EVOLUTION OF TRANSCRIPTIONAL REGULATION AND EPIGENETIC MACHINERY IN MALARIA PARASITES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sandeep P. Kishore

[January 2012]

© 2012 Sandeep P. Kishore

EVOLUTION OF TRANSCRIPTIONAL REGULATION AND EPIGENETIC MACHINERY IN MALARIA PARASITES

Sandeep P. Kishore,

Cornell University 2012

The acquisition of additional transcriptional regulatory units and epigenetic machinery facilitated the transition of ancestral apicomplexans to parasitic life cycles. The C-terminal domain (CTD) of the enzyme RNA polymerase II is responsible for integrating the diverse events of gene expression in eukaryotes and is indispensable for life in yeast, fruit flies and mice. The CTD is comprised of tandemly repeated Y_1-S_2-P_3-T_4-S_5-P_6-S_7 amino acid heptads that are highly conserved across evolutionary lineages, with all mammalian polymerases featuring 52 identical heptad repeats. We show that malaria parasites display an unprecedented plasticity within the length and composition of their CTDs. The CTD in malaria parasites that infect human and non-human primates has expanded compared to closely related parasite species that infect rodents or birds. Based on a recent report suggesting rodent parasites lack epigenetic memory at virulence genes, we hypothesized that the expanded CTD in primate parasites permitted the binding of novel primate parasite-specific transcription factors to facilitate epigenetic memory. A comparative screen of three primate and three rodent parasite genomes indicated the presence a complementary pair of histone modifiers specific to primate parasites (PfSet2, a dominant H3K36 methylase and

PfJmj1, a H3K36 demethylase). Consistent with our hypothesis, in higher eukaryotes Set2 orthologs are known to bind phosphorylated CTD, deposit chromatin marks only in the process of active transcription by RNA polymerase II and bind non-coding RNA, suggesting a primary role in the maintenance of epigenetic memory. We provide with the first conclusive identification of the H3K36-triMe modification in *P. falciparum* across the parasite life cycle. The evolutionary history of PfSet2 and PfJmjC1, and a H4K20 methyltransferase (PfSet8) proteins and their relation to the expansion of the RNA polymerase II CTD in the genus Plasmodium is discussed. We provide the first conclusive evidence of a horizontal gene transfer of Set8 from animal hosts to the ancestral apicomplexan. This work should provide new insights on the nexus of evolutionary history of transcription units in basal eukaryotes.

BIOGRAPHICAL SKETCH

Sandeep P. Kishore is enrolled in the Weill Cornell/ Rockefeller/ Sloan-Kettering Institute Tri-Institutional MD-PhD Program and completed his B.S. at Duke University (2004) and his M.Sc at Oxford University (2006). He is the 2012 Raymond Sarber Graduate Laureate of the American Society of Microbiology.

His scientific research concerns characterizing gene activation in the parasite responsible for the most severe form of human malaria. He has founded and directed the first global health curriculum at Cornell University's medical college and has helped integrate neglected disease issues into current medical school curricula across North America. He intends a career as a public health advocate, integrating insights from clinical medicine, basic science and policy.

For Mom, Dad, Sanju and Tata

ACKNOWLEDGMENTS

I am greatly indebted to Dr. Kirk W. Deitsch for his kind scientific and creative supervision and his thoughtfulness (and patience) in my development. I was granted the twin boons of flexibility and superb mentorship during my doctoral work to develop the ideas described in this work. Exposure to thought leaders in evolutionary microbiology in Tom Templeton at Weill Cornell, Susan Perkins at the American Museum of Natural History and John Stiller at East Carolina University has proven pivotal in my study of apicomplexan parasitism. I am grateful to the insights from my roommate and now collaborator Chethan Pandarinath on developing this work. I am also thankful to Dr. Mark Ptashne and Dr. Ron Dzikowski for their perspectives on gene regulation that influenced my own deliberations on the topic. This work was greatly advanced through the kind advice and counsel of my thesis committee including Dr. George Cross, of The Rockefeller University, Dr. Neal Lue and Dr. Tom Templeton. I am grateful for their counsel on both my research and extended career development. I am also indebted to mentors that nurtured my interest in global public health: Drs. Carl Nathan, Warren Johnson, Daniel Fitzgerald, Marcus Reidenberg, Madelon Finkel and Oliver Fein.

Finally, I thank my loving parents, Dr. Anand and Lakshmi Kishore and my brother Sanjay Kishore for continuous, critical support of my development and training. This would not have been possible with them. I would like thank the Tri-Institutional MD-PhD program, Dr. Olaf Andersen and the Paul and Daisy Soros Fellowship for New Americans for sponsoring and enabling my training.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ATCC - American Type Culture Collection
CAP - catabolite activating protein
cDNA - complementary DNA
CTD - carboxy-terminal domain
DNA - deoxyribonucleic acid
DOT - disrupter of telomeric silencing
ELISA - enzyme-linked immunosorbent assay
H3K36 - histone 3 lysine 36
H3K4-diMe - Histone 3 Lysine 4 di-methyl
H3K9-triMe - Histone 3 Lysine 9 tri-methyl
HDM - histone demethylases
HKMT - histone methyltransferase
HP1 - heterochromatin protein 1
IGT - intracellular gene transfer
JMJC – Jumonji-C domain
JMJN – Jumonji-N domain
LSD - lysine-specific demethylases
MACPF - membrane attack complex and perforin domain
Me - methyl
MRA/MR4 - Malaria Research and Reference Reagent Resource Center
mRNA - messenger RNA
NCBI - National Center for Biotechnology Information
NIH - National Institutes of Health
NYU - New York University
PCR - polymerase chain reaction
PHD - plant homeodomain
Q-RT-PCR - quantitative real-time polymerase chain reaction
RBC - red-blood cell
RNA - ribonucleic acid
SET = Su = Supressing (var)) and commonly found in *Su*(6*ar*)3 -9 in the gene *Enhancer of zeste* (*E*[*z*]) and in the activating gene *trithorax*
SNP - single nucleotide polymorphism
SRI - Set2-Rpb1 interaction
TFIID - Transcription Factor II D
tRNA - transfer RNA

CHAPTER 1:

INTRODUCTION

1. The natural history of malaria parasites and related protists

Microbial eukaryotes (protists), defined as eukaryotes that are neither plants, animals nor fungi, represent 1.5 billion years of evolution and are represented within the near-entirety of eukaryotic taxa (Katz 2006). The explosion of eukaryotic diversity in the Paleozoic era (543 to 248 million years ago) led to key evolutionary innovations within microbial taxa and the independent origins of plants, animals and fungi that would co-evolve with microbes. This, then, provides for this work a focus on protists that are responsible for a significant burden on human health and economic productivity in the developing world: malaria parasites. I study the specific evolution of gene activation machinery and the seemingly more complex machinery of epigenetic modifiers in the genus *Plasmodium* and discuss the co-evolution of this process with animal malaria hosts.

Plasmodium falciparum and other members of the genus Plasmodium are part of a group of protists in the phylum Apicomplexa. These protists are defined by an apical complex (used for invasion of host cells) and feature an apicoplast which is a remnant chloroplast-like organelle. The >5 000 species described thus far include other obligate intracellular parasites such as *Cryptosporidium parvum*, *Theileria annulata*, *Toxoplasma gondii* and *Babesia bovis* (Cavalier-Smith 1993). Of all the Apicomplexa, malaria parasites have attracted much biomedical attention because of their serve effects on health and economic development. The most lethal of human

malaria parasites is *Plasmodium falciparum* and it is the principal topic of this thesis. This parasite has co-evolved with humans for millennia and displays sophisticated patterns of gene expression that enable it to multiply in several distinct morphological stages in remarkably different environments, including clonal replication in the liver and circulating red blood cells (RBCs) of its mammalian host as well as both sexual and asexual replication within its mosquito vector.

Historically, efforts to generate highly effective (>80%) vaccines have proven difficult (Hill 2006; Plowe et al. 2009) and the current most widely used group of medicines, the artemisinin-combination therapies, are facing drug resistance in diverse geographical regions, including Southeast Asia (Dondorp et al. 2010) and West Africa (Noedl et al. 2008). Insights into the evolution of parasitism by *P. falciparum* and related protozoa, including other members of the *Plasmodium* (ancestral apicomplexans), could shed new light on the complex nature of host: parasite interactions and offer new clues to the basic biology of parasitism. This approach can provide both a fundamental understanding of malaria pathogenesis and new leads in research aimed at disrupting it.

A global picture of the tree of eukaryotic life (Figure 1.1) shows where the unicellular protists of the Apicomplexa reside – with an evolutionary origin of approximately 423 million years ago coinciding with the radiation of nematodes, arthropods and marine-faring cnidarians (e.g. sea urchins), long before the appearance of mammals (Okamato 2008).

Apicomplexa are part of the super phylum Alveolata and the Kingdom Chromoalveolata which also includes the ciliates, dinoflagellates (e.g. *Perkinsus*

marinus) and colpodellids (Figure 1.1). Chromoalveolata are characterized by a secondary endosymbiotic event in which an ancestor engulfed an algae (red or green) that itself harbored a cyanobacterium (chloroplast) that has now become known as the plastid (see discussion of horizontal gene transfer). This endosymbiotic event made these organisms capable of autotrophy via photosynthesis, although this capability has been lost in some lineages of the Alveolates (e.g. ciliates). It is believed that members of the Chromoalveolata served as symbionts with ancestral animals (Okamato 2008). The relationship between some modern invertebrates (corals) and the dinoflagellate zooxanthellae has been cited as the type of relationship ancient apicomplexa shared with their hosts; a similar relationship with the recently discovered *Chromera velia* and extant corals is consistent with this (Moore et al. 2008).

However, the details, timing and context of a transition from free-living protists or mutual symbionts to organisms that featured an obligate intracellular lifestyle complete with invasion, host immune evasion and metabolic dependence pathways, however, remains less clear. The modern use of immune evasion by antigenic variation and known invasion pathways, well studied in the lethal parasite *P. falciparum*, help provide an organizing framework for research on this question.

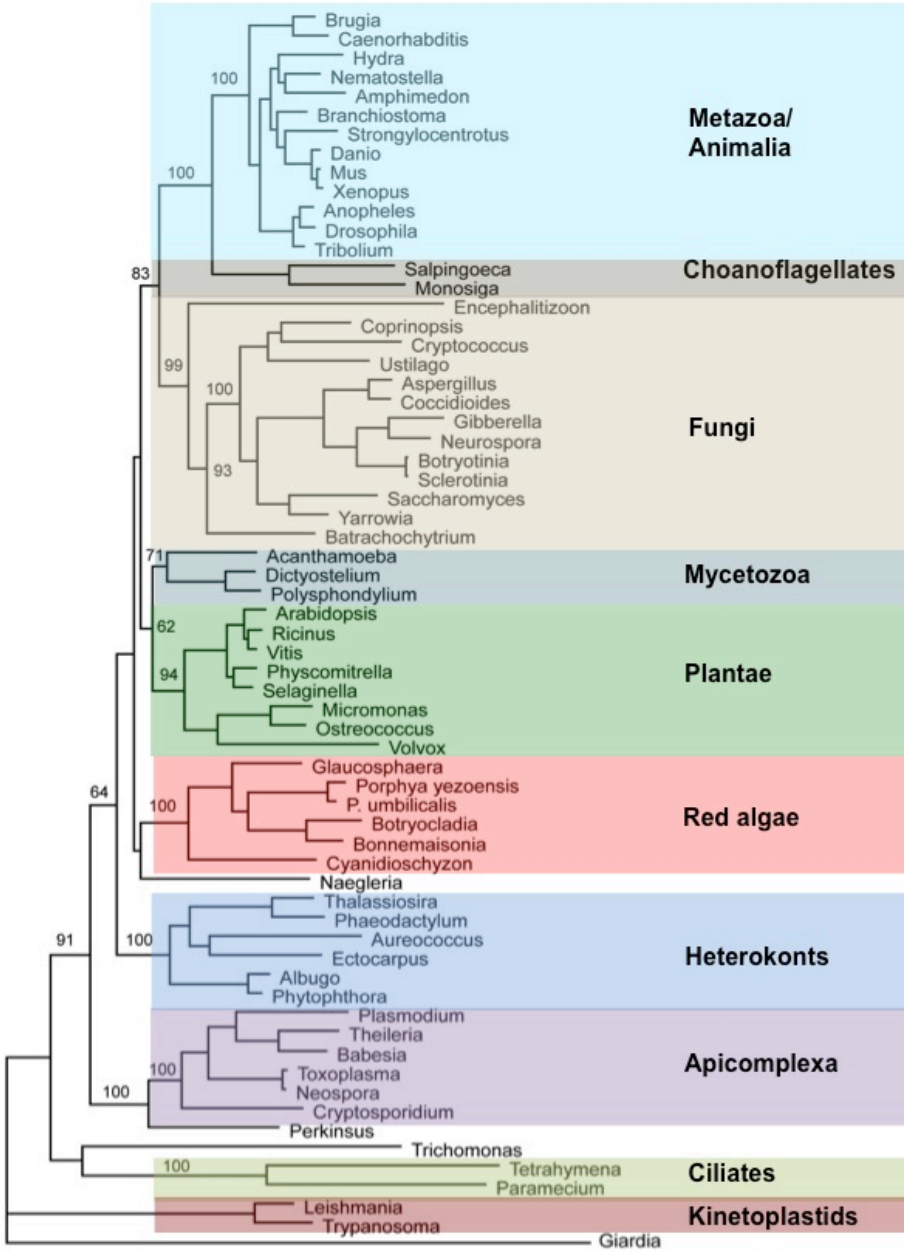

Figure 1.1. Tree of eukaryotic life based on the largest subunit of RNA Polymerase II (RBP1)

Metazoa (in light blue at top) are very distantly related to Apicomplexa (in purple) at the base of the eukaryotic tree of life. Organisms from selected taxa across the tree of life are shown. The tree, based on maximum likelihood, was constructed by John Stiller of East Carolina University (details of construction are shown in Chapter 4, Methods).

2. Malaria parasites utilize antigenic variation to evade host immunity

P. falciparum is a specific type of apicomplexan that invades red blood cells (RBCs) within human hosts. There are four other species of *Plasmodium* that infect humans (*P. knowlesi*, *P. vivax P. malariae* and *P. ovale*), with *P. falciparum* leading to the most severe disease (Mackintosh et al. 2004).Many of the severe consequences of infection by *P. falciparum* result from both the adherence of infected RBCs to the endothelial surface of blood vessels and the ability of the parasite to avoid clearance by the immune system (Berendt et al. 1994; Wassmer et al. 2003). This process is particularly pronounced in the case of cerebral malaria, in which infected cells tightly adhere to brain vasculature leading to infarction or cerebral hemorrhage (Wassmer, Combes, Grau 2003). Both cytoadhesion and immune evasion are closely linked to variant surface antigens expression by the parasite (Miller et al. 1994). Patients experience waves of successive parasitemia that result from the expression of one cytoadhesive protein at a time on the surface of the infected red blood cell and the ability of the parasite to vary expression to evade host immune responses. The cytoadhesive proteins themselves are encoded by *var* genes, a large and highly variable family of virulence genes that exhibit mutually exclusive expression (Baruch et al. 1995). It is the single active *var* gene that determines the antigenic and virulence properties of the infected cells (Deitsch et al. 2004; Scherf et al. 2008). This process facilitates continuous cytoadhesion of parasitized RBCs within the postcapillary vasculature.

In the face of immune pressure, however, the parasite can switch expression to another single *var* gene, thereby avoiding clearance by the host immune system. This process leads to lengthy infections. Once a *var* gene is activated it appears to remain for several cycles in a process that is thought to result from epigenetic memory mediated by chromatin marks on core histones (Scherf, Lopez-Rubio, Riviere 2008). This process of constant expression of a single *var* gene product over several cell cycles facilitates the continuous and coordinated cytoadhesion of parasitized RBCs to blood vessels, a process that is most pronounced and pathogenic in the brain or placenta. Notably, it is believed that each peak of parasitemia in a patient corresponds to expression of a single *var* gene, as ascertained by cytoadherence phenotypes and nuclear run-on assays (Chen et al. 1998; Fernandez et al. 1998; Scherf et al. 1998). In the face of host immune pressure, the "on" *var* gene product, can be switched "off" and another *var* gene in turn activated, thereby avoiding clearance and leading to lengthy infections. This process represents a genetic switch underpinning antigenic variation in *P.falciparum* (Figure 1.2) (Voss et al. 2006; Deitsch et al. 2007).

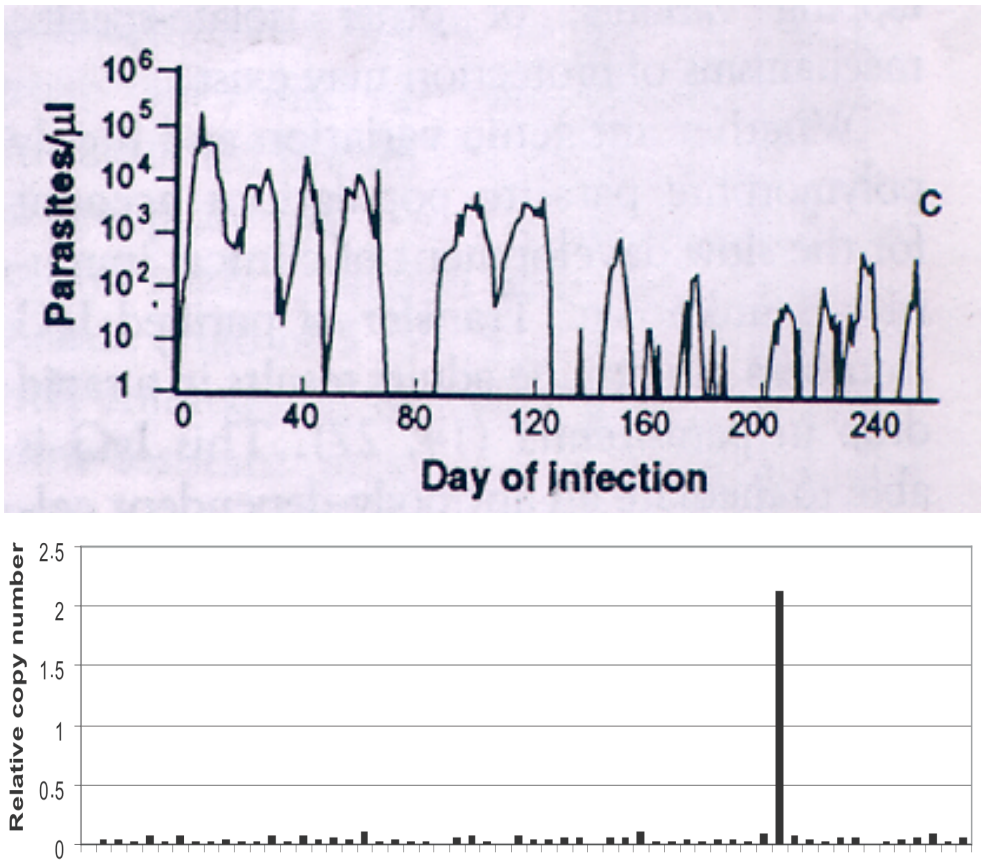

Figure 1.2. Antigenic variation in the primate malaria parasite *Plasmodium falciparum*

The top figure shows an experimental infection in a prisoner where no treatment was given until day 240 (denoted by the C = chloroquine) with regular measurement of parasitemia. The waves of parasitemia provided the first hint that *P. falciparum* utilizes antigenic variation in human infections. In the bottom figure, Dzikowski et al performed a real-time polymerase chain reaction (RT-PCR) across the panel of 60 of the *var* genes that encode the surface-exposed PfEMP-1 proteins implicated in antigenic variation. In laboratory lines generated from human clinical isolates, we observe only one *var* gene expressed in individual parasites and across a population of recently cloned parasites. It is believed that each peak of parasitemia in a patient corresponds to expression of a single *var* gene (see text for details). The asexual replicative cycle of *P. falciparum* is 2 days. The long-ranging periods of infection (20-30 days), in which it is thought that a single *var* gene is active, raises question of how *var* genes are "remembered" through the cell cycle. Adapted from: (top) Miller, LH, MF Good, G Milon. 1994. Malaria pathogenesis. Science 264:1878-1883; (bottom): Dzikowski, R, KW Deitsch. 2008. Active transcription is required for maintenance of epigenetic memory in the malaria parasite Plasmodium falciparum. Journal of molecular biology 382:288-297.

3. Antigenic variation differs across the Plasmodium genus: the case of rodent malaria parasites

Other multicopy gene families exist in various species of *Plasmodium* that also encode variant antigens displayed on the infected RBC surface. The *pir* family of genes across *Plasmodium* includes multi-copy genes implicated in virulence and pathogenesis. The *vir* genes exist in *P. vivax* and homologues exist in *P. yoelii* (*yir*), *P. chabaudi* (*cir*) and *P. berghei* (*bir*). These are considered variant antigens and are expressed at the surface of erythrocytes by late stage asexual parasites. Host immunity has been shown to modulate the transcription of the *yir* genes, consistent with current models of antigenic variation and immune escape.

The pattern of expression of the *yir* genes differ in an important regard from the *var* genes in *P. falciparum*. Unlike the strict mutually exclusive expression displayed by *var* genes, any given individual *P. yoelii* parasite expresses 1-3 *yir* genes. However, as the parasite moves through the asexual cycle (24 hours), a different set of *yir* genes may be randomly activated and displayed on the surface of the subsequent generation of infected cells (Cunningham et al. 2009). Thus there appears to be no epigenetic memory at the *yir* gene loci in rodent parasites. As a result, the aggregation of individual parasites within the parasite population appears to show all the *yir* genes as active – constituting the "smokescreen" approach of the infection (See Figure 1.3). This is thought to slow down the acquisition of protective immunity by the host, ensuring maintenance of the infection while potentially presenting all the virulence genes at once. This infection pattern is typified by a major infection peak that is followed by the death of the host or by the acquisition of sterile immunity. Rodent

malaria parasites are thought to be more ancient than *P. falciparum* and related primate parasites, suggesting that the smokescreen approach was an ancient trait.

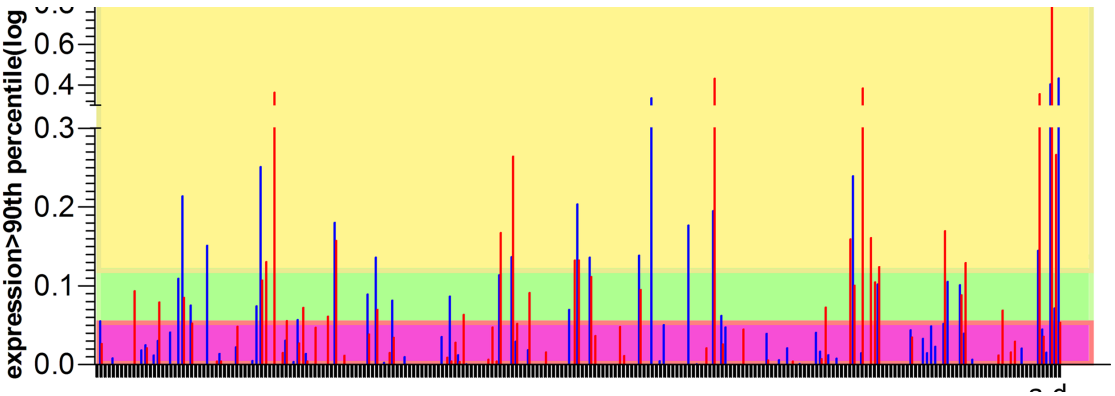

Figure 1.3. Microarray gene expression data from schizont and ring /uninucleate trophozoite stage parasites from a primary *P. yoelii* infection.

Yir gene expression signals in schizont (blue) and ring /uninucleate trophozoite (red) stage parasites for each gene are paired and shown as bars, for a large set of *yir* genes analyzed in this experiment. Positive gene expression values, above the 90th percentile cut-off in the starter population, are shown. This indicates that many *yir* genes are active in a given *P. yoelii* primary infection. Adapted from: (Cunningham et al. 2009) Cunningham, D, J Fonager, W Jarra, C Carret, P Preiser, J Langhorne. 2009. Rapid changes in transcription profiles of the Plasmodium yoelii yir multigene family in clonal populations: lack of epigenetic memory? PLoS One 4(1):e4285.

4. Transcriptional regulation by RNA Polymerase II

Antigenic variation in *Plasmodium* is regulated at the level of transcription (Deitsch et al. 2007; Kyes et al. 2007; Li et al. 2008; Scherf, Lopez-Rubio, Riviere 2008). An understanding of the mechanisms underlying antigenic variation requires an understanding of how parasites regulate transcription by RNA polymerase II. Little is known about transcriptional regulation in *Plasmodium*, but the completion of the genome sequences of several parasite species provided some clues. Using a computation screen for domains shared amongst transcription factors, Aravind reported a so-called "paucity" of transcription factors in *Plasmodium falciparum*. Examples of the gaps include general transcription factors (e.g. TFIID) (Aravind et al. 2003).

RNA polymerase II is a central component of a complex apparatus responsible for controlling the basic biochemical steps required for mRNA production, including transcription initiation, capping, elongation, splicing and polyadenylation (Howe 2002; Proudfoot et al. 2002; Zorio et al. 2004). These diverse processes are integrated by the C-terminal domain (CTD) of the protein (Carty et al. 2002). See Figure 1.4 for classic schematic on transcription from Ptashne and Gann (Ptashne 2002).

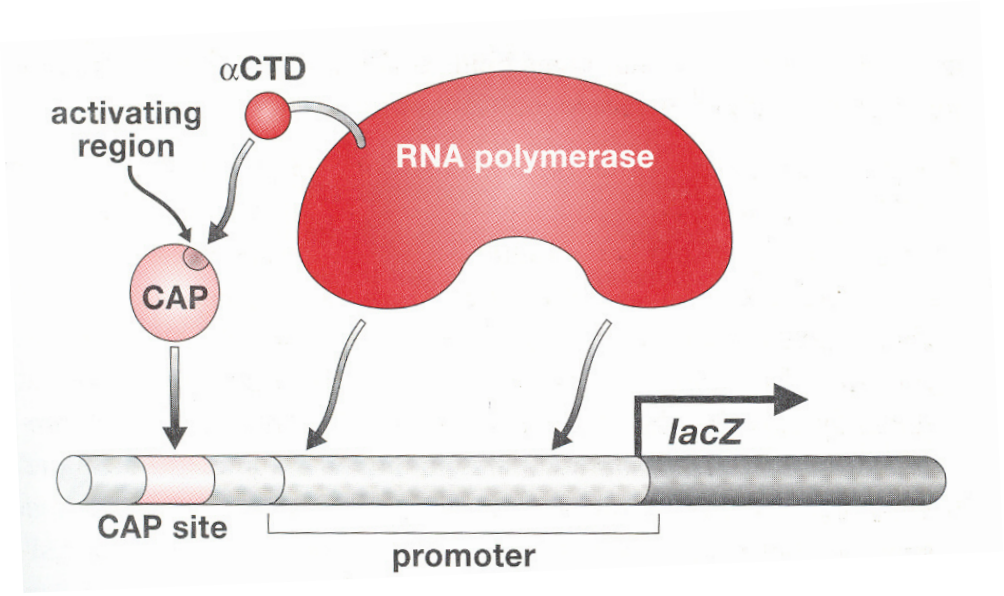

Figure 1.4. The C-terminal domain (CTD) of RNA Polymerase

Cooperative binding of catabolite activating protein (CAP) and RNA polymerase to DNA. The schematic is simplified in depicting the bacterial RNA polymerase and in presenting only one CAP monomer and only one RNA polymerase II CTD. The CAP transcription factor possesses a DNA-binding domain and the activating domain or region as shown by the arrow. In simplest terms, the CTD adheres to the activating domain of transcription factors; in higher eukaryotes the adherence may be through bridging molecules, such as the mediator to facilitate recruitment of transcription factors. Note, that while this figure depicts the basis of bacterial transcription, it was this figure in the text referenced that focused my interest on the CTD of *P. falciparum*. Adapted from: Ptashne, M, Gann, A. 2002. Genes & Signals. Cold Spring Harbor, New York: Cold Spring Harbord Laboratory Press.

