred to as barrier insulators, which limit the spreading of heterochromatin from neighboring regions. Barrier insulators (e.g, HS4 element) exert their function by either recruiting euchromatin promoting factors or by tethering the insulator site to regions of the nucleus enriched in euchromatin promoting factors [414].

Not much is known about the mechanisms that limit heterochromatin spreading in *P. falciparum* and govern its genome-wide restricted distribution. CTCF does not have known homologs in *P. falciparum* [415], and typical chromosomal territories are not formed [229]. Previous studies revealed that heterochromatin distribution in specific CVGs had conserved limits, suggesting the existence of some kind of barrier insulator elements that restricted the spreading of heterochromatin

beyond a certain point [331,345], although such elements were not identified. In *var* genes, insulator-like pairing DNA elements have been identified in both their upstream promoter and their intron promoter. These pairing elements have been proposed to mediate the interaction between both promoters, leading to *var* repression [416].

4.2 Strong and weak insulators, a hypothesis

Given that heterochromatin limits are conserved between different subclones/strains both at chromosomal distribution level and in individual CVG, we hypothesize the existence of two types of general insulator elements in *P. falciparum*. On one hand, the restricted nature of facultative heterochromatin distribution in *P. falciparum*, which spreads only along CVG clusters, and never further, suggests the presence of some kind of "strong" insulating elements that restrict heterochromatin to these domains. On the other hand, the activation/silencing of CVGs, in which heterochromatin only spreads until certain points in each state, suggests the presence of another type of "weak" insulating elements, which limit the spreading of heterochromatin in the active state, but can be overcome in the silenced state of CVGs (Figure 20).

4.3 Strong Insulators

We attempted to identity "strong" insulators that preclude the expansion of heterochromatin outside CVG clusters boundaries. Our data confirms that heterochromatin in the subtelomeres spreads only until a certain point, which is chromosome specific, suggesting the implication of some kind of sequence element implicated in the insulator role. Using our ChIP-Seq data, we searched for sequence elements enriched in heterochromatin-euchromatin limits at the end of subtelomeric heterochromatin domains and in the intrachromosomic heterochromatic islands, but we did not identify any significantly enriched motif. Many factors can explain this lack of success. The fact that several heterochromatic domains expand during gametocytogenesis [254,273] suggests that these insulating elements are dynamically regulated and help define the different active and silenced compartments during different life cycle stages (unlike cohesion and CTCF in

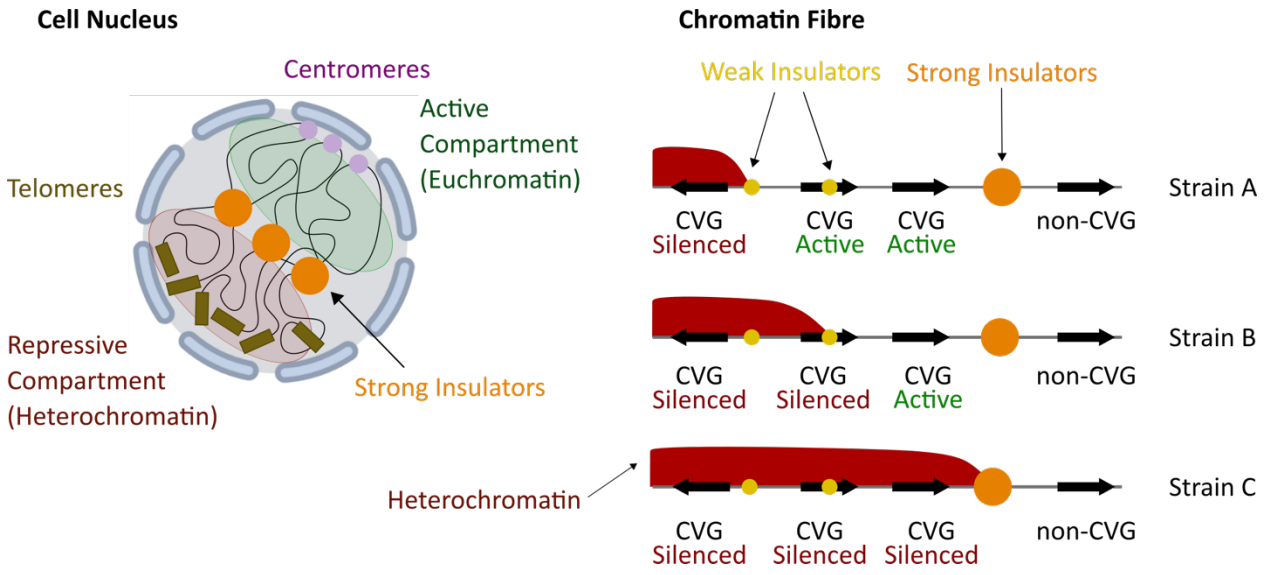


Figure 20. Proposed model for strong and weak insulator elements. Strong insulator elements delineate repressive from active compartments inside the cell nucleus and demarcate the limit of heterochromatin expansion in clonally variant gene (CVG) clusters in chromosomes. Heterochromatin never spreads beyond these strong insulators. Weak insulators act as "barrier" elements, impeding the spread of heterochromatin in the active state of CVGs, but can be overcome when a CVG is silenced.

most eukaryotes, which define chromosome territories but do not separate active and silenced compartments). Moreover, heterochromatin domains cluster together in certain areas of the nucleus [229,231,247], in a specific three-dimensional conformation. The dynamic and three-dimensional nature of the putative insulator elements, which would define the overall chromatin organization in the nucleus, could imply that simple linear sequence searches might not be enough to identify them. Other methods, able to incorporate three-dimensional data, like Chromatin Conformation Capture (3C) and its derivatives, or imaging techniques with whole cells, together with sequence data, may be needed to identify such insulating elements. These techniques could identify heterochromatic loci that are close to euchromatin regions in the three-dimensional space, and are thus in the boundaries of heterochromatin domains, even if they appear to be in the middle of broad heterochromatic regions when looking at chromosomes linearly. Another possibility is that the insulating role of some sequences is not determined by short sequence motifs but by the properties of longer sequences, like the relative AT-richness, or specific repeats.

4.4 Weak Insulators

Next, we attempted to identify sequence elements that could act as “weak” barrier insulators. In our dataset, the silenced state of most CVGs was characterized by the expansion of heterochromatin along the upstream and part of the coding region of CVGs, which could expand from further upstream or downstream regions. The expansion of heterochromatin reached different points along the gene sequence in different CVGs. In some CVGs, heterochromatin expanded only until half of the coding sequence, while in others it expanded beyond its coding sequence, up-to neighboring genes. However, one of the most surprising characteristics of our dataset was that for the same CVG, the heterochromatin expansion limit was always the same in different subclones that shared the same transcriptional state. To put it in another way, for any particular CVG, the boundary of heterochromatin in its silenced and active states was conserved among subclones. In the few cases in which this was not true, the differences in the heterochromatin boundary between two subclones could be usually explained by the different state of a neighboring CVG. Our data strongly suggests the presence of some kind of “weak” insulator elements, which can be overcome when a gene switches its transcriptional state. We hypothesize a model in which when a CVG is in an active state, heterochromatin cannot usually expand beyond a certain point (the weak insulator), but eventually it becomes able to overcome this “barrier” and expand until the next weak insulator (or strong insulator), rendering the gene silenced. Whether the overcoming of this weak insulators occurs spontaneously or is a consequence of the combination of certain effectors, such as the targeted effect of a TF or a certain histone reader, remains a question for further research.

Again, we failed to identify any sequence elements in our dataset enriched in euchromatin-heterochromatin limits of CVGs, that would be compatible with such weak insulator role. Unlike defining the boundaries of broad heterochromatic regions of the genome, which is relatively straightforward, identifying the limits of individual gene heterochromatin boundaries from ChIP-Seq data relies on the correct identification of differential-peaks, which is not trivial. For this reason, we cannot rule out the possibility that we failed to identify any sequence motifs just

because our pool of candidate sequences was not ideal. However, the possibility that these weak insulators, should they really exist, are sequence-based is just one among many, and as it was the case for strong insulating elements, other techniques, especially those that can take into account three-dimensional data, might be better suited for this purpose.

5 On the problem of identifying Differential ChIP-Seq Peaks

The homogeneity of the parasite lines used to generate our dataset, regarding CVG expression, hugely enhanced our ability to compare the heterochromatin distribution patterns between different subclones. One of the typical ways in which differences in the distribution of any chromatin associated element are characterized is a process known as differential peak-calling. Differential peak-calling consists in the identification, by statistical procedures, of regions of the genome that are differentially enriched in the chromatin-associated element of interest between two samples, in our case H3K9me3. Differential peak-calling should result in an unbiased definition of what we consider a significant difference in H3K9me3 enrichment between samples. In our case, differential peaks should roughly correspond to the regions of the genome that are heterochromatic in one sample and euchromatic in another.

Identifying differential peaks between different samples is an inherently difficult problem. Many different software tools exist for this task, which may yield considerably different results for the same dataset. Even when focusing on a single tool, different sets of parameters used to run it may result in different sets of differential peaks. The lack of a gold standard also hinders the ability of establishing what a “correct” differential peak calling should look like. Ideally, a differential H3K9me3 peak should represent a region which is euchromatic in one sample and heterochromatic in another, but we must consider that, as is the case with all bulk sequencing techniques, the coverage we obtain from a ChIP-Seq experiment represents the mean enrichment of H3K9me3 over a given region for a whole parasite population and is thus never that “clean”. The different presence of the mark in different individual parasites may result in intermediate enrichment values for each sample that, when compared, give rise to a whole range of enrichment differences between samples. Other technical reasons, possibly related with the primary sequence of some regions, may also result in intermediate coverage values. Even in our highly homogeneous subclones, intermediate levels of H3K9me3 enrichment were observed for

some CVGs. The other main problem of differential peak-calling is the omnipresent normalization problem. Different samples can have different number of total reads, different signal-to-noise ratios, and other biases, which further complicate the task of faithfully comparing them.

5.1 Differential Peak-Calling in our subcloned samples

When visually inspecting H3K9me3 enrichment on our samples, the striking similarity in the distribution of the mark between subclones was immediately apparent. This similarity made the restricted set of differences in particular CVGs quite obvious by simple visual inspection. Despite that, when running differential peak-calling routines from different software tools (MACS2, MAnorm, SICER) the set of obtained differential peaks did not match, in many cases, the clear differences we observed by visual inspection. It is worth noting that we were restricted in our software choices by the fact that our study design did not include replicates. The discrepancy between the differential peaks found by the peak-calling software and by visual inspection could be due to many factors. Of note, most differential peak calling tools are designed for the analysis of datasets which are very different from ours. Most differential peak-calling analyses try to identify large differences in the targets of a TF or the distribution of a histone PTM under different conditions (e.g., treatment vs control) or between largely different cell types. In those analyses, the TF or histone PTMs analyzed usually have broadly different distributions in different samples. In our samples, the distribution of H3K9me3 is mostly identical and is only different in small regions inside otherwise identical broadly heterochromatic domains, making the search space very different. In addition, our differential peaks could only be present in the subtelomeres and a few internal islands, rendering the distribution of the mark in most of the genome uninformative (and possibly biasing the signal distributions modelled by the peak-caller).

