STRATEGIC GENOMICS: FROM DEVELOPING MALARIA DIAGNOSTICS TO DEMYSTIFYING GENE REGULATION IN APICOMPLEXAN PARASITES

by

JENNA OBERSTALLER

(Under the Direction of Jessica C. Kissinger)

ABSTRACT

The phylum Apicomplexa consists of ~5000 species of parasitic protists. Many of these parasites are of great social and economic importance, including those responsible for malaria (genus *Plasmodium*), cryptosporidiosis (genus *Cryptosporidium*) toxoplasmosis (*Toxoplasma gondii*) and a number of other significant human and veterinary diseases. As such, many apicomplexan genomes have been sequenced to date. This dissertation describes our work to use these genomes to develop improved malaria diagnostic assays, as well as to study transcriptional regulatory phenomena in an organism with few experimental tools, *Cryptosporidium parvum*. We have shown that malaria diagnostic assays developed to conserved, repetitive sequences in several human-infecting malaria parasite genomes are species-specific and more sensitive than existing molecular diagnostics. We also present the first comprehensive study of a major transcription factor family in *Cryptosporidium parvum*, the ApiAP2s, and present evidence that *C. parvum* may not be as reliant on ApiAP2 regulation as previous research has indicated for other apicomplexans.

INDEX WORDS: Apicomplexa, malaria, diagnostics, transcription, *Plasmodium*, *P. vivax*, *P. falciparum*, *P. knowlesi*, *Cryptosporidium parvum*, AP2, E2F, G-box

STRATEGIC GENOMICS: FROM DEVELOPING MALARIA DIAGNOSTICS TO DEMYSTIFYING GENE REGULATION IN APICOMPLEXAN PARASITES

by

JENNA OBERSTALLER

BS, University of South Carolina, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

STRATEGIC GENOMICS: FROM DEVELOPING MALARIA DIAGNOSTICS TO DEMYSTIFYING GENE REGULATION IN APICOMPLEXAN PARASITES

by

JENNA OBERSTALLER

Major Professor:	Jessica C. Kissinger
Committee:	Haini Cai Richard Meagher Douglas Menke Boris Striepen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

DEDICATION

To my mother Rhonda, my father Herbert and the memory of my sister Brandy, for always knowing what I am capable of, and for reminding me when I forgot.

ACKNOWLEDGEMENTS

I would like to thank my committee members, Jessie, Rich, Boris, Doug and Haini for their suggestions over the years that have improved the quality of my work. They have always been focused on helping me be the best scientist I could be, and committee meetings were always comfortable, encouraging and productive. They have been a pleasure to learn from.

Jessie has been a particularly great mentor. She has been so supportive of my professional development, sending me to conferences, pushing for me to have the opportunity to do fieldwork, forwarding me any public health career opportunities that come her way, spending hours with me improving talks to ensure I don't embarrass either one of us. She has allowed me to think independently, intervening only when I need it, pushing me always to think bigger and to not get lost in the details.

So many people in the CTEGD and the Genetics Department have been key to my success at UGA, and I know I'm not going to be able to list them all. But thank you in particular to the past and present members of the Kissinger lab, the Striepen lab, and my Genetics cohort for help with techniques, equipment, and discussions scientific and otherwise that made Science a fun place to be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and literature review

1.1 The Apicomplexa

The phylum Apicomplexa consists of ~5000 species of entirely parasitic protists [1]. Many of these parasites are of great social and economic importance, including those responsible for malaria (genus *Plasmodium*), cryptosporidiosis (genus *Cryptosporidium*) toxoplasmosis (*Toxoplasma gondii*) and a number of other significant human and veterinary diseases. Many apicomplexan genomes have been sequenced to date due to these parasites' significance for public health. Apicomplexans have small genomes (ranging from ~8.5Mb in *Theileria* to ~63Mb in *Toxoplasma*; Figure 1.1) characterized by a very distinct evolutionary history involving multiple and often-ancient gene-transfer events from distantly related species, as well as from their own organellar genomes [2,3]. Most apicomplexans have retained a relict non-photosynthetic chloroplast-like organelle, called the apicoplast, derived from the ancient secondary endosymbiosis of an alga [4]. The phylum Apicomplexa is thought to have diverged anywhere from 350 – 900 million years ago [5,6], which is a more ancient divergence than even that within vertebrates (~550 million years; [7]). Given the grand timescale of divergences between apicomplexans, it is perhaps not surprising that apicomplexans affect a wide range of different hosts and have wide variability within their lifecycles, though apicomplexans in general are characterized by complex lifecycles often involving multiple hosts. The *Plasmodium falciparum*, *Toxoplasma gondii*, and *Cryptosporidium parvum* life cycles will be discussed further as these are the primary apicomplexans discussed in this dissertation.

Plasmodium falciparum life cycle

The haemosporidian parasite *P. falciparum* requires both mosquito and human host to complete its lifecycle, where the parasite undergoes sexual and asexual replication, respectively (Figure 1.2). A female *Anopheles* mosquito vector bites the human host, releasing sporozoites into the bloodstream. These sporozoites travel to the liver, where they undergo several rounds of asexual replication, eventually bursting out into the blood stream in the “merozoite” form. Merozoites invade red blood cells, and again undergo a round of asexual replication, forming a multi-nucleate cell (schizont) that ultimately divides to form several more merozoites. These merozoites lyse out of red blood cells and undergo the red blood cell invasion process over and over again. Parasite numbers are generally too low to produce clinical symptoms in human hosts until the parasites have completed the liver stage and begin replicating and producing higher and higher numbers via invasion of and lysis from red blood cells (known as the intra-erythrocytic cycle, or the blood stage); the synchronized lysis of merozoites from red blood cells produces the cycles of fever that often characterize malaria. Not all merozoites go on to form schizonts; some undergo sexual differentiation into female and male gametes, which are taken up by the female *Anopheles* mosquito to undergo sexual replication, and thus begin the cycle anew.

Toxoplasma gondii life cycle

Though the coccidian parasite *T. gondii* can (and does) parasitize a wide range of vertebrate hosts, the definitive hosts of the parasite (where sexual replication occurs) are members of the cat family (Figure 1.3). Infection in humans generally occurs either by ingesting viable parasite tissue cysts in raw or undercooked meat or by ingesting oocysts shed in the feces of a cat [8].

T. gondii has a complex life cycle consisting of two distinct developmental stages—the tachyzoite (fast-growing) stage, and the bradyzoite (slow-growing) stage. *T. gondii* exists in the host in the tachyzoite stage during the initial acute infection, often causing flu-like symptoms in otherwise-healthy people. After acute infection subsides in a few days to a few months, tachyzoites travel to the muscles and brain where they make the developmental switch to the bradyzoite stage. *T. gondii* bradyzoites continue to reside indefinitely within tissues cysts in the host muscles and brain—once infected, a person will remain infected for the rest of their lives, as no treatments currently exist to combat *T. gondii* once it has settled into its bradyzoite stage. In immuno-competent individuals, latent *T. gondii* infection is asymptomatic. However, in people with immunodeficiencies such as AIDS, rupture of cysts results in disease reactivation, often leading to encephalitis, which can be deadly [8,9]. Acute or reactivated infection is also particularly devastating in pregnant women, as the pregnancies of these infected women often result in miscarriage if acquired early in pregnancy and severe birth defects and ocular infections later in pregnancy [10]. It is estimated that 20 – 40% of people in the US are currently infected, whereas other countries (i.e., El Salvador) show a prevalence of as high as 75% [9,10].

***Cryptosporidium parvum* life cycle**

C. parvum has historically been classified as coccidian (like *T. gondii*), because the life cycle and morphological stages of *Cryptosporidium* resemble that of other coccidia and it infects the gut. However, phylogenies constructed from several evolutionarily conserved genes suggest that *Cryptosporidium* is more closely related to the gregarines than to coccidia [11,12], suggesting that *Cryptosporidium* is actually one of the most basal-branching apicomplexans. The first case of human *Cryptosporidium* infection was reported in 1976 [13], and only seven additional cases were documented before 1982 [14]. Since then, the number of cases identified has increased dramatically, largely due to the recognition of a life-threatening form of infection in patients with AIDS [15]. In addition, seroprevalence rates of 25-35% in the United States indicate that infection with *Cryptosporidum* is very common among healthy persons [16].

C. parvum has a complex, obligate-intracellular life cycle involving both asexual and sexual developmental stages (Figure 1.4). As with the coccidia, transmission of cryptosporidosis happens through the fecal-oral route where an infection is initiated by the ingestion of oocysts. Upon ingestion, oocysts release sporozoites, which primarily infect the microvillus border of the intestinal epithelium, and to lesser extent extraintestinal epithelia, causing acute gastrointestinal disease in a wide range of mammalian hosts. The parasites undergo merogeny (asexual replication), and then sexual multiplication producing microgamonts (male) and macrogamonts (female). Upon fertilization of the macrogamonts by the microgametes, both thick-walled oocysts (which are excreted from the host) and thin-walled oocysts (which remain in the host) are produced. These oocysts sporulate in the host, leaving the thick-walled oocysts viable to infect the next host and the thin-walled oocysts ready to start another round of autoinfection within the existing host. Unlike in the case of *P. falciparum* and *T. gondii*, *C. parvum* can complete its entire lifecycle in a single host, and it undergoes both sexual and asexual replication within its host.

Apicomplexa in the tree of life

Protists are not a monophyletic group, and they have little in common beyond being unicellular. They comprise a tremendous diversity of unicellular eukaryotes with as-yet unresolved relationships within the eukaryotic tree of life [17,18]. Apicomplexans have classically been considered part of the kingdom Chromalveolata, which also comprises sister phyla dinoflagellates and ciliates, as well as the more distantly-related stramenopiles, cryptophytes, and haptophytes (Figure 1.5), though more recent phylogenies have questioned the monophyly of chromalveolates [17]. Throughout this dissertation I will be investigating *Plasmodium* genomes (Chapters 2 and 6) as well as evolution of apicomplexan transcriptional regulatory mechanisms (Chapters 3 and 4), and knowledge of the relationships of select apicomplexans to each other as well as the placement of the phylum in the broader context of the tree of life will be important to understanding my arguments.

1.2 Organization of this dissertation

This dissertation is organized into six chapters. In the following sections of Chapter 1, I discuss relevant background information for this dissertation in two parts. In section 1.3, I review relevant literature pertaining to malaria diagnostic tools. In sections 1.4 and 1.5, I review what is known about apicomplexan gene regulation, and discuss how computational approaches can be used to study genome evolution in both a malaria-diagnostic and a transcriptional regulation context. In Chapter 2, I present our efforts to improve PCR-based molecular diagnostics for the two most widespread and deadly human malaria parasites, *Plasmodium falciparum* and *P. vivax*. We were able to use the available genome sequences to develop more sensitive, multiplex assays for the specific detection of both parasites. Chapters 3 and 4 describe studies undertaken to elucidate transcriptional regulatory mechanisms in *C. parvum*. In Chapter 3, I present a study where we use the genome sequence and the transcriptome of *C. parvum* to predict putative *cis*-regulatory elements upstream of co-expressed genes. In Chapter 4, I present our efforts to define the ApiAP2 transcription factor regulatory network in *C. parvum* by experimentally determining DNA binding motifs for this entire protein family and use of the transcriptome and bioinformatics tools to predict putative regulatory targets. In Chapter 5, I discuss and present ideas for future directions of the work described in this dissertation. Chapter 6 describes efforts to improve PCR-based diagnostics for the zoonotic malaria parasite *Plasmodium knowlesi* using the methodology developed in Chapter 2. Each manuscript in this dissertation is the result of a collaborative team effort. Section 6.2 describes the respective contributions of the authors associated with each chapter.

1.3 Malaria diagnostics

Malaria is one of the most devastating parasitic infections of humans worldwide. The disease in humans is caused by any of five species of the apicomplexan parasite genus *Plasmodium*: *P. falciparum, P. vivax, P. knowlesi, P. ovale* and *P. malariae*. In several parts of the world (in particular Africa and southeast Asia), subsets of these organisms have overlapping species ranges, and individuals in these areas may be singly, or multiply, infected with any of the local species. Initial symptoms of the disease caused by any one of these malaria parasites are largely indistinguishable (from themselves as well as other endemic diseases), though the outcomes of infection vary greatly depending upon the causative parasite(s) [19]. Correct diagnosis of the disease and the responsible parasite(s) early during infection is vital to inform the course of treatment and to improve the outcome of treatment for the affected individual [20].

Widespread methods for diagnosis include microscopy, parasite antigen/enzyme detection tests commonly in the form of rapid diagnostic tests (RDTs), and molecular detection tests such as the Polymerase Chain Reaction (PCR). Each of these diagnostic methods has advantages and limitations. Microscopy remains the gold standard for the diagnosis of malaria in endemic countries where an infrastructure to support this method of diagnosis is available. This is the cheapest method, and a skilled microscopist can differentiate species of malaria parasites and provide quantitative data on the level of parasitemia. One of the limitations of microscopy is that it can fail to identify mixed infections and/or low levels of parasitemia. As malaria often occurs in communities where even microscopic diagnosis is not easily available, RDTs have become an alternative tool for malaria diagnosis. Current RDTs capture products such as *P. falciparum*-specific histidine-rich protein –2 (Pf HRP-2), *Plasmodium* genus-specific aldolase and *Plasmodium* genus-specific lactate dehydrogenase enzyme. As RDTs based on aldolase or lactose dehydrogenase enzyme are often genus-specific, these diagnostics are limited in their utility as they are not able to accurately discriminate species.

Molecular diagnostic methods for malaria diagnosis have at least two advantages compared to the other methods: depending upon the target, they can accurately define the species (or multiple species) of malaria parasite(s), and they have higher sensitivity to detect submicroscopic infections. Molecular diagnostic tools have helped to identify zoonotic transmission of *P. knowlesi* in parts of Southeast Asia [21]. Molecular tools are also helping to identify new species of malaria parasites including *P. falciparum*-like parasite species in non-human primates [22]. Several molecular diagnostic tools for malaria are available, the majority of which are PCR-based assays. As molecular tools are expensive, require sophisticated infrastructure and well-trained personnel, these methods are commonly restricted to reference laboratories. Newer, simpler assays such as the recently developed RDTs are easy to use, provide quick results, and are useful alternatives when there is no access to microscopic diagnosis or molecular diagnostics. [23]. However, one major limitation is that RDTs cannot readily distinguish *P. vivax* from other species. Additionally, the HRP2 antigen can persist in blood after parasite clearance, leading to false positive diagnoses. It has also been reported that up to 40% of *Pf* parasites in some parts of South America have *HRP-2* gene deletions, increasing concerns about false negative diagnoses [23]. Molecular diagnostic techniques based on DNA amplification such as PCR (standard, nested, and real-time) are highly sensitive and able to accurately differentiate between species [24].

Molecular diagnosis of malaria parasites began with the use of the 18S ribosomal RNA (18S rRNA) gene as the target about 20 years ago [25] and this method is widely used in many reference laboratories with various modifications. This target was a logical choice in the pre-genomics era. Its regions of conserved sequence allowed cloning from multiple *Plasmodium* species facilitating the subsequent design of species-specific primers. Also, at that point in time, all eukaryotic organisms that had been examined contained multiple, often hundreds of identical copies of rRNA [26], so it seemed likely that this target would lead to a very sensitive assay.

The genome sequence for the most lethal human malaria parasite *Plasmodium falciparum* was first published in 2002 [27]. In 2008, the genomes of both *P. vivax*, the second most important species of human malaria parasite from a public health standpoint, [28] and *P. knowlesi* [21] were published. The genome for the fourth malaria parasite, *P. ovale*, as well as the genomes of several other strains of the other malaria parasites are now also available (Sanger). Examination of *Plasmodium* genome sequences has subsequently revealed that the 18S rRNA target is present in only 4-8 divergent, non-tandem copies depending upon the species. In addition, the few 18S rRNA sequences that are present are not identical in sequence and are variably expressed during the parasite life cycle in some *Plasmodium* species [29]. As PCR sensitivity is greatly influenced by the starting target molecule copy number, a low target copy number limits the detection capabilities of these assays, especially when the parasitemia is low. The 18S rRNA gene target also presents challenges for effective multiplex platforms which would cut back on costs, test time and reduce contamination possibilities. The design of multiple primers to the same target can result in primer competition and decrease the efficiency of the assay. Previous multiplex assays for simultaneous detection of malaria parasite species using the 18S rRNA target showed decreased sensitivity, particularly in detecting the minor species [30].

Despite available whole-genome sequence data for 4 of the 5 human-infective *Plasmodium* species, existing molecular diagnostics still rely on a 2-step PCR protocol that targets the 18S rRNA gene. New diagnostic targets are needed. Ideal targets will be species-specific, highly sensitive and amenable to both single-step and multiplex PCR. *Plasmodium* genome sequence data provide the starting point from which better diagnostic targets can be developed.

1.4 Gene regulation in the Apicomplexa

We have come a long way in recent years in our understanding of apicomplexan gene regulation. While regulation in these parasites is still largely a black box, pieces of the puzzle are being revealed which implicate extensive transcriptional regulatory mechanisms and, to a lesser extent, post-transcriptional mechanisms in the control of apicomplexan gene expression.

1.4.1 Evidence for transcriptional regulation in apicomplexans: examples from *Plasmodium* and *Toxoplasma gondii*

Microarray studies in *Plasmodium* indicated that more than 80% of the transcripts monitored were regulated, with most having a peak expression within a single timeframe of the developmental cycle—mRNA expression in *Plasmodium* is largely stage-specific, with few transcripts shared between stages [31]. Proteomic analysis largely confirms this observation of stage-specific expression [32]. Serial analysis of gene expression (SAGE) in *T. gondii* also demonstrates that specific sets of genes are coordinately transcribed in a stage-specific manner [33]. These findings suggest a tightly regulated program of development in both *Plasmodium* and *T. gondii*, and the stage-dependent fluctuation of mRNA pools implicates transcription as a major mechanism of control.

Though the evidence for transcriptional regulation in these parasites has been fairly conclusive, the elucidation of factors responsible for this regulation has been more challenging. Mining of apicomplexan genomes has indicated the presence of several pieces of the eukaryotic core transcriptional machinery, including several general transcription factors, RNA pol II, and part of the Mediator complex, which mitigates interactions between the basal transcriptional machinery and sequence-specific transcription factors in other eukaryotes [34]. However, notably absent from any of the studied apicomplexan genomes were recognizable sequence-specific transcription factors, those proteins that bind specific enhancer elements and are responsible for spatial and temporal expression of genes in other eukaryotes.

1.4.2 Discovery of the ApiAP2 family of transcriptional regulators and their roles in stage-specific gene expression

The lack of recognizable specific transcription factors in the midst of extensive transcriptional control initially suggested two possibilities for apicomplexan transcriptional regulation: either the specific transcription factors responsible are so divergent from those found in other eukaryotes that they are unrecognizable; or, apicomplexans are unusually reliant on other means to control transcription, such as epigenetic mechanisms. Balaji et al. (2005) took an aggressive approach to tackle the paucity of transcriptional regulators [35]. They screened all identified coding regions in the *P. falciparum* genome with sensitive bioinformatics approaches to identify all known DNA-binding domains. The team revealed a general lack of known DNA-binding domain proteins with one notable exception. They identified a family of proteins with members present in all apicomplexan genomes examined (*Plasmodium, Cryptosporidium,* and *Theileria*) that could potentially be acting as apicomplexan transcription factors. This family of proteins, called ApiAP2 (Apicomplexan Apetala2), is similar to the AP2 family of transcription factors found in plants. The discovery of the ApiAP2 family of proteins set the stage for several exciting regulatory stories to unfold.

To explore the possibility that the 27 ApiAP2 proteins they uncovered in *P. falciparum* were involved in stage-specific control of gene expression, the Balaji team clustered these proteins into groups based on their expression profiles. *P. falciparum* expression data indicated that distinct groups of 4-6 ApiAP2 proteins are expressed during each of the parasite's developmental stages. The role of the AP2 protein family in plant transcriptional regulation coupled with the differential stage-specific expression of the ApiAP2 proteins in *P. falciparum* suggests a role for this family of proteins in controlling transcriptional regulation during developmental switches in *P. falciparum.*

De Silva et al. (2008) were the first to demonstrate sequence-specific binding of two *P. falciparum* ApiAP2 proteins using a custom-made chip containing oligos of all possible 10-mers and subsequently allowing two asexual-stage ApiAP2 proteins (one protein representing a single ApiAP2-domain architecture, the other a tandem ApiAP2-domain architecture) to hybridize [36]. They used the Finding Informative Regulatory Elements (FIRE) algorithm to compile a list of candidate target genes associated with each of the two identified binding motifs. The majority of the putative target genes for one of the proteins, PF0200c, that were annotated were implicated in processes such as host cell rupture and invasion, a guild of proteins required at a similar time point in development that were likely to be co-regulated. Expression profiles for these putative target genes correlated highly with the expression profile of the protein itself, suggesting that PF0200c could indeed be responsible for the regulation of the putative target genes. The FIRE algorithm predicted 21 significantly enriched motifs in total in the *P. falciparum* genome (including those bound by the experimentally verified ApiAP2 proteins), with these motifs occurring upstream of the majority of the 27 ApiAP2 proteins. While the idea was little more than speculative as binding sites were not known for any of the other ApiAP2 proteins, the authors suggested that these 21 predicted motifs occurring within 27 different proteins hint that these proteins regulate each other—that there are ApiAP2 regulatory cascades in control of stage-specific expression.

While the De Silva team's findings provided the first proof of sequence-specific binding by ApiAP2s as well as strong circumstantial evidence that an asexual-stage ApiAP2 protein coordinately regulated a stage-specific set of genes, Yuda et al. (2009) provided real experimental evidence that an ookinete-specific ApiAP2 protein (AP2-O) is the ookinete transcription factor that directly activates invasion-related genes by binding to a specific upstream motif [37]. By generating AP2-O knockout lines of *P. berghei* parasites and subsequently monitoring gene expression in these mutants through DNA microarrays, they found that all 15 genes demonstrating a 5x or greater drop in expression relative to wild type were ookinete-specific with some involvement in invasion. Furthermore, ChIP analysis using anti-GFP antibodies to GFP-fused AP2-O proteins indicated significant enrichment in the promoter regions of all 15 genes identified—suggesting that AP2-O directly regulates all 15 genes. They determined AP2-O binding sites by performing gel shift analyses using motifs shared by the 15 genes as probes. Further analyses using probes containing point mutations and specific and non-specific competitors allowed them to ascertain that AP2-O does indeed bind the 5′-TAGCTA-3′ motif very specifically. Analyses *in vivo* indicated that the identified motif does function as an ookinete stage-specific *cis*-regulatory element *in vivo* as well. Subsequent ChIP analyses indicated that AP2-O does regulate all ookinete-specific genes, even those few not included in the initial group of 15 identified genes. Thus the Yuda group provided the first evidence of an ApiAP2 protein regulating an entire set of stage-specific genes in *Plasmodium*.

Further building on these beginnings to experimentally implicate ApiAP2s in regulation, another ApiAP2 protein has since been identified as a master sporozoite stage-specific regulator in *Plasmodium* (AP2-Sp)[38]. Still another ApiAP2 protein (PFF0200c) has been implicated as a player in *Plasmodium var* gene regulation by acting not as a transcription factor, but by binding the SPE2 DNA motif and interacting with epigenetic machinery to somehow ensure that only one of the sixty members of this family of surface antigens involved in immune evasion is expressed at a time [39]. Campbell et al. (2010) have since comprehensively characterized the binding specificities and the putative regulatory target genes for 27 of the predicted *P. falciparum* ApiAP2 proteins [40]. These studies have not, as yet, produced definitive functional information for these ApiAP2s beyond what is already known, though they did observe that at least one representative of the ApiAP2 family is transcribed at each timepoint across the intra-erythrocytic developmental cycle. This observation suggests that ApiAP2s could be driving transcription throughout the intra-erythrocytic cycle. They additionally found that *P. falciparum* ApiAP2 proteins are able to bind a diverse array of DNA motifs; individual domains within the same protein may recognize completely different sequences, with individual domains being able to bind up to five tertiary motifs. ApiAP2 proteins have very little in common besides the

ApiAP2 domain, which can occur in various architectures with anywhere from one to six variably spaced domains per protein. The diversity of sequences recognized by *P. falciparum* ApiAP2 domain-containing proteins indicates that this family of proteins may be able to regulate a much more complicated network of genes than previously thought.

There are few studies of the functions of ApiAP2 proteins in *T. gondii* to date. However, sensitive bioinformatics profile searches preliminarily indicate a relative explosion of ApiAP2 proteins present in the genomes of both *T. gondii* and the closely-related organism *Neospora caninum* (personal research), and work from the Michael White Laboratory (USF) has implicated ApiAP2 proteins in regulating progression through the *T. gondii* cell cycle [41].

The world of transcriptional regulation in apicomplexans is not entirely comprised of ApiAP2 transcription factors. There are a limited number of other proteins associated with sequence-specific transcription contained within apicomplexan genomes, such as MYBs, C2H2 zinc fingers, GATA-binding factors, and in *Cryptosporidium*, E2Fs [35,42]. However there are few representatives of these families per genome, and no putative transcription factor families have expanded in apicomplexans as ApiAP2s have. Thus, most studies on apicomplexan transcriptional regulation have focused on ApiAP2s.

1.4.3 Evidence for epigenetic regulation of transcription in apicomplexans: examples from *T. gondii* and *Plasmodium spp.*

The emerging story of the ApiAP2 family of transcriptional regulators is certainly an exciting one. However, even when putative ApiAP2 proteins are considered, there is still somewhat of a dearth of transcription factors in apicomplexan genomes compared to what is expected for their genome content. While the model eukaryote *Saccharomyces cerivisiae* has about 200 transcription factors involved in regulating its ~6000 genes, even *T. gondii*, which has a relative explosion of ApiAP2 proteins compared to most other apicomplexans, has maybe ~60 ApiAP2 proteins to regulate a genome of a slightly larger size ([43], personal research). Thus it is possible that ApiAP2 proteins do not comprise the entire story of apicomplexan transcriptional regulation by sequence-specific transcription factors, and there may still be other heretofore-undiscovered transcriptional regulators. It has also been widely considered that apicomplexans may pick up the regulatory slack of this apparent lack of transcription factors with unusual reliance upon epigenetic regulatory mechanisms [44].

While regulation of apicomplexan transcription by sequence-specific transcription factors has only recently been explored, epigenetic regulation of transcription in apicomplexans has proven relatively straightforward to observe. While transcription factors can take on any number of appearances due to lineage-specific expansions of particular proteins [45], chromatin-remodeling and modifying proteins contain a relatively conserved set of easily-recognizable domains [46] which facilitated their rapid discovery in apicomplexan genomes. *Toxoplasma gondii* and *Plasmodium* possess conserved histones H3 and H4, whose tails in other eukaryotes are susceptible to covalent modifications that have different consequences for gene transcription. Each residue in the histone tails reported to be susceptible to chemical modification in other eukaryotes is present in *T. gondii* and *Plasmodium* histones [47]. Protein domains capable of interacting with specific histone modifications are also present in apicomplexan genomes, such as bromodomains (bind acetylated lysines) or chromodomains (bind methylated lysines) [44]. T. *gondii* and *Plasmodium* have both been shown to possess a complement of chromatin remodeling enzymes (many containing domains discussed above) including histone-modifying enzymes such as acetyltransferases, histone deacetylases, methyltransferases and demethyltransferases, as well as SWI2/SNF2 ATPases responsible for nucleosome repositioning [34].

Thus it appears based on computer-based predictions that apicomplexans possess the basic complement of proteins required for epigenetic regulation of transcription. Saksouk et al. (2005) set out for the first time to investigate whether epigenetic marks could be associated functionally with transcription of stage-specific genes in *T. gondii* [48]. They developed a ChIP assay especially for *T. gondii* to monitor the status of histone modifications for both stage-specific and constitutively expressed genes for tachyzoites and bradyzoites. They found that histones H3 and H4 upstream of tachyzoite-specific genes in tachyzoites, bradyzoite-specific genes in bradyzoites, and housekeeping genes in both tachyzoites and bradyzoites were acetylated—as is often the case in other eukaryotes, acetylation occurred upstream of active genes. Bradyzoite genes were hypoacetylated in the tachyzoite stage, and tachyzoite genes were hypoacetylated in the bradyzoite stage, suggesting a correlation of acetylation with active genes. They additionally demonstrated functionality of several of the histone-modifying enzymes predicted bioinformatically using epitope-tagged proteins and subsequent localization studies, including TgGCN5 (acetyltransferase), TgHDAC3 (deacetyltransferases) and TgCARM1 (methyltransferase). Thus it was shown that *T. gondii* does employ epigenetic mechanisms as a means of regulation.

While the Saksouk team demonstrated the utility of epigenetic regulation in *T. gondii* for control of a specific subset of genes, Gissot et al. (2007) attempted a larger-scale study of epigenetic modifications in *T. gondii* using custom oligonucleotide microarrays to examine a contiguous 1% of the genome [49]. They created the microarray using DNA purified from a ChIP procedure performed with intracellular tachyzoites and antibodies to three modified histones associated with active transcription in other eukaryotes—acetylated histone H4 (H4ac), acetylated histone H3 lysine 9 (H3k9ac) and tri-methylated lysine 4 of histone 3 (H3k4me3). They found that these three peaks occurred coincidently in 52 regions, most often found within 1000 bp of the region of predicted coding regions—the modifications were found in promoter regions. To detect whether these genes were actively transcribed, the authors hybridized intracellular tachyzoite cDNA to their array. Indeed, 51 of the 52 regions enriched for the three chromatin modifications had significant cDNA hybridized to them, suggesting a correlation between the three modifications and actively transcribed genes. Thus epigenetic markers can be used to predict active promoters on a much larger scale in *T. gondii*.

The Gissot team additionally tested for correlation between expression and three other histone modifications thought to be general activation marks in *T. gondii* or other eukaryotes—histone H3 dimethylated at arginine 17 (HeR17me2), histone H3 methylated at lysine 4 (H3K4me1), and histone H3 dimethylated at lysine 4 (H3K4me2). The smaller survey of histone modifications from the Saksouk group (2005) found HeR17me2 associated with the subset of genes they examined, and they predicted the modification would be present at all active *T. gondii* promoters. However, this modification was found at only 4 of the 52 regions identified. The other two modifications they tested for were not associated with any active promoter regions. These findings of a modification present at select, active promoters suggested the existence of possible additional layers of epigenetic regulation unaddressed by studies to date—there are aspects of the histone code in apicomplexans that the field has yet to understand. Indeed, not all of the regions associated with the identified histone modifications corresponded to actively-transcribed genes—one region in particular, associated with a gene encoding a putative UV-induced double-strand break protein had no detectable mRNA expression. The authors draw from findings in humans that the histone modifications might in this case place the gene in a "ready to go" state in the event that it is needed, but until then is not actively transcribed—this mechanism of regulation has been attributed to up to 20% of regions with modified histones found in a survey of human epigenetic regulation [50,51]. Studies in yeast have found that pol II is poised at promoters of rapidly expressed genes bearing these epigenetic marks, and low-similarity homologs of factors responsible for this pausing of the polymerase in yeast are present in *T. gondii*, making this explanation of histone modifications at inactive promoters plausible [50]. Histone acetylation of bradyzoite-specific genes has been reported in the tachyzoite stage of *T. gondii* strains that rapidly differentiate into bradyzoites (though the bradyzoite genes are not expressed in the tachyzoite stage) [52], lending more support to the idea that epigenetic markers indicate a poised state for transcription in *T. gondii*.

Several studies have additionally demonstrated the importance of epigenetic regulation in *Plasmodium*, particularly in regard to *var* genes and the phenomenon of antigenic variation. Antigenic variation describes the occurrence in *Plasmodium* whereby only one surface antigen out of large multi-member gene families is expressed at any one time, which helps the parasite avoid elimination by the host immune system. Until recently, mechanisms controlling this phenomenon were largely mysterious; now epigenetic regulation is known to play a large role. Hypoacetylation has been associated with *var* gene silencing [53]. Chookajorn et al. (2007) additionally implicated histone modifications in the expression of *var* genes using a transgenic line of parasites that expressed one particular *var* gene under drug selection [54]. They found that silent *var* loci promoters were enriched for trimethylation of lysine 9 on histone H3 (H3K9me3), a mark also associated with silencing in other eukaryotes, while there was no association of the mark with the active *var* locus. When they allowed the transgenic parasites to switch *var* genes, the previously active *var* promoter also became enriched for the H3K9me3 marker, indicating a role for this modification in repression of *var* genes. Dzikowski and Deitsch (2008) deepened the epigenetic regulatory story with their finding that briefly preventing an active *var* gene from being transcribed results in the gene reverting quickly to a silenced state, indicating that active transcription is necessary for the maintenance of the memory of *var* gene expression [55].

Ralph et al (2005) showed that nuclear positioning of both sub-telomeric and internally-positioned *var* loci is also involved in *var* gene transcription [56]. FISH analysis indicated that the 60 *var* genes do not scatter randomly throughout the nucleus—they cluster in 6-8 groups on the nuclear periphery. To further investigate whether or not there is differential positioning of *var* loci dependent on transcriptional state, the group did FISH co-localization experiments using probes for the active *var* gene, inactive *var* genes and telomeric clusters (which are associated with silencing in other eukaryotes). They found that inactive *var* genes often co-localized with the telomeric clusters, while the active *var* gene did not. When they tested for co-localization of the same *var* gene in an inactive state, they found it also co-localized with telomeric clusters. These data suggest that active *var* loci do indeed dissociate from inactive *var* loci. The reasons for this occurrence could be to bring the locus into closer association with transcriptional machinery, or perhaps to move the active locus away from silencing proteins enriched in telomeric clusters. Experimental evidence for these speculations is required. In any case, sub-nuclear localization appears to bear heavily upon *var* gene regulation.

Having noted the importance of sub-nuclear localization to *var* gene expression and the existence of nuclear sub-compartment “factories” for transcription and silencing in other organisms, the Scherf lab tried to correlate specific histone modifications with these nuclear sub-compartments in *Plasmodium* [57]. They performed an immunofluorescence assay to examine the sub-nuclear localization of several specific histone modifications. They noticed a particularly striking pattern of localization for the histone H3 lysine 79 trimethylation mark, which were polarized to one end of the nucleus—thus histone modifications showed differential sub-nuclear enrichment. As the H3K79me3 modification is correlated with activation in other eukaryotes, the group investigated the possibility that this particular modification was marking any transcriptionally competent zone, or any of the previously characterized *Plasmodium* nuclear sub-compartments (ie telomere clusters, perinuclear expression site for *var* genes) through FISH co-localization studies. They didn't find any associations between the histone modification and the other previously characterized sub-compartments. Thus the authors put forward that the epigenetic modification could be marking an active transcriptional compartment for another group of genes other than the *var* family, an assertion that is purely speculative without tests for co-localization of other gene families. However it is plausible at this point to consider that particular histone modifications can be used to identify nuclear sub-compartments, and perhaps there is an additional role for these modifications in marking certain genes for targeting to particular nuclear sub-compartments.

These examples from studies in both *T. gondii* and *Plasmodium* indicate expansive roles for epigenetic regulation in the control of apicomplexan gene expression. Certainly the true extent of epigenetic regulation in apicomplexans is only just beginning to be understood, and several of the above studies offer intriguing hints that the story is very complex.

1.4.4 Evidence for post-transcriptional mechanisms of gene regulation

Evidence for post-transcriptional regulation is largely obtained through assays of mRNA levels and comparisons to resultant protein levels. Studies in *Plasmodium* indicate very close correlation between mRNA expression and cognate protein across 9 different life cycle stages [58], which would suggest very little post-transcriptional regulation is going on. The same trends have generally been observed in *T. gondii*.