The CTD, essential for gene expression in animals and fungi, is comprised of a tandem array of heptapeptide repeats, featuring a signature amino acid sequence: Y_1-S_2-P_3-T_4-S_5-P_6-S_7 (Allison et al. 1985; Corden 1990) (see Figure 1.5 for a depiction of the mammalian CTD and basic heptad structure). In yeast, rodents and humans, these repeats coordinate transcriptional events by undergoing reversible phosphorylation at the 2nd and 5th serines of the heptad (Licatalosi et al. 2002; Corden 2007) with more recent data implicating the seventh serine as critical for transcriptional elongation (Chapman et al. 2007; Egloff et al. 2007). The combinatorial phosphorylation patterns create novel docking platforms for the sequential recruitment of specific factors involved in initiation, capping, splicing and termination. I theorized that if the *Plasmodium* CTD is chemically or functionally distinct from those found in higher eukarytoes, there would be a different chemical or functional constraint for different transcription factors and activation domains. The paucity of transcription factors, then, would result from computationally searching for domains that do not exist in malaria parasites (Aravind et al. 2003).

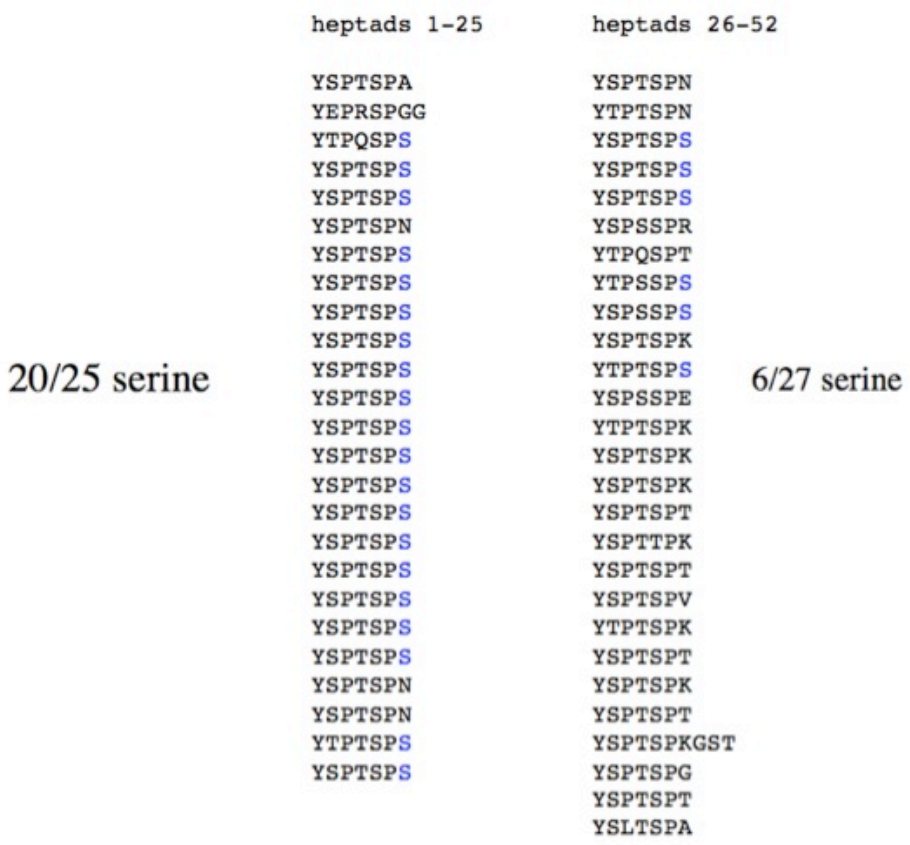

Figure 1.5. The Mammalian CTD with tandem heptads displayed.

All 52 heptads in the mammalian RNA polymerase II CTD are shown here. Each heptad is in a consecutive tandem array with other heptads. The consensus YSPTSPS heptad is found in the proximal portion; while there is wobble in the seventh position of the heptads in the distal portion of the CTD. The non-canonical YSPTSPK heptads in the mammalian polymerase are particularly notable as they are found only in mammalian and plasmodia CTDs.

5. Active transcription by RNA polymerase II is required for epigenetic memory of malaria virulence genes

Previous work in the laboratory identified the regulatory elements found within *var* genes that are required for mutually exclusive expression contributing to antigenic varation. Of note, the *var* intron was originally shown to play a role in transcriptional regulation by our laboratory (Deitsch et al. 2001b). This model was challenged in a subsequent paper by Voss et al (Voss et al. 2006). However, more recent work in the Deitsch laboratory appears to have provided a model that accommodates the data produced by both groups (Dzikowski et al. 2007; Swamy et al. 2011).

By deletion of the intron in a particular *var* gene, it is therefore possible to isolate transgenic parasites that actively transcribe from more than one *var* gene promoter (Dzikowski et al. 2008). Dzikowski et al took advantage of this observation to create a system that forced transfected malaria parasites to express increasing copies of unregulated episomal *var* promoters. This process, we hypothesize, effectively titrates away a *var*-promoter-specific transcription factor which results in repression of transcription of the "active" or "on" *var* gene independent of the classic mutually exclusive expression system (Dzikowski, Deitsch 2008). When the competing episomes were removed, the parasites did not return to expressing the same "active" gene, but displayed random activation indicating that an epigenetic imprint, potentially marked by the process of transcription itself, was lost (Figure 1.6). This led Dzikowski et al to conclude that:

> "…once activated, a *var* gene tends to remain active, suggesting the possibility that the act of transcription itself might help to reinforce the maintenance of the epigenetic marks that control cellular memory. Consistent with this hypothesis, it has recently been shown in other eukaryotic systems that certain chromatin-modifying proteins directly associate with RNA polymerase II and exert their influence on chromatin structure during transcriptional elongation, thus linking transcription to chromatin modification and epigenetic memory…"(Dzikowski, Deitsch 2008).

This notion first helped to identify a functional link between transcriptional regulation and epigenetic machinery and opened the door to the question of how this process and machinery evolved together.

6. Epigenetic machinery and modification in P. falciparum

Histone modifications are known to alter the chromatin structure to a state that is permissive (euchromatin or active) or repressive (heterochromatin or silent) states (Cui et al. 2010). This has led many to affirm the existence of a "histone code" across eukaryotes, including metazoan and protozoa. The modifications in *P. falciparum* principally feature acetylation and methylation at lysine residues across core and variant histones (Cui, Miao 2010). The targets of great interest are lysine-based modifications that are deposited and reversed by specific histone lysine methyltransferases and demethylases (reviewed below). The residues may be mono, di or trimethylated allowing the parasite maximum flexibility in imprinting. Specific

acetylases and deacetylases are also implicated in epigenetic memory (Cui et al. 2008).

The residues that undergo methylation have been most extensively studied in *P. falciparum* through mass spectrometry and the use of commercial antibodies against conserved sequences (Hakimi et al. 2007; Trelle et al. 2009; Cui, Miao 2010). The modifications observed so far include H3K34-triMe, H3K9-triMe, H3K27-triMe, H4-K20-mono/di/triMe and H3K79-triMe (Miao et al. 2006; Cui, Miao 2010). H3K9-triMe modifications are known to associate specifically with heterochromatin in *P. falciparum* and seem to be limited to the regions of the genome where the variant antigens are located (Voss et al. 2007; Lopez-Rubio et al. 2009; Westenberger et al. 2009). The specific chromatin modifiers that participate in this particular process localize to the nuclear periphery (Volz et al. 2010), suggesting a link between histone modifications and subnuclear spaces. In addition to methylation, acetylation has been observed at these lysine residues H3K9, H3K14, H3K18, and H3K27 (Miao et al. 2006) and typically associates with gene activation. Orthologs of both the isoforms of the *P. falciparum* deacetylase Sir2 have also been implicated in coordinating silencing at telomeric and pericentric chromosomal regions (Duraisingh et al. 2005; Merrick et al. 2007; Tonkin et al. 2009). Histone deacetylase inhibitors of Sir2 guide some current drug discovery efforts targeting *P. falciparum* (Andrews et al. 2009).

In addition to the core histones, there are four histone variants that exist in *P.falciparum* as confirmed by mass spectrometry. They include centromeric variants (CenH3) and the Histone 3 variant (H3.3) and H2A.Z and H2Bv (Miao et al. 2006; Cui, Miao 2010). The variant H2A.Z has been implicated in the epigenetic regulation

of *var* genes (Petter et al. 2011). The functions of the modifications at the N-termini of these variants remain poorly understood.

There are two well studied modifications specific to the genes involved in antigenic variation/ virulence (*var*) and those implicated in invasion there are two well-established modifications: Histone 3 Lysine 4-diMe and triMe (H3K4diMe and H3K4triMe) and Histone 3 Lysine 9 triMe (H3K9-triMe) (Jang et al. 1999; Scherf, Lopez-Rubio, Riviere 2008; Crowley et al. 2011). As shown in Figure 1.6, Scherf and others have proposed a model for antigenic variation in *P. falciparum* whereby H3K9-triMe is heavily restricted to multi-copy gene families that are surface-exposed (including the *var* family) and to the repressed members of this family sequestered in heterochromatic regions. Chookajorn et al have independently confirmed this, finding H3K9-triMe at silent *var* genes (Chookajorn et al. 2007b); others have found that H3K9-ac is found at active *var* genes (Chookajorn et al. 2007a; Scherf, Lopez-Rubio, Riviere 2008). Meanwhile, H3K4-triMe is not restricted to the *var* gene family but is generally associated with active expression. In other species, as reviewed below, the H3K4 methyltransferase (Set1) is associated with actively elongating RNA polymerase II and thus links the process of transcription to epigenetic memory (Li et al. 2007; Li et al. 2011). Other modifications that have been described in the *P. falciparum* literature, including H3K36-triMe and H4K20mono/di/triMe, both the specific modifications and the chromatin modifiers that deposit them - might shed additional light on the regulation of the *var* gene family in *P. falciparum* (Hakimi, Deitsch 2007).

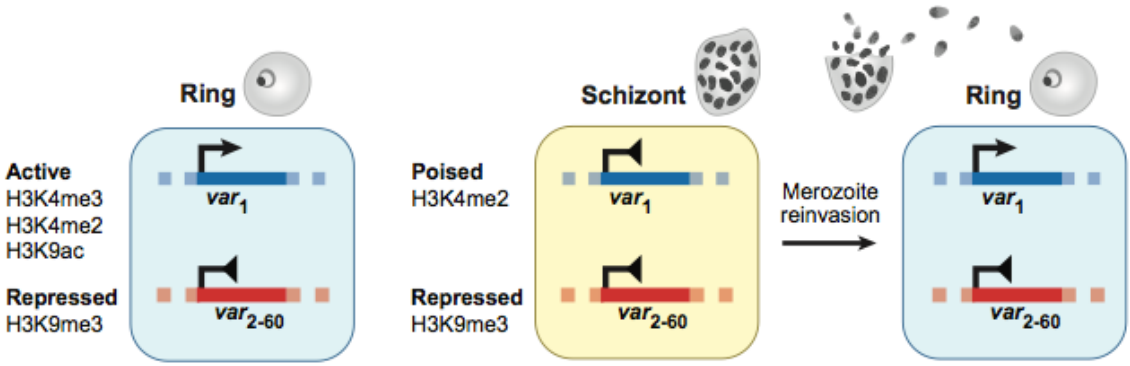

Figure 1.6. Histone marks linked to *var* gene expression.

Adapted from: Scherf, A, JJ Lopez-Rubio, L Riviere. 2008. Antigenic variation in Plasmodium falciparum. Annual review of microbiology 62:445-470.

7. Histone lysine methyltransferases in P. falciparum

Histone methylation is involved in both transcriptional activation and silencing are governed by histone methyltransferases (HKMTs) and histone demethylases (Li, Carey, Workman 2007). The HKMTs include the disruptor of telomeric silencing (DOT) and the SET family of proteins. The SET family was identified as involved in the suppression of genes in Drosophila. The full name stands for: *Su*(*var*)3 -9 in the gene *Enhancer of zeste* (*E*[*z*]) and in the activating gene *Trithorax*. The SET domain comprises a 130 amino acid motif with family members broadly implicated in lysine methylation. Early insights revealed that histone methyltransferases were able to mono -, di- and tri- methylate histones which provided great versatility for transcription. The first to be described was Su(var)3-9 which represented the H3K9 methyltransferase. This enzyme was shown to be involved in gene silencing. The literature over the past decade has shown that SET domains are present in all eukaryotes and are essential for diverse processes in development, immunity and transcriptional regulation. The known SET homologs and known substrates in *Plasmodium falciparum* are shown in Table 1.0. The SET-domain bearing proteins found in *P. falciparum,* along with their architecture are profiled in Figure 1.7.

PFF1440w	PfSet1	H3K4Me1-3
MAL13P1.122	PfSet2	H3K36Me2,3
PF08_0012	PfSet3	H3K9Me2-3
PFF10485c	PfSet4	H3K4
PFL0690c	PfSet5	Unknown
PF13_0293	PfSet6	H3K4
PF11_0160	PfSet7	Unknown
PFD0190w	PfSet8	H4K20Me1-3
PFE0400w	PfSet9	Unknown

Table 1.1.The SET family of methyltransferases and their substrates (Cui, Fan, Miao 2008)

Cui, L, Q Fan, J Miao. 2008. Histone lysine methyltransferases and demethylases in Plasmodium falciparum. International Journal for Parasitology 38:1083-1097.

8. The Set2 methyltransferase links transcription and epigenetic memory

Two SET-domain-bearing proteins link active transcription by RNA polymerase II to epigenetic memory, Set1 and Set2. Both have been heavily studied in budding yeast. Set1 orthologs interact with the CTD indirectly through the recruitment of the PAF complex (Hampsey et al. 2003) and deposit the H3K4-triMe mark at the 5' regions of expressed genes, with enrichment at promoter regions (Hampsey, Reinberg 2003). Set2, by contrast binds the CTD directly via the Set2-RBP1 interaction domain (SRI) and deposits the H3K36-triMe mark along the length of the gene, with its greatest deposition at the 3' end of coding regions (Gerber et al. 2003; Hampsey, Reinberg 2003). Numerous studies have detected Set2 interacting with the RNA polymerase II CTD directly (Gerber, Shilatifard 2003; Vojnic et al. 2006), as shown in Figure 1.7

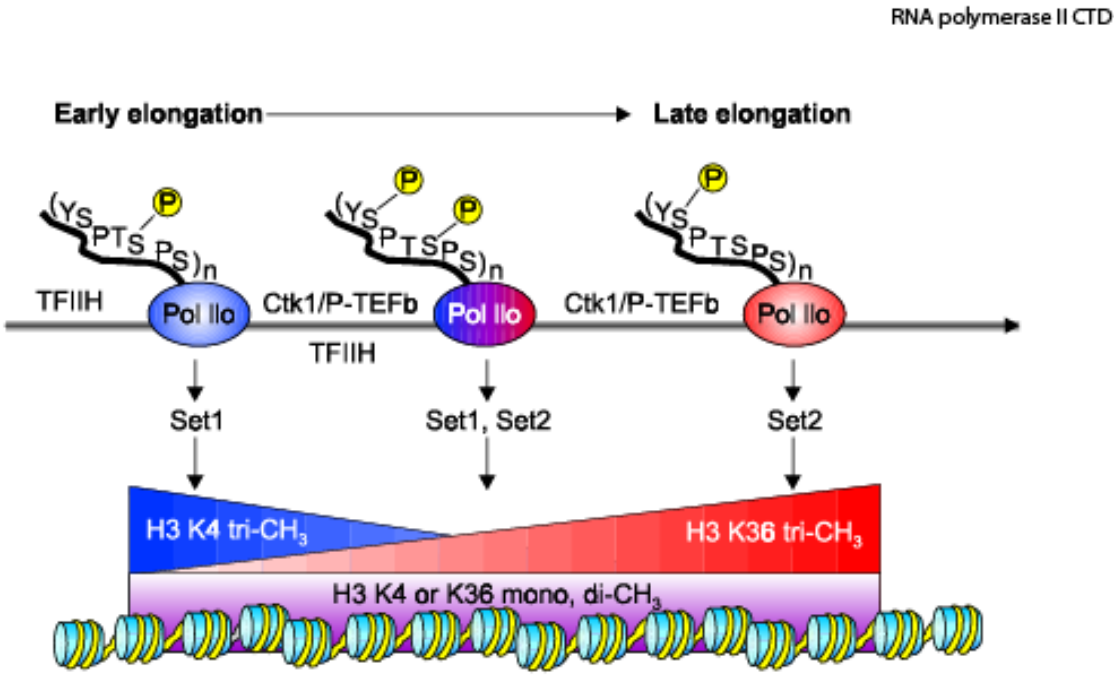

Figure 1.7. Set2 is recruited to actively transcribed genes

Increasing evidence suggests that Set2 is the major histone-modifier associated with the transcription apparatus, a result that may indicate an important role for Set2 and H3 Lysine methylation in regulating RNA polymerase II transcription through changes in chromatin structure (Krogan et al. 2003). This idea is supported by recent results showing that budding yeast strains deficient in Set2 are sensitive to 6-Azauracil (a phenotype correlated with transcription elongation defects) and are synthetically growth deficient with other transcriptional elongation factors (Krogan et al. 2003).

Recent data indicate that the histone H3K36 methyl groups added by Set2 recruit a histone deacetylase that dampens the activity potential of newly transcribed chromatin (Carrozza et al. 2005; Joshi et al. 2005; Keogh et al. 2005; Chu et al. 2006). Set2-mediated methylation of H3K36 in vivo requires the presence of phosphorylated CTD to adhere properly (Krogan et al. 2003). Correspondingly, the SRI domain is essential for co-transcriptional methylation (Kizer et al. 2005). The binding specificity of this 100-amino-acid domain was determined using a series of synthetic CTD peptides of varying length and phosphorylation patterns (Kizer et al. 2005; Li et al. 2005). Truncations of the yeast CTD below 12 heptads compromise the ability of the Set2 to deposit tri-methyl modifications as the Set2 protein lacks a definitive docking site (Xiao et al. 2003).

9. The Set8 methyltransferase controls mitotic progression and heterochromatin propagation in apicomplexans

Another central methylation event in eukaryotic chromatin is the modification of Histone 4 Lysine 20 (H4K20) by the specific histone methyltransferase Set8. In metazoa, H4K20Me1 is exclusively induced by the Pr-Set7 HMTase (canonical Set8) (Fang et al. 2002) and has been linked with transcriptional repression, mitotic regulation and DNA checkpoint repair (Karachentsev et al. 2005; Jorgensen et al. 2011). It is currently thought that the H4K20 methylation in DNA repair and genome integrity can be considered an early evolutionary function whereas the heterochromatin function of the marker arose later in evolution in response to increasing genome complexity (Karachentsev et al. 2005).

In certain Apicomplexa (*N. caninum*, *P. falciparum* and *T.gondii*), mono-, di-, and tri-methylated forms of H4K20 can exist. It was recently shown that H4K20Me1 and H4K20Me3 mark heterochromatic regions in Apicomplexa at rRNA genes, satellite repeats and telomeric regions (Sautel et al. 2007). In addition, H4K20 associates with H3K9-triMe in the apicomplexan *T.gondii*, which may help not only establish but propagate silent marks at the multi-copy gene families such as *var* (to which H3K9 methylation is restricted to) (Lopez-Rubio, Mancio-Silva, Scherf 2009).

10. Lysine demethylases in P. falciparum

In addition to methyltransferases, most eukaryotes also possess demethylases that remove methyl marks from histone residues. This allows histone methyl modifications to be reversible. To date, two families of histone lysine demethylases have also been identified in *P.falciparum*: lysine-specific demethylases 1 (LSD1) and Jumonji-C (JmjC)-domain-containing histone demethylases (JHDMs) (Klose et al. 2006a). LSD1 is known to demethylate mono and dimethylated histones. Demethylases bearing JmjC domains catalyze demethylation of mono, di and trimethylated histone lysine residues. JmjC-domain-bearing family members demethylate H3K36 and H3K9. The protein family was discovered and named "Jumonji" after Higashinakagawa described a gene knockout that led to a cruciform ("jumonji") shape in the neural plate of mice. *P. falciparum* possess two JmjC1 proteins. While the substrate of PfJmjC2 remains unknown, PfJmjC1 is a member of the JHDM3/JMJD2 sub-family known in other species to remove methyl groups from both H3K9 and H3K36. Specifically, it is a member of the JMJD2A family as the protein features a 40 amino acid JmjN domain upstream of the JmjC domain (Cui, Fan, Miao 2008). Phylogenetic analyses also show that apicomplexan JmjC1 proteins are most similar to the JMJD2A subfamily (Cui, Fan, Miao 2008). Of note, work in the fruit fly *Drosophila melanogaster* has shown cross-talk between H3K9-triMe and JMJD2A orthologs. Heterochromatin protein 1 (HP1), which is recruited to H3K9-triMe, stimulates the active removal of H3K36-triMe modifications by *Drosophila* JMJD2A (Zofall et al. 2006).

11. Horizontal (lateral) gene transfer in apicomplexa

The concept of a lateral transfer supposes that a genomic fragment is inherited laterally, or transferred between two otherwise distantly related organisms. This is a mode of inheritance that differs from traditional vertical inheritance pathways. Examples of lateral gene transfer have been reported for alterations to the metabolic repertoire of protists adapting to new environments (Huang et al. 2004; Andersson 2009). Currently, there is no evidence of transposable elements or viral-mediated transfer events in Apicomplexa (Gardner et al. 2002).

To explain what is meant by later transfer in Apicomplexa, it should first be noted that Apicomplexa possess a plant (algal) relict, which itself contains a cyanobacterial relict (chloroplast). There is controversy about whether the algal endosymbiont is a green or red algae. Regardless, this vestige, the plastid, retains transcriptional machinery and ribosomal RNA genes, along with a 35kb circular plastid genome (McFadden et al. 1996; Kohler 2005). Based on evidence that the plastid is necessary for apicoimplexan viabilitiy, researchers are investigating new ways of adapting herbicides and antibiotics for use in apicomplexan infections, including malaria.

It has been theorized that the enzymes involved in fatty acid biosynthesis and the ApiAP2 transcription factors, the largest family of transcription factors in Apicomplexa (most extensively studied in *P. falciparum*), originated in the algal endosymbiont and were acquired by the parasite along with the plastid (Balaji et al. 2005). Ap2-Sp has been shown to be involved in the transcription of sporozoite-specific genes (Yuda et al. 2010) and the AP2 protein PfSIP2 is involved in telomere

regulation (Flueck et al. 2010). Bacterial protein domains such as the discoidin domain came either from the cyanobacterial endosymbiont or from bacteria that were environmentally present in the surroundings of ancient Apicomplexa (Templeton et al. 2004). In a general sense, these transfers from the plastid are from the "inside-out", from an organelle within the parasite to the parasite's nuclear genome.

Alternatively, lateral transfer in Apicomplexa can occur from the "outside-in", derived from the host that harbors the organism; this represents canonical horizontal or lateral gene transfer. The intracellular lifestyles of Apicomplexa, and the observed ability of malaria parasites to take up and integrate DNA fragments from the host cytoplasm into their genome, provide mechanistic support for this type of transfer (Deitsch et al. 2001a). The known examples of lateral transfer from an animal host to Apicomplexa concern extracellular domains involved in cytoadhesion and O-linked glycosylation enzymes (involved in glycosylating the same extracellular domains (e.g. EGF, TSP) (Templeton 2007). The requirement for the parasite to evade host immunity, particularly adaptive immunity involving the thymus and spleen, provides a selective advantage for it to cytoadhere in deep tissues in extant animal hosts.

Thus, as shown in Figure 1.8, the complex picture of lateral gene transfers in Apicomplexa must take into account i) the primary endosymbiont cyanobacterium (chloroplast) harbored inside, ii) the secondary endosymbiont, a red or green alga (plastid) that once imparted photosynthetic abilities to ancestral apicomplexa and iii) the obligate intracellular apicomplexan harbored inside the host as a parasite (and no longer as a symbiont).

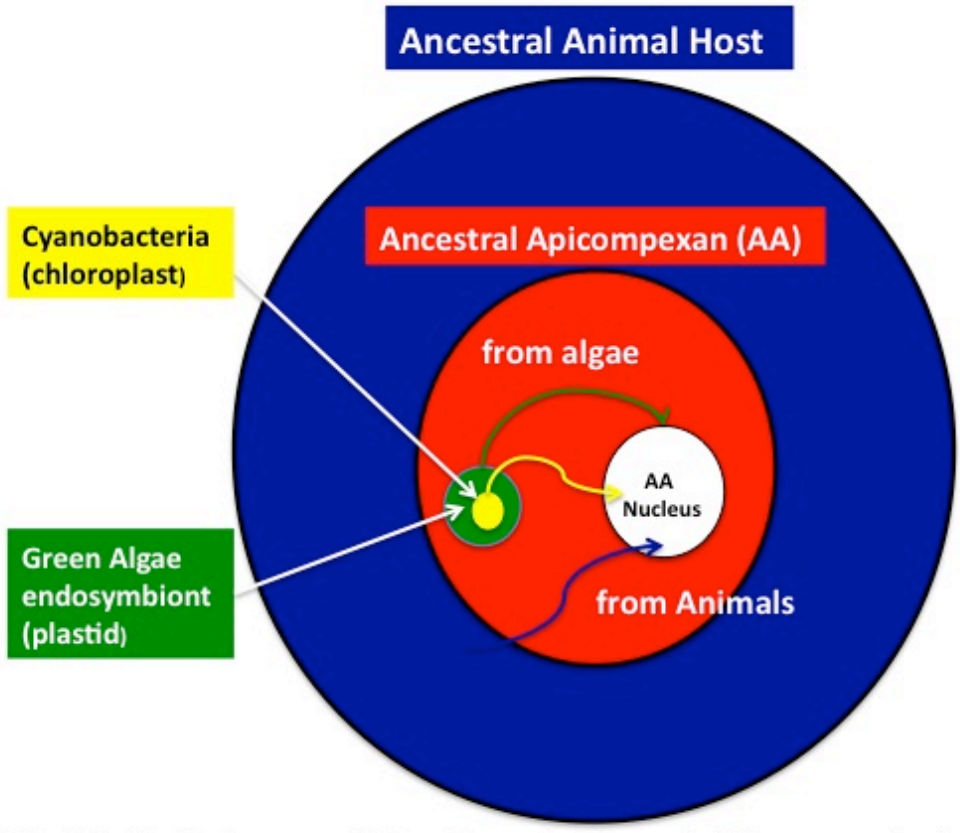

Figure 1.8. The various modes of horizontal gene transfer in Apicomplexa
Horizontal gene transfer is possible from the cyanobacterial relict (chloroplast), the algal endosymbiont (plastid) or from an animal host. Note, a green alga is shown as the endosymbiont for simplicity; this may be a red alga as well.