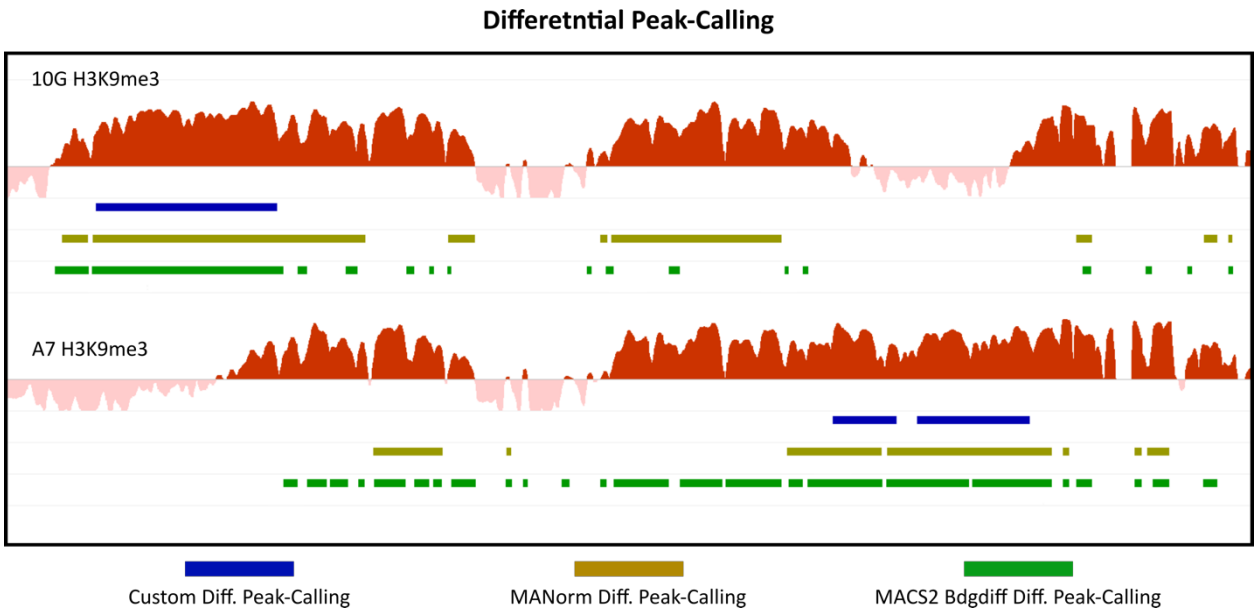


Figure 21. Representative example of differential peak-calling with multiple methods. Data represents log2(H3K9me3/input) coverage tracks for 10G and A7 subclones, respectively, in a representative region of chromosome 2 (Chr 2:830-880kb). Peaks are always marked on the strain with greater H3K9me3 level.

Regardless of the reason for the poor performance of said methods, when applied to our data, we decided to implement a custom solution that could yield a set of differential peaks with a better agreement with the visual inspection of H3K9me3 enrichment differences in our subclones. We restricted the search of differential peaks to heterochromatic regions of the genome (which we identified using MACS2 individual peak-calling, which seemed to perform very well on our samples). Coverage was calculated for each sample in 100 bp bins and only those bins overlapping an individual peak called by MACS2, in any of the two compared samples, where retained. A skewed normal distribution was fitted to these filtered coverage values, generating a probability density function (PDF) for each sample. A cumulative density function (CDF) value was then calculated for each bin and sample, using each sample's PDF, and bins with a difference of CDF value greater than a threshold (0.3) where considered as differential peaks. Finally, differential peaks separated by less than 500 bp were joined together and resulting differential peaks smaller than 1000 bp were filtered out. Despite the simplicity of the method, it did perform well in our samples, and yielded a set of differential peaks that was in better agreement with visual inspection of the data than any of the tested available software (Figure 21). In order to

validate these results, we compared, between the different differential peak-calling methods, the set of genes that overlapped a differential peak and where also differentially expressed according to our microarray transcriptional data. Again, our custom method outperformed the rest of the tested methods, both in terms of total number and proportion of differential peaks that corresponded to a differentially expressed gene (Figure 22). Our experience highlights the need to tailor the analysis methods to the data to be analysed, and to be cautious when applying universal solutions to specific problems.

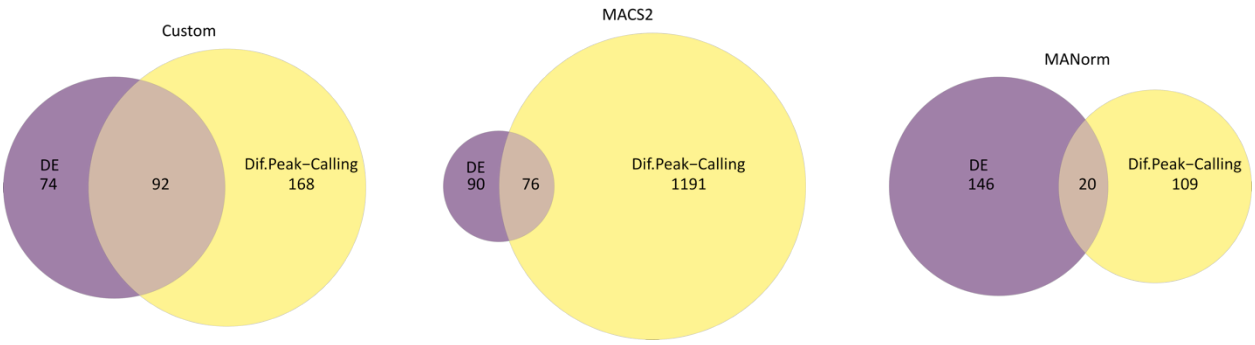


Figure 22. Comparison of genes differentially expressed (DE)(fold-change > 4) between pairs of subclones in our microarray data and genes overlapping a differential peak between the same subclones using different differential peak-calling methods (data has been aggregated for the different pairwise comparisons).

6 Categorizing the transcriptional state of CVGs and non-CVGs

Another major technical challenge we had to overcome was the classification of genes as transcriptionally active or silenced, at the whole genome level. In Article 1, we were interested in characterizing how differences in the enrichment of H3K9me3 and H3K9ac correlated with the transcriptional state of CVGs and non-CVGs. To be able to analyze those differences, the first step was to faithfully classify genes (both CVGs and non-CVGs) as active or silenced. The classification of genes into CVGs or non-CVGs was made using previously published data as described in Article 2. The classification into the active/silenced categories was made based on transcriptional data (both from our microarray-based analyses and publicly available RNA-Seq datasets). When trying to analyze different characteristics from the data (differences in the distribution of H3K9me3/H3K9ac, in this case) it is very important to start with a correctly labeled dataset. Wrongly categorized genes could potentially obscure relations between the variables in the data or even result in the imputation of incorrect relations between variables. For this reason, we have included the *undetermined* label in our classification, and only genes that could be assigned a transcriptional label with high confidence were classified as active or silenced.

Determining whether a gene is active or silenced from transcriptomic data is not as easy as it may sound. Given the highly dynamic range of gene expression, an active gene with a weak promoter, expressed in virtually all parasites in a population, could still be expressed at lower levels than a gene with a strong promoter that is silenced in the majority of the population, and only expressed in a minor subset of parasites. For this reason, looking at relative expression valus between different samples, instead of absolute expression values, may seem to be a good approach. However, differences in expression between genes expressed at very low absolute levels can still yield very high transcriptional fold-changes when comparing between samples and can correspond to activation in only a minor subset of parasites, implying that absolute expression values also must be considered. On top of that, each technique used to measure gene expression levels (microarrays and RNA-Seq, mainly) has its own unique possible biases when measuring

expression changes between samples. For these reasons, and many others, a perfect classification is not a feasible objective, and a level of uncertainty will always remain in the classification.

We developed our classification algorithm trying to balance all these considerations. In short, microarray data was used to classify genes with significant expression fold-changes between samples as active or silenced, respectively. The Cy5 channel of microarrays (surrogate of absolute expression) and publicly available RNA-Seq data were used to classify genes without significant expression differences between samples and to filter differences occurring in very lowly expressed genes. Additional filters where set to deal with problematic cases (such as genes with high fold-change between samples occurring at a point of the IDC different than their point of peak expression or genes with genetic differences between samples). Finally, as we have already mentioned, genes with intermediate levels of expression were classified as *undetermined* in order to avoid misclassifications as much as possible.

Similar approaches could be used to generate high confidence labeled datasets as this one, which would in turn be instrumental to train prediction/classification algorithms to predict the transcriptional state of CVGs from heterochromatin profiling data, a possibility we will discuss in the following section.

7 Prediction of transcriptional states from epigenetic profiling

Thoroughly characterizing the heterochromatin patterns of the active and silenced sates of CVGs might be the first step in accomplishing the inverse task, accurately predicting the transcriptional status of CVGs from heterochromatin profiling data.

Given their importance in host-parasite interactions, the expression of CVGs is highly relevant for malaria pathogenesis. Many CVGs are known virulence factors, including *rif* [313,314,316], *stevor* [318–320], and more importantly *var* [7,208–210,307,407] genes, while the role of many of the rest regarding disease severity is still poorly understood. As such, studying the patterns of expression of CVGs in natural infections, and their relationship with different clinical outcomes and/or epidemiological settings would be of great interest for better understanding the determinants of pathogenesis and severity of the disease. One of the main problems when trying to study CVG expression in natural infections arises from the fact that the only parasite forms found circulating in the blood stream of patients are ring stage parasites and mature stage V gametocytes. Parasites at other life cycle stages remain sequestered in different tissues. Many CVGs are expressed in stages of the IDC different from the ring stage and cannot therefore be assessed directly from blood samples from patients. Some studies, including our results from Article 2, have used ex-vivo culture to overcome this limitation [417,418], but this is a complicated and time-consuming procedure that requires specialized facilities, which hinders its feasibility in large scale studies or in some malaria endemic regions.