While it doesn't appear that post-transcriptional regulatory mechanisms are involved in large-scale control of gene expression (unlike transcriptional regulation), there are several examples from the literature demonstrating the importance of post-transcriptional regulation to individual genes. Studies in *T. gondii* indicate unbalanced mRNA to protein ratios for specific genes such as certain surface antigens and proliferating cell nuclear antigens [34]. Studies in *Plasmodium* indicate that a homolog to an RNA helicase involved in translational repression in other eukaryotes is expressed in blood-stage gametocytes (termed DOZI in *Plasmodium*—development of zygotes inhibited) [59]. Immunoprecipitation of DOZI showed its association with several mRNAs that were predicted to be translationally repressed due to imbalances between mRNA levels and protein. Yuda et al. (2009) showed that ApiAP2 transcription factor AP2-O is regulated by this mechanism using AP2-O/GFP transgenic parasites [37]. Though mRNA expression data indicated that AP2-O is transcribed in the gametocyte stage, fluorescence indicating the presence of the protein was not detected until several hours after fertilization. Immunoprecipitation experiments using anti-GFP antibodies and DOZI::GFP transgenic parasites indicated that AP2-O mRNA was indeed complexed with DOZI, suggesting that AP2-O is translationally repressed until after fertilization.

These few examples of the importance of post-transcriptional regulatory mechanisms to individual apicomplexan genes alert us to the existence of the mechanism in these parasites, and certainly it is possible that other cases exist and that the importance of post-transcriptional regulation has been underestimated.

1.5 Using genomics to improve apicomplexan diagnostics and the study of transcriptional regulation

We are in an age when many apicomplexan genomes are available to us. These genomes have given us (1) the potential to develop diagnostic targets that are species-specific and no longer limited by what we understand biologically about most organisms (as was the case with 18S rRNA); and (2) the power to study transcriptional regulatory phenomena on an organism-wide scale, as our colleagues have undertaken in *Plasmodium* and to a lesser extent, *Toxoplasma* [40,41]. The availability of multiple *Plasmodium* genomes, as well as that of the human host, have been essential components of the methodology we developed to select improved malaria diagnostic targets, discussed in Chapters 2 and 6.

Though much has been learned about apicomplexan gene regulation from *Plasmodium* and *Toxoplasma*, there is a notable void in the field, especially in the even more distantly related apicomplexan, *Cryptosporidium*. Studies aimed at characterizing putative transcription factor function, such as the characterization of AP2-O and AP2-Sp referenced in previous sections, often involve genetic manipulation, a tool that is currently unavailable in the experimentally intractable *Cryptosporidium*. As the published data largely address ApiAP2 regulation only in *Plasmodium spp.*, there have been no extensive comparative studies between organisms, and the question of the evolution of this gene family has not been formally addressed. Certainly there are a myriad avenues left to explore to fully elucidate the roles of ApiAP2 proteins in apicomplexan gene regulation, or where transcriptional regulation fits in with epigenetic and post-transcriptional regulatory mechanisms to control gene expression. Indeed, ApiAP2 proteins have been demonstrated to interact with epigenetic machinery in both *T. gondii* and *Plasmodium* [60,61], suggesting that regulation is most certainly an interplay between transcriptional, epigenetic and possibly also post-transcriptional mechanisms.

Chapters 3 and 4 of this dissertation will largely be focused on the study of apicomplexan transcriptional regulatory mechanisms, though it is important to understand that these mechanisms are only one piece of the regulatory puzzle. It is also important to note that the ApiAP2 family of proteins, so central to studies of apicomplexan transcriptional regulation to date, are not the only sequence-specific transcription factors in apicomplexan genomes, and other transcription factor families may contribute heavily to transcriptional regulation in *C. parvum* in particular (discussed in Chapter 3). The availability of the *Cryptosporidium parvum* genome, as well as several other apicomplexan and chromalveolate genomes, have allowed us to shed light on *C. parvum* transcriptional regulatory mechanisms, as well as to make comparative studies to learn about the evolution of the ApiAP2 transcriptional regulatory network.

REFERENCES

1. Wasmuth J, Daub J, Peregrin-Alvarez JM, Finney CAM, Parkinson J (2009) The origins of apicomplexan sequence innovation. Genome Research 19: 1202-1213.

2. Huang JL, Mullapudi N, Sicheritz-Ponten T, Kissinger JC (2004) A first glimpse into the pattern and scale of gene transfer in the Apicomplexa. International Journal for Parasitology 34: 265-274.

3. Striepen B, Pruijssers AJP, Huang JL, Li C, Gubbels MJ, et al. (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. Proceedings of the National Academy of Sciences of the United States of America 101: 3154-3159.

4. Delwiche CF (1999) Tracing the Thread of Plastid Diversity through the Tapestry of Life. Am Nat 154: S164-S177.

5. Escalante AA, Ayala FJ (1995) Evolutionary origin of *Plasmodium* and other Apicomplexa based on ribosomal-RNA genes. Proceedings of the National Academy of Sciences of the United States of America 92: 5793-5797.

6. Okamoto N, McFadden GI (2008) The mother of all parasites. Future Microbiology 3: 391-395.

7. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064-1071.

8. Dubey JP, Jones JL (2008) *Toxoplasma gondii* infection in humans and animals in the United States. International Journal for Parasitology 38: 1257-1278.

9. Montoya JG, Liesenfeld O (2004) Toxoplasmosis. Lancet 363: 1965-1976.

10. Jones JL, Kruszon-Moran D, Wilson M, McQuillan G, Navin T, et al. (2001) *Toxoplasma gondii* infection in the United States: seroprevalence and risk factors. Am J Epidemiol 154: 357-365.

11. Leander BS, Clopton R, Keeling P (2003) Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. International Journal of Systematic and Evolutionary Microbiology 53: 345-354.

12. Carreno RA, Matrin DS, Barta JR (1999) Cryptosporidium is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. Parasitology Research 85: 899-904.

13. TR N, AM H (1987) Cryptospridiosis in patients with Aids. Journal of infectious diseases 155.

14. Tzipori S (1988) Cryptosporidiosis in perspective. Adv Parasitol 27: 63-129.

15. Spano F, Crisanti A (2000) *Cryptosporidium parvum*: the many secrets of a small genome. Int J Parasitol 30: 553-565.

16. Campbell PN, Current WL (1983) Demonstration of serum antibodies to *Cryptosporidium sp.* in normal and immunodeficient humans with confirmed infections. J Clin Microbiol 18: 165-169.

17. Baldauf SL (2008) An overview of the phylogeny and diversity of eukaryotes. Journal of Systematics and Evolution 46: 263-273.

18. Baldauf SL (2003) The Deep Roots of Eukaryotes. Science 300: 1703-1706.

19. Poon LLM, Wong BWY, Ma EHT, Chan KH, Chow LMC, et al. (2006) Sensitive and inexpensive molecular test for *falciparum* malaria: Detecting *Plasmodium falciparum* DNA directly from heat-treated blood by loop-mediated isothermal amplification. Clinical Chemistry 52: 303-306.

20. Paris DH, Imwong M, Faiz AM, Hasan M, Bin Yunus E, et al. (2007) Loop-mediated isothermal PCR (LAMP) for the diagnosis of *falciparum* malaria. American Journal of Tropical Medicine and Hygiene 77: 972-976.

21. Singh B, Kim Sung L, Matusop A, Radhakrishnan A, Shamsul SS, et al. (2004) A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363: 1017-1024.

22. Prugnolle F, Durand P, Neel C, Ollomo B, Ayala FJ, et al. (2010) African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. Proc Natl Acad Sci U S A 107: 1458-1463.

23. Moody A (2002) Rapid diagnostic tests for malaria parasites. Clin Microbiol Rev 15: 66-78.

24. Mens P, Spieker N, Omar S, Heijnen M, Schallig H, et al. (2007) Is molecular biology the best alternative for diagnosis of malaria to microscopy? A comparison between microscopy, antigen detection and molecular tests in rural Kenya and urban Tanzania. Trop Med Int Health 12: 238-244.

25. Snounou G, Viriyakosol S, Zhu XP, Jarra W, Pinheiro L, et al. (1993) High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Mol Biochem Parasitol 61: 315-320.

26. Mercereau-Puijalon O, Barale JC, Bischoff E (2002) Three multigene families in *Plasmodium* parasites: facts and questions. Int J Parasitol 32: 1323-1344.

27. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498-511.

28. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455: 757-763.

29. Li J, Gutell RR, Damberger SH, Wirtz RA, Kissinger JC, et al. (1997) Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. J Mol Biol 269: 203-213.

30. Mixson-Hayden T, Lucchi NW, Udhayakumar V Evaluation of three PCR-based diagnostic assays for detecting mixed *Plasmodium* infection. BMC Res Notes 3: 88.

31. Llinas M, DeRisi JL (2004) Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. Curr Opin Microbiol 7: 382-387.

32. Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, et al. (2005) A Comprehensive Survey of the *Plasmodium* Life Cycle by Genomic, Transcriptomic, and Proteomic Analyses. Science 307: 82-86.

33. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, et al. (2005) The transcriptome of *Toxoplasma gondii*. Bmc Biology 3: 18.

34. Weiss L, Kim K (2007) *Toxoplasma gondii*: the model apicomplexan: perspectives and methods: Academic Press.

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

36. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. (2008) Specific DNA-binding by Apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 105: 8393-8398.

37. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, et al. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Molecular Microbiology 71: 1402-1414.

38. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites. Molecular Microbiology 75: 854-863.

39. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BTF, et al. (2010) A Major Role for the *Plasmodium falciparum* ApiAP2 Protein PfSIP2 in Chromosome End Biology. Plos Pathogens 6.

40. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. Plos Pathogens 6.

41. Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, et al. (2010) Coordinated Progression through Two Subtranscriptomes Underlies the Tachyzoite Cycle of *Toxoplasma gondii*. Plos One 5.

42. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Research 14: 1686-1695.

43. Ho SW, Jona G, Chen CTL, Johnston M, Snyder M (2006) Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proceedings of the National Academy of Sciences of the United States of America 103: 9940-9945.

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

45. Anantharaman V, Iyer LM, Aravind L (2007) Comparative genomics of protists: New insights into the evolution of eukaryotic signal transduction and gene regulation. Annual Review of Microbiology 61: 453-475.

46. Iyer LM, Anantharaman V, Wolf MY, Aravind L (2008) Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. International Journal for Parasitology 38: 1-31.

47. Sullivan WJ, Hakimi MA (2006) Histone mediated gene activation in *Toxoplasma gondii*. Molecular and Biochemical Parasitology 148: 109-116.

48. Saksouk N, Bhatti MM, Kieffer S, Smith AT, Musset K, et al. (2005) Histone-modifying complexes regulate gene expression pertinent to the differentiation of the protozoan parasite *Toxoplasma gondii*. Molecular and Cellular Biology 25: 10301-10314.

49. Gissot M, Kelly KA, Ajioka JW, Greally JM, Kim K (2007) Epigenomic modifications predict active promoters and gene structure in *Toxoplasma gondii*. Plos Pathogens 3: 709-719.

50. Gissot M, Kim K (2008) How Epigenomics Contributes to the Understanding of Gene Regulation in *Toxoplasma gondii*. Journal of Eukaryotic Microbiology 55: 476-480.

51. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

52. Behnke MS, Radke JB, Smith AT, Sullivan WJ, White MW (2008) The transcription of bradyzoite genes in *Toxoplasma gondii* is controlled by autonomous promoter elements. Molecular Microbiology 68: 1502-1518.

53. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, et al. (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. Cell 121: 25-36.

54. Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ, et al. (2007) Epigenetic memory at malaria virulence genes. Proceedings of the National Academy of Sciences of the United States of America 104: 899-902.

55. Dzikowski R, Deitsch KW (2008) Active transcription is required for maintenance of epigenetic memory in the malaria parasite *Plasmodium falciparum*. Journal of Molecular Biology 382: 288-297.

56. Ralph SA, Scheidig-Benatar C, Scherf A (2005) Antigenic variation in *Plasmodium falciparum* is associated with movement of var loci between subnuclear locations. Proceedings of the National Academy of Sciences of the United States of America 102: 5414-5419.

57. Issar N, Ralph SA, Mancio-Silva L, Keeling C, Scherf A (2009) Differential sub-nuclear localisation of repressive and activating histone methyl modifications in *P. falciparum*. Microbes Infect 11: 403-407.

58. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, et al. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res 14: 2308-2318.

59. Mair GR, Braks JA, Garver LS, Wiegant JC, Hall N, et al. (2006) Regulation of sexual development of *Plasmodium* by translational repression. Science 313: 667-669.

60. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 438: 103-107.

61. Dixon SE, Stilger KL, Elias EV, Naguleswaran A, Sullivan WJ, Jr. (2010) A decade of epigenetic research in *Toxoplasma gondii*. Mol Biochem Parasitol 173: 1-9.

Figure legends

Figure 1.1. Cladogram of apicomplexan relationships. Select apicomplexans for which genomes are available. Genome size, protein count, and number of chromosomes are indicated for each organism (credit: Jeremy DeBarry, Kissinger Research Group 2012).

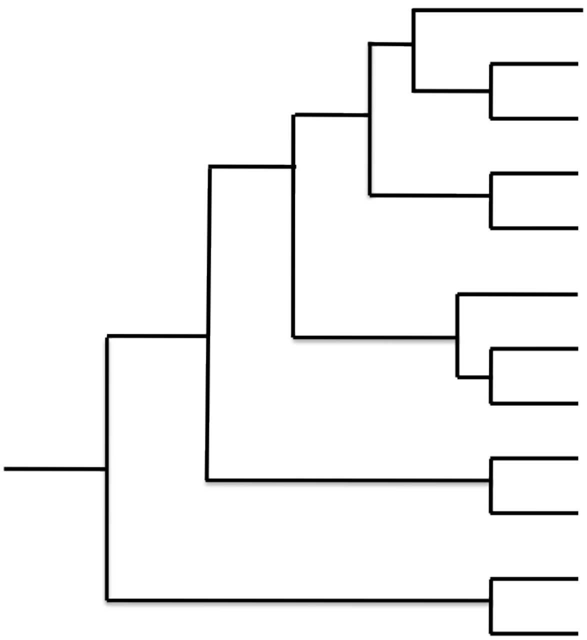
Figure 1.2. Life cycle of *Plasmodium falciparum*. *P. falciparum* requires both a mosquito and a human host to complete its lifecycle, where the parasite undergoes sexual and asexual replication, respectively. A female *Anopheles* mosquito vector bites the human host, releasing *Plasmodium* sporozoites into the bloodstream. Sporozoites travel to the liver, where they undergo several rounds of asexual replication, eventually bursting out into the blood stream in the "merozoite" form. Merozoites invade red blood cells, and again undergo a round of asexual replication, forming a multi-nucleate cell (schizont) which ultimately divides to form 12-16 more merozoites. These merozoites lyse out of red blood cells and undergo the red blood cell invasion process over and over again. Some merozoites skip merogony and instead undergo sexual differentiation into female and male gametes, which are taken up by the female *Anopheles* mosquito to undergo sexual replication and begin the cycle all over again. Adapted from Pleass and Holder, Nature Reviews Microbiology 3: 893-899 (Nov. 2005).

Figure 1.3. Life cycle of *Toxoplasma gondii*. The life cycle includes both sexual and asexual modes of replication. The sexual cycle takes place exclusively in the intestinal enterocytes of many members of the cat family. (a, b) After ingestion of tissue cysts, the parasites invade the enterocytes, undergo several rounds of division and (c) differentiate into microgametocytes and

macrogametocytes. (d) The gametocytes fuse to form a zygote or 'oocyst' that is shed into the environment with the cat's feces. (e) The oocyst undergoes meiosis, producing an octet of highly infectious 'sporozoites' that are resistant to environmental damage and may persist for years in a moist environment. (f) After ingestion (by a secondary host such as a mouse), (g) sporozoites differentiate into the rapidly dividing 'tachyzoite' form, which establishes and sustains the acute infection. (h) During the acute infection, congenital transmission to the developing fetus can occur. (i) In many hosts, a chronic phase of the disease ensues, as the tachyzoite changes into a slowly dividing form known as the 'bradyzoite'. Latent bradyzoite tissue cysts persist for the life of the host, re-emerging occasionally, but do not produce overt disease in healthy individuals. (j) Carnivorous ingestion of tissue cysts can lead to the infection of a naive host, allowing for an indefinite nonsexual propagation of T.gondii. (k) In the cat, this will initiate the sexual cycle. The solid lines indicate parasite differentiation and the dashed lines indicate modes of transmission (Ajioka, JW. et al. Expert Rev. Mol. Med. 2001:1-19).

Figure 1.4. Life cycle of *Cryptosporidium parvum*. Following ingestion, excystation (a) occurs. The sporozoites are released and parasitize epithelial cells (b ,c) of the gastrointestinal tract and some other tissues. The parasites undergo asexual multiplication (merogony) at 24-48 hr (d ,e ,f), and then sexual multiplication producing microgamonts (male) (g) and macrogamonts (female) (h) 48-72 hr. Upon fertilization of the macrogamonts by the microgametes (i), oocysts (j,k) develop that sporulate in the infected host. Two different types of oocysts are produced, the thick-walled (J), which is commonly excreted from the host, and the thin-walled (k) oocyst , which is primarily involved in autoinfection. Reproduced from http://www.dpd.cdc.gov/.

Figure 1.5. The apicomplexan phylum in context of the eukaryotic tree of life. Apicomplexa is boxed in red. Modified from Keeling et al. (2005).



Species	Mb	Protein #	Chr #
Plasmodium falciparum (3D7)	~23	5491	14
Plasmodium knowlesi (H)	~25	5197	14
Plasmodium vivax (Sal_1)	~23	5432	14
Plasmodium berghei (ANKA)	~23	4904	14
Plasmodium chabaudi (chabaudi)	~23	5131	14
Babesia bovis (Texas T2Bo)	~9	3671	4
Theileria annulata (Muguga)	~8.5	3792	4
Theileria parva (Ankara)	~8.5	4023	4
Toxoplasma gondii (ME49)	~63	7993	14
Neospora caninum (Nc_LIV)	~62	7082	14
Cryptosporidium muris (RN66)	~9.1	3934	8
Cryptosporidium parvum (IOWA)	~9.2	3805	8

Figure 1.1. Cladogram of apicomplexan relationships.

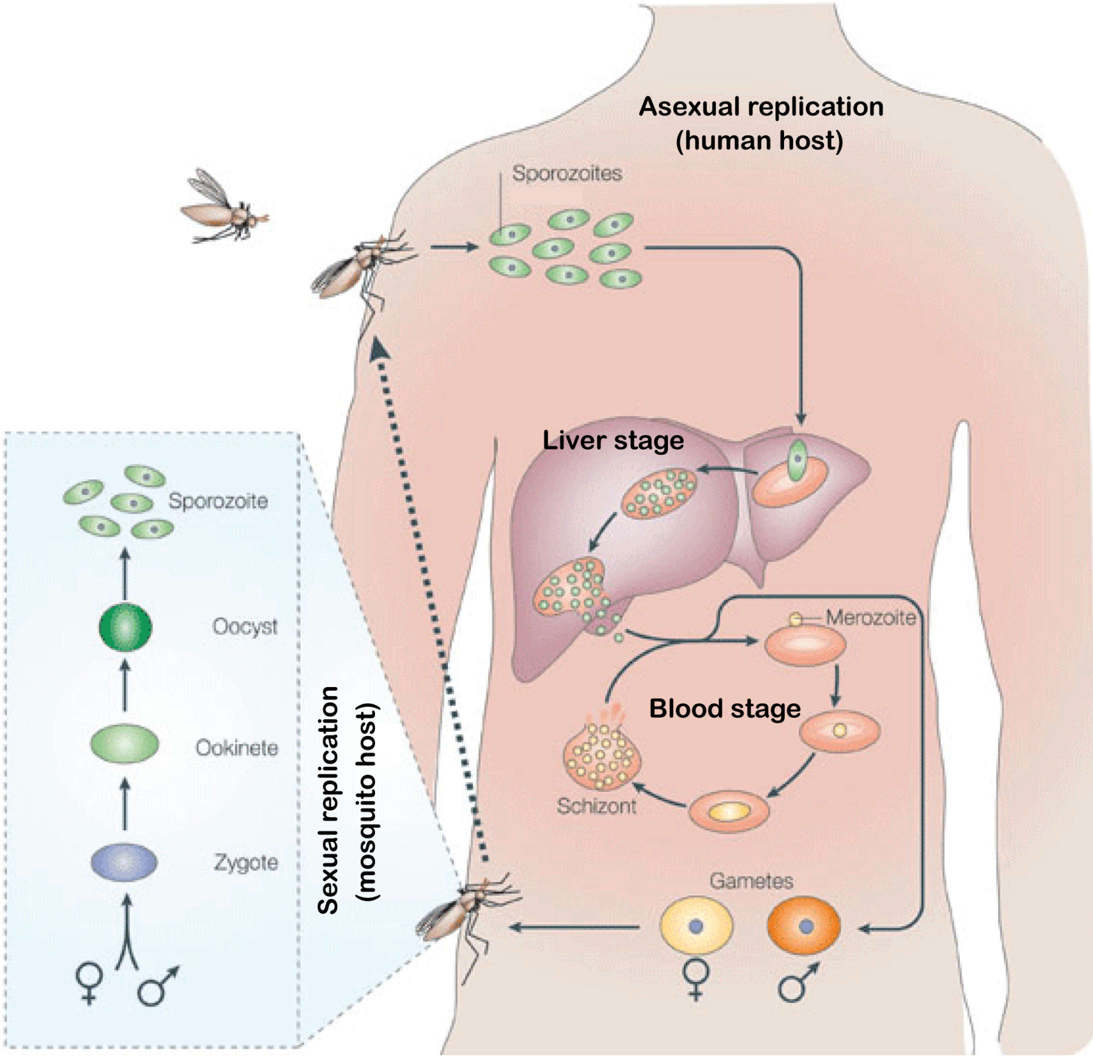


Figure 1.2. Life cycle of *Plasmodium falciparum*.

(Adapted from Pleass and Holder, Nature Reviews Microbiology 3: 893-899, Nov. 2005)

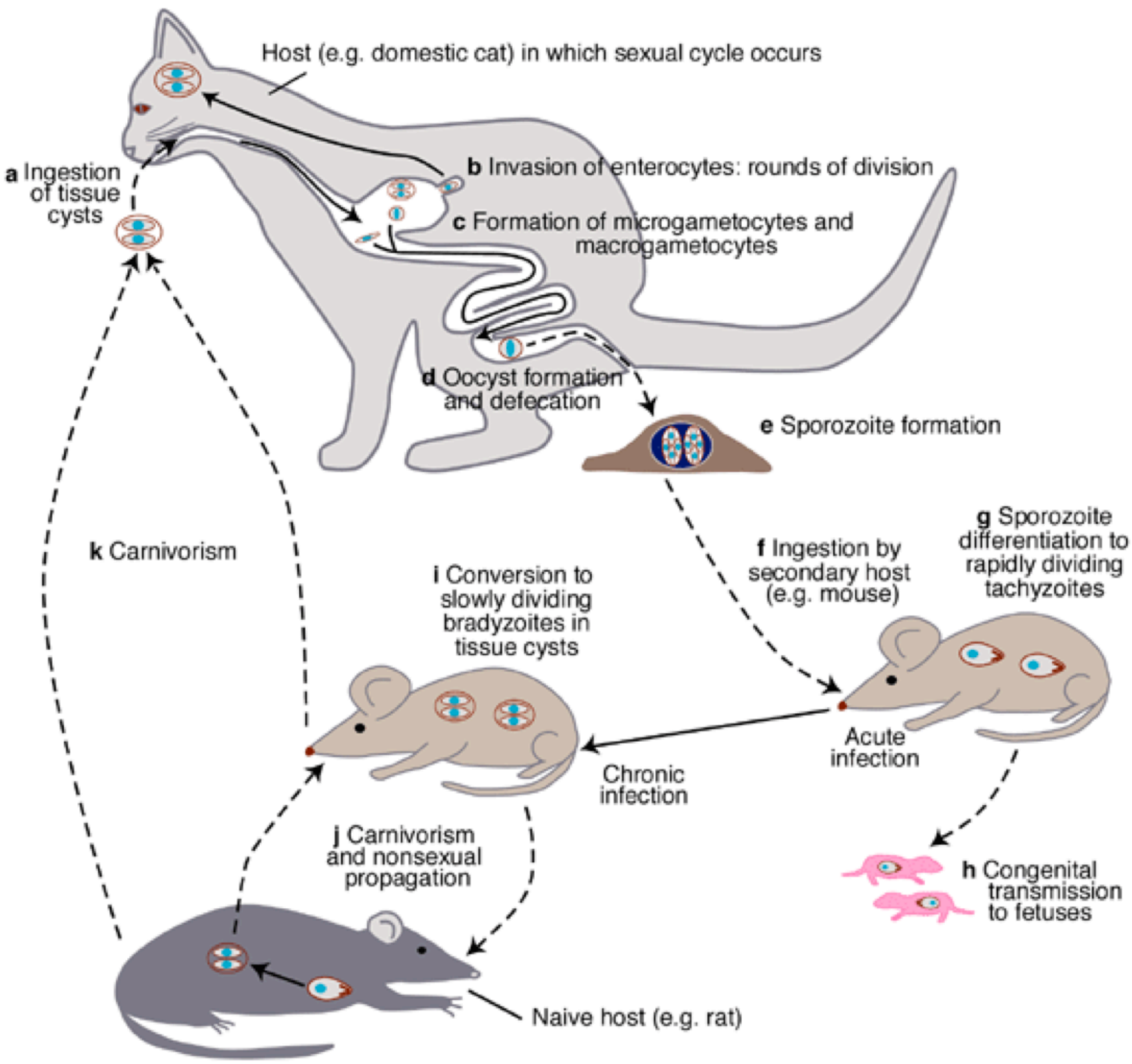


Figure 1.3. Life cycle of *Toxoplasma gondii*.

(Ajioka, JW. et al. Expert Rev. Mol. Med. 2001:1-19)

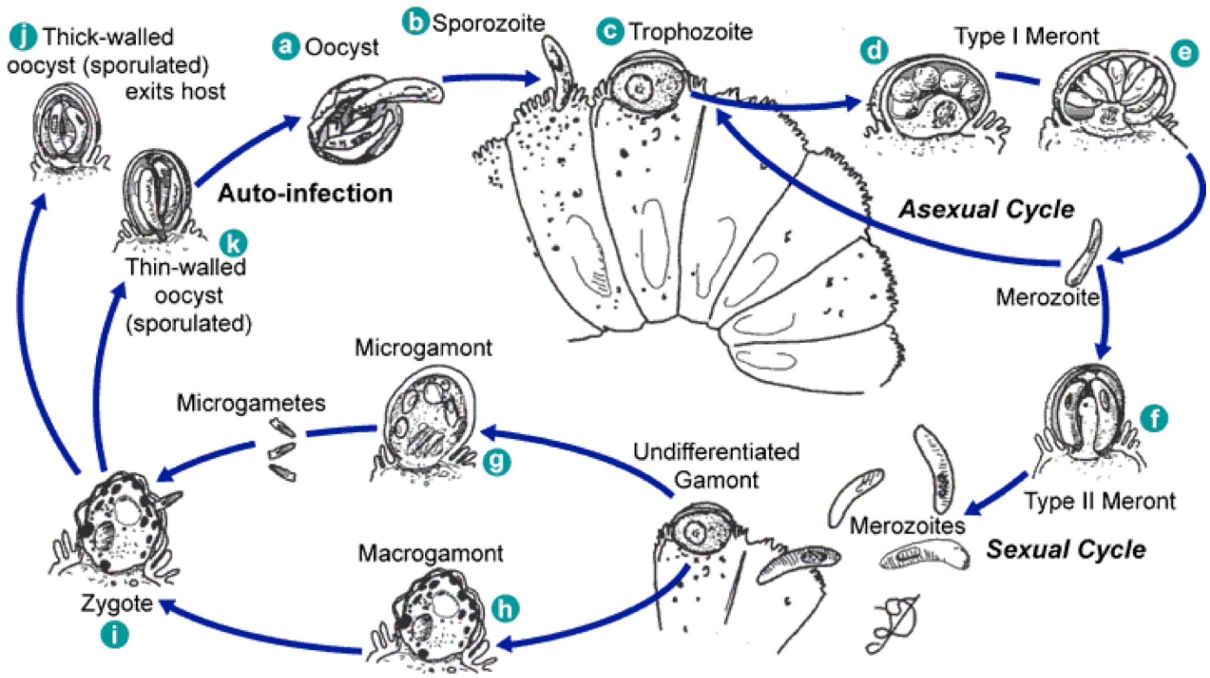


Figure 1.4. Life cycle of *Cryptosporidium parvum*.

(Reproduced from http://www.dpd.cdc.gov/)

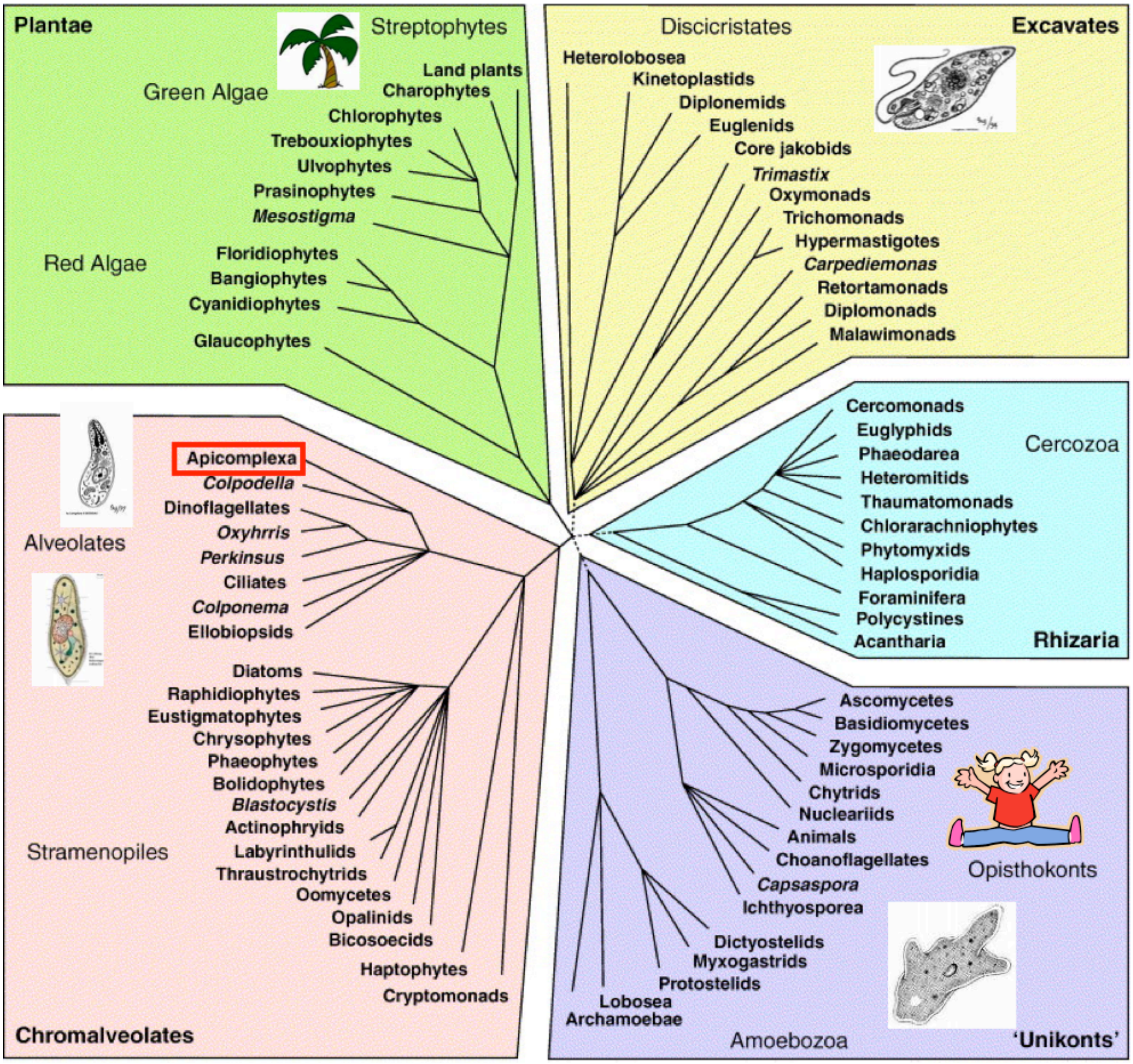


Figure 1.5. The apicomplexan phylum in context of the eukaryotic tree of life.

(Adapted from Keeling et al., Trends in Ecology and Evolution, 20:12 2005)

CHAPTER 2

Applied genomics: Data mining reveals species-specific malaria diagnostic targets more sensitive than 18S rRNA

Allison Demas*, Jenna Oberstaller*, Jeremy DeBarry*, Naomi W. Lucchi, Ganesh Srinivasamoorthy, Deborah Sumari, Abdunoor M. Kabanywanyi, Leopoldo Villegas, Ananias A. Escalante, S. Patrick Kachur, John W. Barnwell, David S. Peterson, Venkatachalam Udhayakumar, and Jessica C. Kissinger. 2011. J Clin Microbiol. 2011 July; 49(7): 2411–2418. Reprinted here with permission of publisher.

*These authors contributed equally to this work.

doi:10.1128/JCM.02603-10

ABSTRACT

Accurate and rapid diagnosis of malaria infections is crucial for implementing species-appropriate treatment and saving lives. Molecular diagnostic tools are the most accurate and sensitive method of detecting malaria parasite species, capable of differentiating between *Plasmodium* species and detecting even sub-clinical infections. Despite available whole-genome sequence data for *Plasmodium falciparum* and *P. vivax*, the majority of PCR-based methods still rely on the 18S ribosomal RNA (18S rRNA) gene targets. Historically, this gene has served as the best target for diagnostic assays. However, it is limited in its ability to detect mixed infections in multiplex assay platforms without the use of nested PCR. New diagnostic targets are needed. Ideal targets will be species-specific, highly-sensitive and amenable to both single-step and multiplex PCR. We have mined the genomes of *P. falciparum* and *P. vivax* to identify species-specific, repetitive sequences that serve as new PCR targets for the detection of malaria. We show that these targets (Pvr47 & Pfr364) exist in 14-41copies and are more sensitive than 18S rRNA when utilized in a single-step PCR reaction. Parasites are routinely detected at levels of 1-10 parasites/μl. The reaction can be multiplexed to detect both species in a single reaction. We have examined 7 *P. falciparum* strains and 91 *P. falciparum* clinical isolates from Tanzania and 10 *P. vivax* strains and 96 *P. vivax* clinical isolates from Venezuela, and we have verified a sensitivity and specificity of ~100% for both targets compared with a nested 18S rRNA approach. We show that bioinformatics approaches can be successfully applied to identify novel diagnostic targets and improve molecular methods for pathogen detection. These novel targets provide a powerful alternative molecular diagnostic method for the detection of *P. falciparum* and *P. vivax* in conventional or multiplex PCR platforms.

INTRODUCTION

Malaria continues to be a leading cause of morbidity and mortality worldwide. It is responsible for 2-300,000 diagnosed cases and 6-900,000 deaths in 2009 alone (41). Early detection and accurate diagnosis are the best tools for saving lives in endemic regions. Correct species identification and accurate diagnosis of mixed infections are of particular importance for proper treatment in regions where multiple parasite species are endemic. Of the five species within the genus *Plasmodium* known to infect humans, *Plasmodium falciparum* (*Pf*), is the most deadly, followed by *Plasmodium vivax* (*Pv*), which also causes significant morbidity and some mortality (2, 10, 14, 23, 29, 39). *Pf* and *Pv* also have wider global distributions than other species. The remaining three species, which are not the subject of this paper, *P. malariae* (*Pm*), *P. ovale* (*Po*), and *P. knowlesi* (*Pk*), each have differing global distributions (with *Pm* being found primarily in South America and Asia, and *Po* and *Pk* being found primarily in Asia) and differing levels of morbidity and mortality.

Light microscopy remains the gold standard of malaria diagnosis in endemic regions. While microscopy is cost-effective and requires little equipment, a well- trained microscopist is essential. A highly trained and experienced microscopist can typically detect parasitemias as low as 90-200 parasites/µl. Misdiagnosis may still occur due to low parasitemia or mixed infection. Immunochromatographic rapid diagnostic tests (RDTs) are increasingly being implemented in case management and control programs. RDTs identify the parasite antigens HRP2, pLDH, or pAldolase, and may be pan-specific (for all *Plasmodium* species) or *Pf* specific, or both, depending on the test. RDTs are not effective for the full diagnosis of mixed infections, as they can only distinguish *Pf* and indicate the presence or absence of another *Plasmodium* species. While they can detect parasitemia as low as 100 parasites/µl, they are not quantitative (21).

