

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Towards an understanding of the gene regulatory network of the intraerythrocytic developmental cycle of Plasmodium falciparum

Permalink

<https://escholarship.org/uc/item/4b54g455>

Author

Irie, Takeshi

Publication Date

2007-09-17

Peer reviewed|Thesis/dissertation

Towards an understanding of the gene regulatory network of the
intraerythrocytic developmental cycle of *Plasmodium falciparum*

by

Takeshi Irie

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

UMI Number: 3288940

Copyright 2007 by
Irie, Takeshi

All rights reserved.

UMI[®]

UMI Microform 3288940

Copyright 2008 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Copyright 2007

by

Takeshi Irie

Acknowledgements

The cast of people that have contributed in various ways to the work described here is innumerable but I will provide an incomplete list of people I would like to thank. First, I would like to start out by thanking my parents for providing the chassis with which I have navigated the entirety of this experience. Their enthusiastic support of my studies provided much of the fuel that has propelled me to my current station. The constitution that I have inherited from them is also a gift that has served me well over these years.

Second, I would like to thank my thesis advisor, Joseph DeRisi. Among scientists, there is an understanding that the apprenticeship of science is akin to a familiar bond, and in that respect, Joe has been a tremendous scientific father to me. He has provided much encouragement and support, and I am very grateful to him for having provided me with a scientific home for these past 6 plus years (three homes, if one count the number of times the lab moved while I have been a member). The DeRisi lab has been an unparalleled learning environment for massively parallel studies, and it is a continuous draw for many great scientists who arrive, attracted by Joe's tremendous enthusiasm for science and his unflinching dedication to his ideals.

I would also like to thank my thesis committee members, James McKerrow for sharing with me his infectious love of parasites and for his enduring support, and Hao Li for critical bioinformatics suggestions as well as for providing me with a valuable mentor, and collaborator in the form of Jeffrey Chuang who is now on the faculty at Boston College. Much of the work of Chapter 2 was only possible because of Jeff's involvement.

I would also mention here that Jim provided some of the earliest and strongest endorsements for the pursuit of the TG binding factor, and although I cannot claim that this project has been carried to completion, his recommendations provided much needed focus to my scientific studies. Keith Yamamoto is a man who operates on a different plane of existence, and I would like to thank him for creating this heaven for scientists that is the Mission Bay campus at UCSF, and also for somehow still finding the time to take an interest in my work. The words of encouragement provided by all of my thesis committee members have been a major motivating force in my pursuits during my graduate work.

Every labmate has contributed in memorable ways to my experience in graduate school, but I certainly cannot omit a few words of gratitude for Zbynek Bozdech and Manuel Llinas, two amazing human beings, whose shared work provided the foundation for what I was able to pursue. Aaron Sarver was also great company during those earlier years at Parnassus, and I have Brian Pulliam to thank for showing me the ropes of computation. More recent members of the lab have provided invaluable support as well. I could not ask for more in a baymate than Jennifer Weisman; she helped me through one of the hardest times during my PhD work. I am also grateful for having had the brief opportunity to share a bay with Charlie Kim, whose broad scientific curiosity is a joy to experience. He also saved my PhD in the 11th hour. The work that I describe in Chapter 3 relied almost entirely on parasite culturing done by Ally Liou; without her assistance, this work would not have progressed as far as it did. Polly Fordyce and Victoria Newman provided very much appreciated timely editorial assistance during the preparation of this thesis. Also, I

very much look forward to seeing how Polly will carry forward the torch of transcription after I have left the lab. Charles Chiu is an inspiring role model for any aspiring physician-scientist; he also has an amazing ability to make everything “basic” (anyone who has worked with him knows what I am talking about).

There are also some non-Derisi-lab folk who I would like to thank as well: Nilesh Shah, Clement Chu, and Anselm Levskaya have been very good late night company. Clement also helped me to get started in protein chromatography. Leslie Spector and Manny De Vera have my thanks for keeping the DeRisi lab running smoothly. Danny Dam and Sue Adams of the Tetrad office as well as Jana Toutolmin and Catherine Norton of the MSTP office have all provided staunch support for me over the years; thank you.

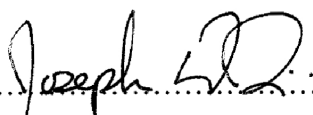
Last but not least, I would like to thank Camille Scribner for blessing my life outside of the lab. As much as one can love science, science never loves you back, and Cam has brought much meaning and joy to my existence outside of the lab. Her creative spirit and wanderlust have propelled us through numerous adventures, and I look forward to her companionship through the coming years.

I also want to thank our axolotl mascot, Two-toed Bob. In the time that I have known him, his stoic countenance has never once revealed any sign of weakness, even after an amputation he suffered at the hands of his tank-mate. In fact, this disaster actually led to the eventual regeneration of his whole arm including his previously missing middle toe! His story is a helpful reminder about the mysterious miracles of the natural world.

Thesis Advisor's Statement About Co-Author Contributions and Previously Published Work

Chapter 2 of this dissertation was carried out in collaboration with Dr. Jeffery Chuang and Dr. Hao Li. Takeshi Irie was the project leader for this work and carried out the majority of the computational experiments, the wet lab experiments, and the data analysis.

The work described in Chapter 3 of this dissertation was executed entirely by Takeshi, including project design, wet lab experiments, and analysis.

..........

Joseph DeRisi, PhD

Thesis Advisor

Table of contents

Copyright	ii
Acknowledgements	iii
Thesis Advisor’s Statement About Co-Author Contributions and Previously Published Work	vi
Table of Contents.....	vii
Absract.....	1
Chapter 1 : Introduction	2
Chapter 2 : <i>cis</i> -Regulatory Elements Controlling mRNA expression in the <i>P. falciparum</i> Intraerythrocytic Developmental Cycle	28
Chapter 2 : Figures	59
Chapter 3 : Towards the biochemical purification of the <i>P. falciparum</i> TG binding factor implicated in the IDC clock.....	70
Chapter 3 : Figures	100
References	104
Statement of library release	117

Abstract

The mechanisms of gene regulation in the malaria parasite *P. falciparum* are not well known. However, the genome sequence and existing gene expression datasets are rich resources that can aid in identifying transcriptional regulatory elements. By comparing promoter sequences and expression data in the parasite's intraerythrocytic developmental cycle (IDC) (Bozdech et al. 2003), we computationally identify 11 cis-regulatory sequence motifs whose appearance in promoters correlates with timing of expression. Defining motif activity profile as correlation of motif with expression, each motif has a sinusoidal activity profile with a period equal to that of the IDC. In several cases, the equivalent motif on the reverse strand has an identical activity profile, while other motifs display orientation specific correlations. These motifs occur in the intergenic regions of a large fraction of the genes in the genome, suggesting that they govern a large proportion of the transcriptome. Target gene predictions support this thesis, as a significant fraction of the 3518 genes transcribed periodically during the IDC can be matched to at least one of the 11 motifs. Furthermore, these motifs appear to have strong co-occurrence biases, implying that regulation is frequently combinatorial. One motif was validated by a biochemical approach, and we call this motif the TG box motif. It is predicted to be involved in transcription of 20% of the periodically transcribed genes of the IDC. We have developed a partial purification protocol which yields a candidate polypeptide which is suggested to be the sequence specific transcription factor with affinity for the TG box motif.

Chapter 1 : Introduction

A historical perspective on malaria

Malaria is a disease that has afflicted humankind since the dawn of civilization. The records of the Greek physician Hippocrates suggest that he witnessed the hallmark intermittent fevers of malaria among his patients. In addition, Chinese texts from the first millennium BC describe these symptoms (Sallares et al. 2004), and evidence of malaria has also been obtained from Egyptian mummies (Miller et al. 1994). Despite this long history between man and malaria, it was only in 1880 that Charles Louis Alphonse Laveran discovered that the disease is caused by a protozoan parasite (Guillemin 2002), now classified collectively in the genus *Plasmodium*. Soon thereafter, in 1898, Ronald Ross reported his discovery of avian plasmodia within *Anopheles* mosquitoes (Guillemin 2002), and Giovanni Battista Grassi showed that human plasmodia are also transmitted between hosts via anopheline vectors. Interestingly, the etymology of the word malaria evinces the now retired belief that polluted swamp air (*mal'aria* is Italian for bad air) causes malaria, a popular belief prior to this microbial theory of disease. This folk understanding of malaria seems not to have been entirely incorrect when one considers that mosquitoes are prevalent in swamps, where they have ample standing water to lay their eggs.

Numerous species of plasmodia are recognized, but only four cause significant morbidity in humans. These include *Plasmodium falciparum*, *P. vivax*, *P. ovale*, and *P. malariae*. Of these, *P. falciparum* is most prevalent and most deadly. It was estimated that there were more than 500 million cases of *P. falciparum* malaria in 2002 (Snow et al. 2005) and 1 million deaths in 2000 (Hay et al. 2005). Although the percentage of the world's population at risk for malaria has decreased in proportion from 77% at the turn of the 20th century to less than 50% in 1994 (Hay et al. 2004), the population growth during this same period has led to a net increase in the global population at risk for malaria from 0.9 to 3 billion (Hay et al. 2004).

Recently, human civilization reached a landmark point such that for the first time, half of the human population lives in urban settings (Bettencourt et al. 2007), and urban lifestyle is negatively correlated with malaria incidence (Hay et al. 2005). In less urban and less economically developed countries, particularly those in sub-Saharan Africa, malaria continues to exert a tremendous human cost. Awaiting urbanization and development is not an adequate solution; a more direct and intentional solution to the problem of malaria control is needed. Many governmental and non-governmental programs are currently engaged in the global battle against malaria, whether it be through the distribution of insecticide impregnated bednets or through chemotherapeutics, but they are struggling due to inadequate financial support (Feachem et al. 2007).

A history of malaria control

Armed with the knowledge of the identity of the plasmodial parasite and its vector host, many have embarked on memorable attempts to eradicate *Plasmodium* and its insect vector. For example, beginning in 1955, the WHO supported efforts to eradicate malaria through mass drug-administration campaigns via distribution of the anti-plasmodial drugs pyrimethamine and chloroquine in table salt (D'Alessandro et al. 2001). At the outset of these programs, it was not appreciated that patients would accumulate sub-therapeutic levels of drug in their blood stream, providing a perfect scenario for parasites' evolution of drug resistance. For vector control, the discovery of the insecticidal properties of dichlorodiphenyltrichloroethane (DDT) during World War II had a significant impact on the global malaria burden. DDT is cheap to synthesize, and is chemically stable. Between 1955 and 1969, the WHO recommended widespread use of DDT for mosquito abatement and some countries (such as Sri Lanka and India) saw a 99% reduction in malaria caseload (Attaran et al. 2000). The efficacy of DDT against various arthropods led to its global adoption in agricultural pest control, leading to large scale spraying. However, DDT also has a long half-life *in vivo* contributing to its deleterious effects in insectivorous avian populations. This fact was widely popularized by Rachel Carson in her 1962 book *Silent Spring*, leading to a backlash that culminated in a global ban on the use of DDT for agricultural purposes. The environmental cost of saving human lives through DDT use remains a controversial topic (Attaran et al. 2000), however in 2006, the WHO announced support for limited indoor spraying for mosquito control (Mandavilli 2006).

In addition to malaria prevention, medical treatment strategies have also been employed for much of human history. The pharmacological armamentarium can be divided into three main classes. These are the quinoline derivatives, the antibacterials, and the artemisinin derivatives.

The quinoline derivatives are chemical cousins of quinine, a traditional anti-malarial drug extracted from the bark of the cinchona tree by natives of the east Andes slopes of the Amazonian jungle. A nearly mythological story relates the miraculous cure in 1638 of the countess of Cinchon by this native remedy (Honigsbaum 2002), and this anecdote led to the popularization of tonic water, an aqueous cinchona extract albeit much diluted in its current form due to dangerous drug interactions with the antihistamine astemizole (Klausner 1996). The acquisition of a reliable source of cinchona bark became a competitive goal of French, Spanish, British, and Dutch empires throughout the 17th and 18th century. In 1820, two French chemists, Pelletier and Caventou, isolated quinine as the active anti-malarial ingredient of the cinchona extract (Kyle et al. 1974). Although they graciously shared their hard-won knowledge, the Dutch pursued cinchona for financial gain and created large plantations of high quinine content cinchona in Java, such that by 1930 they were producing 97% of the world's quinine (Honigsbaum 2002). The occupation of Java by the Japanese Empire during World War II was motivation for the investment by the American military into research programs that led to the development of many synthetic chemical congeners of quinine (Kitchen et al. 2006). World War II had another effect on the history of anti-malarials: chloroquine, the best synthetic quinoline anti-malarial drug made to date, was first synthesized in 1934 by a German scientist

working at Bayer (Greenwood 1995), but this knowledge was not widely distributed until after the war. Today, the quinoline class of drugs includes chloroquine, as well as mefloquine, primaquine, halofantrine, lumefantrine, quinidine, quinacrine, and amodiaquine. The latter two drugs are indicated for the treatment of some arrhythmias, a use consistent with the cardiac effects of quinine at higher doses. The search for safer, more effective chloroquine derivatives continues (Madrid et al. 2005).

In 1946, the nascent Centers for Disease Control (CDC) determined chloroquine (CQ) to be a safe and effective anti-malarial therapy (2004) and subsequent widespread use of CQ was followed by dramatic reduction in malaria caseload. However, since the first report of clinical cases of CQ-resistant malaria in Colombia (YOUNG et al. 1961), resistance has spread throughout the world. CQ-resistant strains are now widely endemic, and CQ is not effective in much of sub-Saharan Africa and south-east Asia (White 2004). The mechanism of drug resistance has been shown to involve the mutation of a transmembrane protein named *Plasmodium falciparum* chloroquine resistance transporter (PfCRT; (Fidock et al. 2000). Although other genetic loci were previously associated with CQ resistance ((Su et al. 1997); (Reed et al. 2000), this study was the first to show that expression of a drug-resistance associated allele of this gene could confer resistance to a previously drug-sensitive strain. Meanwhile, the molecular mechanism of action of the quinoline derivatives remains somewhat controversial. *In vitro* assays indicate that CQ is able to inhibit the polymerization of reactive free heme into the inert polymer hemozoin, a process allowing the parasite to detoxify the large quantities of hemoglobin-derived heme released in the process of consuming hemoglobin. However, this result has

the caveat that the assay was performed with a CQ concentration of 100 μ M, which is significantly higher than the mid nanomolar therapeutic concentration seen in the serum of patients treated with CQ ((Sullivan et al. 1996); (Sullivan et al. 1996)); it may be the case that the diprotonation of CQ in the acidic food vacuole increases the concentration of CQ to millimolar levels by preventing escape (Yayon et al. 1985). Another hypothesis suggests that CQ may modify of the pH of the digestive vacuole, which is an acidic compartment inside the parasite where the breakdown of hemoglobin is performed, although more recent data belie this finding(Hayward et al. 2006).

CQ and hydroxychloroquine are also used in the treatment of various autoimmune diseases including rheumatoid arthritis, and recent work has indicated that its effect in abatement of autoimmunity may be due to its effect on suppression of the innate immune system through the toll-like receptor (TLR) pathway ((Kyburz et al. 2006); (Brentano et al. 2005)). It is unknown whether a component of the host immune response is a requisite growth cue for the malaria parasite, as has recently been documented for the trematode *Schistosoma mansoni* (Davies et al. 2001).

Mefloquine (Larium) is well known for its psychiatric side effects, but is often prescribed as a prophylactic agent for travelers to malarious countries because of its long serum half-life. Resistance to mefloquine is becoming a significant problem, particularly among *P. falciparum* in the Mekong peninsula, and it is associated with amplifications but not mutations of the multi-drug resistance transporter (PfMDR1; (Price et al. 2004)), but cross-resistance exists between members of this drug class as well as with artemisinin

(more below). It is interesting that a calcium channel blocking agent, verapamil, is able to reverse the CQ resistance of some strains (Lakshmanan et al. 2005), and that the proposed target of artemisinin is a calcium channel (more below). There may be room for speculation that some day there will be a unifying theory of the mechanism of action for multiple classes of anti-malarial action through disruption of parasite calcium homeostasis. Ultimately, more work is needed to improve the understanding of the mechanism of action of the quinoline class of drugs.

Some anti-bacterial drugs are effective anti-malarials. This class of drugs includes the antifolate drugs sulfadoxine and dapsone that inhibit the enzyme dihydropteroate synthase (DHPS), as well as proguanil and pyrimethamine which target dihydrofolate reductase (DHFR). Other antibacterials in this class include the DNA gyrase inhibitor ciprofloxacin, the RNA polymerase inhibitor rifampicin, and the translational inhibitors thiotrepton and doxycycline (Dahl et al. 2006). Atovaquone was recently shown to work through disruption of ubiquinone regeneration (Painter et al. 2007). Much recent work in anti-malarial development has focused on the apicoplast as a drug target, given that this organelle is absent in human cells. The antibacterial fosmidomycin inhibits DXP reductoisomerase in the non-mevalonate pathway of isoprenoid synthesis within the apicoplast (Jomaa et al. 1999), and triclosan may target fatty acid metabolism (Surolia et al. 2001), though the veracity of this latter study have recently come under question (Weisman 2007).

Artemisinin represents another promising class of anti-malarial drugs. The original member of this class of drugs, artemisinin, is a terpene extracted from the wormwood, *Artemisia annua*, a plant that grows as a weed in many parts of the world. In 1972, the Chinese military scientists determined the active ingredient of the ancient herbal remedy *qinhaosu* (Chinese for “active principle of green herb”) first described in Chinese medical texts dating to the second century BC (Klayman 1985);(Woodrow et al. 2005). By the time this knowledge was translated to western scientists, the Chinese had already synthesized the first two derivatives, artemether and artesunate, and conducted animal studies as well as trials in human patients(Bruce-Chwatt 1982). The activity of artemisinin was confirmed by Walter Reed Army Medical Center scientists who found artemisinin producing strains of *A. annua* growing near Washington D.C. (Klayman 1985). Since then, other derivatives have been made including arteether, artelinic acid, artemisone, and dihydroartemisinin, which is the active metabolite of all artemisinin derivatives. The mechanism of action of the artemisinin is thought to be through the inhibition of the sarcoendoplasmic reticulum type Ca^{2+} ATPase (SERCA) PfATP6(Eckstein-Ludwig et al. 2003).

Despite the lack of rigorous clinical trials for the artemisinins, an estimated 1-2 million patients had been safely treated with these drugs with particularly good effect against CQ-resistant *P. falciparum* as of 1997(Davis et al. 1997). Although neurotoxic in animal models (Brewer et al. 1994), such toxicity has not been a significant problem in human patients so far though a rare case has been reported (Miller et al. 1997). Consequently, in 2001 the WHO began recommending artemisinin combination therapy (ACT) as first line

of treatment in countries where CQ resistant malaria is endemic; by 2004, 40 countries had begun to recommend ACTs as the first line treatment (Mutabingwa 2005).

Currently, the main problems with artemisinin treatments are the high cost (\$2.40 per dose compared to \$0.1 for quinine; (Towie 2006)) and the short serum half-life (1 hour for artesunate and dihydroartemisinin; (Teja-Isavadharm et al. 2001)). Due to the comparatively high cost, distribution of counterfeit artemisinin on the black market has become a significant problem, and it has been reported that 38-52% of the drugs sold as artemisinin in southeast Asia do not contain active artemisinin (Newton et al. 2007).

Recent explorations of synthetic biology approaches to reducing drug cost are promising (Ro et al. 2006). However, the short serum half-life of drug demands a more frequent dosing regimen, which is associated with lower patient compliance. Poor adherence will expose parasite populations to sub-lethal doses of drug, and this mechanism virtually assures that monotherapy will meet with rapid emergence of drug resistance (White 2004). In fact, *in vitro* studies suggest that a single amino acid change in PfATP6 may be sufficient to decrease the sensitivity by almost 300 fold (Uhlemann et al. 2005). In 2006, *in vivo* resistance to the combination artemisinin-mefloquine was first documented at the Cambodia-Thailand border, with 7.91 times the risk of recrudescence associated with parasite strains having greater than three copies of the PfMDR1 gene (Alker et al. 2007). So far, these studies implicating PfATP6 in the mechanism of artemisinin action and PfMDR1 in resistance have not been confirmed with the experimental gold standard of allelic replacement with mutant drug-resistant alleles. In the meantime, the WHO recommends combination therapies to stave off the arrival of resistance. Several

artemisinin combination therapies (ACTs) have been found to be safe and effective and are currently recommended by the WHO (Table 1; (WHO 2006)). Additional combination therapies (such as dihydroartemisinin-piperaquine) have been deemed safe and effective in recent clinical trials (Ashley et al. 2004), but are not officially recommended at this time due to poor availability of such formulations. Despite these problems, artesunate remains the best choice in the face of drug-resistant severe malaria, and the CDC announced this summer that it has allowed for an investigational new drug application (IND) to be filed for intravenous artesunate treatment for severe malaria, making the drug legally available in the US for the first time (Division of Parasitic Diseases 2007).

Table 1

WHO recommended artemisinin combination therapies
artemether – lumefantrine
artesunate – amodiaquine
artesunate – mefloquine
artesunate – sulfadoxine – pyrimethamine

Although chemotherapy and chemoprophylaxis options have increased somewhat over the years, no effective vaccine has been made to date. There are currently more than thirty vaccine trials currently underway (Van de Perre et al. 2004). However, these trials target epitopes mapping to plasmodial surface proteins known to be highly variable from

strain to strain; it remains to be seen to what degree selection and evolution of vaccine resistance might become problems.

Research in *Plasmodium* and malaria biology

Despite all of the earnest attempts to control malaria, we appear to be somewhat distant from besting this disease. In the meantime, more research must be done to understand both the *Plasmodium* parasite and the disease, malaria. Critical advances in *P. falciparum* *in vitro* culture methods (Trager et al. 1976) laid the foundations for such work as described below.

The parasite life cycle

Like its human host, the *Plasmodium* lifecycle has a sexual and asexual component. The effects on human health are felt during the asexual stage, when the parasites are haploid. As described above, the parasites are transmitted between hosts by the *Anopheles* mosquitoes. The initial infection of the human host is carried out by a parasite in a motile stage called the sporozoite (reviewed in (Prudencio et al. 2006)); it has been estimated that 15-123 sporozoites are sufficient to ensure an infected host. Within 30 minutes of entering human tissue, the sporozoites migrate through the layers of the dermis until finding a capillary, then circulate until they are able to invade a hepatocyte via a gliding motility mechanism (Sultan et al. 1997). The recognition of the hepatocyte is mediated by several parasite surface proteins (Prudencio et al. 2006) including circumsporozoite

protein (CSP), thrombospondin-related anonymous protein (TRAP), and apical membrane antigen 1 (AMA-1). Heparan sulfate proteoglycans (HSPGs) on the surface of the hepatocyte seem to be a requirement for infection. HSPG is present on many cells in the human body, but is more highly sulfated in hepatocytes and the sulfation of HSPG has been shown to be essential for sporozoite adhesion (Prudencio et al. 2006). Inside the hepatocytes, the parasites develop into tissue schizonts which undergo successive rounds of replication to give rise to as many as 10,000 merozoite progeny. In a mouse model of malaria, *P. berghei*, they are released by the budding of host membranes derived merozoite-filled vesicles called merosomes (Sturm et al. 2006); whether this route of exit is used by the human malarias is currently under investigation. The merozoites are markedly different in morphology from the sporozoites.

Merozoites are notable for the organelles at their apical end, including rhoptries, micronemes, and dense granules. These organelles are the defining feature of the family *Apicomplexa*, which include *Plasmodium* as well as the related intracellular parasites, *Toxoplasma gondii*, *Cryptosporidium parvum*, and *Theileria parva*. Much work has been done on the receptors used by *P. falciparum* merozoites for invasion. Through the use of numerous variants of (at least) two families of surface proteins including the Duffy binding protein family (erythrocyte-binding antigens, EBAs), as well as the reticulocyte binding protein family, the parasite is able to avoid immune clearance (Baum et al. 2005).

The merozoite invasion of the host erythrocyte marks the beginning of the asexual intra-erythrocytic developmental cycle (IDC; for background regarding the IDC, see (Waters et

al. 2004) and (Sherman 1998)). Each *P. falciparum* merozoite can invade a new erythrocyte, and within 48 hours, replicate itself into 16-32 progeny merozoites. During this IDC, several morphological changes can be noted by microscopy using Giemsa stain. Initially, the parasites are small and notable for a ring-like appearance where the circular demarcation of the parasite boundary is punctuated by a dot that corresponds to the compact nucleus. This stage is called the ring stage and for the approximately 20 hour duration of this portion of the IDC, the parasite has relatively low metabolic activity. In the second stage, a food vacuole begins to form due to active ingestion and catabolism of host hemoglobin (Goldberg 2005). After sufficient growth, the trophozoites begin the process of DNA replication and division, and this final stage of the erythrocytic parasite is called the schizont stage. Late stage schizonts are notable for fully formed individual merozoites contained inside the erythrocyte, awaiting release, and these parasites are often referred to as segmenters. Mechanisms controlling the progression through IDC are likely to be genetically encoded, but remain a very engaging scientific mystery at the heart of the motivation for my studies.