12. Thesis Objective

My thesis focuses on the evolution of transcriptional regulation and epigenetic machinery in malaria parasites. The overall claim of this thesis is that the acquisition of additional transcriptional regulatory units and epigenetic machinery facilitated the transition of ancestral apicomplexans to parasitic lifestyles and fundamental changes in the development and cellular differentiation in various hosts. These acquisitions were derived from animal hosts (with strong support) and plants or algal endosymbionts (with weaker support). When the regulatory units and machinery were not necessary, we observe lineage-specific deletions.

First, using *P. falciparum* and a collection of rodent, bird, human and non-human primate parasites, I show that the ancestral *Plasmodium* RNA polymerase II CTD was short (eight heptads), serving as the smallest CTD observed in evolution. This short CTD remains in bird and rodent malaria parasites. I then show that the CTD in primate (but not rodent or bird) malaria parasites has twice expanded, independently in evolution, with the expansion of the YSPTSPK heptad in the CTD, a heptad only otherwise found in mammalian polymerases. The expansion appears to be ongoing even at the species level where I provide the first demonstration of a CTD polymorphism within a species (*P. falciparum*) (see Chapter 2).

Next, I link the innovations in CTD expansions to transcriptional regulation and the evolution of epigenetic machinery in *Plasmodium*. I show that a pair of reciprocal Histone 3 Lysine 36 (H3K36) epigenetic modifiers are absent from ancient *Plasmodium* lineages that infect rodents (and feature the shortest CTDs described to date), but present in parasites with expanded CTDs. Of special note, orthologs of the

H3K36 methyltransferase (Set2) are known to bind to the CTD itself in higher eukaryotes during active transcription, depositing H3K36-triMe modifications involved in relaying epigenetic memory for subsequent gene activation. This observation raises the possibility of an expanded CTD-Set2 axis in *P. falciparum* and primate parasites.

In my research, I have documented the expression of these modifiers across the asexual replicative cycle and have developed the first antibody specific to the *P.falciparum* H3K36-triMe. My work conclusively demonstrates the presence of the modification in this organism with early-stage (ring) enrichment coinciding with the expression of *P. falciparum* variant genes (*var*). I have also studied the natural history of the acquisition and deletion of these modifiers in *Plasmodium* and Apicomplexa, finding that Set2 and JmjC1 were present in other Apicomplexa, particularly those with expanded CTDs, and may have been horizontally transferred to an ancestral apicomplexan from an algal endosymbiont. I further suggest that the rodent parasite lineages deleted both chromatin modifiers over time, as this lineage failed to expand the CTD to functionally accommodate the H3K36 modifiers in the first instance (See Chapter 3).

Finally, building on my interest in horizontal acquisitions of chromatin modifiers, I have studed the acquisition of the H4K20 histone lysine methyltransferase implicated in mitosis, cellular differentiation, epigenetic memory and antigenic variation in primate malaria parasites. I provide the first conclusive evidence of a horizontal gene transfer of this methyltransferase gene from an animal source and suggest the source of the transfer was a nematode. I theorize that this event transpired

400 million years ago, when the ancestors of apicomplexans and parasitic nematodes co-existed. Notably, the transfer coincides with the postulated horizontal transfer events from animals of domains linked to parasitic invasion and adherence (domains involved in cytoadhesion O-linked glycosylation).

The work, in total, sheds light on the evolution of transcriptional regulation and epigenetic machinery at the nexus of host: parasite interactions and the transitions attendant to apicomplexan parasitism. The work was done in collaboration with Thomas J. Templeton of Weill Cornell Medical College, Susan L Perkins of the American Museum of Natural History, and John Stiller of East Carolina University.

CHAPTER 2:
THE EVOLUTIONARY PLASTICITY OF THE RNA POLYMERASE II CTDS OF MALARIA PARASITES

1. Introduction

The length of the RNA polymerase II C-terminal domain (CTD) varies in different organisms and roughly tracks with genome size. For example, the CTD of the microsporidian *Encephalitizoon cuniculi* contains 15 repeats while a CTD containing 52 repeats is present in all mammalian species thus far described (Chapman et al. 2007; Egloff et al. 2007). The perfect conservation of the mammalian CTD demonstrates the strict conservation of this portion of RNA polmerase II within broad evolutionary lineages. The length of the CTD is essential for RNA polymerase II stability, transcriptional efficiency and ultimately cell viability. For example, the CTD of the budding yeast *Sacharomyces cerevisiae* consists of 26 heptad repeats and truncation to 17, 14, 13 or 11 repeats leads to respective reductions in specific transcript levels to 58%, 8%, 5% and 2% of that found in wild type cells (Liao et al. 1991). Further reduction of the repeat length to 8-10 repeats yields cells that are temperature-sensitive, and truncation to less than eight repeats is lethal. Similarly, 30 of the 52 heptads are necessary for viability in mice (Meininghaus et al. 2000) and mutants with 31-39 heptads exhibit growth defects (Bartolomei et al. 1988; Litingtung et al. 1999). The importance of CTD length is reflected in the fact that this region of

the protein is strictly conserved in closely related organisms, and no variability within an individual species in the number of heptad repeats has been observed.

Throughout evolution, the seventh position of the conserved heptad repeat displays the most heterogeneity in higher eukaryotes(Guo et al. 2005; Liu et al. 2008) , and in yeast the serine in this position is non-essential(Stiller et al. 2000). The canonical or consensus heptads are Y_1-S_2-P_3-T_4-S_5-P_6-**S**$_7$. Intriguingly, mammals, including rodents and primates, feature seven to nine non-consensus Y_1-S_2-P_3-T_4-S_5-P_6-**K**$_7$ heptads near the distal end of the CTD (Barron-Casella et al. 1992; Guo, Stiller 2005). In mammals, while 20 of the first 25 heptads contain a serine in the seventh position, only 6 of the last 27 heptads do, and 8 of the last 17 heptads all contain a lysine in this position. In mice, however, these lysine-containing repeats are dispensable, with polymerases consisting of only consensus repeats supporting normal growth and viability of cells (Chapman et al. 2005). The role of the lysine heptads in mammalian polymerases remains an outstanding question in RNA polymerase II biology and awaits functional, biochemical and regulatory characterization.

2. Results

2.1 Plasticity of the CTD within the genus Plasmodium

While evolutionarily very distant, the CTD heptad repeats of *P. falciparum* bear an interesting similarity to those found in higher eukaryotes, and the presence of heptads containing lysine in the seventh position is reminiscent of the mammalian CTD. We therefore hypothesized that the heptad structure might be significant for *Plasmodium* RNA polymerase function and thus might provide insight into transcriptional regulation in these organisms. Previously, two laboratories described a greater number of CTD heptad repeats in *P. falciparum* when compared with the

rodent malarias *P. berghei* and *P. yoelii* (Giesecke et al. 1991; Chapman et al. 2008; Egloff et al. 2008), which suggests an unusual variability in CTD structure between these closely related organisms. Since the number of CTD heptad repeats is generally stringently conserved between even distantly related genera, we wondered if the differences between the human and rodent parasites reflected a general plasticity of CTD heptads in the genus *Plasmodium*.

To obtain a broad perspective on CTD sequences from throughout the *Plasmodium* genus, we collected sequences from the largest subunit of RNA polymerase II (RBP1) in parasites infecting humans (*P. falciparum*, *P. vivax* and *P. ovale*), non-human primates (*P. reichenowi*, *P. knowlesi* and *P. fragile*), rodents (*P. berghei* and *P. yoellii*) and birds (*P. gallinaceum*). The sequences were either obtained from genome sequence databases at the NCBI or Sanger Institute, or directly amplified from parasitized blood obtained from field samples. This collection of sequences was then aligned using CLUSTALW and adjusted through visual inspection; finally, repeats were tabulated as described in the Methods portion of this chapter.

Analysis of the amino acid sequences from the various species showed that they all maintained the typical, conserved CTD structure consisting of a linker region (R1), the heptad repeat-containing region (R2), and a tail region following the last repeat (R3) (Chapman et al. 2008). However, while all of the species possessed the heptad repeat structure typical of CTDs from higher eukaryotes, there was a remarkable degree of variability in the number of repeats in region R2 as well as in the amino acid found in the seventh position of each heptad (Figure 2.1). This is in stark

contrast to the extreme conservation of heptad sequences observed in many other evolutionary lineages. For example, mammals from different orders exhibit no variability in either heptad structure or number of repeats, with all species thus far examined displaying indistinguishable CTDs consisting of 52 identical repeats. Second, the *Plasmodium* CTDs are remarkably short, consisting of only eight repeats in the rodent and bird parasites, *P. berghei, P. yoellii*, and *P. gallinaceum,* and reaching a maximum of 13-15 repeats in the primate parasites *P. knowlesi*, *P. vivax*, *P. reichenowi* and *P. falciparum*. The exception to this trend is *P. ovale*, the earliest branching primate parasite most closely related to rodent parasites, which possesses only eight repeats within the CTD. In addition, the different *Plasmodium* species all displayed a propensity toward a non-serine amino acid in the seventh position, and in particular the primate parasites exhibited an expanded set of tandemly arrayed repeats containing lysine in this position. All of the species that display an expanded CTD infect monkeys, apes or humans, thus correlating increased heptad number with parasitism of primates.

When phylogenetic analyses between the various species of *Plasmodium* are overlaid on the CTD data, there is support for two separate expansions of the lysine heptads (indicated by an asterisk in Figure 2.1) in primate parasites. One expansion occurred after *P. fragile*, *P. knowlesi* and *P. vivax* diverged from *P. ovale* and the other expansion occurred in the lineage giving rise to *P. falciparum* and *P. reichinowi*. Codon usage data may further support this argument; *P. falciparum* and *P. reichinowi* share expansion of a single codon for the lysine repeats (AAA), whereas the other primate parasites that have undergone expansion (*P. vivax*, *P. knowlesi* and *P. fragile*)

feature lysine codon AAG as the common or predominant codon in the repeats. Furthermore, the rest of the RNA polymerase II molecule of these parasites is more closely related to the polymerase of the rodent parasite *P. berghei*, which lacks the expanded motifs, than to the primate parasite *P. falciparum*, which features them (data not shown). The separate expansions imply that the increased CTD length provides a selective advantage for parasites infecting primate hosts.

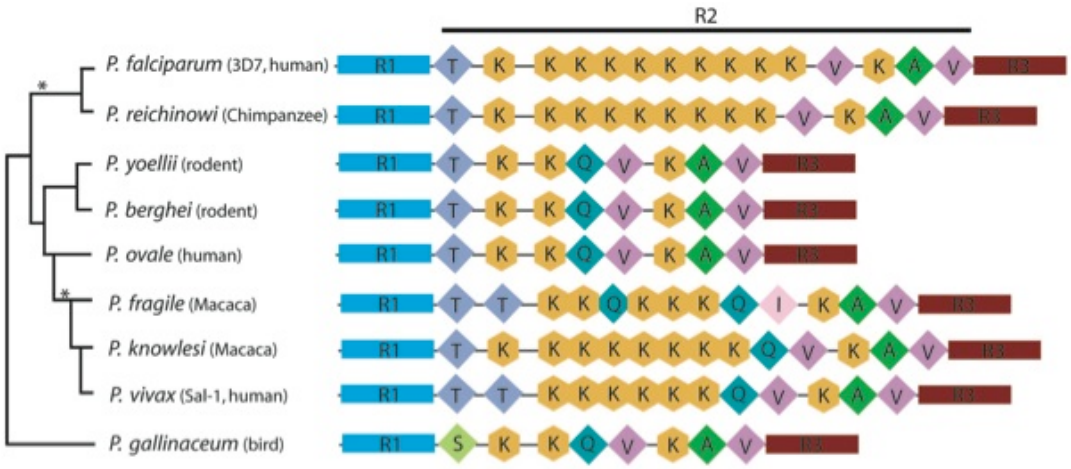

Figure 2.1. Plasticity of the heptads in the RNA polymerase II CTD across the *Plasmodium* genus.
Schematic depiction of the CTD heptads from nine species of *Plasmodium*. The species names are listed at the left of each line along with a tree indicating their evolutionary relationships. R1 is the linker domain, R2 is the heptad-bearing region and R3 is the acidic tail region following the last repeat of the CTD. Diamonds and hexagons symbolize individual heptads, with the amino acid in the seventh position indicated. The plasticity of the CTD is restricted to heptad-containing region, R2. Rodent malaria parasites (*P. yoelli* and *P. berghei*) and the bird parasite (*P. gallinaceum*) exhibit eight repeats with a common stalk (Q,V,K,A,V). Primate malaria parasites feature an expanded set of tandemly arrayed heptads all containing lysine in position 7 of heptads immediately preceding this stalk. Two separate, independent expansions (indicated by asterisks) are proposed to have occurred. Sequences drawn from PlasmoDB (www.plasmodb.org) and contigs at the Sanger Institute: *P. berghei*: ANK,PB000038.00.0; *P. yoelii*17xNL: PY03187; *P. vivax* Sal-1: PV095320; *P. knowlesi* Strain H: PKH_082310; *P. falciparum* 3D7: Pfc0805w. *P. gallinaceum:* EU840284 and *P. reichinowi*: EU850397, *P. fragile:* EU840282 and *P. ovale:* EU887536

2.2. Plasticity of the CTD within a single species of Plasmodium

Such extensive plasticity in CTD heptad repeat numbers and composition between closely related species of a single genus has not been described before. We therefore questioned whether the CTD structure might be rapidly diverging within the *Plasmodium* genus, and consequently whether such plasticity might extend to an individual species. For this analysis we took advantage of recent efforts to sequence the genomes of several independent isolates of *P. falciparum* obtained from infected individuals in different malarious regions around the world. The sequence of RNA polymerase II from an isolate from Honduras (Li et al. 1989) (HB3;NCBI database) and one from Papua New Guinea (D10; Broad Institute database) were analyzed for CTD sequence and structure. Notably, both of these isolates exhibited fewer heptad repeats (HB3: 14 repeats; D10: 13 repeats) than 3D7 (15 repeats), the reference strain used for the initial genome sequencing project. To determine if the change in repeat number was an artifact of genome sequence assembly, we obtained genomic DNA from both 3D7 and D10 (gift from Dr. X. Su, National Institute of Allergy and Infectious Disease), directly amplified the CTD encoding region of the gene using species-specific PCR primers and sequenced the purified amplicons. We found that both sequences precisely matched the sequences reported in the genome nucleotide sequence databases, thus verifying the observed polymorphisms.

To determine the degree of repeat number polymorphisms within *P. falciparum*, we extended the analysis to additional isolates from other geographically diverse locations, including 4 African, 5 Asian and 3 American isolates of *P. falciparum* from the Malaria Research and Reference Reagent Resource Center (MR4)

of the American Type Culture Collection (ATCC). We amplified the CTD from genomic DNA and directly sequenced the purified amplicons. The results displayed an even greater degree of polymorphism than initially observed, with isolates exhibiting CTD lengths ranging from as few as 12 to as many as 17 repeats (Figure 2.2A). Both heptad repeats and diheptad repeats (repeats in tandem) were tabulated. The variations appear to cluster weakly by geographic region. Asian isolates (i.e.. DD2, K1 and D10) tend to be shorter (12-13 repeats) while American isolates (i.e. St. Lucia, HB3 and 7G8) all have 14 repeats. African isolates tend to be the longest with 15 in 3D7, 17 in FCR3 and 14 in RO33;.The African isolate D6, with only 13 repeats, is an exception. Figure 2.2B shows a neighbor-joining tree of the various *P. falciparum* isolates based on 137 SNPs (Volkman et al. 2007), indicating that the CTD polymorphism tracks with continental Asian and American clades.

We next asked whether two other primate parasites, *P. vivax* and *P. cynomolgi,* exhibit plasticity in CTD repeat numbers similar to *P. falciparum*. As with the *P. falciparum* samples, using known database sequence and genomic DNA from MR4, we amplified the CTD from several DNA samples and directly sequenced purified amplicons. Similar to *P. falciparum*, we found plasticity of the CTD specific to the lysine-containing, tandemly arrayed heptad repeats (Figure 2.2A). There is also a correlation between the number of repeats and geographic location, with Asian isolates of *P. vivax* (ONG and Pakchong) exhibiting more repeats (15 and 14, respectively), than the American (Nicaragua, Panama and Salvador-1) isolates (13 repeats). The Smithsonian isolate of *P. cynomolgi* features one more repeat than the Bastianelli and Cambodian isolates.

To determine if the intra-species polymorphisms in CTD length extend to other *Plasmodium* species, we took advantage of 11 rare samples of genomic DNA obtained from various isolates and subspecies of the rodent malaria parasites *P. berghei*, *P. yoelii* and *P. vinckei* as described in Perkins et al., 2007 (Perkins et al. 2007). These rodent parasite species are restricted to West and Central Africa; there are no other continental isolates available. We observed no differences in repeat numbers either across or within the *P. berghei*, *P. yoelii* or *P. vinckei* species or subspecies (Figure 2.2A). The number of repeats was also identical to the number reported for the bird parasite, *P. gallinaceum*. These data suggest that the diversity of heptad repeat numbers may be specific to the primate malaria parasites *P. falciparum, P. vivax* and *P. cynomolgi*. In addition, the polymorphisms do not appear to be an artifact of laboratory culture, as seen in budding yeast (Nonet et al. 1987), because fresh uncultured clinical isolates also display plasticity of repeats (unpublished data).

A

Species	Isolate	Lysine containing heptad repeat region
P. cynomolgi (Macaca)	Cambodian	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	Brasillanum	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	Smithsonian	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
P. vivax (human)	Salvador-1	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	Panama	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	Nicaragua	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	Pakchong	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
	ONG	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPQYSPSSPV
P. falciparum (human)	K1 (Thailand)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	W2 (Indo-China)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	DD2 (Indo-China)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	D10 (Papua New Guinea)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	FCC-2 (China)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	D6 (Sierra Leone)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	St.Lucia (El Salvador)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	HB3 (Honduras)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	7G8 (Brazil)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	RO-33 (Ghana)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	3D7 (?)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK----------YSPTSPVAQNIASP
	FCR3 (The Gambia)	SPVYSVTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPKYSPTSPK YSPTSPVAQNIASP

B

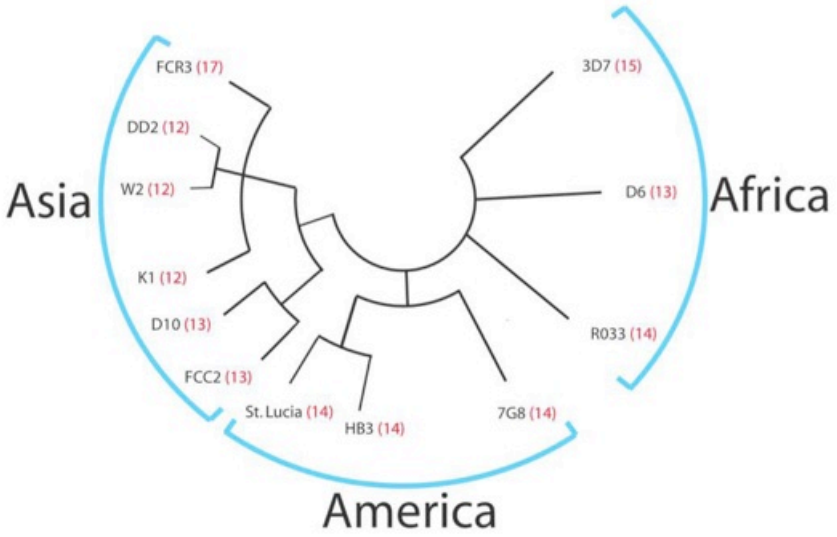

Figure 2.2. Plasticity of the heptads in the RNA polymerase II CTD across *Plasmodium* species.

Amino acid sequences of the R2 domain of 2 isolates of *P. berghei*, 5 isolates of *P. yoelii* and 4 isolates of *P. vinckei* (all rodent parasites), 3 isolates of *P. cynomolgi*, 5 isolates of *P. vivax* and 12 geographical isolates of *P. falciparum*. The lysine in the seventh position of each heptad repeat is shown in red. Numbers of total heptads and diheptads are tabulated on the right. B. Neighbor-joining tree of various *P. falciparum* isolates based on 137 SNPs and arranged according to geographical origin. The number of heptad repeats found in the CTD of each isolate is shown in red. Adapted from Volkman et al (Volkman et al. 2007).

2.3 Evolution of the CTD in protozoa

The heptad repeat structure of the CTD is a ubiquitous feature of RNA polymerase II of metazoa, and it is indispensable for transcriptional activity in these organisms. Among the unicellular eukaryotes, however, some organisms possess RNA polymerase II enzymes that contain CTDs with recognizable heptads, while others do not, suggesting that the CTD evolved within an ancient organism near this point in the eukaryotic lineage. In light of the remarkable diversity of CTD heptads within the *Plasmodium* genus and within different isolates of *P. falciparum* and *P. vivax*, we investigated whether closely related protozoa possess conserved CTD heptads. In particular we were interested in other apicomplexans, as well as other alveolates such as dinoflagellates and ciliates. Our examination of CTDs from protozoa using conventional genome sequence databases indicated that CTD heptads are also present in the apicomplexans *Babesia bovis*; *Theileria spp; Toxoplasma gondii* and *Cryptosporidium spp* as well as in the early branching dinoflagellate *Perkinsus marinus*. The apicomplexans and dinoflagellates examined possessed heptad repeats that varied in size from nine in *Toxoplasma* to as many as 24 in *Cryptosporidium*. The seventh position of the heptads varies across organisms with an alanine in *Babesia* and *Perkinsus* and a histidine in *Cryptosporidium*. The two closely related parasites, *Cryptosporidium parvum* and *C. hominis*, possess the same heptad repeat composition and number throughout their CTDs, whereas the murine parasite *Cryptosporidium muris* displays the same number of heptads but uses a different heptad composition that relies on histidine and arginine in the seventh position. Thus the remarkable variability observed in repeat numbers in *Plasmodium* may be unique to this genus.

While there is extensive variation within Apicomplexa, all protozoa in this group possess heptad repeats. By contrast, ciliates lack CTD heptads altogether despite their grouping with apicomplexans within Alveolata. Interestingly, the RNA polymerases of ciliates possess serine-proline (SP) motifs that are enriched within the C-terminus but not elsewhere in RNA polymerase II (Figure 2.3). A similar pattern of "SP" richness is also seen in the eukaryotic human pathogen *Trichomonas* (Dacks et al. 2002), but is not seen in *Giardia* or the kinetoplastids (*Trypanosoma cruzi*, *T. brucei* and *Leishmania major*). The "SP" motifs are reminiscent of Y_1-S_2-P_3-T_4-S_5-P_6-X_7 heptad motifs found in metazoan and apicomplexan CTDs. This suggests the possibility that they are either a direct result of degeneracy or serve as early, primitive repeats. The ciliate *Stylonchia mytilus*, for instance, features regular, repeating SP motifs often preceded by tyrosine residues (e.g. YSP motifs). These observations suggest that that the SP motifs have evolved as specific sites for phosphorylation of RNA polymerase II. The absence of SP motifs or heptads does not necessarily imply absence of phosphorylation of the RNA polymerase II, however; the CTD of *T. brucei*, for instance, is phosphorylated (Chapman et al. 1994).

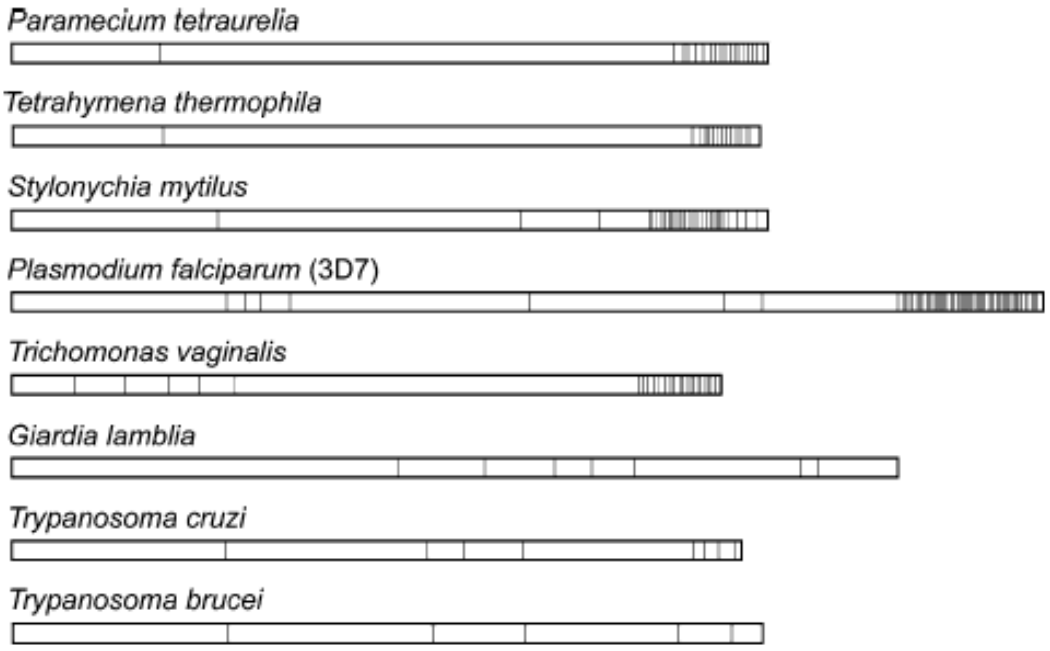

Figure 2.3. Enrichment of "SP" motifs in the tails of RNA Polymerase II from eight different protozoan species.

Serine-proline (SP) motifs that represent potential phosphorylation sites are marked with vertical lines **(|) .** Sequence accession numbers are: *Stylonchia mytilus*: AAK00313.1; *Paramecium tetraurelia*: CAI39063.1; *Tetrahymena thermophila*: GI accession: 118348890; *Trypanasoma brucei*: P17545; *Trypanasoma cruzi*: XP_812569; *Trichomonas vaginalis*: TVU20501; *Giardia lamblia*: XP_001704218

3. Discussion

The extreme plasticity that we observed for the CTD heptad repeats in *Plasmodium* has not been described for any other eukaryotic species. Moreover, the plasticity is not observed across any of the rodent parasites we studied. It is possible that the repetitive nature of the DNA sequence encoding the CTD simply leads to frequent duplications and deletions, and in this regard *Plasmodium* might have a propensity toward replication errors in the absence of purifying selection regulating CTD length. However, rather than displaying random variation in length and composition, the primate malarias appear to have a specific expansion of a particular heptad, leading to a longer CTD. Further, both phylogenetic analyses and codon usage bias support the hypothesis that the expansion of the CTD occurred twice in primate parasites – once in the line giving rise to *P. falciparum* and once in a line giving rise to *P. vivax*, and not any non-primate parasites. The fact that these expansion events appear to have occurred independently and in parallel suggests a strong functional basis for the longer CTD, although the nature of this selective force in primates has not been elucidated; there are no available data correlating CTD length in human parasites with virulence, drug resistance or any other obvious phenotype.