Contrary to patterns of CVG expression, heterochromatin patterns underlying them are stable across the IDC [272,273,331,407]. In the work presented in this thesis we have seen that heterochromatin patterns associated with the silenced or active state of a particular CVG are highly conserved among subclones. This raises the possibility of predicting CVG transcriptional states along the whole IDC solely from their heterochromatin patterns, which could be obtained from circulating ring stage parasites. Heterochromatin profiles could be assessed by ChIP-Seq

profiling of H3K9me3 or PfHP1, and other histone PTMs, especially H3K9ac, could help refine the predictions. H3K9ac could be particularly helpful when intermediate levels of H3K9me3/PfHP1 are found. In that situation, it is difficult to discern whether these intermediate levels are an intrinsic property of the locus, a sequence dependent poor mapping, or indicate the presence of both the active and the silenced states in different parasites in the population. Given that H3K9ac and H3K9me3/PfHP1 are mutually exclusive, the presence of both histone PTMs in a region would clearly indicate that the underlying gene is active in some parasites and silenced in others, clarifying the situation. Although ChIP-Seq is not an easy technique to perform, especially in complicated settings, it would still be a great simplification over the culture of ex-vivo samples and their transcriptional characterization at multiple time-points. Moreover, recent advances in Cleavage Under Targets and Tagmentation (CUT&Tag) techniques, which build on the same principles and yield very similar information as ChIP-Seq about the distribution of a chromatin binding protein of interest, could greatly reduce the amount of input sample and processing time necessary for heterochromatin profiling.

Predicting the transcriptional state of every CVG from heterochromatin profiling in a bulk population would not be an easy task. First, the expression of CVGs in a natural infection may not be homogeneous, resulting in intermediate heterochromatin active/silenced patterns. A certain level of heterogeneity will always be present in natural infections, and indeed our results from Article 2 suggest that human infections are heterogeneous for the expression of many CVGs, from their onset. However, natural selection of the combinations of active/silenced CVGs is expected to result in increasingly homogeneous patterns of CVG expression as the infection progresses. Moreover, in longstanding infections where adaptation has already taken place, CVGs showing high levels of heterogeneity are probably the less relevant for the parasite survival. Another possible problem arises from the fact that even in parasites in which a particular CVG is in the same state, the heterochromatin distribution of the gene might be influenced by neighboring CVGs, rendering its heterochromatin profile in different parasites dissimilar. Although we have observed this phenomenon in our subclones from Article 1, a sufficient number of training samples could identify the relevant region of a gene when evaluating its state (which would

correspond to the one that changes between the active and silenced states but does not change depending on the state of neighboring genes). Additionally, as we have seen in Article 1, there is not a universal pattern of CVG silencing. Different CVGs have different patterns of heterochromatin in their silenced and active states (represented by the different heterochromatin transitions we have identified), and in some CVGs heterochromatin distribution differences between the active and silenced states are minimal, all of which should be considered when making the predictions.

This problem could be tackled both as a prediction problem or a classification one. In both approaches, the input data would be the heterochromatin profiling data. In the prediction approach the output would be a predicted transcription value, while in the classification approach, the output would be a categorical label, such as *highly expressed/expressed /silenced*. In practice, the boundary between both kinds of methods is a bit imprecise, since predicted values can usually be used for classification by applying a decision boundary.

Machine learning algorithms can perform both prediction or classification, and can identify complex relations in the data, making them good candidates for such a task, but they need a sufficient amount of training data to perform well. For the training set, a considerable amount of heterochromatin profiling data with its corresponding transcriptional data (RNA-Seq or microarray data) should be generated. Ideally, a broad range of strains of different genetic backgrounds should be used, in order to better represent the heterogeneity present in natural infections. An important limitation of the dataset in Article 1 is that all parasite subclones compared are of the same genetic background. With a heterogeneous and big enough dataset, it should be possible to predict with considerable accuracy the transcriptional states of all CVGs. Whether H3K9me3/PfHP1 profiling alone could yield an accurate enough prediction or H3K9ac, or some other histone PTM data, would be needed, should be assessed while training the predictor.

At the cost of generating the training set (which could be greatly reduced by re-using already published data), once trained, such a tool would allow easy determination of the whole profile of

CVG expression in natural infections, just from heterochromatin profiling of circulating ring stage parasites. This would allow to interrogate the relation between the transcriptional state of particular CVG variants and patients with different ages, clinical outcomes, pregnancy status or different epidemiological settings (such as level of endemicity or seasonality of infections). A better understanding of this critical set of host-parasite related genes would be of great help in developing better diagnostics, more accurate prognostics, and guiding possible treatments for patients as well as designing better malaria interventions in different epidemiological settings.

Conclusions

Conclusions

- Different types of heterochromatin transitions underlay the switch between the active and silenced state of CVGs.
- Heterochromatin levels at the upstream putative regulatory region and the start of the coding sequence of CVGs negatively correlate with their expression levels.
- Expansion and retraction of heterochromatin domains, and localized opening of heterochromatin at a small region, drive the switches in the expression of CVGs during the IDC, whereas *de novo* nucleation of heterochromatin domains or their complete dismantling are mostly restricted to genes expressed during transmission stages.
- The heterochromatin distribution associated with the active or silenced state of a given CVG is conserved among different subclones.
- Mutually exclusive expression of *var* genes is not a strict mechanism.
- The heterochromatin levels at the upstream region of the *gdv1* antisense lncRNA positively correlate with basal sexual conversion rates.
- Heterochromatin patterns of CVGs are reset and reestablished during transmission stages.
- Different CVGs have a different probability of becoming active after reestablishment of heterochromatin patterns during transmission stages, leading to the same CVG members of each family being the most expressed in the majority of parasites at the onset of a new human infection.
- The inducible expression of *pfap2-g* has demonstrated that its expression is sufficient to trigger sexual conversion and has allowed us to identify the early up and down regulated genes after sexual commitment.
- The *pfap2-g* coding sequence remains heterochromatic even in populations in which 90% of parasites will convert into gametocytes upon induction, indicating that heterochromatin at this position does not preclude gene activation. On the other hand, heterochromatin in its upstream putative regulatory region was depleted in parasites that

did convert into gametocytes upon induction and maintained in those that would not convert upon induction, suggesting a critical role in gene regulation.

- The ApiAP2 transcription factor *pfap2-hs* drives a very compact protective response to heat-shock, through the activation of *hsp70-1, hsp90* and PF3D7_1421800. It is the first transcription factor identified to drive a directed transcriptional response in *P. falciparum*.

Agraïments

Vull començar aquests agraïments pel meu director de tesi, l'Alfred Cortés. Sense ell aquesta tesi no hagués estat possible. Parteix de les seves idees i s'ha nodrit extensivament del seu talent i la seva experiència en la recerca. A l'Alfred li haig d'agrair molt, tant a nivell científic com a nivell personal. Ha estat un director de tesi fantàstic, un autèntic mentor, que no només m'ha dirigit sinó que m'ha ajudat a créixer.

A nivell científic, he après moltíssim de la seva manera de veure la biologia. La seva atenció al detall, a entendre un sistema posant atenció als casos particulars abans d'intentar generalitzar de manera més sistemàtica, ha estat una lliçó especialment valuosa per a mi. També la seva mirada crítica. Llegir un *paper,* pensant que està força ben fet, i que l'Alfred el destrossi amb quatre observacions afilades, sobre les mancances que amaguen les dades, s'ha convertit quasi en una tradició al grup. Crec que puc parlar per tot l'equip si dic que l'Alfred ens ha fet partícips de totes les decisions importants que s'han pres al laboratori, ens ha encoratjat a opinar i expressar els nostres punts de vista sobre la nostra recerca i la dels companys i s'ha implicat personalment en la recerca que hem realitzat tots i cadascun de nosaltres. Si soc una mica millor científic ara que quan vaig començar aquesta tesi, és sobretot gràcies a l'Alfred.

A nivell personal, no podria haver tingut una millor relació amb el meu director. He gaudit molt de l'ambient al laboratori que, ell el primer, ha ajudat a generar. L'Alfred ha estat el primer en demostrar-nos que la professionalitat no està renyida amb el bon humor i la manca de sentit del ridícul (els vídeos d'algunes tesis són llegendaris). M'he sentit valorat i escoltat, i també redirigit quan m'ha fet falta. L'Alfred ha estat una persona propera que ha exercit el seu lideratge amb empatia i que s'ha convertit no només en un director sinó en un amic.

Moltes gràcies, Alfred.

També voldria fer un agraïment molt sentit als meus companys de laboratori, especialment als *Epis*: la Núria (Rovira), la Cristina, l'Oriol, la Eli, l'Anastàsia, en Harvie, la Carla, la Sofia, en Rafa, la Rosa, la Núria (Casas), l'Alba, la Neus, la Sandra i la Íngrid. He passat uns anys fantàstics al *lab*, plens de moments divertits i d'anècdotes, a més de molta ciència. Enlloc es passen tantes hores al dia com al lloc de treball i tots ells l'han convertit en un espai amable i acollidor, on feia una mica menys de peresa venir cada dematí. En el meu rol com a bioinformàtic, he tingut la sort de

poder col·laborar molt estretament amb la majoria d'ells i de tots n'he après moltes més coses de les que podria encabir en uns agraïments. Ells també han fet de mi un millor científic i una persona més sabia. El que m'emportaré de tots ells per sempre però, va molt més enllà del coneixement. M'emporto infinitat de dinars al pati del CEK, amb les seves converses sempre un punt surrealistes (amb una quantitat de temps exagerada dedicada a la disquisició sobre fruites o sobre el sexe al món animal), els sopars a casa de l'un o de l'altre, les sortides de grup (si calia sota la pluja i el vent), les cerveses als vespres, els riures histèrics enmig de la feina (normalment per alguna autèntica tonteria) i en definitiva, tants i tants moments bons, d'autèntica felicitat, d'aquella que rau en les petites coses. No podria haver tingut uns millor companys de laboratori, m'emporto uns amics per a tota la vida.

També voldria tenir unes paraules per a la resta de companys de laboratori, els *Nanos*, els *Micros*, els *Falcis* i tota la gent de la planta 1 del CEK. Amb molts d'ells he compartit també molts dinars, moltes converses i molts moments divertits dins i fora del *lab*. Amb altres no hi he pogut compartir tant de temps, sempre una mica aïllat allà al meu ordinador, però tots ells han contribuït, i molt, a crear un ambient de treball fantàstic, que he gaudit sincerament tots aquests anys de tesi. Moltes gràcies a tots.

Vull donar les gràcies a la meva família petita, el Pare, la Mare i el Max, mon germà. He tingut la gran sort de néixer en una casa on la ciència era un tema de conversa habitual a la taula, on la televisió era prohibida durant els sopars. Aquelles xerrades informals i les hores interminables de documentals amb el Max, d'animals sobretot, però també de física, d'astronomia i del que ens caigués a les mans, tenen molt a veure, sense dubte, amb aquesta tesi.

Evidentment, no seria on soc ara si no fos per la infinita paciència i sacrifici dels meus pares. M'ho han donat tot, i mai m'han demanat res a canvi (de moment). Han aguantat estoicament els meus canvis de rumb i els meus moments difícils i han estat els primers en compartir les meves alegries. Han esta incondicionalment al meu costat. I el més important, m'han estimat i m'han fet sentir estimat. A sobre, he tingut la sort de tenir una família que no només estimo amb bogeria sinó

que admiro profundament. Sempre sereu els meus primers referents. Moltes gràcies per tant, de cor.