The IDC is particularly unusual for its hybrid developmental and cell cycle. Unlike a classical cell cycle in which controlled DNA replication occurs once and only once (Alberts 2002), the *Plasmodium* IDC incorporates multiple rounds of DNA replication in a process called cryptomitosis (Merckx et al. 2003), in which the nuclear membrane remains intact, and the daughter genomes of the first replication continue to undergo further rounds of replication in a seemingly unsynchronized manner (Doerig et al. 2004). This asynchronous replication produces a distribution of genome copy numbers that

includes odd numbers, rather than the doublings of $1n$, $2n$, $4n$, $16n$, and $32n$ that would be expected for a fully synchronized process. The parasites utilize cyclins and cyclin-dependent kinases (CDKs) that are homologous to those of other eukaryotic organisms (Doerig et al. 2002).

At a frequency of less than 1% in *in vitro* culture (Eichner et al. 2001), some of these asexual forms undergo a switch to a distinct developmental program called gametocytogenesis leading to the development of male and female gametocytes. The commitment to these sexual stages is suspected to occur during the trophozoite stage of the previous cycle (Smith et al. 2000), but the full course of development to gametocytes takes 14 days. Gametocytes are also known to sequester, preferring to reside in bone marrow (Rogers et al. 2000). Recent work has shown that a calcium dependent protein kinase, CDPK4 (Billker et al. 2004), and a male-specific mitogen activated protein kinase 2, Pfmap-2 (Rangarajan et al. 2005), are both required genetically for normal male gametocytogenesis.

Upon completion of development, the gametocytes circulate until taken up by another anopheline host, at which point the final transformation happens: during this process known as gametogenesis, both male and female gametocytes shed their erythrocyte exterior and the male gametocyte undergoes exflagellation. It has been found that the drop in temperature, rise in pH, and the presence of xanthurenic acid in the mosquito midgut all contribute to the rapid transformation that leads to mating competency (Billker et al. 1998). Within 30 minutes, the two haploids mate, leading to fertilization and the

development of the zygote. The zygote then develops into the oocyst, which embeds in the wall of the host gut and undergoes asexual divisions once again, in a process called sporogony to generate 1000~10,000 sporozoites (Beier et al. 1998). With the help of the *Anopheles* vector, these sporozoites are competent to invade a new host, thus closing the transmission cycle.

A very brief description of the pathophysiology of malaria

If one parasite were estimated to multiply by tenfold, a population of 10 parasites expands to approximately 10^8 parasites within 7 cycles, or 14 days (White 2004). Assuming 2L of packed erythrocytes per person, this parasitemia of 50 per uL is the detection limit by microscopy. Non-immune patients may report symptoms two days prior to this parasitemia, but between parasitemias of 10^{10} and 10^{11} , most patients are typically symptomatic. Typical symptoms include the famous periodic fever, which corresponds to the host response to synchronized merozoite release as first described by Camillo Golgi (Muscatello 2007). Other common symptoms include mild anemia, fatigue, nausea, vomiting, headaches, myalgia, chills, sweats, diarrhea, and cough. Golgi's suggestion that malarial fevers of differing periodicity correspond to infections by different species has been validated: *P. falciparum*, *P. vivax*, and *P. ovale* have an approximately 48 hour IDC duration, while *P. malariae* has a 72 hour IDC (Bray et al. 1982).

Clinical observations reveal that parasites of the later stages are often not detectable in blood smears from patients, due to the adhesive properties of the trophozoite and schizont stages. Stage specific expression and surface presentation of key adhesion molecules by the parasites allow them to bind and sequester in the microvasculature through a process called cytoadherence. Sequestration from circulation, allows parasites to develop without the risk of splenic clearance, which is a major mechanism by which the host controls *Plasmodium* infection; splenectomized patients are at significantly increased risk for malaria (Bach et al. 2005). The adherent properties of plasmodia contribute to the complications of severe malaria including cerebral malaria (MacPherson et al. 1985) and placental malaria (Fried et al. 1996), in which parasites adhere to vasculature of the brain and placenta respectively.

This host-parasite interaction has been defined at a molecular level; cytoadherence depends upon variants of the protein family *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) ((Baruch et al. 1995); (Su et al. 1995); (Smith et al. 1995)), encoded by the *var* genes present in roughly 100 copies per haploid genome (Su et al. 1995). The host molecules currently known to interact with PfEMP1 are several, including intracellular adhesion molecule 1 (ICAM-1; (Smith et al. 2000)), chondroitin sulfate A (Pouvelle et al. 2000), and CD 36 (Baruch et al. 1995). Each clone of parasite has been observed to rely on the expression of one *var* gene in a process called allelic exclusion ((Voss et al. 2006); (Deitsch et al. 2001)), and switching expression between the variant copies allows parasites (Horrocks et al. 2004) to utilize different surface proteins to accomplish adhesion, thereby evading the host immune response (Peters et al.

2002). Infected erythrocytes are known to use PfEMP1 to adhere to uninfected erythrocytes as well, in a process called resetting which has been suggested to mask the surface proteins of *Plasmodium* that are immunogenic, and simultaneously decrease the transit time for newly released merozoites to find fresh hosts (Rowe et al. 1997).

The genome era of *Plasmodium* research

Genome sequencing and functional genomics of *P. falciparum*

Given the tremendous global health impact of the disease, *P. falciparum* was an obvious candidate for whole genome sequencing, and the completed sequence and annotation were released in a series of articles in 2002 ((Gardner et al. 2002); (Hall et al. 2002); (Hyman et al. 2002)). The haploid genome is 23 Mb and organized into 14 chromosomes, coding for approximately 5300 genes (summary of genome sequence : (Gardner et al. 2002)). The sequence is 80% AT rich, rising to ~90% in non-coding regions. There are also two organellar genomes, both of which are circular. The mitochondrial genome is ~6 kb and encodes only three open reading frames (ORFs), while the apicoplast genome is approximately 35 kb. Genome sequences, and data from functional genomics published to date are organized and curated at www.plasmodb.org (Bahl et al. 2003).

With the information provided by the genome sequence, new approaches are available for the study of *Plasmodium* biology. A pioneering microarray study of the transcriptome of *P. falciparum* was published in 2000 by Rhian Hayward in collaboration with my thesis

advisor, Joseph DeRisi (Hayward et al. 2000). This study was done on a spotted microarray (Churchill 2002) of a genomic library from *P. falciparum* using a “shotgun” microarray strategy. Briefly, the library was constructed using a mung bean nuclease treatment, which preferentially cleaves AT-rich sequences, leaving the coding regions relatively intact. PCR products were generated from this library using universal primers flanking the inserts, and these PCR products were printed onto glass slides treated with poly-lysine to capture nucleic acid material using electrostatic interactions. cDNAs representing transcripts of different stages of parasites, then hybridizations were performed on this approximately 3648 feature microarray. Clones of interest were analyzed by followup sequencing. As part of my rotation project during the summer of 2000, I sequenced some of the clones from this library that had not been previously annotated. This study corroborated previous work revealing stage-specific gene expression patterns for a small number of genes, and also found many new genes exhibiting distinct expression patterns between trophozoite gametocyte stages.

The publication of the *P. falciparum* genome sequence allowed a more sophisticated approach to construction of the second generation *P. falciparum* microarray. This microarray was designed such that the identity of the printed spots would be known *a priori*, due to intentional design of oligonucleotide probes. This strategy was a significant advance compared to the “shotgun” microarray as it obviated the need for sequencing of clones post-hybridization for the determination of the association of an ORF to a spot on the array. The algorithm for microarray design was developed by Jing Zhu (Bozdech et al. 2003) and is based on computed thermodynamic energies of hybridization of 70mer

oligonucleotide probes. This array included oligos for 4,488 of the predicted 5,409 ORFs, allowing for a more thorough investigation of the IDC transcriptome culminating in the publication of a high resolution (1-hour resolution) time course of the 48-hour *P. falciparum* IDC (Bozdech et al. 2003). This landmark study afforded many observations regarding transcription in the *P. falciparum* IDC, including evidence that approximately 60% of the genome is transcriptionally regulated during the asexual phase of the lifecycle, and that some 80% of these genes (or just under 50% of the entire genome) are expressed in a periodic manner through the course of the 48 hours of development. For comparison, analysis of the *Saccharomyces cerevisiae* genome displaying periodic transcription during a cell cycle comprised of only 3.8% of the genome. An analogous study was published by another group using a gene-specific microarray platform implemented by Affymetrix using proprietary photolithographic printing methods, confirming many of the observations made in our laboratory (Le Roch et al. 2003). Their study extended the breadth of its investigation to include expression from the gametocyte and sporozoite stages, although the IDC data were obtained with coarser time resolution.

It is worth noting that the availability of genome sequence has also facilitated several non-microarray based functional genomics studies including proteomics (Florens et al. 2002), yeast two hybrid (LaCount et al. 2005), as well as cDNA and EST sequencing projects (Watanabe et al. 2004), however they will not be discussed in detail here.

The goal of this work

When I commenced the work described in the following chapters, it was my hope that these studies would provide a better understanding of how the transcriptional control of the development of *P. falciparum* is governed during the intraerythrocytic developmental cycle (IDC). In the near term, the goal of this project has been to discern the components of the molecular clock that dictate the periodicity of the intermittent malarial fevers.

To begin, one can gain some insight about how to approach the problem by examining work done on various other biologically encoded clocks. The cell cycle is a well described molecular clock that has been analyzed with respect to gene expression in several model eukaryotes, including *Saccharomyces cerevisiae* (Spellman et al. 1998), *Schizosaccharomyces pombe* (Rustici et al. 2004), and human fibroblast cells (Cho et al. 1998; Cho et al. 2001). In all of these examples, the developmental program utilizes a series of transcriptional cascades, in which stage-specific transcription factors (TFs) govern the expression of TFs controlling the succeeding stages. In addition, the dimorphic bacterial cell cycle of *Caulobacter crescentus* (Holtzendorff et al. 2004) and the metazoan circadian cycle (McDonald et al. 2001; Miller et al. 2007) also employ TFs as components of their oscillating regulators. Partly modeled on these systems, a completely synthetic transcriptional oscillator has been constructed *de novo* (Elowitz et al. 2000). On the other hand, there is a major post-transcriptional aspect of the cell cycle, wherein regulated proteolysis of cyclins determines the activity of their cyclin-dependent kinase (CDK) partners that have numerous substrates in the cell (Alberts 2002). Also, the cyanobacterial circadian clock was elegantly recapitulated *in vitro* using only the three central proteins, with the phosphorylation status of the KaiC protein acting as the

temporal state indicator (Nakajima et al. 2005). Although these two models differ in their molecular implementation, they can in theory be used to similar effect towards the construction of biological rhythm, and in fact they are not mutually exclusive. Given the timely availability of the transcriptome data of the *P. falciparum* IDC when I embarked, I chose to tackle the problem of IDC regulation at the level of transcription.

Transcription is a good starting place for approaching this problem, given the primacy of transcription in eukaryotic gene expression. Some work has been done on the study of transcription in *P. falciparum* demonstrating the conservation of many of the basic principles learned from the study of other model eukaryotes (Alberts 2002). For example, genome sequence analysis reveals the presence of components for the three major polymerase complexes, RNA polymerase I, II, and III (RNAP I, II, III; (Coulson et al. 2004; Callebaut et al. 2005; LaCount et al. 2005)). While RNAP I transcribes rRNA, and RNAP III transcribes tRNA and snoRNAs, the bulk of cellular mRNA is generated by RNAP II. Classically, RNAP II promoters contain a TATA-box in close proximity to the site of transcription initiation, and the binding of TATA-box binding protein (TBP) to a TATA-box leads to the recruitment of the RNAP II complex to these core promoters (Alberts 2002). Binding of a *P. falciparum* TBP in proximity to the transcription initiation sites of two promoters has been experimentally confirmed (Ruvalcaba-Salazar et al. 2005). The specificity of RNAP II is also mediated through interactions with general transcription factors (GTFs), and some of these were identified in the *P. falciparum* genome by bioinformatics as well (Callebaut et al. 2005). Interestingly, these authors report that one of the GTFs, TFIIA, is duplicated in the *P. falciparum* genome

(Callebout 2005); although duplication of this gene has not been observed in other genomes, existence of the duplication and subsequent acquisition of promoter-specific functions of the TBPs of animals has been documented (Holmes et al. 2000; Davidson 2003). Another law of eukaryotic transcription is that short *cis*-regulatory elements called enhancers located at some distance from the core promoter are bound by sequence-specific DNA binding proteins, called transcription factors (TFs), and these TFs can recruit the RNAP II complexes to specify transcription (Miesfeld et al. 1987). However, bioinformatics analyses have failed to find this class of regulators at expected abundances, implying they are relatively under-represented in the *P. falciparum* genome, or that they are evolutionary divergent from other eukaryotic transcription factors (Coulson et al. 2004). Regulation of mRNA stability represents an important means by which transcript abundance is controlled, and global stage-specific differences in mRNA half-lives was recently documented (Shock et al. 2007) but this mechanism of gene regulation will not be discussed in detail here.

Another major mechanism of gene expression is through chromatin regulation. Several recent findings regarding *var* gene expression highlight the primacy of epigenetic mechanisms of antigenic gene regulation in this parasite:

- 1) PfSir2, the histone deacetylase homolog of the yeast telomere silent information regulator, was demonstrated by ChIP to be associated with silenced subtelomeric *var* loci, while acetylated histone H4 was associated with expressed loci (Freitas-Junior et al. 2005).

- 2) Genetic ablation of PfSir2 led to upregulation of a subset of var and rifin genes encoded in subtelomeric regions (Duraisingh et al. 2005).
- 3) Trimethylation of histone H3 at lysine K9 was enriched at silenced var loci relative to actively transcribed ones (Chookajorn et al. 2007).

Furthermore, work on the related Apicomplexan, *Toxoplasma gondii*, demonstrated that the histone deacetylase TgGCN5 from this organism associates with promoters in a stage-specific manner, and that inhibition of the histone arginine methyltransferase TgCARM1 using a chemical inhibitor AMI-1 results in decreased global H3 methylation which correlates with a change in the developmental course of the parasite (Saksouk et al. 2005). Recruitment of chromatin by sequence-specific transcription factors is a long established mechanism of eukaryotic gene expression (Zaret et al. 1984). The possibility that a sequence-specific DNA binding protein interacts with nucleosomal proteins, thereby specifying the chromatin and transcriptional status, will be an intriguing hypothesis for future work contingent on the identification of the TGBF.

The approach

I have taken a two-pronged approach towards understanding the molecular clock of *P. falciparum*. The first approach is a computational analysis of the *cis*-regulatory elements correlating with transcription during the IDC. The second approach is a biochemical validation of one of these motifs demonstrating recruitment of a *trans*-acting factor that regulates the genome through binding of this motif. The computational analyses and the

initial validation experiments are described in detail in chapter 2. Work towards a purification strategy for the eventual identification of this factor and this work is described in chapter 3.

Computational biology has seen a renaissance in parallel with the advent of large-scale DNA sequencing technologies. Many years of experimental work on determining the mechanisms of transcriptional regulation have shown that much of transcription is typically regulated by proteins, called transcription factors (TFs), which bind DNA with high affinity for specific sequences of small size. These transcription factor binding sites (TFBSs) are typically found enriched in the upstream regulatory sequences of genes, called promoters, and the binding of TFs to their TFBSs leads to the recruitment of the RNA polymerase to the promoter and subsequent transcription of that gene. These TFs have a remarkable ability to bind specific sequences of short length, with zinc-finger TFs discriminating sequences as short as three nucleotides (Rebar et al. 2002). It has also been observed that promoters regulated by a particular TF typically have multiple copies of a transcription factor binding site (TFBS) (Struhl 1982). Bioinformatics tools have been developed to capitalize on this property, and statistical analysis of the frequencies of small subsequences has been used to predict functional TFBSs. One early algorithm employed simple counting approaches to look for over-representation of discrete subsequences within promoter sequences of interest (van Helden et al. 1998). Another approach modeled TFBSs as a matrix of nucleotide probabilities, called a position weight matrix (PWM), to capture the information of TF affinities for related variants of discrete subsequences (Stormo et al. 1989). A widely used algorithm, MEME, was developed

based upon the PWM model using resampling of promoter sequences to build PWMs capturing the over-representation of the TFBSs (Bailey et al. 1995). With the availability of large genome-wide transcription datasets from microarray studies, an algorithm named REDUCE was developed that uses linear regression of motif copy number to microarray gene expression data (Bussemaker et al. 2001). I have used this last algorithm extensively in the analysis of the IDC transcriptome dataset.

One motif predicted using the REDUCE algorithm was validated through the use of classical biochemical methods. The electromobility shift assay (EMSA) is a solution-based assay used to demonstrate the existence of a TF with affinity for a particular sequence of DNA (Fried et al. 1981; Garner et al. 1981). Briefly, a binding site for a TF is radioactively labeled to create a probe, and then incubated with a TF suspected to bind this sequence. The mixture is then separated with electrophoresis through a polyacrylamide or agarose gel. If the TF binds to the DNA, it will retard the migration of the DNA, revealing a band distinct from that of the unbound DNA. The source of the TF can be a purified protein if such can be obtained, or a complex mixture derived from a cellular extract. This strategy has been utilized in the study of *Plasmodium* promoters in the past, taking native promoters and dissecting the sequence into smaller regions retaining sequence specific binding activity. The utilization of the computationally guided approach towards probe selection has allowed me to validate a short oligonucleotide likely to be a close approximation of the idealized binding site for one of the TFs regulating the IDC.

After developing an assay demonstrating the presence of a factor with sequence-specific DNA binding activity within extracts prepared from *P. falciparum*, I have explored the use of various chromatographic techniques to pursue the assembly of a purification strategy which would allow for the identification of this candidate TF predicted by bioinformatics analysis to regulate a significant portion of the transcription of the *P. falciparum* IDC.

Chapter 2 :

cis*-Regulatory Elements Controlling mRNA expression in the *P. falciparum

Intraerythrocytic Developmental Cycle

Abstract

The mechanisms of gene regulation in the malaria parasite *P. falciparum* are not well known. However, the genome sequence and existing gene expression datasets are rich resources that can aid in identifying transcriptional regulatory elements. By comparing promoter sequences and expression data in the parasite's intraerythrocytic developmental cycle (IDC) (Bozdech et al. 2003), we computationally identify 11 *cis*-regulatory sequence motifs whose appearance in promoters correlates with timing of expression. Defining motif activity profile as correlation of motif with expression, each motif has a sinusoidal activity profile with a period equal to that of the IDC. In several cases, the equivalent motif on the reverse strand has an identical activity profile, while other motifs display orientation specific correlations. These motifs occur in the intergenic regions of a large fraction of the genes in the genome, suggesting that they govern a large proportion of the transcriptome. Target gene predictions support this thesis, as a significant fraction of the 3518 genes transcribed periodically during the IDC can be matched to at least one of the 11 motifs. Furthermore, these motifs appear to have strong co-occurrence biases, implying that regulation is frequently combinatorial. One motif was validated by a biochemical approach, and we call this motif the TG box motif. It is predicted to be involved in transcription of 20% of the periodically transcribed genes of the IDC.

Introduction

Malaria is a disease caused by parasitic protozoa, of which *Plasmodium falciparum* is the most deadly species, and is spread by mosquitoes of the genus *Anopheles*. While the disease has been largely controlled in temperate regions of the world, it remains a massive health problem in many tropical regions (Hay et al. 2004). Previously effective prevention and treatment strategies have also waned in efficacy, e.g. because of the realization of environmental side effects of DDT (Carson 1962) and the nearly concomitant emergence of drug resistance to chloroquine (Young et al. 1961).

Functional genomics and bioinformatics may be particularly useful for studying *P. falciparum*, since their parasitic nature and unusual base composition can make them refractory to classical genetic manipulations. An important recent genome-scale study is the characterization of the stage-specific transcription patterns for the genes associated with the intraerythrocytic developmental cycle (IDC) (Bozdech et al. 2003). At least 2714 genes, or approximately 50% of all annotated *P. falciparum* genes, have been found to be expressed in the IDC in a simple sinusoidal program, in each case with a ~ 48 hour period, with variation in phasing. Functional groups of genes in the IDC have also been shown to be transcribed synchronously (Bozdech et al. 2003).

Despite these recent strides, the molecular mechanisms of transcriptional regulation are not yet well understood in plasmodia. A handful of studies have experimentally dissected individual promoters to identify *cis*-regulatory sequence elements (Horrocks et al. 1999)

(Chow et al. 2003) (Lanzer et al. 1992; Voss et al. 2003). One recent work has also identified binding sites for the *P. falciparum* TATA-binding protein in the promoters of two independent genes (Ruvalcaba-Salazar et al. 2005). However, compared to the abundant gene expression data, knowledge of the *cis*-regulatory program is minute.

Several works have suggested that the regulatory program of the IDC may be relatively simple, requiring only a handful of transcription factors (Bozdech et al. 2003; Coulson et al. 2004). This hypothesis is supported by the comparison of conserved transcription associated proteins across multiple eukaryotic genomes. Such analysis showed that approximately 10% of the gene repertoire of a typical eukaryote is dedicated to transcription associated proteins (Coulson et al. 2003), whereas *P. falciparum* utilizes a much smaller proportion, estimated at 3% (Coulson et al. 2004). Furthermore, only 1.3% of the *P. falciparum* genome matches to the registered transcription factors in the TRANSFAC database, whereas *S. cerevisiae* has a 4.0% match rate. Experimental evidence for conservation of paired transcriptional *cis*- and *trans*-regulatory components between *P. falciparum* and non-plasmodial species exists only for the TATA box (Ruvalcaba-Salazar et al. 2005) and the Myb family of transcription factors (Gissot 2005). Hence, understanding the mechanisms of gene regulation in the *P. falciparum* IDC, in addition to being important, may also be tractable.

Given these motivations, we used computational approaches to identify candidate *cis*-regulatory motifs associated with the IDC and validate them experimentally. In contrast to previous studies of single promoters, we made use of the genome-wide sequence and

IDC expression data to identify candidate core *cis*-regulatory motifs in the non-coding regions of the *P. falciparum* genome. These motifs were interesting because they occur in high copy number and tend frequently to co-occur. Furthermore, the motifs we have detected agree well with previously described protein binding sites. We experimentally validated one of these motifs for binding of a nuclear factor through the IDC time course, and predict a broad role of this factor in the regulation of a large set of genes. The similarity of our predicted motif with a trophozoite stage enhancer in the promoter of *var* gene introns suggests that the transcriptional silencing of *var* genes is integrated into the global transcriptional program of the IDC.

Materials & Methods:

Expression data pre-processing.

We used the quality controlled expression data set from the 48-hour IDC timecourse (available at <http://malaria.ucsf.edu>) for analysis. Negative expression ratios were filtered, and all data were \log_2 transformed. Missing values were filled using KNN Impute with $k=10$ (Troyanskaya et al. 2001), and instances of data for multiple oligos mapping to single ORF were averaged, using the CAST algorithm to exclude outlier oligos below a cutoff threshold of 0.8 (Ben-Dor et al. 1999).

Promoter sequence identification.

For each ORF, the 1kb upstream of the ATG was designated as the promoter sequence, using the ORF coordinates from the PlasmODB annotation version 4.1. In cases where another ORF or contig break occurs in this 1kb (Milgram 2004), the promoter sequence was truncated at that point. 11 sequences were excluded because their predicted ORFs spanned contig breaks, and 19 sequences were excluded because their intergenic UTR lengths were 50 bases or less. The final sequence file used for analysis contained 5304 sequences; 3518 of these sequences were associated with expression data. 3' intergenic sequences were extracted as well using the 1kb and ORF truncation/contig breakpoint criteria, with 2975 sequences associated with expression data. Promoters of up to 3 kb or up to 5kb were also extracted, with results similar to those for the 1kb analysis (data not shown).

Motif detection and analysis.

Promoter sequences and processed expression data from the IDC time series were analyzed for motifs using the REDUCE algorithm (Bussemaker et al. 2001). Given a single expression array, REDUCE searches for sequence motifs whose frequency of occurrence in promoters correlates with the expression level of the adjacent gene.

REDUCE outputs a p-value for each motif, under the expectation that correlation values for non-functional motifs will be Gaussian distributed.

Motifs detected using REDUCE were characterized over the IDC time course via their activity level:

$$S_{mt} = -(\log_{10} p)_{mt} * \text{sign}(m),$$

where p is the REDUCE p-value, t is the time in the IDC, m is the motif, and $\text{sign}(m)$ takes a value of +1 or -1, depending on whether the motif correlates with enhanced or repressed gene expression. S therefore measures the magnitude and direction of the motif's effect on expression. Motif activity profiles (the activity of a motif over all time points t) were sorted using an ordering by phase, analogous to the phase ordering of gene expression profiles described in Bozdech et al. 2003 (Bozdech et al. 2003).