Additionally, the HRP2 antigen can persist in blood after parasite clearance, leading to false positive diagnoses. It has also been reported that up to 40% of *Pf* parasites in some parts of South America have *HRP-2* gene deletions, increasing concerns about false negative diagnoses (8).

Molecular diagnostic tools are the most accurate and sensitive method of detecting malaria parasite species. Their current use however, is restricted to reference laboratories or research studies, since there are limitations associated with the use of molecular tools in endemic regions for routine diagnostic use (including infrastructure problems, prohibitive costs, a refrigerated or frozen supply cold chain, and the requirement of trained personnel). Despite these limitations, molecular methods are the best methods for detecting multiple species and sub-clinical infections (4, 7), making them invaluable for malaria parasite detection. Molecular methods will become increasingly important given the proposed eradication/elimination goals and the need to detect sub-clinical infections (12).

Polymerase chain reaction-based amplification methods (PCR), including multiplex PCR, real-time PCR and, more recently, the loop-mediated DNA amplification method (LAMP), have been developed to detect malaria parasite species (11, 24, 25, 31, 32, 36, 38). Molecular methods offer the advantage of highly specific differentiation of *Plasmodium* species. Recently, molecular techniques confirmed the natural infection of humans with the zoonotic *P. knowlesi* in Southeast Asia (34). This simian malaria parasite species had not previously been found in humans in great numbers, and a similar morphology resulted in an incorrect *P. malariae* diagnosis by microscopy.

The most widely used molecular target for the detection of *Plasmodium* and diagnosis of malaria was developed prior to the completion of any *Plasmodium* genome sequence. The target is the 18S ribosomal RNA (18S rRNA) gene(s) (11, 16, 30, 33, 35). This target was a logical choice given its high sequence conservation, the availability of universal primer sequences for its amplification and the fact that it was known to exist in multiple copies in all organisms that had been examined at the time. The availability of complete *Plasmodium* genome sequences presents a great opportunity for improving the existing molecular diagnostic tools by identifying new targets for more sensitive and specific detection. The *Pf* genome was completed in 2002 (9), and *Pv* and *P. knowlesi* have since been sequenced (5, 26). Despite the existence of genomic information for three of the five human-infecting malaria parasites for many years, the majority of molecular diagnostic tools still rely on 18S rRNA. Subsequent examination of *Plasmodium* genome sequences has revealed that the 18S rRNA target is present in only 4-8, divergent, non-tandem copies depending upon the species, in contrast to other eukaryotic genomes that have hundreds of tandem copies of rRNA gene clusters (18, 19). In addition, the few 18S rRNA sequences that are present are not identical in sequence and are variably expressed during the parasite life cycle (15). As PCR sensitivity is greatly influenced by the starting target molecule copy number, a low target copy number limits the detection capabilities of these assays, especially if the parasitemia is low.

The 18S rRNA gene target also presents challenges for effective multiplex platforms. The design of multiple primers to the same target can result in primer competition and decrease the efficiency of the assay. While multiplex assays do exist for simultaneous detection of malaria parasite species (25, 32, 38), they show decreased sensitivity, particularly in detecting the minor species (20). Rubio *et al.* designed a semi-nested two-tube multiplex PCR, with an initial genus-specific amplification followed by a secondary amplification using a universal *Plasmodium* primer and species-specific reverse primers. Padley *et al.* designed a one-tube multiplex assay, using species-specific primers. However both of these methods have been shown to perform less effectively than the standard nested PCR method (20). Taylor *et al.* designed a multiplex real time platform, relying on the increased sensitivity of both novel targets and fluorescent probes. However, this assay was most effective in duplex format, and not as a true four-species multiplex.

To address the limitations of existing molecular diagnostic tools, we have mined *Plasmodium* genome sequence data and identified new target DNA sequences for improved molecular diagnostic applications. Here we detail the method used to identify these targets in *Pf* and *Pv*, and we show that they provide increased sensitivity in a single-step PCR reaction and efficacy in multiplex assays.

MATERIALS AND METHODS

Data harvesting

Assembled genome sequence data for *Pf* (3D7 strain) and *Pv* (Sal-1 strain) were obtained from PlasmoDB (release 5.5). The *Pf* genome data consist of 14 sequences (23,264,338 bp) and the *Pv* genome data consist of 2,747 sequences (27,007,990 bp). Differences in the numbers of sequences between species reflect the more advanced state of *Pf* assembly relative to *Pv*. There are 14 highly assembled chromosomes for each species, and 2,733 unassigned contigs for *Pv*.

Consensus repeat sequence (CRS) screens and copy number determination

The pipeline shown in Figure 1, and described below, was constructed using custom PERL scripts. RepeatScout (version 1.0.5, default parameters) (28) was used to identify genomic CRS. 418 *Pf* and 428 *Pv* CRS were generated. The Tandem Repeat Finder Program (TRF) version 4.0 (3) was used to eliminate CRS with internal tandem repeats that could potentially interfere with PCR amplification. Repeats containing vector sequences introduced during genome sequencing were identified by a comparison with the NCBI UniVec database (build 5.2; http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) (with WU-BLAST (blastn ver. 2.0; http://blast.wustl.edu) with an E-value cutoff of 1E-10. To ensure that targets were not also present in the human genome, CRS were compared to human genome sequences (RefSeq, Primary Reference Assembly, Build 37 version 1) with BLAST (1) (version 2.2.22, blastn), with an E-value cutoff of 1E-10. Screens were applied in parallel to all CRS. Any sequence failing a screen was removed from further consideration. A total of 165 *Pf* sequences and 331 *Pv* sequences passed all screens. All *Pf* and *Pv* CRS were compared (WU-BLAST) to all available *Plasmodium* sequence data and the results were manually inspected to ensure species specificity. To allow sufficient space for primer design and the evaluation of repeat family conservation, CRS smaller than 300 bp were not considered further. CRS were used to calculate the copy number of each repeat. Each screened repeat was used to search (WU-BLAST), against the species' genome from which it was derived. Repeat copies were required to hit to the CRS with an E-value of less than 1E-50 for *Pv*. The stringency for *Pf* was relaxed to 1E-10 because lower E-value requirements did not produce sufficient candidates for screening. A minimum distance of 100 bp between copies was required to remove potential amplification complications. Repeat families with at least 6 copies were considered for further testing, yielding a total of 21 *Pf* and 68 *Pv* candidates.

Target validation

Primers were designed to test six *Pf* and seven *Pv* CRS families. Primers were designed manually to candidate targets and screened for GC-content, melting temperature, secondary structure, and primer dimer-forming potential using Primer Explorer version 2.0 (http://primerexplorer.jp/e/). Primer pairs were optimized using gradient PCR cycling on BioRad iCycler machines to determine the optimum annealing temperature, with additional adjustments to primer concentration (concentrations from 0.25µM to 1.0µM were tested) and master mix components ($MgCl_2$ concentrations from 2.0mM – 4.0mM were tested) see below for final conditions. Primers were further tested for species specificity using laboratory cultures of *Pf* (3D7), or DNA stocks of *Pv* (SV4), *Pm, Po,* and *Pk*.

***Plasmodium* parasites**

Pf strains 3D7, W2, V1-S, Dd2, HB3, D6, and FCR3 were cultured in our laboratory. DNA stocks of *Pv* (Sal-1, SV4, and NAM/CDC)*, Po, Pm, Pk* and filter paper blood spots of additional *Pv* strains (from Thailand, N. Korea, Vietnam, India, Miami, New Guinea, S. Vietnam, and Brazil) were all provided by John Barnwell (CDC). DNA was isolated using commercially available QIAamp DNA mini kits (Qiagen, Valencia CA, USA), following the manufacturer's instructions.

Nested PCR

Nested PCR for malaria parasite detection (as described by Singh et al. (33)) was used as the standard method for comparison.

Amplification of CRS targets by PCR

Amplification of CRS targets was performed in a 25µl reaction containing 1X Taq Buffer (contains 10mM Tris-HCl, 50mM KCl, 1.5mM $MgCl_2$; New England Biolabs, Ipswich MA, USA), 4mM $MgCl_2$, 200µM each dNTP, 500nM each oligonucleotide primer, 1.25 units of Taq DNA Polymerase (New England Biolabs), and 1µl of DNA template. Oligonucleotide primers for *Pf* candidate Pfr364 and *Pv* candidate Pvr47 are shown in Table 1. Separate reactions were performed for *Pf* and *Pv* under the following cycling parameters: initial denaturation at 95°C for 2 minutes, and then 35 cycles of 95°C for 30 seconds, 57°C (for *Pf*) or 54°C (for *Pv*) for 30 seconds, and 72°C for 45 seconds, followed by final extension at 72°C for 5 minutes. PCR products were visualized by gel electrophoresis on a 2% agarose gel.

Serial dilutions of quantified parasite DNA, isolated from laboratory cultures, were used to determine the detection limits (DNA concentrations ranging from 10,000 parasites/µl to 0.01 parasites/µl were tested). Final validation of targets was performed with *Pf* and *Pv* clinical samples from Tanzania (n=91, median parasitemia 3200 parasites/µl) and Venezuela (n=96, no parasitemia data), respectively, as well as with additional geographically-diverse strains for both targets (Pfr364: *Pf* strains W2, V1-S, Dd2, HB3, D6, and FCR3; Pvr47: *Pv* isolates from Thailand, N. Korea, Vietnam, India, Miami, New Guinea, S. Vietnam, and Brazil).

Multiplex PCR

The multiplex PCR platform was optimized by gradient PCR cycling to determine the annealing temperature, with additional adjustments to primer concentrations (0.25 to 1.0µM were tested) and master mix components ($MgCl_2$ from 2.0mM to 4.0mM, dNTPs from 200µM to 400µM each, and Taq DNA Polymerase from 1.25 units to 2.5 units were all tested). Multiplex PCR for detecting *Pf* and *Pv* was performed under the following amplification conditions: in a 25µl reaction containing 1X Taq Buffer (New England Biolabs, Ipswich MA, USA; contains 10mM Tris-HCl, 50mM KCl, 1.5mM $MgCl_2$), 4mM $MgCl_2$, 400µM each dNTP, 1000nM each *Pf* primer, 600-800nM each *Pv* primer, 2.5 units of Taq DNA Polymerase (New England Biolabs, Ipswich, MA), and 1µl of DNA template. The alternate *Pf* oligonucleotide primer sequences (Table 1) were used in the multiplex assay. The *Pv* primers were the same as used in the conventional PCR described above. The reaction was carried out under the following cycling parameters: initial denaturation at 95°C for 2 minutes, and then 35 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 45 seconds, followed by final extension at 72°C for 5 minutes. All possible combinations of dilutions ranging from 10,000 parasites/µl to 0.01 parasites/µl for each species were tested. PCR products were visualized by gel electrophoresis on a 2% agarose gel.

Sensitivity and Specificity Calculations

Sensitivity and specificity (95% confidence interval) were calculated using the nested 18S rRNA PCR as the gold standard for distinguishing a true positive from a false positive (Table 2).

RESULTS

Repeat mining and screening of diagnostic candidates

A semi-automated bioinformatics pipeline was constructed for genome repeat mining and *in silico* candidate screening (Figure 1), see Materials and Methods. Six *Pf* and seven *Pv* putative targets were identified for validation. Over 50 primer pairs were designed to these targets and empirically tested in conventional PCR amplification assays and multiplex assays. Of these targets, the most effective were *Pf* candidate Pfr364 and *Pv* candidate Pvr47, as these targets consistently performed with the greatest sensitivity and specificity. The functions of Pfr364 and Pvr47 are not known. Neither sequence is annotated or protein encoding. However, regions of Pfr364 are expressed according to PlasmoDB. Full-length sequence alignments and repeat coordinates can be found in Supplemental Files S1-S2 and Supplemental Table S1.

Diagnostic targets: copy number and distribution

At least one putative target from each species was found to significantly improve existing diagnostic capabilities. Pfr364 exists in 41 copies each of which is localized to the SB2 subtelomeric repeat region found on most chromosome ends (Figure 2). The SB2 region of *Pf* chromosomes is variably sized (1-3 kb, though it may contain up to 6 kb of additional sequence) and is composed of different repeat types (9). Many regions were found to contain two proximal copies of Pfr364 and chromosome 6 contains 3 copies at it 3 end (data not shown). Multiple alignment reveals significant sub-family structure resulting in two related alignment groups, which we have designated as subfamilies 1 and 2 (Figure 3A; Supplemental File S2 and Table S1). Interestingly, when multiple copies of Pfr364 are found at chromosome ends, there is one member of each subfamily present (Figure 2).

Pvr47 is found in 14 copies (Figure 3B; Supplemental File S1 and Table S1). All members are located on contigs that have not yet been assigned to chromosome scaffolds. The majority of these members map to small (<16kb) subtelomeric contigs that could not be assembled onto chromosomes due to their repetitive nature (5). Two of these family members are located proximal to annotated *vir* genes, while a third is located proximal to the subtelomeric transmembrane protein Pvstp1 (6).

Detection of *P. vivax* and *P. falciparum*

Primers designed to Pfr364 and Pvr47 (Table 1) specifically identified *Pf* and *Pv* respectively. Other *Plasmodium* species, including *Pm, Po,* and *Pk*, were not amplified. No amplification was observed using human non-malaria DNA (Data not shown). Using known quantities of laboratory-cultured parasites, we were able to consistently detect parasites (p) in concentrations as low as 10-0.1 parasites/µl, compared to 10-1 parasites/µl detected with the standard method (Figure 4 and Table 3). *Pf* candidate Pfr364 detected between 10-0.1 parasites/µl of DNA (detected 0.1 parasites/µl twice and 10parasites/µl once). For each repeat target, single amplified products were clearly defined on a 2% agarose gel stained with ethidium bromide.

Specificity and sensitivity

The targets were further validated in three ways. First, using microscopically-determined *Pv* samples from Venezuela (n=96) and *Pf* samples from Tanzania (n=91). In comparison to standard nested 18S rRNA PCR, Pvr47 had 98.9% sensitivity and 100% specificity, and Pfr364 had 100% sensitivity and 100% specificity. Second, target amplification was assessed in 7 *Pf* strains and 10 *Pv* strains from around the world. The target was successfully amplified in each case (data not shown). Finally, PlasmoDB was queried to assess the number and distribution of SNPs in the 41 *Pf* repeats using the data of (13, 22, 40). These data represent information from 21 *Pf* strains. There are an average of 50 polymorphic sites along the ~1500 nt length of each of the Pfr364 repeats for an average of 3% each. An average of 2 different nucleotides are observed at each polymorphic position.

Multiplex assay

The multiplex PCR assay with combined Pvr47 and Pfr364 specifically detected *Pv* and *Pf* and correctly identified both single and mixed species infections. An alternative *Pf* primer was used to make the PCR products similar in size to increase efficiency (See Materials and Methods, Table 1). The limit of detection for the multiplex platform was determined using "mock-mixed" infections of *Pf* and *Pv* laboratory cultures. This method had a limit of detection of 10 p/µl for each species (Figure 5,). *Pf* DNA was also detected at 1.0 p/µl when *Pv* was present at the same concentration (*Pv* was not detected). Clinical mixed *Pf*/*Pv* samples from Venezuela (n=11) were detected with 90.9% sensitivity and 100% specificity, in comparison to the standard nested PCR method, which was performed as separate reactions for the different species.

DISCUSSION

Here we show the value of applying bioinformatics methods and mining genomic data to answer biological questions that address practical needs. This approach can be applied to additional pathogens, or to improve existing molecular diagnostic tools (LAMP, Real Time PCR, etc). Increasing the sensitivity and specificity of molecular assays will facilitate greater high-throughput detection of pathogens.

Discovery of the exact locations of Pvr47 repeats will depend on the continued refinement of the *Pv* genome assembly and improved annotation. The presence of some members near genes known to be located in subtelomeric regions (see above), combined with the known subtelomeric location of Pfr364, points to an interesting role in *Plasmodium* chromosome end biology for use in diagnostic target development. There has been no comprehensive, systematic study of the genomic repeats of the genus *Plasmodium*. Our understanding of the organization and content of subtelomeric regions is largely restricted to what is known in *Pf,* where it has been shown that these regions contain genes responsible for host immune evasion and antigenic variation (9). Given the biological importance of these regions, and the useful diagnostic targets that they contain, it is critical that we increase our understanding of their repeat content and organization.

While there is evidence for their location and distribution, the biological functions of Pfr364 and Pvr47 are not yet established. Combined with their repetitive, potentially non-genic nature, this necessitates a thorough evaluation of their robustness as diagnostic targets. Sequences with no coding potential often evolve more quickly relative to coding regions (17). However, we show that these families are highly conserved (< 3% variation at the nt level in *Pf*, indicative of selection. Further, assays designed to the targets were able to detect infections across as large a range of field isolates (7 *Pf* strains and 10 *Pv* strains) as the standard nested 18S rRNA PCR. These observations suggest that these targets are as robust to evolutionary change as the 18S rRNA target, despite the uncertainty of their biological roles.

Pfr364 and Pvr47 are not necessarily the most abundant repeats in these genomes. We tested only a handful of repetitive sequences resulting from our data mining for their potential as diagnostic targets. It is possible that more sensitive targets exist. As we've noted above, there has been no comprehensive investigation of the genomic repeat content of these organisms and our analysis is still ongoing.

Amplification of the novel targets presented here was highly sensitive and specific. Both assays have a detection limit ten-fold lower than the historic standard and utilize a single, as opposed to nested, PCR reaction. This is an important improvement, as single-round, un-nested PCR reactions have fewer steps, decrease the chances of contamination or error, decrease the overall cost in materials, and require less time to complete. The standard nested protocol requires two separate reactions, and the amplified product of the first reaction must be transferred to a second tube prior to the second reaction. Opening the tubes increases the risks of contamination and human error, and also increases the time and costs for necessary reagents and consumables.

The targets produced clean products, clearly visible on an agarose gel stained with ethidium bromide, at 716 bp and 333 bp for *Pf* and *Pv*, respectively. There were no non-specific bands in clinical samples, including negative samples, as was sometimes found with the standard PCR method (data not shown). While DNA amplified from laboratory cultures using the standard nested PCR method showed clean bands on the agarose gel, clinical samples often produced non-specific bands of similar size to the expected bands when the 18S rRNA gene-based method was used. This can be especially confusing when interpreting the results, and additional time was required to fully separate the bands by electrophoresis. The non-specific bands appeared when *Pf* samples were tested to amplify field samples with the *Pv*-specific primers (unpublished observation). Additionally, sometimes several rounds of repetition of the standard method by Singh et al. (33) were necessary to confirm the results of clinical samples tested. We found that PCR amplification with the newly identified targets yielded consistent clear results with no spurious bands among the clinical samples tested in this study.

One-step multiplex reactions will offer a great improvement to existing *Plasmodium* diagnostics. Efficient, high-throughput pathogen detection will decrease the time to results and appropriate treatment. Mixed infections naturally occur in regions where multiple parasite species are found, and present a challenge for diagnosis. To validate our multiplex method, we tested all possible combinations of varying DNA concentrations (from 10,000 parasites/μl to 1 parasite/μl) to cover all the range of naturally-occurring mixed infections (Data not shown). The limit of detection (10 parasites/μl for *Pf* and *Pv*) compares favorably to other multiplex methods. In mixed-species infections, one major species will frequently dominate over another that is present in relatively low concentrations during PCR amplification (27, 37). On the contrary, the current method detects both major and minor species of mixed infection providing another advantage of using this method for diagnosis.

In conclusion, the findings from this study demonstrate that using bioinformatics to identify novel genetic targets for diagnostic application is a valid approach. This methodology will be extended to identify additional targets from other *Plasmodium* species for diagnostic assays when the genome sequences become available. Our results demonstrate that the newly identified Pfr364 and Pvr47 targets are valuable tools to improve and simplify molecular diagnostic methods for field use.

ACKNOWLEDGEMENTS

The authors would like to thank Jatan Patel and Zubin Mehta for their bioinformatics assistance in screening putative target sequences. This work was supported by a CDC-UGA seed grant (OPHR #8212) awarded to JCK and VU. This study was supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer. AD was supported by an EID Fellowship from the Association of Public Health Laboratories and the CDC. NWL and AD (after the EID Fellowship) were supported by Atlanta Research and Education Foundation. Atlanta, GA.

REFERENCES

1. WHO (2008) World malaria report 2008. Geneva: World Health Organization.

2. Genton B, D'Acremont V, Rare L, Baea K, Reeder JC, et al. (2008) *Plasmodium vivax* and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea. PLoS Med 5: e127.

3. Barcus MJ, Basri H, Picarima H, Manyakori C, Sekartuti, et al. (2007) Demographic risk factors for severe and fatal *vivax* and f*alciparum* malaria among hospital admissions in northeastern Indonesian Papua. American Journal of Tropical Medicine and Hygiene 77: 984-991.

4. Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, et al. (2009) Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. Lancet Infectious Diseases 9: 555-566.

5. Kochar DK, Das A, Kochar SK, Saxena V, Sirohi P, et al. (2009) Severe *Plasmodium vivax* malaria: a report on serial cases from Bikaner in northwestern India. Am J Trop Med Hyg 80: 194-198.

6. Price RN, Douglas NM, Anstey NM (2009) New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. Curr Opin Infect Dis.

7. Tjitra E, Anstey NM, Sugiarto P, Warikar N, Kenangalem E, et al. (2008) Multidrug-resistant *Plasmodium vivax* associated with severe and fatal malaria: a prospective study in Papua, Indonesia. PLoS Med 5: e128.

8. Moody A (2002) Rapid diagnostic tests for malaria parasites. Clin Microbiol Rev 15: 66-78.

9. Gamboa D, Ho MF, Bendezu J, Torres K, Chiodini PL, et al. (2010) A Large Proportion of *P. falciparum* Isolates in the Amazon Region of Peru Lack pfhrp2 and pfhrp3: Implications for Malaria Rapid Diagnostic Tests. Plos One 5: 8.

10. Bronzan RN, McMorrow ML, Kachur SP (2008) Diagnosis of malaria: challenges for clinicians in endemic and non-endemic regions. Mol Diagn Ther 12: 299-306.

11. Erdman LK, Kain KC (2008) Molecular diagnostic and surveillance tools for global malaria control. Travel Med Infect Dis 6: 82-99.

12. Harris I, Sharrock WW, Bain LM, Gray KA, Bobogare A, et al. (2010) A large proportion of asymptomatic *Plasmodium* infections with low and sub-microscopic parasite densities in the low transmission setting of Temotu Province, Solomon Islands: challenges for malaria diagnostics in an elimination setting. Malar J 9: 254.

13. Snounou G, Viriyakosol S, Zhu XP, Jarra W, Pinheiro L, et al. (1993) High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Mol Biochem Parasitol 61: 315-320.

14. Padley D, Moody AH, Chiodini PL, Saldanha J (2003) Use of a rapid, single-round, multiplex PCR to detect malarial parasites and identify the species present. Ann Trop Med Parasitol 97: 131-137.

15. Rubio JM, Post RJ, van Leeuwen WM, Henry MC, Lindergard G, et al. (2002) Alternative polymerase chain reaction method to identify *Plasmodium* species in human blood samples: the semi-nested multiplex malaria PCR (SnM-PCR). Trans R Soc Trop Med Hyg 96 Suppl 1: S199-204.

16. Rougemont M, Van Saanen M, Sahli R, Hinrikson HP, Bille J, et al. (2004) Detection of four *Plasmodium* species in blood from humans by 18S rRNA gene subunit-based and species-specific real-time PCR assays. J Clin Microbiol 42: 5636-5643.

17. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, et al. (2000) Loop-mediated isothermal amplification of DNA. Nucleic Acids Res 28: E63.

18. Han ET, Watanabe R, Sattabongkot J, Khuntirat B, Sirichaisinthop J, et al. (2007) Detection of four *Plasmodium* species by genus- and species-specific loop-mediated isothermal amplification for clinical diagnosis. J Clin Microbiol 45: 2521-2528.

19. Taylor SM, Juliano JJ, Trottman PA, Griffin JB, Landis SH, et al. (2010) High-Throughput Pooling and Real-Time PCR-Based Strategy for Malaria Detection. Journal of Clinical Microbiology 48: 512-519.

20. Singh B, Kim Sung L, Matusop A, Radhakrishnan A, Shamsul SS, et al. (2004) A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363: 1017-1024.

21. Snounou G, Viriyakosol S, Jarra W, Thaithong S, Brown KN (1993) Identification of the four human malaria parasite species in field samples by the polymerase chain reaction

and detection of a high prevalence of mixed infections. Mol Biochem Parasitol 58: 283-292.

22. Rougemont M, Van Saanen M, Sahli R, Hinrikson H, Bille J, et al. (2004) Detection of four *Plasmodium* species in blood from humans by 18S rRNA gene subunit-based and species-specific real-time PCR asays. Journal of Clinical Microbiology 2004: 5636 - 5643.

23. Li J, Wirtz RA, McConkey GA, Sattabongkot J, Waters AP, et al. (1995) *Plasmodium*: genus-conserved primers for species identification and quantitation. Exp Parasitol 81: 182-190.

24. Singh B, Bobogare A, Cox-Singh J, Snounou G, Abdullah MS, et al. (1999) A genus- and species-specific nested polymerase chain reaction malaria detection assay for epidemiologic studies. American Journal of Tropical Medicine and Hygiene 60: 687-692.

25. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498-511.

26. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455: 757-763.

27. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455: 799-U797.

28. Mercereau-Puijalon O, Barale JC, Bischoff E (2002) Three multigene families in *Plasmodium* parasites: facts and questions. International Journal for Parasitology 32: 1323-1344.

29. Long EO, Dawid IB (1980) Repeated genes in eukaryotes. Annual Review of Biochemistry 49: 727-764.

30. Li J, Gutell RR, Damberger SH, Wirtz RA, Kissinger JC, et al. (1997) Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. J Mol Biol 269: 203-213.

31. Mixson-Hayden T, Lucchi N, Udhayakumar V Evaluation of three PCR-based diagnostic assays for detecting mixed *Plasmodium* infection. BMC Research Notes 3: 88.

32. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-i358.

33. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27: 573-580.

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

35. del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, et al. (2001) A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. Nature 410: 839-842.

36. Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, et al. (2007) Genome variation and evolution of the malaria parasite *Plasmodium falciparum* (vol 39, pg 120, 2007). Nature Genetics 39: 567-567.

37. Mu JB, Myers RA, Jiang HY, Liu SF, Ricklefs S, et al. (2010) *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. Nature Genetics 42: 268-U113.

38. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, et al. (2007) A genome-wide map of diversity in *Plasmodium falciparum*. Nat Genet 39: 113-119.

39. Li W-H, Graur D (1991) Fundamentals of Molecular Evolution. Sunderland, Massachusetts, USA: Sinauer Associates, Inc. 284 p.

40. Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. Applied and Environmental Microbiology 64: 3724-3730.

41. Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and Environmental Microbiology 62: 625-630.

Figure and table legends

Figure 2.1. Schematic of diagnostic target screening and development pipeline. All genomic sequences for *Pv* and *Pf* were downloaded from PlasmoDB (http://PlasmoDB.org). Data were mined for repeats using the RepeatScout algorithm to construct consensus repeat sequences (CRS) for each identified repeat family. CRS were then screened in parallel for tandem repeats, similarity to human sequences, and vector sequences. Any CRS failing these screens were removed from further consideration. CRS that were non species-specific or less than 300 bp long were eliminated. Family copy number for remaining candidates was determined via comparison of the CRS against the appropriate genome data. Candidate repeat families containing 6 or more copies, separated by at least 100 bp were considered for further testing. For additional information and clinical sample validation, see Materials and Methods.

Figure 2.2. Spatial distribution of Pfr364 family members across the 14 *P. falciparum* chromosomes. Tick marks indicate 200 kb of sequence. Pfr364 family members occur in two proximal copies at most chromosome ends. Black lines represent the outermost copies (subfamily “1”), gray lines represent the innermost copies (subfamily “2”). Chromosome 6 has three copies at its 3 end (only two are shown). Circos 0.51 (http://mkweb.bcgsc.ca/circos/) was used to generate this map.

Figure 2.3. Alignments of Pfr364 and Pvr47 family members with PCR primers. **Panel A**. Pfr364 with primers. Arrows represent locations of PCR primers in context of full alignment. The full alignment is 1,538 positions in length; here a partial alignment is shown. Vertical black lines indicate where the sequence alignment has been truncated for to enable viewing of all 4 primer locations. The alignment shows two subfamilies within Pfr364. We have designated the

upper 22 sequences as subfamily “1” and the lower 19 sequences as subfamily “2”. Forward and reverse primer pairs used for multiplex and conventional PCR are respectively the last two sequence pairs in the alignment. **Panel B.** Pvr47 with primers. Arrows represent location of PCR primers in context of full alignment. The full alignment is 1,070 positions; here only positions 433 to 776 are shown. Forward and reverse primers are respectively the last two sequences in the alignment.

Figure 2.4. Limit of detection for conventional PCR assays. Primers to novel targets *P. falciparum* Pfr364 (A) and *P. vivax* Pvr47 (B) were used to amplify parasite DNA of the appropriate species. DNA was quantified and ten-fold serial dilutions from 10,000 parasites/µl (lane 1) to 0.01 parasites/µl (lane 7) were used to determine the limit of detection. A 100 bp standard ladder (L) and no template control (NTC) were included.

Figure 2.5. Evaluation of Pfr364 and Pvr47 primers on geographically diverse field isolates. **Panel A.** Pfr364 primers tested on various *Pf* isolates. Lanes: 1.) 3D7; 2.) w2; 3.) V1-S; 4.) Dd2; 5.) Hb3; 6.) D6; 7.) FCR3. A 100 bp standard ladder (L) and no template control (N) were included. **Panel B.** Pvr47 primers tested on various *Pv* field isolates. Lanes: 1.) Thailand; 2.) N. Korea; 3.)Vietnam; 4.) India; 5.) NAM/CDC; 6.) Miami; 7.) New Guinea; 8.) Sal-1; 9.) S. Vietnam; 10.) Brazil. The Pfr364 and Pvr47 primers clearly detect all of the tested isolates. Pfr364 primers detected an additional 91/91 (100%) *Pf* isolates from Tanzania, and Pvr47 primers detected an additional 95/96 (98.9%) *Pv* isolates from Venezuela (not shown).

Figure 2.6. Multiplex PCR. The multiplex method clearly identified mock mixed *Pf* and *Pv* infections (lane Pf/Pv). Single species infections (lanes *Pf* and *Pv*) were also detected. The *Pf* band appears at 220 bp, and *Pv* band at 333 bp. A 100 bp standard ladder (L) and no template control (NTC) were used.

Table 2.1. New diagnostic target primer sequences.

Primer sequences designed to targets Pfr364 and Pvr47. The alternate primer pair for *Pf* was used in multiplex reactions only. The *Pv* primer set was the same for both single-species PCR and multiplex.

Table 2.2. Sensitivity and specificity of new PCR assays compared to standard nested 18S rRNA PCR.

Sensitivity and specificity of new PCR assays as compared to standard nested 18S rRNA PCR (33). For conventional PCR, sensitivity and specificity were calculated using 96 *P. vivax* samples from Venezuela and 91 *Pf* samples from Tanzania. For the multiplex PCR, 11 mixed infection (*Pf* / *Pv*) samples from Venezuela were used. In both cases, DNA from non-malarious patients was included as a negative control.

Table 2.3. Detection limits of new diagnostic targets.

Detection limits (in parasites/µl) calculated using 10-fold serial dilutions of *Pf* and *Pv* DNA (See Figure 3.4).

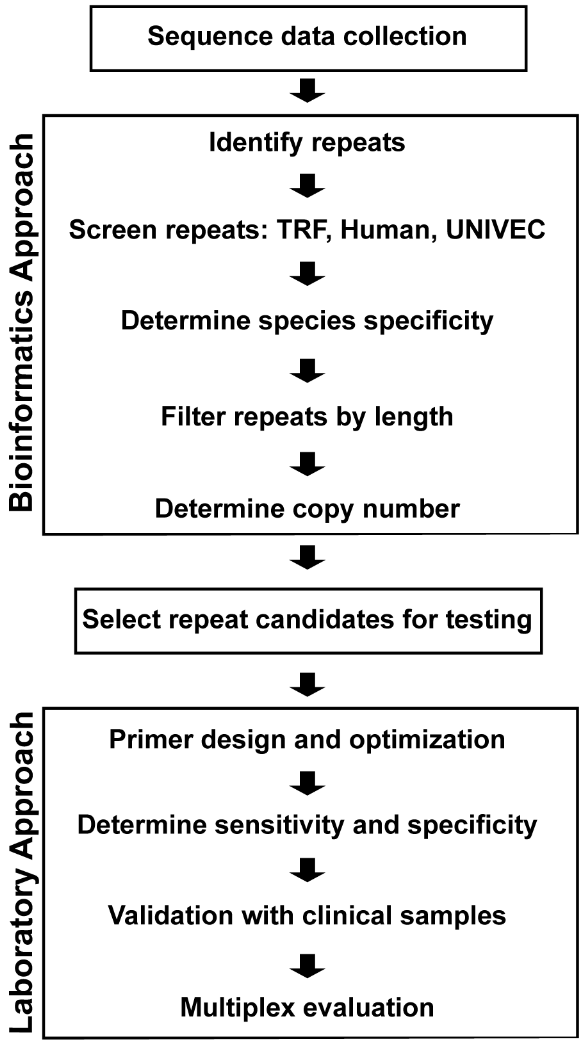


Figure 2.1. Schematic of diagnostic target screening and development pipeline.

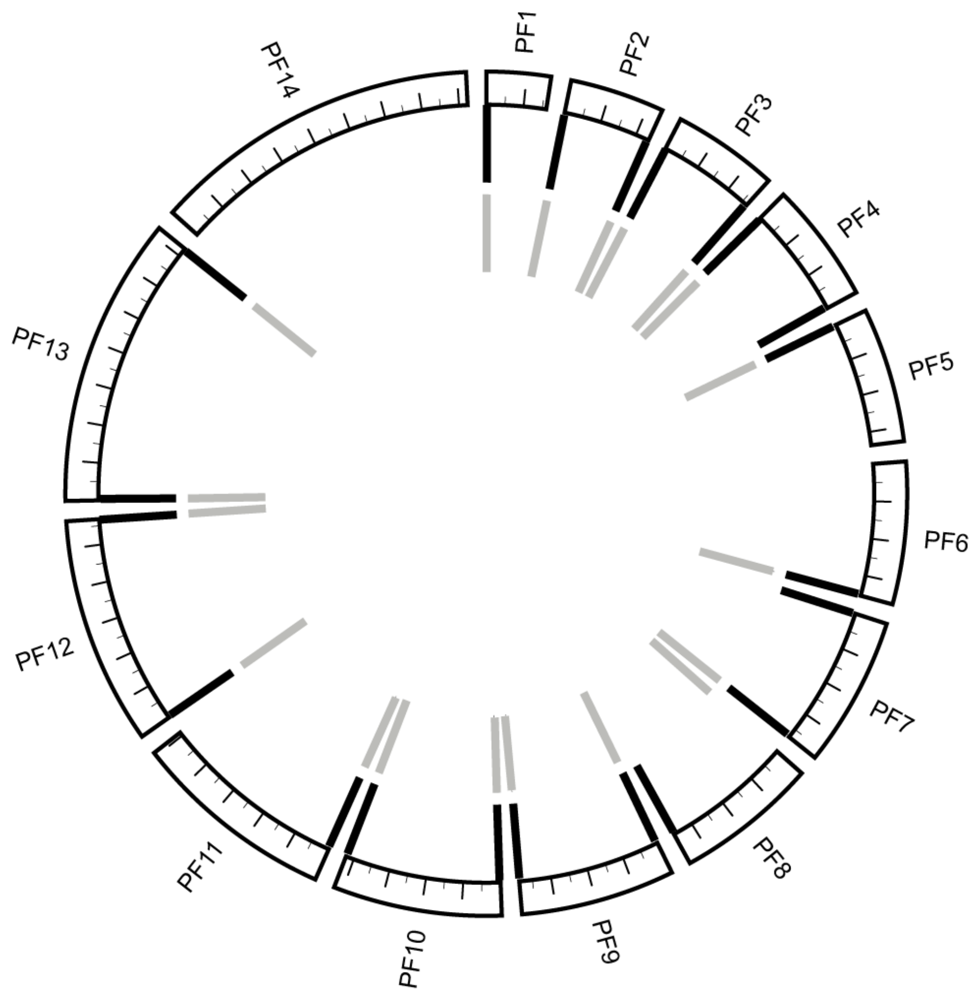


Figure 2.2. Spatial distribution of Pfr364 family members across the 14 *P. falciparum* chromosomes.