En els últims anys, n'he guanyat una altra de família. Des del primer dia, m'han fet sentir com un més i s'han esforçat per animar-me, per ajudar-me i per fer-me sentir com a casa. Moltes gràcies Maite, Diego, Didac, Lubi, Pere, Ting, Camí, Andrew i a tota la família, heu estat un gran suport i vosaltres també m'heu fet sentir molt estimat. Moltes gràcies també a la Tawny, la seva alegria i entusiasme han estat un gran acompanyament durant aquest camí.

També voldria tenir unes paraules per a tota la gent que m'ha acompanyat durant tots aquests anys i que d'una manera o altra m'ha ajudat a arribar fins aquí. Els *garriguencs*, els tiets i tota la cosinada, els amics de la *uni,* els *possibles desfassats...* i tanta gent a qui no acabaria mai d'anomenar. Moltes gràcies a tots per existir i per compartir el vostre temps amb mi!

Finalment, no podria acabar aquests agraïments sense unes paraules molt especials per a la Mar, la meva dona, la meva companya de vida. Ella és qui ha estat al meu costat durant el dia a dia durant tots aquests anys. M'ha animat quan he estat trist, no ha perdonat ni una celebració pels bons moments i s'ha enfadat amb mi quan no he estat a l'alçada de mi mateix. Sense el seu suport, la seva alegria i la seva empenta tampoc seria aquí. Ningú ha "patit" aquesta tesi més que ella i, tot i així, ningú m'ha animat més que ella a tirar-la endavant. Moltes gràcies amor, a tu també te la dedico amb tot el cor, espero que t'hagi agradat.

Bibliography

1. Franchini, J. Ch. Alphonse Laveran, 1845-1922: His Life and Works, Read on the Fiftieth Anniversary of the Malaria Parasite. *Ann Med Hist* **3**, 280 (1931).
2. Ross, R. Observations on a Condition Necessary to the Transformation of the Malaria Crescent. *Br Med J* **1**, 251 (1897).
3. Grassi, B., Bignami, A. & Bastianelli, G. Medical Zoology: Further Researches upon the Cycle of Human Malaria in the Body of the Mosquito. *Ind Med Gaz* **34**, 104 (1899).
4. Shortt, H. E. & Garnham, P. C. C. Pre-erythrocytic Stage in Mammalian Malaria Parasites. *Nature 1948 161:4082* **161**, 126–126 (1948).
5. Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193**, 673–675 (1976).
6. Krotoski, W. A. *et al.* Demonstration of hypnozoites in sporozoite-transmitted Plasmodium vivax infection. *Am J Trop Med Hyg* **31**, 1291–1293 (1982).
7. Smith, J. D. *et al.* Switches in expression of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101–110 (1995).
8. Su, X. zhuan *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82**, 89–100 (1995).
9. Baruch, D. I. *et al.* Cloning the P. falciparum Gene Encoding PfEMPl, a Malarial Variant Antigen and Adherence Receptor on the Surface of Parasitized Human Erythrocytes. **92**, 77–97 (1995).
10. Crabb, B. S. *et al.* Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell* **89**, 287–296 (1997).
11. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
12. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–252 (2014).
13. World Health Organization. WHO Malaria World Report. (2021).
14. Ponnudurai, T. *et al.* Sporozoite load of mosquitoes infected with Plasmodium falciparum. *Trans R Soc Trop Med Hyg* **83**, 67–70 (1989).
15. Sidjanski, S. & Vanderberg, J. P. Delayed Migration of Plasmodium Sporozoites from the Mosquito Bite Site to the Blood. *Am J Trop Med Hyg* **57**, 426–429 (1997).
16. Amino, R. *et al.* Quantitative imaging of Plasmodium transmission from mosquito to mammal. *Nat Med* **12**, 220–224 (2006).
17. Sturm, A. *et al.* Manipulation of Host Hepatocytes by the Malaria Parasite for Delivery into Liver Sinusoids. *Science (1979)* **313**, 1287–1290 (2006).
18. Weiss, G. E., Crabb, B. S. & Gilson, P. R. Overlaying Molecular and Temporal Aspects of Malaria Parasite Invasion. *Trends Parasitol* **32**, 284–295 (2016).
19. Maier, A. G., Matuschewski, K., Zhang, M. & Rug, M. Plasmodium falciparum. *Trends Parasitol* **35**, 481–482 (2019).
20. Meibalan, E. & Marti, M. Biology of malaria transmission. *Cold Spring Harb Perspect Med* (2017) doi:10.1101/cshperspect.a025452.

21. Alano, P. Plasmodium falciparum gametocytes: Still many secrets of a hidden life. *Molecular Microbiology* Preprint at https://doi.org/10.1111/j.1365-2958.2007.05904.x (2007).
22. Knuttgen, H. J. The bone marrow of non-immune Europeans in acute malaria infection: a topical review. *Annals of Tropical Medicine and Parasitology* Preprint at https://doi.org/10.1080/00034983.1987.11812158 (1987).
23. Sinden, R. E. & Smalley, M. E. Gametocytogenesis of Plasmodium falciparum in vitro: The cell-cycle. *Parasitology* (1979) doi:10.1017/S003118200005335X.
24. Bousema, T., Okell, L., Felger, I. & Drakeley, C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nat Rev Microbiol* **12**, 833–840 (2014).
25. Greischar, M. A., Reece, S. E., Savill, N. J. & Mideo, N. The Challenge of Quantifying Synchrony in Malaria Parasites. *Trends Parasitol* **35**, 341–355 (2019).
26. White, N. J. Severe malaria. *Malar J* **21**, 284 (2022).
27. Hampsey, M. Molecular Genetics of the RNA Polymerase II General Transcriptional Machinery. *Microbiology and Molecular Biology Reviews* (1998) doi:10.1128/mmbr.62.2.465-503.1998.
28. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* (1981) doi:10.1016/0092-8674(81)90413-X.
29. Spitz, F. & Furlong, E. E. M. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/nrg3207 (2012).
30. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/nrg3682 (2014).
31. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012) doi:10.1038/nature11082.
32. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* (2014) doi:10.1016/j.cell.2014.11.021.
33. Vietri Rudan, M. & Hadjur, S. Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. *Trends in Genetics* Preprint at https://doi.org/10.1016/j.tig.2015.09.004 (2015).
34. Lam, M. T. Y. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* (2013) doi:10.1038/nature12209.
35. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* (2011) doi:10.1038/nsmb.2085.
36. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* (2010) doi:10.1038/nature09033.
37. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* (2014) doi:10.1038/nature12787.
38. de Santa, F. *et al.* A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* (2010) doi:10.1371/journal.pbio.1000384.

39. Cech, T. R. & Steitz, J. A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* **157**, 77–94 (2014).
40. Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311**, 532–537 (1984).
41. Lai, W. K. M. & Pugh, B. F. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/nrm.2017.47 (2017).
42. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
43. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
44. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
45. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
46. 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**, 900–905 (2004).
47. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology 2017 19:4* **19**, 229–244 (2017).
48. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* vol. 20 207–220 Preprint at https://doi.org/10.1038/s41576-018-0089-8 (2019).
49. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
50. Doskočil, J. & Šorm, F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochim Biophys Acta* **55**, 953–959 (1962).
51. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**, 8689–8694 (2010).
52. Bird, A. P. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol* **118**, 49–60 (1978).
53. Watt, F. & Molloy, P. L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* **2**, 1136–1143 (1988).
54. Ben-Hattar, J. & Jiricny, J. Methylation of single CpG dinucleotides within a promoter element of the Herpes simplex virus tk gene reduces its transcription in vivo. *Gene* **65**, 219–227 (1988).
55. Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev* **3**, 612–619 (1989).
56. Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**, 909–912 (2016).

57. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**, 335–340 (1997).
58. Ferguson-Smith, A. C., Sasaki, H., Cattanach, B. M. & Surani, M. A. Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature* **362**, 751–755 (1993).
59. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
60. Lock, L. F., Takagi, N. & Martin, G. R. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* **48**, 39–46 (1987).
61. Mohandas, T., Sparkes, R. S. & Shapiro, L. J. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**, 393–396 (1981).
62. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**, 555–567 (2013).
63. Bender, C. M. *et al.* Roles of cell division and gene transcription in the methylation of CpG islands. *Mol Cell Biol* **19**, 6690–6698 (1999).
64. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
65. Arents, G. & Moudrianakis, E. N. Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. *Proc Natl Acad Sci U S A* **90**, 10489–10493 (1993).
66. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
67. Buschbeck, M. & Hake, S. B. Variants of core histones and their roles in cell fate decisions, development and cancer. *Nat Rev Mol Cell Biol* **18**, 299–314 (2017).
68. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**, 178–189 (2015).
69. Sitbon, D., Boyarchuk, E., Dingli, F., Loew, D. & Almouzni, G. Histone variant H3.3 residue S31 is essential for Xenopus gastrulation regardless of the deposition pathway. *Nat Commun* **11**, (2020).
70. Chang, F. T. M. *et al.* CHK1-driven histone H3.3 serine 31 phosphorylation is important for chromatin maintenance and cell survival in human ALT cancer cells. *Nucleic Acids Res* **43**, 2603–2614 (2015).
71. Hake, S. B. *et al.* Serine 31 phosphorylation of histone variant H3.3 is specific to regions bordering centromeres in metaphase chromosomes. *Proc Natl Acad Sci U S A* **102**, 6344–6349 (2005).
72. Martire, S. *et al.* Phosphorylation of histone H3.3 at serine 31 promotes p300 activity and enhancer acetylation. *Nat Genet* **51**, 941–946 (2019).
73. Malik, H. S. & Henikoff, S. Major evolutionary transitions in centromere complexity. *Cell* **138**, 1067–1082 (2009).
74. Gambogi, C. W. & Black, B. E. The nucleosomes that mark centromere location on chromosomes old and new. *Essays Biochem* **63**, 15–27 (2019).