Motif comparisons and clustering

REDUCE motifs longer than 3 nucleotides were compared pairwise for sequence similarity, in order to group motifs which are variants of the same transcription factor binding site. For each pair of motifs a and b , with lengths l_a and l_b , the two motifs were aligned in all possible frames in which the motifs overlapped. For a given frame, the score was evaluated as $\sum_i \sigma(a_i, b_i)$, with i indicating a position in the alignment for which both sequences have a base, and with the function s as described below. The overall score was taken to be the maximum score over all possible alignment frames, with a Bonferroni correction for the number of frames ($-\log(l_a+l_b-1)$). The values of $s(a,b)$ were chosen to prevent biases toward long alignments of random sequences, i.e. $s(a,b) = \log [P(a,b)/q_a q_b]$ (Durbin 1998). Here q_a is the frequency of base a in random promoter sequence ($q_A = 0.4224$, $q_C = 0.0620$, $q_T = 0.4504$, $q_G = 0.0652$) and $P(a,b)$ is the frequency

of the pairing of a and b in motifs which are variants of the same binding site. For motifs describing variants of the same transcription factor binding site, it was assumed that mismatches would occur at a rate $g = 0.1$ for unrelated motifs ($P(a,b) = gq_aq_b$) and that matches would occur at a frequency proportional to that of the base in question ($P(a, a) = fq_a$). These constraints, along with the requirement that the $P(a,b)$ values sum to unity, were sufficient to determine the score function.

Motif similarity scores were then used to cluster motifs. Similarity scores were converted to distance scores via a linear transformation that mapped maximal similarity scores to a distance of zero and the minimal similarity to 1, and UPGMA clustering was performed via the program NEIGHBOR in the PHYLIP package (Felsenstein 2005). Note that although we assumed a particular value of g , the inferred clustering tree was insensitive to variations in g over a wide range (0.05 to 0.4).

Prediction of target genes and flanking sequence information.

The target genes of the significant motifs were predicted using the MODEM algorithm, which uses a maximum-likelihood approach to distinguish the subset of motif instances likely to be true targets for a transcription factor (Wang et al. 2005). MODEM takes a consensus core motif, promoter sequences, and a gene expression microarray dataset as inputs, and outputs a set of likely targets, as well as a position-specific weight matrix (PWM) describing the flanking sequence around true targets. We modified the MODEM protocol to use the full time-course microarray data. Rather than use expression data from

an individual experiment, we used pseudo-expression values that measured the coherence of the motif activity with gene expression, i.e. pseudo-expression values were set to be the dot product of the motif activity profile and the mRNA expression time course for each gene. In the MODEM inputs, the mean of the true target pseudo-expression values was initialized to be three standard deviations greater than the mean of the background.

The following 11 core motif sequences were analyzed by this method: TGTG, TATAA, TATAGA, TGTCT, TGCA, TTGT, AATT, TTTT, TGTAG, CTTA, and GATAC. These representative sequences were chosen because, in any given sequence family, these sequences had the highest value of S_{mt} over all time points. Note that, for four motifs (TGTG, TATAA, TATAGA, and TGTCT), the motif and its reverse complement had a similar activity profile and were clustered into the same family. For these motifs, both forward and reverse strand genomic sequence data were input into MODEM. For motifs without this strand symmetry property (e.g. AAAA, TTGT, TTTT, CTTA), only the forward strand data were used. For motifs that were equivalent to their reverse complement (TGCA and AATT), the sequence from the strand with the better PWM score was used.

The flanking sequences of the predicted true targets were used to analyze co-occurrence of adjacent motifs. For each core motif, we tallied instances of other motifs in the adjacent 10 bases upstream and downstream. All co-occurring motifs (not just the representative sequences) were counted. A co-occurrence count between the core motif and each other family was assessed, based on the number of core motif true targets for

which at least one member of the co-occurring family was adjacent to the core motif. The significance of these co-occurrence counts was assessed by comparison to motif frequencies in 10000 randomly sampled 10mers within promoter sequences, using Poisson statistics.

Positional Biases

To identify positional biases of the 11 motif families, we tested whether any of the members of a motif family occurred at each position in each promoter. This yielded occurrence counts as a function of position for each motif family. Frequencies were calculated by normalizing for the number of promoters having sequence at each position. This number of promoters varies slightly by position because some promoters were truncated to be less than 1 kb (see above). Positional biases were assessed by comparing the occurrence rates of motifs near the ATG to those in the remaining sequence out to 1000 bp. For example, the near region might consist of the binned region between 10 and 50 bp, with the remaining sequence being the region from 50 to 1000 bp. Statistics were tabulated starting at 10bp to avoid issues associated with finite motif lengths. Statistical significance of the rate biases was assessed by considering the binomial probability (approximated by a Gaussian distribution) that the occurrence statistics, or anything more extreme, in the near region would be generated by chance given the occurrence rates in the far region.

Parasite culturing and nuclear protein extraction

Strain 3D7 *P. falciparum* cultures were grown in RPMI media as described (Bozdech et al. 2003). We made nuclear extracts from synchronized parasites harvested at 5 different time points of the *Plasmodium* IDC and tested them for sequence specific DNA binding activity.

Electromobility shift assays

We designed the probe sequences by using MODEM to define flanking sequences associated with the TGTGTG core motif. Based on this computational analysis, the core motif was flanked with poly-AT extensions such that the final oligo sequence was TATATATGTGTGTATATA for the forward strand and TATATACACACATATATA for the reverse strand (wt probe in Figure 5). A non-specific probe was designed with the same GC content with changes in the core motif with the forward strand TATATAACTGACTATATA and the reverse strand TATATAGTCAGTTATATA (mut probe in Figure 5). Note that the flanking sequences of this control probe have been kept the same as the original probe and only the core motif has been scrambled.

Oligonucleotides were purchased from IDT, and each oligo was labeled with T4 polynucleotide kinase (NEB) in 10 ml reactions for 1 hour according to manufacturer's instructions. Complementary strands were mixed, then heated to 100°C and slowly cooled to room temperature. Double stranded DNA (dsDNA) was purified from unannealed DNA on a native 15% polyacrylamide/1X TBE gel, and labeled dsDNA probe was eluted to TE pH 8.0 and kept at 4°C. Binding reactions were done using 5µg nuclear extract

from each time point, as determined by Bradford assay (Biorad). Binding buffer was adapted from Voss et al. 2002 and consisted of 20mM Hepes pH 7.6, 10 mM KCl, 0.67mM MgCl₂, 1 mM EDTA, 1 mM DTT, 0.67 mM PMSF, 5% glycerol, 0.0067% NP-40, 0.25 µg/µl BSA, 0.1 µg/µl salmon sperm DNA, and 0.2 µg/µl poly dI-dC. Binding reactions were done by adding nuclear extracts to pre-mixed components including labeled probe, and allowed to bind for 15 minutes at room temperature, before loading to a 5% polyacrylamide/0.25X TBE gel. Typical binding reactions were done with 10,000CPM of probe per 15 µl binding reaction, which represents approximately 1fmol of dsDNA probe.

Results

Identification of Motifs and Activity Profiles

At least 2714 of the 5409 ORFs in the *P. falciparum* genome are transcribed during the IDC and exhibit a simple mRNA expression profile, in which the expression level peaks once during the 48 hour cycle. (Bozdech et al. 2003). This relatively simple expression pattern suggests that the regulatory program of the IDC may also be simple. We set out to understand this program from the viewpoint of *cis*-regulatory elements, by integrating the genomic sequence and IDC expression data to find the candidate transcription factor binding site sequences associated with *P. falciparum* gene expression during intraerythrocytic development.

To identify regulatory sequences, we analyzed a set of promoters from the *P. falciparum* genome sequence. A schematic for the information flow is available in Figure 1. We focused on the 1kb of non-coding sequence upstream of each annotated start codon, though sequences were truncated if they overlapped with any coding sequence. We then searched these promoters for sequence motifs likely to regulate the mRNA expression levels of the downstream genes. This search was performed using the REDUCE algorithm, a method which has been successful in integrating sequence and microarray expression data to predict regulatory motifs in *Saccharomyces cerevisiae* (Bussemaker et al. 2001). REDUCE identifies sequence motifs whose pattern of occurrence in promoters correlates significantly with the expression level of the adjacent gene (see Methods). Expression data were obtained from Bozdech et al (Bozdech et al. 2003).

Motifs were identified for the 46 hourly time points in the IDC for which microarray expression data were available. Over the complete set of time points, this method uncovered 128 significant motifs. No significant motifs were recovered when 3' intergenic sequences were analyzed for correlations with expression, suggesting that 5' sequences determine most of the changes in steady state mRNA levels in the IDC. Using the expression information at different time points, we were able to identify the changes in motif activity over time (see Materials & Methods).

Next, these motifs were clustered on the basis of their activity profiles (their activity scores S over all timepoints), analogous to the common technique of clustering genes by their microarray expression profiles. The resultant pattern of motif activity (Figure 2B)

was reminiscent of the gene expression pattern (Figure 2A, obtained with permission from (Bozdech et al. 2003)). That is, activity profiles were sinusoidal with a period equal to that of the IDC. All stages of the IDC were predicted to have at least one motif associated with positive activity.

Motifs with similar activity profiles were often found to be similar in sequence, suggesting that some predicted motifs are biologically tolerated variants of the same transcription factor binding site. To confirm this observation, we clustered the motifs by their sequence (see Methods) into major motif families. This motif clustering was performed on the 100 motifs with reasonable copy number (> 50 genes) and length (> 3 bases long). Four motifs were not classified into these families, as they appeared to be outliers. The resulting sequence similarity tree (Figure 3A) yielded a set of 11 major motif families, comprising 96 individual motifs. As expected, we found that motifs within sequence clusters tended to have similar associated transcriptional activity profiles (Figure 3B-3D). For each family, we identified a representative motif, chosen on the basis of having the strongest REDUCE p-value, when all timepoints for all motifs within the family were considered. For each of these representative motifs, the maximal REDUCE p-value was $p = 10^{-4}$ or better.

The eleven representative motifs from these families are TGTG, TATAA, TATAGA, TGTCT, TGCA, TTGT, AATT, TTTT, TGTAG, CTTA, and GATAC, in the order of their statistical significance (Table 1). These predicted motifs were robust to the sequence set used : we repeated predictions using 3kb or 5kb of upstream sequence and in each

case recapitulated a similar set of motifs. The activity profiles of the motifs in three of these families are shown in Figure 3B-D. The complete list of representative motifs, as well as the activity profile for each motif, is available in Table 1. Note that, for seven of these families, motifs have been clustered together with motifs that appear to be equivalent binding sites on the reverse strand. For these seven cases, the forward and reverse complement versions were found to have similar activity profiles (see next section).

Reverse complementarity

It is known that many transcription factors can promote transcription when bound on either the forward or reverse strand of a promoter sequence. Consistent with this mechanism, seven of our core motifs were found to have similar activity profiles on the forward and reverse strands. Two examples are shown in Figure 3: TATAGA/TCTATA (Figure 3B), TGTG/CACA (Figure 3C). In both of these cases, the main motif clearly has a similar activity profile to its reverse complement, supporting the veracity of the motifs. The complete set of strand-symmetric motifs (with an example reverse complement motif having a similar activity profile) are: TGTG/ACACA, TATAA/TTATA, TATAGA/TCTATA, TGTCT/AGACA, TGTAG /TCTACA, GATAC/TATC, and CTTA/TAAG.

The remaining four core motifs did not show strand symmetry. For example, the variants of the TTTT motif as well as the TTGT motif had much stronger REDUCE p-values

when the T-tract occurred on the forward strand. This effect is shown for TTGT in Figure 3D. The activity values are strong in the forward direction (reddish curves), but the motif activity values are significantly less for the reverse complement versions, such as ACAAC (greenish curves). Two motifs (AATT and TGCA) were equivalent to their reverse complement, making the strand symmetry issue undefined in these two motifs.

Positional Biases

Several of our predicted motifs also showed positional biases within promoters, as would be expected if they play a functional role. For example, the motif AATT has 10922 copies in the region between 10 and 50 bp, which is ~ 600 more than what would be expected, based on the statistics of the region between 50 and 1000 bp ($p=1.0\times 10^{-8}$). Similarly, the motifs TTTT ($p<1.0\times 10^{-300}$), TGTG ($p=1.3\times 10^{-5}$), TTGT ($p=1.3\times 10^{-41}$), and TGTAG ($p=2.9\times 10^{-6}$) all have an excess of copies in the region between 10 and 50 bp. Conversely, the motif TATAA is strongly underrepresented between 10 and 50 bp, having 2200 fewer copies than would be expected based on the more distant regions ($p=2.7\times 10^{-62}$). The most striking example of a positional bias is for AATT, which is strongly overrepresented in the region between 10 and 50 bp, but is underrepresented in the region between 50 and 150 bp (Supplemental Figure 1).

Motif Targets

High Frequency

One unusual aspect of these predicted motifs is their high frequency throughout the genome. Table 1 shows the number of occurrences of exact matches to the representative members of each motif family, and the number of genes in which they appear. The motifs TATAA, TTGT, AATT, and TTTT each appear in nearly all of the 3518 ORFs for which we analyzed the promoter sequences. TTTT is the most abundant of these, occurring 239556 times, in 3511 promoters, an average of ~70 times per promoter. This contrasts with yeast, for which the cell cycle regulatory motifs typically occur only 1-2 times per promoter (Bussemaker et al. 2001). Other malaria motifs also appear more often per promoter than the yeast motifs, such as TGTG (~3 copies), TATAA (~14 copies), TTGT (~6 copies), AATT (~14 copies), and CTTA (~3 copies). The high occurrence frequencies of motifs suggest that copy number could have a stronger role in transcription in malaria than it does in yeast. Still, a few IDC motifs (TATAGA, TGTCT, TGTAG, and GATC) have only 1-2 copies per gene, indicating that this copy number phenomenon may only be relevant for a few transcription factors.

Target Refinement Using MODEM

It is well known that a core motif alone does not contain enough information to sensitively and specifically identify the target genes of a transcription factor. Given the high copy number of these motifs, it seemed particularly important to refine to the set of most likely target genes. We therefore separated motif instances into true targets and false positives using a modification of the MODEM algorithm (Wang et al. 2005), which

applies a likelihood-maximization approach to separate motifs on the basis of flanking sequence and gene expression data. Normally the MODEM algorithm uses expression data from one experiment, but we altered it to incorporate the expression data for the complete IDC timecourse (see Materials & Methods). An auxiliary output of the MODEM algorithm is a position specific weight matrix for sequences flanking motif occurrences within true target genes, as well as another matrix for sequences flanking false positives.

The MODEM analysis significantly reduced the number of likely target genes. The large discrepancy between observed motif instances and true targets suggests other features, such as combinatorial control or chromatin structure, could be particularly important for determining which motif instances are active. A complete list of MODEM predicted targets is available upon request.

Motif Co-occurrence

Since MODEM also yields flanking sequence information for motif instances, we were able to test whether motifs act combinatorially. The small number of transcription factor binding sites and small number of known transcription factors would suggest that combinatorial control is necessary to create specific transcription patterns.

Consistent with the hypothesis of combinatorial control, we observed that flanking sequences of our representative motifs are often enriched for instances of other motifs.

An example is shown for the sequence flanking the TGTG representative motif (Figure 4B, 4C). The TGTG motif has a strong bias to have extra instances of itself in close proximity on both the left ($p=10^{-19}$) and the right ($p=10^{-9}$) side. The TGCA motif (Figure 4C) is also associated with the TGTG core motif, but only on one side ($p=10^{-10}$). Conversely, an analysis of the TGCA motif (Figure 4D, 4E) shows it to be associated with the TGTG motif on the left side.

Other examples of motif associations include the association of the TTGT motif with copies of the TTTT motif on both sides, as well as itself. The AATT motif is associated with poly-A tracts on the left side and poly-T tracts to the right side. The TGTCT motif has high co-occurrence with poly-T motifs only.

Experimental verification

Although several experiments in the literature confirm the existence of our predicted motifs within fragments of DNA shown to bind nuclear factors, the presence of multiple motifs within these previously defined sequences precludes precise interpretation of such experiments. We chose to validate the TG repeat motif for experimental validation because of its breadth of target genes and its strong correlation with expression. By using a synthetically designed binding site free of other motifs, we were able to observe sequence-specific binding of a nuclear factor throughout all IDC stages tested (Figure 5), confirming that the TGTGTGT sequence is sufficient for sequence specific binding of a nuclear factor.

Discussion:

The prevailing patterns of the *P. falciparum* IDC transcriptional program have been described, however the *cis*- and *trans*-acting regulatory determinants of the transcriptional program are poorly understood. Many of the components involved in transcription in other eukaryotes are missing in *Plasmodium* (Coulson et al. 2004; Callebaut et al. 2005), suggesting that there may be global characteristics of transcription specific to *Plasmodium*. While the global transcriptional program encompasses the majority of the genome, the incorrect timing of expression of specific genes could have important phenotypic consequences, as exemplified by several studies showing that proper temporal encoding at the promoters may be necessary for proper localization of secretory cargos (Kocken et al. 1998; Triglia et al. 2000; Rug et al. 2004). Understanding the program of transcriptional regulation of the *Plasmodium* IDC may afford us an insight into how the organism has adapted to its particular parasitic lifecycle.

Our approach to this problem is through the identification of potential *cis*-regulatory motifs on a genome-wide scale. Based on correlations of motif abundance in promoter regions with expression, we identified 11 motif families. This low abundance of regulatory elements is consistent with previous suggestions that the IDC is regulated by a small number of transcriptional regulators (Bozdech et al. 2003; Coulson et al. 2004). We have also shown that 3' UTR sequences do not correlate with expression, implying that the bulk of regulation of the relative mRNA levels in *P. falciparum* occurs through the

families of 5' motifs we have characterized. We note that 5' intergenic sequences tend to be longer than 3' intergenic sequences, with 5' sequences having a median length of ~1.2kb and 3' UTR sequences having a median length of ~500bp (Supplemental Figure 2), consistent with a greater proportion of information being encoded in promoters rather than 3'UTRs. On the other hand, Coulson et al. reported the expansion of protein families encoding RNA binding proteins in *P. falciparum* (Coulson et al. 2004). A recent genome-wide study of mRNA half-lives in *P. falciparum* revealed global differences in the distribution of mRNA half-lives in different stages of the IDC suggesting that mRNA stability may also contribute significantly to the regulation of steady state mRNA levels (Shock 2007). The relative contribution of production versus destruction of mRNA in gene expression in *P. falciparum* remains an open question.

The activity profiles of the motifs are sinusoidal and consistent with a simple program of gene regulation. Motifs clustered by their activity profiles were found to have similar sequences (Figure 2) and conversely, motifs clustered by sequence had similar activity profiles (Figure 3). Eleven major motif families emerged from these data, all but one with best p-value scores better than 10^{-10} suggesting that these motifs are functionally relevant (Supplemental Figure 3). The 11th motif family GATAC is predicted as several independent oligonucleotide words, however the significance score for this motif family was weakest of all of the motif families (p-value $< 10^{-4}$). Six of these motif predictions were bolstered by detection of orientation independence of correlation with transcription, and several motif families show a positional bias. Also, many of the motifs can be found

within probe sequences used for EMSA by previous researchers, though many of these experiments have not analyzed binding as a function of developmental stage.

Each motif profile is a score correlating motif copy number with relative abundance of the transcript. Our analysis does not allow us to infer directly whether a motif is acting as a positive or negative regulatory motif. However, if we assume that the motif is acting positively, we can infer from the motif profiles that the period of the IDC with upward slope represents the window of time in which activators are inducing target genes.

Conversely, if we assume that the motif is a repressor-binding site, the duration of time when the slope is negative should represent the period during which the motif is silencing transcription. If all motifs are assumed to be activators, we find that the points of upward inflection span the period of time when the parasite is known to be the most transcriptionally active (Martin et al. 2005) suggesting that these motifs may represent most of the key regulatory motifs governing the IDC. Interestingly, this interpretation suggests that the TATAA motif is the last motif to activate transcription of the IDC, at a stage when the activity of most other motifs begins to decline, suggesting that the late stage transcripts may be more dependent on TBP (see below).

Descriptions of motifs

Polymer-like motifs

Many of the binding sites we predicted are consistent with some known *cis*-regulatory sites. One of our predictions is the poly-T motif. Several previous reports have shown

that the number and length of homopolymeric stretches of DNA are associated with changes in transcription of downstream genes (Porter 2002; Polson et al. 2005), and we add to this concept the idea that homopolymeric stretches encode stage-specific temporal information as well, in an orientation-dependent manner. The motif AATT also is predicted in our analysis to be associated with expression when present at the boundary between poly-A and poly-T tracts. The mechanism of transcriptional regulation by these homopolymeric DNA could be through sequence specific DNA binding proteins, or it may be mediated by structural differences of the DNA which provide for differential nucleosome binding (Zaret et al. 1984). Recently, it was shown that in *S. cerevisiae*, a 7bp poly-T stretch following a binding site for the transcription factor Reb1 defines a combinatorial *cis*-regulatory element that specifies placement of variant histone H2A.Z at promoter regions in the genome (Raisner et al. 2005). Our detection of the poly-T motif suggests this mechanism may be at play in *P. falciparum* as well, and the transcript abundance of the H2A variant histone of *P. falciparum* (PFC0920w) exhibits trophozoite stage enrichment in the IDC transcriptome dataset that would be consistent with a role of the H2A variant in regulation via this motif if one assumes a 4-6 hours of lag for protein accumulation following changes in transcription (Le Roch et al. 2004).

The G-doped poly-T motif is one of the motifs detected in our analysis, and its correlation with expression seems to be strongly strand-specific (Figure 3D). We propose two possible mechanisms of activity for this motif. In the first model, this sequence represents a functional homolog of the G-doped poly-T tracts reported recently in yeast studies that found these sequences to be associated with nucleosome-free regions

of the genome (Rando et al. 2006). In the second model, the strand-specificity of this motif is a consequence of its function at the level of RNA. A TATTTTTTGTTT sequence found at the 5' termini of two transcripts from the *Plasmodium yoelii* mitochondrial genome has been suggested to be a mitochondrial promoter sequence (Suplick et al. 1990), and it is a possibility that our detection of this motif stems from the fact that the IDC microarray data derives from a total RNA hybridization series which presumably includes RNA from the mitochondria.

TGTCT

The core motif TGTCT is detected with expression correlation in both orientations suggesting that it is a binding site for a transcription factor. Flanking nucleotide analysis of this motif using MODEM yields base biases indicating that these sequences are often flanked on both sides by poly-T tracts (Supplemental Figure 4). The TGTCT sequence is present in a promoter fragment from *P. gallinaceum* that binds a sequence specific DNA binding activity, suggesting that the binding site may be conserved across distantly related plasmodia.

TATAGA, TGTAG

The motif TATAGA is predicted with high confidence, though our current MODEM analysis indicates that the number of target genes with expression profiles similar to the motif profile is a small set (33 genes). This prediction set is a conservative estimate: this

motif may represent an activator of late stage expression based on the motif profile since the time point of upward inflection has a more sharp transition than at the downwardly inflected time point. If the motif is primarily active only as an activator, the use of pseudo-expression projections of the entire time course of motif and expression profiles may reduce the signal to noise ratio for the MODEM analysis, allowing only the most resilient correlations to be observable. A MODEM analysis using only expression data from timepoint 35 yielded far more target genes than currently reported here. Thus, these 33 genes represent the most conservative predictions set. Variants of this motif family are noted to occur within a block of DNA conserved in promoters of *var* genes located in subtelomeric as well as chromosomally internal regions as well. Since these classes of promoters have been described as being significantly divergent at the nucleotide level, this block of conservation has been suggested to play a regulatory role (Voss et al. 2000). Furthermore, a third class of *var* gene promoters was found among parasite strains derived from cases of placental malaria, and while these promoters are divergent at the nucleotide level from the abovementioned two classes of *var* gene promoters, instances of this motif are abundant in this promoter class as well (Vazquez-Macias et al. 2002). Our top scoring predicted target genes do not include the *var* genes, which is not surprising given that the sequences used for the microarray design derive from the reference 3D7 strain, and the *var* loci are significantly polymorphic from the HB3 strain used for the expression study (Bozdech et al. 2003). However, many other genes known to be involved in merozoite invasion are predicted to be targets of this motif. Glycophorin binding protein (Gbp-130, plasmODB ID: PF10_0159) is one example of a gene with an essential role in invasion, and it is significant to note that experimental

dissection of this promoter has shown that a 5bp change in the promoter can abolish expression as well as the binding of the nuclear factor (Horrocks et al. 1999). This 5bp mutation changes the first nucleotide of an instance of this motif and it would be interesting to verify the stage specificity of this transcriptional activity. This motif may have been overlooked in the prediction set of van Noort et al. because it may regulate *P. falciparum* specific antigenic genes of recent origin not conserved in *P. yoelii* (van Noort et al. 2006).

The TGTAG motif family has a motif profile similar in timing to the TATAGA motif profile. The KAHRP gene, involved in the production of knob structures on the infected erythrocyte surface, has a promoter that has been analyzed and a region of the promoter containing the TGTAG motif has been shown to bind nuclear factors with stage and sequence specificity (Lanzer et al. 1992). However, this oligo contains instances of other motifs we have predicted, precluding a clear interpretation of which particular sequence may be dictating binding (see below).