We could detect no variation in the CTDs of any of the rodent parasites, and their length was identical to that of *P. gallinaceum*, a parasite of birds, suggesting that the non-primate parasites might be fixed in their repeat numbers at eight heptads. In addition, the eight specific heptads found in bird and rodent parasites are also conserved in the CTDs of primate parasites (Figure 2.1), indicating that this represents the ancestral structure of the CTD prior to the expansion observed in primate parasites.

Interestingly, eight repeats is consistent with the minimal number of heptads required for viability in yeast and the minimal number needed in the mouse RNA polymerase II to form a secondary structure resembling a full length CTD. (Liao et al. 1991; Bienkiewicz et al. 2000) We therefore surmise that the eight repeats might be the minimal number required for a functional RNA polymerase II, and thus this represents the shortest CTD possible within this genus. However, in yeast the functional unit of the CTD was shown to be pairs of tandemly arranged heptad repeats, and that at least seven such "diheptads" were required for viability (Stiller et al. 2004). Thus, it appears that both the number and tandem arrangement of repeats is important for function. It is not yet clear which specific RNA polymerase functions require the tandem arrangement of the heptads, though the tandem arrangement is present in all eukaryotes from yeast to mammals. Notably, the rodent and bird parasites do not possess the minimal number of diheptad repeats required for viability in yeast, while the heptad expansion found in the CTD of primate parasites is specific for tandemly arranged repeats. This expansion greatly increases the number of diheptads (Figure 2A) resulting in CTDs that meet or exceed what is required in higher eukaryotes. Therefore, we speculate that the CTD expansion found in primate parasites results from a functional requirement shared with higher eukaryotes.

Interestingly, the expanded repeats all feature lysine in the seventh position. The expansion of lysine containing heptads in primate parasites highlights the potential significance of repeats that deviate from the consensus sequence YSPTSPS, particularly in the seventh position. The seventh position of the conserved heptad repeat displays the most heterogeneity in higher eukaryotes (Guo, Stiller 2005; Liu,

Greenleaf, Stiller 2008) and in yeast, the serine in this position is non-essential. (Stiller, McConaughy, Hall 2000). In mammals, while 20 of the first 25 heptads contain a serine in the seventh position, only 6 of the last 27 heptads do, and 8 of the last 17 heptads all contain a lysine in this position. In mice, however, these lysine-containing repeats are dispensable, with polymerases consisting of only consensus repeats supporting normal growth and viability of cells (Chapman, Conrad, Eick 2005). Thus the role of the lysine-enrichment in the distal end of the RNA polymerase II remains an outstanding question in CTD biology (Phatnani et al. 2006; Egloff, Murphy 2008). It has been hypothesized that lysine containing heptads might confer additional specificity in recruitment of specific proteins to the transcription complex (Phatnani, Greenleaf 2006). Naturally, this raises the question of why plasmodia species use lysine-containing repeats within their CTDs, particularly as our analysis indicates that the extensive reliance on lysine containing heptads is unique to plasmodia and not present in other organisms, including other apicomplexan parasites. In the case of *P. falciparum* the lysine (codon: AAA or AAG) enrichment could simply be a natural consequence of selective pressures of its A/T-rich (80%) genome. However, *P. vivax* also displays a similar number of lysine containing repeats but is nearly balanced in its A/T content (57%), implying functional significance for these heptads. In light of recent data concerning lysine modifications on non-histone targets involved in gene regulation, it is possible that the lysines in the CTD might be differentially acetylated, particularly by adjoining transcription factors (e.g.TAF1 subunit of TFIID) (Mizzen et al. 1996).

The study of RNA polymerase II CTDs in protozoans, particularly with the ever expanding number of genome sequences that are becoming available, is providing valuable insights into the evolution of this important protein domain. The fact that different species of *Plasmodium* naturally display variable CTDs, and that both *P. falciparum* and *P. berghei*, haploid single celled protists, are genetically tractable systems, means that it might be possible to use these organisms to test many of the current hypotheses regarding the roles of heptad repeats in CTD function. Thus *Plasmodium*, an organism of great importance to human health and economic development, could also serve as a model system for understanding the evolution of transcription in eukaryotic biology.

4. Methods

Amplification of Plasmodium CTDs

Genomic DNA for *P. falciparum, P. vivax and P. simium* strains were obtained from the Malaria Research and Reference Reagent Resource Center (MR4; www.mr4.org) and Dr. X. Su of the National Institute of Allergy and Infectious Disease. Genomic DNA from *P. ovale* was from New York University by J. Carlton and genomic DNA for isolates and sub-species of rodent malarias (*P. berghei, P. yoelii* and *P. vinckei*) were from stocks held at the University of Edinburgh by R. Carter (Perkins, Sarkar, Carter 2007)Oligonucleotide primers complementary to sequences flanking the first repeat in the CTD and the carboxy-terminus of the largest subunit of RNA polymerase II (*Rbp1*) gene were used to PCR-amplify the CTD. All PCR reactions were carried out on a PTC-2000® Peltier thermal cycler using *Taq* polymerase® (Invitrogen) under

the following conditions: 95 °C for 5 min followed by 35–38 cycles of 94 °C for 30 s, 56 °C for 60 s, 60 °C for 1 min, and a final extension step of 65 °C for 5 min. Reaction products were purified and directly sent for automated sequencing by the university core facility. The forward primer that was used to amplify the RNA polymerase II CTD from *P. falciparum* isolates was 5'-CCTAAACCTCAAATTAATCATAATATTTATTCA-3' and the reverse primer was 5'-CATATTTTCCTTCATTTCGTCCTCGTATAT-3'. The forward primer for *P. vivax* and *P. fragile* RNA polymerase II CTD was 5' – TCCCC(A/C)TT(C/T)TCTCC(A/T/C)TTTGAT-3' and the reverse primer was: 5'- CATTTCGTCCTC(C/G)TC(C/T)ATGTTGTA-3'

The forward primer for *P. berghei* RNA polymerase II CTD was 5'-CCAAAACCTCAGATGCAAAATAATATATATTCT-3' and the reverse primer was: 5'-TTATTCTTCCTGCATCTCCTCTTCATCCAT-3'; forward primer for *P. yoelii yoelii* RNA polymerase II CTD was 5'-CCAAAACCTCAGATGCAAAATAATATATATTCT-3' and the reverse primer was as above for *P. berghei*.

Tabulation of Repeat Numbers

Repeats were defined conventionally, as heptads featuring a serine in the second and fifth position (S_2 and S_5) (Liu, Greenleaf, Stiller 2008) and were counted manually. The lone exception to this rule was made for the repeat YAIASPK in the non-human primate *Plasmodium* species as it aligned with the repeat YSITSPK in rodent and human parasites. We note that this definition differs from recent work of Chapman et al. (Liu, Greenleaf, Stiller 2008)who include phasing in their definition. This group

identifies five repeats in *P. yoelii*; whereas we identify eight. Diheptads or heptad pairs refers to two heptad repeats found in tandem with no intervening amino acid residues. A single heptad can be counted as part of two diheptads. For example, three consecutive heptad repeats would constitute 2 diheptads.

Note: The work described in Chapter 2 has been published and can be found here:

Kishore, SP, SL Perkins, TJ Templeton, KW Deitsch. 2009. An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. Journal of molecular evolution 68:706-714.

Susan L. Perkins is a Curator at the American Museum of Natural History and provided technical assistance and samples. Thomas J. Templeton is at Weill Cornell Medical College and assisted with the computational analyses.

Data deposition footnote:

Sequences generated by this study: RNA Polymerase II CTD (regions R2 and R3)

	Accession Code:	**Source**
P. gallinaceum RNA polymerase II (whole)	EU840284	Sanger
P. berghei isolate ANKA	EU827171	Edinburgh
P. berghei isolate NK65	EU827172	Edinburgh
P. yoelii yoelii 17X	EU827173	Edinburgh
P. yoelii yoelii 33X	EU827174	Edinburgh
P. yoelii killicki	EU827175	Edinburgh
P. yoelii nigeriensis	EU827176	Edinburgh
*P. yoelii (*new subspecies)	EU827177	Edinburgh
P. vinckei VIBA CyO	EU827178	Edinburgh
P. vinckei VIBA Cy P1	EU827179	Edinburgh
P. vinckei lentum 194ZZ	EU827180	Edinburgh
P. vinckei (new subspecies)	EU827181	Edinburgh
P. reichinowi	EU850397	Sanger
P.vivax isolate ONG	EU840276	MRA-341G
P. vivax isolate Pakchong	EU840277	MRA-342G
P. vivax isolate Nicaragua	EU840278	MRA-343G
P. vivax isolate Panama	EU840283	MRA-340G
P. cynomolgi isolate Smithsonian	EU840279	MRA-351G
P. cynomolgi isolate Cambodian	EU840280	MRA-579G
P. cynomolgi isolate Bastianelli	EU840281	MRA-350G
P. fragile	EU840282	MRA-352G
P. falciparum isolate FCR3	EU827182	MRA-731
P. falciparum isolate D6	EU827183	MRA-285
P. falciparum isolate RO33	EU827184	MRA-200
P. falciparum isolate 7G8	EU827185	NIH
P. falciparum isolate HB3	EU827186	MRA-155
P. falciparum isolate Santa (St.) Lucia	EU827187	MRA-331
P. falciparum isolate FCC-2	EU827188	MRA-733
P. falciparum isolate D10	EU827189	NIH
P. falciparum siolate K1	EU827190	MRA-159
P. falciparum isolate W2	EU827191	MRA-157
P. falciparum isolate DD2	EU827192	MRA-156
P. ovale	EU887536	NYU

Edinburgh = R. Carter, University of Edinburgh
MRA = Malaria Research and Reference Reagent Resource Center (www.mr4.org)
NIH = National Institutes of Health
NYU = J. Carlton, New York University

CHAPTER 3:

ACQUISITION AND DELETION OF CHROMATIN MODIFIERS LINKED TO THE RNA POLYMERASE II CTD

1. Introduction

It has long been noted that patients infected with the most lethal of human malaria parasites, *Plasmodium falciparum*, experience successive waves of parasitemia (Miller, Good, Milon 1994). This is thought to result from the expression of one cytoadhesive protein at a time on the surface of the infected red blood cell (RBC) and the ability of the parasite to vary expression over the course of an infection to evade host immune responses. These cytoadhesive proteins are encoded by the *var* genes, members of a large and highly variable cluster of virulence genes that exhibit mutually exclusive expression. It is the single active *var* gene that determines the antigenic and virulence properties of the infected cells (Deitsch, Calderwood, Wellems 2001b; Deitsch 2005; Voss et al. 2006; Scherf, Lopez-Rubio, Riviere 2008). Once a *var* gene is activated it appears to remain transcriptionally active for several replicative cycles in a process that is thought to result from epigenetic memory mediated by chromatin marks on core histones. This process of constant expression of a single *var* gene product over several cell cycles facilitates the continuous and coordinated cytoadhesion of parasitized RBCs to blood vessels, a process that is most pronounced and pathogenic in the brain or placenta. In the face of host immune pressure, the "on" *var* gene product can be switched "off" and another *var* gene in turn activated, thereby avoiding clearance and leading to lengthy infections. This process represents a genetic switch underpinning antigenic variation in *P. falciparum*.

While the phenotypic changes resulting from alterations in *var* gene expression are well characterized, the molecular and evolutionary basis of the genetic switch controlling *var* gene expression remains unknown Other primate parasites (e.g. *P. vivax, P. knowlesi*) also appear to feature this genetic switch. We have previously demonstrated that active transcription by RNA polymerase II is required to preserve the epigenetic memory that maintains a single *var* gene in the active state through multiple rounds of cellular replication. In recent work in the laboratory, Dzikowski et al state "the act of transcription itself might help to reinforce the maintenance of the epigenetic marks that control cellular memory" (Dzikowski, Deitsch 2008).

These observations and insights sharply focused our attention on the *P.falciparum* RNA polymerase II and specifically the tail of the catalytic subunit (RBP1) that is responsible for the integration of diverse transcriptional events (initiation, elongation, termination, polyadenylation and splicing). We observed that the C-terminal domain (CTD) of RBP1 features an expanded motif in primate parasites when compared with rodent or avian parasites. This motif, composed of heptads (YSPTSPK) is otherwise only found in mammalian polymerases, (described in detail in Chapter 2), which is suggestive of the underappreciated host: parasite interactions underlying primate parasitism. As there are definitive data showing that chromatin modifying proteins directly associate with the RNA polymerase II CTD in other eukaryotic systems (Xiao et al. 2003) to help link transcription to chromatin modification and epigenetic memory, we theorized that the expanded CTD might have been accompanied by chromatin modifying factors that were CTD-binders themselves.

Here, I describe the identification of two chromatin modifiers found in primate parasites with expanded CTDs that are absent in rodent parasites with truncated CTDs. The modifiers reciprocally modify the Histone 3 Lysine 36 residue through the addition/removal of methyl groups during active transcription by RNA polymerase II. Consistent with this idea, one of the histone modifiers, a methyltransferase, is recruited to the RNA polymerase complex by directly binding to the CTD in budding yeast and a diverse array of eukaryotes. I study the expression of the methyltransferase and demethylase across the *P. falciparum* asexual cycle and provide conclusive evidence for the presence of the H3K36-triMe modification via the generation of new rabbit polyclonal antisera. This report resolves conflicting data and confusion as to whether the mark exists in *P. falciparum*. Finally, I provide insights on the natural history of the acquisition and deletion of histone modifiers in *Plasmodium* and related apicomplexa. I believe the findings have broad implications regarding the evolution of primate parasitism and may help shed light on the complex strategies employed by protozoa to mediate immune evasion through epigenetic memory.

2. Results

2.1 Identification of primate parasite specific epigenetic modifiers

Antigenic variation of malaria parasites has been tied to transcription and histone modifications, yet the mechanism and the molecular players involved remain incompletely characterized. In the work described in chapter 2 of this thesis, I found that primate parasites, with the exception of *P. ovale*, exhibit an expanded CTD of RNA polymerase II, featuring an unusual motif only otherwise found in mammals.

Considering that all species of *Plasmodium* that infect mammals have

remarkably similar lifecycles, including obligatory stages in the mosquito, vertebrate liver and RBCs, the function of the expanded CTD was not readily apparent. However, it is likely that the function of the additional heptad repeats will involve interactions with proteins that are recruited to the RNA polymerase complex, in particular transcription factors or chromatin modifiers. To search for candidate proteins that might interact with the expanded CTD, I used a collection of 202 transcription factors, including epigenetic modifiers, identified in *P. falciparum* and published by Bischoff et al in 2010, and did a comparative screen using orthologs in the primate parasites (*P. vivax*, *P. knowlesi*) and three rodent parasites (*P. berghei*, *P. chabaudi* and *P. yoelii*), all available via PlasmoDB.(Bischoff et al. 2010). I asked whether there were transcription factors specific to primate parasites and absent from the three rodent parasites (see Computational Screening in the Methods section of this chapter). I identified three proteins that were present in *P. vivax*, *P. knowlesi* and *P. falciparum* and absent from the rodent lineages, including a histone methyltransferase (PfSet2; MAL13P1.122) that has been shown to bind directly to the CTD in other organisms. This protein was accompanied by a demethylase (PfJmjC1: MAL8P1.111), suggesting the tandem evolution of a complementary pair of histone modifiers specific to primate parasites. The other gene product is a gametocyte-restricted zinc-finger transcription factor (PF11_0357).

The PfSet2 protein features 4 plant homeodomains (PHD) and a SET domain. Orthologs of Set2 feature a Set2-RBP1 interaction (SRI) domain that interfaces with the RNA polymerase II CTD (Li et al. 2002; Li et al. 2003; Xiao et al. 2003; Vojnic et al. 2006). Set2 associates with the RNA polymerase II complex as it transcribes

through the 3' end of genes and the resulting histone methylation marks are typically associated with active transcription (Li et al. 2003). Notably, truncations of the CTD below 12 repeats led to less recruitment of Set2 in yeast (and thus less histone methylation) (Xiao et al. 2003). All parasites studied to date that feature an expanded CTD have at minimum 12 heptads, though not are all in tandem. The PjJmjC1 protein features a JmjC1 domain and a JmjN domain that facilitate demethylation of tri-methylated H3K36 and H3K9. In other species, the JmjC1 proteins specifically demethylate Set2-methylated histones. I believe the JmjC1 protein is a member of this family (JDJM2) (Klose, Kallin, Zhang 2006a; Klose et al. 2006b). That there exists a tandem, complementary pair of histone modifiers that act on the same histone mark in primate parasites and that orthologs of one of the modifiers is recruited to the RNA polymerase II CTD itself motivated our interest in studying the expression profile and function of these genes *P. falciparum.*

2.2 Expression of the PfSet2 methyltransferase and PfJmjC1 demethylase through the P. falciparum asexual cycle

I first wondered when these genes were expressed during asexual replication. I hypothesized that Set2 would be expressed when *var* genes are actively transcribed (in the early ring stage of the asexual cycle) and that the JmjC1 gene product would not be expressed at the same time; therefore I speculated that the JmjC1 gene would be expressed in the late (schizont) stage of replication. *P. falciparum* features a 48-hour replicative cycle in the first 24 hours referred to as the ring or early trophozoite stage in which peak *var* gene transcription is observed. The remaining 24 hours consist of

the late trophozoite and schizont stages in which DNA synthesis and S-phase transpires.

P. falciparum parasite cultures were synchronized using percoll/sorbitol gradients to generate cultures that consisted exclusively of schizonts. Parasites were then collected 14 hours (ring stage) and 48 hours (schizont stage) later. Total RNA was collected, reverse transcribed into complementary DNA (cDNA). A quantitive real time polymerase chain reaction (qRT-PCR) was then performed using primers specific for Set2 and JmjC1 and the housekeeping gene arginine t-RNA synthetase (P61), for which I normalized the expression data. Our laboratory and others have identified P61 as a transcript that does not change significantly over the 48-hour replicative cycle (Salanti et al. 2003; Dzikowski et al. 2006; Frank et al. 2006). As shown in Figure 3.1, I found that PfSet2 expression is highest in the ring stage (2.67 +/- 0.065) versus late stage and coincides with known time points of highest *var* gene transcription. PfJmjC1 was highest in the schizont stage (4.54 +/-1.43) compared with the ring stage.

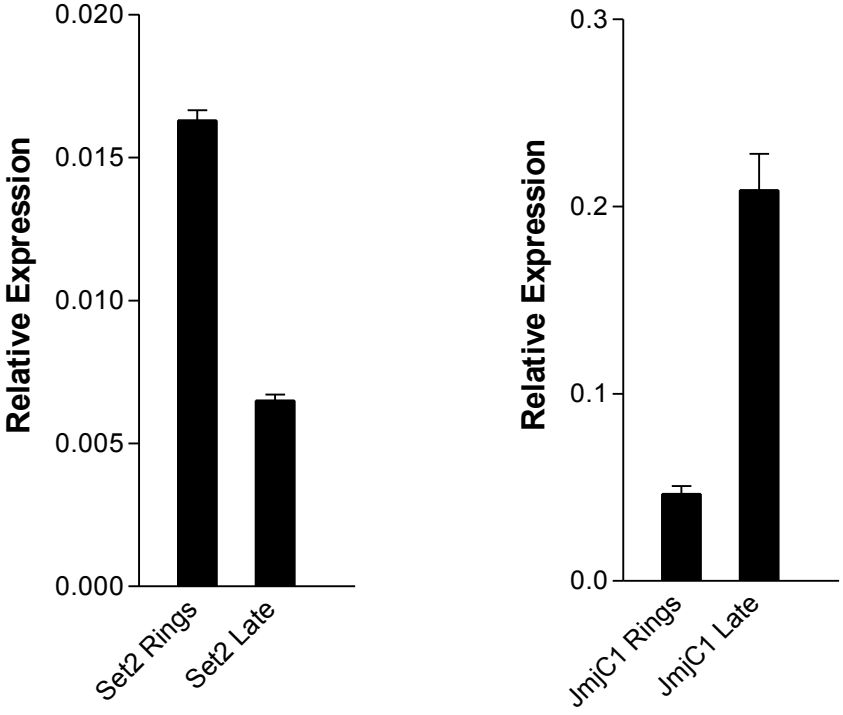

Figure 3.1. Complementary Expression of H3K36-triMe Methyltransferase and Demethylase across the *P. falciparum* Asexual Life Cycle
Steady-state mRNA levels expressed from PfSet2 and PfJmjC1 in the ring (early) and schizont (late) stage of P. falciparum analyzed via real-time PCR. Analysis of the levels of transcription are normalized to housekeeping gene (arginine tRNA-synthetase), indicated by the ratio (relative expression) on the Y-axis. The experiments were done in triplicate and replicated three times. As shown on the left panel, the analysis reveals that the PfSet2 expression is 2.67 +/- 0.065 fold higher in early versus late stage whereas as shown on the right panel PfJmjc1 expression is 4.54 +/- 1.43 fold higher in late versus ring stage.

2.3 Deposition of the H3K36-triMe across the P.falciparum asexual cycle

Previous published work has not definitively demonstrated whether the H3K36 mark is present in *P. falciparum*. Commercial antibodies are available, but these antibodies have been generated using yeast or mammalian histone H3 epitopes where there is considerable sequence divergence upstream of the K36 residue. Use of these commercial antibodies to probe Western blots of *P. falciparum* proteins has led to considerable confusion as whether the H3K6-triMe mark exists in *P. falciparum.* While one laboratory has claimed that a recombinant SET domain from PfSet2 trimethylates H3K36 in *P. falciparum* histones *in vitro*, other laboratories, using another commercial antibody and a proteomics approach, have suggested that the mark is absent (Cui, Fan, Miao 2008; Issar et al. 2009; Trelle et al. 2009).

To directly resolve this issue directly, I worked with collaborators to develop polyclonal antisera to an epitope spanning the K36 site with collaborators; the epitope included trimethylated H3K36 peptides (epitopes described in the Methods section) synthesized by Rockefeller University's proteomics core. The peptides were conjugated and injected in two rabbits via Covance using a 77-day immunization protocol (see Methods). An ELISA was performed against the target antigen using pre-immunization sera and the production bleed sera that were affinity purified against the target antigen. The antibody was highly sensitive, requiring a 1:3,600,000 dilution to achieve a 50% reduction in binding.

The specificity of the antiserum was tested by asking whether the polyclonal antibodies recognized unmethylated, monomethylated and dimethylated H3K36

peptides otherwise identical to the trimethylated peptide used to immunize the rabbits. The polyclonal antibody cocktail was first pre-absorbed with the un, mono and di-methylated peptides. 15 micrograms of each of the four purified peptides were deposited on nylon filter paper via slot blot and detected with the antibody solution. The results are shown in Figure 3.2. The antibody preparation detected only the trimethylated H3K36 peptides.

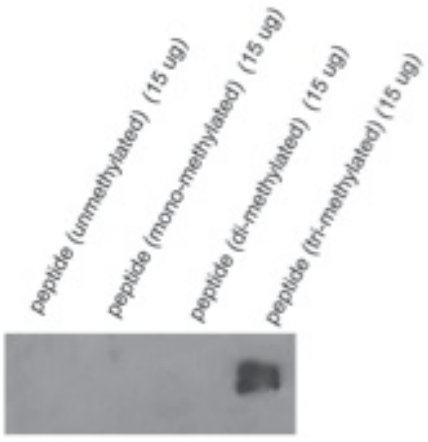

Figure 3.2. Verification of *P. falciparum* H3K36-triMe rabbit polyclonal antibody

The specificity of the antiserum was tested by asking whether the polyclonal antibodies recognized unmethylated, monomethylated and dimethylated H3K36 peptides otherwise identical to the trimethylated peptide used to immunize the rabbits. The polyclonal antibody antiserum was first preabsorbed with the un, mono and di-methylated peptides. 15 micrograms of each of the four purified peptides were deposited on nylon filter paper via slot blot and detected with the antibody solution. The antibody preparation detected only the trimethylated H3K36 peptides.

This antiserum was then utilized to determine whether the mark was detectable in *P. falciparum* and, if so, at which stage of the cycle. Protein extracts were obtained from early (ring) stages coinciding with *var* gene transcription (14 hours post invasion of parasites into erythrocytes), and with early trophozoite (20 hours post invasion) and late (schizont) stages before erythrocyte lysis.

As shown in Figure 3.3, western blots determined that the mark appears strongest in the early stage (14 hpi), begins to fade in early trophozoites (20 hpi) and is absent from schizonts (36 – 40 hpi). A histone H3 control is provided for comparison and shows that the total Histone H3 pool is high in the early stage, is lower in early trophozoites and stable in schizonts. This is consistent with data that show a depletion of histones in the trophozoite stage by Le Roch et al (Le Roch et al. 2004),

These data, showing that the mark is strongest in the early stage of the asexual cycle, are in agreement with the RT-PCR expression data. An interesting additional finding, one that I do not explore further here, is that there appears to be a second, additionally modified Histone H3 species (represented by the slightly higher band) in schizonts.

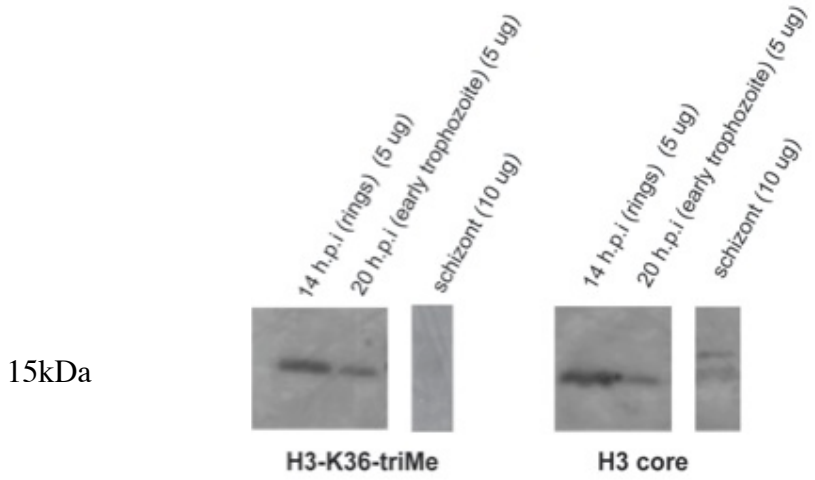

Figure 3.3. The H3K36-triMe mark across the *P. falciparum* asexual life cycle
An immunoblot showing the detection of the H3K36-triMe mark in early life cycle stages of *P. falciparum* (rings and early trophozoites). The mark did not appear strongly in schizonts (36-40 h.p.i) with the H3K36-triMe antibody, but total H3 was only faintly detected using an H3 core antibody. 5 micrograms of lysate of all stages were prepared as described in the Methods.