75. Bönisch, C. & Hake, S. B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res* **40**, 10719–10741 (2012).
76. Rangasamy, D., Greaves, I. & Tremethick, D. J. RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. *Nat Struct Mol Biol* **11**, 650–655 (2004).
77. Hou, H. *et al.* Histone variant H2A.Z regulates centromere silencing and chromosome segregation in fission yeast. *J Biol Chem* **285**, 1909–1918 (2010).
78. Greaves, I. K., Rangasamy, D., Ridgway, P. & Tremethick, D. J. H2A.Z contributes to the unique 3D structure of the centromere. *Proc Natl Acad Sci U S A* **104**, 525–530 (2007).
79. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**, 819–830 (2014).
80. Hu, G. *et al.* H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **12**, 180–192 (2013).
81. Ku, M. *et al.* H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biol* **13**, (2012).
82. Dryhurst, D. *et al.* Characterization of the histone H2A.Z-1 and H2A.Z-2 isoforms in vertebrates. *BMC Biol* **7**, (2009).
83. Matsuda, R. *et al.* Identification and characterization of the two isoforms of the vertebrate H2A.Z histone variant. *Nucleic Acids Res* **38**, 4263–4273 (2010).
84. Suto, R. K., Clarkson, M. J., Tremethick, D. J. & Luger, K. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nat Struct Biol* **7**, 1121–1124 (2000).
85. Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B. & Borggrefe, T. The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin* **12**, (2019).
86. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nature Reviews Genetics 2022 23:9* **23**, 563–580 (2022).
87. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
88. Dai, Z., Ramesh, V. & Locasale, J. W. The evolving metabolic landscape of chromatin biology and epigenetics. *Nat Rev Genet* **21**, 737–753 (2020).
89. Wiese, M. & Bannister, A. J. Two genomes, one cell: Mitochondrial-nuclear coordination via epigenetic pathways. *Mol Metab* **38**, (2020).
90. Hempel, K., Lange, H. W. & Birkofer, L. [Epsilon-N-trimethyllysine, a new amino acid in histones]. *Naturwissenschaften* **55**, 37–37 (1968).
91. Paik, W. K. & Kim, S. E-N-dimethyllysine in histones. *Biochem Biophys Res Commun* **27**, 479–483 (1967).
92. Murray, K. THE OCCURRENCE OF EPSILON-N-METHYL LYSINE IN HISTONES. *Biochemistry* **3**, 10–15 (1964).
93. Tsukada, Y. I. *et al.* Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439**, 811–816 (2006).

94. Whetstine, J. R. *et al.* Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell* **125**, 467–481 (2006).
95. Cloos, P. A. C. *et al.* The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. *Nature* **442**, 307–311 (2006).
96. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
97. Kim, J. *et al.* Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep* **7**, 397–403 (2006).
98. Yang, Y. *et al.* TDRD3 is an effector molecule for arginine-methylated histone marks. *Mol Cell* **40**, 1016–1023 (2010).
99. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
100. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
101. Wang, Y. *et al.* Regulation of Set9-mediated H4K20 methylation by a PWWP domain protein. *Mol Cell* **33**, 428–437 (2009).
102. Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442**, 86–90 (2006).
103. Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442**, 96–99 (2006).
104. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**, 229–244 (2018).
105. Kabi, M. & Filion, G. J. Heterochromatin: did H3K9 methylation evolve to tame transposons? *Genome Biol* **22**, 1–3 (2021).
106. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
107. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
108. Brasher, S. v. *et al.* The structure of mouse HP1 suggests a unique mode of single peptide recognition by the shadow chromo domain dimer. *EMBO J* **19**, 1587 (2000).
109. Cowieson, N. P., Partridge, J. F., Allshire, R. C. & McLaughlin, P. J. Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis. *Curr Biol* **10**, 517–525 (2000).
110. Eissenberg, J. C. *et al.* Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **87**, 9923 (1990).
111. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
112. Tschiersch, B. *et al.* The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J* **13**, 3822 (1994).

113. Daujat, S. *et al.* H3K64 trimethylation marks heterochromatin and is dynamically remodeled during developmental reprogramming. *Nat Struct Mol Biol* **16**, 777–781 (2009).
114. Schotta, G. *et al.* A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev* **18**, 1251–1262 (2004).
115. Jack, A. P. M. *et al.* H3K56me3 is a novel, conserved heterochromatic mark that largely but not completely overlaps with H3K9me3 in both regulation and localization. *PLoS One* **8**, (2013).
116. Canzio, D. *et al.* Chromodomain-mediated oligomerization of HP1 suggests a nucleosome-bridging mechanism for heterochromatin assembly. *Mol Cell* **41**, 67–81 (2011).
117. Schuettengruber, B. & Cavalli, G. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**, 3531–3542 (2009).
118. Aranda, S., Mas, G. & di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci Adv* **1**, (2015).
119. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol* **10**, 697–708 (2009).
120. Vermeulen, M. *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).
121. Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* **47**, 1149–1157 (2015).
122. Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends Genet* **27**, 389–396 (2011).
123. Benayoun, B. A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673–688 (2014).
124. Lismer, A. *et al.* Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. *Dev Cell* **56**, 671-686.e6 (2021).
125. Siklenka, K. *et al.* Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science* **350**, (2015).
126. Zhang, B. *et al.* Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553–557 (2016).
127. Dahl, J. A. *et al.* Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548–552 (2016).
128. Hughes, A. L., Kelley, J. R. & Klose, R. J. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim Biophys Acta Gene Regul Mech* **1863**, (2020).
129. Kizer, K. O. *et al.* A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Mol Cell Biol* **25**, 3305–3316 (2005).
130. Bannister, A. J. *et al.* Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem* **280**, 17732–17736 (2005).

131. Weinberg, D. N. *et al.* The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**, 281–286 (2019).
132. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
133. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123**, 581–592 (2005).
134. Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J* **7**, 1395–1402 (1988).
135. Fujisawa, T. & Filippakopoulos, P. Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nat Rev Mol Cell Biol* **18**, 246–262 (2017).
136. Morgan, M. A. J. & Shilatifard, A. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat Genet* **52**, 1271–1281 (2020).
137. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A* **51**, 786–794 (1964).
138. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**, 897–903 (2008).
139. Cluntun, A. A. *et al.* The rate of glycolysis quantitatively mediates specific histone acetylation sites. *Cancer Metab* **3**, (2015).
140. de Ruijter, A. J. M., van Gennip, A. H., Caron, H. N., Kemp, S. & van Kuilenburg, A. B. P. Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J* **370**, 737–749 (2003).
141. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).
142. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**, 897–903 (2008).
143. Schübeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* **18**, 1263–1271 (2004).
144. Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**, 3131–3141 (2009).
145. Berk, A. *et al.* Molekulare Zellbiologie. *Molekulare Zellbiologie* (1996) doi:10.1515/9783110810578/HTML.
146. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science (1979)* **147**, 1462–1465 (1965).
147. Translation of mRNA - The Cell - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK9849/.
148. Rocheleau, C. E. *et al.* Wnt signaling and an APC-related gene specify endoderm in early C. elegans embryos. *Cell* **90**, 707–716 (1997).

149. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806–811 (1998).
150. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* **2**, 986–991 (2001).
151. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* **12**, 99–110 (2011).
152. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).
153. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
154. Szittya, G. & Burgyán, J. RNA interference-mediated intrinsic antiviral immunity in plants. *Curr Top Microbiol Immunol* **371**, 153–181 (2013).
155. Nayak, A., Tassetto, M., Kunitomi, M. & Andino, R. RNA interference-mediated intrinsic antiviral immunity in invertebrates. *Curr Top Microbiol Immunol* **371**, 183–200 (2013).
156. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**, 246–258 (2011).
157. Luteijn, M. J. & Ketting, R. F. PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nat Rev Genet* **14**, 523–534 (2013).
158. Wu, H., Yang, L. & Chen, L. L. The Diversity of Long Noncoding RNAs and Their Generation. *Trends Genet* **33**, 540–552 (2017).
159. Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2* **22**, 96–118 (2020).
160. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670 (2007).
161. Bond, A. M. *et al.* Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**, 1020–1027 (2009).
162. Feng, J. *et al.* The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* **20**, 1470–1484 (2006).
163. Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**, 390–403 (2010).
164. Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**, 1831–1842 (2009).
165. Lee, J. T. & Bartolomei, M. S. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152**, 1308–1323 (2013).
166. Bartolomei, M. S., Zemel, S. & Tilghman, S. M. Parental imprinting of the mouse H19 gene. *Nature* **351**, 153–155 (1991).
167. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).

168. Yang, F. *et al.* The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol* **16**, 1–17 (2015).
169. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**, 198–206 (2014).
170. Mao, Y. S., Sunwoo, H., Zhang, B. & Spector, D. L. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat Cell Biol* **13**, 95–101 (2011).
171. Böhme, U., Otto, T. D., Sanders, M., Newbold, C. I. & Berriman, M. Progression of the canonical reference malaria parasite genome from 2002-2019. *Wellcome Open Res* **4**, 58 (2019).
172. Su, X. Z., Lane, K. D., Xia, L., Sá, J. M. & Wellems, T. E. Plasmodium Genomics and Genetics: New Insights into Malaria Pathogenesis, Drug Resistance, Epidemiology, and Evolution. *Clin Microbiol Rev* **32**, (2019).
173. Bushell, E. *et al.* Functional Profiling of a Plasmodium Genome Reveals an Abundance of Essential Genes. *Cell* **170**, 260-272.e8 (2017).
174. 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**, (2005).
175. Bischoff, E. & Vaquero, C. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of Plasmodium falciparum. *BMC Genomics* **11**, (2010).
176. Horrocks, P., Dechering, K. & Lanzer, M. Control of gene expression in Plasmodium falciparum. *Mol Biochem Parasitol* **95**, 171–181 (1998).
177. Toenhake, C. G. & Bártfai, R. What functional genomics has taught us about transcriptional regulation in malaria parasites. *Brief Funct Genomics* **18**, 290–301 (2019).
178. 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**, 2463–2475 (2016).
179. Templeton, T. J. *et al.* Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res* **14**, 1686–1695 (2004).
180. Coulson, R. M. R., Hall, N. & Ouzounis, C. A. Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res* **14**, 1548–1554 (2004).
181. Balaji, S., Madan Babu, 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).
182. 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).

183. Russell, K. *et al.* Homopolymer tract organization in the human malarial parasite Plasmodium falciparum and related Apicomplexan parasites. *BMC Genomics* **15**, (2014).
184. van Noort, V. & Huynen, M. A. Combinatorial gene regulation in Plasmodium falciparum. *Trends Genet* **22**, 73–78 (2006).
185. Santos, J. M. *et al.* Red Blood Cell Invasion by the Malaria Parasite Is Coordinated by the PfAP2-I Transcription Factor. *Cell Host Microbe* **21**, 731-741.e10 (2017).
186. le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503–1508 (2003).
187. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol* **1**, (2003).
188. Llinás, M., Bozdech, Z., Wong, E. D., Adai, A. T. & DeRisi, J. L. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res* **34**, 1166–1173 (2006).
189. Poran, A. *et al.* Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature 2017 551:7678* **551**, 95–99 (2017).
190. Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, (2018).
191. Zhang, C. *et al.* Systematic CRISPR-Cas9-Mediated Modifications of Plasmodium yoelii ApiAP2 Genes Reveal Functional Insights into Parasite Development. *mBio* **8**, (2017).
192. 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 (2022).
193. 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**, (2010).
194. Shang, X. *et al.* A cascade of transcriptional repression determines sexual commitment and development in Plasmodium falciparum. *Nucleic Acids Res* **49**, 9264–9279 (2021).
195. Singh, S. *et al.* The PfAP2-G2 transcription factor is a critical regulator of gametocyte maturation. *Mol Microbiol* **115**, 1005–1024 (2021).
196. Yuda, M., Kaneko, I., Iwanaga, S., Murata, Y. & Kato, T. Female-specific gene regulation in malaria parasites by an AP2-family transcription factor. *Mol Microbiol* **113**, 40–51 (2020).
197. 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).
198. Martins, R. M. *et al.* An ApiAP2 member regulates expression of clonally variant genes of the human malaria parasite Plasmodium falciparum. *Sci Rep* **7**, (2017).
199. Flueck, C. *et al.* A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathog* **6**, (2010).