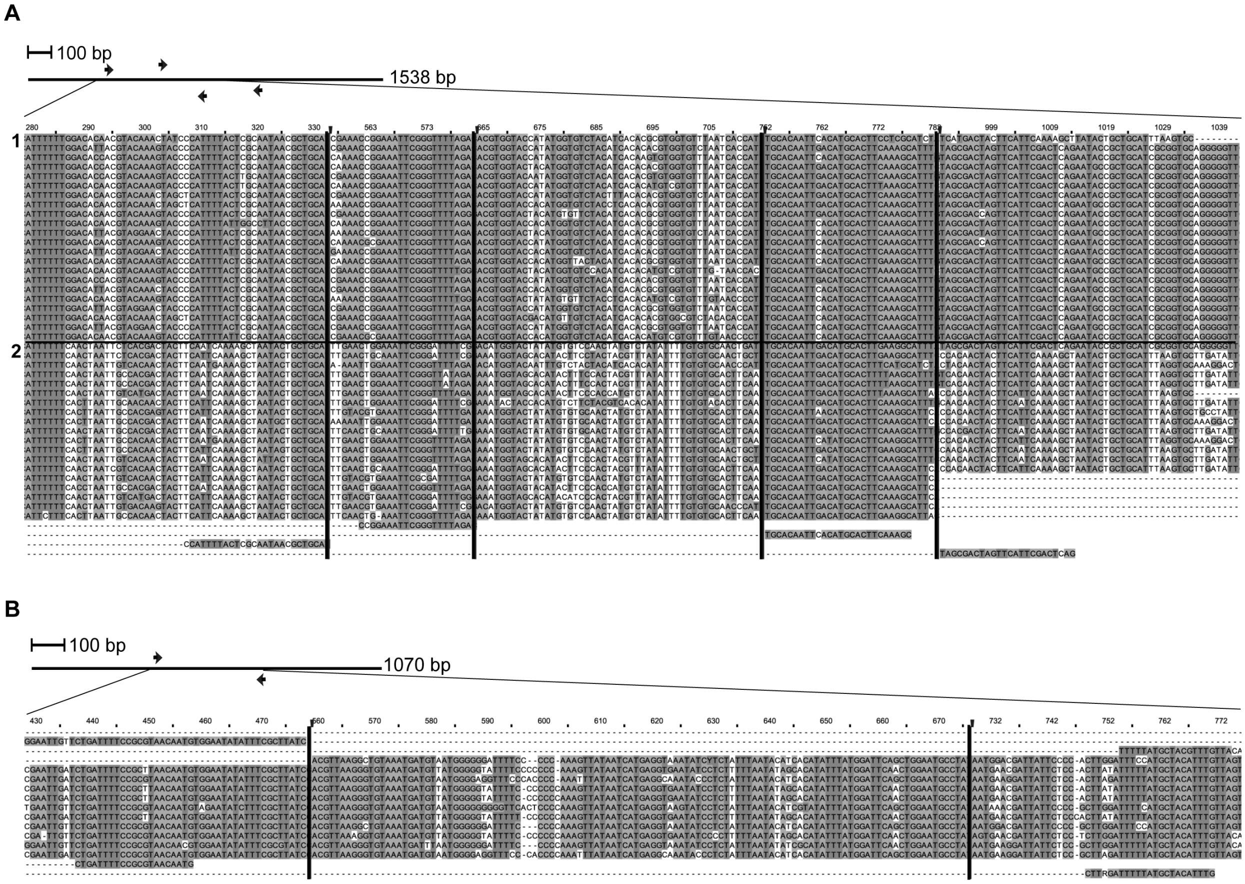


Figure 2.3. Alignments of Pfr364 and Pvr47 family members with PCR primers.

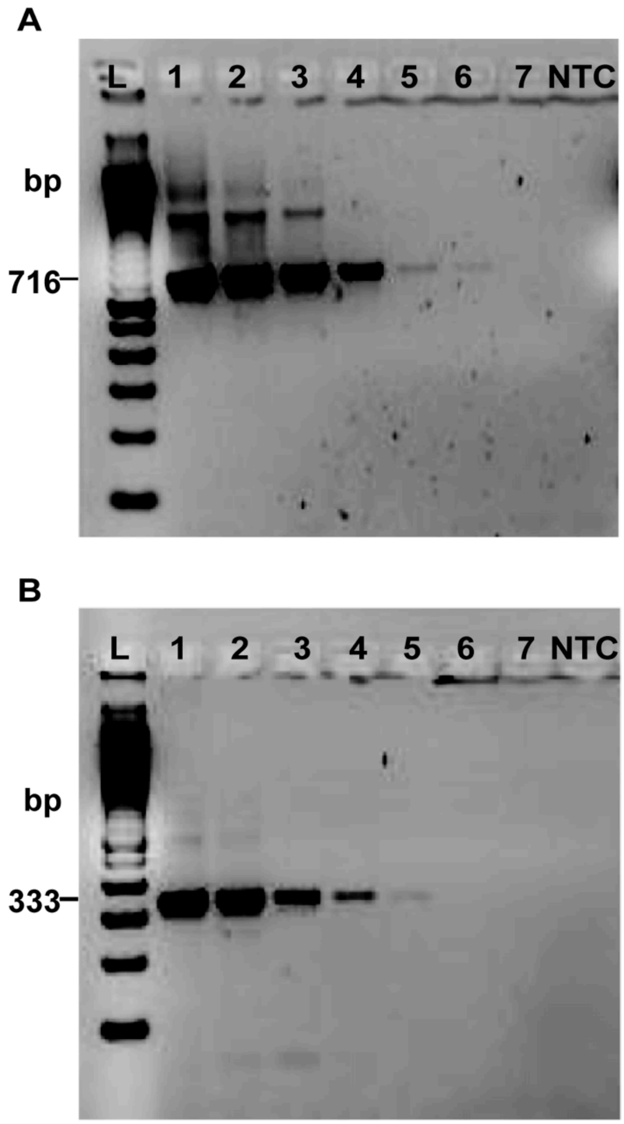


Figure 2.4. Limit of detection for conventional PCR assays.

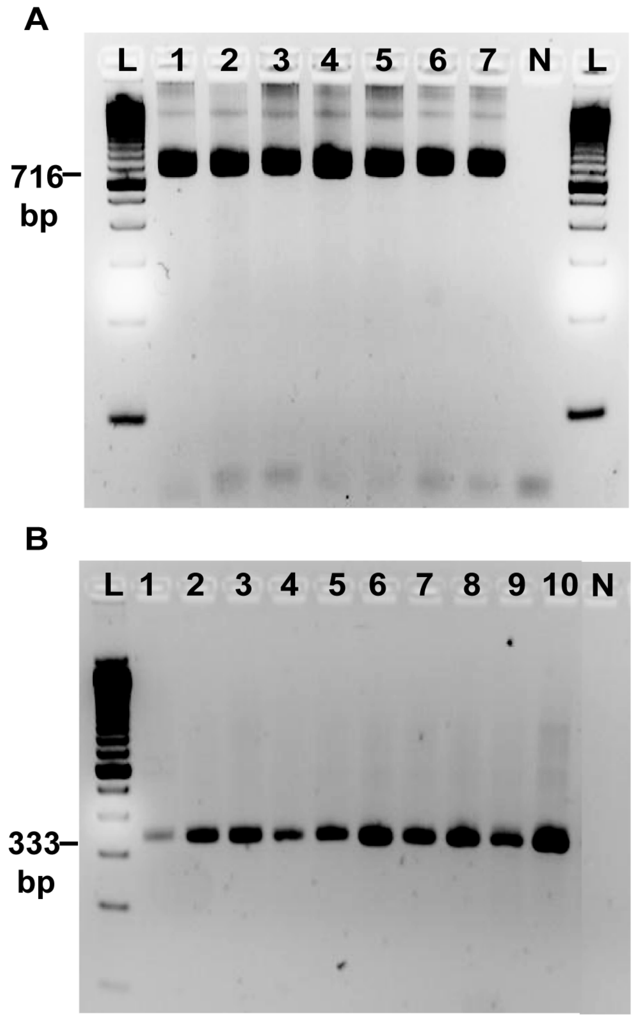


Figure 2.5. Evaluation of Pfr364 and Pvr47 primers on geographically diverse field isolates.

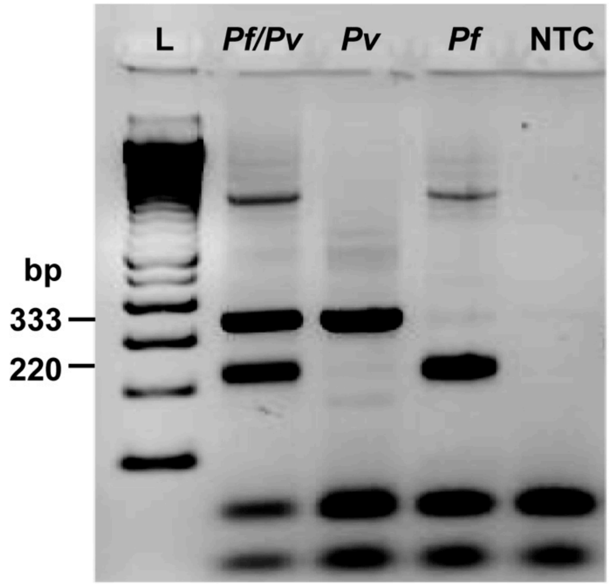


Figure 2.6. Multiplex PCR.

Table 2.1. New diagnostic target primer sequences.

Primers	*P. falciparum* Pfr364	*P. vivax* Pvr47
Forward	5′-CCATTTTACTCGCAATAACGCTGCAT	5′-CTGATTTTCCGCGTAACAATG
Reverse	5′-CTGAGTCGAATGAACTAGTCGCTAC	5′-CAAATGTAGCATAAAAATCYAAG
Alt-Forward	5′-CCGGAAATTCGGGTTTTAGAC	
Alt-Reverse	5′-GCTTTGAAGTGCATGTGAATTGTGCAC	

Table 2.2. Sensitivity and specificity of new PCR assays compared to standard nested 18S rRNA PCR.

18s rRNA nested PCR (n) ***P. falciparum***	**New primers**	
	Positive	Negative
Positive (91)	91	0
Negative (9)	0	9
Sensitivity	**100%**	
Specificity	**100%**	
18s rRNA nested PCR (n) ***P. vivax***	**New primers**	
	Positive	
Positive (96)	95	Negative
Negative (13)	0	1
Sensitivity	**98.9%**	13
Specificity	**100%**	

Table 2.3. Detection limits of new diagnostic targets.

Replicate	***P. falciparum***	***P. vivax***
Replicate 1	0.1	10
Replicate 2	0.1	1
Replicate 3	10	10

CHAPTER 3

Upstream sequence analysis of clustered post-infection expression profiles of 3281 *Cryptosporidium parvum* genes

---

Jenna Oberstaller, Sandeep J. Joseph, Mary Mauzy, Cheryl Lancto, Shin Enomoto, Mitchell Abrahamsen, Mark Rutherford and Jessica C. Kissinger. To be submitted to PloS Pathogens.

ABSTRACT

There are very few molecular genetic tools available to study the apicomplexan parasite *Cryptosporidium parvum*. The organism is not amenable to continuous *in vitro* cultivation, and purification of intracellular developmental stages in sufficient numbers for most downstream molecular applications is quite difficult. In this study, we clustered whole-genome gene expression profiles generated from 7 post-infection time points of 3281 genes. We used fuzzy *c*-means (FCM) clustering to identify genes that show similar expression patterns throughout the first 72 h of the intracellular life cycle in HCT-8 epithelial cell culture. We used the algorithms MEME, AlignACE and FIRE to identify conserved, overrepresented DNA motifs in the upstream promoter region of genes with similar expression profiles. Several DNA motifs were identified in the upstream sequences of gene clusters that might serve as potential *cis*-regulatory elements. The most highly overrepresented motifs were the E2F motif (5′-TGGCGCCA-3′), the G-box motif (5′-G.GGGG-3′), a well-documented ApiAP2 binding motif (5′- TGCAT -3′), and an as-yet unknown motif 5′-[A/C]AACTA-3′. The E2F and ApiAP2 motifs were previously documented as overrepresented in *C. parvum* noncoding regions. We generated a recombinant *C. parvum* DNA-binding protein domain from putative ApiAP2 transcription factor cgd8_810 and determined its binding specificity using protein-binding microarrays. We demonstrate that cgd8_810 can putatively bind the overrepresented G-box motif, potentially implicating this ApiAP2 in regulation of many gene clusters. This study generates valuable and much-needed insight into gene regulation and developmental gene expression in an important human pathogen.

INTRODUCTION

The apicomplexan parasite *Cryptosporidium parvum* primarily infects the microvillous border of the intestinal epithelium, and to a lesser extent extraintestinal epithelia, causing acute gastrointestinal disease in a wide range of mammalian hosts. The first case of human *Cryptosporidium* infection was reported in 1976 [1], and only seven additional cases were documented before 1982 [2]. Since then the number of cases identified has increased dramatically, largely due to the recognition of a life-threatening form of infection in patients with AIDS [3]. *Cryptosporidium* was also recently implicated as a significant pathogen contributing to diarrhea in children in developing countries [4]. In addition, seroprevalence rates of 25-35% in the United States indicate that infection with *Cryptosporidium* is very common among healthy persons [5].

C. parvum has a complex, obligate-intracellular life cycle involving both asexual and sexual developmental stages. Transmission of *Cryptosporidium* happens through the fecal-oral route where an infection is initiated by the ingestion of oocysts, which releases sporozoites capable of invading intestinal epithelial cells. The parasite's obligate intracellular developmental stages are exceedingly difficult to study, as the parasite cannot be isolated from the host cells in sufficient numbers for most downstream molecular applications. *C. parvum* is also not amenable to either continuous *in vitro* cultivation or genetic dissection [6,7].

As there are no tools to propagate *Cryptosporidium* outside of animals, genetically manipulate, or isolate large quantities of genetic material or proteins from *C. parvum* during the obligate intracellular life stages, transcriptional regulation in this parasite is largely a black box. Indeed, transcriptional regulation across the entire apicomplexan phylum is still poorly understood, though the combination of computational and bench analyses have yielded significant discoveries in parasites distantly related to *C. parvum*. Genome-wide scans for possible DNA-binding domains across the phylum revealed several families of DNA-binding proteins, but noted a significant expansion of the Apicomplexan AP2 (ApiAP2) family of transcriptional regulators [8], and subsequent experimental analyses have confirmed the regulatory roles of several of these proteins [9,10,11]. By generating several recombinant ApiAP2 proteins and testing them on protein-binding microarrays (PBMs) [12], Campbell *et al*. (2010) [13] determined binding specificities for 20 of the 27 identified members of this family in *P. falciparum*. These binding sites matched several previously determined *Plasmodium cis*-elements. Militello *et al*. (2004) computationally predicted a *cis*-regulatory element in the upstream sequences of 8 out of 18 *P. falciparum* heat shock genes (called the G-Box) and subsequently demonstrated the importance of this element through transient transfections and mutational analyses [14]. Similarly, Young *et al*. (2008) also predicted several *cis* regulatory elements by looking upstream of *Plasmodium* genes clustered based on similarity of gene expression profile (21 clusters total) and demonstrated the regulatory importance of one of the predicted elements (PfM18.1, 5′-GTGCA-3′) *in vitro* [15]. The work of Campbell *et al.* (2010) identifying specific *trans* factors that bind to these two motifs [13] indicates the power of computational methods in predicting *cis* regulatory elements in *Plasmodium*.

Computational methods have been used successfully to predict regulatory elements across the apicomplexan phylum, though unlike in *Plasmodium* we rarely know which, if any, *trans* factors bind these elements. In *Toxoplasma gondii*, Mullapudi *et al*. (2009) identified putative *cis*-regulatory elements present upstream of functionally related groups of genes and subsequently characterized the function of some of these conserved elements using reporter assays in the parasite [16]. Behnke *et al*. (2010) used tachyzoite gene expression profiles to predict regulatory elements in their upstream sequences [17]. Guo and Silva (2008) mined the non-coding sequences in two *Theileria* genomes and predicted the presence of five putative *cis*-regulatory elements [18]. Two previous studies characterized regulatory elements in the upstream sequences in *C. parvum*. They grouped genes based on function and looked for conserved DNA motifs in the promoter regions, then correlated these conserved motifs with the RT-PCR expression profiles of the genes examined [19,20].

Many classical techniques for the experimental analysis of promoters and gene expression, such as those utilizing microarrays, are not possible due to the low abundance of *C. parvum* mRNA relative to that of the host cell. These types of analyses have been successfully performed in other apicomplexan parasites including *Plasmodium falciparum* [21,22,23] and *Toxoplasma gondii* [17,24,25]. Alternate approaches to investigate fundamental gene regulatory mechanisms in this important parasite are required. We began our expression analyses several years before RNAseq became a viable option to study gene expression. However, the availability of several genome sequences [26,27] enabled the design of primers and the quantification of expression for each gene using semi-quantitative-PCR [28]. Transcriptome data provide the basis for future efforts to determine the role specific genes play in virulence and the regulatory pathways that are vital to this pathogen. Further, these transcriptome data provide the foundation for studying gene regulatory mechanisms, as these data can be used in conjunction with the genome sequence to identify putative *cis*-acting promoter elements that control the developmental regulation of the complex *C. parvum* life cycle. Transcriptome analyses have been performed in other apicomplexans including *Plasmodium falciparum* [15], *Toxoplasma gondii* [16,17] and even in *C. parvum*, though with limited expression data [19,20] as a means to investigate transcriptional regulation.

In the current study, we utilize data from a brute-force study that generated whole genome expression data for *C. parvum* using RealTime-PCR by extracting total RNA at several post-infection time points [28]. Out of 3805 annotated protein-encoding genes, expression data were generated for 3281. We normalized the data and clustered gene expression profiles using fuzzy *c*-means (FCM) clustering into groups showing similar expression patterns throughout the first 72 hours of the intracellular life cycle in HCT-8 epithelial cell cultures. We used motif-finding algorithms to identify conserved, overrepresented DNA motifs in the upstream promoter region of genes with similar expression profiles. A recombinant *C. parvum* DNA-binding protein domain from putative ApiAP2 transcription factor cgd8_810 was generated and tested on protein-binding microarrays to determine its binding specificity. We demonstrate that cgd8_810 can putatively bind the overrepresented G-box motif, potentially implicating this ApiAP2 in the regulation of many gene clusters. The additional information gleaned from our clustered *C. parvum* expression profiles allows us to more accurately predict genes that may be co-regulated by factors binding shared upstream putative *cis* elements. While we cannot directly test binding of these proteins to putative *cis* elements *in vivo* due to the experimental intractability of *Cryptosporidium*, our *in vitro* evidence that a putative *C. parvum* transcription factor can bind a predicted overrepresented motif provides support for our methodology.

MATERIALS AND METHODS

Data generation

We utilized expression data generated for 3281 *C. parvum* genes (data from [28]). Briefly, HCT8 cell infection was carried out according to previously reported conditions [29,30,31]. 2-2.5 x 10^7 oocysts were added to each culture dish at time (t) = 0 hr and incubated at 37ºC and 5% CO_2. After visual confirmation of excystation and attachment in culture (t= 2 hr), infected cell cultures were lysed in TRIzol (Invitrogen) at 2, 6, 12, 24, 36, 48, and 72 hours post infection, and RNA was isolated and Dnase-treated following manufacturer protocol. cDNA synthesis was accomplished using Superscript III cDNA synthesis kits using a modified version of the manufacturer's protocol. Real Time PCR was performed on the cDNA with 3,302 primer pairs designed to most *C. parvum* genes. 4 biological replicates of each gene for each time point were successfully obtained for 3,281 genes.

Real Time PCR (RT-PCR) data normalization

The fluorescence data from the PCR instrument was exported and fitted to a four parameter logistic curve as a function of the PCR cycle I:

$$F(c) = F_b + \frac{F_{\max}}{\left(1 + e^{\left(\frac{C_h - c}{\beta}\right)}\right)} \tag{1}$$

Where F_b is the fluorescence base, F_{max} is the fluorescence maximum, C_h is the PCR cycle at mid-point between F_b and F_{max} and β is the slope of the logistic curve.

The initial fluorescence (IO) was calculated by substitution of c=0 in the above equation (1) into the fitted curve:

$$IO = F(0) - F(b) = \left(\frac{F_{\max}}{1 + e^{\left(\frac{C_h}{\beta}\right)}}\right) \tag{2}$$

The relative transcript abundance for each gene at each time point and for each replicate was obtained by normalizing the IO values (obtained from (2)) of a gene to that obtained from 18s rRNA [6,32] IO values.

In order to get a representative measure of the transcript abundance for each gene at a time point, we took the median of the four (replicates) normalized IO values for each gene at a time point. We standardized this representative normalized IO expression value to the maximum expressed time point for each gene, in a modified ΔΔCt fashion [33,34].

Cluster analysis

In order to identify likely groups of co-expressed genes, two clustering algorithms, Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) and Fuzzy c-means (FCM) clustering methodologies were implemented using the normalized and standardized expression data obtained from real time PCR.

The HOPACH method combines the strengths of both partitioning and agglomerative clustering methods and was implemented using the HOPACH package [35] available from the Bioconductor repository [36]. Euclidean distance was used as the distance metric. The HOPACH algorithm uses the median silhouette (MSS) criteria [37] to automatically determine the main clusters. The main purpose of implementing this clustering procedure was to estimate the number of clusters inherent in the data. FCM, the soft partitioning clustering method, was implemented using the Mfuzz package [38], which is based on the open-source statistical language R and available from the Bioconductor repository. The FCM clustering algorithm requires two main parameters (c, the number of clusters, and m, the fuzzification parameter) and uses Euclidean distance as the distance metric. FCM assigns to each gene expression vector a membership value in the range [0,1] for each of the c clusters. The membership value indicates how well the gene expression vector is represented by the cluster to which it is assigned. Large membership values indicate high correlation of the gene expression vector to its cluster center. The FCM algorithm iteratively assigns the gene expression vector to the cluster with the nearest cluster center while minimizing an objective function. The fuzzification parameter, m, plays an important role in deriving robust clusters that are not greatly influenced by noise and random artifacts in the data.

If m is increased, poorly classified gene expression vectors which have small cluster membership values contribute less to the calculation of cluster centers. Two other parameters, e, the minimal change in the objective function for terminating the clustering process and T_{max}, the maximal number of iterations, are also specified. In this study, we specified the default value for e (0.001) and for T_{max} (100,000 iterations).

In order to select the optimal values of c and m, we used a combination of heuristics as well as a data driven approach by implementing FCM while increasing c and m. We performed separate FCM cluster analysis by gradually increasing c from 50 to 250 in increments of 50 (c= 50, 100, 150, 200 & 250) and specifying m = 1.05, 1.15, 1.25, 1.35, 1.45 & 1.55. For each FCM cluster analysis, we determined the overall mean of the membership values of a particular FCM cluster analysis (a single combination of c and m). We noted the number of genes included in clusters (not all genes cluster under all conditions) and the largest and smallest cluster size for each of the FCM cluster analysis.

Biological process GO term enrichment of each the clusters were tested using the GOEAST tool [39] assuming our experiment being a customized microarray platform. The p-value of GOID enrichment was calculated as the hypergeometric probability of getting so many genes (number of genes in each of the clusters) under the null hypothesis that they were picked out randomly from the total pool of 3281 genes. In order to control error rates for multiple hypothesis testing, the p-values were adjusted using Benjamini Hochberg method [40], where a false discovery rate (FDR)-adjusted p-value < 0.15 was considered significant.

Upstream sequence analysis

Generation of an upstream sequence database

Whole genome sequence (v 4.2) and gene-predictions of the all protein-encoding genes for *Cryptosporidium parvum* were obtained from CryptoDB (http://cryptodb.org/cryptodb/). Scripts were written in Perl to extract the upstream sequence. We defined the upstream region of a gene as 1kb of sequence upstream of the ATG, or until a gene is encounteredon the same strand, or on the opposite strand, whichever sequence length is smaller. To exclude the possibility of including coding regions in this set due to mis-annotation, a BLASTX was performed against the NCBI NR database using the set of upstream sequences as the query. Upstream sequences that contained significant portions of 100% identity to coding sequences were pruned.

Identification of conserved motifs in upstream regions of genes present in each cluster

Upstream regions of the genes present in each cluster were analyzed for *de novo* patterns using 3 pattern finding algorithms: 1) Multiple EM for Motif Elicitation (MEME) [41]; 2) AlignACE [42] and 3) Finding Informative Regulatory Elements (FIRE) algorithm [43].

MEME was run using the parameters minw =7, maxw=20, in two modes (zoops & anr) and the significant motifs (E-value >= 1e-01) for each cluster were examined. A background model is used by MEME to calculate the log likelihood ratio and statistical significance of the motif. The model used in this study was a zero-order Markov chain derived from all the non-coding sequences of *C. parvum*.

The AlignACE Gibbs-sampler motif finding algorithm parameters were set to 7 aligned columns, 10 expected sites and GC%=27 (the background GC frequency of all the upstream sequence for *C. parvum)*. We used the motif comparison tool, STAMP [44] to compare the motifs identified by MEME and AlignACE . Those motifs that have a STAMP E-value less than 1e-05 were considered to be similar.

FIRE, the *de novo* motif discovery program was implemented by specifying the motif seed length, *k* being 5, 6, 7 and 8. Those motifs (statistically significant with a z-score > 4.0) on a robustness index ranging from 1 to 10 and also present in at least 60% of the upstream sequences of a cluster were considered significant in this study.

Identification of conserved motifs in upstream regions of functionally related genes

Upstream regions for each of nine *Cryptosporidium* oocyst wall protein (COWP) genes, 105 genes belonging to clusters 7, 44 and 162 peaking primarily at 72 hours post-infection, and 68 *P. falciparum* and 60 *C. parvum* ribosomal protein genes were separately mined for overrepresented motifs using MEME (max motif width 12bp, 5 motifs max, mode = anr). Similarity of motifs to each other was determined via the STAMP tool [44].

ApiAP2 domain binding site determination

N-terminal GST fusion proteins were made as previously described [13], using the pGEX4T-1 vector (GE Healthcare) and the predicted AP2 domains and flanking residues from cgd8_810 (residues 543-676) and the previously examined domain cgd2_3490 (residues 299-463) as a control [45]. Many flanking residues were included to ensure capture of the domain. The domain and flanking sequence were PCR-amplified and cloned into the BamHI restriction site in pGEX4T-1. Proteins were expressed and purified as in [45]. Briefly, *E. coli* BL21 (RIL Codon PLUS, Stratagene) cells were induced with 200 mM IPTG at 25C. Proteins were then purified using Uniflow Glutathione Resin (Clontech) and eluted in 10mM reduced glutathione, 50mM Tris HCL, pH 8.0. Proteins were verified with western blots using an anti-GST antibody

(Invitrogen), and purity was verified by silver stain. A minimum of two protein-binding microarray experiments were performed with each purified protein construct to determine their binding specificities as previously described [13,45]

RESULTS

Real Time PCR gene expression data normalization

Available relative transcript abundance data for 3281 genes (data from [28]) were normalized to 18S rRNA and standardized as described in Materials and Methods. The normalized expression profiles of all the 3281 genes were sorted according to peak expression at each time point (Figure 1A). There is a cascade of tightly regulated expression of genes across the 72-hour intracellular life cycle of *C. parvum*.

Determining co-expressed genes using cluster analysis

The underlying assumption of putative *cis*-regulatory element discovery is that many co-expressed genes (genes that have similar expression profiles) are likely controlled by common regulatory elements. In order to identify tightly clustered groups of co-expressed genes, two clustering algorithms, Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) and Fuzzy c-means (FCM) clustering methodologies were implemented using the normalized and standardized expression data obtained from semi-quantitative real time PCR. To identify putative *cis*-regulatory elements for these clusters, we searched the upstream regions of all genes in a group/cluster for conserved, overrepresented sequence motifs. One of the major challenges in cluster analysis is to determine the number of clusters present in a given dataset. Most clustering methods are restricted to a one-to-one mapping scheme where one gene is assigned to only a single cluster, known as hard clustering (examples are *k*-means, Self Organizing Maps (SOM) and hierarchical clustering), while soft clustering (such as FCM) can assign genes with a metrics (membership) value indicating the strength of its association with a cluster (see Materials & Methods). Moreover, it is important to have tight clusters of gene profiles that are strongly associated with each other to be most informative for identifying putative *cis*-regulatory elements. The FCM "fuzzification" parameter, *m*, determines the influence of noise (genes that do not tightly fit the expression pattern of the cluster) on the cluster analysis. For *m*=1, FCM will be equivalent to *k*-means clustering. Increasing *m* reduces the influence of genes with low membership values, which are most likely those genes that are only loosely associated with a cluster. One can assess the stability of clusters by tracking the variation of membership values as *m* and cluster number are increased. Considering inherent biological properties of gene expression as well as the importance of identifying tight and stable clusters, we thought soft clustering using FCM would be the most appropriate method for this study.

Determination of the optimal FCM parameter set

As indicated above, setting the appropriate value for two major parameters, *c* and *m*, is crucial to identifying appropriate clusters. Our initial effort to determine the optimal number of clusters using HOPACH cluster analysis resulted in 207 main clusters, of which 124 clusters contained more than two genes (Data not shown). Results of additional FCM clustering by increasing *c* and *m* (see Materials & Methods) are also shown in Table 3.1. For all analyses with minimal *m, m*=1.05, almost all genes were included in the clustering process, particularly for *c*=150, 200 and 250. This is equivalent to hard clustering, and false positives in clusters are more likely. The highest membership values were obtained for the analysis with *m*=1.05 and increasing values of *c*, where there were corresponding increments in the overall membership values. As *m* was increased, the number of genes included in clusters decreased (any genes with membership values < .5 were excluded). There is also a gradual reduction in the overall average of the membership value for each FCM analysis as *m* increases, indicating fuzzification influences the membership values of genes, and very similarly expressed genes that form stable clusters will be least affected as *m* is increased. For smaller *c* values, there were larger cluster sizes, but as *c* was increased those main clusters split into smaller clusters (sub-clusters). An ideal parameter set would allow sufficient fuzzification while also including an optimal number of genes in the analysis. By tracking the number of genes included in clusters and the range of cluster sizes for each of the FCM cluster analyses (Table 3.1), we estimated the ideal parameter set would be one of the four combinations of $m = 1.15$ or 1.25, and $c = 150$ or 200. In order to fix the optimal parameter set, we looked for the significant presence of the core motifs of three previously predicted *C. parvum cis*-regulatory elements [20,48] in the upstream sequences of the genes clustered by the four possible FCM analyses. We performed MEME analysis on the upstream sequences of all clusters (150 and 200) and tracked the number of clusters with significant presence of the three core motifs (5′-GCATGC-3′; 5′-GGCGGG-3′; and 5′-GGGGGG-3′). The parameter set m=1.25 & c=200 produced the most clusters wherein all three core motifs were conserved and overrepresented in upstream regions relative to other FCM parameter combinations.

All 200 expression profiles generated using FCM cluster analysis were sorted by peak expression at each time point and are displayed in heatmap format, where each row represents a cluster (Figure 3.1B). Representative expression profiles for each of these clusters recapitulate the tightly regulated expression cascade of all 3281 genes across the 72-hour intracellular life cycle of *C. parvum*. Seventy-four clusters showed at least one biological process GO term enrichment based on the hypergeometric statistical test. Not all genes have predicted GO terms, which explains the limited number of clusters with significant GO term enrichment. This reflects the lack of available experimental data in *C. parvum* relative to other apicomplexan parasites. We predicted at least one conserved and significantly over represented DNA motif in the upstream regions of genes in 198 of 200 clusters.

Conserved DNA sequence motifs and their possible biological relevance

Using three *de novo* pattern-finding algorithms, MEME, AlignACE and FIRE, we mined the upstream region (see Materials & Methods) of all genes present in each of the 200 clusters identified in this study. Twenty-five statistically significant conserved motifs were identified by at least one of the three algorithms (Table 3.2). All the three pattern-finding algorithms identified motifs 1, 2 & 3, while only MEME and AlignACE identified motifs 4, 5 and 6. Motifs 7 to 25 were identified by FIRE only. We observed a disparity in the different types of motifs identified across all the algorithms; some motifs were identified by all the methods, while other motifs were identified by only one or two algorithms, a finding explained by the differences in these algorithms' underlying assumptions. MEME and AlignACE discover degenerate motif candidates using an expectation maximization strategy and Gibbs sampling, respectively, from a set of sequences. FIRE uses model-independent mutual information and continuous (e.g., expression log ratios from a single microarray experiment) or discrete (e.g., a clustering partition) data to identify motifs. Due to the theoretical similarity behind the MEME and AlignACE motif discovery methods, there should be a correlation between the motifs identified by them. This was exactly what we observed. The first six motifs (motifs 1 to 6) were identified by both MEME [41] and AlignACE [42] . One of the possible limitations of FIRE is that it may overlook certain highly degenerate motifs, as it initially begins by searching non-degenerate motif representations [43]. Perhaps for these reasons, FIRE did not identify motifs 4, 5 & 6, nor was there a consensus between FIRE and the other two algorithms concerning all clusters identified as having overrepresentations of motifs 1, 2 and 3.

Overrepresented motif families

We further divided the 25 identified motifs into motif families based on sequence similarity (Table 3.2). Motifs 1, 7, 8, 11 and 23 are highly similar to the ApiAP2 binding site 5′-GCATGCA-3′, a well-documented motif in Apicomplexa. We have designated it "AP2_1". The AP2_1 motif was previously noted to be overrepresented in the noncoding regions of *C. parvum* chromosome 6 [49], the only part of the genome the authors examined. It was also previously identified as a potential *cis*-regulatory element in *C. parvum* [19] in the upstream sequences of a subset of glycolysis pathway genes. De Silva et al. (2008) showed that orthologous ApiAP2 proteins from *P. falciparum* (PF14_0633) and *C. parvum* (cgd2_3490) both bind the 5′-TGCATGCA-3′ core sequence [45]. The AP2_1 motif is known to be enriched upstream of *P. falciparum* sporozoite-specific genes, which suggested a role in sporozoite-specific transcriptional regulation. Yuda et al. (2010) subsequently proved that the *Plasmodium berghei* ortholog of ApiAP2 PF14_0633 (PBANKA_132980) binds the AP2_1 motif and is essential for regulation of sporozoite-specific genes [10,15]. Outside the *Plasmodia*, this motif is also overrepresented in the non-coding regions of other apicomplexan parasites, including *T. gondii* (TRP-2 motif) [16] and *E. tenella* [49]. In this study, 55 clusters of co-expressed genes were predicted to have statistically significant overrepresentation of the AP2_1 motif in the upstream regions of their genes (Figure 3.2). The majority of these clusters have little expression at 2, 6, and 24 hours post-infection. We investigated the possible biological relevance of these gene clusters using hypergeometric tests for biological process GO term enrichment. Glycolysis, cellular polysaccharide metabolic process, carbohydrate metabolism, post-translational protein modification, protein phosphorylation and regulation of biological quality are all significantly enriched .