2.4 Evolution of Set2 and JmjC1 in Apicomplexa: acquisition and deletion

With the exception of the rodent plasmodia, all apicomplexan parasites, regardless of host, possess Set2. One ciliate (*Tetrahymena thermophila*) also possesses Set2, raising the possibility that Set2 was acquired by this phylogenetic lineage before the divergence of Apicomplexa from the ciliates.

Given the diversity of SET-domain-bearing proteins in Apicomplexa, I first performed an extended analysis to clearly define the Set2 and JmjC1 clade clearly in the tree of life. Aravind and colleagues have argued that a major family of transcription factors (AP-2) in Apicomplexa were laterally transferred from the red or green algal endosymbiont harbored intracellularly by all members of this group of parasites (Balaji et al. 2005). This notion led me to question whether the chromatin modifiers Set2 and JmjC1 may have also originated from such a lateral transfer event.

To examine this, I used the canonical SET and JmjC domains from the budding yeast to identify orthologs from a broad sampling of plant, animal and apicomplexan organisms across the tree of life including moss, red and green algae, diatoms, other Apicomplexa, humans, fission yeast, nematodes and fruit flies. These sequences were used to build a maximum-likelihood tree to help discriminate which proteins were orthologs of Set2 (see Figure A in the Appendix).

From this tree, I could now clearly define *Plasmodium falciparum* Set1, Set2 and Set3 clades from a broad array of eukaryotes across the tree of life. After discriminating which sequences were Set2 orthologs, including the Set2 ortholog from the ciliate *Tetrahymena thermophila,* I built a tree based on Set2 sequences only, as shown in Figure 3.4.

First, it appears that there is a correlation between the algal and apicomplexan Set2 clade. Apicomplexan Set2 paralogs do not group with the ostensible sister group, ciliates (represented by *Tetrahymena*). Second, in this tree, apicomplexan Set2 groups specifically with green paralogs that have C-terminal SET domains. The ciliate *Tetrahymena thermophila* Set2 exhibits has an N-terminal SET domain. Thus, this provides a tree that is also consistent with domain architecture. Finally, it appears that the apicomplexan Set2 groups in a clade that is distinct from the canonical Set2 clade (ciliates, humans, budding yeast); this implies that if a horizontal gene transfer occurred, it was of a gene that may not have had the same function as canonical Set2.

I next completed a similar analysis for the JmjC1 orthologs. In this tree, I also recover an apicomplexan/ green algal association for these proteins and I find that there is no suggestion that JmjC in apicomplexans is related to ciliates as shown in Figure 3.5.

These findings are difficult to interpret under the assumption that Set2 was inherited from a common ancestor of both ciliates and Apicomplexa. The algal/plant affinity of apicomplexan Set2 and JmjC1 raises the possibility of a horizontal gene transfer event. However, the SET domain is relatively short for phylogenetic inference across eukaryotic diversity and the low confidence values presented preclude us from making a definitive conclusion on the horizontal transfer at this time.

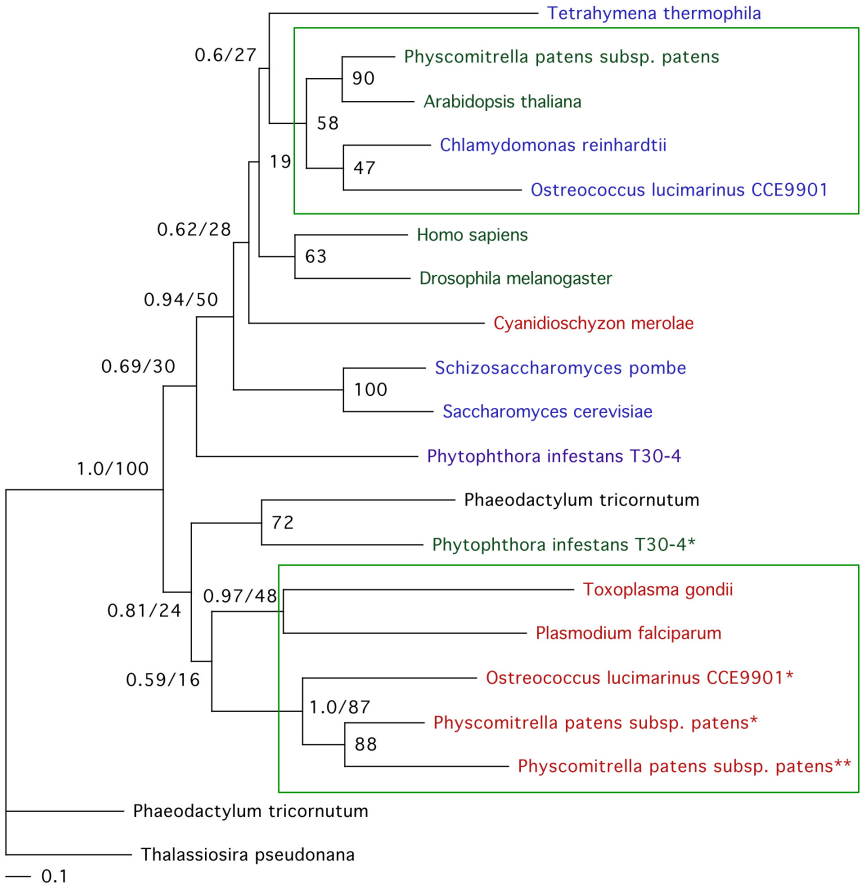

Figure 3.4. The Set 2 clade across the tree of life
The Set2 orthologs were analyzed by maximum likelihood analysis and Bayesian inference. Three observations are worth noting. First, it appears that there is a correlation between the algal and apicomplexan Set2 clade orthologs. The green boxes indicate the plant/algal groups. Apicomplexan Set2 paralogs do not group with the ostensible sister group, ciliates (represented by *Tetrahymena*). Second, in this tree, apicomplexan Set2 groups specifically with green paralogs that have C-terminal SET domains. The ciliate *Tetrahymena thermophila* Set2 exhibits has an N-terminal SET domain. Thus, this provides a tree that is also consistent with domain architecture. Finally, it appears that the apicomplexan Set2 groups in a clade that is distinct from the canonical Set2 clade (ciliates, humans, budding yeast); this implies that if a horizontal gene transfer occurred, it was of a gene that may not have had the same function as canonical Set2.

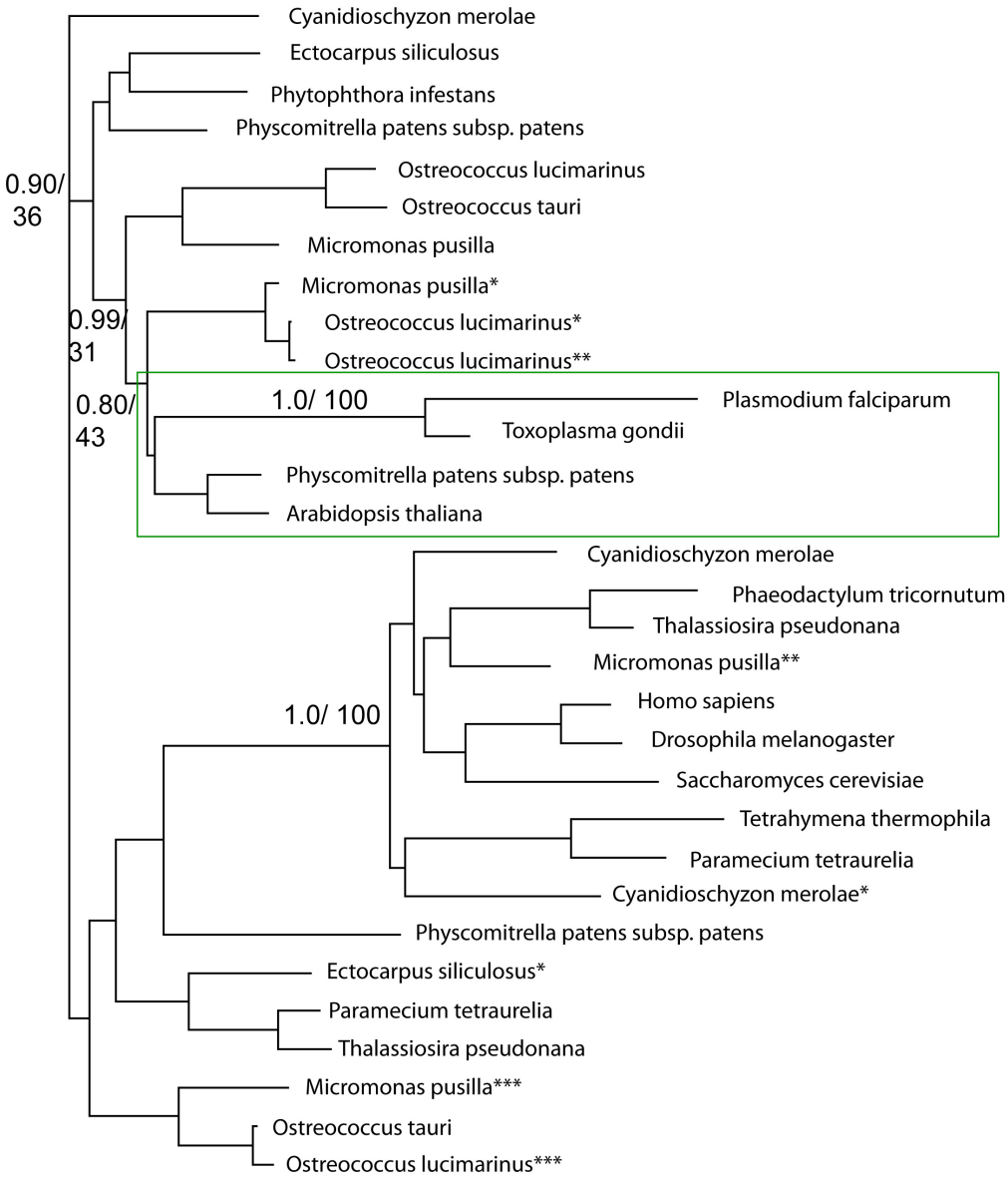

Figure 3.5. JmjC1 domains across the tree of life
Protein sequences for the JMJC domain were curated across the tree of life provides phylogenetic support for the grouping of apicomplexan JmjC1-bearing proteins within the plant clade. This is consistent with the grouping for apicomplexan Set2 proteins as shown in the previous figure for Set2 orthologs. However, the weak bootstrap values for the affinity with the plant clade does not make the tree conclusive in showing that the apicomplexan origin of JmjC1 was laterally transferred from plant or alga.

The absence of the Set2 and JmjC1 orthologs from rodent plasmodia but the presence of these proteins in primate malaria parasites raised the possibility that these proteins were lost in a lineage-specific fashion in rodent parasites. The chromosome maps on the publicly available *P. falciparum* genome database (PlasmoDB) display near perfect synteny in the regions surrounding both Set2 and JmjC1, providing additional support for the deletion of these two genes from the rodent parasite genomes (Figure 3.6). To verify the putative deletion, PCR was used to amplify across the syntenic regions of the genome of *P. berghei*. Amplification from the flanking genes yielded the predicted product and validated that neither Set2 nor JmjC1 exist within the predicted positions of the genome. Secondly, I designed degenerate primers to Set2 and PfJmjC1 sequences and queried whether these genes were present anywhere in the genome. The reactions yielded no product. Taken together, these data confirm the deletion and suggest that the genes are not present elsewhere in the genome. These data also suggest that Set2 and JmjC1 were lost from malaria rodent parasite lineages, which is a more parsimonious solution than multiple, independent acquisitions of the genes in various plasmodia lineages apart from the rodent parasites. Of note, Cryptosporidia possess Set2 but lack JmjC1 orthologs, raising the possibility of a gene-specific loss in Cryptosporidia. The loss of one, but not both of the H3K36 modifiers, may also suggest that Set2 and JmjC1 were not initially functionally paired together in early diverging Aapicomplexa.

Figure 3.6. The deletion of Set2 (above) and JmjC1 (below) from rodent malaria parasites

The gene maps on the publicly available *P. falciparum* genome database (PlasmoDB) verify the deletion. I first amplified a product using the flanking genes in *P. berghei* that are syntenic to the flanking genes in *P. falciparum*. Amplification from the flanking genes yields the predicted product. Secondly, I designed degenerate primers to Set2 and PfJmjC1 sequences and queried whether these genes were present anywhere in the genome. The reactions yielded no product. Taken together, these data confirm the deletion and suggest that the genes are not present elsewhere in the genome.

3. Discussion

The findings presented here may explain how *var* gene expression patterns are "remembered" between cell cycles. In higher eukaryotes, Set2 is a protein that is known to bind the RNA polymerase II CTD via the SRI domain during active transcription (Li, Moazed, Gygi 2002; Li et al. 2003), resulting in the deposition of H3K36-triMe modifications on the nucleosomes found near the 3' end of expressed genes. In *Plasmodium*, Set2 is only found in primate parasites and is always accompanied by a methylase (JmjC1) that reciprocally modifies the same residue. The observation that expanded CTDs, Set2 and JmjC1 are all exclusively found in primate parasites suggests that they may have evolved as components of a genetic switch to help reinforce epigenetic memory over time in parasites. This model is consistent with our laboratory's previous finding that epigenetic memory of virulence genes is maintained via active transcription by RNA polymerase II (Dzikowski, Deitsch 2008).

Recently, I started a collaboration with Ron Dzikowski who used the *P. falciparum* H3K36-triMe antibody to study where the H3K36-triMe mark localized across the parasite life cycle. As shown in Figure 3.7, in rings, as predicted by our RT-PCR and immunoblot data, there was a high degree of central nuclear staining. There appeared to be no staining in the trophozoite stage (not shown), consistent with evidence demonstrating that histones are selectively degraded in this stage (Le Roch et al. 2004). Surprisingly, however, we observed perinuclear association of the H3K36-triMe modification in the late (segmented) schizonts and in each of the early-forming

daughter cells (merozoites). In this result, the staining of the H3K36-triMe modification in this result is consistent with the staining of a GFP-tagged PfSet2 H3K36-triMe methyltransferase as recently shown by Volz et al during the late stage (Volz et al. 2010). This is notable, because as it is thought that both silenced and active *var* genes, in addition to genes involved in invasion, reside at the nuclear periphery. Potential *var* gene regulators and epigenetic modifiers are also thought to associate at the periphery (Scherf, Lopez-Rubio, Riviere 2008). Volz et al have reported that the Set2 protein co-localizes with the H3K9acetyl (H3K9ac) modifications, coinciding with the mark of active expression (Volz et al. 2010).

If the same biological principles hold in *P.falciparum* as with higher eukaryotes, I presume the H3K36-triMe is a byproduct of active transcription by RNA polymerase II during the segmented schizont stage. The key genes known to be activated in this temporal period and which are regulated by epigenetic memory appear to be those that participate in invasion (e.g. *RH4*, *clag3.1*, *clag3.2* and *eba-140*) (Jiang et al. 2010; Crowley et al. 2011). This raises the possibility that the H3K36-triMe modification may also be implicated in parasite invasion pathways.

Our early schizont samples exhibit reduced *Set2* message and a corresponding depletion of the H3K36-triMe modifications. Dzikowski's group, however, used late schizonts showing segmentation. The *Set2* message appears to increase during this very late period according to one time-course dataset (Winzeler) on the genome resource PlasmoDB. This resurgence of Set2 in late schizonts could explain the reason why I observed H3K36-triMe in late schizonts.

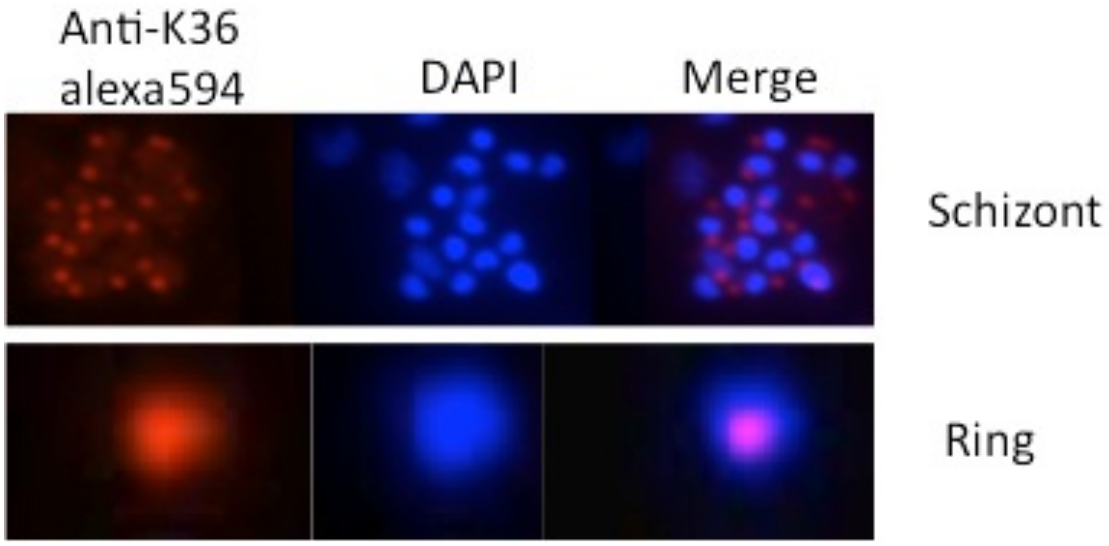

Figure 3.7. H3K36-triMe during the asexual *P.falciparum* life cycle

The H3K36-triMe modification was followed in late segmented schizonts (top panel) and early stage rings (bottom panel). The H3K36-triMe modification is detected by immunofluorescence (left panels; Alexa594 which emits orange-red); a DNA (nuclear stain) is provided by the (middle panels; DAPI which emits blue); and the merged view is provided by the right hand panels. As shown in the top panels, the H3K36-triMe stains the perinuclear regions in late stages (schizonts) (upper right hand) and in the center of the nucleus in early stages (rings) (lower right hand). Genes involved implicated in erythrocyte invasion and which are known to be subject to epigenetic memory are typically activated at this time in the life cycle in late segmented schizonts and display a similar pattern of staining (perinuclear). The trophozoite stage is not shown as no staining was observed. Histones are known to be degraded during the trophozoite stage which could explain this finding (Le Roch et al. 2004). This unpublished work was done in collaboration with Ron Dzikowski and Noa Dahan-Pasternak of Hebrew University.

I believe the identification of two complementary chromatin modifiers in *P. falciparum* and related primate parasites may suggest a pathway of epigenetic memory whereby once activated, a *var* gene is "remembered" over subsequent cell cycles through active transcription involving the deposition of trimethylated H3K36. To directly test this hypothesis, our laboratory will next characterize whether "on" but not "off" malaria virulence genes associate with the H3K36-triMe mark through chromatin immunoprecipitation experiments.

Cross talk with histone methyltransferases such as Set8, which deposits H4K20 methylations in mitotic chromosomes to faithfully transmit epigenetic information, might be one mechanism to accomplish this. In higher eukaryotes, H3K36triMe is the mark of active genes and serves as the reciprocal mark to the H3K9triMe, a mark of silencing. In *Plasmodium*, the H3K9-triMe is restricted to subtelomeric regions and to multi-copy gene families including the *var*, *rifin* and *stevor* families as well as to genes involved in parasite invasion (e.g. as above, *RH4, clag3.1, clag3.2, eba-140*) that are subject to epigenetic memory. I hypothesize that the H3K36-triMe similarly marks the same regions as H3K9-ac (Lopez-Rubio, Mancio-Silva, Scherf 2009).

Therefore, the identification of genes implicated in epigenetic memory and cytoadhesion in these protists, therefore, could relate to under-appreciated host: parasite interactions including new modes of immune evasion. The fact that Set2 is present but JmjC1 is absent in *Cryptosporidium* raises the possibility that Set2 and JmjC1 were not acting together initially. In *Plasmodium*, specifically, the failure of the rodent parasite CTD to expand the YSPTSPK heptad and the presumed requirement

for an expanded CTD for recruitment of Set2, would have rendered Set2 and JmjC1 superfluous. The recent independent expansions of the CTD in two different primate plasmodia (see Chapter 2) underscores the potential significance of accommodating Set2 for primate parasitism. The observation in budding yeast that at least 12 repeats are required for CTD adhesion to Set2 and corresponding methylation events is consistent with the idea that an expanded CTD provides binding space for Set2 and CTD binders (Xiao et al. 2003). Consistent with these observations, the deletion of both Set2 and JmjC1 in rodent plasmodia suggests tight and specific association between these chromatin modifiers in *Plasmodium;* this association is likely to reflect a recent integration into the genome, rather than ancient, conserved functions that would have been harder to lose. The absence of epigenetic memory at rodent parasite surface-exposed variant antigens (Cunningham et al. 2009) (best studied in the *yir* family *in P. yoelii*) suggests a functional consequence of this deletion.

With respect to the source of apicomplexan Set2 and JmjC1, green algal and plant affinities of these apicomplexan proteins suggests the source of a potential horizontal transfer in the ancestral apicomplexan. The fact that both ciliate Set2 and JmjC1 sequences do not group with Apicomplexa and, in the case of ciliate Set2, feature a different placement of the SET domain (N-terminal instead of C-terminal) challenges the assumption that Set2 was vertically inherited from a common ancestor of Apicomplexa and ciliates. The postulation that the major apicomplexan transcription factors (AP-2) were derived from the algal endosymbiont raises interesting questions about the adaptation of transcriptional and epigenetic machinery acquired via horizontal transfer (Balaji et al. 2005). Still, a definitive analysis is

lacking as the weak bootstrap values and the possibility of gene duplication of SET domains complicate the picture in determining vertical versus lateral transfer.

In this chapter, I examined the expression of a pair of genes encoding a histone methyltransferases and de-methyltransfearse that, together, are predicted to modulate the deposition of the H3K36-triMe mark on nucleosomes in *P. falciparum*. In addition, I utilized a new polyclonal antibody to detect the presence of the predicted mark in cultured parasites. The resulting data have broad evolutionary implications concerning the acquisition of a tandem, complementary pair of histone modifiers that associate with the CTD of RNA polymerase II. These data also resolve the confusion as to whether the H3K36-triMe mark exists and I suggest that H3K36-triMe is an as yet under-appreciated modification in primate parasites and is particularly notable due to its association with RNA polymerase II. I believe the findings have broad implications regarding the evolution of primate parasitism and help define the complex strategies employed by protozoa to mediate immune evasion through epigenetic memory.

4. Methods

Real-Time Polymerase Chain Reaction (RT-PCR)

RNA extraction and realtime RT-PCR for assaying expression of the var gene family. RNA was extracted from synchronized ring stage parasites 14 h post-invasion. RNA extraction was performed with the TRIZOL LS Reagent (Invitrogen). RNA to be used for cDNA synthesis was purified on a PureLink column (Invitrogen) according to manufacturer's protocol. Isolated RNA was then treated with Deoxyribonuclease I

(Invitrogen) to degrade contaminating genomic (gDNA). cDNA synthesis was performed with Super- script II Rnase H reverse transcriptase (Invitrogen) with random primers (Invitrogen) as described by the manufacturer. cDNA was synthesized from 800 ng total RNA in a reaction volume of 40 ll. For each cDNA synthesis reaction, a control reaction without reverse transcriptase was performed with identical amounts of template. For realtime quantitative RT-PCR reactions to detect transcription from all var genes present in the 3D7 genome, I employed the primer set for PfSet2 Forward 5' GGGAAAATGCAAATTGGAA 3' and Reverse 5' TTTCCAAAACTCAGCTGAAACA 3' and for PfJmjC1
Forward 5' TGGGGTCAACCCAAAATATG 3' and
Reverse 5' AAGGGCATGGAAAATCTCCT 3. The housekeeping control was P61 Forward (5' TGTACCACCAGCCTTACCAG 3') and P61 Reverse (5' TTCCTTGCCATGGTTCAAT 3') arginine-tRNA synthetase. Amplification efficiency was verified by performing amplifications using different concentrations of genomic DNA as templates. Reactions were performed at a final primer concentration of 0.5 lM using Biorad ITAQ SYBR green Supermix in 20-ll reactions on an ABI Prism 7900HT. All runs were done in triplicate and yielded virtually identical Ct (cycle threshold) values. The delta Ct for each individual primer pair was determined by substracting the measured Ct value from the Ct value of the control seryl-tRNA synthetase (User bulletin 2, Applied Biosystems, http://www.appliedbiosystems.com). Delta Cts were then converted to relative copy numbers with the formula 2^delta Ct.

Antibody Production

Unmethylated, mono, di and trimethylated peptides that include and flank the H3K36 residue were synthesized and conjugated at Rockefeller University using the following epitope: RKSAPISAGI-K(Me0.1,2,3)-KPHRYRPGT. Two rabbits were immunized with the peptide described above. We developed a rabbit polyclonal antibody via a 77-day immunization protocol (Covance) to trimethylated *P. falciparum* H3K36 and assayed for presence of the histone mark through the cell cycle. Briefly, the primary immunization was done on day 1 with 250 micrograms of antigen, followed by a boost on day 21, 42 and 63 with 125 micrograms of antigen each. The ELISA Titers were completed on days 74 and the terminal bleed was completed on day 77. I established that the peptides were specific through peptide competition assays and verified specificity through assaying detection of 15 micrograms of purified unmethylated, monomethylated, dimethylated and tri-methylated peptides through dot-blot assays.

Immunodetection

A total 2x 10^8 cells were lysed in RIPA buffer (Sigma; 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% NP-40, 0.25% sodiumdeoxycholate, 0.1% SDS), plus mammalian proteinase inhibitor cocktail (Sigma) supplemented with 200 mg ml-1 PMSF and 4 milligrams ml-1 pepstatin. 1 – 15 micrograms of lysates were separated on a 15% SDS-PAGE gel and transferred onto a nitrocellulose membrane. Membranes were blocked overnight with 5% milk. Antibody specificity was confirmed by dot blot and peptide competition assays. Primary antibodies were detected with horseradish-peroxidase-conjugated sheep anti-rabbit antibodies.

Phylogenetic Analyses

The queries from canonical orthologs of SET and JmjC domains in the Set2 and JmjC1 from budding yeast below were used to search red, brown and green algae, moss, stramenophiles, oomycetes and the animals, fruit flies, nematodes, rodents and human genomes for orthologs. All the proteins bearing the domain were curated and then analyzed via maximum likelihood analysis (Prof. John Stiller, East Carolina University).