TATA

One of the most significant and most abundant motifs resembles a canonical TATA box and is present at 49002 copies in 3512 genes for an average of 14 copies per gene, and the temporal pattern of the motif profiles of this motif family correlate with the expression profile of the *P. falciparum* TATA-binding protein (PfTBP, plasmodb ID: PFE0305w), as well as a second TBP homolog PF14_0267. Although TBP is typically considered to be a general transcription factor, our analysis suggests that TBP may act as

a developmental regulator. Recently, PftBP was shown to bind to sequences proximal to transcription start sites of two promoters (Ruvalcaba-Salazar et al. 2005), and our EMSA of one of these sequences indicates that the sequence specific binding activity is most abundant in schizont stage parasites (data not shown). The high prevalence of the TATAA motif variants suggests that there are additional determinants of promoter specificity, and some of that specificity may be achieved through TATAA copy number. We also find that the density of the TATAA motif is correlated with expression, as MODEM predicted true targets of TATAA tend to have another copy of the motif within 10bp of each other ($p < 10^{-20}$). If our interpretation that the trough of the motif profiles correlate with the time in the IDC when the motifs begin to contribute to the relative transcript abundance of target genes, PftBP or its homolog PF14_0267 likely regulates a subset of the genome specifically based on TATAA frequency and density.

TGTG, TGCA

The TGTG motif has the strongest correlation of frequency with expression in the IDC. We obtained further information about this motif by examining the correlation of extended versions of this motif using MODEM (see Materials & Methods). The extended motif predicted by MODEM is suggested to be involved in the periodic transcription of approximately 20% of the 3518 genes with periodic transcription during the IDC (Table 1). By averaging the flanking sequence biases of the best scoring instances of the motifs in the predicted target genes, we obtained an extended motif which is the average of these high scoring predicted target sites. This extended motif has a repeated polymer of AT repeats flanking the core motif (Figure 4A). We

experimentally validated this *in silico* predicted synthetic motif and showed that a nuclear factor with sequence specificity for this double stranded DNA sequence exists throughout most of the IDC, consistent with our analysis suggesting that it represents an activator sequence with activity from 12 hours post invasion (hpi) to 35hpi. The idealized version of the motif was computed through the EM procedure of MODEM, and this analysis predicts that repeats of TG on both sides of the core motif have stronger correlation with expression. Thus the nuclear factor we showed to bind to our probe may have preference for binding to multimeric repeats of the TGTG sequence (Figure 4A). The detection of TG repeats in the TGTG PWM is also corroborated by the motif co-occurrence analysis, which shows that instances of TGTG associated with expression have an over-representation of secondary copies of TGTG in both left and right flanking regions (Figure 4B). Calderwood et al. showed that an imperfect TG polymer within the intronic promoter of the *var* gene has a trophozoite stage specific enhancer activity (Calderwood et al. 2003). By our conservative estimate, this motif is responsible for the regulation of at least 736 genes in the genome, suggesting that the *var* gene silencing program is an integrated part of the IDC regulatory program. One study The PfCAM promoter used in many *P. falciparum* transfection studies also has 3 copies of this motif, and the PfDHFR promoter has two copies (Crabb et al. 1996), although none of these instances are polymeric. This observation is consistent with our experimental results showing that one monomer is sufficient for sequence specific binding of a nuclear factor. It will be interesting to determine experimentally, whether longer polymers of TG repeat units will allow for cooperative binding of the factor, resulting in stronger promoters.

The flanking motif analysis of the TGTG motif also indicates that many functional instances are present with proximal copies of other motifs. The TGTG motif is often flanked by another copy of the TGTG motif within 10 bp (Figure 4B, 4C). The TGCA motifs is also associated with the TGTG motif, but with much greater preference on one side (Figure 4C). The KAHRP gene promoter fragment mentioned above may be an example of a complex binding site, with instances of TGTG, TGCA, and TGTAG encoded within a 14-bp window. Although several stage specific bands were observed using this oligo in an EMSA, the presence of multiple predicted binding sites within this oligo make the precise explanation of this result in the context of our motif predictions difficult. It should be noted that a band specific to the ring stage and a separate band specific to schizont stages were observed with this native promoter fragment (Lanzer et al. 1992), and it would be interesting to revisit this experiment by altering nucleotides to test the contribution of each of the three motifs to nuclear factor binding independently.

The TGCA motif is predicted independently of the TGTG motif by REDUCE, and flanking sequence analysis using the TGCA motif as the core motif yields the reciprocal enrichment of the TGTG motif on the left side of the TGCA motif (Figure 4D). Instances of TGCA are also over-represented on the left side of the TGCA motifs, and consistent with this observation, a tandem repeat of TGCA has been isolated from a *var* gene promoter and shown to bind a nuclear factor with temporal and sequence specificity (Voss et al. 2003).

The similarity of the motif profiles of the TGTG and the TGCA motifs, coupled with their high co-occurrence may call into question whether these motifs are in fact independent motifs. However, we prefer the interpretation that the upward inflection point of the motif profiles indicates the time in the IDC when the motifs become active, and in this model, the TGTG motif controls activation approximately 9 hours earlier than the TGCA motif (Supplemental Figure 5). There are two possible scenarios for how these two sequences interact. In the first scenario, these motifs represent binding sites for two different transcription factor complexes, and the overlapped instances represent examples of combinatorial control. An example of this mode of regulation is the overlapped specificity and competition between Sum1p and Ndt80 in *S.cerevisiae* in middle sporulation genes (Wang et al. 2005). In this model, a hybrid site such as that observed in the *var* gene promoter (TGTGCA) is bound a TGTG binding factor initially, but later by a TGCA binding factor possessing higher affinity for the hybrid site. The second explanation for these observations is that the TGTG motif and TGCA motif are recognized by one transcription factor with different affinities. In this scenario, temporal information is encoded in binding site sequence variations, coupled with increasing concentrations of the binding factor throughout the IDC. Consistent with this hypothesis, the TGBP is expressed more abundantly later in the IDC (data not shown, Materials & Methods), and the timepoint at which the TGTG motif and TGCA motifs no longer correlate with transcript increase are approximately the same. A testable hypothesis from this model is that the TGTG motif and the TGCA motif will compete for binding, but that the TGTG motif will have a higher affinity than the TGCA motif. The lambda phage, one of the earliest genetic circuit to be studied in molecular detail, utilizes such an

affinity-based encoding strategy (Ptashne et al. 2002); *Drosophila* embryo segmentation also achieves encoding of 8 stripes by two transcription factor gradients (Clyde et al. 2003). This mode of encoding regulatory information is efficient in that fewer transcription factors are required to encode multiple output states. Given the concordance of transcription factor scarcity and low diversity of expression profiles, this model certainly is attractive. We are currently testing these predictions.

Hypotheses on regulatory design strategy

We also note that many of the motifs we detect are part of simple sequence repeats (AATT, poly-T, G doped poly-T, and TG repeat). It is interesting to speculate that the use of repetitive sequences as regulatory elements affords *P. falciparum* with a high degree of evolvability with respect to attaining desirable gene expression patterns, through DNA replication slippage or recombination at these repeat tracts. Simple repeat sequence polymorphisms have been associated with variation in gene expression patterns in antigenic phase variation in pathogenic bacteria (reviewed in (Bergman et al. 2006)) and with oncogenesis in humans (reviewed in (Brandt et al. 2006)). Larger repeat units ranging from 60bp to 300bp have been described in some *P. falciparum* promoters of genes with clinical consequence including those involved in antigenicity (Vazquez-Macias et al. 2002), RBC invasion (Horrocks et al. 1999), and gametocytogenesis (Alano et al. 1996). Naturally occurring polymorphisms in copy number of these blocks have been observed in clinical isolates for two types of these repeats (Vazquez-Macias et al. 2002) (Sallicandro et al. 2000). PfMDR1, a gene associated with drug resistance, has

also been reported to have a polymorphic promoter and specifically, one common deletion spans a region which appears to be part of a G-doped poly-T tract (Myrick et al. 2005). Our analysis suggests that further study of *Plasmodium* promoter polymorphisms may shed light on the basis of variation in clinical outcome in malaria infections.

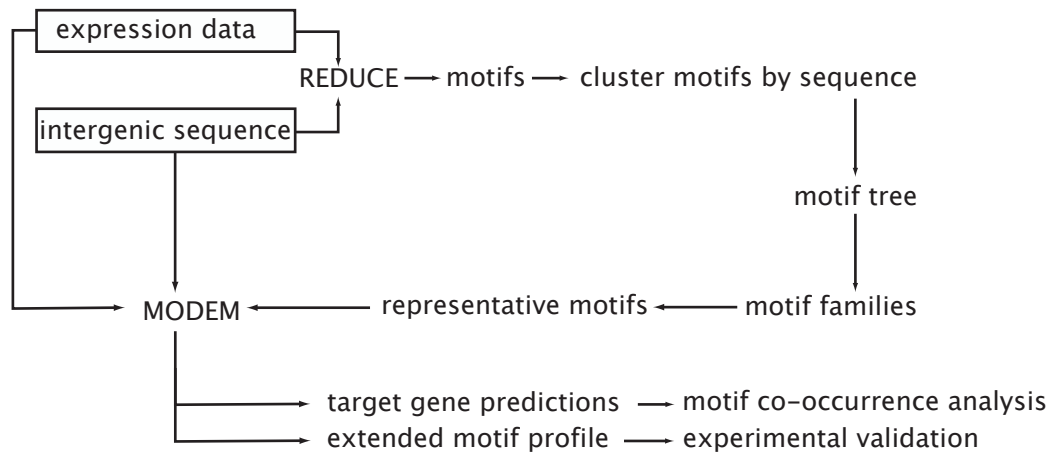


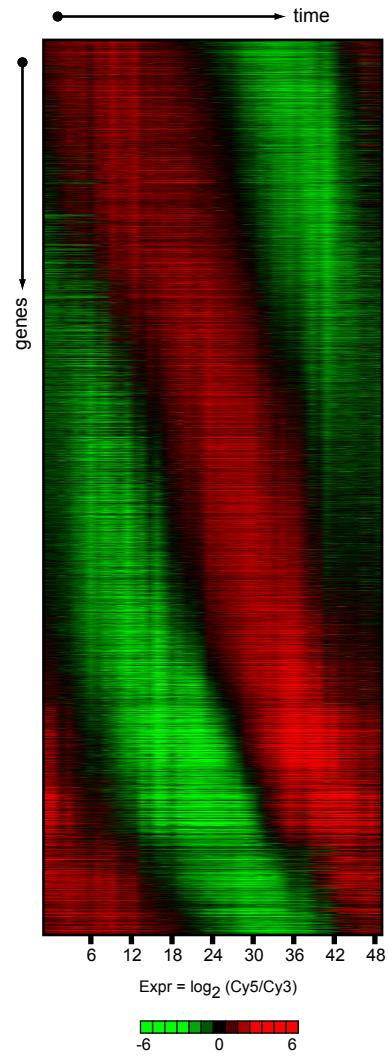
Figure 1. Information flow chart.

The expression dataset used for analysis is from Bozdech, et al. 2003.

REDUCE (Bussemaker, et al. 2001) is an algorithm for regulatory element detection which incorporates expression data using a linear regression.

MODEM (Wang, et al. 2005) is a companion algorithm for taking core motifs along with expression and sequence data to perform an integrated EM procedure to obtain further information content in flanking sequences of high scoring motifs and target gene predictions.

Figure 2A.



2B.

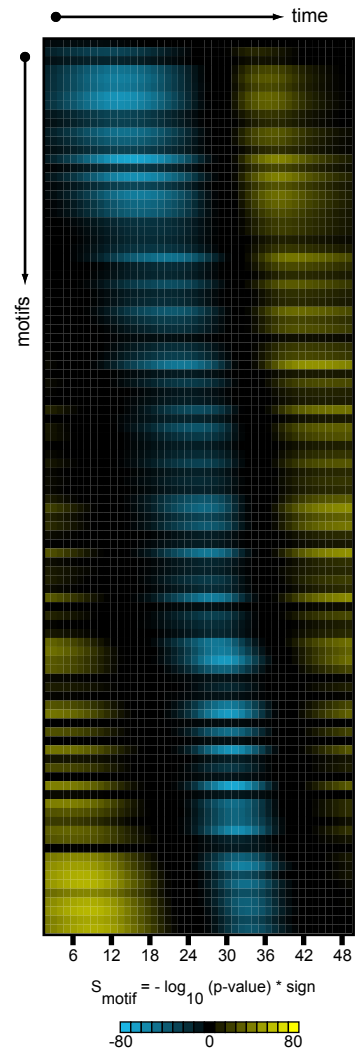


Figure 2 Motifs detected with the REDUCE algorithm. (A) expression profiles of ~3500 Plasmodium genes during the 48 hour IDC, from Bozdech et al. PLoS 2003. (B) motif score profiles for 100 detected motifs during the IDC.

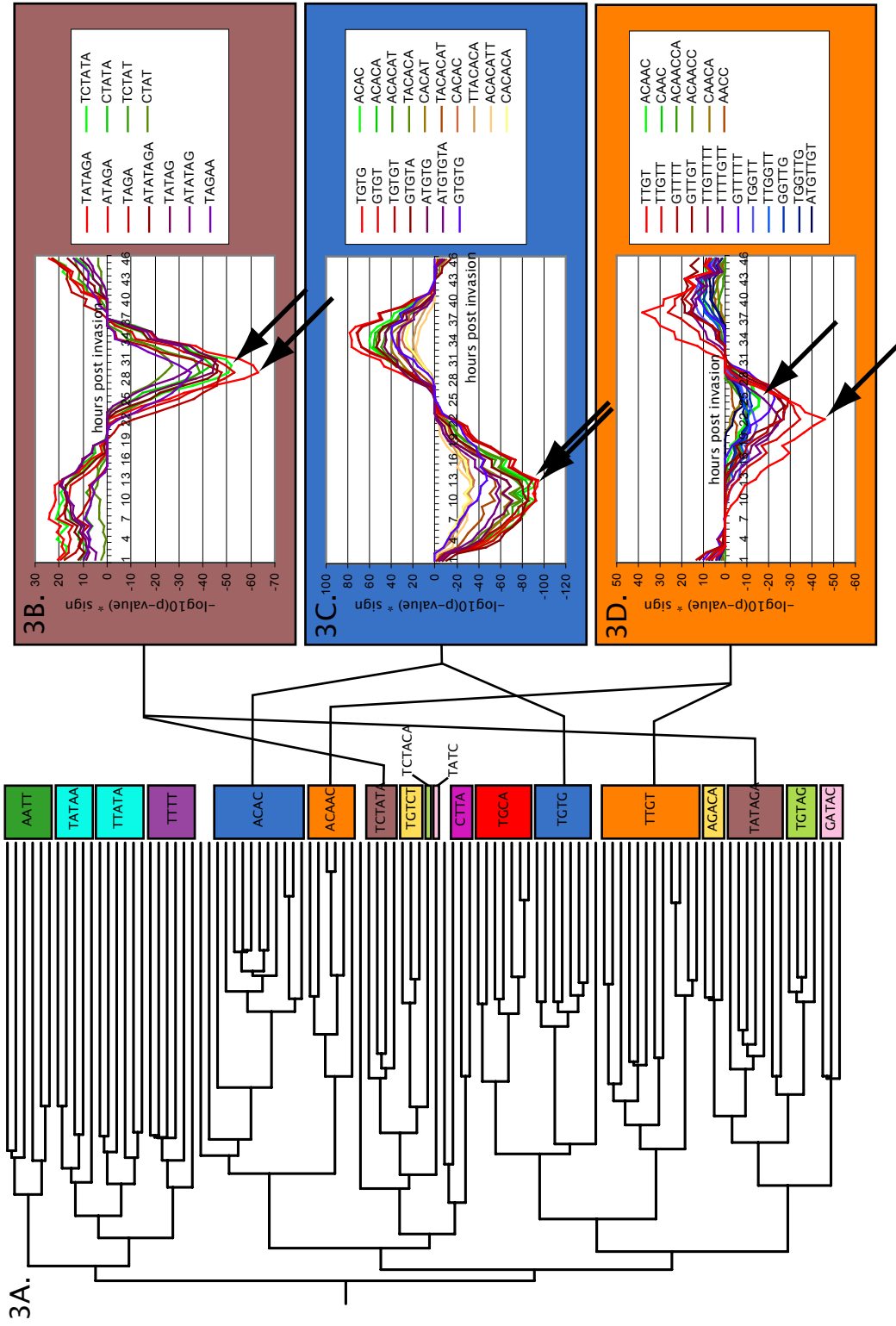


Figure 3. Motif classification tree. (A) Pairwise sequence comparisons were used to categorize motifs into families. Reverse complements were not compared directly. The TATAGA (B) and TCTG (C) motif families are associated with gene expression in both orientations, while the TTGT motif (D) exhibits partial orientation preference. Timing of anticipated activation indicated with arrows for best scoring motif for each strand.

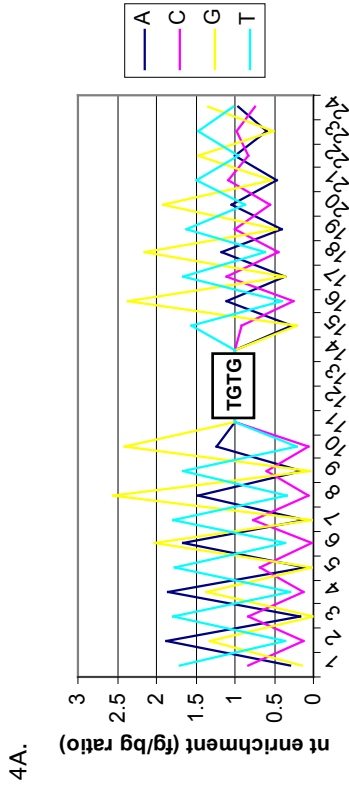


Figure 4
 A) Flanking sequences of predicted targets of TGTG motif have nucleotide bias indicative of motif repetition. Core TGTG sequence shown in box.
 B) Flanking sequences enriched for predicted targets of TGTG have statistically significant co-occurrence of a secondary motif. Similar analysis on flanking sequence to the C) right of TGTG, D) left of TGCA, and E) right of TGCA core motifs.

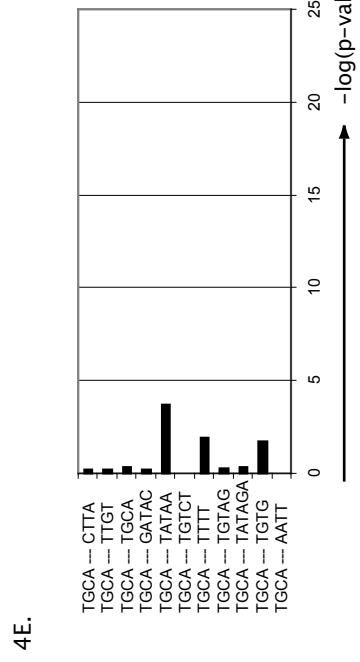
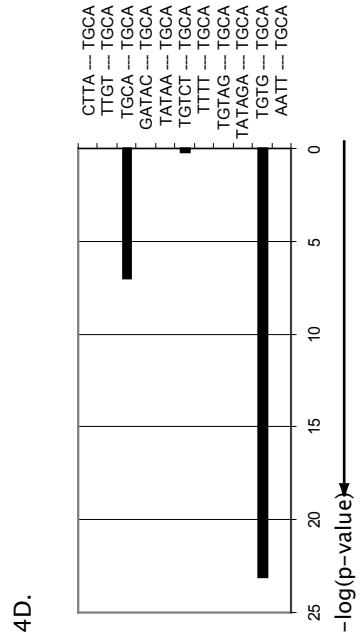
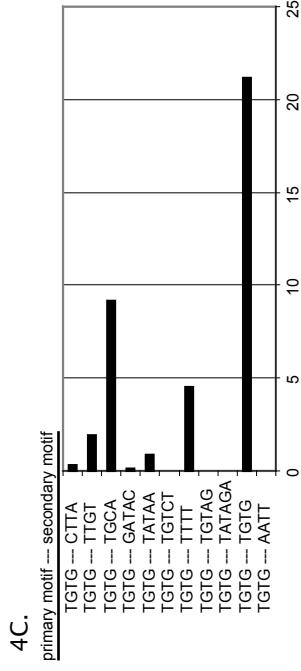
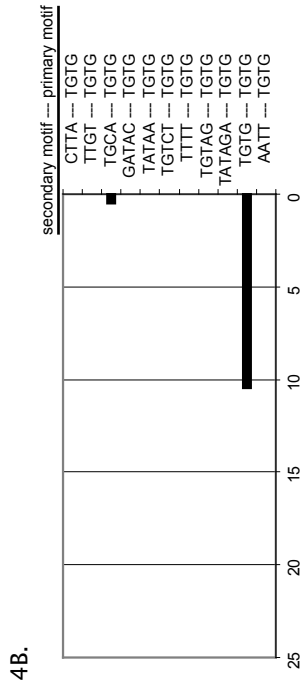


Figure 5

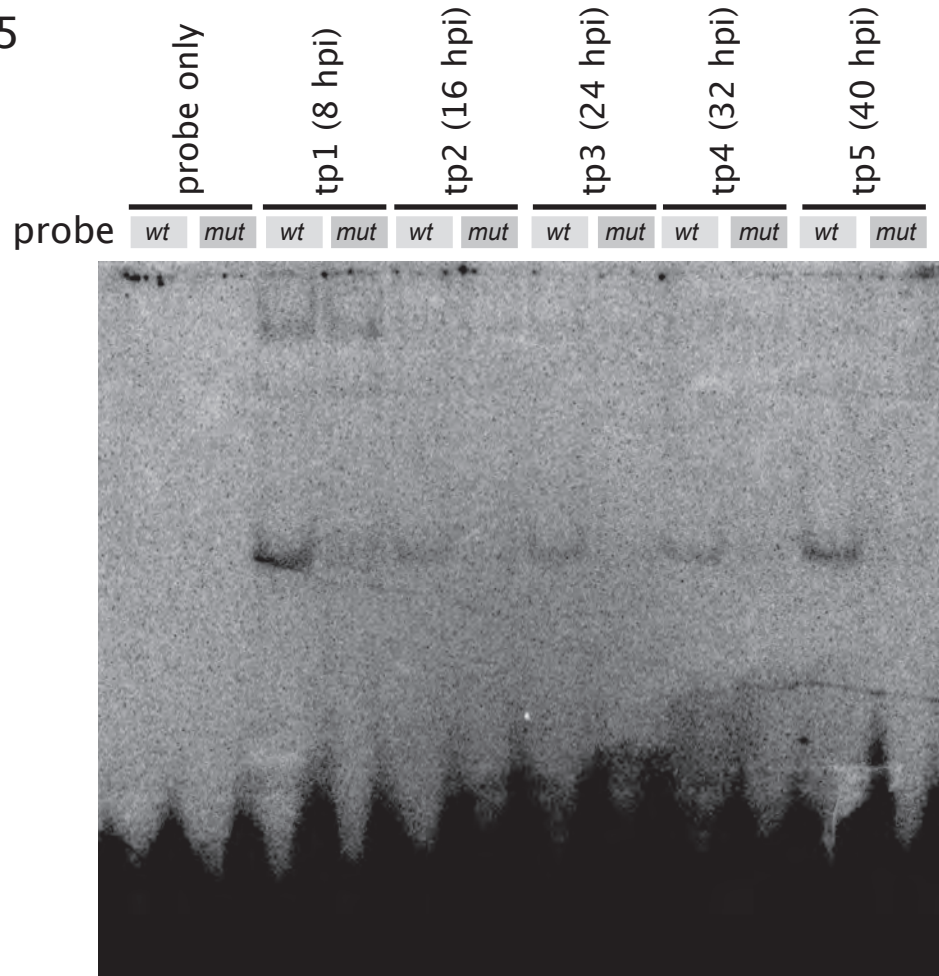


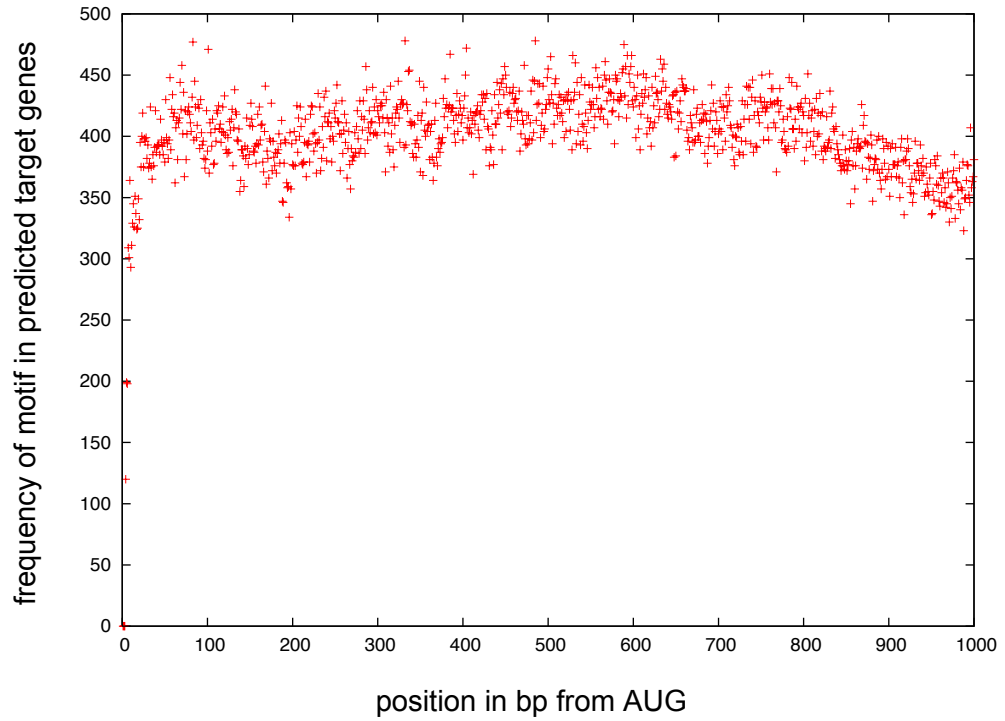
Figure 5

P. falciparum nuclear extracts were prepared from synchronized parasites obtained with 8 hour time resolution from one IDC. Time points are represented here as hours post invasion (hpi). Extracts were tested for DNA binding activity against our predicted motif, as well as a mutant sequence in which only the core motif was altered.