Motifs 2 and 6 (5′-G[T/G/A]GGGG-3′) identified in this study are very similar to the G-box motif previously reported in *C. parvum* in the upstream region of a sub-set of genes involved in DNA metabolism, as well as 8 out of 18 *P. falciparum* heat shock genes and 11 out of 12 *C. parvum* heat shock genes [20,48]. Motif 6 is significantly overrepresented in the upstream sequences of the genes in 16 clusters (Table 3.2). While MEME analysis identified over representation of this motif in the upstream regions of the genes in all 16 clusters, AlignACE detected the same in 7 clusters while FIRE analysis did not detect the G-box upstream of any of the clusters. Expression profiles for 8 out of the 12 *C. parvum* heat shock genes were grouped into 8 different clusters. Only one indicated the significant presence of the G-Box motif in the upstream sequences of the genes in that cluster. Promoter regions of the genes contained in the remaining clusters contained G-box motifs but their presence was not statistically significant within their respective clusters. PBM results for putative *C. parvum* ApiAP2 transcription factor cgd8_810 indicate it binds the G-box motif (Figure 3.3A). G-box-like motifs are overrepresented in the upstream sequences of 54 *C. parvum* gene expression clusters (Figure 3.3B and 3.3C), and again we note that these clusters are for the most part not active 2 hours post-infection. Some of the important biological process GO terms enriched in the these gene clusters were DNA packaging, nucleosome organization, organophosphate metabolic process, alcohol metabolic process, mRNA metabolic process, ubiquitin-dependent protein catabolic process, phospholipid biosynthetic process, membrane lipid biosynthetic process and DNA metabolism. Out of 54 clusters, promoter regions of genes present in 16 clusters have both

AP2_1-like and G-box-like motifs conserved, which suggests the possibility of joint involvement in regulation of these genes (Table 3.4).

Motif 3 (core sequence pattern 5′-[C/G]GCGC[G/C]-3′) and motif 4 (core sequence pattern 5′-GGCGGG-3′) are highly similar to the binding site of the E2F-DP transcription factor (TF), which represent an important class of TFs that function as major regulators of the cell cycle and apoptosis [50]. The E2F motif was previously noted to be overrepresented in the noncoding regions of *C. parvum* chromosome 6 [49], though it was not identified as an E2F motif. E2F transcription factors have been studied extensively in a broad range of organisms including higher eukaryotes, such as mammals [51], worm [52], frog [53], fly [54] and plants [55]. The typical conserved sequence of the E2F/DP binding site is 12 bp in length, which consists of a 6 bp CG core flanked by T and A-enriched sequence. This conserved central CG motif ([C/G]GCGC[G/C]) is symmetric, and amino acids that contact these bases are conserved amongst all known E2F and DP proteins [50]. Ramirez-Parra *et al*. (2003) found that consensus motifs 5′-TTTCCCGCC-3′ and 5′-TTTGGCGGG-3′ are the most abundant motifs in the *Arabidopsis* genome, and these sites were previously known to be able to direct binding of E2F/DP [56]. In *C. parvum*, Templeton *et al*. (2004) reported the existence of two E2F/DP winged-helix DNA-binding domain transcription factor pairs not found in *P. falciparum* [57,58]. However, the specific roles these TFs play in *C. parvum* are unknown. Out of 200 clusters, overrepresented E2F motifs were found upstream of genes present in 163 clusters, making E2F-like motifs the most abundant putative TF binding sites in *C. parvum* (Figure 3.4). Clusters containing overrepresented E2F-like motifs in their upstream regions do not show any particular expression patterns and genes with peak expression can be observed at all examined time points. *C. parvum* possesses two putative E2F transcription factors (Table 3.3) that are maximally expressed at 2 and 12 hours post-infection, though they are expressed at some level at all time points. We find that in 45% of 20 clusters containing upstream overrepresented E2F motifs maximally expressed at 2 hours, E2F is the only overrepresented upstream motif. These data suggest that E2F regulation could be sufficient to drive expression of this subset of clusters. GO enrichment analysis revealed that these clusters are statistically over-enriched for a number of biological processes, including structure-specific DNA binding, gene expression, translation, DNA metabolic process, response to DNA damage stimulus, DNA repair, regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, RNA processing, RNA binding, ribonucleoprotein assembly, nucleocytoplasmic transport, golgi vesicle transport, cell redox homeostasis, establishment of protein localization to lipids, secretion by cell, lipid transport, carbohydrate transport and glycolysis. E2F-like motifs have previously been found overrepresented *in C. parvum* at the promoter regions of subsets of genes associated with DNA replication and glycolysis [19].

Motif 5, which is rich in G's and A's, is similar to the 5′-GAGA-3′ motif identified in *Drosophila* [59], as well as the NTBA (NucleoTide Biosynthesis-A) motif identified upstream of the nucleotide biosynthetic genes in *T. gondii* [16]. No proteins similar to the known GAGA-binding family of transcription factors are annotated in the *C. parvum* genome. The GAGA motif was found significantly overrepresented in the upstream regions of genes in 12 co-expressed gene clusters in the current study (Table 3.4). Important GO terms (biological process) enriched in these clusters are nucleosome assembly, DNA dependent transcription initiation, protein dephosphorylation and ubiquitin-dependent protein catabolic process.

Motifs 13, 16 and 20 are similar to the CAAT-box (5′-CAAT-3′) motif (CITE). The *C. parvum* genome contains 2 putative CAAT-binding transcription factors (Table 3.3). These motifs are overrepresented in the upstream regions of genes present in 21 gene clusters that are maximally expressed across the life cycle (Figure 3.6), with biological process enrichments in processes including ATP synthesis coupled proton transport, intracellular protein transport, oxidative phosphorylation, cytoskeletal organization and microtubule-based movement.

The remaining 12 motifs fall into various families that do not appear to be significantly similar to known regulatory motifs. Motif 14, with the A-rich core 5′-[A/C]AACTA-3′, is the second-most overrepresented motif in the upstream regions of the genome, found upstream of genes in 122 of 200 clusters. These clusters are maximally expressed at all time points across the life cycle (Figure 3.7). This motif appears in conjunction with many different motifs upstream of clusters with very different expression profiles (Table 3.4). The ubiquity of this motif and the wide variation between combinations of motifs and expression profiles makes it very difficult to attribute any particular expression pattern to motif 14.

Motifs 10, 12, 15, 17, 18, and 19 comprise the family we term "Unknown set 1" and do not appear similar to any known *cis*-regulatory motifs. Several of these motifs are overrepresented upstream of single clusters (Figure 3.8). Motif 17 was found only in the promoter region of the genes present cluster 6, which is enriched with ribosomal proteins. Biological process GO enrichment analysis of the genes in that cluster revealed highly significant enrichment of GO terms associated with gene expression, translation, translational elongation and tRNA aminoacylation. Motifs 18 and 19 were each found conserved in the upstream regions of genes in two separate clusters, 21 and 143 respectively. The genes in cluster 21 did not show any specific GO-enrichment. Genes in cluster 143 were enriched with ubiquitin-dependent protein catabolic process and protein dephosphorylation GO terms.

Motifs 9 and 24 comprise the "Unknown set 2" motif family, with the consensus sequence 5′-[C/T][C/T]T[A/G]CA-3′. Unknown set 2 motifs are found upstream of 15 clusters maximally expressed across all time points (Figure 3.9). Motifs 21, 22 and 25 are unrelated, do not have overt similarity to any known *cis*-regulatory motifs, and are overrepresented upstream of 3, 5, and 12 clusters respectively. Again, clusters containing these overrepresented motifs upstream are expressed at any of the tested time points (Figure 3.10). Major biological process GO terms found enriched in these clusters include nucleosome assembly (motif 21), translation (motif 22), regulation of gene expression, DNA-dependent transcription initiation, post-translational protein modification and protein dephosphorylation (motif 25).

Evidence for biological relevance of select clusters and motifs

Ribosomal proteins

To evaluate the biological relevance of our clustering methods, we more closely examined several groups of functionally related or timepoint-specific genes, starting with ribosomal proteins. We examined expression data for 68 of *C. parvum*'s 81 predicted ribosomal proteins and 68 intraerythrocytic stage-expressed *P. falciparum* ribosomal proteins (*P. falciparum* expression data from [21,22,23]). Sixty of 68 *C. parvum* ribosomal proteins clustered into 22 groups; 8 had expression profiles too dissimilar to be clustered. The majority of ribosomal proteins have a bimodal expression pattern, peaking at both 6 and 24 hours, which corresponds to stages in the life cycle thought to be translationally active [28]. Ribosomal proteins have been documented to be tightly co-regulated in other organisms such as yeast [60], and their grouping together in our analyses in very similar clusters gives support that our clustering methodology is biologically relevant.

Upstream sequence analysis of potentially co-regulated ribosomal proteins indicates that E2F-like and GAGA-like motifs are overrepresented (Figure 3.11), and we confirm the presence of the G-box motif that was previously noted upstream of *P. falciparum* ribosomal proteins [61]. Campbell et al. (2010) identified the G-box binding ApiAP2 transcription factor PF13_0235 as the putative regulator of *P. falciparum* ribosomal proteins [13], noting that the mRNA expression profiles of this protein correlated very tightly with ribosomal protein expression. The G-box motif is also conserved upstream of three other *Plasmodium* species' ribosomal genes, as well as piroplasm ribosomal genes [61]. The putative E2F transcription factor expression profiles do not closely correlate with the expression of these *C. parvum* ribosomal proteins, though E2Fs are expressed as some level at all time points. There are no predicted *trans* factors for the GAGA-like motif in *C. parvum*. *T. gondii* ribosomal proteins were found to have the AP2_1-like motif overrepresented upstream (referred to as TRP-2 in *T. gondii*) [62]. The overrepresentation of different motifs upstream of ribosomal protein regulons across the phylum raises the possibility that there may have been multiple transcription factor substitutions in ribosomal protein transcriptional regulation over time.

Cryptosporidium oocyst wall proteins (COWPs)

COWP genes form two distinct classes based on expression profile: one class (4 genes) which peaks at 48 hours, then declines to 72hours, which we have termed Class I; and the other class (5 genes) which rises steadily from 36 hours to peak at 72 hours (Class II) (Figure 3.12). Though subclasses of COWPs have not been previously described, our expression data generally fall in line with what has previously been shown for COWPs [63] with the exception of COWP1 and COWP6, which belong to the respective opposite subclass. Three E2F-like motifs, one

GAGA-like motif and one motif with the consensus 5′-GCACAC-3′, similar to several *P. falciparum* ApiAP2 binding sites are overrepresented upstream of Class I COWP genes, which we have designated "AP2_2". Class II COWP genes share the E2F motif overrepresented upstream, but otherwise have very different motifs: AP2_1-like motifs, a CAAT-box-like motif, and an unknown motif with the consensus 5′-A[T/A]G[T/A]GGA.A-3′ which is not similar to any of our 11 overrepresented motif families.

Transcripts peaking at 72 hours post-infection

As discussed in the introduction, *C. parvum* does not complete its entire life cycle *in vitro*. Culture fails from 72 to 96 hours post-infection. We examined clusters peaking only at 72 hours to study what the parasite is doing at this critical timepoint. No GO-terms are over-enriched for genes in clusters 44 or 7, though genes involved in proteolysis and carbohydrate metabolic process are over-enriched in cluster 162. AP2_1-like, E2F-like, AP2_2-like and G-box-like motifs are over-enriched upstream of genes in these clusters. Known AP2_1-binding ApiAP2 cgd2_3490 is maximally expressed at 72 hours post-infection, as is the G-box-binding ApiAP2 cgd8_810 we present in this paper. No AP2_2-like binding proteins have been indicated in *C. parvum*, but it is reasonable to believe that ApiAP2s orthologous to the CACACA-binding ApiAP2s in *P. falciparum* could also bind this motif, given the conservation of binding site found between another *P. falciparum*/*C. parvum* ApiAP2 ortholog pair [45]. These results suggest ApiAP2 proteins are regulators in the late stages of the parasite life cycle.

DISCUSSION

Little is known about transcriptional regulation in apicomplexans in general and *Cryptosporidium* in particular, though recent studies in *Plasmodium* and *T. gondii* begin to suggest the tremendous complexity of transcriptional regulatory mechanisms in these parasites [13,45,64]. Studies in *C. parvum* are further crippled by the lack of a continuous culture system and the lack of tools for genetic manipulation. In this study, we have used bioinformatic tools in conjunction with the *C. parvum* transcriptome and genome sequence to advance our understanding of regulatory mechanisms in this experimentally intractable parasite.

Clustering of gene expression profiles is commonly used to reveal patterns of gene regulation. Such analyses provide valuable information regarding which genes are expressed at a particular time point/stage of the life cycle. Mauzy et al. (2012) used the DIANA algorithm available in the "cluster" package in R [46,47] to cluster the 3,281 *C. parvum* genes into nine groups based on similarity of expression profile [28]. These large clusters, consisting for the most part of hundreds of genes each, allowed them to observe general trends of genes expressed at each stage of the life cycle. Among other findings, they note that transcripts expressed at each time point make biological sense in the context of what is known about *C. parvum* life cycle stages. For example, genes involved in protein synthesis and degradation, nutrient availability, and ribosome biogenesis are highly expressed in the trophozoite stage (~6 hours post-infection), where the parasite is growing, absorbing nutrients and preparing for the first round of cell division. While these observations are certainly useful for a global understanding of the *C. parvum* transcriptome and establishing the accuracy of the dataset, the hundreds of genes comprising each of these clusters are not likely to be truly co-regulated in the organism, and the entire diversity of *C. parvum* gene expression profiles cannot be accurately captured in only nine clusters.

We have clustered *C. parvum* genes into 200 putatively co-regulated clusters. Many lines of evidence support the biological relevance of many of these clusters, namely: (1) Expression profiles for each cluster are very tightly co-expressed, and there are statistically significant overrepresented motifs upstream of the genes comprising 198 of 200 clusters; (2) identified overrepresented motifs fall into 11 motif families, many of which could potentially be bound by known *C. parvum* transcription factors, as well as one previously unknown G-box-binding ApiAP2 transcription factor, cgd8_810; and (3) the two examples of functionally related and known co-expressed genes that we examined (COWP genes and ribosomal proteins) often fall into the same clusters.

Many of the overrepresented motifs we present are still unknown. The binding specificities of most of the putative *C. parvum* transcription factors are not known, particularly the many possible zinc fingers and ApiAP2 proteins; these unknown motifs could represent binding sites for these factors. It is also a possibility that these motifs are not true transcription factor binding sites, or they might represent some other *cis* element important for other mechanisms of gene regulation, such as binding sites for proteins involved in epigenetic regulation. It is another possibility that these motifs do not play any biological role and represent some sort of repeat element. Further studies to determine binding sites for the other suspected *C. parvum* transcription factors or other DNA-binding proteins coupled with experiments to determine their binding sites throughout the genome (ie, utilizing ChIP-seq) are needed to distinguish between these possibilities.

We also note that the E2F motif is particularly overrepresented throughout the upstream regions of the *C. parvum* genome. This is very interesting, given the absence of E2F transcription factors in other apicomplexans. It is an intriguing possibility that *C. parvum* is unusually reliant on its two E2F transcription factors for transcriptional regulation. Clusters containing overrepresented E2F motifs in the upstream regions of their genes are maximally expressed at any of the post-infection timepoints. The E2Fs themselves are expressed at some level at all timepoints, though cgd1_1570 is maximally expressed at 2 hours post-infection, and cgd6_1430 is maximally expressed at 12 hours post-infection. E2F proteins could thus be available to regulate at any timepoint. However, presence of the motif does not necessarily indicate that the transcription factor binds it. Indeed, Flueck et al. (2010) recently identified that *P. falciparum* ApiAP2 protein PFF0200c only binds a small subset of its possible motifs *in vivo* [65]. ChIP-seq experiments to determine whether or not most of these overrepresented motifs act as true E2F binding sites will help to elucidate the importance of E2F transcription factors in *C. parvum* transcriptional regulation.

Our data suggest that in most cases, a single overrepresented motif is not sufficient to explain cluster expression patterns. A notable exception is in the case of E2F motif-containing clusters that peak at 2 hours post-infection, where the E2F motif is the only overrepresented motif detected upstream in 45% of these clusters. Both E2F transcription factors are expressed at this time point and could possibly be driving expression of these clusters. However, peak expression at 2 hours post-infection is not usually so easily explained, and the presence of the E2F motif is not the only determinant of peak expression at 2 hours post-infection; clusters containing any of our identified overrepresented motifs can peak at this time. Another 45% of E2F-motif-containing clusters also have Unknown motif 14 overrepresented upstream, and the ubiquity of this motif in regions upstream of clusters having a wide variety of expression patterns makes the influence it has on gene expression, if any, very difficult to decipher. We see any manner of combinations of motifs overrepresented upstream of clusters with highly variable expression patterns, which suggests a very complicated interplay between motifs and transcription factors that act together to determine these intricate and precise expression patterns.

Functionally related or known co-expressed genes appear together in clusters in the case of ribosomal proteins and COWP genes. Clustering further allowed us to distinguish between two potentially co-regulated classes of COWP: the earlier-expressed Class I, which peaks at 48 hours, then declines to 72hours; and the later-expressed Class II, which rises steadily from 36 hours to peak at 72 hours. The E2F binding motif (motif 3) is overrepresented upstream of Class I COWPs, while a known ApiAP2 binding site (motif 1) is overrepresented upstream of Class II. It is possible that this differential regulation indicates functional differences between the two classes of COWP. It should be noted that the expression data for COWPs generated from our study differ slightly from what has previously been described [63]. The membership between Class I and Class II differs slightly between datasets, with COWP1 and COWP6 changing classes. Despite these differences, both datasets suggest two differentially regulated classes of COWPs. Electron microscopy data indicate that the *C. parvum* thick-walled oocyst is divided into three layers: a ~10nm outer layer; a 2.5nm electron-lucent middle layer; and a thick, multi-zoned inner layer of 37.4 nm [66]. No mechanism has yet been indicated for how the oocyst wall is formed. Protein localization data indicate that COWP1 (a member of the earlier-expressed class of COWP) localizes to the inner oocyst wall [67]. An antibody to COWP8 (a member of the later-expressed Class II) is only reactive to ruptured oocysts [63], indicating this COWP is not expressed on the oocyst surface, but there is no precise localization data for

COWP8. To our knowledge no other COWP protein localization data is available. With these limited data, it is tempting to speculate that the earlier class of COWPs represents components of the inner oocyst membrane, while the later-expressed class of COWPs builds on this earlier structure to help form the remaining layers. Future localization studies on the remaining COWPs will help investigate this hypothesis.

Truly proving regulatory functions of putative transcription factors and overrepresented upstream motifs in *Cryptosporidium* is a tremendous challenge without a continuous *in vitro* culture system or molecular genetic tools. Despite these limitations, bioinformatic tools have allowed us to shed more light on transcriptional regulation in *C. parvum*, as well as to generate some testable hypotheses that will further elucidate regulatory mechanisms and other aspects of *C. parvum* biology.

REFERENCES

1. TR N, AM H (1987) Cryptospridiosis in patients with Aids. Journal of infectious diseases 155.

2. Tzipori S (1988) Cryptosporidiosis in perspective. Adv Parasitol 27: 63-129.

3. Spano F, Crisanti A (2000) *Cryptosporidium parvum*: the many secrets of a small genome. Int J Parasitol 30: 553-565.

4. Gatei W, Wamae CN, Mbae C, Waruru A, Mulinge E, et al. (2006) Cryptosporidiosis: prevalence, genotype analysis, and symptoms associated with infections in children in Kenya. American Journal of Tropical Medicine and Hygiene 75: 78-82.

5. Campbell PN, Current WL (1983) Demonstration of serum antibodies to *Cryptosporidium sp.* in normal and immunodeficient humans with confirmed infections. J Clin Microbiol 18: 165-169.

6. Abrahamsen MS, Schroeder AA (1999) Characterization of intracellular *Cryptosporidium parvum* gene expression. Mol Biochem Parasitol 104: 141-146.

7. Girouard D, Gallant J, Akiyoshi DE, Nunnari J, Tzipori S (2006) Failure to propagate *Cryptosporidium spp*. in cell-free culture. J Parasitol 92: 399-400.

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

9. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, et al. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Mol Microbiol 71: 1402-1414.

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

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

12. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc 4: 393-411.

13. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. Plos Pathogens 6.

14. Militello KT, Dodge M, Bethke L, Wirth DF (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. Mol Biochem Parasitol 134: 75-88.

15. Young JA, Johnson JR, Benner C, Yan SF, Chen K, et al. (2008) In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. BMC Genomics 9: 70.

16. Mullapudi N, Joseph SJ, Kissinger JC (2009) Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. Genome Biol 10: R34.

17. Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, et al. (2010) Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*. PLoS One 5: e12354.

18. Guo X, Silva JC (2008) Properties of non-coding DNA and identification of putative cis-regulatory elements in *Theileria parva*. BMC Genomics 9: 582.

19. Mullapudi N, Lancto CA, Abrahamsen MS, Kissinger JC (2007) Identification of putative cis-regulatory elements in *Cryptosporidium parvum* by de novo pattern finding. BMC Genomics 8: 13.

20. Cohn B, Manque P, Lara AM, Serrano M, Sheth N, et al. (2010) Putative cis-regulatory elements associated with heat shock genes activated during excystation of *Cryptosporidium parvum*. PLoS One 5: e9512.

21. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu JC, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. Plos Biology 1: 85-100.

22. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. Nucleic Acids Res 34: 1166-1173.

23. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503-1508.

24. Gaji RY, Behnke MS, Lehmann MM, White MW, Carruthers VB (2011) Cell cycle-dependent, intercellular transmission of *Toxoplasma gondii* is accompanied by marked changes in parasite gene expression. Mol Microbiol 79: 192-204.

25. Lescault PJ, Thompson AB, Patil V, Lirussi D, Burton A, et al. (2010) Genomic data reveal *Toxoplasma gondii* differentiation mutants are also impaired with respect to switching into a novel extracellular tachyzoite state. PLoS One 5: e14463.

26. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, et al. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304: 441-445.

27. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, et al. (2004) The genome of *Cryptosporidium hominis*. Nature 431: 1107-1112.

28. Mauzy MJ, Enomoto S, Lancto CA, Abrahamsen MS, Rutherford MS (2012) The *Cryptosporidium parvum* Transcriptome during In Vitro Development. Plos One 7.

29. Upton SJ, Tilley M, Brillhart DB (1995) Effects of select medium supplements on in vitro development of *Cryptosporidium parvum* in HCT-8 cells. J Clin Microbiol 33: 371-375.

30. Schroeder AA, Brown AM, Abrahamsen MS (1998) Identification and cloning of a developmentally regulated *Cryptosporidium parvum* gene by differential mRNA display PCR. Gene 216: 327-334.

31. Sifuentes LY, Di Giovanni GD (2007) Aged HCT-8 cell monolayers support *Cryptosporidium parvum* infection. Appl Environ Microbiol 73: 7548-7551.

32. Cai X, Woods KM, Upton SJ, Zhu G (2005) Application of quantitative real-time reverse transcription-PCR in assessing drug efficacy against the intracellular pathogen *Cryptosporidium parvum* in vitro. Antimicrob Agents Chemother 49: 4437-4442.

33. Giulietti A, Overbergh L, Valckx D, Decallonne B, Bouillon R, et al. (2001) An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. Methods 25: 386-401.

34. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonak J, et al. (2006) The real-time polymerase chain reaction. Mol Aspects Med 27: 95-125.

35. Laan MJvd, Pollard KS (2001) Hybrid Clustering of Gene Expression Data with Visualization and the Bootstrap. UC Berkeley Division of Biostatistics Working Paper Series U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 93.

36. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

37. Pollard KS, Laan MJvd (2002) A Method to Identify Significant Clusters in Gene Expression Data. UC Berkeley Division of Biostatistics Working Paper Series Working Paper 107.

38. Futschik ME, Carlisle B (2005) Noise-robust soft clustering of gene expression time-course data. J Bioinform Comput Biol 3: 965-988.

39. Zheng Q, Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36: W358-363.

40. Hochberg Y, Benjamini Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of Royal Statistical Society 57: 289-300.

41. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

42. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16: 939-945.

43. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28: 337-350.

44. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Research 35: W253-W258.

45. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. (2008) Specific DNA-binding by Apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 105: 8393-8398.

46. R Development Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

47. Maechler M, Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2011) Cluster analysis basics and extensions. R package version 1.14.1.

48. Mullapudi N, Lancto CA, Abrahamsen MS, Kissinger JC (2007) Identification of putative cis-regulatory elements in *Cryptosporidium parvum* by de novo pattern finding. Bmc Genomics 8.

49. Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M, et al. (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. Genome Res 13: 1787-1799.

50. Zheng N, Fraenkel E, Pabo CO, Pavletich NP (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev 13: 666-674.

51. Cartwright P, Muller H, Wagener C, Holm K, Helin K (1998) E2F-6: a novel member of the E2F family is an inhibitor of E2F-dependent transcription. Oncogene 17: 611-623.

52. Page BD, Guedes S, Waring D, Priess JR (2001) The C. elegans E2F- and DP-related proteins are required for embryonic asymmetry and negatively regulate Ras/MAPK signaling. Mol Cell 7: 451-460.

53. Suzuki A, Hemmati-Brivanlou A (2000) Xenopus embryonic E2F is required for the formation of ventral and posterior cell fates during early embryogenesis. Mol Cell 5: 217-229.

54. Sawado T, Yamaguchi M, Nishimoto Y, Ohno K, Sakaguchi K, et al. (1998) dE2F2, a novel E2F-family transcription factor in *Drosophila melanogaster*. Biochem Biophys Res Commun 251: 409-415.

55. Guo J, Song J, Wang F, Zhang XS (2007) Genome-wide identification and expression analysis of rice cell cycle genes. Plant Mol Biol 64: 349-360.

56. Ramirez-Parra E, Frundt C, Gutierrez C (2003) A genome-wide identification of E2F-regulated genes in *Arabidopsis*. Plant J 33: 801-811.

57. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res 14: 1686-1695.

58. Rider SD, Jr., Zhu G (2010) *Cryptosporidium*: genomic and biochemical features. Exp Parasitol 124: 2-9.

59. Biggin MD, Tjian R (1988) Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. Cell 53: 699-711.

60. Planta RJ, Goncalves PM, Mager WH (1995) Global regulators of ribosome biosynthesis in yeast. Biochem Cell Biol 73: 825-834.

61. Essien K, Stoeckert CJ, Jr. (2010) Conservation and divergence of known apicomplexan transcriptional regulons. Bmc Genomics 11: 147.

62. Van Poppel NF, Welagen J, Vermeulen AN, Schaap D (2006) The complete set of *Toxoplasma gondii* ribosomal protein genes contains two conserved promoter elements. Parasitology 133: 19-31.

63. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Research 14: 1686-1695.

64. Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, et al. (2010) Coordinated Progression through Two Subtranscriptomes Underlies the Tachyzoite Cycle of *Toxoplasma gondii*. Plos One 5.

65. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BTF, et al. (2010) A Major Role for the *Plasmodium falciparum* ApiAP2 Protein PfSIP2 in Chromosome End Biology. Plos Pathogens 6.

66. Reduker DW, Speer CA, Blixt JA (1985) Ultrastructure of *Cryptosporidium parvum* Oocysts and Excysting Sporozoites as Revealed by High Resolution Scanning Electron Microscopy1. Journal of Eukaryotic Microbiology 32: 708-711.

67. SPANO F, PURI C, RANUCCI L, PUTIGNANI L, CRISANTI A (1997) Cloning of the entire COWP gene of *Cryptosporidium parvum* and ultrastructural localization of the protein during sexual parasite development. Parasitology 114: 427-437.

Figure and table legends

Figure 3.1. *C. parvum* gene expression across the *in vitro* infective stage. **A.** The normalized expression profiles of all 3281 genes used in our study were sorted according to peak expression at each time point. Each row represents the expression profile of each gene at 2 hr, 6 hr, 12 hr, 24 hr, 36 hr, 48 hr and 72 hr post-infection. **B.** The normalized expression profiles of a representative gene for all 200 clusters identified using FCM analysis. Each row in the heatmap represents a cluster.

Figure 3.2. Expression profiles of clusters containing overrepresented AP2_1-like motifs in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.3. Expression profiles of clusters containing overrepresented E2F-like motifs in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.4. Expression profiles of clusters containing overrepresented G-box-like motifs in the upstream regions of their genes. A. The ApiAP2 transcription factor binds the G-box motif.

B-C. Similarity of motifs to each other was determined via the STAMP tool. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.5. Expression profiles of clusters containing overrepresented GAGA-like motifs in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.6. Expression profiles of clusters containing overrepresented CAAT-box-like motifs in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.7. Expression profiles of clusters containing overrepresented motif 14 in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.8. Expression profiles of clusters containing overrepresented Unknown set 1 motifs in the upstream regions of their genes. Similarity of motifs to each other was

determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.9. Expression profiles of clusters containing overrepresented Unknown set 2 motifs in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.10. Expression profiles of clusters containing overrepresented Unknown motifs 21, 22 or 25 in the upstream regions of their genes. Similarity of motifs to each other was determined via the STAMP tool. Expression data shown for 2, 6, 12, 24, 36, 48 and 72 hours post-infection. Green = peak expression; red = min expression. The normalized expression profiles of clusters were sorted according to peak expression at each time point.

Figure 3.12. Overrepresented motifs upstream of ribosomal proteins in *P. falciparum* and *C. parvum*. *Pf* graph shows expression profiles for 68 co-expressed IDC ribosomal proteins (data from Bozdech et al. 2003). *Cp* graph shows 30 co-expressed ribosomal proteins from clusters #6 and #20. Five representative upstream regions are shown for each organism out of 68 and 60 searched for *Pf* and *Cp* respectively. Upstream regions for each of these genes (see Materials and Mathods) were mined for overrepresented motifs using MEME (max motif width 12bp, 5 motifs max, mode = anr). Similarity of motifs to each other was determined via the STAMP tool. As

previously documented, the upstream regions of *Pf* ribosomal proteins contain overrepresented G-box motifs (Essien and Stoeckert, 2010). *Cp* ribosomal proteins instead have E2F-like and GAGA-like motifs overrepresented upstream.

Figure 3.12. Overrepresented motifs upstream of COWPs by subclass. A. Expression profiles of Class I and Class II COWPs. The five COWPs that fall into class 1 peak at 48 hrs post-infection, then decline to 72 hrs. The remaining four COWPs that comprise Class II begin rising at 48 hrs and peak at 72 hrs. **B1.** The upstream regions of each of the Class I COWPs contain 5 overrepresented motifs that fall into 3 groups. Upstream regions for each of these genes (see Materials and Mathods) were mined for overrepresented motifs using MEME (max motif width 12bp, 5 motifs max, mode = anr). Similarity of motifs to each other was determined via the STAMP tool. Three motifs overrepresented upstream of Class I COWPs are closely related to E2F binding sites. A GAGA-like motif and an ApiAP2 motif identified in *P. falciparum* (Campbell et al. 2010; here we designate this motif AP2_2) are also overrepresented upstream of Class I COWPs. **B2.** The upstream regions of each of the Class II COWPs contain 5 overrepresented motifs that are unrelated to each other. Two motifs are similar to a documented ApiAP2 binding site across apicomplexans. E2F-like and CAAT-box-like motifs are also overrepresented. The remaining motif is unknown and does not appear related to any of the 25 motifs overrepresented upstream of genes throughout the genome.

Figure 3.13. Overrepresented motifs upstream of genes in clusters peaking primarily at 72hrs post-infection. Nine representative upstream regions are shown out of 105 searched. Upstream regions for each of these genes (see Materials and Mathods) were mined for

overrepresented motifs using MEME (max motif width 12bp, 5 motifs max, mode = anr). Similarity of motifs to each other was determined via the STAMP tool. The upstream regions of genes in clusters peaking primarily at 72 hours share 4 overrepresented motifs. Two of these motifs are similar to previously identified ApiAP2 binding sites. One binding site is E2F-like. The remaining site is similar to the G-box noted in other apicomplexans, which we have demonstrated is an ApiAP2 binding site in *C. parvum*.

Table 3.1. FCM cluster analysis parameter exploration. The fuzzification parameter, *m*, and the number of clusters, *c*, were varied from 1.05 to 1.55 and 50 to 250, respectively. Black text indicates the average membership value of genes to their assigned clusters. Purple text indicates the total number of genes included in clustering at each parameter set. Red and blue text indicate the maximum and minimum cluster sizes, respectively.

Table 3.2. List of all 25 overrepresented motifs identified in this study. IUPAC codes are used to represent each motif. The algorithm(s) that identified each motif is indicated, as well as the number of clusters identified with the motif overrepresented upstream.

Table 3.3. Possible *C. parvum* transcription factors. Domains commonly associated with transcription factors and their counts in *C. parvum* as determined by text searches at Cryptodb.org. *Presence of several of these domains, particularly the C2H2 Zinc finger and Myb, do not necessarily indicate the protein acts as a transcription factor.

Table 3.4. Occurrence of all 25 identified motifs overrepresented upstream of 200 clusters.

X's indicate the presence of each motif. Motifs are grouped by similarity to each other.

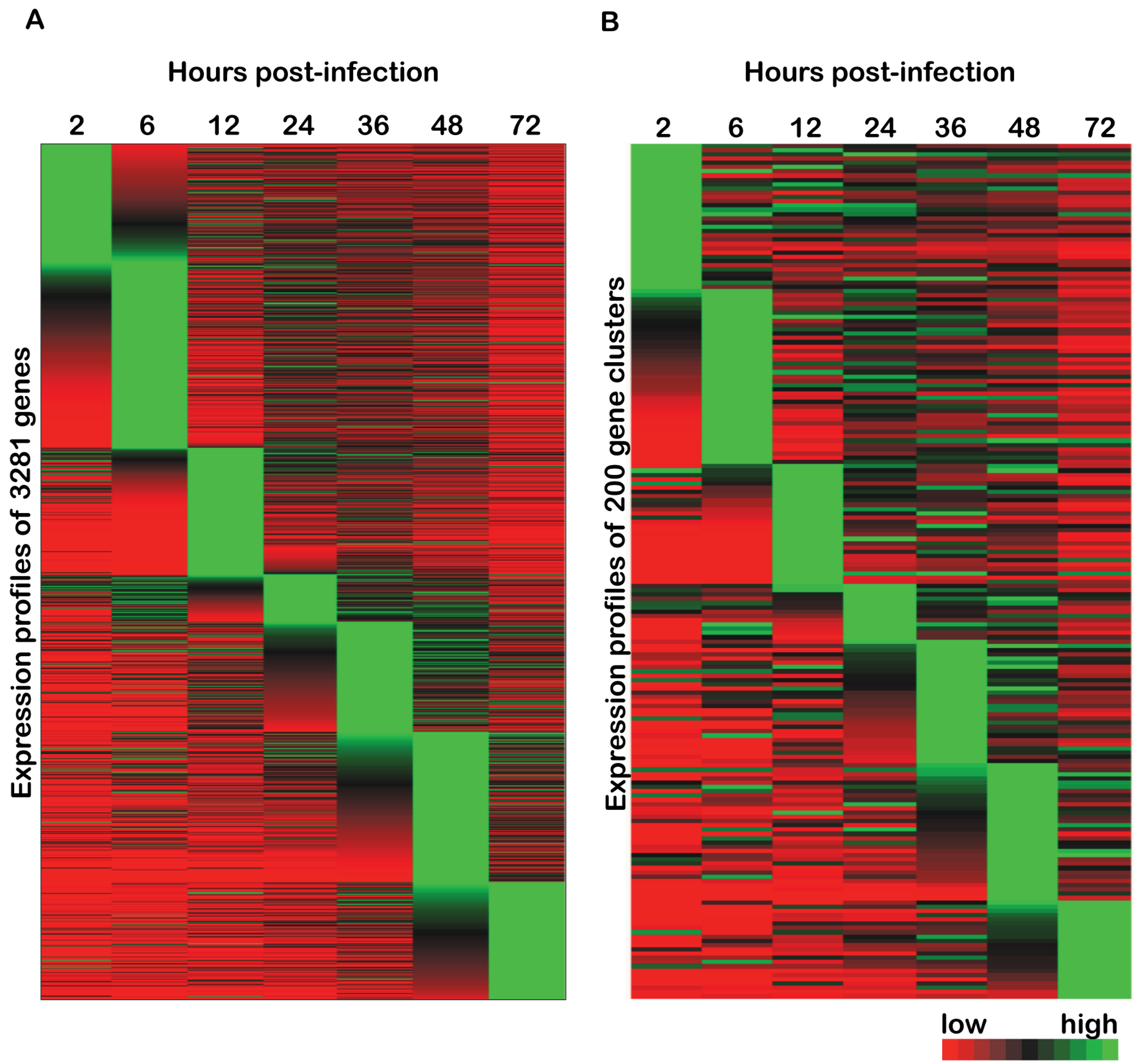


Figure 3.1. ***C. parvum*** **gene expression across the** ***in vitro*** **infective stage.**

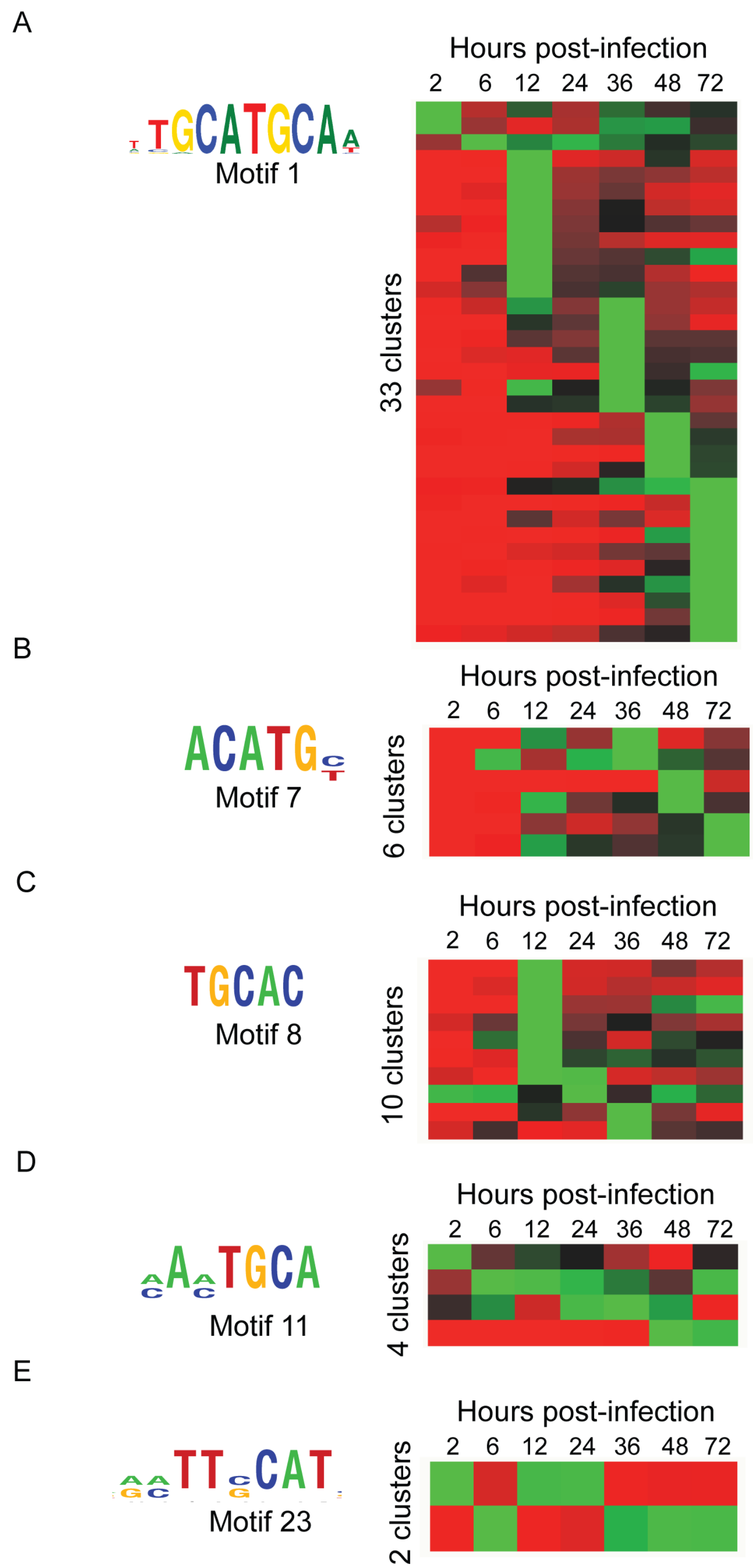


Figure 3.2. Expression profiles of clusters containing overrepresented AP2_1-like motifs in the upstream regions of their genes.