Blast Queries:
SET2
REYIGKCASFCILNHHLKLISQLIRAESSNSEAEDLVDTERMNGMINCGENCW
NRAVCTECCDLSCRCGELCQNRRFQKHQDACVYPVPTRGKGWGLCAGQFIP
KGTFIIQYTGEVFDINSSEGIKRCKDYSRSTCTYLMKIDRNEVIDPTYKGNLARF
INHSCDPNCITQKWHVLGEICIGIFSIKDIQEDEELTFDYQFDSFKTPLTKCLCQ
AAKCKGYLGYIPTDFTVEEWEERLDNLPCSICDGNTEDDDDKLLLCDRCNNG
FHIFCLKPPLTEIPEEQWFCADCINQMNNVNHEKIALEKKERLKKKKSKNITLE
SNDEEIFKYSEQYEQFYSFMKNIENQAIEE

JMJC1
RKGSVSKSTKLKLKNFESSFNIDDFEQFRTEYTIDLSDFQNTERLKFLEEYYWK
TLNFTTPMYGADTPGSIFPEGLNVWNVAKLPNILDHMETKVPGVNDSYLYAG
LWKASFSWHLEDQDLYSINYIHFGAPKQWYSIPQEDRFKFYKFMQEQFPEEA
KNCPEFLRHKMFLASPKLLQENGIRCNEIVHHEGEFMITYPYGYHAGFNYGYN
LAESVNFALEEWLPIGKKAGKCHCISDSVEIDVKKLAKSWRDNNKESKGTPPL
NQLPNPAMPLLHRPTLKEMESSSLRS

Sequences were trimmed by hand to include only SET domains and nearby flanking regions. The sequences were They were aligned using ClustalX and Muscle programs. (Jeanmougin et al. 1998) (Edgar 2004). Alignments were analyzed via maximum-likelihood phylogenetic reconstruction using the program PhyML (Guindon et al. 2003) The following parameters were used in ML analyses: an LG substitution matrix and gamma (4 categories) + invariable site model for rate variation, with the α

parameter and the fraction of invariable sites calculated from the data. One hundred non-parametric bootstrap replicates were preformed to provide relative support values for membership of each sequence in a given protein subfamily.

CHAPTER 4:

HORIZONTAL GENE TRANSFER AND THE EVOLUTION OF PARASITISM IN *PLASMODIUM FALCIPARUM* AND OTHER APICOMPLEXANS

1. Introduction:

The process whereby -living or symbiotic organisms made the transition to full-fledged obligate, intracellular parasites in evolutionary history remains unclear. This is perhaps best exemplified by the case of the Apicomplexa, a group of 5000 species that includes several major human disease-causing agents such as *Plasmodium falciparum*, the most lethal of human malaria parasites (Cavalier-Smith 1993). Presumably, this transition involved the development of novel cellular differentiation pathways for invading different hosts, invasion schema and immune evasion strategies including antigenic variation. The newly acquired lifestyle complexity would likely also require the acquisition of new mechanisms to control gene expression. For example, more sophisticated transcriptional regulation and epigenetic machinery would enable the evolution of complex life cycles involving multiple hosts and stages, and drive developmental changes accompanying the transition to parasitism. This could involve either de novo innovations or, as has proposed earlier, machinery could be acquired by horizontal transfer and adapted to enable and enhance parasitism. In the case of Apicomplexa, two major potential sources for gene transfer are established: i) one in which the relic of an algal endosymbiont transfers machinery (fatty acid synthases or AP-2 transcription factors) (Gardner et al. 2002; Balaji et al. 2005) or ii) an animal host transfers machinery (domains involved in cytoadhesion and O-linked

glycosylation) (Templeton et al. 2004; Templeton 2007). Here, we study the putative acquisition of epigenetic machinery in the ancestor of all apicomplexans with a focus on histone lysine modifiers, which are central to pathways of cellular differentiation, cellular invasion and immune evasion.

Histone lysine methyltransferases, characterized by a SET domain, play a fundamental role in gene activation and epigenetic regulation across all eukaryotes. These domains modify histone lysine residues at Histone 3 Lysine 4, 9, 36, and Histone 4 Lysine 20. These modifications are crucial for the establishment and maintenance of epigenetic memory, including in *P. falciparum,* and are involved in imprinting invasion and immune evasion genes (Chookajorn et al. 2007b; Dzikowski, Deitsch 2008; Scherf, Lopez-Rubio, Riviere 2008; Lopez-Rubio, Mancio-Silva, Scherf 2009).

Of these SET-bearing domain modifiers, the epigenetic modifier Set8 is known to participate in mitosis and is thought to facilitate the transmission of heterochromatic marks through the cell cycle across eukaryotic life, including in the Apicomplexa (Sautel et al. 2007). In recent work, Saulet et al. sampled a small set of animal sources and found strong homology to apicomplexan Set8 (Sautel et al. 2007). The study, however, did not explicitly address the likelihood of horizontal gene transfer rigorously or, if the event transpired, when in the course of apicomplexan evolution the transfer likely occurred.

Here, using a wide array of SET-domains retrieved from eukaryotes across the tree of life, I provide phylogenetic evidence that the apicomplexan clade groups with the animal Set8 clade with very strong support, substantiating the hypothesis of a

lateral transfer event from an animal host. This grouping has strong statistical support and is contrary to what is expected from established phylogenies based on housekeeping proteins. Moreover, apicomplexan Set8 shows no phylogenetic affinity toward sequences from other organisms studied, including plants and algae, thereby ruling out lateral transfer from the secondary endosymbiont (the plastid). The acquisition of Set8 instead appears to have occurred at approximately the same stage of apicomplexan evolutionary history as the transfer of the extracellular adhesion and O-linked glycosylation genes.

These results suggest that during the transition to parasitism, in addition to acquiring the adhesion and glycosylation capabilities that underlie many of the intercellular interactions required for invasion and survival within their hosts, apicomplexan parasites also acquired further additional transcriptional regulation capabilities, in particular the epigenetic control utilized in cellular differentiation pathways and immune evasion. Moreover, both of these acquisitions appear to have occurred through horizontal gene transfer events at a similar time in evolutionary history, marking a key point in the transition of apicomplexans from free-living organisms to a parasitic lifestyle.

2. Results

2.1 Apicomplexan Set8 is derived from animal Set8

Using a canonical SET domain from homo sapiens Set8 (PR-7) as the query, I searched for the closest sequences to the Set8 domain across all major eukaryotic phyla including Apicomplexa, dinoflagellates, ciliates, gregarines, diplomonads, rhodophytes, plants, Microsporidia/Encephalitzoa, Kinetoplastids, *Entamoeba,*

heterokonts/stramenopiles, mycetes (*Dictyostelium*), and various metazoan taxa including sponges, nematodes, trematodes, arthropods, trochozoans, echinoderms, birds, reptiles and primates. Figure 4.1A displays an expected tree of eukaryotic life derived from combined maximum-likelihood and Bayesian inference using the largest subunit of RNA polymerase II (RPB1). This tree shows the generally accepted eukaryotic relationships and the great evolutionary distance between apicomplexans and the animal kingdom.

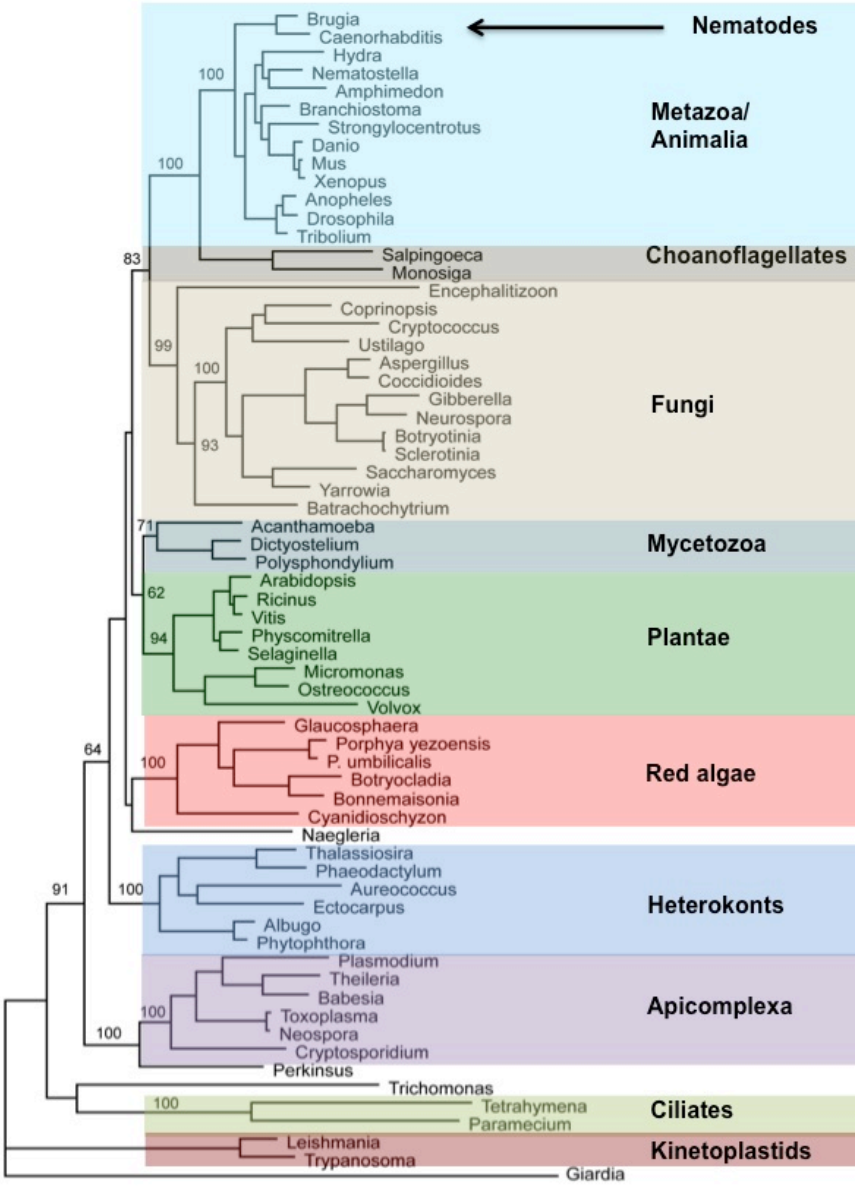

Figure 4.1A. Tree of eukaryotic life based on the largest subunit of RNA Polymerase II (RBP1)

Metazoa (in light blue) inclusive of nematodes, are very distantly related to Apicomplexa (in purple) at the base of the tree. Organisms from selected taxa across the tree of life are shown. Members of these taxa were used for an analysis of Set8 phylogeny across the eukaryotic tree of life (Figure 4.1B). A similar figure is shown in Chapter 1 (Figure 1.1). The tree, based on maximum likelihood, was constructed by John Stiller of East Carolina University (details of construction are shown in Chapter 4, Methods).

Comparable analyses were performed using Set8 orthologs across the tree of life with the results shown in Figure 4.1B. I find that the position of the apicomplexan Set8 does not match the expected tree. Namely, instead of grouping with closely related protozoa such as ciliates or heterokonts, apicomplexan Set8 sequences nest strongly within the animal Set8 clade. Consistent with this, none of the other kingdoms or taxa show high affinity for apicomplexan Set8, including other protists, plants and heterokonts. Moreover, the other SET-domain-bearing proteins included in the upper clade in Figure 4.1B do not appear to be Set8. *Tetrahymena* proteins appear in this clade, for instance; *Tetrahymena* is known to lack Set8 (Sautel et al. 2007).

In the survey of eukaryotes, based on the partial genomes known to date, I failed to identify Set8 in organisms closely related to the apicomplexans, namely the dinoflagellates, ciliates, colpodellids, gregarines and *Chromera velia*, a photosynthetic autotroph discovered recently. In addition, the sequence appears to be missing from other protists including kinetoplastids, *Trichomonas* and *Giardia* protists. Set8 appears to be missing in many fungi (including budding yeast) as well, and also in certain mycetozoa, including *Acanthamoeba*. This point suggests that Set8 evolved in animals and then was transferred to apicomplexans rather than lost independently and repeatedly across all the intervening taxa between apicomplexans and animals.

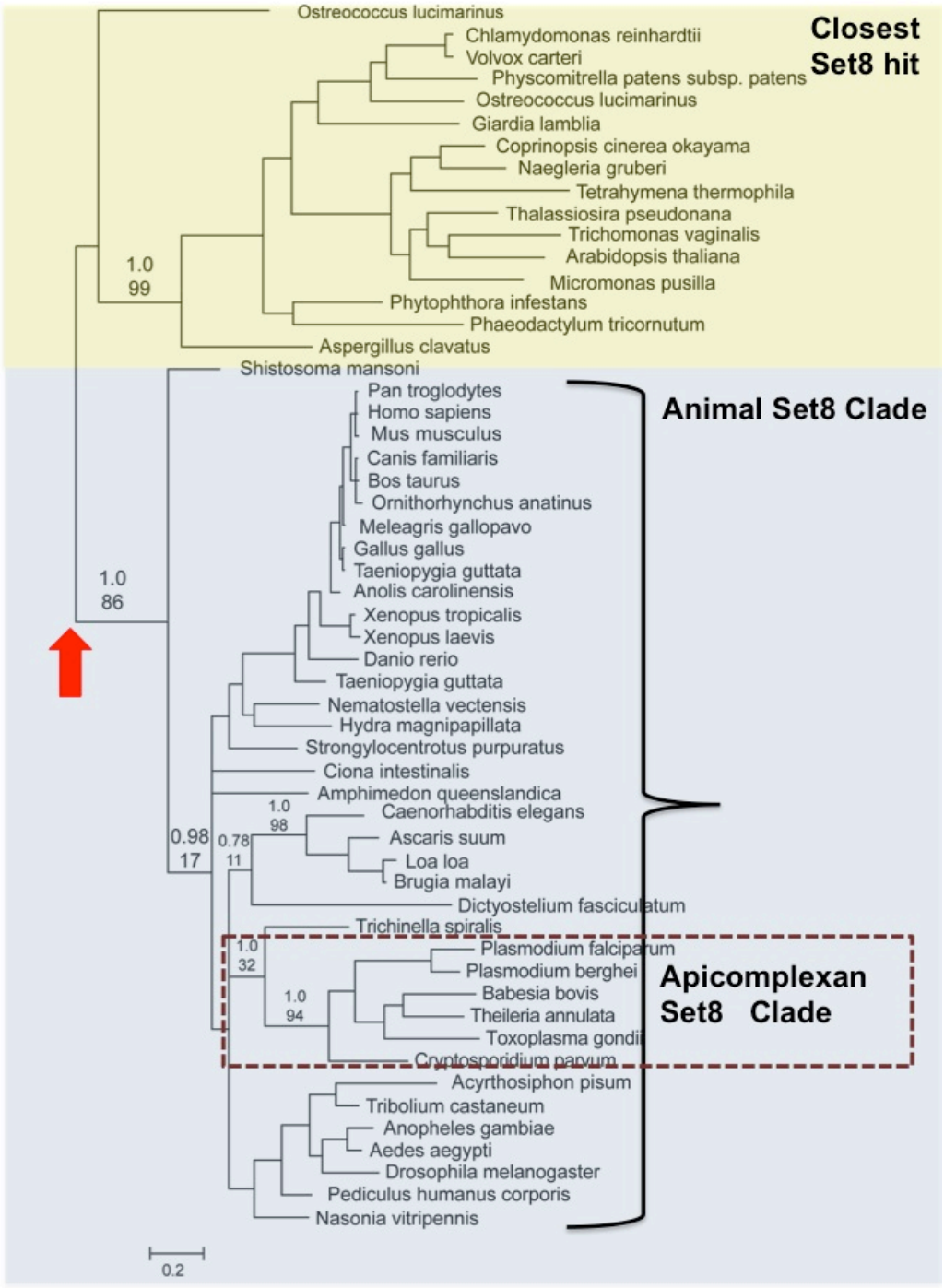
Ostreococcus lucimarinus
Chlamydomonas reinhardtii
Volvox carteri
Physcomitrella patens subsp. patens
Ostreococcus lucimarinus
Giardia lamblia
Coprinopsis cinerea okayama
Naegleria gruberi
Tetrahymena thermophila
Thalassiosira pseudonana
Trichomonas vaginalis
Arabidopsis thaliana
Micromonas pusilla
Phytophthora infestans
Phaeodactylum tricornutum
Aspergillus clavatus
Closest Set8 hit
1.0
99
Shistosoma mansoni
Pan troglodytes
Homo sapiens
Mus musculus
Canis familiaris
Bos taurus
Ornithorhynchus anatinus
Meleagris gallopavo
Gallus gallus
Taeniopygia guttata
Anolis carolinensis
Xenopus tropicalis
Xenopus laevis
Danio rerio
Taeniopygia guttata
Nematostella vectensis
Hydra magnipapillata
Strongylocentrotus purpuratus
Ciona intestinalis
Amphimedon queenslandica
Caenorhabditis elegans
Ascaris suum
Loa loa
Brugia malayi
Dictyostelium fasciculatum
Trichinella spiralis
Plasmodium falciparum
Plasmodium berghei
Babesia bovis
Theileria annulata
Toxoplasma gondii
Cryptosporidium parvum
Acyrthosiphon pisum
Tribolium castaneum
Anopheles gambiae
Aedes aegypti
Drosophila melanogaster
Pediculus humanus corporis
Nasonia vitripennis
Animal Set8 Clade
Apicomplexan Set8 Clade
1.0
86
1.0
98
0.98
17
0.78
11
1.0
32
1.0
94
0.2

Figure 4.1B. Tree of life with Set8 orthologs curated across the tree of life

This tree based on Set8 shows a phylogenetic incongruence with respect to the tree with RPB1 in Figure 4.1A. Shown in grey is the animal clade of Set8 denoted by a strongly supported node (red arrow). Apicomplexan Set8 groups within the metazoan/animal clade with strong posterior probability support that is unexpected given the phylogeny shown in Figure 4.1A. Set8 is absent in other lower eukaryotes based on current genome sequences. Where the closest Set8 hits were found, the proteins group together in the yellow shaded box (with strong support). These are likely not Set8 as *Tetrahymena*, a representative ciliate known to lack Set8, is featured in this clade. Based on these observations, I conclude that the apicomplexan Set8 is derived from animals. A horizontal gene transfer provides the simplest, most parsimonious explanation.

Per Figure 4.1B, the support for the apicomplexan Set8 grouping with animals is very strong (Bayesian probability of 1.0 and maximum-likelihood bootstrap of 86%), while the closest sequences to Set8 in other species listed above grouped out in a separate clade altogether (also with very strong support). Of note, the slime mold *Dictyostelium* also groups in the animal Set8 clade. As other members of the Ameobozoa phylum (e.g. *Entamoeba* and *Acanthamoeba*) lack Set8, this provides evidence of a similar horizontal transfer to slime molds. Given the vast evolutionary distance between animals and Apicomplexa/slime molds, and the absence of Set8 in virtually other all eukaryotic lineages strongly supports an an animal derivation of apicomplexan Set8. The most parsimonious solution is that of a single acquisition in ancestral Apicomplexa rather than the invention of Set8 in the common ancestor of Apicomplexa and animals and the subsequent loss from all other eukaryotes except slime molds.

2.2 A nematode appears to be the source of the host transfer to an ancient apicomplexan over 400 million years ago.

The finding that apicomplexan Set8 is likely of animal origin raises the question of where the animal sequence came from and when. The time of the transfer appears to be after the divergence of dinoflagellates (480 million years ago) and the colopodellids (423 million years ago) (Okamato 2008). The complete genome for several ciliates is now available (e.g. *Paramecium*, *Tetrahymena*). and I am confident that Set8 is absent; I also could not find Set8 among any of the current partial genome

sequences from the other taxa.

However, to explore which specific animal taxon was the likely source of the gene, I used Set8 sequences from extant apicomplexans (*P. falciparum, T. annulata, T.gondii, C. parvum* and *B. bovis*) in phylogenetic analyses involving both maximum-likelihood and Bayesian inference, with the animal and *Dictyostelium* sequences. To avoid misleading associations, I discarded the SET sequences that grouped outside the animal Set8 clade in Figure 4.1B. I recognized that the highly divergent SET sequences from other eukaryotes could be causing "phylogenetic artifacts" within the animal Set8 clade. By removing the more divergent out-group sequences, I am more likely to recover relationships within the in-group accurately

As shown in Figure 4.2, I consistently observed that apicomplexan sequences branch within the nematodes, and specifically with Set8 from *Trichonella spiralis*. The maximum-likelihood bootstrap support is weak (32%) but the Bayesian posterior probability is strong (0.99). Importantly, this tree recovers the same relationship between nematodes and Apicomplexa as did the expanded analysis, indicating this topology is stable regardless of the taxa sampled. Based on this consistent association, it appears that the nematode association is significant. In particular, the specific association with *Trichinella spiralis* should be noted in both trees.

Furthermore, the fact that both the slime mold *Dictyostelium* and the Apicomplexa both branch with nematodes, but separately, is telling. If the positions of the *Dictyostelium* and apicomplexan sequences were classic long-branch attraction artifacts, I would expect them to attract each other as the two most divergent branches of the tree. Moreover, neither is attracted to the base of the nematode clade, but rather

to individual sub-clades within the nematodes. Thus, although statistical support for a specific relationship between apicomplexan and nematode sequences is not strong, there appears to be basis for concluding it is not a phylogenetic artifact.

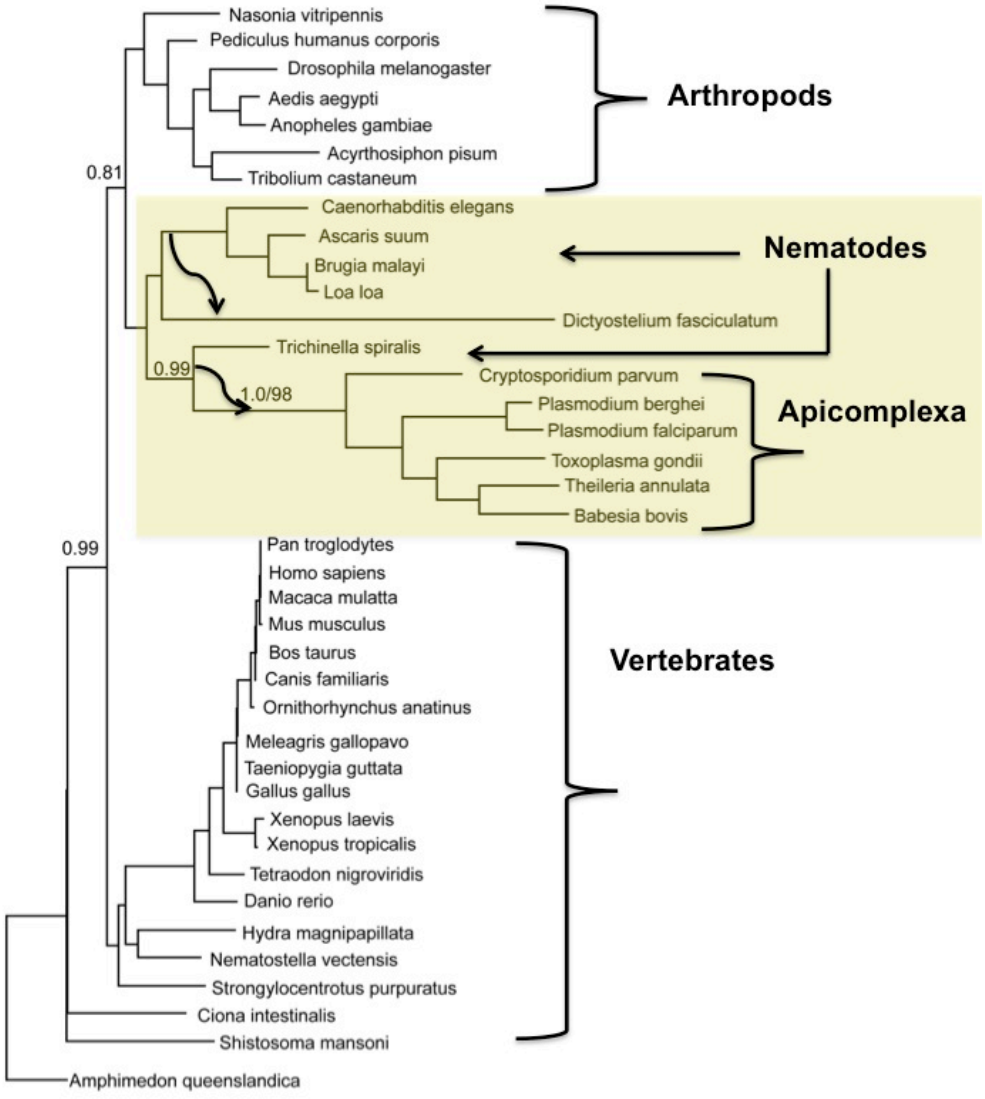

Figure 4.2. A deeper look at Set8: Apicomplexa, animals and slime molds

A phylogenetic analysis with only animal, apicomplexan and slime mold Set8 was next performed to better resolve the relationships between these taxa and to query potential sources of the Set 8 horizontal gene transfer. I find that Apicomplexa, nematodes and arthropods strongly group together (0.81 posterior probability score) in a sub-clade apart from vertebrates. Set8 orthologs from an animal, the parasitic nematode Trichinella spiralis, groups with strong posterior probability to the Apicomplexans. Similarly, soil nematodes group with the soil-based slime mold Dictosylium, implying a horizontal gene transfer event in this taxon as well. These proposed events are denoted by the black arrows. As all apicomplexans studied feature Set8, the most likely scenario is a horizontal gene transfer to the ancestral apicomplexan that existed circa 400 million years ago.

Therefore, I next examined whether the horizontal transfer of a nematode derived Set8 to slime molds and an ancestral apicomplexan is biologically plausible. As shown in Figure 4.3, I first observed that nematodes, as well as arthropods were extant at the time of ancestral apicomplexa. While arthropods are known definitive hosts for many apicomplexans today, apicomplexan parasitism of nematodes has not been documented (or perhaps is under-studied). While *Trichenella* is a modern parasite of mammals, the likelihood that this nematode or its ancestors transferred Set8 to an ancestral apicomplexan during a mammalian infection is highly unlikely as mammals had not appeared at this time. It is thus more likely that the ancestral apicomplexan parasitized nematodes and acquired Set8 in a gene transfer. The observation that extant malaria parasites are capable of internalizing and incorporating exogenous animal DNA spontaneously in animal cells is consistent with the potential mechanism of a transfer (Deitsch et al. 2001c; Templeton et al. 2004). As Figure 4.3 shows, of special note is the observation that two other instances of horizontal gene transfer from animals are theorized to have occurred at the same evolutionary moment (into ancestral apicomplexans). Domains involved in cytoadhesion and invasion as well as O-linked glycosylation are proposed to have also been transferred at this time. The animal source of this transfer remains unidentified (Anantharaman et al. 2007; Templeton 2007).

Meanwhile, some of the soil-based nematodes (e.g. *C. elegans*) could have transferred Set8 to the immediate ancestor of social amoebae and slime molds. Set8 orthologs are found in other slime molds (e.g. *Polysphondylium pallidum)* but are missing in other amoebas. Thus, the most parsimonious solution is an animal-derived

horizontal gene transfer event specific to slime molds. *C. elegans* is known to feed on *Dictyostelium,* ingesting *Dictyostelium* spores and dispersing them (Kessin et al. 1996). This suggests that the incorporation of nematode epigenetic machinery into slime mold genome is biologically plausible. Moreover, slime molds are known to adapt to strong predation pressure exerted by nematodes. This adaptation could provide a source of strong positive selection on enhanced control of gene expression in slime molds (Kessin et al. 1996). The slime molds are also well known to scavenge bacteria from the soil and recent reports indicate that bacteria are consumed by slime molds and incorporated into slime mold fruiting bodies and spores during asexual reproduction (Brock et al. 2011). There are 16 known cases of horizontal gene transfer from bacterial sources into *Dictyostelium* (Eichinger et al. 2005; Sucgang et al. 2011).