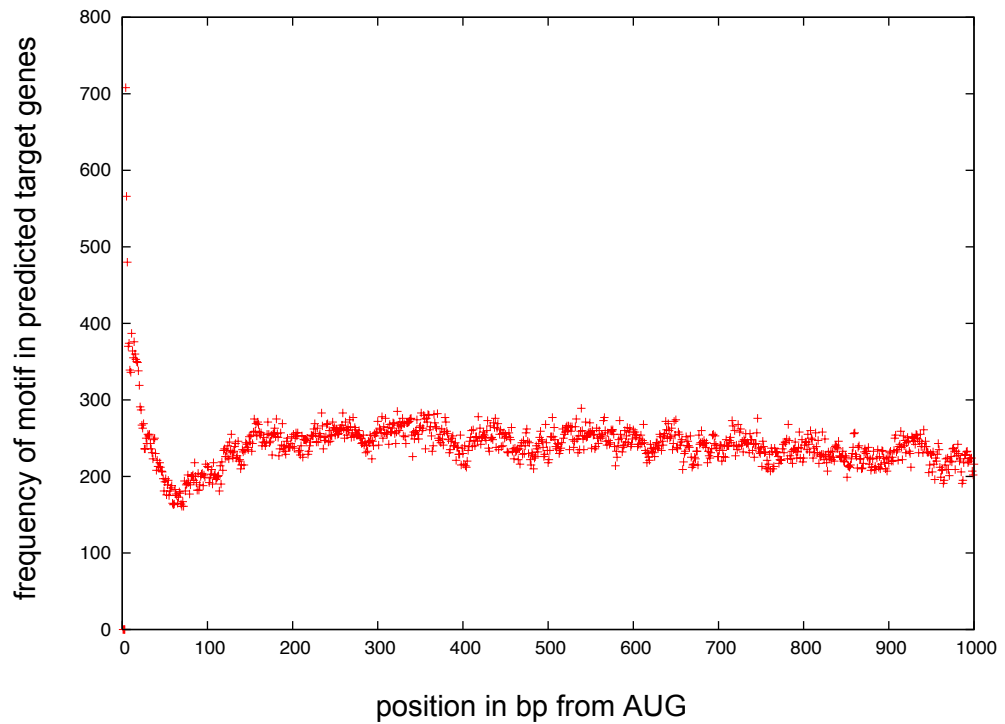
Table 1

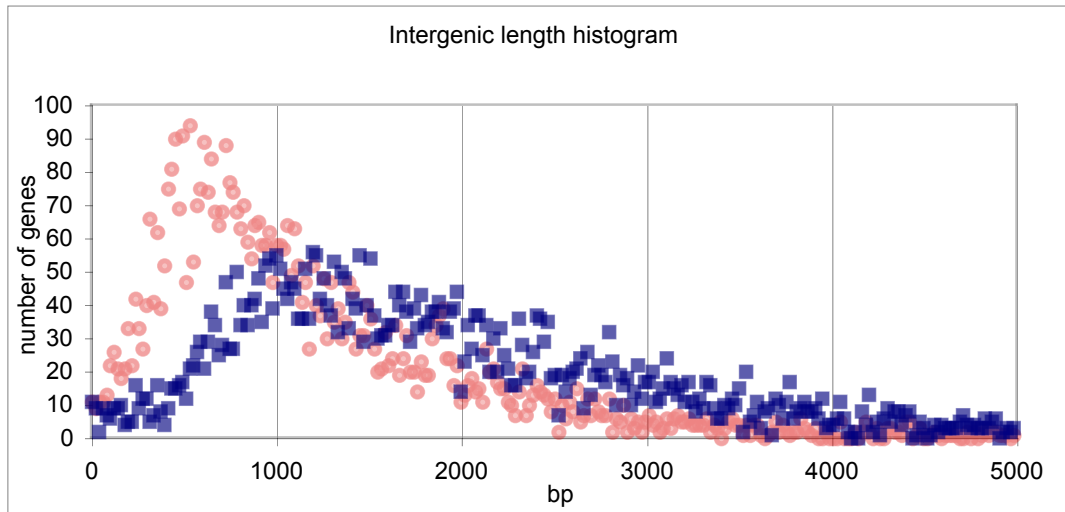
Motif	number of occurrences	number of genes with motif	analysis mode	number of MODEM targets
TGTG	8074	2883	both strands	736
TATAA	49002	3512	both strands	1736
TATAGA	2093	1461	both strands	33
TGTCT	806	643	both strands	124
			self-	
TGCA	4330	2362	complement	275
TTGT	21721	3468	one strand	533
			self-	
AATT	50407	3511	complement	1132
TTTT	239556	3511	one strand	950
TGTAG	1542	1231	both strands	133
CTTA	10199	3239	one strand	20
GATAC	929	787	both strands	99

Supplemental Figure 1A. TATAA position histogram



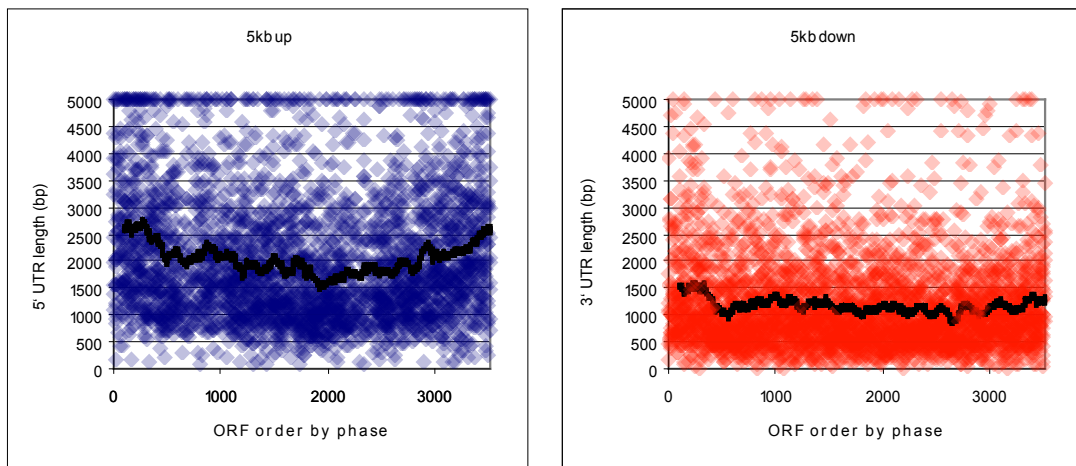
Supplemental Figure 1B. AATT position histogram





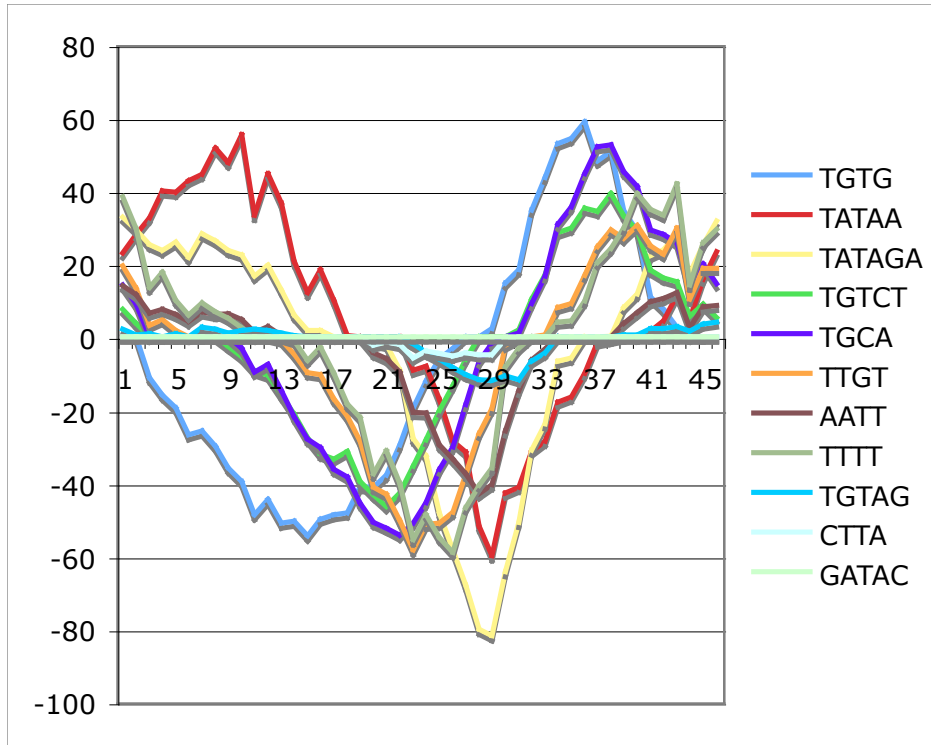
Supplemental Figure 2A.

Intergenic sequences upto the next adjacent ORF or contig break was obtained for each gene (5' in blue, 3' in red). Sequence lengths were binned and plotted as a histogram (lengths >5000bp not shown).



Supplemental Figure 2B.

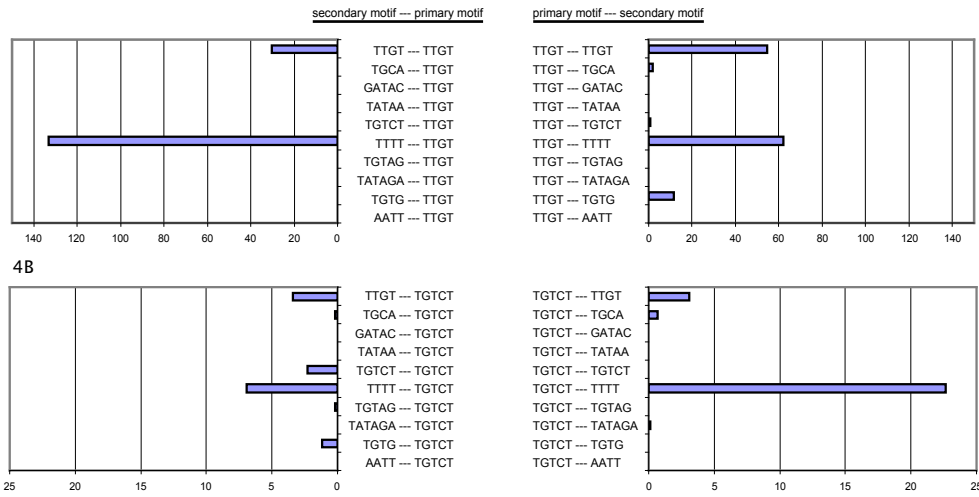
Intergenic sequence lengths of ORFs as a function of expression order in IDC. Each point represents a gene, black line represents mean value of sliding window of 100. Blue, 5' intergenic length. Red, 3' intergenic length.



Supplemental Figure 3

The motif profiles of the 11 representative motifs are shown. The x-axis is hours post invasion of the *P. falciparum* IDC, and the y-axis represents motif scores (see Materials & Methods).

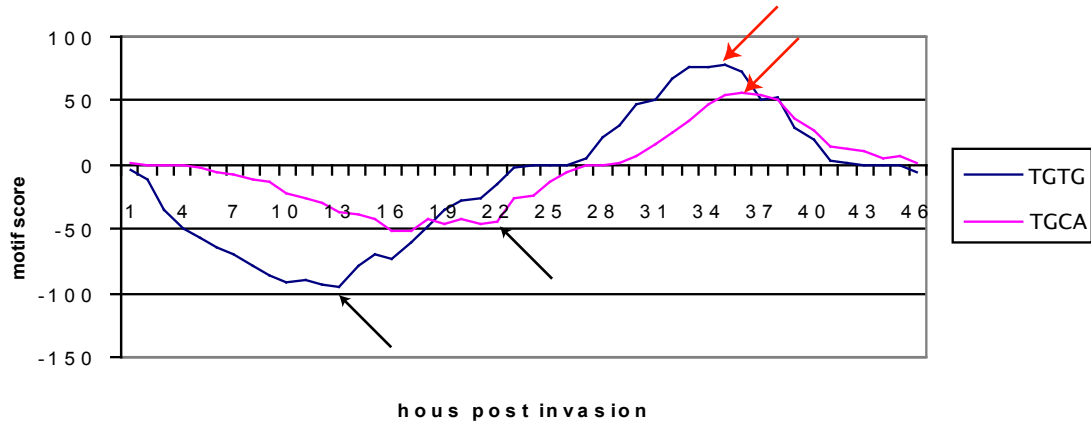
Supplemental Figure 4A



Supplemental Figure 4

Flanking motif co-occurrence analysis was carried out on two core motifs : A) TTGT and B) TGCT. Here, only the 10 best of the 11 motif families were included for the analysis of these two motifs. Note the two sets of plots for the two core motifs are not plotted to scale.

TGTG and TGCA motif profiles



Supplemental Figure 5

Motif profiles for the TGTG motif and TGCA motif have approximately a 9 hour discrepancy in correlation with upregulation (black arrows). The x-axis represents hours post invasion, the y-axis represents motif scores (see Material & Methods). The two motifs have less discrepancy for motif scores the time at which they are correlated with down-regulation (red arrows).

Chapter 3 :

Towards the biochemical purification of the *P. falciparum* TG binding factor implicated in the IDC clock

Abstract

Gene regulatory networks have been found to be central to the regulation of several periodic biological behaviors. We have been interested in determining the molecular components of the periodic transcriptome of the 48-hour intraerythrocytic developmental cycle of the malaria parasite, *Plasmodium falciparum*. Using bioinformatics methods, we discovered 11 *cis*-regulatory elements enriched in the promoter regions of the genes demonstrating cyclic transcript abundance, and one motif that we call the TG box motif was suggested to be involved in the coordination of 20% of these genes. The presence of a sequence-specific DNA binding activity in *P. falciparum* protein extracts was demonstrated. The breadth of the genome regulated by this factor underscores the importance of this protein in the lifecycle of the parasite, and here we describe our pursuit of the biochemical purification this factor we call the TG binding factor (TGBF) for its eventual identification. Our current best strategy employs ammonium sulfate precipitation, anion exchange, cation exchange, and sequence-specific DNA affinity chromatography. One highly purified fraction retained DNA binding activity that may be sequence-specific, and this fraction contained one polypeptide of approximately 12 kD

that was purified to greater than 90% homogeneity. Future improvements to the strategy will be discussed.

Introduction

Malaria is caused by parasitic protozoa of the genus *Plasmodium*. The four species of *Plasmodia* that cause human malaria include *P. falciparum*, *P. vivax*, *P. ovale*, and *P. malariae*, with *P. falciparum* causing the most morbidity and mortality. The famous physician-scientist Camillo Golgi is credited for having first suggested that the differences in period lengths of the recurrent fevers of malaria are associated with infection by different species of plasmodia (Muscatello 2007). This hypothesis has been confirmed; *P. falciparum*, *P. vivax*, and *P. ovale* have an approximately 48-hour intraerythrocytic developmental cycles (IDCs), while *P. malariae* has a 72-hour IDC (reviewed in Bray 1982). Furthermore, several mouse malaria species have 24-hour IDCs including *P. chabaudi* and *P. vinckei*, while *P. berghei* and *P. yoelii* are reported to have IDCs of 21 and 18 hours respectively (Landau et al. 1998). The molecular nature of the clock that regulates the periodicity of the plasmodial IDC remains elusive.

Recent work from the DeRisi lab has shown that during the *P. falciparum* IDC, approximately 50% of the genome is transcribed in a periodic manner, with most genes demonstrating one peak of mRNA abundance. However, the molecular mechanism for this global periodicity of transcriptional regulation remains unknown. A pertinent observation made by a bioinformatics survey of transcription factors is that there is a

paucity of DNA binding proteins in the *P. falciparum* genome (Coulson 2004). Speculating that the molecular components of the IDC clock coincide with the gene regulatory network governing the IDC, we have pursued the determination of the architecture of this network. In work described in chapter 2, we have applied the bioinformatics method REDUCE to the IDC transcriptome dataset to predict 11 *cis*-regulatory motif families which correlate with the expression of the downstream genes. One motif, which we call the TG box motif, was predicted to be involved in the transcription of 20% of the periodic IDC transcriptome, and validated by electrophoretic mobility shift assay (EMSA) to bind a factor in *P. falciparum* nuclear extract with sequence-specificity. A similar bioinformatics study corroborated our detection of this motif (van Noort et al. 2006). As of yet, none of the *trans*-acting factors governing the periodic transcription of the IDC have been determined. This chapter details work towards the determination of the molecular identity of the *trans*-acting factor specific for the TG box motif.

Having previously developed an electrophoretic mobility shift assay (EMSA) for a factor with specificity for the TG box motif, we have employed this assay for the eventual biochemical purification of this factor, which we refer to as TG binding factor (TGBF). Several different variants in purification strategy have been tried. After investigation of many candidate strategies, the currently most promising route for the purification employs ammonium sulfate precipitation, followed by anion exchange chromatography, cation exchange chromatography, and DNA affinity chromatography. The cation exchange chromatography and the DNA affinity chromatography steps are likely to

benefit significantly from further optimization. One of the elution fractions from the cation exchange chromatography has been purified further via sequence-specific DNA affinity chromatography, yielding an elution which contains a protein purified to greater than 95% purity, as determined by absence of other bands in a silver stain of PAGE. Furthermore, this fraction demonstrated DNA binding activity against a probe containing the TG box motif. However, the sequence specificity of this fraction could not be determined with confidence at this time. Directions for improvements for future attempts will be discussed.

Materials and Methods

Tissue culture

P. falciparum cultures were maintained in human erythrocytes obtained from volunteers, including myself, as per UCSF IRB requirements. Infected erythrocytes were grown as a suspension culture of approximately 2% hematocrit in RPMI media supplemented with Albumax II as described (Bozdech et al. 2003). Typically, parasites were harvested at a time in the IDC close to the developmental boundary of the ring and trophozoite stages.

Oligonucleotides used for electrophoretic mobility shift assay and for construction of affinity chromatography substrate

f1v2monoClamp probe was made by labeling f1v2monoClampF and f1v2monoClampRC separately by T4 PNK in 20 μ l volume reactions containing 5 pmol oligo, 20 nmol γ 32 P ATP (MPI catalog number 3502002), and 20 units T4 polynucleotide kinase (PNK; NEB M0201S) using the buffer supplied. Phosphorylation reactions were allowed to proceed for 1 h at 37 $^{\circ}$ C, then the labeled complementary oligos were mixed into one tube, placed on a 90 $^{\circ}$ C heat block, and allowed to cool over 3 hours to room temperature to anneal. This annealed probe was purified on a 15% polyacrylamide gel in 1X TBE buffer and eluted and stored in TE pH 8.0 at 40 $^{\circ}$ C. Specific activity was determined by a scintillation counter to be typically in the range of 4-10k CPM per fmol.

f1v2monoClamp2 probe was constructed similarly, from separately labeled f1v2monoClamp2F and f1v2monoClamp2RC. The nomenclature of the probes used in these studies is organized such that the probes are named by removing the last 1-2 characters of the oligo name specifying the strand of the probe design. Thus, probe f1v2monoClamp2 derives from the annealing of f1v2monoClamp2F and f1v2monoClamp2RC. The sequences of various oligonucleotides used are detailed in the section below. Unlabeled competitor probes were generated in a manner similar to the radiolabeled probe, by annealing 5pmol of unlabeled oligo, and annealing as per above in 1X T4 PNK buffer (NEB M0201S). All experiments were documented by drying gels and exposing to phosphorimager plates, followed by scanning on a Typhoon 9400 (GE Healthcare Life Sciences) imaging system. All image quantitations were done using ImageJ version 1.36b and pixel intensities of scans were squared, to correct for normalization of the .gel file standard.

The EMSA reported in Figure 1 was done in binding buffers containing 20 mM Hepes pH 7.6, 41 mM KCl, 1 mM EDTA, 1mM DTT, 0.5 mM PMSF, 5% glycerol, 0.03% NP-40, 0.125 µg/µl BSA, 0.15 ug/ul dI-dC (Roche cat #1 219 847), and 15 fmol/µl noMotifCompetitor oligo (www.idtdna.com). Separation was done on a 5% polyacrylamide gel in 0.25X TBE at 85V for 3 hours. Approximately 2.5 fmol of double stranded f1v2monoClamp probe were used for this assay in the presence of 500 fmol of various unlabeled competitors. All reagents including radiolabeled probe were pre-mixed prior to addition of 5 µg of nuclear extract, which was prepared in a fashion similar to the method described in Chapter 2.

Later competitor EMSA experiments were performed with extract pre-incubated with higher amounts of cold competitor for 15 minutes at room temperature, prior to addition and 15 min incubation with probe, as this gave better signal to noise discrimination of the TGBF. Some assays were also done in 5% polyacrylamide in 0.5X TBE gels since these conditions allowed better migration of complexes in cases when chromatographic fractions were in salt concentrations greater than 0.2 M KCl.

The EMSA depicted in Figure 2 are from fractions from various steps from the purification series depicted in Figure 2. Competitor EMSAs were done by pre-binding of extract to 5000 fold molar excess of probe (f1v2monoClamp2), or a mutant probe (f1v2monoClampSc2), to allow the determination of the presence of TGBF sequence-specific binding, in an internally consistent manner (comparisons between samples

should be made with caution, as different quantities of protein in slightly different binding buffers were loaded in this experiment). Inputs to EMSA are described in volume equivalents, relative to the flowthrough fraction from the HiPrep 16/10 Q FF (described below). Volume equivalents were determined by dividing the volume of extract used per EMSA by the total volume of the fraction, then normalizing to the Q sepharose fastflow/flowthrough fraction. Two fractions of flowthrough were tested, as the UV monitor of the AKTA suggested some breakthrough may have occurred during this chromatographic run. Separation was on a 5% polyacrylamide/0.5X TBE gel run at 90V for 2 hours.

Fraction	Relative volume equivalents assayed	approximate [KCl] in binding reaction
35% ammonium sulfate/supernatant	0.38	0.05M KCl buffer A
35% ammonium sulfate/pellet	0.54	0.225M KCl buffer A
HiPrep 16/10 Q FF/flowthrough	1.0	0.15M KCl buffer A
HiPrep 16/10 Q FF/flowthrough 2	1.0	0.15M KCl buffer A
HiPrep 16/10 Q FF/0.3M-0.5M KCl elution	1.2	~0.2M KCl buffer A
HiPrep 16/10 Q FF/0.5M-0.8M KCl elution	10.	0.1M KCl buffer A

The EMSA depicted in Figure 3 was done with samples dialyzed to 0.2M KCl buffer C (see section below on Whole cell extraction of parasite protein), with input protein quantities normalized by BCA estimates to 1 μ g. The EMSA depicted in Figure 4 was done with samples in 0.4M KCl buffer C in a 5% polyacrylamide/0.5X TBE gel, except the eluate fraction which was in approximately 0.25M KCl buffer C. Input volumes were determined in relative volume equivalents.

Fraction	Relative volume equivalents assayed	approximate [KCl] in binding reaction
Streptavidin-only/load	1	0.2M
Streptavidin-only/flowthrough (=affinity/load)	~1	0.2M
Streptavidin-only/eluate	~10	0.2M
Streptavidin-DNA/flowthrough	~2	0.4M
Streptavidin-DNA/wash 1	~2	0.4M
Streptavidin-DNA/wash 2	~2	0.4M
Streptavidin-DNA/elution	~4	0.25M

The poly dI-dC concentrations varied slightly between the fractions in this series, with wash 1 and wash 2 containing 10 ng/ μ l, and the elution fraction 0.2 ng/ μ l. Separation was done in 5% polyacrylamide/0.5X TBE. The experiments of Figure 3 and Figure 4 were performed with 5000 fold molar excess of cold competitor probes prior to addition of radiolabeled probe.