200. Carrington, E. *et al.* The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the malaria parasite Plasmodium falciparum. *iScience* **24**, (2021).
201. Sierra-Miranda, M. *et al.* PfAP2Tel, harbouring a non-canonical DNA-binding AP2 domain, binds to Plasmodium falciparum telomeres. *Cell Microbiol* **19**, (2017).
202. Miao, J. *et al.* A unique GCN5 histone acetyltransferase complex controls erythrocyte invasion and virulence in the malaria parasite Plasmodium falciparum. *PLoS Pathog* **17**, (2021).
203. 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).
204. Iwanaga, S., Kaneko, I., Kato, T. & Yuda, M. Identification of an AP2-family protein that is critical for malaria liver stage development. *PLoS One* **7**, (2012).
205. Yuda, M. *et al.* Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol Microbiol* **71**, 1402–1414 (2009).
206. Rovira-Graells, N. *et al.* Transcriptional variation in the malaria parasite Plasmodium falciparum. *Genome Res* **22**, 925 (2012).
207. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
208. Scherf, A. *et al.* Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J* **17**, 5418–5426 (1998).
209. Chen, Q. *et al.* Developmental selection of var gene expression in Plasmodium falciparum. *Nature 1998 394:6691* **394**, 392–395 (1998).
210. Su, X. zhuan *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell* **82**, 89–100 (1995).
211. Rathjen, T., Nicol, C., McConkey, G. & Dalmay, T. Analysis of short RNAs in the malaria parasite and its red blood cell host. *FEBS Lett* **580**, 5185–5188 (2006).
212. Xue, X., Zhang, Q., Huang, Y., Feng, L. & Pan, W. No miRNA were found in Plasmodium and the ones identified in erythrocytes could not be correlated with infection. *Malar J* **7**, (2008).
213. Baum, J. *et al.* Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res* **37**, 3788–3798 (2009).
214. 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 (2015).
215. Subudhi, A. K. *et al.* Natural antisense transcripts in Plasmodium falciparum isolates from patients with complicated malaria. *Exp Parasitol* **141**, 39–54 (2014).
216. 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).
217. 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**, (2014).

218. López-Barragán, M. J. *et al.* Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. *BMC Genomics* **12**, (2011).
219. 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 (2001).
220. Frank, M. *et al.* Strict pairing of var promoters and introns is required for var gene silencing in the malaria parasite Plasmodium falciparum. *J Biol Chem* **281**, 9942–9952 (2006).
221. Gannoun-Zaki, L., Jost, A., Mu, J., Deitsch, K. W. & Wellems, T. E. A silenced Plasmodium falciparum var promoter can be activated in vivo through spontaneous deletion of a silencing element in the intron. *Eukaryot Cell* **4**, 490–492 (2005).
222. Epp, C., Li, F., Howitt, C. A., Chookajorn, T. & Deitsch, K. W. Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. *RNA* **15**, 116–127 (2009).
223. 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 (2016).
224. 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**, (2020).
225. Ruiz, J. L. *et al.* Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum. *Nucleic Acids Res* **46**, 9414–9431 (2018).
226. Broadbent, K. M. *et al.* A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol* **12**, (2011).
227. Filarsky, M. *et al.* GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* **359**, 1259–1263 (2018).
228. 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**, 9655–9660 (2009).
229. Lemieux, J. E. *et al.* Genome-wide profiling of chromosome interactions in Plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic. *Wiley Online Library* **90**, 519–537 (2013).
230. Flueck, C. *et al.* Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog* **5**, (2009).
231. Lopez-Rubio, J. J., Mancio-Silva, L. & Scherf, A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* **5**, 179–190 (2009).
232. Trelle, M. B., Salcedo-Amaya, A. M., Cohen, A. M., Stunnenberg, H. G. & Jensen, O. N. Global histone analysis by mass spectrometry reveals a high content of

acetylated lysine residues in the malaria parasite Plasmodium falciparum. *J Proteome Res* **8**, 3439–3450 (2009).
233. Miao, J. *et al.* The malaria parasite Plasmodium falciparum histones: organization, expression, and acetylation. *Gene* **369**, 53–65 (2006).
234. Bártfai, R. *et al.* H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS Pathog* **6**, (2010).
235. Sullivan, W. J. Histone H3 and H3.3 variants in the protozoan pathogens Plasmodium falciparum and Toxoplasma gondii. *DNA Seq* **14**, 227–231 (2003).
236. Hoeijmakers, W. A. M. *et al.* Plasmodium falciparum centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. *Cell Microbiol* **14**, 1391–1401 (2012).
237. 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 (2013).
238. 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**, 1061–1073 (2013).
239. Petter, M. *et al.* Expression of P. falciparum var genes involves exchange of the histone variant H2A.Z at the promoter. *PLoS Pathog* **7**, (2011).
240. Sullivan, W. J., Naguleswaran, A. & Angel, S. O. Histones and histone modifications in protozoan parasites. *Cell Microbiol* **8**, 1850–1861 (2006).
241. Kensche, P. R. *et al.* The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Res* **44**, 2110–2124 (2016).
242. 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**, 2463–2475 (2016).
243. 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**, (2009).
244. 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**, 716–724 (2011).
245. Bunnik, E. M. *et al.* DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics* **15**, (2014).
246. Ponts, N. *et al.* Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res* **20**, 228–238 (2010).
247. Ay, F. *et al.* Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* **24**, 974–988 (2014).

248. Bunnik, E. M. *et al.* DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum. *BMC Genomics* **15**, (2014).
249. Ponts, N. *et al.* Genome-wide mapping of DNA methylation in the human malaria parasite Plasmodium falciparum. *Cell Host Microbe* **14**, 696 (2013).
250. Hammam, E. *et al.* Discovery of a new predominant cytosine DNA modification that is linked to gene expression in malaria parasites. *Nucleic Acids Res* **48**, 184 (2020).
251. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
252. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
253. Saraf, A. *et al.* Dynamic and combinatorial landscape of histone modifications during the intraerythrocytic developmental cycle of the malaria parasite. *J Proteome Res* **15**, 2787–2801 (2016).
254. Coetzee, N. *et al.* Quantitative chromatin proteomics reveals a dynamic histone post-translational modification landscape that defines asexual and sexual Plasmodium falciparum parasites. *Sci Rep* **7**, (2017).
255. Gupta, A. P. *et al.* Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum. *PLoS Pathog* **9**, (2013).
256. Green, J. L. *et al.* Ubiquitin activation is essential for schizont maturation in Plasmodium falciparum blood-stage development. *PLoS Pathog* **16**, (2020).
257. Bui, H. T. N. *et al.* Mapping and functional analysis of heterochromatin protein 1 phosphorylation in the malaria parasite Plasmodium falciparum. *Sci Rep* **9**, (2019).
258. Gupta, A. P. *et al.* Histone 4 lysine 8 acetylation regulates proliferation and host-pathogen interaction in Plasmodium falciparum. *Epigenetics Chromatin* **10**, (2017).
259. Kaur, I. *et al.* Widespread occurrence of lysine methylation in Plasmodium falciparum proteins at asexual blood stages. *Sci Rep* **6**, (2016).
260. Cobbold, S. A., Santos, J. M., Ochoa, A., Perlman, D. H. & Llinas, M. Proteome-wide analysis reveals widespread lysine acetylation of major protein complexes in the malaria parasite. *Sci Rep* **6**, (2016).
261. Dastidar, E. G. *et al.* Comprehensive histone phosphorylation analysis and identification of Pf14-3-3 protein as a histone H3 phosphorylation reader in malaria parasites. *PLoS One* **8**, (2013).
262. Treeck, M., Sanders, J. L., Elias, J. E. & Boothroyd, J. C. The phosphoproteomes of Plasmodium falciparum and Toxoplasma gondii reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host Microbe* **10**, 410–419 (2011).
263. von Grüning, H. *et al.* A Dynamic and Combinatorial Histone Code Drives Malaria Parasite Asexual and Sexual Development. *Molecular and Cellular Proteomics* **21**, 100199 (2022).
264. Mancio-Silva, L., Lopez-Rubio, J. J., Claes, A. & Scherf, A. Sir2a regulates rDNA transcription and multiplication rate in the human malaria parasite Plasmodium falciparum. *Nat Commun* **4**, (2013).

265. Tang, J. *et al.* Histone modifications associated with gene expression and genome accessibility are dynamically enriched at Plasmodium falciparum regulatory sequences. *Epigenetics Chromatin* **13**, (2020).
266. Karmodiya, K. *et al.* A comprehensive epigenome map of Plasmodium falciparum reveals unique mechanisms of transcriptional regulation and identifies H3K36me2 as a global mark of gene suppression. *Epigenetics Chromatin* **8**, (2015).
267. Cui, L. *et al.* Histone acetyltransferase inhibitor anacardic acid causes changes in global gene expression during in vitro Plasmodium falciparum development. *Eukaryot Cell* **7**, 1200–1210 (2008).
268. Coleman, B. I. *et al.* A Plasmodium falciparum Histone Deacetylase Regulates Antigenic Variation and Gametocyte Conversion. *Cell Host Microbe* **16**, 177 (2014).
269. Hoeijmakers, W. A. M. *et al.* Epigenetic reader complexes of the human malaria parasite, Plasmodium falciparum. *Nucleic Acids Res* **47**, 11574 (2019).
270. Fan, Q., An, L. & Cui, L. PfADA2, a Plasmodium falciparum homologue of the transcriptional coactivator ADA2 and its in vivo association with the histone acetyltransferase PfGCN5. *Gene* **336**, 251–261 (2004).
271. 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).
272. 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 (2007).
273. Fraschka, S. A. *et al.* Comparative Heterochromatin Profiling Reveals Conserved and Unique Epigenome Signatures Linked to Adaptation and Development of Malaria Parasites. *Cell Host Microbe* **23**, 407-420.e8 (2018).
274. Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* **499**, 223–227 (2013).
275. 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**, (2019).
276. Cui, L., Fan, Q., Cui, L. & Miao, J. Histone lysine methyltransferases and demethylases in Plasmodium falciparum. *Int J Parasitol* **38**, 1083–1097 (2008).
277. 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).
278. Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* **499**, 223–227 (2013).
279. Chen, P. B. *et al.* Plasmodium falciparum PfSET7: enzymatic characterization and cellular localization of a novel protein methyltransferase in sporozoite, liver and erythrocytic stage parasites. *Sci Rep* **6**, (2016).
280. Gomes, A. R. *et al.* A genome-scale vector resource enables high-throughput reverse genetic screening in a malaria parasite. *Cell Host Microbe* **17**, 404–413 (2015).