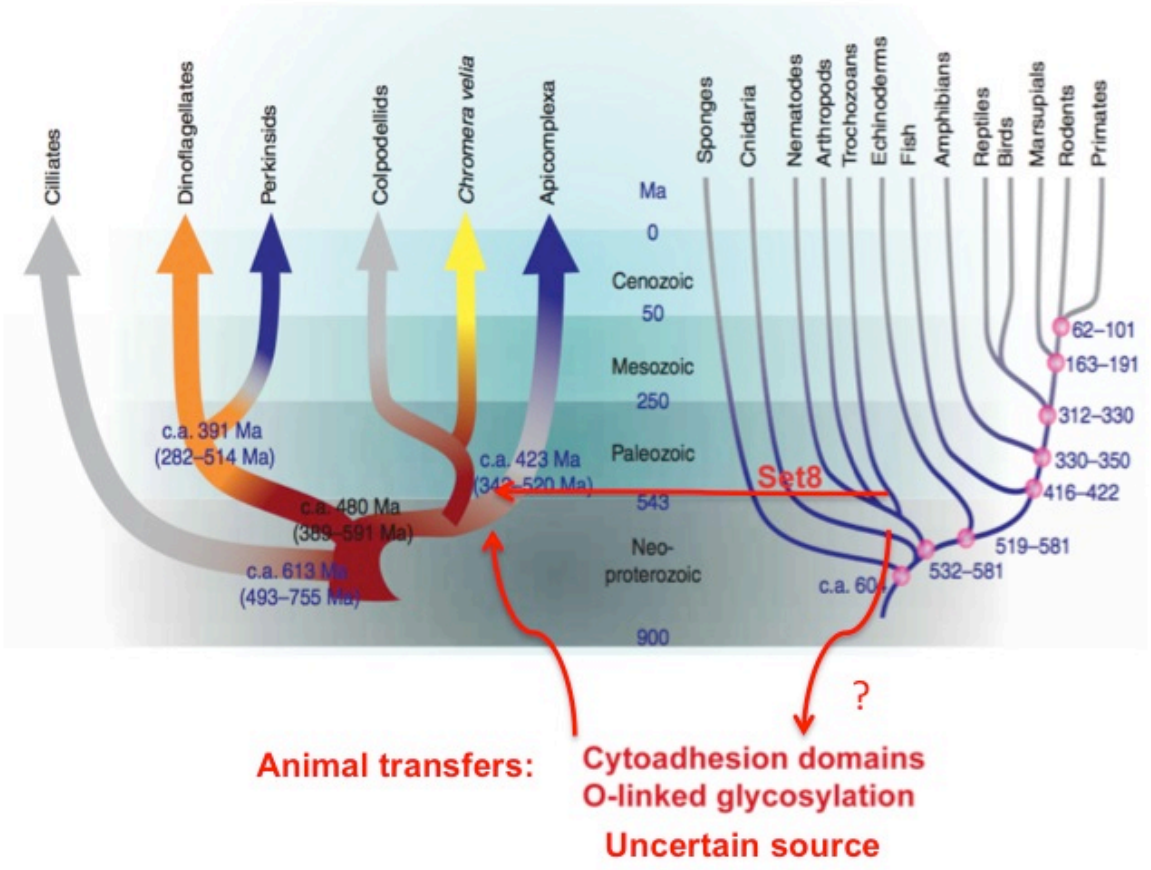

Figure 4.3. The timing and source of horizontal gene transfer of Set8

In this figure, adapted from Okamato et al, 2008, the evolution of chromoalveolates is shown next to the evolution of animals. Sponges were the first animals to appear nearly 600 million years ago. Based on the left tree, we may pinpoint the time that ancestral apicomplexa appeared to be around 423 million years ago, during the transition from the Neoproterozoic to the Paleozoic era. This timing coincides with the appearance of nematodes and arthropods (as shown on the right). Others have proposed that cytoadhesion protein/invasion proteins and enzymes involved in O-linked glycosylation where horizontally transferred from an unidentified animal source to an ancestral apicomplexan. Based on our findings in this report, I propose that Set8 was horizontally transferred from a nematode host 400 million years ago.

Adapted from: Okamato, N. 2008. The mother of all parasites. Future Microbiology 3:391-395.

3. Discussion

The data presented here, to our knowledge, provides the first conclusive support for the identification of lateral animal transfer of a chromatin/epigenetic modifier into the Apicomplexa. The grouping of the apicomplexan Set8 sub-clade within the animal clade is strongly supported and has been tested by rigorous phylogenetic comparison of the apicomplexan sequences to Set8 homologs from selected organisms from across the tree of life. The likely source of the gene was a nematode and the transfer was into an ancestral apicomplexan that lived 400 million years ago.

What is the functional consequence of this transfer? Given the adaptations of apicomplexans toward intracellular niches, the acquisition of a major epigenetic regulator that governs and faithfully affirms chromatin-silencing marks would be a critical step forward by protozoans for parasitism in diverse hosts with increasingly complex immune systems and host machinery. I propose that the organism acquired Set8 to enable higher order gene regulation, including cellular differentiation (for invading different hosts) as well as immune evasion. Acquisition of this type of regulation would enable the evolution of complex life cycles that involve multiple hosts as well as dormant cyst stages.

H4K20Me1, the epigenetic mark deposited on histones by canonical Set8, is known to participate in gene silencing, heterochromatin formation and mitotic regulation and DNA damage checkpoint repair. H4K20 methylation in DNA repair and genome integrity can be considered an early evolutionary function whereas the

heterochromatin function of the marker arose later in evolution in response to increasing genome complexity (Sautel et al. 2007). In extant apicomplexans *T.gondii* and *P. falciparum*, the H4K20 methylation marks have been mapped to heterochromatic regions, particularly those associated with H3K9triMe (a known epigenetic marker of immune and invasion genes). Furthermore, in metazoans, Set8 propagates the silenced chromatin to newly synthesized DNA by marking H4K20 according to the map of unacetylated H4K16 on mitotic chromosomes, a known silencing mark in the apicomplexan *P. falciparum*. Thus, the animal Set8 homolog and the functional modifications H4K20-mono/di/triMe -- are established modifications known to participate in epigenetic memory, antigenic variation and immune evasion strategies by selected Apicomplexa.

As Set8 orthologs are present in all the apicomplexans sampled, the transfer of Set8 appears to have accompanied the adaptation to parasitism by the ancestral apicomplexans The associations between modern invertebrates and their dinoflagellate zooxanthellae may serve as a model for the type of relationship shared by the ancient apicomplexan that acquired Set8 shared with its host. The loss of photosynthetic ability, however, would have resulted in an obligate intracellular organism that is metabolically dependent on the animal host and an obligate intracellular organism. Thus, the hallmarks of parasitism, including nutritional dependence, invasion, and immune escape likely evolved as part of the same evolutionary process by which photosynthesis was lost.

The concomitant transfer of domains involved in cytoadhesion and immune escape including TSP1, Sushi/CCP, Notch/Lin (NL1), NEC (neurexin-collagen

domain), fibronectin type 2 (F2), MAM domain and scavenger receptor (SR) is consistent with this notion (See Figure 4.3) (Templeton 2007). Based on the use of common adhesive domains involved in functions from adhesion to invasion and motility, and the timing of the proposed horizontal transfer of these animal domains, we may further surmise that they came from nematodes as well. The fact that zooxanthellae live between cell layers, but are not intracellular, points to the additional transition of extracellular habitation to intracellular habitation which the transfer of these proteins could have also facilitated.

Here, our report on the adaptation of an animal SET-domain-bearing methyltransferase involved in epigenetic memory extends our early investigations on the transcriptional machinery in *Plasmodium*. Our laboratory identified an expanded motif of the tail of the RNA polymerase II CTD is only otherwise found animal (specifically, mammalian) host polymerases (see Chapter 2) (Kishore et al. 2009). Intriguingly, this expansion is present only in primate parasites. In other eukaryotes, the CTD is connected to epigenetic memory through direct recruitment of SET-bearing methyltransferases that mark chromatin during active transcription. I highlight a new potential link between the expanded CTD motif and its involvement in the recruitment of SET-bearing methyltransferases (Chapter 3) in *P. falciparum*.

Here, I also provide the first report of horizontal gene transfer of animal provenance into slime molds. It is worth noting that among the amoebas, multicellular development and complex cellular communication and differentiation is found only in slime molds, the group that also acquired Set8 from animals (Strassmann et al. 2000; Brock et al. 2011). I cannot imply causality but simply raise this point to buttress the

argument that acquisitions of epigenetic machinery accompanied higher order cellular processes. Thus, I believe synergies between transcription machinery and epigenetic modifiers at the nexus of innate immunity and immune evasion offer much to explore in the natural history of host:parasite interactions. I believe this work will help to shed more light on the curious fate of lateral transfers from animal hosts into apicomplexans and other organisms, particularly SET-bearing chromatin modifiers and those molecules involved in epigenetic memory and immune evasion. I hope the work will help reveal novel, underappreciated acquisitions by protists at the dawn of apicomplexan parasitism that were functionally exploited as the immunological arms race between animal hosts and their parasites intensified.

4. Methods

Identification and Curation of Set8 Orthologs

The queries from canonical orthologs of Set8 from homo sapiens were used in BlastP searches of NCBI non-redundant protein databases for red, brown and green algae, moss, stramenophiles, oomycetes and animals, fruit flies, nematodes, rodents and human. Comparative evolutionary analyses were carried out on all sequences using the following approach. Sequences were aligned using the program MUSCLE (Edgar 2004) and trimmed to include only those domains that could be aligned clearly across all taxa included in the analyses. Phylogenetic relationships among sequences were inferred by both Bayesian inference (MrBayes) (Ronquist et al. 2003) and maximum-likelihood (PhyML) (Guindon, Gascuel 2003) in collaboration with John Stiller at East Carolina University. Relative support for tree nodes was inferred from Bayesian

posterior probabilities estimated from all trees sampled once the average standard deviation of split frequencies had converged on a stable value, as well as through 100 nonparametric ML bootstrap replicates.

BLAST query utilized:
PRSet7 (human Set8)
PVRRSSRKSKAELQSEERKRIDELIESGKEEGMKIDLIDGKGRGVIATKQFSRG
DFVVEYHGDLIEITDAKKREALYAQDPSTGCYMYYFQYLSKTYCVDATRETN
RLGRLINHSKCGNCQTKLHDIDGVPHLILIASRDIAAGEELLYDYGDRSKASIE
AHPWLKH

CHAPTER 5:
CONCLUSION

1.Evidence of RNA Polymerase II CTD plasticity at the genus and species level in Plasmodium

In this work, I provided the first evidence of plasticity of the RNA polymerase II CTD in the *Plasmodium* genus and showed that the plasticity extends to the species level. The plasticity concerns a motif that is only otherwise observed in mammalian RNA II polymerases, namely the YSPTSP**K** non-canonical heptads. The outstanding questions in this project principally involve determining the significance of the lysine in the seventh position of the heptad. Experiments to be undertaken that will substitute the lysine residue in the seventh position with a serine residue will help in this regard. Next, asking whether the lysine in the seventh position is itself modified may yield important clues to the function of this residue in primate malaria parasites. Finally, when more sequences become available from additional isolates, our laboratory and others may be able to determine if the size of the CTD correlates with transmission intensity, altitude, geography, vector species and drug resistance. The limited current sample size precludes such an analysis.

2. Linking the RNA Polymerase II CTD plasticity to epigenetic memory

I next identified a pair of histone modifiers of H3K36-triMe that are present in primate malaria parasites but are absent in rodent malaria parasites. I developed a novel rabbit

polyclonal antibody to detect this modification in *P. falciparum* and my data suggest that this modification *P. falciparum* that may be implicated in epigenetic memory of *P. falciparum* genes involved in immune evasion and invasion pathways. More importantly, I formed a hypothesis suggesting *how* this histone modification is deposited in the first instance and *how* it is linked to the epigenetic memory of immune escape and invasion genes as they relate to primate parasitism. I proposed that orthologs of Set2 adhere to the CTD during active transcription and deposit histone modifications to ensure epigenetic memory for gene activation in subsequent cell cycles. The outstanding questions here, which are currently being studied by my laboratory colleagues, include specifying what genes the H3K36-triMe modification associates with across the genome (through a ChIP-Seq) experiment. Secondly, determining whether Set2 orthologs in *P. falciparum* actually adhere to the *P. falciparum* CTD is useful. If binding is observed, the nature of the domain responsible for the interaction should be mapped. This domain may be compared with similar interaction domains in yeast or mammals for key evolutionary insights on the nature of histone modifications and transcriptional regulation. Finally, were such an interaction domain to be mapped, the domain could be used as a dominant negative construct through transgenic experiments to query for phenotypic effects – including loss of epigenetic memory at immune escape and invasion genes.

3. Horizontal gene transfer of chromatin modifiers in ancestral apicomplexans

In this work, I found three instances of horizontal gene transfer of chromatin modifiers that have implications for the origins of apicomplexan parasitism. The first two instances concern the reciprocal H3K36-triMe modifiers (Set2, methyltransferase

and JmjC1, demethylase) that were horizontally transferred from an algal endosymbiont and were adapted for immune escape and invasion. I also studied the natural history of the acquisition and deletion of these modifiers in *Plasmodium* and Apicomplexa, finding that Set2 and JmjC1 were present in other Apicomplexa, particularly those with expanded CTDs, and were lost specifically in rodent plasmodia lineages (which feature fewer than12 heptads). Truncated CTDs of budding yeast that feature fewer than 12 heptads fail to both recruit Set2 and to deposit the H3K36-triMe modification.

The third instance of horizontal gene transfer concerns the acquisition of the H4K20 histone lysine methyltransferase, which is implicated in mitosis, cellular differentiation, epigenetic memory and antigenic variation in primate malaria parasites. In this work, I provided the first conclusive evidence of a horizontal gene transfer for this gene and suggest that the source was a nematode. I further relate this event to coincidental events that transpired 400 million years ago, including the transfer of domains involved in cytoadhesion and genes involved in O-linked glycosylation, to argue that these events accompanied the transition to apicomplexan parasitism. To this end, determining the potential animal source of the O-linked glycosylation enzymes and cytoadhesion domains will yield valuable information for this analysis. The acquisition of the histone modifiers, Set2 and Set8, particularly from other organisms and their presumed adaptation for parasitic traits including immune escape and invasion, could guide new drug discovery efforts targeting these molecules. The use of small molecules targeting towards histone methyltransferases could also be exploited toward this end.

The overall claim of this thesis is that the acquisition of additional transcriptional regulatory units and epigenetic machinery facilitated the transition of ancestral apicomplexans to parasitic lifestyles and produced fundamental changes in development and cellular differentiation in various hosts. The motivation for the work is that by pinpointing the factors that accompanied or enabled apicomplexan parasitism, it may be possible to exploit or inhibit the same factors in the hopes of disabling parasitism. The work has hopefully provided new insights and tools towards this end.

APPENDIX

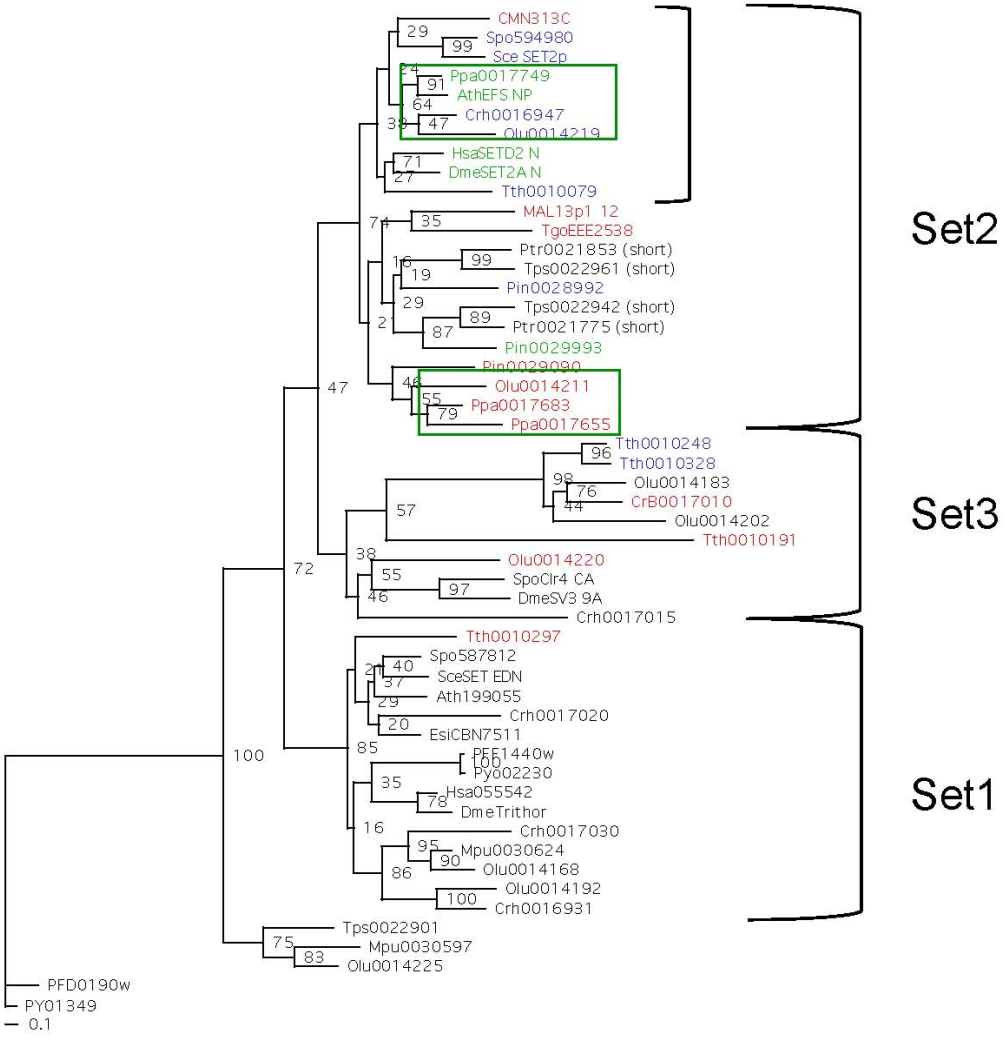

Figure A: The Set 1, 2 and 3 clades across the tree of life
This tree of peptide sequences surveyed across the tree of life provides phylogenetic verification of Set1, Set2 and Set3 clades. The apicomplexan (TgoEEE2538, *Toxoplasma gondii* and MAL13p1.122, *Plasmodium falciparum*) clearly group together within a Set2 clade but not within the animal Set2 clade denoted by *Homo sapiens* (Hsa) and *Drosophila melanogaster* (SET2AN). The highlighted green boxes indicate plant groups [Ath, *Arabidopsis thaliana* and Ppa, *Physicometrella patents* (moss) and Olu, *Ostreococcus luminaris*]. Overlaid on the phylogeny is the location of the SET domain within the protein (last third/ C-terminal, red; first third/ N-terminal blue; middle third, green).

Cmn = *Cyanidioschyzon merolae*, Spo=*Schizosaccharomyces pombe*, Sce = *Saccharomyces cerevisiae,* Ppa = *Physcomitrella patens subsp. patens*, Ath = *Arabidopsis thaliana*, Crh = *Chlamydomonas reinhardtii*, Olu = *Ostreococcus luminaris*, Hsa = *Homo sapiens*, Dme = *Drosophila melanogaster*, Tth = *Tetrahymena thermophila*, MAL = *Plasmodium falciparum*, Tgo = *Toxoplasma gondii*, Ptr = *Phaeodactylum tricornutum*, Tps = *Thalassiosira pseudonana*, Pin = *Phytophthora infestans* , Crb = *Chlamydomonas reinhardtii*

Accession Numbers of histone modifiers used:

PfSet2 orthologs:
XP_001007909.1| SET domain containing protein [Tetrahymena thermophila]
XP_001774900.1| predicted protein [Physcomitrella patens subsp. patens]
Arabidopsis
XP_001694743.1| histone methyltransferase [Chlamydomonas reinhardtii]
XP_001421914.1| predicted protein [Ostreococcus lucimarinus CCE9901]
NP_054878.5| histone-lysine N-methyltransferase SETD2 [Homo sapiens]
NP_572888.2| Set2, isoform A [Drosophila melanogaster]
CMN313C |[Cyanidioschyzon merolae]
NP_594980.1| histone lysine methyltransferase Set2 (predicted)
[Schizosaccharomyces pombe 972h-]NP_012367.2| Set2p [Saccharomyces cerevisiae S288c]
XP_002899256.1| histone-lysine N-methyltransferase, putative [Phytophthora infestans T30-4]
XP_002177534.1| predicted protein [Phaeodactylum tricornutum CCAP 1055/1]
XP_002999311.1| histone-lysine N-methyltransferase, putative [Phytophthora infestans T30-4]*
EEE25384.1| SET domain-containing protein, putative [Toxoplasma gondii GT1]
XP_001349960.1| SET domain protein, putative [Plasmodium falciparum 3D7]
XP_001421172.1| predicted protein [Ostreococcus lucimarinus CCE9901]*
XP_001768386.1| predicted protein [Physcomitrella patens subsp. patens]*
XP_001765507.1| predicted protein [Physcomitrella patens subsp. patens]**
XP_002185333.1| predicted protein [Phaeodactylum tricornutum CCAP 1055/1]*
XP_002296152.1| SET-domain containing protein [Thalassiosira pseudonana CCMP1335]

PfJmjC1 orthologs:
CBN79746.1| conserved unknown protein [Ectocarpus siliculosus]
XP_002902031.1| histone demethylase, putative [Phytophthora infestans T30-4]
XP_001778258.1| predicted protein [Physcomitrella patens subsp. patens]
XP_001421122.1| predicted protein [Ostreococcus lucimarinus CCE9901]
XP_003082696.1| transcription factor jumonji (ISS) [Ostreococcus tauri]
XP_003061284.1| JmjN/JmjC protein [Micromonas pusilla CCMP1545]
XP_003055255.1| JmjN/JmjC protein [Micromonas pusilla CCMP1545]*
XP_003078252.1| DNA-binding protein jumonji/RBP2/SMCY, contains JmjC domain (ISS) [Ostreococcus tauri]*
XP_001416696.1| predicted protein [Ostreococcus lucimarinus CCE9901]*
XP_001349427.1| JmjC domain containing protein [Plasmodium falciparum 3D7]
EEE19207.1| hypothetical protein TGGT1_055200 [Toxoplasma gondii GT1]
XP_001770596.1| predicted protein [Physcomitrella patens subsp. patens]*
NP_172338.4| transcription factor jumonji and C5HC2 type zinc finger domain-containing protein [Arabidopsis thaliana]

XP_001701660.1| jumonji protein [Chlamydomonas reinhardtii]
XP_002182492.1| predicted protein [Phaeodactylum tricornutum CCAP 1055/1]
XP_002289184.1| predicted protein [Thalassiosira pseudonana CCMP1335]
XP_003059818.1| predicted protein [Micromonas pusilla CCMP1545]**
BAA31652.2| KIAA0677 protein [Homo sapiens]
NP_788344.2| histone demethylase 4B, isoform A [Drosophila melanogaster]
P39956.1|RPH1_YEAST RecName: Full=DNA damage-responsive transcriptional repressor RPH1
XP_001025065.1| jmjC domain containing protein [Tetrahymena thermophila]
XP_001458550.1| hypothetical protein [Paramecium tetraurelia strain d4-2]
XP_001691536.1| predicted protein [Chlamydomonas reinhardtii]

XP_001771543.1| predicted protein [Physcomitrella patens subsp. patens]**
CBN78828.1| conserved unknown protein [Ectocarpus siliculosus]*
XP_002183428.1| predicted protein [Phaeodactylum tricornutum CCAP 1055/1]
XP_002287652.1| predicted protein [Thalassiosira pseudonana CCMP1335]*
XP_003056317.1| predicted protein [Micromonas pusilla CCMP1545]***
XP_003080681.1| retinoblastoma binding protein 2 (ISS) [Ostreococcus tauri]**
XP_001419094.1| predicted protein [Ostreococcus lucimarinus CCE9901]**

PfSet8 orthologs:
AAM47033.1| [Homo sapiens]
XP_001351351.1| [Plasmodium falciparum 3D7]
ABO38805.1| [Toxoplasma gondii]
XP_001610397.1| [Babesia bovis T2Bo]
XP_625644.1| [Cryptosporidium parvum Iowa II]
XP_954446.1| [Theileria annulata]
NP_650354.1 [Drosophila melanogaster]
XP_001694743.1| [Chlamydomonas reinhardtii]
XP_003062434.1| [Micromonas pusilla CCMP1545]
XP_001422533.1| [Ostreococcus lucimarinus CCE9901]
XP_001421914.1 [Ostreococcus lucimarinus CCE9901]*
XP_001774900.1 [Physcomitrella patens subsp. patens]
XP_001029767.1| [Tetrahymena thermophila]
XP_415116.1 [Gallus gallus]
XP_313242.4| [Anopheles gambiae str. PEST]
NP_001022796.1| [Caenorhabditis elegans]
NP_510241.1| [Caenorhabditis elegans]
XP_674522.1| [Plasmodium berghei strain ANKA]
XP_003222818.1 [Anolis carolinensis]
XP_853243.1 [Canis familiaris]
NP_001039795.1| [Bos taurus]
XP_001270666.1| [Aspergillus clavatus NRRL 1]
XP_002427372.1 [Pediculus humanus corporis]

XP_001942662.2| [Acyrthosiphon pisum]
XP_969412.1| [Tribolium castaneum]
XP_001658906.1| [Aedes aegypti]
XP_509461.3| [Pan troglodytes]
NP_084517.2 [Mus musculus]
XP_001507341.1 [Ornithorhynchus anatinus]
XP_002189578.1| [Taeniopygia guttata]
XP_002154569.1| [Hydra magnipapillata]
NP_001072815.1| [Xenopus (Silurana) tropicalis]
XP_001606626.1| [Nasonia vitripennis]
NP_001082246.1| [Xenopus laevis]
XP_002576803.1| [Schistosoma mansoni]
XP_002131628.1| [Ciona intestinalis]
XP_003137927.1| [Loa loa]
ADY48236.1| [Ascaris suum]
XP_797080.1| [Strongylocentrotus purpuratus]
NP_001093559.1| [Danio rerio]
XP_002910994.1| [Coprinopsis cinerea okayama7#130]
AAD24840.1| [Arabidopsis thaliana]
XP_001584257.1| [Trichomonas vaginalis G3]
XP_001707053.1| [Giardia lamblia ATCC 50803]
XP_002683109.1| [Naegleria gruberi]
XP_001469585.1| [Leishmania infantum JPCM5]
XP_002909073.1| [Phytophthora infestans T30-4]
XP_001351351.1| [Plasmodium falciparum 3D7]
EGG21088.1| [Dictyostelium fasciculatum]
XP_002291362.1| [Thalassiosira pseudonana CCMP1335]
XP_002177534.1| [Phaeodactylum tricornutum CCAP 1055/1]
CAG05293.1| [Tetraodon nigroviridis]
XP_003387132.1| [Amphimedon queenslandica]
XP_955728.1| [Encephalitozoon cuniculi GB-M1]
XP_002576803.1| [Schistosoma mansoni]
AAX30591.2| [Schistosoma japonicum]
XP_002954751.1| [Volvox carteri f. nagariensis]
XP_001629745.1| [Nematostella vectensis]
XP_002154569.1| [Hydra magnipapillata]
XP_003370634.1| [Trichinella spiralis]
XP_001891577.1| [Brugia malayi]
XP_002427372.1| [Pediculus humanus corporis]
XP_969412.1| [Tribolium castaneum]
XP_797080.1| [Strongylocentrotus purpuratus]
AAD24840.1| [Arabidopsis thaliana]

XP_001629745.1| [Nematostella vectensis]
XP_001606626.1| [Nasonia vitripennis]
XP_001097869.2| [Macaca mulatta]
XP_003222818.1| [Anolis carolinensis]
XP_415116.1| [Gallus gallus]
XP_003210954.1| [Meleagris gallopavo]
XP_002189578.1 | [Taeniopygia guttata]

REFERENCES

Allison, LA, M Moyle, M Shales, CJ Ingles. 1985. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell 42:599-610.