Construction of DNA affinity chromatography template

A substrate for DNA affinity chromatography was made via a ligation and PCR strategy resulting in the construction concatamers of TG boxes. Briefly, 600 pmol of family1-5merV2-F and family1-5merV2-RC-4lig were phosphorylated independently with 50 units of T4 PNK in 1X ligase buffer in 100 μ l reactions, then incubated for 2 hours at 37⁰C.

The two oligos were mixed and annealed by heating to 100⁰C followed by slow cooling over 3 hours. These oligos are designed to anneal such that in their optimal conformation, they create a duplex with a 5 bp overhang. The annealed oligos were loaded on a microcon-30 microconcentrator (Millipore 42409) at 4⁰C to buffer exchange with fresh cold 1X ligase buffer, to replete ATP and remove ADP. Annealed oligos were ligated in a 200 µl volume in 1X ligase buffer with 4000 units of ligase for three days at 16⁰C. Ligated material was electrophoresed on a 1% agarose gel in 1X TBE, and products in the range of 1.5kb were excised and electroeluted. This material was ethanol precipitated, then resuspended in 400 µl in TE pH 8.0. The material was used as input to a standard 20 ml PCR reaction volume, using the bio-family1-5merV2-F and family1-5merV2-RC-4lig as primers using a cycling program of 94⁰C 30'' - 56⁰C 30'' - 72⁰C 30'', repeated for 40 cycles, followed by a final 72⁰C extension for 5 minutes. PCR product was pooled, and extracted with a 25:24:1 mix of phenol:chloroform:isoamyl alcohol (Ambion AM9732) prior to ethanol precipitation. The DNA pellet was resuspended in 1 ml of TE pH 8.0 yielding a total of 8 mg of biotinylated DNA consisting of concatamerized TG box motifs. This material was bound to Dynal streptavidin beads (Invitrogen 112-06D). 5 ml of resin containing 5 mg of streptavidin sites bound approximately 360 µg of DNA. After use, resin was washed in buffer consisting of 10 mM Tris pH 8.0, 1M NaCl, 1 mM EDTA, to remove residual binding proteins.

Whole cell extraction of parasite protein

Cultures of infected erythrocytes (iRBCs) were grown at 2% hematocrit as described above, drained of excess media, then washed with pre-warmed media in a 1:1 ratio of resuspended iRBC to fresh media. The iRBC suspensions were then centrifuged at 500 x g in a Beckman GS-6R rotor at room temperature for 5 minutes without brake. Selective lysis of erythrocyte membranes was achieved by resuspending the iRBC pellet in ice cold 0.05% saponin (Sigma S-2149) in 1X PBS solution, using a ratio of 0.05% saponin/1X PBS to pellet exceeding 10:1. This material was centrifuged again in the GS-6R rotor at 2000 RPM at 4⁰C for 10 minutes without brake, pelleting free parasites. To reduce hemoglobin contamination, these parasite pellets were washed by two more rounds of resuspension in ice cold 0.05% saponin/1X PBS solution followed by centrifugation, then frozen in liquid nitrogen and stored until later protein extraction.

Starting material consisted of parasite pellets from 2.3 L of culture volume, with an average stage distribution of 5.5% rings, 9.6% trophozoites, and 0.5% schizonts for a total average parasitemia of 15%. Using 90 fL as mean corpuscular volume of erythrocytes, this leads to an approximation of 8×10^{10} parasites. Frozen parasite pellets were thawed into 20 ml of ice cold extraction buffer consisting of 20 mM Hepes pH 7.6, 1 M KCl, 1 mM EDTA, 1 mM DTT (Sigma D9779-10G), 1 mM PMSF, 1X complete protease inhibitor cocktail (Roche 11873580001), 0.05% Igepal (Sigma 56741-250ML-F), and 10 mM spermine (Sigma 85605-5G). The parasite pellets were resuspended thoroughly, and sonicated using a Fisher Scientific Sonic Dismembrator Model 500, set to 20% amplitude, with alternation of 1 second pulses with 1 second rests for 20 seconds. This material was allowed to stand for 20 minutes at 4⁰C with periodic agitation, and

centrifuged for 10 minutes at 20,000 x g in a Sorvall SLA-1500 rotor. The supernatant was dialyzed into 500 ml of precipitation buffer consisting of 20 mM Hepes pH 7.5, 100 mM KCl, 1mM EDTA, 1mM DTT, 1mM PMSF, and 1X complete protease inhibitor in a 12-30 ml Slide-a-lyzer cassette with a 20 kD molecular weight cutoff (MWCO) membrane (Pierce 66030). The pellet was re-extracted two more times as described above, but with only 5 ml of extraction buffer in these subsequent extractions, and these extracts were also pelleted and supernatants pooled with the initial extract into the dialysis cassette. Similar experiments suggested that further extraction from pellets yield negligible TG binding factor (TGBF).

After more than 2 hours of dialysis into precipitation buffer with three buffer changes, extracts were re-centrifuged at 20,000 x g for 1 hour to yield an insoluble brown pellet surrounded by a glassy halo. Prior iterations of this method indicated that the pellet did not contain detectable TGBF. The supernatant was brought to 35% saturation of ammonium sulfate over 30 minutes on ice, with stirring to prevent local over-precipitation. Upon addition of the last of the ammonium sulfate, the extracts were allowed to precipitate for another 30 minutes prior to centrifugation in a Sorvall SLA-1500 rotor at 20,000 x g for 1 hour at 4⁰C.

The ammonium sulfate precipitate was resuspended in a total of 20 ml of buffer A (20 mM Hepes pH 7.5, 0.1 M KCl, 1 mM EDTA) with 1 mM DTT, 1 mM PMSF, and 1X complete protease inhibitor cocktail, then dialyzed in a 20kD MWCO slide-a-lyzer cassette for 30 minutes into 500 ml of 0.1M KCl buffer A three times, then for 30

minutes into 500 ml of 0.3M KCl buffer A twice, with 1 mM DTT, 1 mM PMSF added fresh. 1X complete protease inhibitor cocktail was added to the extract upon completion of dialysis, prior to loading on a 20 ml HiPrep 16/10 Q FF (GE 17-5190-01) pre-equilibrated to 0.3M KCl buffer A with 1mM DTT at 10⁰C on an AKTA Explorer system. Extract was loaded at 0.5 ml/min, and 1 column volume of flowthrough was collected in two halves, using the UV monitor to determine return to baseline. Washing was done with 2 column equivalents of 0.3M KCl buffer A, and elution was performed with step gradients between 0.3M-0.5M KCl and 0.5M-0.8M KCl, yielding 20 ml and 5 ml of material respectively. The latter fraction was dialyzed to 0.2M KCl buffer A for analysis. 1X complete protease inhibitor cocktail was added to elutions soon after they were obtained. At this point, the 0.3M-0.5M KCl elution fraction from the anion exchange was confirmed to contain the detectable majority of the TGBF.

Prior to a second round of spermine precipitation of the 0.3M-0.5M KCl elution, the fraction was dialyzed using a 20kD MWCO Slide-a-lyzer cassette into a buffer consisting of 20 mM Hepes pH 7.5, 0.1M KCl, 1 mM EDTA, and 5% glycerol. The precipitation was done by addition of spermine to 10 mM, followed by a 30 minute incubation and centrifugation at 20,000 x g for 1 hour. The supernatant was dialyzed using a 20kD MWCO Slide-a-lyzer cassette into a buffer of 25 mM citrate pH 5.0, 1mM EDTA, and 10% glycerol in a 20 kD slide-a-lyzer cassette (Pierce 66012) for two hours. Dialysates were clarified by centrifugation at 20,000 x g for 1 hour, prior to loading on to a 6 ml Resource S column (GE 17-1180-01). Extract was loaded at 0.2 ml/min and 1 column volume was collected. Washing was in 2 column volumes of 0.0M NaCl buffer B, and

elution fractions were obtained as NaCl step gradients in buffer B at 0.15M NaCl, 0.3M NaCl, 0.6M NaCl, and 1.0M NaCl. A final elution was also obtained using a non-acidic buffer, 0.2M KCl buffer C (20 mM Hepes pH 8.0, 0.2M KCl, and 10% glycerol). All Resource S fractions were dialyzed in 20kD MWCO Slide-a-lyzer cassettes in 500 ml of buffer C with 1 mM DTT added fresh for 1 hour. Protein quantitations were done by BCA (Pierce), and EMSA was performed on 1 μ g of protein in the presence of 0.1 μ g dI-dC similar to described above. 1 μ g of protein from each fraction was also analyzed by Coomassie staining of PAGE, with 1 μ l of Benchmark ladder (Invitrogen 10747-012) and 1 μ l of Novex ladder (Invitrogen LC5801) for comparison.

Finally, the 0.6-1.0M NaCl fraction and the 0.2M KCl buffer C fractions from the Resource S chromatography were pooled to a 5 ml volume containing approximately 100 mg protein, and this starting material for the affinity purification series of experiments constitutes 1X for the determination of volume equivalents used for subsequent analysis. The pooled S fractions were first loaded on a 10 ml bed volume of immobilized streptavidin-agarose resin (Pierce 20353). The flowthrough from this pre-clearing step was approximately 12 ml, and a 300 μ l aliquot was re-concentrated to 1X using a microcon-30 and used for EMSA and PAGE analysis. Elutions were obtained from the streptavidin-resin column using 1.0M KCl buffer C, and the 22 ml of elution from this column was dialyzed to 0.2M KCl buffer C in a 20kD MWCO Slide-a-lyzer in 500 ml for 2 hours, prior to concentration with a 50kD MWCO Amicon Ultra concentrator (Millipore UFC 9 050 08) to a final volume of 0.5 ml. By adding 2 ml of 1.6M KCl buffer C to the flowthrough of the pre-clearing step, the KCl concentration was adjusted

to 0.4M prior to binding on approximately 300 µg of concatamerized TG box bindings sites bound to 10 mg of streptavidin on Dynal beads (see Materials & Methods). Binding was allowed to proceed for 12 hours at 4⁰C with nutation, then washed twice with 2.5 ml of 0.4M buffer C in the presence of 25 µg of poly-dIdC, then washed twice more in 0.4M buffer C without poly-dIdC, and eluted to 1.0M buffer C to a final volume of 300 µl. 5 µl of the final eluate was analyzed by EMSA by dilution with 15 µl of 0.0M KCl buffer D (see above section).

Approximately 240 µl of the elution from the affinity series was precipitated by methanol precipitation (Wessel 1984) and resuspended in 30 µl of 1X sample loading buffer (Invitrogen NP0008). 15 µL was loaded on the PAGE shown in Figure 4F. The quantities of the other fractions loaded on the PAGE in Figure 4F are 1/20 of the amounts used for the EMSA gels of Figure 4B and 4C, except for the Streptavidin-only/load fraction (the pooled S fractions) which was 1/400 of the amount used in EMSA. The Benchmark ladder (Invitrogen 10747-012) is loaded at an equivalent of 0.02 µl, and at this volume, bands in the ladder are approximately 2ng.

Fraction	Relative volume equivalents assayed
Streptavidin-only/load	0.05
Streptavidin-only/flowthrough (=affinity/load)	~1
Streptavidin-only/eluate	~10
Streptavidin-DNA/flowthrough	~2
Streptavidin-DNA/wash 1	~2

Streptavidin-DNA/wash 2	~2
Streptavidin-DNA/elution	~30

Protein analysis

All protein gels were Invitrogen NuPAGE Bis-Tris gels of 4-12% acrylamide (Invitrogen NP0323, NP0321), run in 1X MOPS buffer. Silver staining of the gels was done using the Invitrogen SilverQuest Kit (Invitrogen LC6070). Protein quantitations were done using the BCA assay (Pierce 23223, 23224), and standard curves were made by triplicate measurements of BSA (NEB) dilution series.

Chemicals

All common chemicals were obtained from Sigma or Fisher unless otherwise specified.

Results

Cold competitor EMSA with an improved probe containing the TG box motif

Previously, we had demonstrated sequence-specific binding of the TG box motif by a factor present in *P. falciparum* protein extracts (see Chapter 2). During the course of multiple experiments using this probe, it was observed that the short length of the probe and the high AT-content lead to the disassembly of the duplex over several days. To stabilize the probe in a duplex, we added GC-rich flanking sequences to both sides of the probe used in Chapter 2 producing in the f1v2monoClamp probe. A competitor EMSA

was performed with the f1v2monoClamp probe in which extract was bound to probe in the presence of 200 fold molar excess of various competitor DNAs (Figure 1). These results confirmed the presence of a sequence-specific DNA binding activity with preference for our MODEM predicted binding site, through a slightly different experimental approach and a slightly different probe than used previously (Chapter 2). Furthermore, the stabilized probe proved reliable for a longer period than the probes used in Chapter 2.

Although computational analysis had suggested the possibility that the TG box motif may be part of a polymeric repetitive DNA of TG repeats, we found that a pure TG polymer consisting of 15 repeats of TG was not able to compete against the monomeric probe under these conditions (Figure 1, lane 1). This result does not rule out the possibility that *in vivo* binding sites may be arranged with multiple TG box instances in close proximity (see Discussion).

Given the recent report that *Plasmodium falciparum* TATA-binding protein (PfTBP) binds to a non-canonical TATA box (Ruvalcaba-Salazar et al. 2005) similar in sequence to our predicted TG box motif, we also tested the ability of this non-canonical TATA box sequence (SalazarG29 probe, see Materials and Methods Oligo Table) to compete against our predicted binding site probe. At 200-fold molar excess, this non-canonical TATA box probe did not compete noticeably against our probe for the TG box motif (Figure 1, non-canonical TATA box).

More thorough investigation of the properties of the f1v2monoClamp probe indicated that the extra nucleotides appended to the motif for *in vitro* stability affected the labeling of the oligos in a sequence-dependent manner, specifically that the T4 PNK had lower activity against oligos with a 5' end cytosine followed by three guanines (data not shown). Sequence preference has been previously documented for T4 PNK (Lillehaug et al. 1975). In order to obtain more equivalent labeling of the oligos, a variant of the initial clamped probe was designed such that all 5' nucleotides are guanosine; these probes exhibited more similar labeling properties. Experiments in which extracts were allowed to pre-bind to competitor probes prior to the addition of radiolabeled probe demonstrated better signal to noise discrimination for the TGBF and subsequent experiments employed this variant of the competitor EMSA.

Towards a purification strategy for the identification of the TGBF

Using this cold competitor EMSA, we pursued the development of a purification strategy aimed towards the identification of the TG binding factor (TGBF). In the initial approach, parasites were released from erythrocytes via saponin lysis, treated with hypotonic lysis, and centrifuged to separate nuclear and cytoplasmic fractions. Using the cytoplasmic fraction from this method as starting material for ammonium sulfate precipitation at 45% saturation, the pellet was resuspended and separated by anion exchange chromatography and DNA affinity chromatography producing a fraction that retained sequence-specific binding activity for the TG box motif. This fraction was notable for three faint bands of protein visible by silver staining when separated by PAGE, however there was not

sufficient yield for identification of the protein by mass spectrometry (Falick 2006). Subsequent work focused on use of whole cell extraction methods, yielding more starting material with greater ease.

The most promising approach attempted so far is described here in detail. Frozen parasite pellets obtained from saponin lysis of 2.3L of infected RBCs from parasite culture grown at 2% hematocrit were thawed into a high salt extraction buffer containing 10 mM spermine for whole cell protein extraction. This material was then precipitated at 35% ammonium sulfate saturation, which reproducibly precipitates all detectable TGBF from extracts prepared by this method (data not shown). The pellet from this precipitation was then resuspended in 0.3M KCl buffer A (see Materials and Methods) and quantitated by BCA protein assay to be approximately 200 mg of protein. The sample was then loaded on to a HiPrep Q Sepharose Fast Flow column. Elution between 0.3M and 0.5M KCl in buffer A yields most of the TGBF, and eliminates a highly abundant non-specific DNA binding activity enriched in the ring stage extracts (Chapter 2, Figure 4) into the flow through fraction (Figure 2B). The 0.5M-0.8M elution retains no detectable TGBF, indicating that this elution protocol is nearly optimal in this buffer system. Total protein yield in the 0.3-0.5M KCl fraction was approximated however, interference from contaminants made this value unreliable. Further work needs to be done to quantify this fraction carefully.

The efficacy of nucleic acid precipitation by spermine has been noted to be significantly improved at lower salt concentrations (Murphy et al. 1999). To ensure complete nucleic

acid removal prior to cation exchange chromatography, the 0.3-0.5M KCl eluate from the anion exchange step was dialyzed to 20 mM Hepes pH 7.5, 0.1M KCl, 1 mM EDTA, 5% glycerol, and precipitated again with 10 mM spermine for 30 minutes. 1 hour of centrifugation at 20,000 x g yielded a glassy yellow precipitate, however agarose gel analysis of this resuspended pellet suggested that there was no remaining nucleic acid in this sample, indicating that this step may have been unnecessary.

In preparation for loading on a cation exchange column, the extract was dialyzed into an acidic buffer consisting of buffer B (see Materials & Methods). After equilibration, extract was clarified by centrifugation, then loaded onto a 6 ml Resource S column pre-equilibrated to buffer B. Elutions were performed by step elutions of 0.0M-0.15M NaCl, 0.15M-0.3M NaCl, 0.3M-0.6M NaCl, and 0.6M-1.0M NaCl in buffer B. A final elution was performed in 0.2M KCl buffer C. Fractions were dialyzed to 0.2M KCl buffer C, quantitated by BCA, and analyzed by EMSA to determine approximate specific activities for fractions. Previous work had indicated that elution of TGBF occurs on a 5 ml HiTrap S column in this buffer system between 0.4 and 0.6M NaCl. However, on this particular Resource S column, the TGBF eluted in a distribution across most fractions; the elution fraction between 0.3M and 0.6M NaCl possessed the highest specific activity (Figure 3D), as may be anticipated based on previous experiments. The elution fraction between 0.15M and 0.3M NaCl possessed the second highest specific activity among the eluates. Subsequent, PAGE analysis using 1 μ g of protein, as estimated by BCA, revealed that the BCA protein quantitation for the 0.15M-0.3M NaCl fraction was a significant overestimate (Figure 3F), possibly due to low concentration of protein, or the presence of

a contaminant. A qualitative estimate of the protein quantity of the 0.15M-0.3M fraction based on Coomassie band intensities suggests that the BCA-based quantitation of specific activity is likely a 5-10 fold underestimate. The protein concentration of the 0.0M-0.15M NaCl fraction also was overestimated by BCA, though this fraction does not seem to have detectable TGBF (Figure 3F).

To assure the best probability of success with the DNA affinity chromatography method, the most promising fraction spanning the 0.15M-0.3M NaCl elution was frozen for later study, while a pilot study was carried out with the 0.6M-1.0M fraction pooled with the 0.2M buffer C elution fraction. This pooled fraction was loaded onto a 10 ml bed volume of streptavidin-conjugated agarose beads. Previous attempts had indicated that many *P. falciparum* proteins bind to streptavidin-conjugated resins even in the absence of immobilized DNA, and this step removes much protein with non-specific affinity for the DNA-free resin. Following pre-clearing of extract on the streptavidin-only resin, the flowthrough was mixed with 5 mg of Dynal beads previously bound with approximately 300 µg of concatamerized TG box motifs (Materials & Methods) and equilibrated to 0.4M KCl buffer C at 4⁰C. Binding was carried out at 4⁰C for 12 hours with nutation, and two washes were performed with 2.5 ml of 0.4M KCl buffer C containing 25 µg poly dI-dC, followed by two more washes of 2.5 ml of 0.4M KCl buffer C without poly dI-dC. Elution was done in a final volume of 300 µl of 1.0M buffer C, which represents a 16 fold concentration of the material initially loaded on to the streptavidin-agarose pre-clearing column. Thus, 5 µl of the final 300 µl eluate was analyzed by EMSA in Figure 4C lanes 7-9, which constitute 4-fold volume equivalents, as compared to 20 µl of the

original 5 ml starting volume loaded into lanes 1-3 of the EMSA in Figure 4B. This elution fraction demonstrated DNA binding activity (Figure 4C), and quantitative analysis of the appropriate regions of the image indicated that the relative intensities are consistent with the presence of TGBF in this fraction (Figure 4 D, E). On the other hand, the high background and low intensities of the bands in these gels due to sample underloading qualify this conclusion, and this experiment demands replication. It should be noted that the relatively high background in lane 2 and lane 3 of Figure 4B contribute to the seemingly poor sequence-specificity of the competitor probe assays in the quantitative analysis shown in Figure 4D.

To estimate the purity of the elution obtained from the affinity chromatography series, approximately 100 μ L of the eluate volume was concentrated by methanol precipitation, separated by PAGE, and silver stained according to manufacturer's suggestions (10 minute developing) demonstrating the existence of a protein of approximately 12kD purified to 90% purity in this fraction. The yield of this protein was approximately 1 ng, based on comparison to the standard protein ladder.

Discussion

Previous bioinformatics work described in Chapter 2 indicated that a high density of TG box motifs correlates with transcription during the IDC. Two models may explain this correlation. In the first scenario, a TG polymer acts as the functional regulatory sequence, leading to detection of extended TG box motif instances (Chapter 2 Figure 4A) correlated

with transcription. Previous work in mammalian transcription has documented the role of the TG polymer repeat on gene expression (Hamada et al. 1984), although the mechanism of regulation has not been explored in detail. An alternate model is that the TG box functions as a short motif of two to three TG repeats, and that multiple instances of the TG box motif cluster together in native promoters to contribute to transcription.

Homotypic clustering of transcription factor binding sites is a long recognized feature of functional promoters (Donahue et al. 1983). To test the first model, we performed a competitor EMSA using pure TG polymers to compete against a probe consisting of a monomeric TG box motif. While 200-fold molar excess of unlabeled monomeric TG box could compete against probe for the binding activity, 200-fold molar excess of unlabeled pure TG polymers of similar length did not compete to an appreciable degree. This result favors the second model, that in vivo binding sites are clusters of short monomeric TG box motifs. Their bioinformatics detection in high density suggests they may interact cooperatively on transcription. In fact, *var* gene intronic promoters have been observed to have TG box motifs organized in imperfect repeats (Calderwood et al. 2003). Further evidence for cooperative binding of the TGBF at promoters could be obtained through quantitative binding studies of dimeric TG box motifs, as compared against monomeric motifs. However, such studies require purified protein to be able to rule out effects of uncharacterized players in crude cell extracts.

Interest in the abovementioned model of transcriptional regulation in *P. falciparum*, as well as several other observations motivated us to seek the purification of the TG binding factor (TGBF). For one, TG box motifs correlate with 20% of the periodic transcription

during the IDC. Another motivation is the observation that this motif plays a part in the transcription of *var* genes, the main antigenic gene family in *P. falciparum* (Calderwood et al. 2003; Voss et al. 2003). Regulation of this class of genes displays several interesting properties, including allelic exclusion (Freitas-Junior et al. 2005; Voss et al. 2006) reminiscent of vertebrate immunoglobulin and odorant receptor genes, *var* switching (wherein the actively transcribed version changes between generations;), and mitotic recombination (Freitas-Junior et al. 2000). Given our interest in understanding the molecular mechanism of the IDC clock, as well as the nature of the link between IDC gene expression and *var* gene expression, we pursued a purification strategy for the TGBF employing, ammonium sulfate precipitation, anion exchange chromatography, cation exchange chromatography, and DNA affinity chromatography.

We believe that the early steps of the purification described in Figure 2 have been reasonably optimized, since this method reproducibly deposits TGBF in the elution of the anion exchange with little detectable activity in the other fractions. Determination of TGBF yields and specific activities for these steps are forthcoming.

Currently, the cation exchange chromatography and the DNA affinity chromatography steps require further optimization. To date, cation exchange chromatography has not resulted in reproducibly high yields after elution, and in the particular experimental series reported here, the use of a potentially damaged Resource S column may have led to the broad elution of TGBF observed. Residual spermine may have also competed for binding on the column leading to overloading, and contributing to the broad elution profile. On

the other hand, the second elution fraction between 0.15M and 0.3M NaCl yielded a high specific activity fraction, and Coomassie stained PAGE of this fraction showed it to be devoid of a doublet bands seen in the higher salt elution fractions of the series (lane 7-9 in Figure 3F). These high molecular weight proteins have been a persistent contaminant in many of the purifications attempted so far, and the lack of correlation of TGBF activity in EMSA with these bands in PAGE suggests that these are highly abundant contaminants that can be removed using this method.