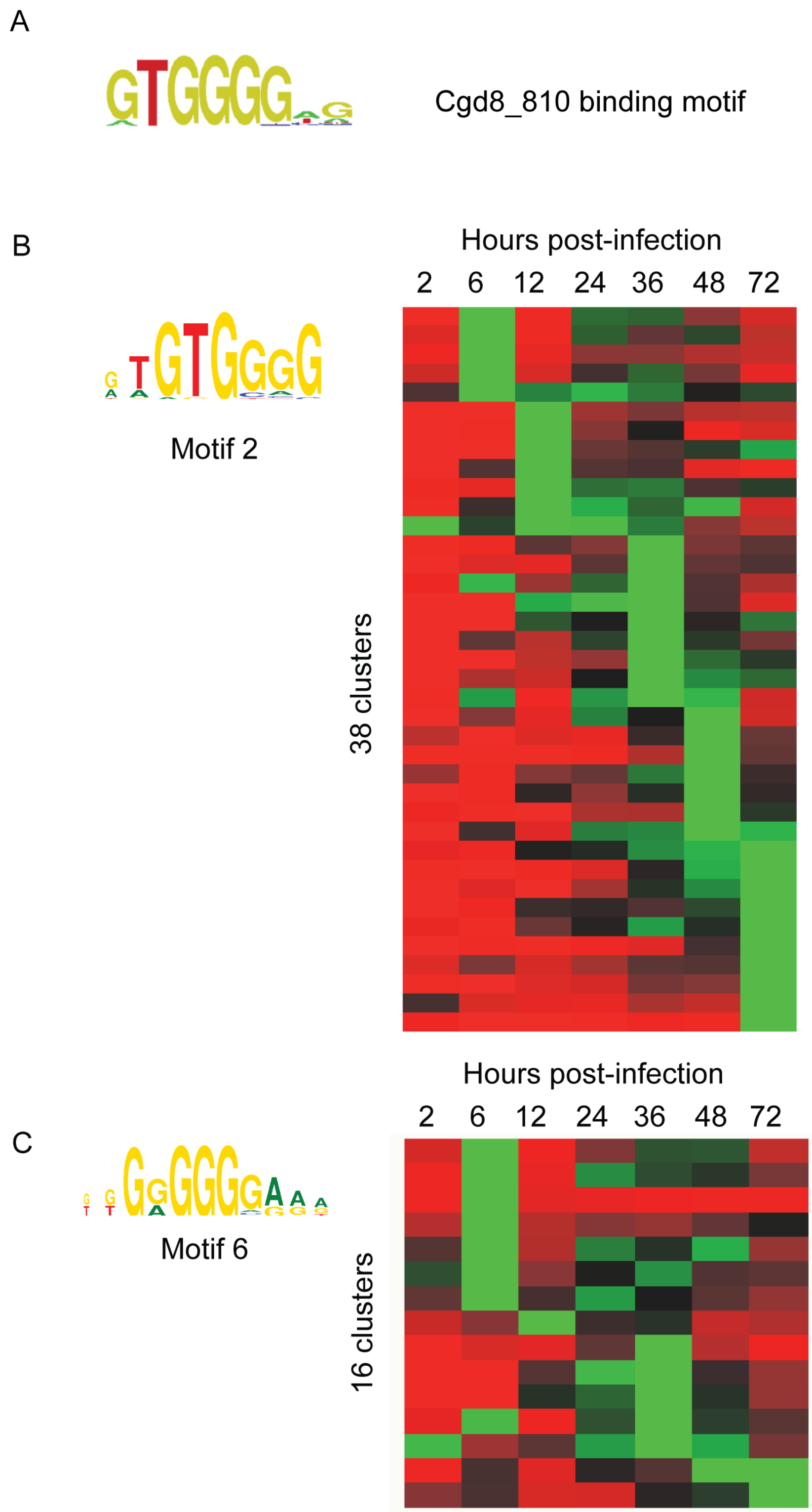
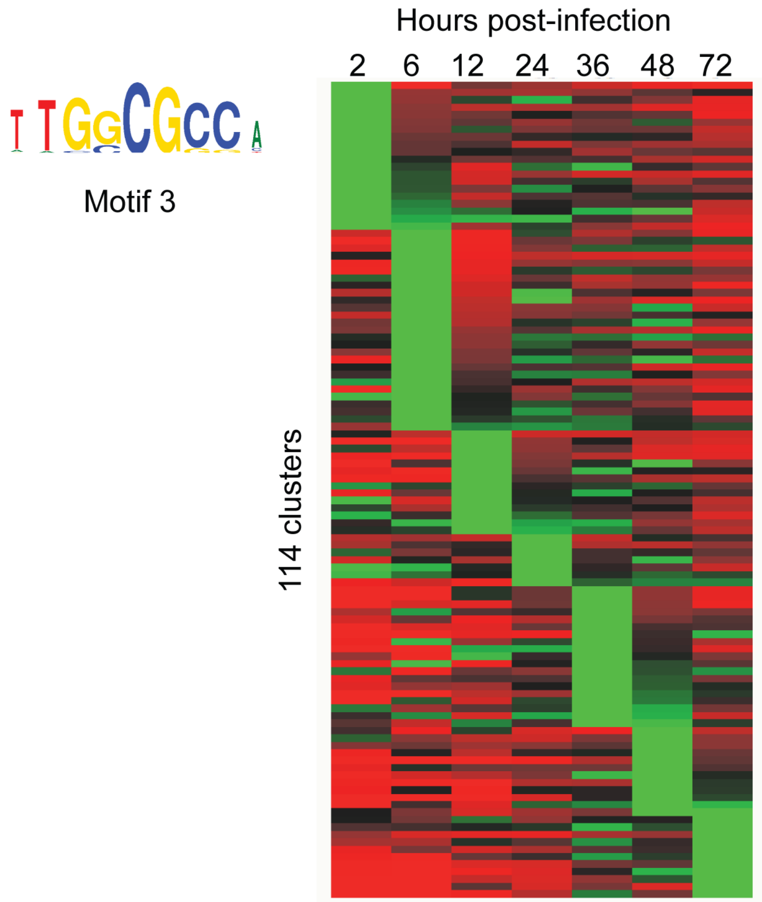


Figure 3.3. Expression profiles of clusters containing overrepresented G-box-like motifs in the upstream regions of their genes.

A



B

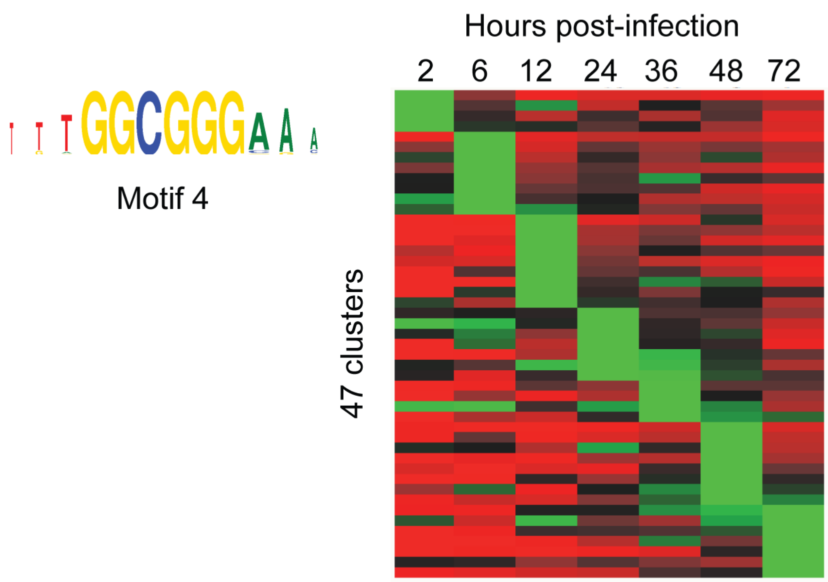


Figure 3.4. Expression profiles of clusters containing overrepresented E2F-like motifs in the upstream regions of their genes.

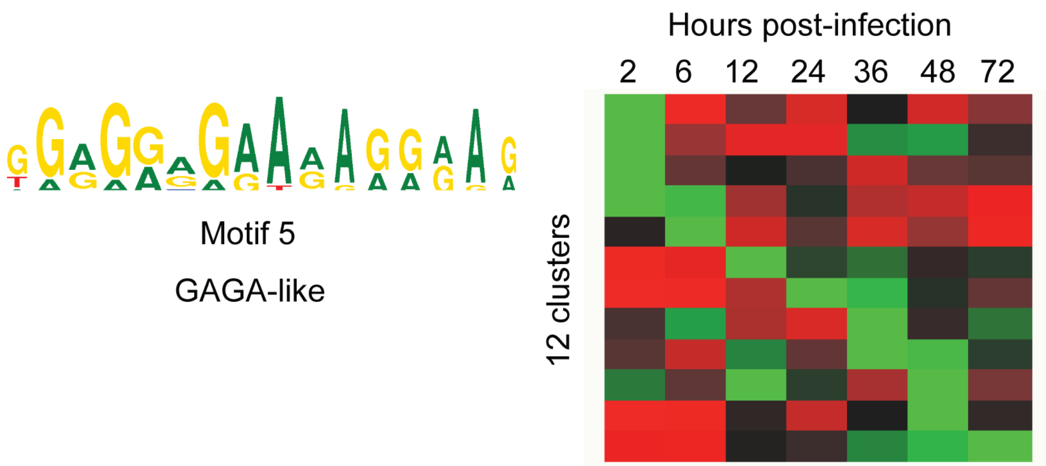


Figure 3.5. Expression profiles of clusters containing overrepresented GAGA-like motifs in the upstream regions of their genes.

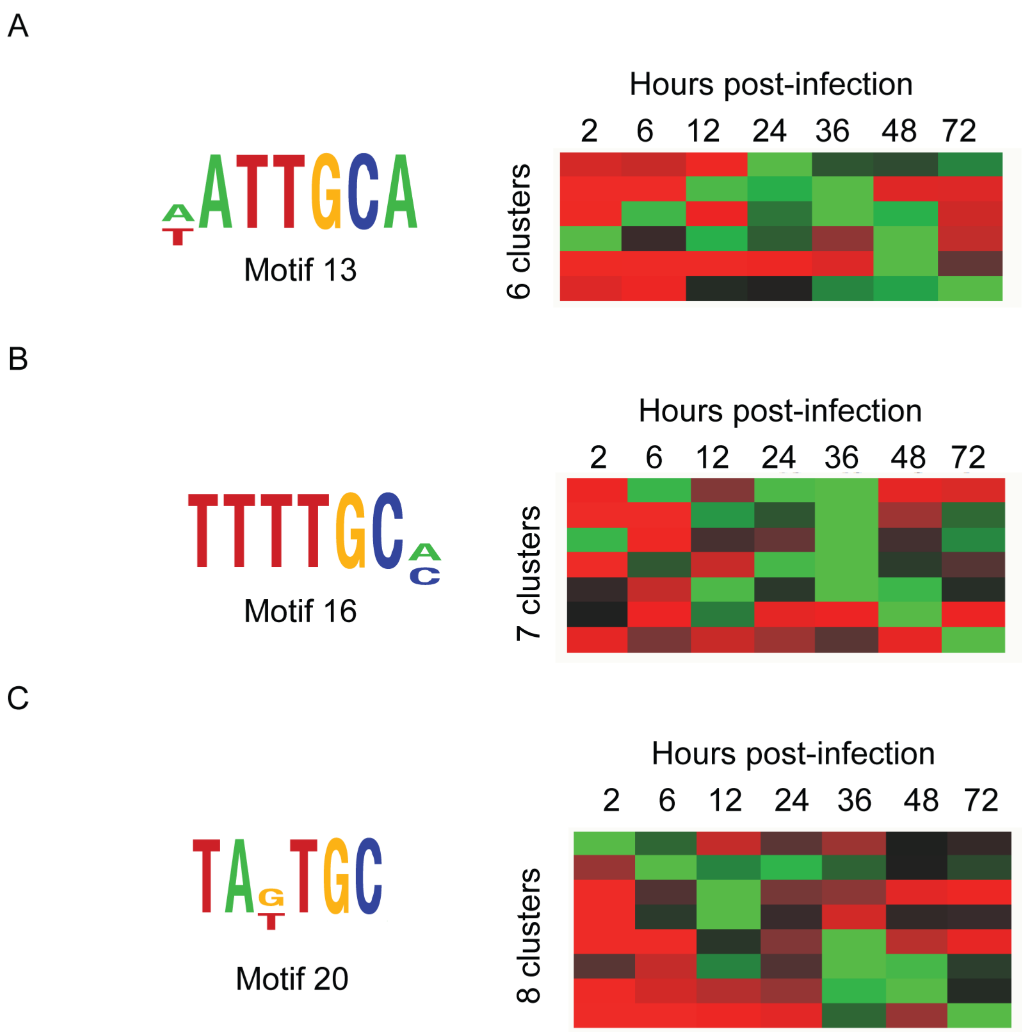


Figure 3.6. Expression profiles of clusters containing overrepresented CAAT-box-like motifs in the upstream regions of their genes.

A

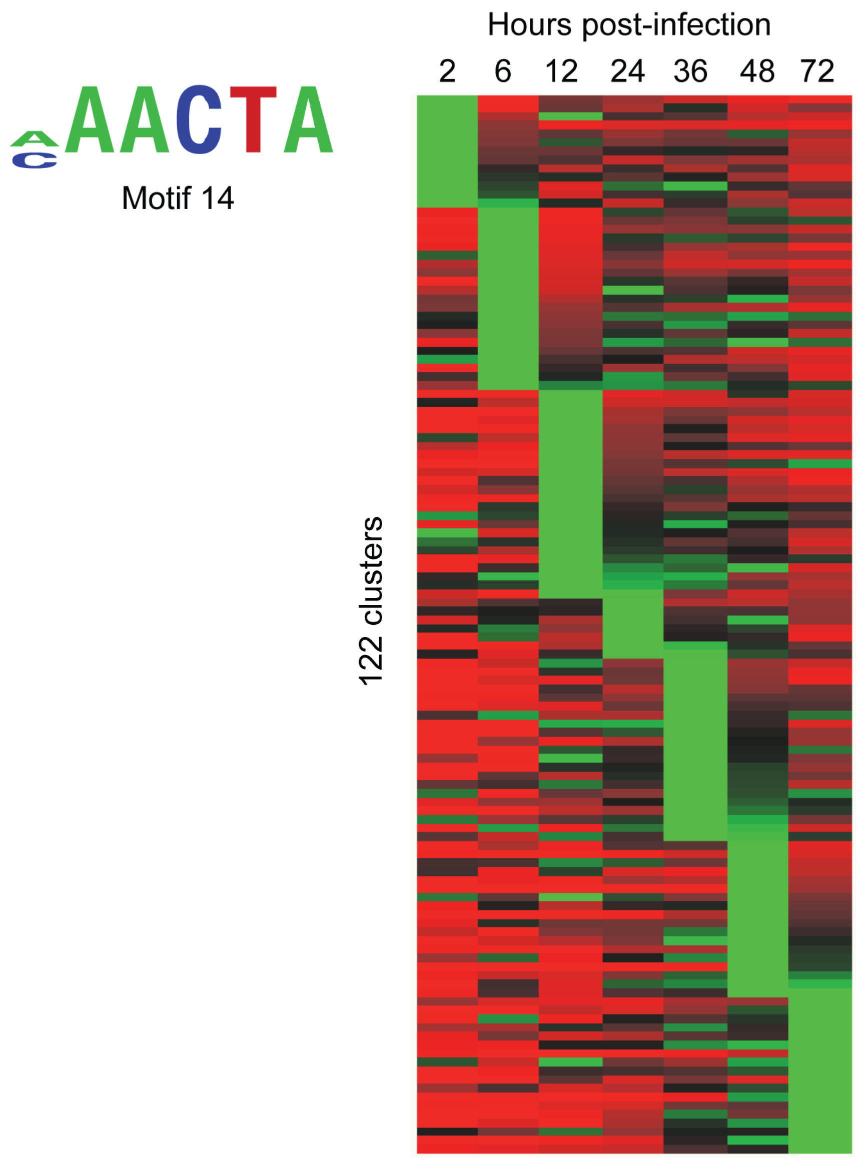


Figure 3.7. Expression profiles of clusters containing overrepresented motif 14 in the upstream regions of their genes.

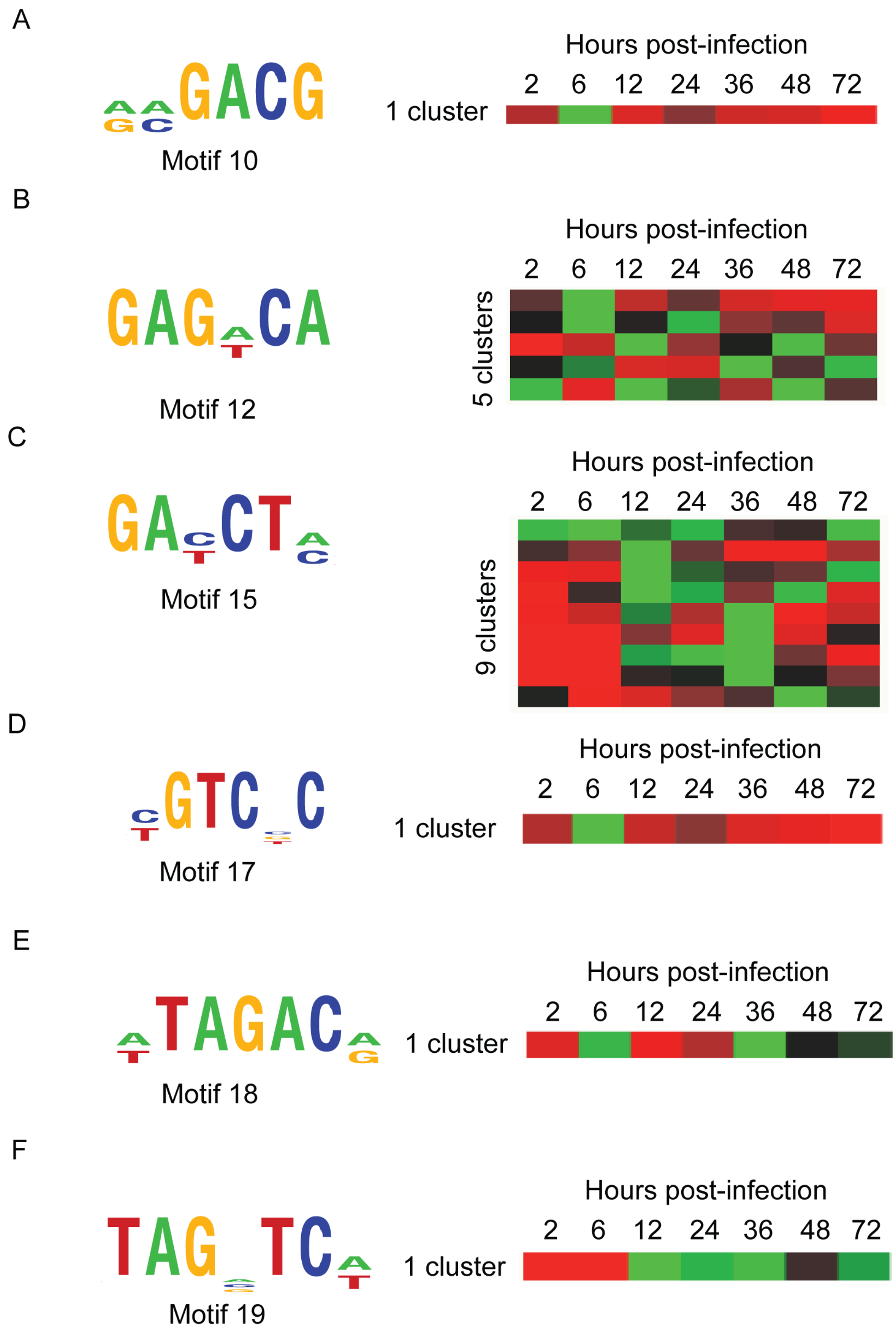


Figure 3.8. Expression profiles of clusters containing overrepresented Unknown set 1 motifs in the upstream regions of their genes.

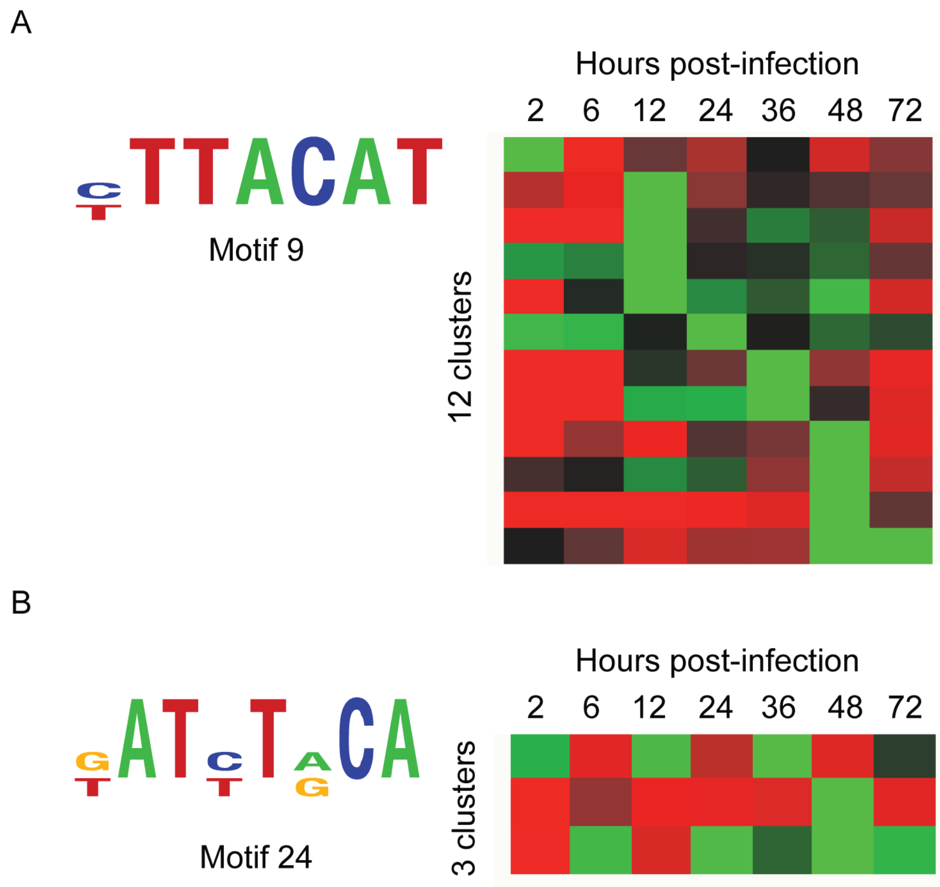


Figure 3.9. Expression profiles of clusters containing overrepresented Unknown set 2 motifs in the upstream regions of their genes.

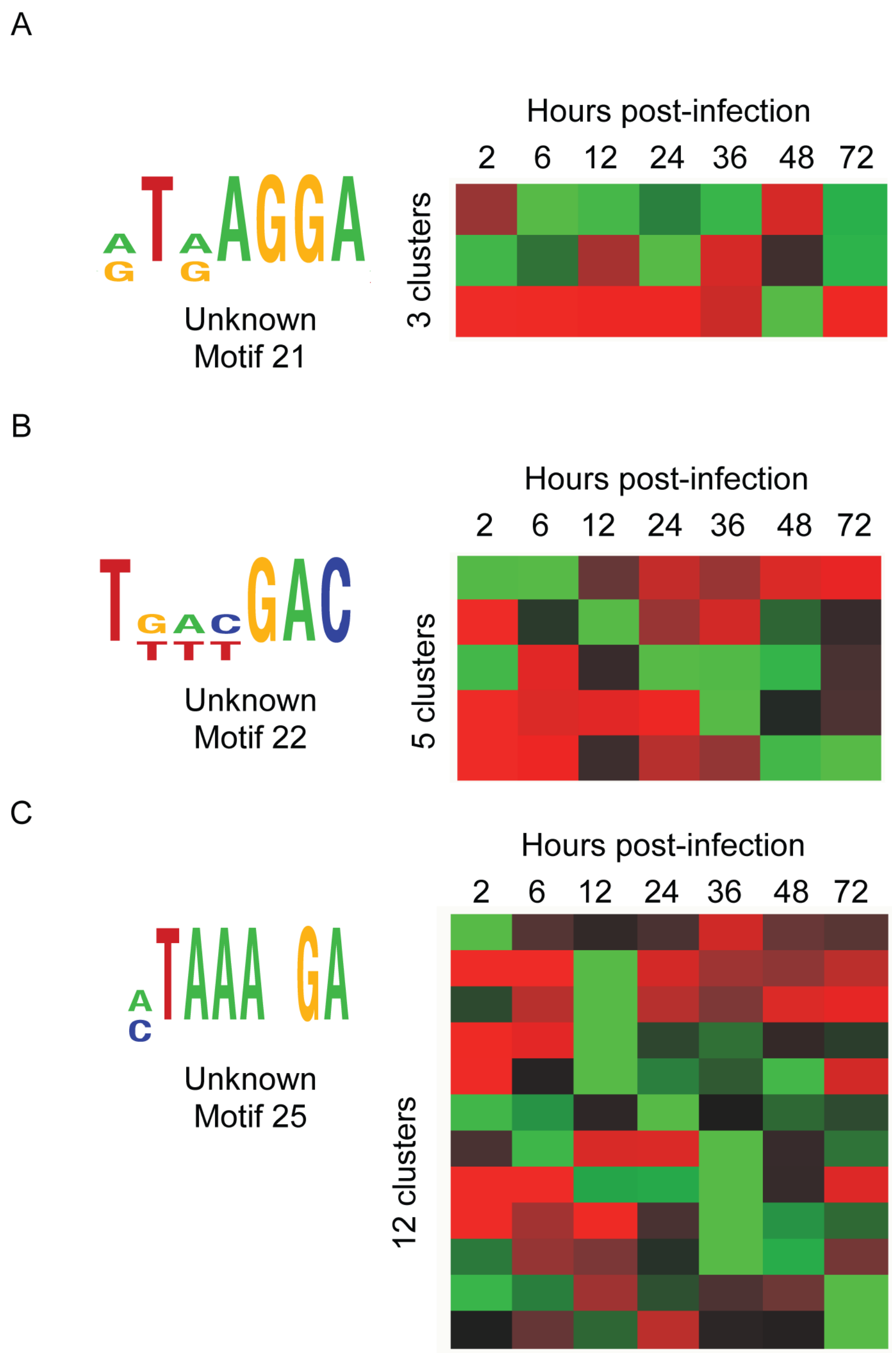


Figure 3.10. Expression profiles of clusters containing overrepresented Unknown motifs 21, 22 or 25 in the upstream regions of their genes.

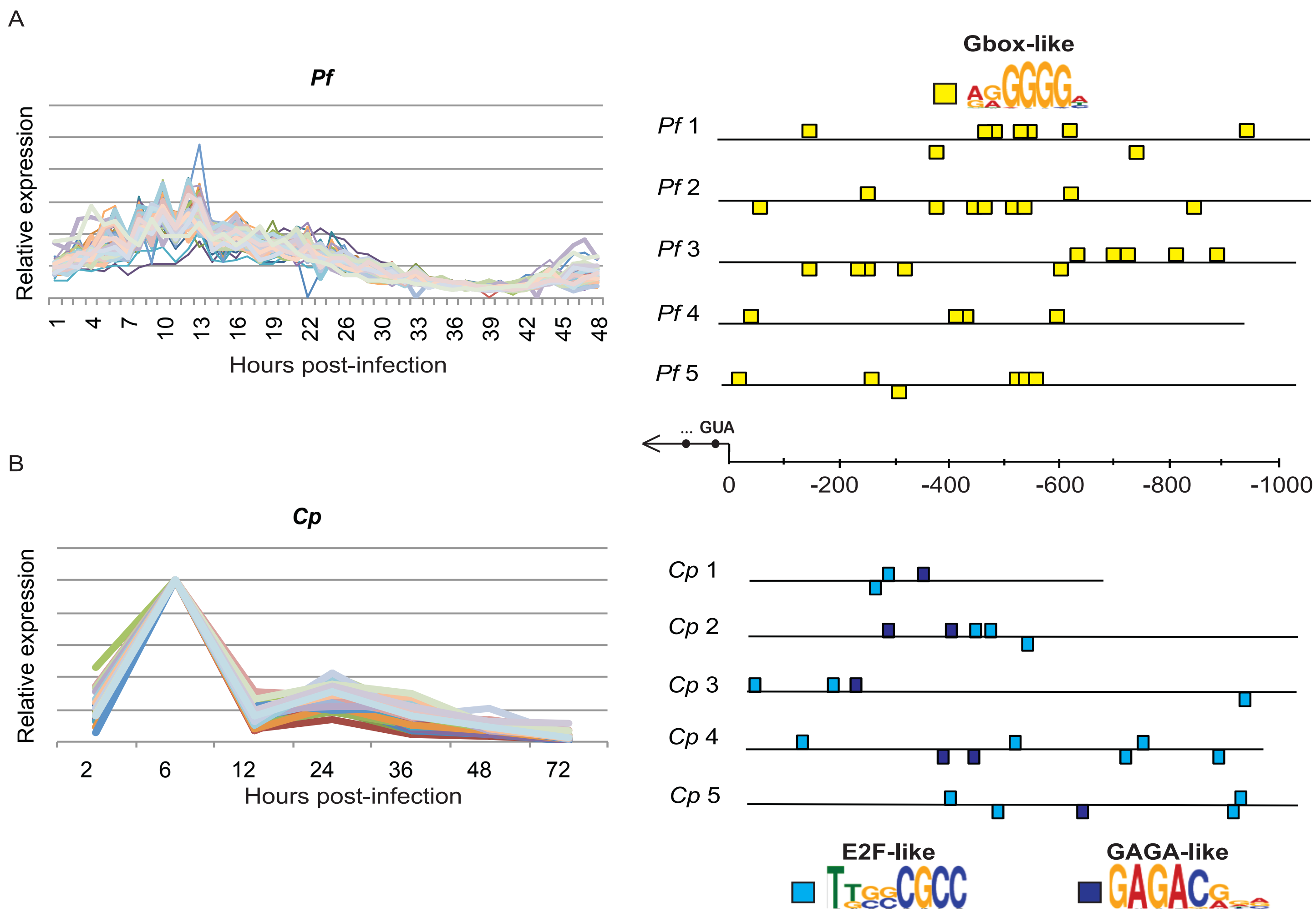


Figure 3.11. Overrepresented motifs upstream of ribosomal proteins in *P. falciparum* and *C. parvum*.