Anantharaman, V, LM Iyer, S Balaji, L Aravind. 2007. Adhesion molecules and other secreted host-interaction determinants in Apicomplexa: insights from comparative genomics. International Review of Cytology 262:1-74.

Andersson, JO. 2009. Gene transfer and diversification of microbial eukaryotes. Annual review of microbiology 63:177-193.

Andrews, KT, TN Tran, NC Wheatley, DP Fairlie. 2009. Targeting histone deacetylase inhibitors for anti-malarial therapy. Current Topics in Medicinal Chemistry 9:292-308.

Aravind, L, LM Iyer, TE Wellems, LH Miller. 2003. Plasmodium biology: genomic gleanings. Cell 115:771-785.

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

Barron-Casella, E, JL Corden. 1992. Conservation of the mammalian RNA polymerase II largest-subunit C-terminal domain. Journal of Molecular Evolution 35:405-410.

Bartolomei, MS, NF Halden, CR Cullen, JL Corden. 1988. Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. Molecular and Cellular Biology 8:330-339.

Baruch, DI, BL Pasloske, HB Singh, X Bi, XC Ma, M Feldman, TF Taraschi, RJ Howard. 1995. Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82:77-87.

Berendt, AR, GD Tumer, CI Newbold. 1994. Cerebral malaria: the sequestration hypothesis. Parasitology Today 10:412-414.

Bienkiewicz, EA, A Moon Woody, RW Woody. 2000. Conformation of the RNA polymerase II C-terminal domain: circular dichroism of long and short fragments. Journal of Molecular Biology 297:119-133.

Bischoff, E, C Vaquero. 2010. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of Plasmodium falciparum. BMC Genomics 11:34.

Brock, DA, TE Douglas, DC Queller, JE Strassmann. 2011. Primitive agriculture in a social amoeba. Nature 469:393-396.

Carrozza, MJ, B Li, L Florens, T Suganuma, SK Swanson, KK Lee, WJ Shia, S Anderson, J Yates, MP Washburn, JL Workman. 2005. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell 123:581-592.

Carty, SM, AL Greenleaf. 2002. Hyperphosphorylated C-terminal repeat domain-associating proteins in the nuclear proteome link transcription to

DNA/chromatin modification and RNA processing. Molecular & Eellular Proteomics : MCP 1:598-610.
Cavalier-Smith, T. 1993. Kingdom protozoa and its 18 phyla. Microbiological reviews 57:953-994.
Chapman, AB, N Agabian. 1994. Trypanosoma brucei RNA polymerase II is phosphorylated in the absence of carboxyl-terminal domain heptapeptide repeats. The Journal of Biological Chemistry 269:4754-4760.
Chapman, RD, M Conrad, D Eick. 2005. Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. Molecular and Cellular Biology 25:7665-7674.
Chapman, RD, M Heidemann, TK Albert, R Mailhammer, A Flatley, M Meisterernst, E Kremmer, D Eick. 2007. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science 318:1780-1782.
Chapman, RD, M Heidemann, C Hintermair, D Eick. 2008. Molecular evolution of the RNA polymerase II CTD. Trends in Genetics 24:289-296.
Chen, Q, V Fernandez, A Sundstrom, M Schlichtherle, S Datta, P Hagblom, M Wahlgren. 1998. Developmental selection of var gene expression in Plasmodium falciparum. Nature 394:392-395.
Chookajorn, T, MS Costanzo, DL Hartl, KW Deitsch. 2007a. Malaria: a peek at the var variorum. Trends in Parasitology 23:563-565.
Chookajorn, T, R Dzikowski, M Frank, F Li, AZ Jiwani, DL Hartl, KW Deitsch. 2007b. Epigenetic memory at malaria virulence genes. Proceedings of the National Academy of Sciences of the United States of America 104:899-902.
Chu, Y, A Sutton, R Sternglanz, G Prelich. 2006. The BUR1 cyclin-dependent protein kinase is required for the normal pattern of histone methylation by SET2. Molecular and Cellular Biology 26:3029-3038.
Corden, JL. 1990. Tails of RNA polymerase II. Trends Biochem Sci 15:383-387.
Corden, JL. 2007. Transcription. Seven ups the code. Science 318:1735-1736.
Crowley, VM, N Rovira-Graells, LR de Pouplana, A Cortes. 2011. Heterochromatin formation in bistable chromatin domains controls the epigenetic repression of clonally variant Plasmodium falciparum genes linked to erythrocyte invasion. Molecular Microbiology 80:391-406.
Cui, L, Q Fan, J Miao. 2008. Histone lysine methyltransferases and demethylases in Plasmodium falciparum. International journal for parasitology 38:1083-1097.
Cui, L, J Miao. 2010. Chromatin-mediated epigenetic regulation in the malaria parasite Plasmodium falciparum. Eukaryotic cell 9:1138-1149.
Cunningham, D, J Fonager, W Jarra, C Carret, P Preiser, J Langhorne. 2009. Rapid changes in transcription profiles of the Plasmodium yoelii yir multigene family in clonal populations: lack of epigenetic memory? PloS One 4:e4285.
Deitsch, K, C Driskill, T Wellems. 2001a. Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. Nucleic Acids Research 29:850-853.
Deitsch, K, M Duraisingh, R Dzikowski, A Gunasekera, S Khan, K Le Roch, M Llinas, G Mair, V McGovern, D Roos, J Shock, J Sims, R Wiegand, E

Winzeler. 2007. Mechanisms of gene regulation in Plasmodium. The American Journal of Tropical Medicine and Hygiene 77:201-208.
Deitsch, KW. 2005. Malaria virulence genes controlling expression through chromatin modification. Cell 121:1-2.
Deitsch, KW, MS Calderwood, TE Wellems. 2001b. Malaria. Cooperative silencing elements in var genes. Nature 412:875-876.
Deitsch, KW, JM Carlton, JC Wootton, TE Wellems. 2001c. Host sequences in Plasmodium falciparum and Plasmodium vivax genomic DNA: horizontal transfer or contamination artifact? FEBS Letters 491:164-165.
Deitsch, KW, L Hviid. 2004. Variant surface antigens, virulence genes and the pathogenesis of malaria. Trends in Parasitology 20:562-566.
Dondorp, AM, S Yeung, L White, C Nguon, NP Day, D Socheat, L von Seidlein. 2010. Artemisinin resistance: current status and scenarios for containment. Nature Reviews Microbiology 8:272-280.
Duraisingh, MT, TS Voss, AJ Marty, MF Duffy, RT Good, JK Thompson, LH Freitas-Junior, A Scherf, BS Crabb, AF Cowman. 2005. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. Cell 121:13-24.
Dzikowski, R, KW Deitsch. 2008. Active transcription is required for maintenance of epigenetic memory in the malaria parasite Plasmodium falciparum. Journal of Molecular Biology 382:288-297.
Dzikowski, R, M Frank, K Deitsch. 2006. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. PLoS Pathogens 2:e22.
Dzikowski, R, F Li, B Amulic, A Eisberg, M Frank, S Patel, TE Wellems, KW Deitsch. 2007. Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. EMBO reports 8:959-965.
Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.
Egloff, S, S Murphy. 2008. Cracking the RNA polymerase II CTD code. Trends in Genetics.
Egloff, S, D O'Reilly, RD Chapman, A Taylor, K Tanzhaus, L Pitts, D Eick, S Murphy. 2007. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science 318:1777-1779.
Eichinger, L, JA Pachebat, G Glockner, MA Rajandream, R Sucgang, M Berriman, J Song, R Olsen, K Szafranski, Q Xu, B Tunggal, S Kummerfeld, M Madera, BA Konfortov, F Rivero, AT Bankier, R Lehmann, N Hamlin, R Davies, P Gaudet, P Fey, K Pilcher, G Chen, D Saunders, E Sodergren, P Davis, A Kerhornou, X Nie, N Hall, C Anjard, L Hemphill, N Bason, P Farbrother, B Desany, E Just, T Morio, R Rost, C Churcher, J Cooper, S Haydock, N van Driessche, A Cronin, I Goodhead, D Muzny, T Mourier, A Pain, M Lu, D Harper, R Lindsay, H Hauser, K James, M Quiles, M Madan Babu, T Saito, C Buchrieser, A Wardroper, M Felder, M Thangavelu, D Johnson, A Knights, H Loulseged, K Mungall, K Oliver, C Price, MA Quail, H Urushihara, J Hernandez, E Rabbinowitsch, D Steffen, M Sanders, J Ma, Y Kohara, S Sharp,

M Simmonds, S Spiegler, A Tivey, S Sugano, B White, D Walker, J Woodward, T Winckler, Y Tanaka, G Shaulsky, M Schleicher, G Weinstock, A Rosenthal, EC Cox, RL Chisholm, R Gibbs, WF Loomis, M Platzer, RR Kay, J Williams, PH Dear, AA Noegel, B Barrell, A Kuspa. 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435:43-57.

Fang, J, Q Feng, CS Ketel, H Wang, R Cao, L Xia, H Erdjument-Bromage, P Tempst, JA Simon, Y Zhang. 2002. Purification and functional characterization of SET8, a nucleosomal histone H4-lysine 20-specific methyltransferase. Current Biology 12:1086-1099.

Fernandez, V, CJ Treutiger, GB Nash, M Wahlgren. 1998. Multiple adhesive phenotypes linked to rosetting binding of erythrocytes in Plasmodium falciparum malaria. Infection and Immunity 66:2969-2975.

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

Frank, M, R Dzikowski, D Costantini, B Amulic, E Berdougo, K Deitsch. 2006. Strict pairing of var promoters and introns is required for var gene silencing in the malaria parasite Plasmodium falciparum. The Journal of Biological Chemistry 281:9942-9952.

Gardner, MJ, N Hall, E Fung, O White, M Berriman, RW Hyman, JM Carlton, A Pain, KE Nelson, S Bowman, IT Paulsen, K James, JA Eisen, K Rutherford, SL Salzberg, A Craig, S Kyes, MS Chan, V Nene, SJ Shallom, B Suh, J Peterson, S Angiuoli, M Pertea, J Allen, J Selengut, D Haft, MW Mather, AB Vaidya, DM Martin, AH Fairlamb, MJ Fraunholz, DS Roos, SA Ralph, GI McFadden, LM Cummings, GM Subramanian, C Mungall, JC Venter, DJ Carucci, SL Hoffman, C Newbold, RW Davis, CM Fraser, B Barrell. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419:498-511.

Gerber, M, A Shilatifard. 2003. Transcriptional elongation by RNA polymerase II and histone methylation. J Biol Chem 278:26303-26306.

Giesecke, H, JC Barale, G Langsley, AW Cornelissen. 1991. The C-terminal domain of RNA polymerase II of the malaria parasite Plasmodium berghei. Biochemical and biophysical research communications 180:1350-1355.

Guindon, S, O Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology 52:696-704.

Guo, Z, JW Stiller. 2005. Comparative genomics and evolution of proteins associated with RNA polymerase II C-terminal domain. Molecular Biology and Evolution 22:2166-2178.

Hakimi, MA, KW Deitsch. 2007. Epigenetics in Apicomplexa: control of gene expression during cell cycle progression, differentiation and antigenic variation. Current Opinion in Microbiology 10:357-362.

Hampsey, M, D Reinberg. 2003. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. Cell 113:429-432.

Hill, AV. 2006. Pre-erythrocytic malaria vaccines: towards greater efficacy. Nature Reviews Immunology 6:21-32.
Howe, KJ. 2002. RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim Biophys Acta 1577:308-324.
Huang, J, N Mullapudi, T Sicheritz-Ponten, JC Kissinger. 2004. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. International Journal for Parasitology 34:265-274.
Issar, N, SA Ralph, L Mancio-Silva, C Keeling, A Scherf. 2009. Differential sub-nuclear localisation of repressive and activating histone methyl modifications in P. falciparum. Microbes and infection / Institut Pasteur 11:403-407.
Jang, YK, L Wang, GB Sancar. 1999. RPH1 and GIS1 are damage-responsive repressors of PHR1. Molecular and Cellular Biology 19:7630-7638.
Jeanmougin, F, JD Thompson, M Gouy, DG Higgins, TJ Gibson. 1998. Multiple sequence alignment with Clustal X. Trends in Biochemical Sciences 23:403-405.
Jiang, L, MJ Lopez-Barragan, H Jiang, J Mu, D Gaur, K Zhao, G Felsenfeld, LH Miller. 2010. Epigenetic control of the variable expression of a Plasmodium falciparum receptor protein for erythrocyte invasion. Proceedings of the National Academy of Sciences of the United States of America 107:2224-2229.
Jorgensen, S, M Eskildsen, K Fugger, L Hansen, MS Larsen, AN Kousholt, RG Syljuasen, MB Trelle, ON Jensen, K Helin, CS Sorensen. 2011. SET8 is degraded via PCNA-coupled CRL4(CDT2) ubiquitylation in S phase and after UV irradiation. The Journal of Cell Biology 192:43-54.
Joshi, AA, K Struhl. 2005. Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Molecular cell 20:971-978.
Karachentsev, D, K Sarma, D Reinberg, R Steward. 2005. PR-Set7-dependent methylation of histone H4 Lys 20 functions in repression of gene expression and is essential for mitosis. Genes and Development 19:431-435.
Katz, LA. 2006. Genomics and Evolution of Microbial Eukaryotes. Oxford: Oxford University Press.
Keogh, MC, SK Kurdistani, SA Morris, SH Ahn, V Podolny, SR Collins, M Schuldiner, K Chin, T Punna, NJ Thompson, C Boone, A Emili, JS Weissman, TR Hughes, BD Strahl, M Grunstein, JF Greenblatt, S Buratowski, NJ Krogan. 2005. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell 123:593-605.
Kessin, RH, GG Gundersen, V Zaydfudim, M Grimson. 1996. How cellular slime molds evade nematodes. Proceedings of the National Academy of Sciences of the United States of America 93:4857-4861.
Kishore, SP, SL Perkins, TJ Templeton, KW Deitsch. 2009. An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. Journal of Molecular Evolution 68:706-714.
Kizer, KO, HP Phatnani, Y Shibata, H Hall, AL Greenleaf, BD Strahl. 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone

H3 K36 methylation with transcript elongation. Molecular and Cellular Biology 25:3305-3316.
Klose, RJ, EM Kallin, Y Zhang. 2006a. JmjC-domain-containing proteins and histone demethylation. Nature Reviews Genetics 7:715-727.
Klose, RJ, K Yamane, Y Bae, D Zhang, H Erdjument-Bromage, P Tempst, J Wong, Y Zhang. 2006b. The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. Nature 442:312-316.
Kohler, S. 2005. Multi-membrane-bound structures of Apicomplexa: I. the architecture of the Toxoplasma gondii apicoplast. Parasitology research 96:258-272.
Krogan, NJ, M Kim, A Tong, A Golshani, G Cagney, V Canadien, DP Richards, BK Beattie, A Emili, C Boone, A Shilatifard, S Buratowski, J Greenblatt. 2003. Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol 23:4207-4218.
Kyes, SA, SM Kraemer, JD Smith. 2007. Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. Eukaryotic Cell 6:1511-1520.
Le Roch, KG, JR Johnson, L Florens, Y Zhou, A Santrosyan, M Grainger, SF Yan, KC Williamson, AA Holder, DJ Carucci, JR Yates, 3rd, EA Winzeler. 2004. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Research 14:2308-2318.
Li, B, M Carey, JL Workman. 2007. The role of chromatin during transcription. Cell 128:707-719.
Li, B, L Howe, S Anderson, JR Yates, 3rd, JL Workman. 2003. The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. The Journal of Biological Chemistry 278:8897-8903.
Li, F, L Sonbuchner, SA Kyes, C Epp, KW Deitsch. 2008. Nuclear non-coding RNAs are transcribed from the centromeres of Plasmodium falciparum and are associated with centromeric chromatin. The Journal of Biological Chemistry 283:5692-5698.
Li, J, D Moazed, SP Gygi. 2002. Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. The Journal of Biological Chemistry 277:49383-49388.
Li, M, HP Phatnani, Z Guan, H Sage, AL Greenleaf, P Zhou. 2005. Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. Proceedings of the National Academy of Sciences of the United States of America 102:17636-17641.
Li, WB, DJ Bzik, HM Gu, M Tanaka, BA Fox, J Inselburg. 1989. An enlarged largest subunit of Plasmodium falciparum RNA polymerase II defines conserved and variable RNA polymerase domains. Nucleic Acids Research 17:9621-9636.
Li, XY, S Thomas, PJ Sabo, MB Eisen, JA Stamatoyannopoulos, MD Biggin. 2011. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biology 12:R34.

Liao, SM, IC Taylor, RE Kingston, RA Young. 1991. RNA polymerase II carboxy-terminal domain contributes to the response to multiple acidic activators in vitro. Genes and Development 5:2431-2440.
Licatalosi, DD, G Geiger, M Minet, S Schroeder, K Cilli, JB McNeil, DL Bentley. 2002. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Molecular cell 9:1101-1111.
Litingtung, Y, AM Lawler, SM Sebald, E Lee, JD Gearhart, H Westphal, JL Corden. 1999. Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II. Mol Gen Genet 261:100-105.
Liu, P, AL Greenleaf, JW Stiller. 2008. The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. Molecular Biology and Evolution 25:719-727.
Lopez-Rubio, JJ, L Mancio-Silva, A Scherf. 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host & Microbe 5:179-190.
Mackintosh, CL, JG Beeson, K Marsh. 2004. Clinical features and pathogenesis of severe malaria. Trends in Parasitology 20:597-603.
McFadden, GI, ME Reith, J Munholland, N Lang-Unnasch. 1996. Plastid in human parasites. Nature 381:482.
Meininghaus, M, RD Chapman, M Horndasch, D Eick. 2000. Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription. Journal of Biological Chemistry 275:24375-24382.
Merrick, CJ, MT Duraisingh. 2007. Plasmodium falciparum Sir2: an unusual sirtuin with dual histone deacetylase and ADP-ribosyltransferase activity. Eukaryotic Cell 6:2081-2091.
Miao, J, Q Fan, L Cui, J Li. 2006. The malaria parasite Plasmodium falciparum histones: organization, expression, and acetylation. Gene 369:53-65.
Miller, LH, MF Good, G Milon. 1994. Malaria pathogenesis. Science 264:1878-1883.
Mizzen, CA, XJ Yang, T Kokubo, JE Brownell, AJ Bannister, T Owen-Hughes, J Workman, L Wang, SL Berger, T Kouzarides, Y Nakatani, CD Allis. 1996. The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. Cell 87:1261-1270.
Moore, RB, M Obornik, J Janouskovec, T Chrudimsky, M Vancova, DH Green, SW Wright, NW Davies, CJ Bolch, K Heimann, J Slapeta, O Hoegh-Guldberg, JM Logsdon, DA Carter. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature 451:959-963.
Noedl, H, Y Se, K Schaecher, BL Smith, D Socheat, MM Fukuda. 2008. Evidence of artemisinin-resistant malaria in western Cambodia. The New England Journal of Medicine 359:2619-2620.
Nonet, M, D Sweetser, RA Young. 1987. Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. Cell 50:909-915.
Okamato, N. 2008. The mother of all parasites. Future Microbiology 3:391-395.

Perkins, SL, IN Sarkar, R Carter. 2007. The phylogeny of rodent malaria parasites: simultaneous analysis across three genomes. Infection, Genetics and Evolution 7:74-83.
Petter, M, CC Lee, TJ Byrne, KE Boysen, J Volz, SA Ralph, AF Cowman, GV Brown, MF Duffy. 2011. Expression of P. falciparum var genes involves exchange of the histone variant H2A.Z at the promoter. PLoS Pathogens 7:e1001292.
Phatnani, HP, AL Greenleaf. 2006. Phosphorylation and functions of the RNA polymerase II CTD. Genes and Development 20:2922-2936.
Plowe, CV, P Alonso, SL Hoffman. 2009. The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. The Journal of Infectious Diseases 200:1646-1649.
Proudfoot, NJ, A Furger, MJ Dye. 2002. Integrating mRNA processing with transcription. Cell 108:501-512.
Ptashne, M, Gann, A. 2002. Genes & Signals. Cold Spring Harbor, New York: Cold Spring Harbord Laboratory Press.
Ronquist, F, JP Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.
Salanti, A, T Staalsoe, T Lavstsen, AT Jensen, MP Sowa, DE Arnot, L Hviid, TG Theander. 2003. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in pregnancy-associated malaria. Molecular Microbiology 49:179-191.
Sautel, CF, D Cannella, O Bastien, S Kieffer, D Aldebert, J Garin, I Tardieux, H Belrhali, MA Hakimi. 2007. SET8-mediated methylations of histone H4 lysine 20 mark silent heterochromatic domains in apicomplexan genomes. Molecular and Cellular Biology 27:5711-5724.
Scherf, A, R Hernandez-Rivas, P Buffet, E Bottius, C Benatar, B Pouvelle, J Gysin, M Lanzer. 1998. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. The EMBO Journal 17:5418-5426.
Scherf, A, JJ Lopez-Rubio, L Riviere. 2008. Antigenic variation in Plasmodium falciparum. Annual Review of Microbiology 62:445-470.
Stiller, JW, MS Cook. 2004. Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. Eukaryotic Cell 3:735-740.
Stiller, JW, BL McConaughy, BD Hall. 2000. Evolutionary complementation for polymerase II CTD function. Yeast 16:57-64.
Strassmann, JE, Y Zhu, DC Queller. 2000. Altruism and social cheating in the social amoeba Dictyostelium discoideum. Nature 408:965-967.
Sucgang, R, A Kuo, X Tian, W Salerno, A Parikh, CL Feasley, E Dalin, H Tu, E Huang, K Barry, E Lindquist, H Shapiro, D Bruce, J Schmutz, A Salamov, P Fey, P Gaudet, C Anjard, MM Babu, S Basu, Y Bushmanova, H van der Wel, M Katoh-Kurasawa, C Dinh, PM Coutinho, T Saito, M Elias, P Schaap, RR Kay, B Henrissat, L Eichinger, F Rivero, NH Putnam, CM West, WF Loomis, RL Chisholm, G Shaulsky, JE Strassmann, DC Queller, A Kuspa, IV

Grigoriev. 2011. Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum. Genome Biology 12:R20.
Swamy, L, B Amulic, KW Deitsch. 2011. Plasmodium falciparum var gene silencing is determined by cis DNA elements that form stable and heritable interactions. Eukaryotic Cell 10:530-539.
Templeton, TJ. 2007. Whole-genome natural histories of apicomplexan surface proteins. Trends in Parasitology 23:205-212.
Templeton, TJ, LM Iyer, V Anantharaman, S Enomoto, JE Abrahante, GM Subramanian, SL Hoffman, MS Abrahamsen, L Aravind. 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Research 14:1686-1695.
Tonkin, CJ, CK Carret, MT Duraisingh, TS Voss, SA Ralph, M Hommel, MF Duffy, LM Silva, A Scherf, A Ivens, TP Speed, JG Beeson, AF Cowman. 2009. Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in Plasmodium falciparum. PLoS Biology 7:e84.
Trelle, MB, AM Salcedo-Amaya, AM Cohen, HG Stunnenberg, ON Jensen. 2009. Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite Plasmodium falciparum. Journal of Proteome Research 8:3439-3450.
Vojnic, E, B Simon, BD Strahl, M Sattler, P Cramer. 2006. Structure and carboxyl-terminal domain (CTD) binding of the Set2 SRI domain that couples histone H3 Lys36 methylation to transcription. The Journal of Biological Chemistry 281:13-15.
Volkman, SK, PC Sabeti, D DeCaprio, DE Neafsey, SF Schaffner, DA Milner, Jr., JP Daily, O Sarr, D Ndiaye, O Ndir, S Mboup, MT Duraisingh, A Lukens, A Derr, N Stange-Thomann, S Waggoner, R Onofrio, L Ziaugra, E Mauceli, S Gnerre, DB Jaffe, J Zainoun, RC Wiegand, BW Birren, DL Hartl, JE Galagan, ES Lander, DF Wirth. 2007. A genome-wide map of diversity in Plasmodium falciparum. Nature Genetics 39:113-119.
Volz, J, TG Carvalho, SA Ralph, P Gilson, J Thompson, CJ Tonkin, C Langer, BS Crabb, AF Cowman. 2010. Potential epigenetic regulatory proteins localise to distinct nuclear sub-compartments in Plasmodium falciparum. International Journal for Parasitology 40:109-121.
Voss, TS, J Healer, AJ Marty, MF Duffy, JK Thompson, JG Beeson, JC Reeder, BS Crabb, AF Cowman. 2006. A var gene promoter controls allelic exclusion of virulence genes in Plasmodium falciparum malaria. Nature 439:1004-1008.
Voss, TS, CJ Tonkin, AJ Marty, JK Thompson, J Healer, BS Crabb, AF Cowman. 2007. Alterations in local chromatin environment are involved in silencing and activation of subtelomeric var genes in Plasmodium falciparum. Molecular Microbiology 66:139-150.
Wassmer, SC, V Combes, GE Grau. 2003. Pathophysiology of cerebral malaria: role of host cells in the modulation of cytoadhesion. Annals of the New York Academy of Sciences 992:30-38.
Westenberger, SJ, L Cui, N Dharia, E Winzeler. 2009. Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-

poor intergenic regions and chromatin remodeling of core and subtelomeric genes. BMC Genomics 10:610.
Xiao, T, H Hall, KO Kizer, Y Shibata, MC Hall, CH Borchers, BD Strahl. 2003. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes and Development 17:654-663.
Yuda, M, S Iwanaga, S Shigenobu, T Kato, I Kaneko. 2010. Transcription factor AP2-Sp and its target genes in malarial sporozoites. Molecular Microbiology 75:854-863.
Zofall, M, SI Grewal. 2006. Swi6/HP1 recruits a JmjC domain protein to facilitate transcription of heterochromatic repeats. Molecular Cell 22:681-692.
Zorio, DA, DL Bentley. 2004. The link between mRNA processing and transcription: communication works both ways. Experimental Cell Research 296:91-97.