The sequence-specific DNA affinity chromatography step is also not fully optimized yet, and previous quantitation of EMSA gel results of similar experiments revealed that the recovery of TGBF in the elution fraction is less than 10% of the starting material. Several problems could contribute to poor recover of TGBF in the elution of such affinity chromatography attempts. One explanation is that *P. falciparum* extracts contain excess amounts of non-specific DNA binding proteins with higher DNA affinity, or slower off rates, than TGBF, such that the binding sites are occupied by contaminant proteins prior to binding of TGBF in significant quantity. In fact, TGBF from crude nuclear extracts did not bind to the DNA affinity chromatography substrate at all, and use of non-specific competitor nucleic acids in the mobile phase of affinity chromatography during binding as well as washes represents an attractive solution to this problem (Kadonaga 1991) which we have explored with some promising results. Proteins with affinity for the streptavidin resin may also represent another class of undesired contaminant proteins. Others have commented that actin can bind non-specifically to streptavidin (Gadgil et al. 2001), as can abundant nuclear proteins such as mRNA processing factors, splicing

proteins, and ribosomal proteins (de Boer et al. 2003), suggesting that pre-clearing extracts on streptavidin-resin may be an effective route to depleting undesired high abundance proteins that can bind streptavidin. It is of potential relevance that a protein with sequence similarity to bacterial BirA was recently observed in the *P. falciparum* genome (Müller et al. 2007), and if this protein is a functional homolog of BirA, there may be multiple covalently biotinylated proteins in *Plasmodium* extract that would be expected to bind to the streptavidin-based affinity resins as well.

Other possible explanations for the current poor yield of the sequence-specific DNA affinity chromatography could be due to resin re-use, which would be affected by residual contaminant proteins bound to the column, or by the action of contaminating nuclease activity reducing the binding capacity of the column with each subsequent use. Reliance on freshly made affinity reagents may increase the reproducibility and binding capacity in future experiments. Kinetic parameters, ionic strength, temperature, and pH also represent orthogonal axes affecting binding of proteins to DNA, as each protein has a somewhat distinct optimal condition for binding to its target DNA. Continued exploration of these parameters presents valid directions for further study.

Finally, the 12kD protein eluted from the affinity chromatography series shown in Figure 4F deserves some attention. A low molecular weight protein in the size range observed from this experiment has been previously observed in a DNA affinity chromatography experiment (data now shown). Furthermore, the 0.3M-0.6M NaCl elution from the Resource S chromatography was further separated by DNA affinity chromatography as

well, without the benefit of a pre-clearing step on streptavidin-agarose. This experiment revealed the presence of numerous bands by silver stain, as compared to the elution shown in Figure 4F, confirming the utility of removing proteins that interact non-specifically with the affinity resin. However, the approximately 12kD low molecular weight band constituted the most prominent band. Quantitative analysis of the EMSA shown in Figure 4C suggested the presence of the TGBF in the elution fraction (Figure 4E), but the low signal intensities preclude confident interpretation of this data, and this result demands replication. In future iterations, EMSA of a larger proportion of the final elution may increase the signal to a range that credibly demonstrates that this band represents the TGBF. Currently, it remains a real possibility that the DNA binding activity in this fraction reflects a non-specific contaminant. To improve signal to noise in the EMSAs for highly pure protein, reduction of amounts of competitor nucleic acids used in the final EMSA should be considered. Using crude extracts, a 5000 fold molar excess of unlabeled probe gives clear evidence for sequence-specific DNA binding activity (Figure 2B), but when the protein sample is sufficiently purified, it may be the case that using only 1000 fold molar excess will be more optimal for demonstrating sequence specificity, while suppressing the signal to a lesser degree. The binding buffers used throughout these studies also contain 1000-fold molar excess of a single-stranded DNA oligo stripped of all of our predicted binding sites, the purpose of which is to suppress probe binding by the *P. falciparum* Replication Protein A (PfRPA) that has been described to be a major contaminating non-specific nucleic acid protein in *P. falciparum* extracts (Voss et al. 2002). In the case of highly purified protein, as in Figure 4F, the presence of this competitor oligo is likely superfluous, and detracts from the signal.

Future replicates of this experiment will avoid the use of extraneous competitors for EMSA analysis of highly purified fractions obtained this late in the purification. Some transcription factors have proven difficult to purify due to loss of activity upon separation from obligate co-factors, and this possibility should be considered here as well.

Obtaining sufficient quantities of material for analysis is a major obstacle for all biochemical purifications. Early transcription factor purification strategies which led to successful identification have utilized as much as 120 g of whole cell extract from 6×10^{10} cells in the case of NF-1 purification (Rosenfeld et al. 1986), and 3 g of nuclear extract in the case of hypoxia inducible factor-1 (HIF-1; (Wang et al. 1995)). Octamer binding protein was first purified from 100 mg of nuclear protein extracted from 1×10^{10} cells (Wang et al. 1987), and human lymphoid specific octamer-binding protein (OTF-2) identification required 770 mg of nuclear extract (Scheidereit et al. 1987). Advances in mass spectrometry have lowered the amounts of material needed to perform successful purifications, and the approximate number of parasite nuclei used here (8×10^{10} parasites) is in the range of many of these studies, suggesting that this approach could eventually yield the identity of the TGBF. However, more work is required to optimize the later steps of the purification strategy described here, as the yield of material obtained at the final step is not sufficient to compellingly correlate the 12kD protein purified with sequence-specific DNA binding activity. On the other hand, the identity of the protein purified in Figure 4F may yet be determined by mass spectrometry, as the band has been cut from the gel, destained, and preserved for later analysis at the time of this writing. The same size band from the DNA affinity chromatography performed on the 0.3M-0.6M

NaCl eluate of the Resource S column has been preserved as well. Although the similarity of the TG box motif and the non-canonical TATA box-like motif described recently (TGTAAG; (Ruvalcaba-Salazar et al. 2005)) had raised the possibility that the motifs were variants of the same transcription factor binding site, the experiments of Figure 1 in combination with the predicted molecular weights of the two predicted PFTBP ORFs (PFE0305w : 38kD, PF14_0267 : 42kD) make it unlikely that this 12kD band will be one of these PFTBP proteins.

Although a classical biochemical approach has been undertaken here towards the identification of the TGBF, other approaches may represent valid routes to the answer as well. The yeast one hybrid screen represents a powerful tool for the discovery of transcription factors, and the paucity of DNA binding proteins in the *P. falciparum* genome makes attainable a candidate expression approach. Recent works by other lab members are aimed at exploring these strategies. In particular, the demonstration of efficient *in vitro* transcription/translation of *P. falciparum* proteins by wheat germ extract has provided an important step forward for the latter strategy (Rathod 2007). However, there are some potential benefits of pursuing a biochemical purification strategy over these other approaches. For one, proper biological activity may require post-translational modifications that may be difficult or impossible to recapitulate in heterologous expression systems. Another consideration is that many transcription factors derive sequence specificity from the dimerization of two polypeptides, and one hybrid screens and candidate expression strategies may overlook DNA binding activities that require the interaction of two components. On the other hand, biochemical purification is fraught

with caveats as well. Biochemical interaction does not necessarily correlate with *in vivo* function, and the biological function of candidate transcription factors identified through any of these methods should be corroborated with *in vivo* evidence, be it conditional genetic ablation of the candidate regulator or a chromatin immunoprecipitation assay. The road to understanding the *Plasmodium* molecular clock remains an exciting one whose terminus has not yet come into sight.

Figure 1

200 fold
cold competitor
co-bound to probe

none

TG polymer

mut polymer

TG monomer

mut monomer

non-canonical TATA box

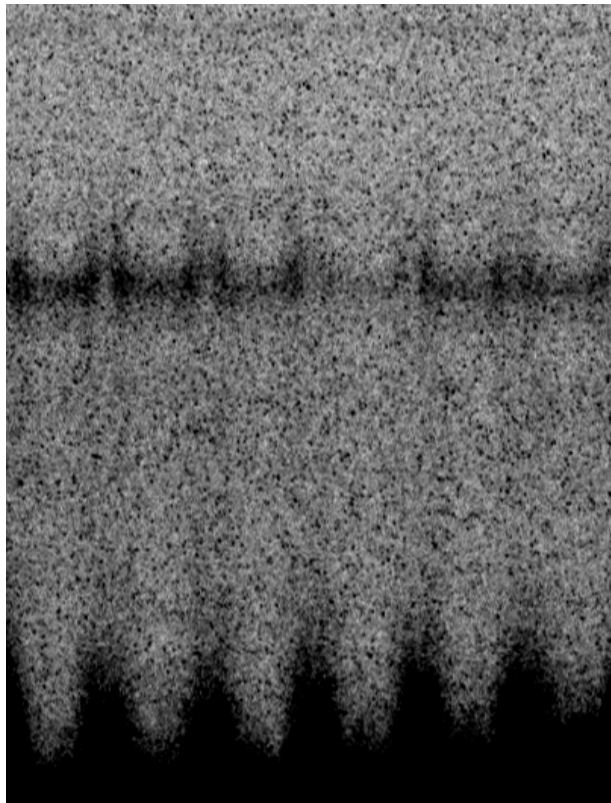


Figure 2A
TGBF Purification Scheme - Part I

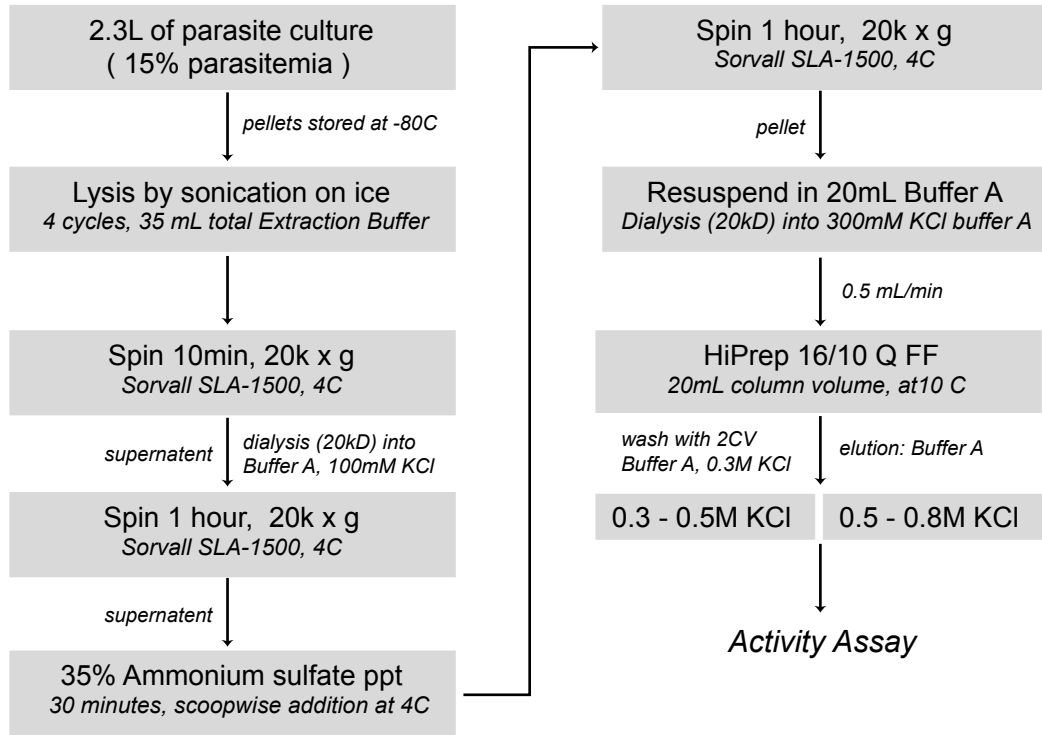


Figure 2B

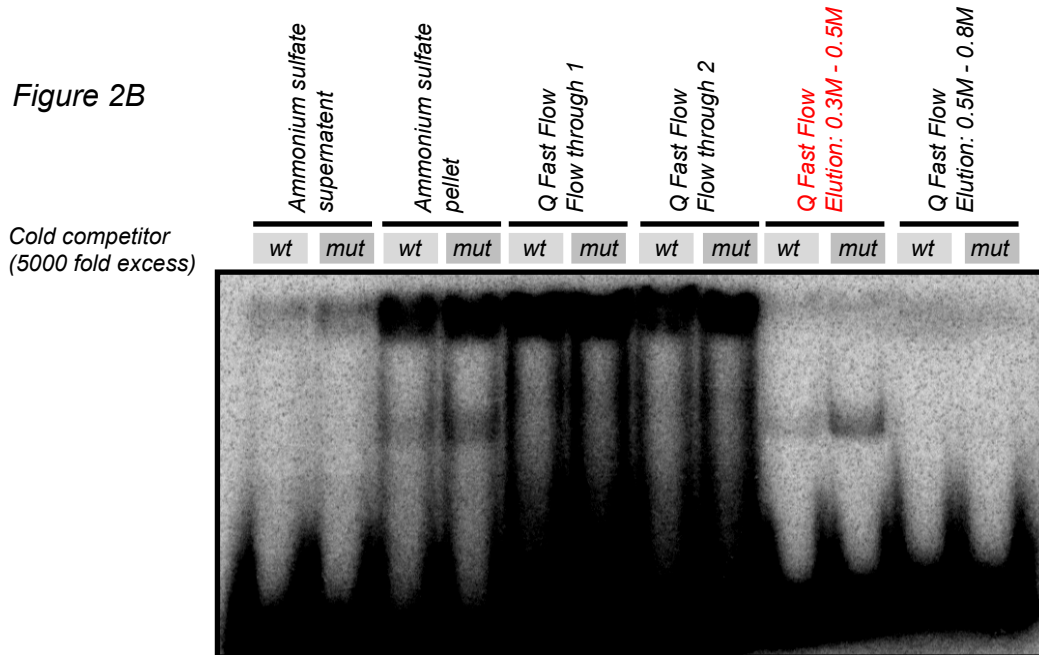


Figure 3A
TGBF Purification Scheme - Part II

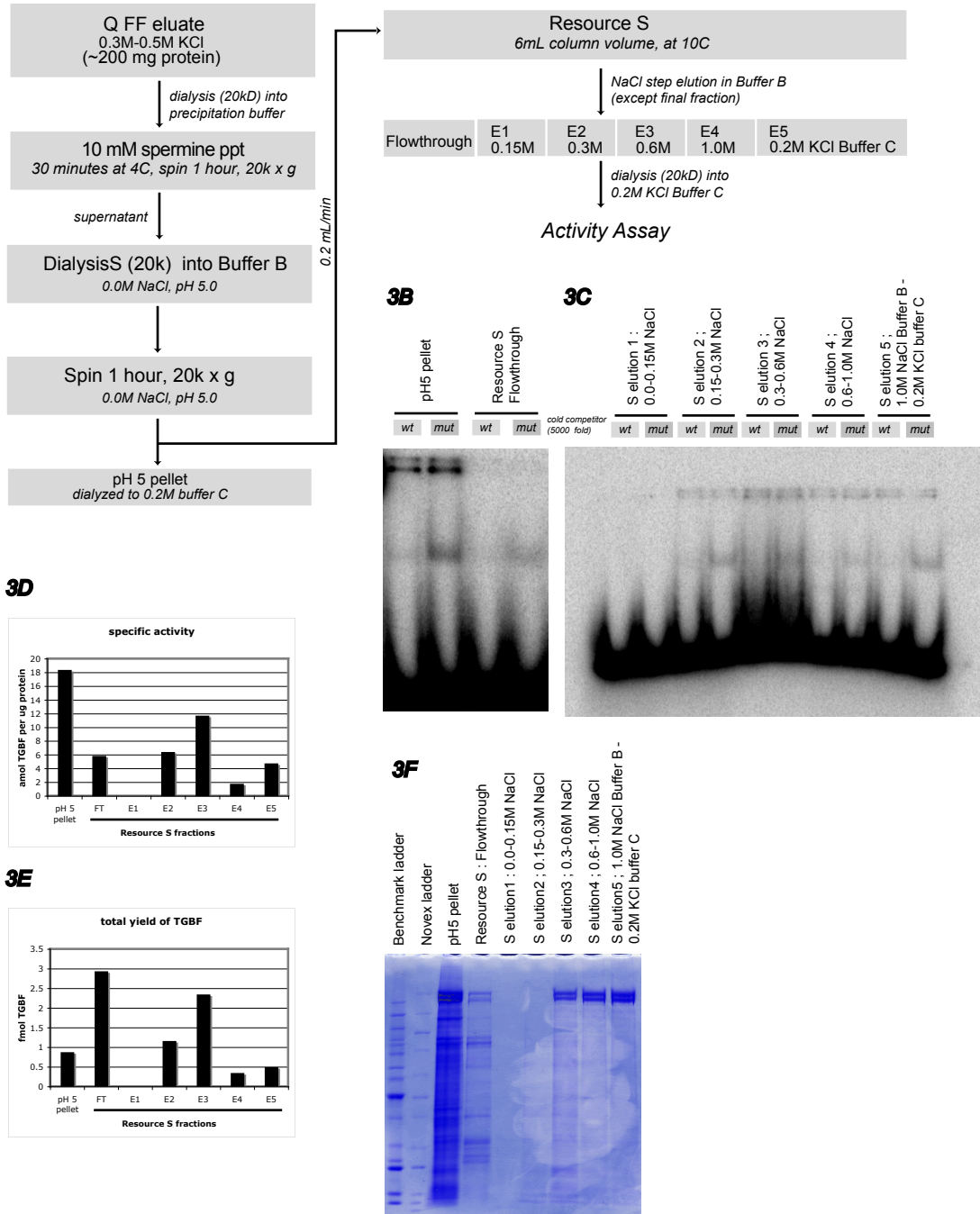
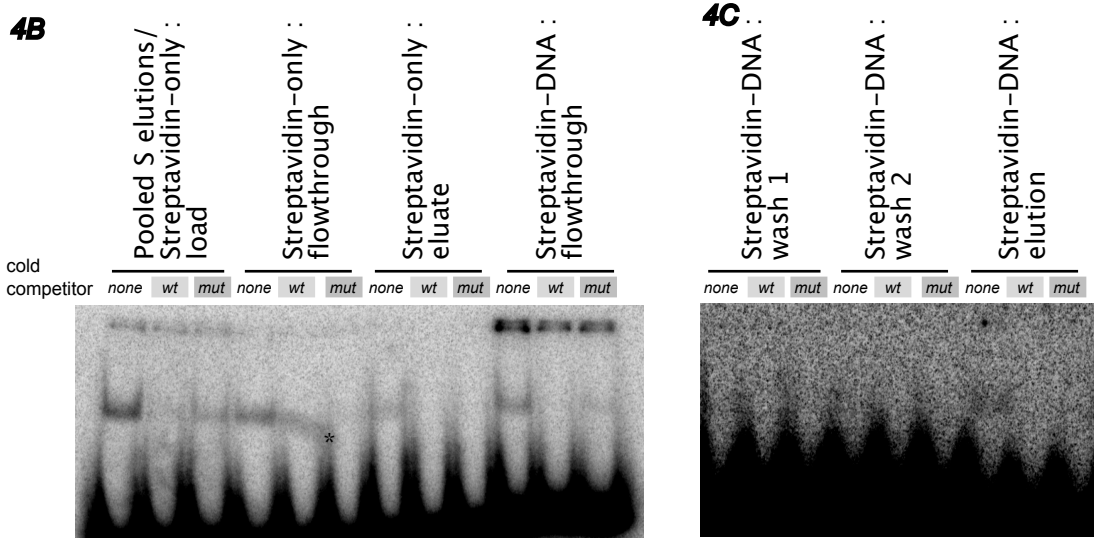
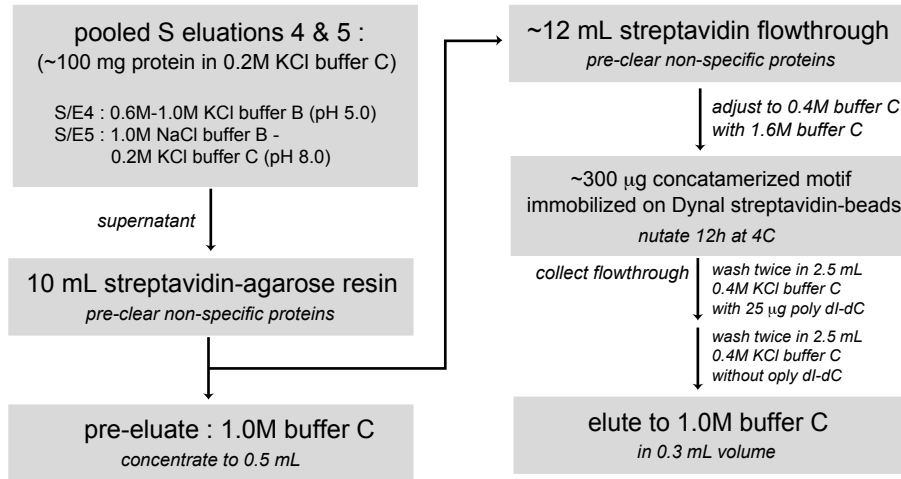
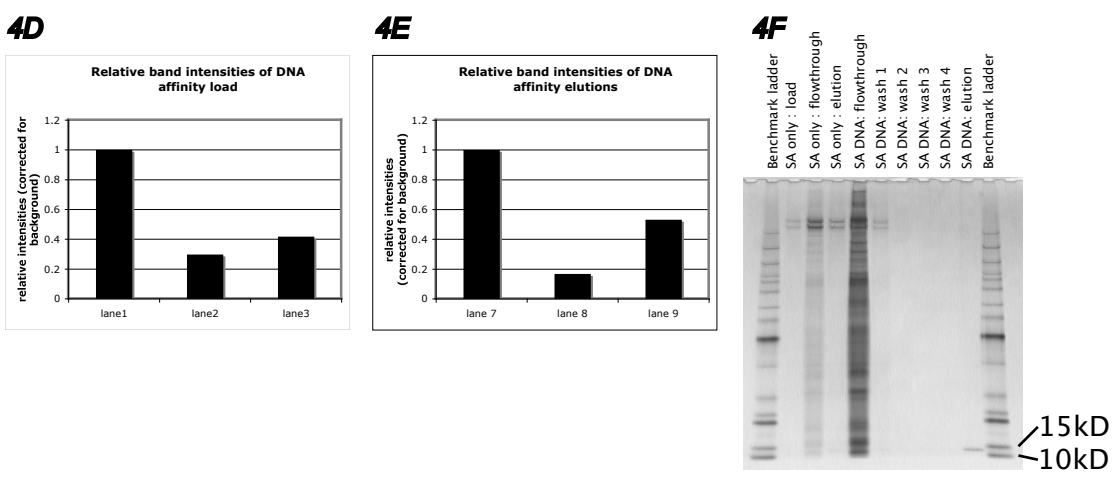


Figure 4A
TGBF Purification Scheme - Part III



* artefact



References

- Alano, P., F. Silvestrini, et al. (1996). "Structure and polymorphism of the upstream region of the pfg27/25 gene, transcriptionally regulated in gametocytogenesis of *Plasmodium falciparum*." Mol Biochem Parasitol **79**(2): 207-17.
- Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.
- Alker, A. P., P. Lim, et al. (2007). "Pfm^{dr1} and in vivo resistance to artesunate-mefloquine in falciparum malaria on the Cambodian-Thai border." Am J Trop Med Hyg **76**(4): 641-7.
- Ashley, E., S. Krudsood, et al. (2004). "Randomized, controlled dose-optimization studies of dihydroartemisinin-piperaquine for the treatment of uncomplicated multidrug-resistant falciparum malaria in Thailand." J Infect Dis **190**(10): 1773-82.
- Attaran, A. and R. Maharaj (2000). "Ethical debate: doctoring malaria, badly: the global campaign to ban DDT." BMJ **321**(7273): 1403-5.
- Bach, O., M. Baier, et al. (2005). "Falciparum malaria after splenectomy: a prospective controlled study of 33 previously splenectomized Malawian adults." Trans R Soc Trop Med Hyg **99**(11): 861-7.
- Bahl, A., B. Brunk, et al. (2003). "PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data." Nucleic Acids Res **31**(1): 212-5.
- Bailey, T. L. and C. Elkan (1995). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol **3**: 21-9.
- Baruch, D. I., B. L. Pasloske, et al. (1995). "Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes." Cell **82**(1): 77-87.
- Baum, J., A. G. Maier, et al. (2005). "Invasion by *P. falciparum* merozoites suggests a hierarchy of molecular interactions." PLoS Pathog **1**(4): e37.
- Beier, J. and J. Vanderberg (1998). Chapter 4, Malaria: Parasite Biology, Pathogenesis, and Protection.
- Ben-Dor, A., R. Shamir, et al. (1999). "Clustering gene expression patterns." J Comput Biol **6**(3-4): 281-97.
- Bergman, M., G. Del Prete, et al. (2006). "Helicobacter pylori phase variation, immune modulation and gastric autoimmunity." Nat Rev Microbiol **4**(2): 151-9.