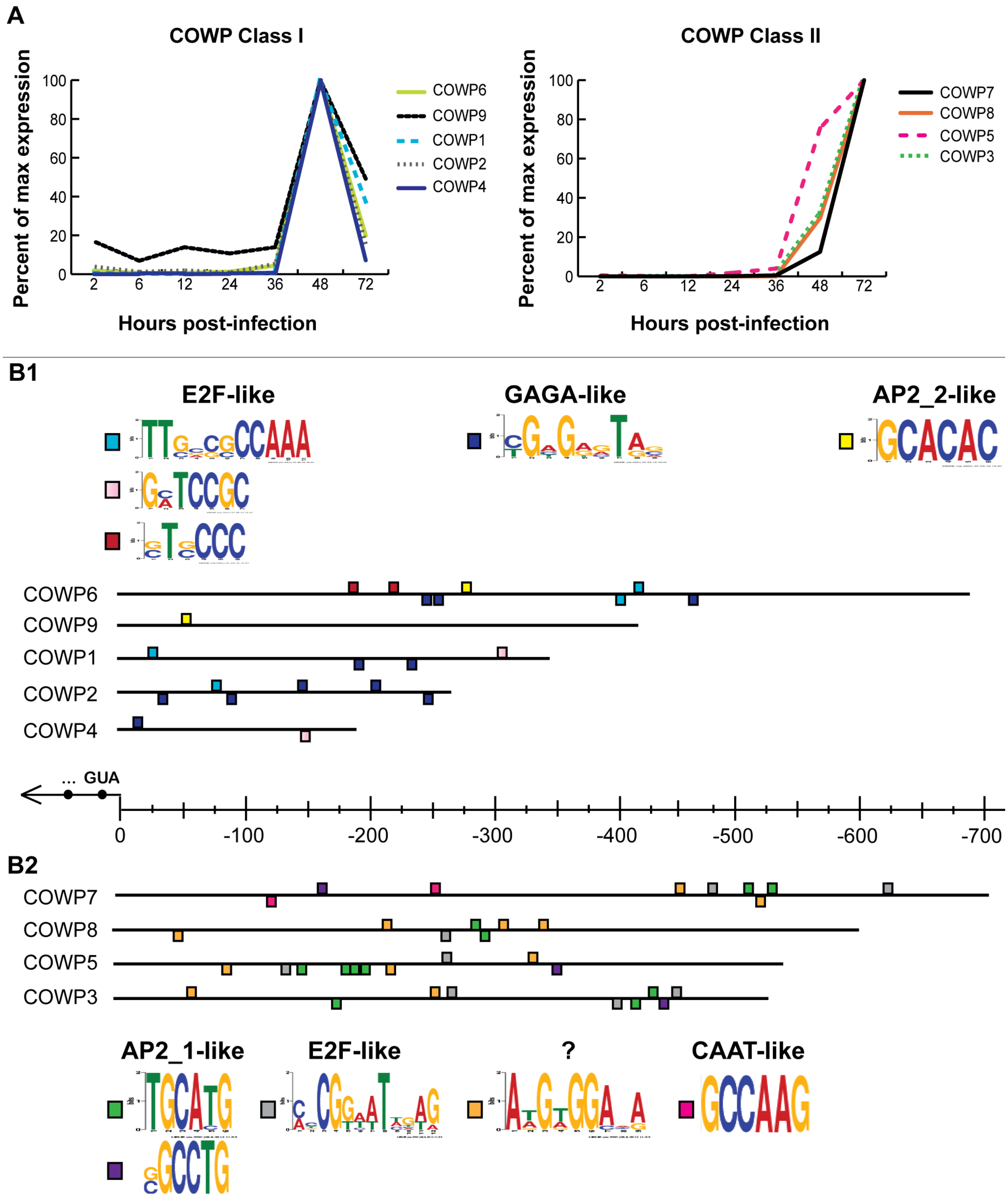


Figure 3.12. Overrepresented motifs upstream of COWPs by subclass.

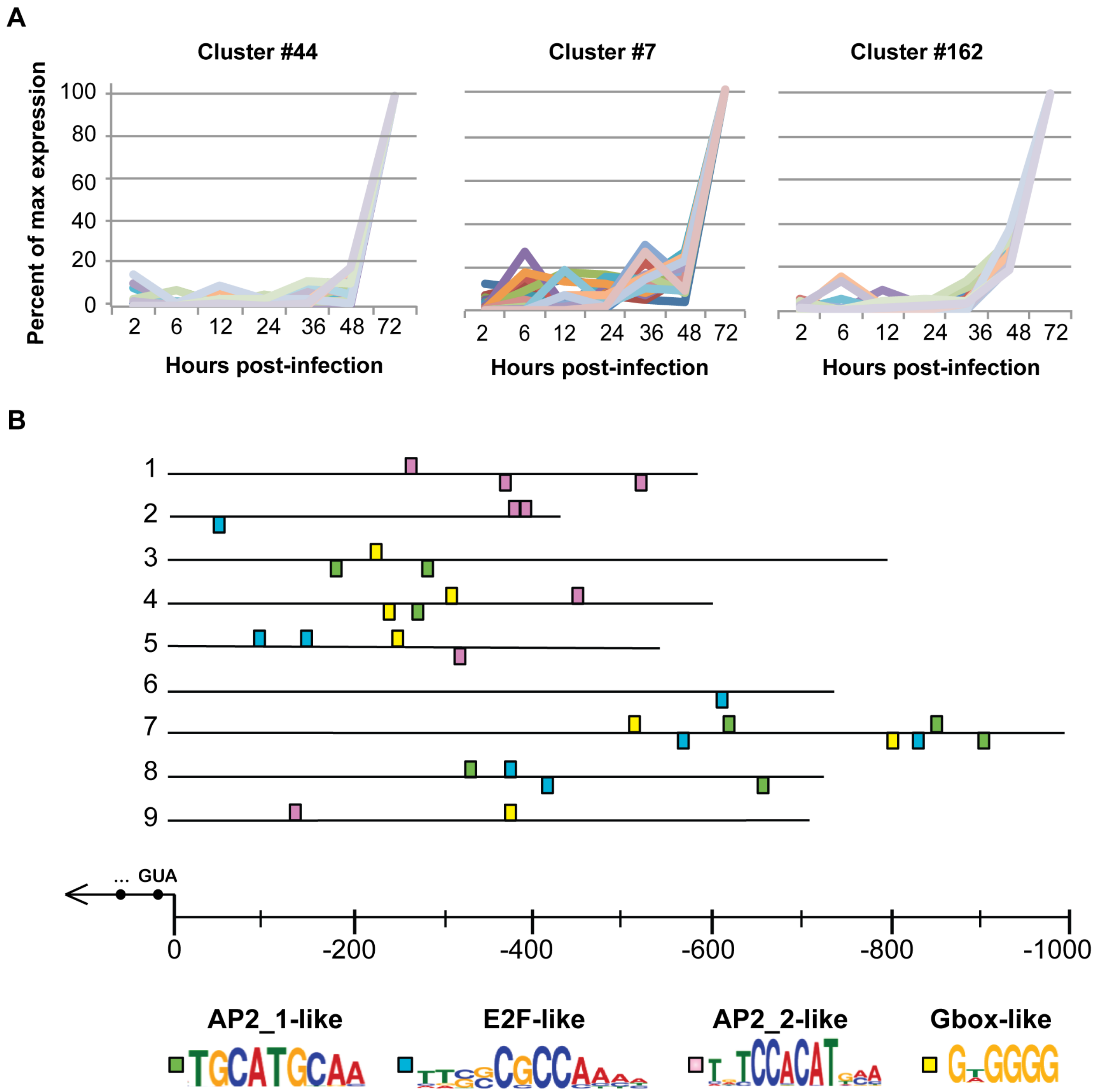


Figure 3.13. Overrepresented motifs upstream of genes in clusters peaking primarily at 72hrs post-infection.

Table 3.1. FCM cluster analysis parameter exploration.

Fuzzification parameter, *m* & No. of clusters, *c*	50	100	150	200	250
1.05	0.9813 3280 176 24	0.9897 3277 105 10	0.9937 3281 82 5	0.9960 3281 68 5	0.9972 3281 51 1
1.15	0.9201 3180 174 26	0.9299 3179 111 9	0.9421 3200 63 5	0.9493 3207 55 5	0.9547 3230 36 3
1.25	0.8583 2913 173 23	0.8678 2909 103 9	0.8762 2893 77 6	0.8889 2925 53 3	0.8980 2933 53 3
1.35	0.8071 2492 157 14	0.8122 2407 93 6	0.8226 2381 68 3	0.8347 2400 50 3	0.8412 2444 36 2
1.45	0.7553 1988 137 4	0.7707 1843 81 2	0.7811 1829 55 2	0.7969 1806 55 2	0.8201 1827 41 1
1.55	0.7127 1446 125 0	0.7320 1282 68 0	0.7534 1283 52 0	0.7829 1276 46 1	0.8096 1272 35 1

Table 3.2. List of all 25 overrepresented motifs identified in this study.

	Motif Number	Consensus Motif pattern* 5′ -> 3′	Algorithms that identified the motifs			No. of clusters whose upstream sequences showed significant over-representation
			MEME	AlignACE	FIRE	
AP2_1-like	Motif 1	BGCATGCAH	+	+	+	33
	Motif 7	ACATGY	–	–	+	6
	Motif 8	HTGCACH	–	–	+	10
	Motif 11	MAMTGCA	–	–	+	4
	Motif 23	DRMTTSCATB	–	–	+	2
G-box-like	Motif 2	DTGTGGGG	+	+	+	38
	Motif 6	KKGRGGGGRR	+	+	–	16
E2F-like	Motif 3	DTTGSCGCCH	+	+	+	114
	Motif 4	TTTGGCGGGAAV	+	+	–	47
GAGA-like	Motif 5	GDGRRRRARARRR ARA	+	+	–	12
CAAT-box-like	Motif 13	WATTGCA	–	–	+	6
	Motif 16	TTTTGCM	–	–	+	7
	Motif 20	BTAKTGCD	–	–	+	8
Unknown set 1	Motif 10	RMGACG	–	–	+	1
	Motif 12	GAGWCA	–	–	+	5
	Motif 15	GAYCTMD	–	–	+	9
	Motif 17	VYGTCBC	–	–	+	1
	Motif 18	WTAGACR	–	–	+	1
	Motif 19	HTAGVTCW	–	–	+	1
Unknown set 2	Motif 9	YTTACAT	–	–	+	12
	Motif 24	KATYTRCAH	–	–	+	3
Other unknown	Motif 14	MAACTA	–	–	+	122
	Motif 21	VRTRAGGAD	–	–	+	3
	Motif 22	HTKWYGAC	–	–	+	5
	Motif 25	WMTAANGA	–	–	+	12

Table 3.3. Possible *C. parvum* transcription factors.

Domain	**# of *C. parvum* proteins**
ApiAP2	19
E2F/TDP	2
Myb	9*
Zinc finger	
GATA DNA-binding	1
C_2H_2	up to 27*
bZIP	
CAAT-binding	3
other	1

Table 3.4. Occurrence of all 25 identified motifs overrepresented upstream of 200 clusters.

Cluster #	Motifs																								
	AP2_1-like					G-box-like		E2F-like		GAGA-like	Unknown set 2		Unknown set 1						CAAT-box-like			Other unknown			
	1	7	8	11	23	2	6	3	4	5	9	24	10	12	15	17	18	19	13	16	20	14	21	22	25
1								x																	
2								x														x			
3								x														x			
4								x																	
5																						x			
6													x			x						x			
7						x		x														x			
8									x		x														
9								x																	
10									x																
11								x																	
12																									
13	x									x															
14		x						x														x			
15								x														x			
16										x				x								x			
17									x																
18								x														x			
19									x													x		x	
20								x	x													x			
21										x				x			x					x			x
22			x					x			x												x		x
23																									
24						x														x		x			
25								x																	
26																						x			
27									x													x			
28						x			x																x
29																						x			
30	x								x								x								
31		x				x			x	x													x		
32								x														x			
33	x					x		x														x			
34									x																
35								x	x					x								x			
36								x																	
37								x														x			
38						x		x												x		x			
39	x					x			x										x			x			
40						x																x			
41								x												x		x			
42	x					x			x													x			
43								x														x			x
44	x					x																x			

45						x		x																	
46								x	x																
47							x															x			
48									x													x			
49								x														x			
50						x		x														x			
51									x													x			
52						x									x							x			
53								x																	
54						x																			
55						x			x													x		x	
56																									
57								x			x											x			
58						x		x														x			
59	x																								
60								x																	
61										x	x											x			
62								x														x			
63						x																			
64								x			x														
65							x		x																
66									x													x			
67									x													x			
68	x							x														x			
69									x																
70									x													x			
71						x					x				x				x			x			x
72								x	x													x			
73								x														x			
74								x																	
75								x																	
76					x			x																	
77																									x
78	x		x			x																x			
79								x																	
80								x																	
81				x				x		x															x
82								x		x															
83						x		x				x										x			
84								x													x	x			
85																						x			
86	x							x														x			
87									x													x			
88		x						x												x					
89								x						x											
90								x														x			
91								x																	
92							x															x			
93	x							x																	
94								x														x			
95								x														x			
96								x																	
97								x																	
98								x																	

99	x			x	x	x		x							x						x	x	x		
100								x																	
101	x		x						x													x			x
102	x																					x			
103	x			x																		x			
104	x								x		x											x			
105									x																
106								x														x			
107	x					x		x														x			
108								x														x			
109							x	x																	
110								x																	
111	x														x							x			
112	x					x																x			
113								x											x						
114								x																	
115								x	x												x	x			
116	x					x		x														x			
117																						x			
118								x														x			
119						x													x			x			
120									x													x			
121	x						x								x							x			
122	x								x																
123			x						x												x	x		x	
124	x					x		x			x								x			x			
125	x					x																x			
126								x												x		x			
127						x																			
128							x	x														x			
129								x																	
130									x													x			
131	x					x			x												x	x			
132								x														x			x
133									x																
134				x				x																	
135																						x			
136									x																
137							x	x																	
138									x													x			
139								x														x			
140									x	x												x			
141								x														x			
142								x																	
143			x			x				x					x			x				x			x
144						x		x																	
145	x																					x			
146							x															x			
147																									
148																									
149																									
150																						x			
151								x		x					x							x			x
152									x																

153								x														x			
154									x													x		x	
155							x	x																	
156								x														x			
157								x																	
158								x																	
159								x																	
160									x													x			
161							x	x														x			x
162	x																								
163								x													x				
164								x																	
165			x																			x			
166		x													x							x			
167							x	x														x			
168								x																	
169								x													x				
170	x		x					x			x											x			
171								x						x								x			
172						x																x			
173									x													x			
174	x							x														x			
175								x														x			
176								x														x			
177						x			x																
178							x		x													x			
179								x														x			
180								x	x													x			
181								x														x			
182							x																		
183	x							x																	
184	x								x													x			
185								x																	
186			x					x																	
187								x														x			
188											x								x			x			
189	x		x				x								x							x			
190								x																	
191						x		x												x					
192		x																				x			
193									x													x			
194						x		x																	
195								x		x												x			
196								x		x		x								x	x	x			
197								x														x			
198											x	x										x			
199							x	x														x			
200	x								x																

CHAPTER 4

Evolution of the ApiAP2 regulatory network in apicomplexan parasites

Jenna Oberstaller, Yoanna Pumpalova, Ariel Schieler, Manuel Llinás, Jessica C. Kissinger. To be submitted to PloS Pathogens.

ABSTRACT

Transcriptional regulation in apicomplexan parasites is still poorly understood. Seven years ago, genome-wide scans of taxa in the Apicomplexa for possible DNA-binding domains revealed the Apicomplexan AP2 (ApiAP2) family of transcriptional regulators. Subsequent experimental analyses have confirmed the regulatory roles of several of these proteins. Little is known about the evolutionary history of this family of proteins in the Apicomplexa. We have used Hidden Markov Models (HMMs) and phylogenetic tools to examine the distribution and evolutionary relationships of the AP2 DNA-binding domains across the Apicomplexa and an outgroup dinoflagellate *Perkinsus marinus*. We find that these AP2 domains fall into distinct evolutionary groups: more ancient classes of domains that span multiple taxa, and other classes of domains that are lineage-specific. We used this information to select and generate recombinant AP2 protein domains representing most of the family from a basal-branching apicomplexan, *Cryptosporidium parvum*. We determined the binding specificities of these domains and searched for the identified binding motifs upstream of co-regulated *C. parvum* gene clusters to identify putative regulatory targets and define the ApiAP2-based transcriptional regulatory network in this organism. We previously reported 25 overrepresented motifs upstream of co-regulated *C. parvum* genes; here we report ApiAP2s putatively bind several of these motifs. We also note that there is much redundancy in *C. parvum* ApiAP2 binding site recognition, particularly the 5′-TGCAT-3′, 5′-CACACA-3′, and G-box motifs (5′-G[T/C]GGGG-3′). The DNA-binding specificities and potential regulatory targets for several ApiAP2 domains were recently identified in the distantly related apicomplexan *Plasmodium falciparum*. We compare select orthologous and lineage-specific ApiAP2 domains and their gene targets between these organisms to gain insight into how this regulatory network may have evolved across the phylum.

INTRODUCTION

Apicomplexan parasites are the causative agents of some of the world's most devastating diseases, including malaria (caused by *Plasmodium*), toxoplasmosis (*T. gondii*) and cryptosporidiosis (*Cryptosporidium*). While RNA polymerase-associated factors and basal transcription factors were clearly identified in the apicomplexan phylum[1], examination of apicomplexan proteomes yielded a surprising dearth of known sequence-specific transcription factors characteristically found in other eukaryotes [2,3]. These findings were highly unexpected, given that these parasites have complex life cycles consisting of several developmental stages, and accordingly, extensive evidence existed for transcriptional control [4,5,6]. The lack of recognizable specific transcription factors in the midst of extensive transcriptional regulation initially suggested that the specific transcription factors responsible were so far diverged from those found in other eukaryotes that they were unrecognizable. Balaji et al. (2005) took an aggressive approach to tackle the issue of absent transcriptional regulators through sensitive sequence analysis for all known DNA-binding domains in the *P. falciparum* genome [3]. The team identified a family of proteins with members present in all apicomplexan genomes tested (*Plasmodium, Cryptosporidium,* and *Theileria*) that could potentially be acting as apicomplexan transcription factors. This family of proteins, called ApiAP2, is similar to the AP2 family of transcription factors first identified in plants.

The Apicomplexa have a very distinct evolutionary history involving expansive and often-ancient gene-transfer events from distantly related species, notably algae. At the initial discovery of ApiAP2s in 2005, the authors postulated that the factor is of plant origin, recruited to the apicomplexan nuclear genome after the secondary endosymbiosis of an alga (largely believed to be rhodophyte in origin) whose only remnants are the apicoplast organelle and many transferred genes [3]. The community has largely accepted this origin theory without further investigation. However, as more genome sequences have been generated, AP2 domains have been found throughout the tree of life, notably in several bacteria and their phages [3,7,8], and sequence similarity between these domains does not link ApiAP2 domains to plant AP2 domains to the exclusion of these other groups. In all other AP2 families identified, the AP2 domain is associated with homing endonuclease or integrase domains of mobile elements. There is no evidence in apicomplexan genomes of currently active mobile elements. However there is evidence that apicomplexan genomes used to contain them, and these elements have since been lost [9,10].

Since the discovery of the ApiAP2 proteins, much work has been done both computationally and experimentally to implicate these proteins in regulation. Two ApiAP2 proteins have been identified as master stage-specific regulators in *Plasmodium* (AP2-Sp and AP2-O (AP2-sporozoite and AP2-ookinete), [11,12]. Another ApiAP2 protein (PFF0200c) has been implicated as a player in *Plasmodium var* gene regulation by acting not as a transcription factor, but by binding the SPE2 DNA motif and interacting with epigenetic machinery to somehow ensure that only one of the sixty members of this family of surface antigens involved in immune evasion is expressed at a time [13]. Campbell and colleagues (2010) recently comprehensively characterized the ApiAP2 binding motifs for all 27 members of this family in *Plasmodium falciparum* and used these data in conjunction with expression data for genes expressed during the *P. falciparum* intra-erythrocytic blood stage to predict regulatory targets of these ApiAP2s [14]. ApiAP2 regulation has also been investigated to a lesser degree in *T. gondii*, where several ApiAP2 proteins have been implicated in progression through the cell cycle [15].

Though much has been learned from *Plasmodium* and *Toxoplasma*, there is a notable void in the field, especially in the even more distantly related apicomplexan, *Cryptosporidium*.

Studies aimed at characterizing protein function, such as the characterization of AP2-O and AP2-Sp referenced above, often involve genetic manipulation, a tool that is currently unavailable in the experimentally intractable *Cryptosporidium*. As the published data largely address ApiAP2 regulation only in *Plasmodium spp.*, there have been no extensive comparative studies between organisms, and the question of the evolution of this gene family has not been formally addressed. For instance, studies to date have not definitively addressed whether orthologous ApiAP2 proteins regulate similar sets of genes across apicomplexans, though the little data that exists suggest that ApiAP2 regulons may be quite different between organisms. The DeSilva group (2008) investigated the binding specificity of a single *C. parvum* domain that was highly conserved with a *Plasmodium* ApiAP2 of interest, PF14_0633 [16]. They found that the binding specificities were absolutely conserved between these two domains. However, of the 127 putative targets of PF14_0633 regulation, only 26 of these targets are conserved in *C. parvum*. These data suggest that while binding specificity of these orthologous domains is absolutely conserved between these distantly related organisms, the transcriptional network itself has evolved considerably since *Plasmodium* and *Cryptosporidium* diverged ~420 mya [17,18]. More study of *Cryptosporidium* ApiAP2 binding specificities and putative regulatory target genes is required to uncover the extent of this transcriptional rewiring.

In this study, we have used Hidden Markov Models (HMMs) and phylogenetic tools to examine the distribution and evolutionary relationships of the AP2 DNA-binding domains across the Apicomplexa and an outgroup perkinsid oyster parasite, *Perkinsus marinus*. We find that these AP2 domains fall into distinct evolutionary groups: more ancient classes of domains that span multiple taxa, and other classes of domains that are lineage-specific. We used this information to select and generate recombinant AP2 protein domains representing most of the family from the basal-branching apicomplexan *Cryptosporidium parvum*. We determined the binding specificities of these domains experimentally and searched for the identified binding motifs upstream of co-regulated *C. parvum* gene clusters to identify putative regulatory targets and define the ApiAP2-based transcriptional regulatory network in this organism. We previously reported 25 overrepresented motifs upstream of co-regulated *C. parvum* genes; here we report ApiAP2s putatively bind several of these motifs. We also note that there is much apparent redundancy in *C. parvum* ApiAP2 binding sites beyond that which has been noted in *Plasmodium* [14], particularly the 5′-TGCAT-3′, 5′-CACACA-3′, and G-box motifs (5′-G[T/C]GGGG-3′). Using the *P. falciparum* ApiAP2 DNA-binding specificities and potential regulatory targets from Campbell et al. (2010), we compare select orthologous and lineage-specific ApiAP2 domains and their gene targets between these Apicomplexa to gain insight into how this regulatory network may have evolved across the phylum.

RESULTS

Perkinsid and apicomplexan AP2 domain families evolved independently after speciation

No published studies to date have further investigated ApiAP2 origins after their initial discovery, and ApiAP2 origins are not the focus of this paper. However, Balaji et al suggest that eukaryotic versions of the AP2-integrase domain evolved from mobile elements and then underwent lineage-specific expansions and recruitment to roles in transcription. Within chromalveolates, if the AP2 domain were transferred from the plastid (which according to the chromalveolate hypothesis [19] was derived from an endosymbiotic event in the chromalveolate ancestor), it would be expected that we would find the domain in other chromalveolates. Consistent with this hypothesis, previous studies have indicated the presence of the AP2 domain in a number of chromalveolates [20]. No studies have specifically looked for the ApiAP2 domain across the chromalveolates, nor in the purported source, the algal endosymbiont. We thus examined the distribution of AP2 domains across several chromalveolates including apicomplexans, perkinsids, dinoflagellates, ciliates, and stramenopiles, as well as in the chromalveolate endosymbionts rhodophytes and chlorophytes using custom-built ApiAP2 HMMs as well as existing AP2 HMMs available from Pfam (www.pfam.org) (see Materials and Methods) (Table 4.1). Phylogenies constructed from the identified domain sequences indicate that perkinsid *Perkinsus marinus* AP2 domains are closely related to ApiAP2 domains, and both are more distantly related to other chromalveolate/endosymbiont AP2 domains (Figure 4.1). Further, phylogenies of perkinsid and ApiAP2 domains alone show that perkinsid domains group together exclusively, while ApiAP2 domains are more diverse and group across taxa (Figure 4.2). This observation suggests that perkinsid AP2 and ApiAP2 domains amplified independently in their respective lineages. Deep evolutionary relationships are difficult, if not impossible to recover due to the short length (~60 amino acids) of the domain. These domains (which can exist from one to four or more per protein) are often the only globular domains in the protein [3], and ApiAP2 proteins are highly divergent in both sequence and length (ranging from ~400 to thousands of amino acids) across apicomplexans outside of the ApiAP2 domain (data not shown).

ApiAP2 domains fall into evolutionary clades

Homology analyses between *P. marinus* and apicomplexan AP2 domains suggest distinct homolog groups spanning different taxa. As many as 13 domains are apicomplexan-specific. Additionally, there are several intra-phylum lineage-specific domains (Figure 4.3, table 4.2). Domain counts and composition of homolog groups varied depending on the stringency of e-value parameters used to assign orthologs to clusters; thus we indicate ranges of domains determined by OrthoMCL clustering at 1e-4 to 1e-11 rather than precise counts in Table 4.2. We determined that 1e-6 is the most stringent e-value at which homology between apicomplexan and perkinsid AP2 can be detected, and thus we chose homolog groups determined at 1e-6 for further analyses.

The 23 *C. parvum* ApiAP2 domains were further classified as ancestral, pan-apicomplexan, or lineage-specific based on their phyletic distribution with OrthoMCL clustering using an e-value cutoff of 1e-6 (Figure 4.4). Those domains that fell into a homolog group with any outgroup *P. marinus* domain were classified as ancestral; these domains likely predate the divergence between Perkinsids and the Apicomplexa. Four *C. parvum* domains (cgd4_1110_D1, cgd4_1110_D3, cgd8_3130 and cgd8_3230) fell into this category. Domains that span all or most apicomplexan lineages, but were absent in *Perkinsus* were classified as pan-apicomplexan (10 domains). It is necessarily true that some pan-apicomplexan domains may have been present in the perkinsid/apicomplexan ancestor as well, and were subsequently lost in *Perkinsus*. But because there is no extant evidence of the domain in *Perkinsus* and there is more ambiguity in when these domains arose, we maintain separate “ancestral” and “pan-apicomplexan” designations. Lineage-specific domains have no orthologs outside their respective taxa (nine domains), though again it is a formal possibility that these could also be true “ancestral” domains that were lost in other lineages.

***C. parvum* ApiAP2 domains bind diverse sequences**

De Silva et al. (2008) determined the DNA binding specificity of *C. parvum* ApiAP2 domain cgd2_3490, and we previously reported the DNA-binding specificity of cgd8_810 ([16]; Chapter 3). We created constructs for the 21 remaining *C. parvum* ApiAP2 domains as well as cgd2_3490 as a control to determine binding specificities on protein-binding microarrays (PBMs). Our results agree with the previously reported 5′-TGCAT-3′ core binding motif for cgd2_3490, providing support for our methods. We detected binding specificity for 16 of these domains (Figure 4.4).

As previously determined for *P. falciparum* ApiAP2 domains, we find that *C. parvum* ApiAP2 domains also have the capability to bind a diversity of sequences similar to what is seen in *P. falciparum*. Though the *C. parvum* ApiAP2 family can recognize a variety of sequences, we found that out of the 16 domains for which we detected binding motifs, 11 of these bind one of three motif types: the 5′-TGCAT-3′ motif (four different domains), the 5′-CACACA-3′ motif (four domains), or the G-box motif (5′-G[T/C]GGGG-3′; 3 domains). *P. falciparum* also has four CACACA-binding ApiAP2 domains, but this is the only markedly redundant *P. falciparum* ApiAP2 binding motif [14].

Secondary and tertiary motif recognition

Multiple binding specificities above threshold were previously reported for several *P. falciparum* ApiAP2 domains [14]. Many of these secondary or tertiary binding sites had little similarity, indicating an additional layer of complexity to ApiAP2 regulation. *C. parvum* ApiAP2s also display multiple motif recognition, though in the majority of cases secondary motifs are highly similar to or are reverse complements of the primary motif (Table 4.3). We found that only one *C. parvum* domain, cgd1_3520, is able to recognize two completely different motifs, both the 5′-TGCAT-3′ motif and the G-box.

Binding sites between putative *Pf* and *Cp* orthologs are often conserved

It was noted previously that orthologous ApiAP2 domains across *P. falciparum*, *P. berghei*, and *C. parvum* (gene ids PF14_0633, PBANKA_132980 and cgd2_3490 respectively) have nearly identical binding specificities for the 5′-TGCATGCA-3′ motif [11,16]. Our phylogenetic analyses support the orthology of this domain group, and we found an additional putative *C. parvum* ortholog to PF14_0633 (cgd1_3520) that also recognizes this motif (Figure 4.5). We found that putative ortholog pair cgd8_3130 and PF14_0533, as well as putative ortholog pair cgd8_3230 and PFE0840c-D2 bind the same, or highly similar, motifs. In another interesting case, *C. parvum* ApiAP2 domain cgd4_3820 recognizes the sequence 5′-GGTGCACC-3′, while its putative *P. falciparum* ortholog PFF-0200c_D2 individually has no known binding site. However, a construct of both PFF0200c ApiAP2 domains joined by a short conserved linker region does bind the same site as cgd4_3820.

We find binding specificity is not conserved between putative orthologs cgd4_1110_D3 and PFE0840c_D2, and binding specificity between putative orthologs cgd5_4250 and PF14_0079 is weakly conserved. We additionally found no binding specificity above threshold in *C. parvum* for two domains (cgd6_5320_D3 and cgd6_5320_D4) whose putative orthologs (PF11_0404 and PFL_1900w, respectively) do have binding motifs. These ill-conserved binding specificities may indicate that these domains are not true orthologs. Alternatively, the lack of conservation may be a true snapshot of evolving binding specificities, especially given the significant support of conserved binding specificities for the other putative ortholog groups.

Though putative orthologous ApiAP2 domains often have similar binding specificities, evolutionary distance does not always predict binding specificity. We constructed a maximum likelihood tree of all predicted *P. falciparum* and *C. parvum* ApiAP2 domains and superimposed their binding motifs to examine the relationship between evolutionary distance and binding motif (Figure 4.6). We found that ApiAP2 domains that recognize similar motifs are usually interspersed throughout the tree. Putative orthologs PF14_0633, cgd2_3490, cgd1_3520, and cgd8_3230 all bind 5′-TGCAT-3′-like motifs, and they are clustered together on the tree, though we also find TGCAT-binding ApiAP2s that are more distantly related to this group. The G-box and CACACA-binding ApiAP2s are more distantly related. These phyletic distributions could be explained by duplication of domains and divergence of their binding sites both within each species and between speciation events. Determining the families of ApiAP2 binding motifs for intermediate taxa, such as *T. gondii* or the piroplasms, may further elucidate the relationship between ApiAP2 binding sites and evolutionary history.

Multiple ApiAP2 domains can bind *C. parvum* overrepresented upstream motifs

We previously reported 11 families of overrepresented motifs found in the upstream regions of *C. parvum* genes. Two of these motif families are known non-AP2 binding motifs (the E2F-like and CAAT-box-like motifs). While it was already known that another of these motif families, comprising those motifs resembling 5′-TGCAT-3′ (designated "AP2_1" in the previous chapter), is an ApiAP2 binding site, we found three additional *C. parvum* ApiAP2s that bind this motif (Figure 4.7). Clusters of co-expressed genes containing these motifs in their upstream regions show maximal expression, individually, at any of the surveyed timepoints across the lifecycle. The same is true of the genes encoding the four TGCAT-binding ApiAP2s (see cgd8_3230, cgd1_3520, cgd5_4250, and cgd2_3490 in Figure 4.9). We additionally reported that ApiAP2 cgd8_810 binds the overrepresented G-box motif (Chapter 3), and we find that cgd6_5320_D2 and cgd2_2990 also recognize the G-box. Cgd6_5320 and cgd2_2990 both have bimodal expression patterns, peaking at 6 and 24 hours post-infection, while cgd8_810 is expressed at multiple later time points. Clusters containing overrepresented G-box motifs in the upstream regions of their genes are also maximally expressed, individually, at any of the surveyed timepoints across the lifecycle. These results suggest that regulation of these differentially expressed gene clusters might be handled by the respective co-expressed ApiAP2.

Stage-specific regulation could explain redundant *C. parvum* ApiAP2s binding specificities, at least in the case of the AP2_1-like and G-box-like motifs.

We did not detect ApiAP2 binding motifs similar to any of the other nine unaccounted-for overrepresented upstream motifs. It is also interesting to note that the 5′-CACACA-3′ motif is not overrepresented upstream of the 200 co-regulated *C. parvum* gene clusters we previously identified, though four different ApiAP2s can bind this motif. We were able to predict putative regulatory targets for two of these CACACA-binding ApiAP2s, cgd8_3130 and cgd4_600. The other CACACA-binding ApiAP2s, cgd5_2570 and cgd6_2600, have no predicted targets below statistical threshold. Most of these putative targets have a bimodal expression pattern, peaking at 12 and 36 hours post-infection (data not shown). ApiAP2s Cgd8_3230 and cgd4_600 are expressed during these time points, and thus could plausibly be involved in regulation of these genes.

ApiAP2 network evolution: Comparisons between predicted *C. parvum* and *P. falciparum* regulatory targets for orthologous and lineage-specific ApiAP2s

Behnke et al. (2010) found that genes expressed throughout the *T. gondii* cell cycle define two subtranscriptomes expressed in two separate waves: genes responsible for basal processes such as DNA replication, protein translation and glycolysis; and genes specific to apicomplexan processes, such as those involved in invasion or immune evasion [15]. They noted that 24 ApiAP2 proteins are expressed in a cascade across the cell cycle. These findings raise the intriguing possibility that evolutionary history of ApiAP2 domains is somehow correlated with the evolutionary history of their regulatory targets—ie, that ancestral or pan-apicomplexan ApiAP2 domains might be responsible for regulating basal housekeeping processes, while lineage-specific ApiAP2 domains might regulate apicomplexan-specific processes. To further investigate this possibility, we used a modified version of the algorithm Campbell et al. (2010) developed to predict regulatory targets (which incorporates genome-wide expression data and presence of ApiAP2 binding sites in upstream regions) for a number of *C. parvum* ApiAP2 domains [14]. We selected lineage-specific and shared ApiAP2 domains from both *C. parvum* and *P. falciparum* and evaluated the category composition of their predicted target genes (see Materials and Methods). We did not find a significant correlation between evolutionary class of ApiAP2 and putative targets in either organism (Figure 4.8), providing further evidence that the ApiAP2 network has been shuffled and evolved considerably over time.

The ApiAP2 expression cascade is conserved in *C. parvum*

It is known that there are a number of other possible transcription factor families in the *C. parvum* genome (reviewed in Chapter 3), some of which are absent in other apicomplexans (E2F, for example). The ratio of available *C. parvum* transcription factors to regulate target genes is much higher than the *P. falciparum* ratio (around 1:340 and 1:800 respectively), due both to the lower gene count in *C. parvum* and a higher absolute number of possible transcription factors [21]. We have also determined that the E2F binding motif is one of the most overrepresented motifs in the upstream regions of the *C. parvum* genome (Chapter 3). Given these observations, it might be expected that *C. parvum* is less reliant on the ApiAP2 family for transcriptional regulation than *P. falciparum* and other apicomplexans. However, expression data for each predicted *C. parvum* ApiAp2 indicate that the expression cascade observed across the *P. falciparum* blood stage [14] and across the *T. gondii* cell cycle [15] is conserved in *C. parvum* (Figure 4.9), though putative orthologous ApiAP2s do not necessarily appear at similar positions in the cascades. This overall conservation suggests that *C. parvum* ApiAP2s are significant players in transcriptional regulation despite presence of other transcription factors.

DISCUSSION

Understanding gene regulatory mechanisms in apicomplexan parasites in general and *Cryptosporidium* in particular has proven to be a challenge. There is no continuous *in vitro* cultivation system available for *Cryptosporidium*, and molecular genetic tools to investigate gene regulatory mechanisms are nonexistent. Even when considering model organisms for which there are myriad genetic tools, few large transcription factor family networks have been characterized in depth [14,22,23,24,25,26,27]. Here we have provided the first comprehensive analysis of a transcription factor family in *C. parvum*, using gene expression data, binding specificity, and phylogenetic tools.

We have placed ApiAP2 regulation in a kingdom-wide context using evolutionary analyses of distribution and relationships between AP2 domains. Phylogenies constructed from AP2 domains spanning chromalveolates and endosymbionts indicate a distinct divide between AP2s found in the plant lineage, stramenopiles, ciliates, and dinoflagellates and those found in the Apicomplexa. The perkinsid domains group more closely with the apicomplexans than the other chromalveolates. However, and quite interestingly, they also group more closely with one another than with any of the other domains surveyed. Some orthologous domains span several apicomplexan taxa, indicating domains predating speciation events. The observation that *P. marinus* domains group together exclusively suggests that the amplification of AP2s in *P. marinus* and apicomplexans occurred independently.