281. Sanderson, T. & Rayner, J. C. PhenoPlasm: a database of disruption phenotypes for malaria parasite genes. *Wellcome Open Res* **2**, (2017).
282. 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**, (2014).
283. Ngwa, C. J., Gross, M. R., Musabyimana, J.-P., Pradel, G. & Deitsch, K. W. The Role of the Histone Methyltransferase PfSET10 in Antigenic Variation by Malaria Parasites: a Cautionary Tale. *mSphere* **6**, (2021).
284. Chen, P. B. *et al.* Plasmodium falciparum PfSET7: enzymatic characterization and cellular localization of a novel protein methyltransferase in sporozoite, liver and erythrocytic stage parasites. *Scientific Reports 2016 6:1* **6**, 1–14 (2016).
285. Matthews, K. A. *et al.* Disruption of the plasmodium falciparum life cycle through transcriptional reprogramming by inhibitors of jumonji demethylases. *ACS Infect Dis* **6**, 1058–1075 (2021).
286. Cui, L. *et al.* PfGCN5-mediated histone H3 acetylation plays a key role in gene expression in Plasmodium falciparum. *Eukaryot Cell* **6**, 1219–1227 (2007).
287. Chaal, B. K., Gupta, A. P., Wastuwidyaningtyas, B. D., Luah, Y. H. & Bozdech, Z. Histone deacetylases play a major role in the transcriptional regulation of the Plasmodium falciparum life cycle. *PLoS Pathog* **6**, (2010).
288. Bhowmick, K. *et al.* Plasmodium falciparum GCN5 acetyltransferase follows a novel proteolytic processing pathway that is essential for its function. *J Cell Sci* **133**, (2020).
289. Grant, P. A. *et al.* Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. *Genes Dev* **11**, 1640–1650 (1997).
290. Baptista, T. *et al.* SAGA Is a General Cofactor for RNA Polymerase II Transcription. *Mol Cell* **68**, 130-143.e5 (2017).
291. Bonnet, J. *et al.* The SAGA coactivator complex acts on the whole transcribed genome and is required for RNA polymerase II transcription. *Genes Dev* **28**, 1999–2012 (2014).
292. Tonkin, C. J. *et al.* Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum. *PLoS Biol* **7**, 0771–0788 (2009).
293. Freitas, 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).
294. Deitsch, K. W. Malaria virulence genes controlling expression through chromatin modification. *Cell* **121**, 1–2 (2005).
295. Merrick, C. J. *et al.* Functional Analysis of Sirtuin Genes in Multiple Plasmodium falciparum Strains. *PLoS One* **10**, (2015).
296. Cui, L. & Miao, J. Chromatin-Mediated epigenetic regulation in the malaria parasite Plasmodium falciparum. *Eukaryot Cell* **9**, 1138–1149 (2010).
297. 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 (2009).

298. Brancucci, N. M. B. *et al.* Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host Microbe* **16**, 165–176 (2014).
299. Josling, G. A. *et al.* A Plasmodium Falciparum Bromodomain Protein Regulates Invasion Gene Expression. *Cell Host Microbe* **17**, 741–751 (2015).
300. Quinn, J. E. *et al.* The Putative Bromodomain Protein PfBDP7 of the Human Malaria Parasite Plasmodium Falciparum Cooperates With PfBDP1 in the Silencing of Variant Surface Antigen Expression. *Front Cell Dev Biol* **10**, (2022).
301. Musselman, C. A., Lalonde, M. E., Côté, J. & Kutateladze, T. G. Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol* **19**, 1218 (2012).
302. Chookajorn, T. *et al.* Epigenetic memory at malaria virulence genes. *Proc Natl Acad Sci U S A* **104**, 899–902 (2007).
303. Fraschka, S. A. *et al.* Comparative Heterochromatin Profiling Reveals Conserved and Unique Epigenome Signatures Linked to Adaptation and Development of Malaria Parasites. *Cell Host Microbe* **23**, 407 (2018).
304. Bachmann, A. *et al.* Mosquito Passage Dramatically Changes var Gene Expression in Controlled Human Plasmodium falciparum Infections. *PLoS Pathog* **12**, (2016).
305. Mira-Martínez, S. *et al.* Expression of the Plasmodium falciparum Clonally Variant clag3 Genes in Human Infections. *J Infect Dis* **215**, 938–945 (2017).
306. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415**, 673–679 (2002).
307. Montgomery, J. *et al.* Differential var gene expression in the organs of patients dying of falciparum malaria. *Mol Microbiol* **65**, 959–967 (2007).
308. Scherf, A., Juan Lopez-Rubio, J. & Riviere, L. L. Antigenic Variation in Plasmodium falciparum. doi:10.1146/annurev.micro.61.080706.093134.
309. Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* **407**, 1018–1022 (2000).
310. Ralph, S. A., Scheidig-Benatar, C. & Scherf, A. Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. *Proc Natl Acad Sci U S A* **102**, 5414–5419 (2005).
311. Duraisingh, M. T. *et al.* Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. *Cell* **121**, 13–24 (2005).
312. 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–E991 (2015).
313. Goel, S. *et al.* RIFINs are adhesins implicated in severe Plasmodium falciparum malaria. *Nat Med* **21**, 314–321 (2015).
314. Mwakalinga, S. B. *et al.* Expression of a type B RIFIN in Plasmodium falciparum merozoites and gametes. *Malar J* **11**, 429 (2012).
315. Fernandez, V., Hommel, M., Chen, Q., Hagblom, P. & Wahlgren, M. Small, Clonally Variant Antigens Expressed on the Surface of the Plasmodium falciparum–Infected Erythrocyte Are Encoded by the rif Gene Family and Are the Target of Human Immune Responses. *J Exp Med* **190**, 1393 (1999).

316. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum. *Proc Natl Acad Sci U S A* **96**, 9333–9338 (1999).
317. Chookajorn, T. *et al.* Epigenetic memory at malaria virulence genes. *Proc Natl Acad Sci U S A* **104**, 899–902 (2007).
318. Niang, M. *et al.* STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting. *Cell Host Microbe* **16**, 81–93 (2014).
319. Sanyal, S. *et al.* Plasmodium falciparum STEVOR proteins impact erythrocyte mechanical properties. *Blood* **119**, (2012).
320. Lavazec, C., Sanyal, S. & Templeton, T. J. Expression switching in the stevor and Pfmc-2TM superfamilies in Plasmodium falciparum. *Mol Microbiol* **64**, 1621–1634 (2007).
321. Sargeant, T. J. *et al.* Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol* **7**, (2006).
322. Proellocks, N. I. *et al.* A lysine-rich membrane-associated PHISTb protein involved in alteration of the cytoadhesive properties of Plasmodium falciparum-infected red blood cells. *FASEB J* **28**, 3103–3113 (2014).
323. Oberli, A. *et al.* A Plasmodium falciparum PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface. *FASEB J* **28**, 4420–4433 (2014).
324. Sam-Yellowe, T. Y. *et al.* A Plasmodium gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. *Genome Res* **14**, 1052–1059 (2004).
325. Bethke, L. L. *et al.* Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of Plasmodium falciparum. *Mol Biochem Parasitol* **150**, 10–24 (2006).
326. Kats, L. M. *et al.* An exported kinase (FIKK4.2) that mediates virulence-associated changes in Plasmodium falciparum-infected red blood cells. *Int J Parasitol* **44**, 319–328 (2014).
327. Nunes, M. C., Goldring, J. P. D., Doerig, C. & Scherf, A. A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell. *Mol Microbiol* **63**, 391–403 (2007).
328. Winter, G. *et al.* SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *J Exp Med* **201**, 1853–1863 (2005).
329. Mphande, F. A. *et al.* SURFIN4.1, a schizont-merozoite associated protein in the SURFIN family of Plasmodium falciparum. *Malar J* **7**, (2008).
330. Cortés, A. *et al.* Epigenetic silencing of Plasmodium falciparum genes linked to erythrocyte invasion. *PLoS Pathog* **3**, 1023–1035 (2007).
331. Crowley, V. M., Rovira-Graells, N., de Pouplana, L. R. & 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).

332. Comeaux, C. A., Coleman, B. I., Bei, A. K., Whitehurst, N. & Duraisingh, M. T. Functional analysis of epigenetic regulation of tandem RhopH1/clag genes reveals a role in Plasmodium falciparum growth. *Mol Microbiol* **80**, 378–390 (2011).
333. Ishino, T., Chinzei, Y. & Yuda, M. Two proteins with 6-cys motifs are required for malarial parasites to commit to infection of the hepatocyte. *Mol Microbiol* **58**, 1264–1275 (2005).
334. Hocking, S. E. *et al.* Highly Variable Expression of Merozoite Surface Protein MSPDBL2 in Diverse Plasmodium falciparum Clinical Isolates and Transcriptome Scans for Correlating Genes. *mBio* **13**, (2022).
335. Cortés, A. & Deitsch, K. W. Malaria Epigenetics. *Cold Spring Harb Perspect Med* **7**, 1–23 (2017).
336. Chan, J. A., Fowkes, F. J. I. & Beeson, J. G. Surface antigens of Plasmodium falciparum-infected erythrocytes as immune targets and malaria vaccine candidates. *Cell Mol Life Sci* **71**, 3633–3657 (2014).
337. Wright, G. J. & Rayner, J. C. Plasmodium falciparum Erythrocyte Invasion: Combining Function with Immune Evasion. *PLoS Pathog* **10**, (2014).
338. Stubbs, J. *et al.* Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes. *Science* **309**, 1384–1387 (2005).
339. Maier, A. G. *et al.* Plasmodium falciparum erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med* **9**, 87–92 (2003).
340. Duraisingh, M. T. *et al.* Phenotypic variation of Plasmodium falciparum merozoite proteins directs receptor targeting for invasion of human erythrocytes. *EMBO J* **22**, 1047–1057 (2003).
341. Reed, M. B. *et al.* Targeted disruption of an erythrocyte binding antigen in Plasmodium falciparum is associated with a switch toward a sialic acid-independent pathway of invasion. *Proc Natl Acad Sci U S A* **97**, 7509–7514 (2000).
342. Amambua-Ngwa, A. *et al.* Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet* **8**, (2012).
343. Desai, S. A. Why do malaria parasites increase host erythrocyte permeability? *Trends Parasitol* **30**, 151–159 (2014).
344. Nguitragool, W. *et al.* Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. *Cell* **145**, 665–677 (2011).
345. Rovira-Graells, N. *et al.* Deciphering the principles that govern mutually exclusive expression of Plasmodium falciparum clag3 genes. *Nucleic Acids Res* **43**, 8243 (2015).
346. Sharma, P. *et al.* An epigenetic antimalarial resistance mechanism involving parasite genes linked to nutrient uptake. *J Biol Chem* **288**, 19429–19440 (2013).
347. Mira-Martínez, S. *et al.* Epigenetic switches in clag3 genes mediate blasticidin S resistance in malaria parasites. *Cell Microbiol* **15**, 1913–1923 (2013).
348. Sinha, A. *et al.* A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. *Nature* **507**, 253–257 (2014).

349. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in human malaria parasites. *Nature* **507**, 248 (2014).
350. Llorà-Batlle, O. *et al.* Conditional expression of PfAP2-G for controlled massive sexual conversion in Plasmodium falciparum. *Sci Adv* **6**, (2020).
351. Bancells, C. *et al.* Revisiting the initial steps of sexual development in the malaria parasite Plasmodium falciparum. *Nat Microbiol* **4**, 144–154 (2019).
352. Josling, G. A. *et al.* Dissecting the role of PfAP2-G in malaria gametocytogenesis. *Nat Commun* **11**, (2020).
353. 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**, (2015).
354. Brancucci, N. M. B. *et al.* Lysophosphatidylcholine Regulates Sexual Stage Differentiation in the Human Malaria Parasite Plasmodium falciparum. *Cell* **171**, 1532-1544.e15 (2017).
355. Carter, R. & Miller, L. H. Evidence for environmental modulation of gametocytogenesis in Plasmodium falciparum in continuous culture. *Bull World Health Organ* **57 Suppl 1**, 37–52 (1979).
356. Portugaliza, H. P. *et al.* Artemisinin exposure at the ring or trophozoite stage impacts Plasmodium falciparum sexual conversion differently. *Elife* **9**, 1–22 (2020).
357. Leas: Proceedings of the December 4-6, 1962, Fall... - Google Académico. https://scholar.google.com/scholar_lookup?title=Proceedings of the December 4-6%2C 1962%2C Fall Joint Computer Conference&author=MO Dayhoff&author=RS. Ledley&publication_year=1962&book=Proceedings of the December 4-6%2C 1962%2C Fall Joint Computer Confere.
358. Dayhoff, M. Atlas of protein sequence and structure. (1972).
359. Needleman, S., biology, C. W.-J. of molecular & 1970, undefined. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Elsevier*.
360. Feng, D., evolution, R. D.-J. of molecular & 1987, undefined. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Springer*.
361. Sievers, F., methods, D. H.-M. sequence alignment & 2014, undefined. Clustal Omega, accurate alignment of very large numbers of sequences. *Springer*.
362. biology, F. C.-J. of molecular & 1968, undefined. The origin of the genetic code. *Elsevier* **38**, 367–379 (1968).
363. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).
364. Kleppe, K., Ohtsuka, E., Kleppe, R., … I. M.-J. of molecular & 1971, undefined. Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases. *Elsevier*.
365. research, R. S.-N. acids & 1979, undefined. A strategy of DNA sequencing employing computer programs. *academic.oup.com* **6**, (1979).
366. Rice, P., Longden, I., genetics, A. B.-T. in & 2000, undefined. EMBOSS: the European molecular biology open software suite. *academia.edu*.
367. Scharf, M. *et al.* GeneQuiz: a workbench for sequence analysis. *aaai.org* (1994).

368. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *genome.cshlp.org* doi:10.1101/gr.361602.
369. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
370. Wooley, J. & Ye, Y. Computational Methods for Protein Structure Prediction & Modeling V1-Xu Xu and Liang. *studfile.net*.
371. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).
372. Leinonen, R., Sugawara, H., … M. S.-N. acids & 2010, undefined. The sequence read archive. *academic.oup.com*.
373. Leinonen, R., Akhtar, R., Birney, E., … L. B.-N. acids & 2010, undefined. The European nucleotide archive. *academic.oup.com*.
374. Field, D. *et al.* Genomic standards consortium projects. *Stand Genomic Sci* **9**, 599–601 (2015).
375. Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549–2557 (2001).
376. Yang, I. *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Springer*.
377. Yang, Y. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *academic.oup.com*.
378. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* **74**, 829–836 (1979).
379. Schmid, R. *et al.* Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics* **11**, (2010).
380. Wang, D. *et al.* Extensive increase of microarray signals in cancers calls for novel normalization assumptions. *Comput Biol Chem* **35**, 126–130 (2011).
381. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 2003 34:3* **34**, 267–273 (2003).
382. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).
383. Consortium, T. G. O. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25 (2000).
384. Carbon, S. *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2021).
385. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4* **9**, 357–359 (2012).
386. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/arxiv.1303.3997.
387. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

388. Zhang, Y., Park, C., Bennett, C., Thornton, M. & Kim, D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res* **31**, 1290–1295 (2021).
389. Musich, R., Cadle-Davidson, L. & Osier, M. v. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Front Plant Sci* **12**, 692 (2021).
390. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology 2011 29:1* **29**, 24–26 (2011).
391. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res* **12**, 996–1006 (2002).
392. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
393. Jeon, H., Lee, H., Kang, B., Jang, I. & Roh, T. Y. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. *Genomics Inform* **18**, 1–9 (2020).
394. 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).
395. Xu, S., Grullon, S., Ge, K. & Peng, W. Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells. *Methods Mol Biol* **1150**, 97 (2014).
396. Shao, Z., Zhang, Y., Yuan, G. C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol* **13**, R16 (2012).
397. Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform* **17**, 953–966 (2016).
398. Casas-Vila, N., Pickford, A. K., Portugaliza, H. P., Tintó-Font, E. & Cortés, A. Transcriptional Analysis of Tightly Synchronized Plasmodium falciparum Intraerythrocytic Stages by RT-qPCR. *Methods Mol Biol* **2369**, 165–185 (2021).
399. 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 (2009).
400. Duraisingh, M. T. & Skillman, K. M. Epigenetic Variation and Regulation in Malaria Parasites. *https://doi-org.sire.ub.edu/10.1146/annurev-micro-090817-062722* **72**, 355–375 (2018).
401. le Roch, K. G. *et al.* A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, Plasmodium falciparum. *BMC Genomics* **9**, (2008).
402. Deitsch, K., … M. D.-A. J. & 2007, undefined. Mechanisms of gene regulation in Plasmodium. *pennstate.pure.elsevier.com* (2006).
403. Ganesan, K. *et al.* A genetically hard-wired metabolic transcriptome in Plasmodium falciparum fails to mount protective responses to lethal antifolates. *PLoS Pathog* **4**, (2008).
404. Mancio-Silva, L. *et al.* Nutrient sensing modulates malaria parasite virulence. *Nature* **547**, 213–216 (2017).

405. Harris, C. T. *et al.* Metabolic competition between lipid metabolism and histone methylation regulates sexual differentiation in human malaria parasites. *bioRxiv* 2022.01.18.476397 (2022) doi:10.1101/2022.01.18.476397.
406. Cabral, F. J., Fotoran, W. L. & Wunderlich, G. Dynamic Activation and Repression of the Plasmodium falciparum rif Gene Family and Their Relation to Chromatin Modification. *PLoS One* **7**, e29881 (2012).
407. Jiang, L. *et al.* Epigenetic control of the variable expression of a Plasmodium falciparum receptor protein for erythrocyte invasion. *Proc Natl Acad Sci U S A* **107**, 2224–2229 (2010).
408. Xiao, B. *et al.* Epigenetic editing by CRISPR/dCas9 in Plasmodium falciparum. *Proc Natl Acad Sci U S A* **116**, 255–260 (2019).
409. Liang, X. *et al.* A Leak-Free Inducible CRISPRi/a System for Gene Functional Studies in Plasmodium falciparum. *Microbiol Spectr* **10**, (2022).
410. Chen, D. & Lei, E. P. Function and regulation of chromatin insulators in dynamic genome organization. *Curr Opin Cell Biol* **58**, 61 (2019).
411. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51 (2017).
412. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).
413. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**, 3573 (2017).
414. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics 2006 7:9* **7**, 703–713 (2006).
415. McCulloch, R. & Navarro, M. The protozoan nucleus. *Mol Biochem Parasitol* **209**, 76–87 (2016).
416. Avraham, I., Schreier, J. & Dzikowski, R. Insulator-like pairing elements regulate silencing and mutually exclusive expression in the malaria parasite Plasmodium falciparum. doi:10.1073/pnas.1214572109.
417. Tarr, S. J. *et al.* Schizont transcriptome variation among clinical isolates and laboratory-adapted clones of the malaria parasite Plasmodium falciparum. *BMC Genomics* **19**, (2018).
418. Mackinnon, M. J. *et al.* Comparative Transcriptional and Genomic Analysis of Plasmodium falciparum Field Isolates. *PLoS Pathog* **5**, 1000644 (2009).
419. Roberds, A., Ferraro, E., Luckhart, S. & Stewart, V. A. HIV-1 Impact on Malaria Transmission: A Complex and Relevant Global Health Concern. *Front Cell Infect Microbiol* **11**, 266 (2021).
420. Svoboda, P. Key Mechanistic Principles and Considerations Concerning RNA Interference. *Front Plant Sci* **11**, 1237 (2020).
421. Connacher, J., von Grüning, H. & Birkholtz, L. Histone Modification Landscapes as a Roadmap for Malaria Parasite Development. *Front Cell Dev Biol* **10**, (2022).
422. Giner-Lamia, J., Hernández-Prieto, M. & E. Futschik, M. ChIP-seq Experiment and Data Analysis in the Cyanobacterium Synechocystis sp. PCC 6803. *Bio Protoc* **8**, (2018).

Abbreviations

ADP	Adenosine Diphosphate
AP2	Apetala 2
BLAST	Basic Local Alignment Search Tool
ChIP-Seq	Chromatin Immunoprecipitation and sequencing
CVG	Clonally Variant Gene
DBBJ	DNA Data Bank of Japan
DHA	Dihydroartemisinin
DNA	Deoxyribonucleic Acid
EMBL	European Molecular Biology Laboratoy
EMBOSS	European Molecular Biology Open Software Suite
ENA	European Nucleotide Archive
FC	Fold-change
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HS	Heat Shock
IDC	Intraerythrocytic Development Cycle
IP	Immunoprecipitation
MS	Mass Spectrometry
MSA	Multiple Sequence Alignment
NAT	Naturally Antisense Transcripts
NCBI	National Center for Biotechnology Information
NCC	Next-cycle Conversion

NDR	Nucleosome Depleted Region
NGS	Next-generation Sequencing
PAM	Point Accepted Mutation
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
Pf	*Plasmodium falciparum*
PIC	Pre-Initiation Complex
Pol II	Polymerase II
PSAC	Plasmodial Surface Anion Channel
PTM	Post Translational Modification
QN	Quantile Normalization
RBC	Red Blood Cell
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
RPKM	Reads per Kilobase Million Reads
SAM	S-Adenosyl Methionine
SCC	Same-cycle Conversion
SRA	Sequence Read Archive
TAF	TATA binding protein Associated Factors
TAREs	Telomere-associated Repetitive Elements
TBP	TATA Binding Protein
TF	Transcription Factor
TSS	Transcription Start Site
WGS	Whole Genome Sequencing
WHO	World Health Organization