- Bettencourt, L. M., J. Lobo, et al. (2007). "Growth, innovation, scaling, and the pace of life in cities." Proc Natl Acad Sci U S A **104**(17): 7301-6.
- Billker, O., S. Dechamps, et al. (2004). "Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite." Cell **117**(4): 503-14.
- Billker, O., V. Lindo, et al. (1998). "Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito." Nature **392**(6673): 289-92.
- Bozdech, Z., M. Llinás, et al. (2003). "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*." PLoS Biol **1**(1): E5.
- Bozdech, Z., M. Llinas, et al. (2003). "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*." PLoS Biol **1**(1): E5.
- Bozdech, Z., J. Zhu, et al. (2003). "Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray." Genome Biol **4**(2): R9.
- Brandt, B., S. Meyer-Staeckling, et al. (2006). "Mechanisms of egfr gene transcription modulation: relationship to cancer risk and therapy response." Clin Cancer Res **12**(24): 7252-60.
- Bray, R. S. and P. C. Garnham (1982). "The life-cycle of primate malaria parasites." Br Med Bull **38**(2): 117-22.
- Brentano, F., O. Schorr, et al. (2005). "RNA released from necrotic synovial fluid cells activates rheumatoid arthritis synovial fibroblasts via Toll-like receptor 3." Arthritis Rheum **52**(9): 2656-65.
- Brewer, T. G., J. O. Peggins, et al. (1994). "Neurotoxicity in animals due to arteether and artemether." Trans R Soc Trop Med Hyg **88 Suppl 1**: S33-6.
- Bruce-Chwatt, L. (1982). "Qinghaosu: a new antimalarial." British medical journal (Clinical research ed) **284**(6318): 767-8.
- Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-71.
- Calderwood, M. S., L. Gannoun-Zaki, et al. (2003). "*Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron." J Biol Chem **278**(36): 34125-32.
- Callebaut, I., K. Prat, et al. (2005). "Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes." BMC Genomics **6**: 100.

- Carson, R. (1962). Silent Spring. Boston, Houghton Mifflin.
- Cho, R., M. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell **2**(1): 65-73.
- Cho, R., M. Huang, et al. (2001). "Transcriptional regulation and function during the human cell cycle." Nat Genet **27**(1): 48-54.
- Chookajorn, T., R. Dzikowski, et al. (2007). "Epigenetic memory at malaria virulence genes." Proc Natl Acad Sci U S A **104**(3): 899-902.
- Chow, C. S. and D. F. Wirth (2003). "Linker scanning mutagenesis of the *Plasmodium gallinaceum* sexual stage specific gene pgs28 reveals a novel downstream cis-control element." Mol Biochem Parasitol **129**(2): 199-208.
- Churchill, G. (2002). "Fundamentals of experimental design for cDNA microarrays." Nat Genet **32** **Suppl**: 490-5.
- Clyde, D. E., M. S. Corado, et al. (2003). "A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*." Nature **426**(6968): 849-53.
- Coulson, R., N. Hall, et al. (2004). "Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*." Genome Res **14**(8): 1548-54.
- Coulson, R. and C. Ouzounis (2003). "The phylogenetic diversity of eukaryotic transcription." Nucleic Acids Res **31**(2): 653-60.
- Crabb, B. S. and A. F. Cowman (1996). "Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*." Proc Natl Acad Sci U S A **93**(14): 7289-94.
- D'Alessandro, U. and H. Buttiëns (2001). "History and importance of antimalarial drug resistance." Trop Med Int Health **6**(11): 845-8.
- Dahl, E., J. Shock, et al. (2006). "Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*." Antimicrob Agents Chemother **50**(9): 3124-31.
- Davidson, I. (2003). "The genetics of TBP and TBP-related factors." Trends Biochem Sci **28**(7): 391-8.
- Davies, S. J., J. L. Grogan, et al. (2001). "Modulation of blood fluke development in the liver by hepatic CD4+ lymphocytes." Science **294**(5545): 1358-61.
- Davis, T., G. Edwards, et al. (1997). "Artesunate and cerebellar dysfunction in *falciparum* malaria." N Engl J Med **337**(11): 792; author reply 793.
- de Boer, E., P. Rodriguez, et al. (2003). "Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice." Proc Natl Acad Sci USA **100**(13): 7480-5.

- Deitsch, K. W., M. S. Calderwood, et al. (2001). "Malaria. Cooperative silencing elements in var genes." Nature **412**(6850): 875-6.
- Division of Parasitic Diseases, N. C. f. Z., Vector-Borne, and Enteric Diseases (2007).
- Doerig, C. and D. Chakrabarti (2004). Chapter 10, Malaria Parasites Genomes and Molecular Biology. Great Britain, Caister Academic Press.
- Doerig, C., J. Endicott, et al. (2002). "Cyclin-dependent kinase homologues of Plasmodium falciparum." Int J Parasitol **32**(13): 1575-85.
- Donahue, T. F., R. S. Daves, et al. (1983). "A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast." Cell **32**(1): 89-98.
- Duraisingh, M., T. Voss, et al. (2005). "Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum." Cell **121**(1): 13-24.
- Durbin, R., Eddy, S., Krogh A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, Cambridge University Press.
- Eckstein-Ludwig, U., R. J. Webb, et al. (2003). "Artemisinins target the SERCA of Plasmodium falciparum." Nature **424**(6951): 957-61.
- Eichner, M., H. H. Diebner, et al. (2001). "Genesis, sequestration and survival of Plasmodium falciparum gametocytes: parameter estimates from fitting a model to malariatherapy data." Trans R Soc Trop Med Hyg **95**(5): 497-501.
- Elowitz, M. and S. Leibler (2000). "A synthetic oscillatory network of transcriptional regulators." Nature **403**(6767): 335-8.
- Falick, A. (2006).
- Feachem, R. G. and O. J. Sabot (2007). "Global malaria control in the 21st century: a historic but fleeting opportunity." Jama **297**(20): 2281-4.
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. U. o. W. Distributed by the author. Department of Genome Sciences, Seattle.
- Fidock, D. A., T. Nomura, et al. (2000). "Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance." Mol Cell **6**(4): 861-71.
- Florens, L., M. P. Washburn, et al. (2002). "A proteomic view of the Plasmodium falciparum life cycle." Nature **419**(6906): 520-6.
- Freitas-Junior, L., E. Bottius, et al. (2000). "Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum." Nature **407**(6807): 1018-22.

- Freitas-Junior, L., R. Hernandez-Rivas, et al. (2005). "Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites." Cell **121**(1): 25-36.
- Fried, M. and D. M. Crothers (1981). "Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis." Nucleic Acids Res **9**(23): 6505-25.
- Fried, M. and P. E. Duffy (1996). "Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta." Science **272**(5267): 1502-4.
- Gadgil, H., S. Oak, et al. (2001). "Affinity purification of DNA-binding proteins." J Biochem Biophys Methods **49**(1-3): 607-24.
- Gardner, M., S. Shallom, et al. (2002). "Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14." Nature **419**(6906): 531-4.
- Gardner, M. J., N. Hall, et al. (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." Nature **419**(6906): 498-511.
- Garner, M. M. and A. Revzin (1981). "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system." Nucleic Acids Res **9**(13): 3047-60.
- Goldberg, D. E. (2005). "Hemoglobin degradation." Curr Top Microbiol Immunol **295**: 275-91.
- Greenwood, D. (1995). "Conflicts of interest: the genesis of synthetic antimalarial agents in peace and war." J Antimicrob Chemother **36**(5): 857-72.
- Guillemin, J. (2002). "Choosing scientific patrimony: Sir Ronald Ross, Alphonse Laveran, and the mosquito-vector hypothesis for malaria." Journal of the history of medicine and allied sciences **57**(4): 385-409.
- Hall, N., A. Pain, et al. (2002). "Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13." Nature **419**(6906): 527-31.
- Hamada, H., M. Seidman, et al. (1984). "Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence." Mol Cell Biol **4**(12): 2622-30.
- Hay, S., C. Guerra, et al. (2005). "Urbanization, malaria transmission and disease burden in Africa." Nat Rev Microbiol **3**(1): 81-90.
- Hay, S., C. Guerra, et al. (2004). "The global distribution and population at risk of malaria: past, present, and future." The Lancet infectious diseases **4**(6): 327-36.

- Hayward, R., J. Derisi, et al. (2000). "Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria." Mol Microbiol **35**(1): 6-14.
- Hayward, R., K. Saliba, et al. (2006). "The pH of the digestive vacuole of *Plasmodium falciparum* is not associated with chloroquine resistance." J Cell Sci **119**(Pt 6): 1016-25.
- Holmes, M. C. and R. Tjian (2000). "Promoter-selective properties of the TBP-related factor TRF1." Science **288**(5467): 867-70.
- Holtzendorff, J., D. Hung, et al. (2004). "Oscillating global regulators control the genetic circuit driving a bacterial cell cycle." Science **304**(5673): 983-7.
- Honigsbaum, M. (2002). The fever trail : in search of the cure for malaria. New York, Farrar Straus and Giroux.
- Horrocks, P. and M. Lanzer (1999). "Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in *Plasmodium falciparum*." Mol Biochem Parasitol **99**(1): 77-87.
- Horrocks, P., R. Pinches, et al. (2004). "Variable var transition rates underlie antigenic variation in malaria." Proc Natl Acad Sci U S A **101**(30): 11129-34.
- Hyman, R., E. Fung, et al. (2002). "Sequence of *Plasmodium falciparum* chromosome 12." Nature **419**(6906): 534-7.
- Jomaa, H., J. Wiesner, et al. (1999). "Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs." Science **285**(5433): 1573-6.
- Kadonaga, J. (1991). "Purification of sequence-specific binding proteins by DNA affinity chromatography." Meth Enzymol **208**: 10-23.
- Kitchen, L. W., D. W. Vaughn, et al. (2006). "Role of US military research programs in the development of US Food and Drug Administration--approved antimalarial drugs." Clin Infect Dis **43**(1): 67-71.
- Klausner, M. (1996). MedWatch Safety Summaries - Hismanal.
- Klayman, D. (1985). "Qinghaosu (artemisinin): an antimalarial drug from China." Science **228**(4703): 1049-55.
- Kocken, C., A. van der Wel, et al. (1998). "Precise timing of expression of a *Plasmodium falciparum*-derived transgene in *Plasmodium berghei* is a critical determinant of subsequent subcellular localization." J Biol Chem **273**(24): 15119-24.
- Kyburz, D., F. Brentano, et al. (2006). "Mode of action of hydroxychloroquine in RA-evidence of an inhibitory effect on toll-like receptor signaling." Nature clinical practice Rheumatology **2**(9): 458-9.

- Kyle, R. and M. Shampe (1974). "Discoverers of quinine." JAMA **229**(4): 462.
- LaCount, D. J., M. Vignali, et al. (2005). "A protein interaction network of the malaria parasite *Plasmodium falciparum*." Nature **438**(7064): 103-7.
- Lakshmanan, V., P. G. Bray, et al. (2005). "A critical role for PfCRT K76T in *Plasmodium falciparum* verapamil-reversible chloroquine resistance." Embo J **24**(13): 2294-305.
- Landau, I. and P. Gautret (1998). Chapter 28, Malaria: Parasite Biology, Pathogenesis, and Protection, ASM Press.
- Lanzer, M., D. de Bruin, et al. (1992). "A sequence element associated with the *Plasmodium falciparum* KAHRP gene is the site of developmentally regulated protein-DNA interactions." Nucleic Acids Res **20**(12): 3051-6.
- Le Roch, K., J. Johnson, et al. (2004). "Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle." Genome Res **14**(11): 2308-18.
- Le Roch, K. G., Y. Zhou, et al. (2003). "Discovery of gene function by expression profiling of the malaria parasite life cycle." Science **301**(5639): 1503-8.
- Lillehaug, J. R. and K. Kleppe (1975). "Kinetics and specificity of T4 polynucleotide kinase." Biochemistry **14**(6): 1221-5.
- MacPherson, G. G., M. J. Warrell, et al. (1985). "Human cerebral malaria. A quantitative ultrastructural analysis of parasitized erythrocyte sequestration." Am J Pathol **119**(3): 385-401.
- Madrid, P. B., J. Sherrill, et al. (2005). "Synthesis of ring-substituted 4-aminoquinolines and evaluation of their antimalarial activities." Bioorg Med Chem Lett **15**(4): 1015-8.
- Mandavilli, A. (2006). "Health agency backs use of DDT against malaria." Nature **443**(7109): 250-1.
- Martin, R. E., R. I. Henry, et al. (2005). "The 'permeome' of the malaria parasite: an overview of the membrane transport proteins of *Plasmodium falciparum*." Genome Biol **6**(3): R26.
- McDonald, M. J. and M. Rosbash (2001). "Microarray analysis and organization of circadian gene expression in *Drosophila*." Cell **107**(5): 567-78.
- Merckx, A., K. Le Roch, et al. (2003). "Identification and initial characterization of three novel cyclin-related proteins of the human malaria parasite *Plasmodium falciparum*." J Biol Chem **278**(41): 39839-50.

- Miesfeld, R., P. Godowski, et al. (1987). "Glucocorticoid receptor mutants that define a small region sufficient for enhancer activation." Science **236**(4800): 423-7.
- Milgram, A. (2004).
- Miller, B., E. McDearmon, et al. (2007). "Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation." Proc Natl Acad Sci USA **104**(9): 3342-7.
- Miller, L. G. and C. B. Panosian (1997). "Ataxia and slurred speech after artesunate treatment for falciparum malaria." N Engl J Med **336**(18): 1328.
- Miller, R., S. Ikram, et al. (1994). "Diagnosis of Plasmodium falciparum infections in mummies using the rapid manual ParaSight-F test." Trans R Soc Trop Med Hyg **88**(1): 31-2.
- Müller, S. and B. Kappes (2007). "Vitamin and cofactor biosynthesis pathways in Plasmodium and other apicomplexan parasites." Trends Parasitol **23**(3): 112-21.
- Murphy, J. C., J. A. Wibbenmeyer, et al. (1999). "Purification of plasmid DNA using selective precipitation by compaction agents." Nat Biotechnol **17**(8): 822-3.
- Muscatello, U. (2007). "Golgi's contribution to medicine." Brain Res Rev.
- Mutabingwa, T. (2005). "Artemisinin-based combination therapies (ACTs): best hope for malaria treatment but inaccessible to the needy!" Acta Trop **95**(3): 305-15.
- Myrick, A., O. Sarr, et al. (2005). "Analysis of the genetic diversity of the Plasmodium falciparum multidrug resistance gene 5' upstream region." Am J Trop Med Hyg **72**(2): 182-8.
- Nakajima, M., K. Imai, et al. (2005). "Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro." Science **308**(5720): 414-5.
- National Center for Infectious Diseases, D. o. P. D. (2004). History | CDC Malaria.
- Newton, P. N., M. D. Green, et al. (2007). "Counterfeit artemisinin derivatives and Africa: update from authors." PLoS Med **4**(3): e139.
- Painter, H., J. Morrissey, et al. (2007). "Specific role of mitochondrial electron transport in blood-stage Plasmodium falciparum." Nature **446**(7131): 88-91.
- Peters, J., E. Fowler, et al. (2002). "High diversity and rapid changeover of expressed var genes during the acute phase of Plasmodium falciparum infections in human volunteers." Proc Natl Acad Sci U S A **99**(16): 10689-94.

- Polson, H. E. and M. J. Blackman (2005). "A role for poly(dA)poly(dT) tracts in directing activity of the *Plasmodium falciparum* calmodulin gene promoter." Mol Biochem Parasitol **141**(2): 179-89.
- Porter, M. E. (2002). "Positive and negative effects of deletions and mutations within the 5' flanking sequences of *Plasmodium falciparum* DNA polymerase delta." Mol Biochem Parasitol **122**(1): 9-19.
- Pouvelle, B., P. A. Buffet, et al. (2000). "Cytoadhesion of *Plasmodium falciparum* ring-stage-infected erythrocytes." Nat Med **6**(11): 1264-8.
- Price, R., A. Uhlemann, et al. (2004). "Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number." Lancet **364**(9432): 438-47.
- Prudencio, M., A. Rodriguez, et al. (2006). "The silent path to thousands of merozoites: the *Plasmodium* liver stage." Nat Rev Microbiol **4**(11): 849-56.
- Ptashne, M. and A. Gann (2002). Genes & signals. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
- Raisner, R. M., P. D. Hartley, et al. (2005). "Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin." Cell **123**(2): 233-48.
- Rando, O. J. and R. Wu (2006).
- Rangarajan, R., A. Bei, et al. (2005). "A mitogen-activated protein kinase regulates male gametogenesis and transmission of the malaria parasite *Plasmodium berghei*." EMBO Rep **6**(5): 464-9.
- Rathod, P. (2007).
- Rebar, E. J., Y. Huang, et al. (2002). "Induction of angiogenesis in a mouse model using engineered transcription factors." Nat Med **8**(12): 1427-32.
- Reed, M., K. Saliba, et al. (2000). "Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum*." Nature **403**(6772): 906-9.
- Ro, D. K., E. M. Paradise, et al. (2006). "Production of the antimalarial drug precursor artemisinic acid in engineered yeast." Nature **440**(7086): 940-3.
- Rogers, N., B. Hall, et al. (2000). "A model for sequestration of the transmission stages of *Plasmodium falciparum*: adhesion of gametocyte-infected erythrocytes to human bone marrow cells." Infect Immun **68**(6): 3455-62.
- Rosenfeld, P. and T. Kelly (1986). "Purification of nuclear factor I by DNA recognition site affinity chromatography." J Biol Chem **261**(3): 1398-408.

- Rowe, J. A., J. M. Moulds, et al. (1997). "P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1." Nature **388**(6639): 292-5.
- Rug, M., M. Wickham, et al. (2004). "Correct promoter control is needed for trafficking of the ring-infected erythrocyte surface antigen to the host cytosol in transfected malaria parasites." Infect Immun **72**(10): 6095-105.
- Rustici, G., J. Mata, et al. (2004). "Periodic gene expression program of the fission yeast cell cycle." Nat Genet **36**(8): 809-17.
- Ruvalcaba-Salazar, O., M. del Carmen Ramírez-Estudillo, et al. (2005). "Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions." Mol Biochem Parasitol **140**(2): 183-96.
- Ruvalcaba-Salazar, O. K., M. del Carmen Ramirez-Estudillo, et al. (2005). "Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions." Mol Biochem Parasitol **140**(2): 183-96.
- Saksouk, N., M. M. Bhatti, et al. (2005). "Histone-modifying complexes regulate gene expression pertinent to the differentiation of the protozoan parasite Toxoplasma gondii." Mol Cell Biol **25**(23): 10301-14.
- Sallares, R., A. Bouwman, et al. (2004). "The spread of malaria to Southern Europe in antiquity: new approaches to old problems." Medical history **48**(3): 311-28.
- Sallicandro, P., M. Paglia, et al. (2000). "Repetitive sequences upstream of the pfg27/25 gene determine polymorphism in laboratory and natural lines of Plasmodium falciparum." Mol Biochem Parasitol **110**(2): 247-57.
- Scheidereit, C., A. Heguy, et al. (1987). "Identification and purification of a human lymphoid-specific octamer-binding protein (OTF-2) that activates transcription of an immunoglobulin promoter in vitro." Cell **51**(5): 783-93.
- Sherman, I. (1998). Malaria: Parasite Biology, Pathogenesis, and Protection.
- Shock, J. L., K. F. Fischer, et al. (2007). "Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle." Genome Biol **8**(7): R134.
- Smith, J. D., C. E. Chitnis, et al. (1995). "Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic

- and cytoadherent phenotypes of infected erythrocytes." Cell **82**(1): 101-10.
- Smith, J. D., A. G. Craig, et al. (2000). "Identification of a Plasmodium falciparum intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria." Proc Natl Acad Sci U S A **97**(4): 1766-71.
- Smith, T. G., P. Lourenco, et al. (2000). "Commitment to sexual differentiation in the human malaria parasite, Plasmodium falciparum." Parasitology **121** (Pt 2): 127-33.
- Snow, R., C. Guerra, et al. (2005). "The global distribution of clinical episodes of Plasmodium falciparum malaria." Nature **434**(7030): 214-7.
- Spellman, P., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.
- Stormo, G. D. and G. W. Hartzell, 3rd (1989). "Identifying protein-binding sites from unaligned DNA fragments." Proc Natl Acad Sci U S A **86**(4): 1183-7.
- Struhl, K. (1982). "Regulatory sites for his3 gene expression in yeast." Nature **300**(5889): 285-6.
- Sturm, A., R. Amino, et al. (2006). "Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids." Science **313**(5791): 1287-90.
- Su, X., L. Kirkman, et al. (1997). "Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa." Cell **91**(5): 593-603.
- Su, X. Z., V. M. Heatwole, et al. (1995). "The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes." Cell **82**(1): 89-100.
- Sullivan, D., I. Gluzman, et al. (1996). "Plasmodium hemozoin formation mediated by histidine-rich proteins." Science **271**(5246): 219-22.
- Sullivan, D., I. Gluzman, et al. (1996). "On the molecular mechanism of chloroquine's antimalarial action." Proc Natl Acad Sci USA **93**(21): 11865-70.
- Sultan, A. A., V. Thathy, et al. (1997). "TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites." Cell **90**(3): 511-22.
- Suplick, K., J. Morrissey, et al. (1990). "Complex transcription from the extrachromosomal DNA encoding mitochondrial functions of Plasmodium yoelii." Mol Cell Biol **10**(12): 6381-8.

- Surolia, N. and A. Surolia (2001). "Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum." Nat Med **7**(2): 167-73.
- Teja-Isavadharm, P., G. Watt, et al. (2001). "Comparative pharmacokinetics and effect kinetics of orally administered artesunate in healthy volunteers and patients with uncomplicated falciparum malaria." Am J Trop Med Hyg **65**(6): 717-21.
- Towie, N. (2006). "Malaria breakthrough raises spectre of drug resistance." Nature **440**(7086): 852-3.
- Trager, W. and J. B. Jensen (1976). "Human malaria parasites in continuous culture." Science **193**(4254): 673-5.
- Triglia, T., J. Healer, et al. (2000). "Apical membrane antigen 1 plays a central role in erythrocyte invasion by Plasmodium species." Mol Microbiol **38**(4): 706-18.
- Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-5.
- Uhlemann, A., A. Cameron, et al. (2005). "A single amino acid residue can determine the sensitivity of SERCAs to artemisinins." Nat Struct Mol Biol **12**(7): 628-9.
- Van de Perre, P. and J. P. Dedet (2004). "Vaccine efficacy: winning a battle (not war) against malaria." Lancet **364**(9443): 1380-3.
- van Helden, J., B. Andre, et al. (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." J Mol Biol **281**(5): 827-42.
- van Noort, V. and M. Huynen (2006). "Combinatorial gene regulation in Plasmodium falciparum." Trends Genet **22**(2): 73-8.
- Vazquez-Macias, A., P. Martinez-Cruz, et al. (2002). "A distinct 5' flanking var gene region regulates Plasmodium falciparum variant erythrocyte surface antigen expression in placental malaria." Mol Microbiol **45**(1): 155-67.
- Voss, T., T. Mini, et al. (2002). "Plasmodium falciparum possesses a cell cycle-regulated short type replication protein A large subunit encoded by an unusual transcript." J Biol Chem **277**(20): 17493-501.
- Voss, T. S., J. Healer, et al. (2006). "A var gene promoter controls allelic exclusion of virulence genes in Plasmodium falciparum malaria." Nature **439**(7079): 1004-8.
- Voss, T. S., M. Kaestli, et al. (2003). "Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters." Mol Microbiol **48**(6): 1593-607.


- Voss, T. S., J. K. Thompson, et al. (2000). "Genomic distribution and functional characterisation of two distinct and conserved Plasmodium falciparum var gene 5' flanking sequences." Mol Biochem Parasitol **107**(1): 103-15.
- Wang, G. and G. Semenza (1995). "Purification and characterization of hypoxia-inducible factor 1." J Biol Chem **270**(3): 1230-7.
- Wang, J., K. Nishiyama, et al. (1987). "Purification of an octamer sequence (ATGCAAAT)-binding protein from human B cells." Nucleic Acids Res **15**(24): 10105-16.
- Wang, W., J. Cherry, et al. (2005). "Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation." Proc Natl Acad Sci USA **102**(6): 1998-2003.
- Wang, W., J. M. Cherry, et al. (2005). "Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation." Proc Natl Acad Sci U S A **102**(6): 1998-2003.
- Watanabe, J., Y. Suzuki, et al. (2004). "Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, Plasmodium species." Nucleic Acids Res **32**(Database issue): D334-8.
- Waters, A. and C. Janse (2004). Malaria Parasites Genomes and Molecular Biology.
- Weisman, J. L. (2007).
- White, N. J. (2004). "Antimalarial drug resistance." J Clin Invest **113**(8): 1084-92.
- WHO (2006). Guidelines for treatment of malaria.
- Woodrow, C. J., R. K. Haynes, et al. (2005). "Artemisinins." Postgrad Med J **81**(952): 71-8.
- Yayon, A., Z. Cabantchik, et al. (1985). "Susceptibility of human malaria parasites to chloroquine is pH dependent." Proc Natl Acad Sci USA **82**(9): 2784-8.
- YOUNG, M. and D. MOORE (1961). "Chloroquine resistance in Plasmodium falciparum." Am J Trop Med Hyg **10**: 317-20.
- Young, M. D. and D. V. Moore (1961). "Chloroquine resistance in Plasmodium falciparum." Am J Trop Med Hyg **10**: 317-20.
- Zaret, K. and K. Yamamoto (1984). "Reversible and persistent changes in chromatin structure accompany activation of a glucocorticoid-dependent enhancer element." Cell **38**(1): 29-38.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

8/17/2007
Date