These studies are only cursory looks into the evolutionary origins of the ApiAP2 domain, as only a handful of chromalveolate genomes were surveyed, and neither bacterial nor other integrase-associated AP2 domains were included in these analyses other than those from the ciliate *T. thermophila* [7]. Thus, these results should not be used to make definitive statements about ApiAP2 origins. However these results do suggest that by whichever manner ApiAP2 domains came to reside in the apicomplexan/perkinsid ancestor, whether by mobile element invasion, transfer from an algal endosymbiont, or some mixture of these events, perkinsid and apicomplexan AP2 domains likely share an origin, and perkinsid and apicomplexan AP2 domains have amplified independently since they separated. Based on our homology analyses, we propose that there were 1-3 progenitor domains arising from the acquisition event (vertical or lateral) in the perkinsid/apicomplexan ancestor. The domain in the perkinsid and apicomplexan lineages then amplified independently from that point. The apicomplexan ancestor possessed a minimum of 11-13 domains. Proposing a more exact count of ancestral domains remains difficult and will likely require more diverse sampling across the phylum and structural analysis of the domain. Though domains spanning most apicomplexans or spanning perkinsids and apicomplexans are likely to be ancestral, domains spanning other combinations of taxa may be either ancestral or the result of recent amplification. After the apicomplexan ancestor with its complement of ApiAP2 domains speciated, domains amplified independently or were lost in separate lineages. The most striking amplifications have occurred in the coccidian and *Plasmodium* lineages, with anywhere from 44 to 69 of the ~90 coccidian domains and 18 to 29 of the ~50 *Plasmodium* domains being lineage-specific.

The current lack of molecular genetic tools in *Cryptosporidium* imposes a critical barrier to further functional characterization of predicted ApiAP2 transcription factors and their putative regulatory targets. Our target predictions are based on expression data from the limited *in vitro* lifecycle, and it must be considered that the transcriptome may differ significantly from what is expressed *in vivo*. It is also important to note that although proteomics data from the very early stages of the *C. parvum* life cycle are available [28,29], there are no proteomics data for the majority of the life cycle. Thus we do not know how closely mRNA expression indicates protein expression in *C. parvum*, and the expectation that ApiAP2 mRNA expression profiles should correlate highly with those of predicted target genes may be flawed. The correlation between mRNA and protein expression in *P. falciparum* was found to be moderately positive, though a delay between first-detected mRNA expression and protein expression has been observed for several genes, indicating the importance of post-transcriptional regulatory mechanisms in *Plasmodium* as well [5,30]. *C. parvum* ApiAP2 mRNA expression does not correlate well with predicted target gene expression profiles in many cases (data not shown), unlike what has previously been indicated for *P. falciparum* ApiAP2s [14].

Though caution should be exercised in interpretation of our ApiAP2 network analyses, these data still indicate the power of computational tools to study transcriptional regulation in the absence of molecular genetic tools. Here, we have presented evidence that ApiAP2s are major players in *C. parvum* transcriptional regulation, namely: (1) the ApiAP2 regulatory cascade is conserved in *C. parvum*, and (2) *C. parvum* ApiAP2s bind a diverse set of motifs, many of which are overrepresented upstream of many co-expressed gene clusters. In conjunction with our phylogenetic analyses, these results contribute to the beginnings of a framework for understanding ApiAP2 regulation in other apicomplexans. Over the years, binding motifs have been identified for several members of a single ApiAP2 ortholog group (PF14_0633 in *P. falciparum*, cgd2_3490 in *C. parvum*, TGME49_110950 in *T. gondii*, and AP2-Sp in *P. berghei*; [12,15,16]), all of which bind the 5′-TGCAT-3′ motif. Our results build on these previous observations—putative orthologous domains on a network scale have conserved binding specificities between two of the most distantly related apicomplexans, *P. falciparum* and *C.*

parvum, which indicates that binding specificities (and by extension, regulatory data where expression data exists) can often be predicted by orthology.

We have presented broad-scale comparisons of ApiAP2 network composition between *P. falciparum* and *C. parvum* and suggest that there is no relationship between evolutionary class of ApiAP2 domain and evolutionary class of predicted targets. We previously reported evidence of a transcription factor substitution in the ribosomal protein regulon between a *P. falciparum* G-box-binding ApiAP2 and *C. parvum* E2F (Chapter 3). Ribosomal gene regulon transcription factor substitution has been noted in yeast [31,32], and thus is not particular to apicomplexans. Campbell et al. (2010) reported extensive divergence between predicted orthologous ApiAP2 regulons in *P. falciparum*, *P. vivax* and *P. yoelli*, indicating that there is extensive network divergence even on relatively small evolutionary time scales (~100 million years). Conservation of transcription factor binding in the face of extensive regulon divergence has been noted across several organisms [33,34,35,36]. Additional analyses focusing on comparisons between specific orthologous ApiAP2 regulons, such as those on the ribosomal proteins, should be undertaken to further investigate the extent and patterns of network divergence across Apicomplexa.

Many *C. parvum* ApiAP2s bind redundant motifs, and the majority of *C. parvum* ApiAP2 domains bind only one motif. Thus *C. parvum* ApiAP2 regulation does not appear to be as multi-faceted as suggested in *P. falciparum* [14]. The presence of more non-ApiAP2 transcription factors in the *C. parvum* genome may explain the decreased diversity of ApiAP2 binding. We noted previously that the E2F motif is the most abundantly overrepresented in the upstream regions of the *C. parvum* genome, being found upstream of 161 of 200 predicted co-regulated gene clusters (Chapter 3). E2Fs are notably absent in *Plasmodium* and other apicomplexans [21]. It is possible that the two predicted E2F transcription factors are responsible for a disproportionate amount of the transcriptional regulation, such that *C. parvum* is less reliant on ApiAP2s. The apparent redundancy in *C. parvum* ApiAP2 binding motifs may also be important to stage-specific transcriptional regulation, as ApiAP2s binding the same or similar motifs are expressed at various points across the lifecycle.

MATERIALS AND METHODS

Identification of AP2 and ApiAP2 domains

To identify ApiAP2 domains for phylogenetic analyses, we developed a Hidden Markov Model (HMM) that appears to be more sensitive to the specific detection of ApiAP2s than the Pfam-available HMM designed for the detection of AP2 domains (www.pfam.org). We first ran the existing AP2 HMM on the annotated protein sequences of apicomplexans *T. gondii*, *N. caninum, P. falciparum, P. vivax, C. parvum, T. annulata,* and *T. parva*. We next constructed an alignment with the T-coffee package [37] of the most significant domain hits from this run (1e-4 or lower). The ApiAP2 HMM was built from this alignment using HMMER (version 2.0). We used this new HMM in conjunction with the Pfam AP2 HMM to search annotated protein sequences to examine the distribution of the AP2 domain across several chromalveolates, including apicomplexans *Plasmodium falciparum*, *P. knowlesi*, *P. vivax*, *P. yoelli, Theileria parva*, *T. annulata*, *Babesia bovis*, *Neospora caninum*, *Toxoplasma gondii, Cryptosporidium muris* and *C. parvum*; the perkinsid oyster parasite *Perkinsus marinus*; dinoflagellates *Karenia brevis* and *Alexandrium tamarense*; ciliates *Tetrahymena thermophila* and *Paramecium tetraurelia*; and stramenopiles *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. Purported algal endosymbionts *Cyanidioschyzon merolae*, *Porphyra purpurea*, *P. yezeoensis* (representative rhodophytes), and *Chlamydomonas reinhardtii* and *Micromonas sp. RCC299* (representative chlorophytes) were also examined. As no annotated protein sequences were available at the time of our analyses, perkinsid and dinoflagellate analyses were run on 6-frame translations of clustered ESTs. AP2 protein and domain counts for each organism were determined using a permissive e-value cutoff of 10.

Phylogenetic analysis of AP2 and ApiAP2 domains

Determination of homolog groups

All phylogenetic analyses were carried out on AP2 and ApiAP2 domain sequences only, as full-length proteins are generally too divergent to be able to detect meaningful evolutionary relationships between them (data not shown). Alignments of AP2 domain sequences were performed using the T-coffee package [37] and edited using Jalview [38]. Unrooted neighbor-joining trees were constructed from top-scoring unambiguously aligned domain sequences across chromalveolates and outgroup green alga *C. reinhardtii* using PHYLIP [39]. Bootstrap support was obtained from 100 replicates. Further analyses were carried out on perkinsid and ApiAP2 domains alone as above using *P. marinus* as an outgroup.

To identify homologous clusters of ApiAP2 domains, a local install of the OrthoMCL algorithm [40] was run on all identified ApiAP2 domains in apicomplexans and perkinsids using an e-value ranging from 1e-04 to 1e-11. Domains displaying similarity at these e-values were clustered into homolog groups. Homolog groups found at 1e-06 were used for subsequent analyses, as this is the highest stringency at which orthology could be detected between apicomplexan and *P. marinus* ApiAP2 domains. Relationships were visualized using Circos [41].

Determination of *C. parvum* ApiAP2 binding motifs

N-terminal GST fusion proteins were made using the pGEX4T-1 vector (GE Healthcare) and the 23 predicted *C. parvum* ApiAP2 domains and their flanking residues. Many flanking residues were included to ensure capture of each domain. Domain boundaries were determined using custom-built HMMs run on all annotated *C. parvum* proteins (downloaded from CryptoDB.org, version 4.6). The domains and flanking sequence were PCR-amplified and cloned into the BamHI restriction site in pGEX4T-1. Proteins were expressed and purified as previously described [16]. Briefly, *E. coli* BL21 (RIL Codon PLUS, Stratagene) cells were induced with 200 mM IPTG at 25C. Proteins were then purified using Uniflow Glutathione Resin (Clontech) and eluted in 10mM reduced glutathione, 50mM Tris HCL, pH 8.0. Proteins were verified with western blots using an anti-GST antibody (Invitrogen), and purity was verified by silver stain.

A minimum of two protein-binding microarray experiments were performed with each purified protein construct to determine their binding specificities as previously described [14,16]. Motifs bound at a threshold of .45 or greater were considered significant. Similarity between *C. parvum* ApiAP2 binding sites was determined using the web-based STAMP tool [42]. Comparisons between orthologous *C. parvum* and *P. falciparum* ApiAP2 binding sites (using *P. falciparum* ApiAP2 binding motif data from [14,16]), as well as comparisons between *C. parvum* ApiAP2 binding sites and *C. parvum* overrepresented upstream motifs (Chapter 3) were also made using STAMP.

Predictions of putative ApiAP2 target genes

Definition of C. parvum upstream regions

Upstream regions were designated as in the previous chapter. Briefly, we downloaded the *C. parvum* genome (v 4.2) and nucleotide sequences for all protein-encoding genes from CryptoDB (http://cryptodb.org/cryptodb/, [43]). Custom Perl scripts were used to extract (1) 1kb of sequence upstream of each translation start site, or (2) the upstream sequence until a gene was encountered on either strand. The translational start site was used because we do not have UTR information for predicted genes. The *C. parvum* genome is only 9.1 Mb and is highly compact with very few introns and small intergenic spaces. To exclude the possibility of including coding regions in this set due to mis-annotation, a BLASTX was performed against the NCBI NR database using the set of upstream sequences as the query. Upstream sequences that contained significant portions of 100% identity to coding sequences were eliminated.

Target gene prediction

We modified the target prediction algorithm used in [14] for use with our data to identify putative AP2 target genes. This algorithm takes position weight matrices derived from PBM scores for each AP2 domain and searches for matches in the upstream sequence database. Each AP2 is assigned a score for each gene based on motifs found. The glmnet package in R [44] is then implemented to make a regression between this AP2 motif score and the expression pattern for each gene (*C. parvum* expression data from [45]) to determine how much the AP2 motif contributes to each gene's expression. An average expression pattern for genes possessing a particular AP2 motif upstream is then iteratively built, and genes that match this average expression pattern within a statistical threshold are designated as putative regulatory targets. *P. falciparum* regulatory targets were previously defined using a false discovery rate of 1% [14]. As we have comparatively few time points over which we have expression information (7 for *C. parvum* vs. 47 for *P. falciparum*) and thus have less statistical power, we considered genes falling within a false discovery rate of 20% as putative regulatory targets.

Evaluating evolutionary history of AP2 domains vs. evolutionary history of their putative target genes

Putative target genes of shared ("ancestral" or "pan-apicomplexan") and lineage-specific ApiAP2 domains were compared against lists of three different evolutionary classes of apicomplexan genes as determined by OrthoMCL: (1) those shared between all of 12 apicomplexans (the 11 used for all other analyses, as well as *P. berghei*); (2) genes shared between apicomplexans of at least two different genera, and (3) genus-specific genes (genes which have no orthologs outside of their respective genus). Putative targets were then classified as "shared" or "lineage-specific".

REFERENCES

1. Meissner M, Soldati D (2005) The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. Microbes and Infection 7: 1376-1384.

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

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

4. Llinas M, DeRisi JL (2004) Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. Current Opinion in Microbiology 7: 382-387.

5. Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, et al. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82-86.

6. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, et al. (2005) The transcriptome of *Toxoplasma gondii*. Bmc Biology 3: 18.

7. Magnani E, Sjolander K, Hake S (2004) From endonucleases to transcription factors: Evolution of the AP2 DNA binding domain in plants. Plant Cell 16: 2265-2277.

8. Wuitschick JD, Lindstrom PR, Meyer AE, Karrer KM (2004) Homing endonucleases encoded by germ line-limited genes in *Tetrahymena thermophila* have APETELA2 DNA binding domains. Eukaryotic Cell 3: 685-694.

9. Roy SW, Penny D (2006) Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. Genome Research 16: 1270-1275.

10. Roy SW, Penny D (2007) Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. Molecular Biology and Evolution 24: 1926-1933.

11. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, et al. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Molecular Microbiology 71: 1402-1414.

12. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites. Molecular Microbiology 75: 854-863.

13. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BTF, et al. (2010) A Major Role for the *Plasmodium falciparum* ApiAP2 Protein PfSIP2 in Chromosome End Biology. Plos Pathogens 6.

14. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. Plos Pathogens 6.

15. Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, et al. (2010) Coordinated Progression through Two Subtranscriptomes Underlies the Tachyzoite Cycle of *Toxoplasma gondii*. Plos One 5.

16. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. (2008) Specific DNA-binding by Apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 105: 8393-8398.

17. Berney C, Pawlowski J (2006) A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc Biol Sci 273: 1867-1872.

18. Okamoto N, McFadden GI (2008) The mother of all parasites. Future Microbiology 3: 391-395.

19. Keeling PJ (2009) Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis. Journal of Eukaryotic Microbiology 56: 1-8.

20. Iyer LM, Anantharaman V, Wolf MY, Aravind L (2008) Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. International Journal for Parasitology 38: 1-31.

21. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Research 14: 1686-1695.

22. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, et al. (2008) Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell 133: 1266-1276.

23. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and Complexity in DNA Recognition by Transcription Factors. Science 324: 1720-1723.

24. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, et al. (2008) A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters. Molecular Cell 32: 878-887.

25. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, et al. (2008) Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. Cell 133: 1277-1289.

26. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Research 19: 556-566.

27. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104.

28. Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, et al. (2008) Determining the protein repertoire of *Cryptosporidium parvum* sporozoites. Proteomics 8: 1398-1414.

29. Snelling WJ, Lin Q, Moore JE, Millar BC, Tosini F, et al. (2007) Proteomics analysis and protein expression during sporozoite excystation of *Cryptosporidium parvum* (Coccidia, Apicomplexa). Mol Cell Proteomics 6: 346-355.

30. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, et al. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res 14: 2308-2318.

31. Planta RJ, Goncalves PM, Mager WH (1995) Global regulators of ribosome biosynthesis in yeast. Biochem Cell Biol 73: 825-834.

32. Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci U S A 102: 7203-7208.

33. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, et al. (2007) Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science 317: 815-819.

34. Moses AM, Pollard DA, Nix DA, Iyer VN, Li X-Y, et al. (2006) Large-Scale Turnover of Functional Transcription Factor Binding Sites in *Drosophila*. PLoS Comput Biol 2: e130.

35. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 39: 730-732.

36. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science 328: 1036-1040.

37. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.

38. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2, a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

39. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

40. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

41. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: An information aesthetic for comparative genomics. Genome Research.

42. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Research 35: W253-W258.

43. Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC (2004) CryptoDB: the *Cryptosporidium* genome resource. Nucleic Acids Research 32: D329-D331.

44. R Development Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

45. Mauzy MJ, Enomoto S, Lancto CA, Abrahamsen MS, Rutherford MS (2012) The *Cryptosporidium parvum* Transcriptome during In Vitro Development. Plos One 7.

Figure and table legends

Figure 4.1. Unrooted neighbor-joining tree of AP2 domains. AP2 relationships across chromalveolates and algae. Color is as indicated. Bootstrap support is indicated by symbols on nodes and described in the key. Constructed from 64 of the top-scoring unambiguously aligned domain sequences using green alga *C. reinhardtii* as an outgroup. There is clear separation between apicomplexan and perkinsid AP2 domains and the rest of the chromalveolates.

Figure 4.2. Unrooted neighbor-joining tree of AP2 Domains. Color is as indicated. Bootstrap support is indicated by symbols on nodes and described in the key. An alignment of 102 top-scoring domains from *Perkinsus* and apicomplexans (~8 domains per taxa) as determined by HMM analyses was constructed and edited as described in Materials and Methods. *P. marinus* was used as an outgroup.

Figure 4.3. Circos diagram of ApiAP2 domain homolog groups across the Apicomplexa. Rectangular boxes represent chromosomes. Organisms are color-coded. Orthologs are connected by colored lines in the interior of the diagram. **A.** ApiAP2 domain homolog groups across apicomplexans as determined by OrthoMCL clustering at 1e-11. All identified ApiAP2 domains were put through clustering. CM=*C.* muris, CP=*C. parvum,* PF=P. *falciparum,* PV=*P.* vivax, PK=*P. knowlesi,* TA=*T. annulata,* TP=*T. parva,* BB=B. *bovis*, , NC=*N. caninum,* TG=*T. gondii* . Inner circle represents genes on the plus (purple) and minus (green) strands containing ApiAP2 domains. **B.** Only links between shared domains are shown (black). **C.** Links between domains spanning all or most taxa are omitted. Links indicate homology between domains only and widths of links are not to scale. Refer to Table 4.2 for counts of domains falling into these groups.

Figure 4.4. *C. parvum* ApiAP2 domain binding sites as determined by protein-binding microarray. Boxes are color coded according to evolutionary groups based on OrthoMCL clustering at 1e-6 as discussed in Materials and Methods.

Figure 4.5. *C. parvum* ApiAP2 binding sites compared to *P. falciparum* ortholog binding sites. Boxes are color coded according to evolutionary groups based on OrthoMCL clustering at 1e-6 as discussed in Materials and Methods.

Figure 4.6. Maximum likelihood tree of *P. falciparum* and *C. parvum* ApiAP2 domains and their corresponding DNA-binding sites. Domain sequences were extracted from full-length proteins using HMM-defined coordinates and aligned using T-coffee. The alignment was edited using JalView. A maximum likelihood tree was constructed from the edited alignment using PhyML with an LG protein evolution model, then visualized using FigTree.

Figure 4.7. Overrepresented *C. parvum* motifs bound by ApiAP2 domains. Overrepresented motifs known to be recognized by non-AP2 proteins are not included.

Figure 4.8. Evolutionary classification of predicted target genes for select lineage-specific and shared ApiAP2s. Evolutionary classifications for target genes as determined in Materials and Methods. Blues indicate "shared" classes of genes. Orange indicates *Cryptosporidium*-specific genes. Pink indicates *Plasmodium*-specific genes.

Figure 4.9. Cascade of ApiAP2 protein expression across the *P. falciparum* and *C. parvum* life cycles. A. AP2 expression cascades. **B.** The *C. parvum* life cycle stages. Following ingestion, excystation (a) occurs. The sporozoites are released and parasitize epithelial cells (b ,c) of the gastrointestinal tract and some other tissues. The parasites undergo asexual multiplication (merogony) 24-48 hr post-infection (d ,e ,f) and then sexual multiplication producing microgamonts (male) (g) and macrogamonts (female) (h) 48-72 hr. Upon fertilization of the macrogamonts by the microgametes (i), oocysts (j,k) develop that sporulate in the infected host. Two different types of oocysts are produced: the thick-walled (J), which is commonly excreted from the host, and the thin-walled (k) oocyst , which is primarily involved in autoinfection. Modified from http://www.dpd.cdc.gov/DPDx/HTML/Cryptosporidiosis.htm.

Table 4.1. Distribution and quantification of AP2 proteins and domains across chromalveolates and algal endosymbionts. Counts of AP2 domain-containing proteins and the number of AP2 domains per species as determined by sensitive sequence profile analysis. Analyses on most species were run on full annotated protein sets. **Dinoflagellate analyses were run on clustered EST data. **P. marinus* analyses were run on clustered EST data. These counts represent profile matches at or below a permissive e-value of 10. ~85% of the hits were at or below 1e-3. Ranges are given for *P. marinus* domains because though up to 14 domains matched the profile at or below an e-value of 10, only 3 of these domains were full-length.

Table 4.2. ApiAP2 domain counts by evolutionary class. ApiAP2 domain evolutionary classes across apicomplexans as determined by OrthoMCL clustering at e-values ranging from 1e-6 to 1e-11. All identified ApiAP2 domains and perkinsid AP2 domains were subjected to clustering.

Table 4.3. *C. parvum* ApiAP2 secondary motifs. All motifs detected above the enrichment score threshold of .45 are shown, along with relationship of these motifs to primary motifs depicted in Figure 4.4.

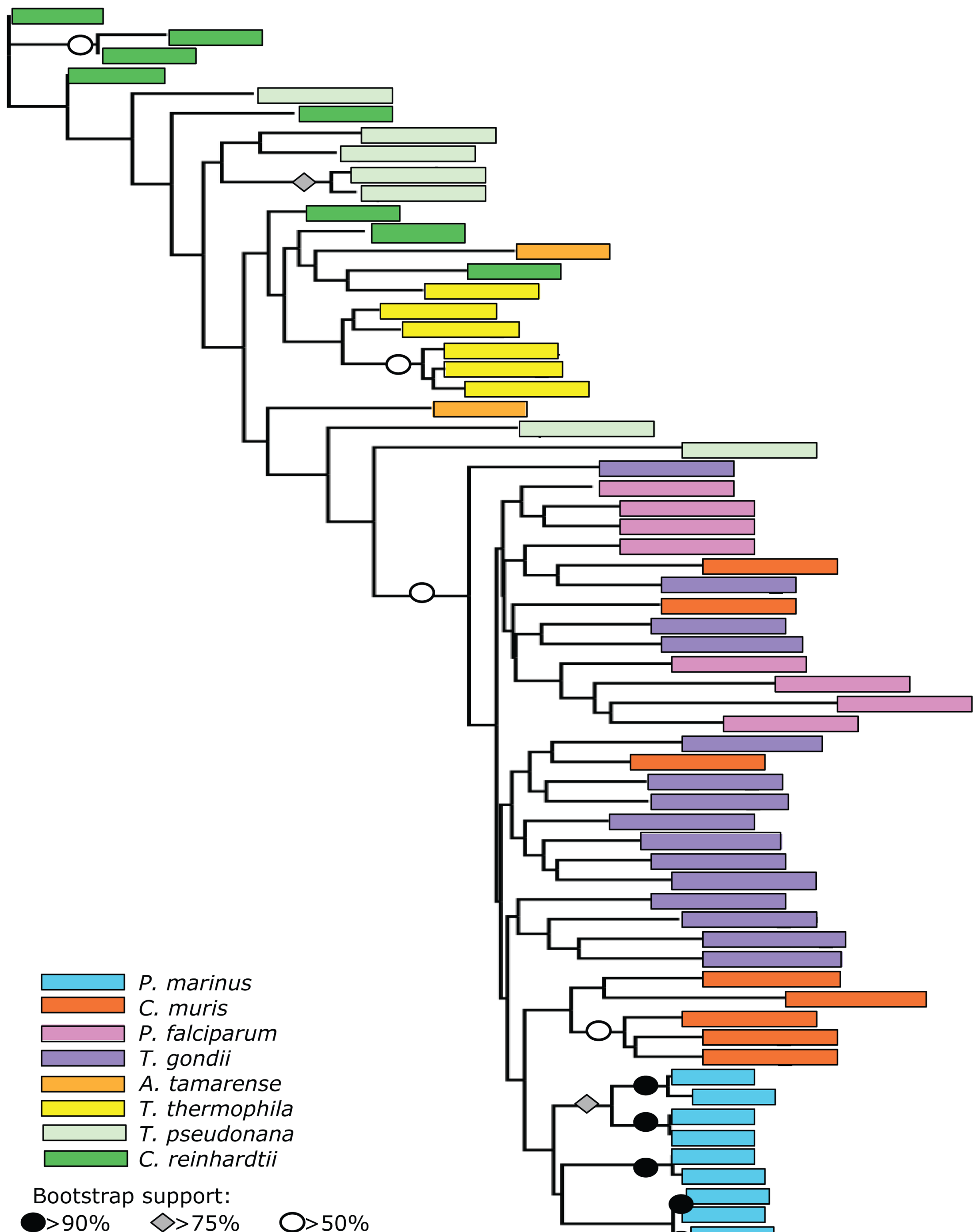


Figure 4.1. Unrooted neighbor-joining tree of AP2 domains.

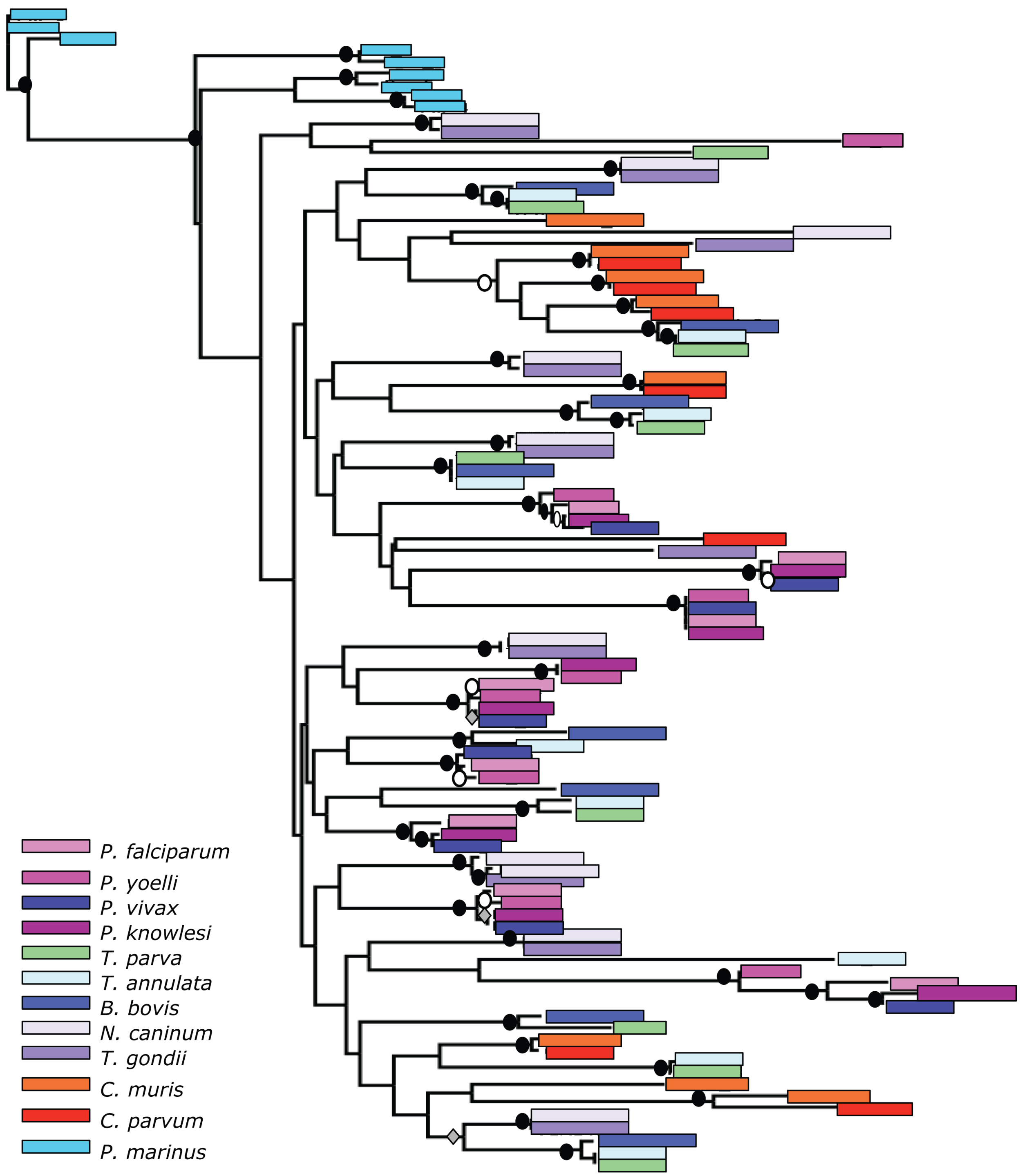


Figure 4.2. Unrooted neighbor-joining tree of AP2 domains.

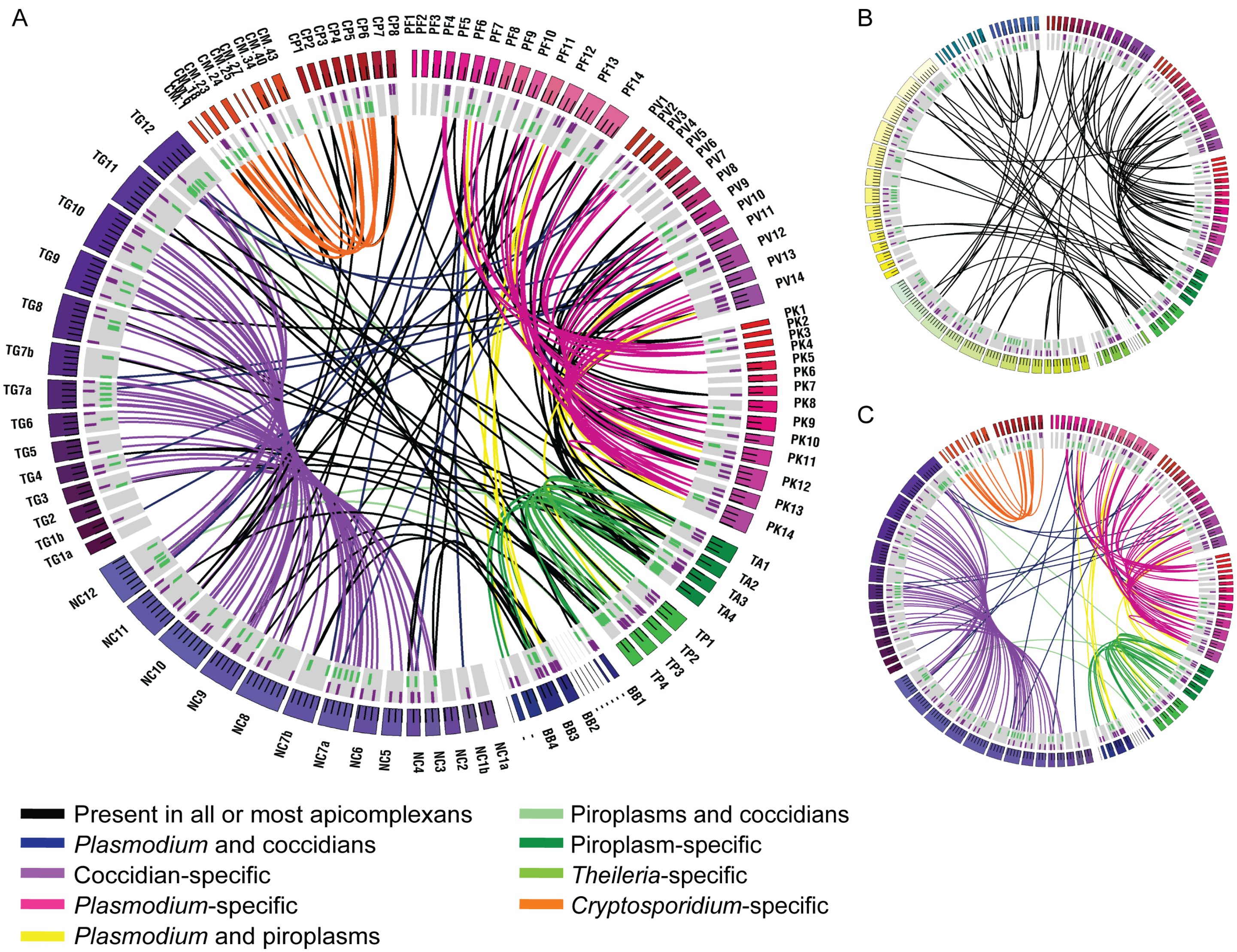


Figure 4.3. Circos diagram of ApiAP2 domain homolog groups across the Apicomplexa.

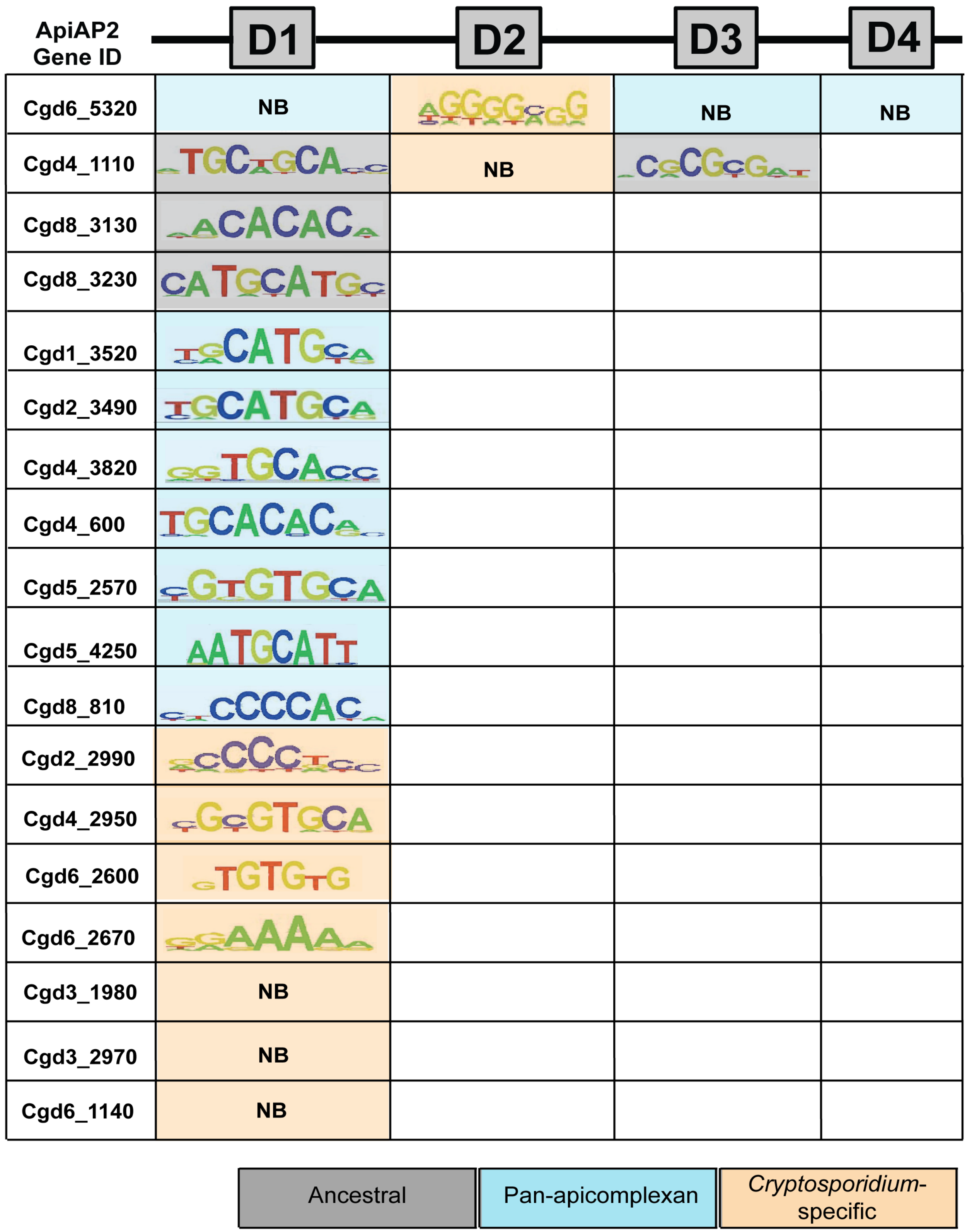


Figure 4.4. ***C. parvum*** **ApiAP2 domain binding sites as determined by protein-binding microarray.**

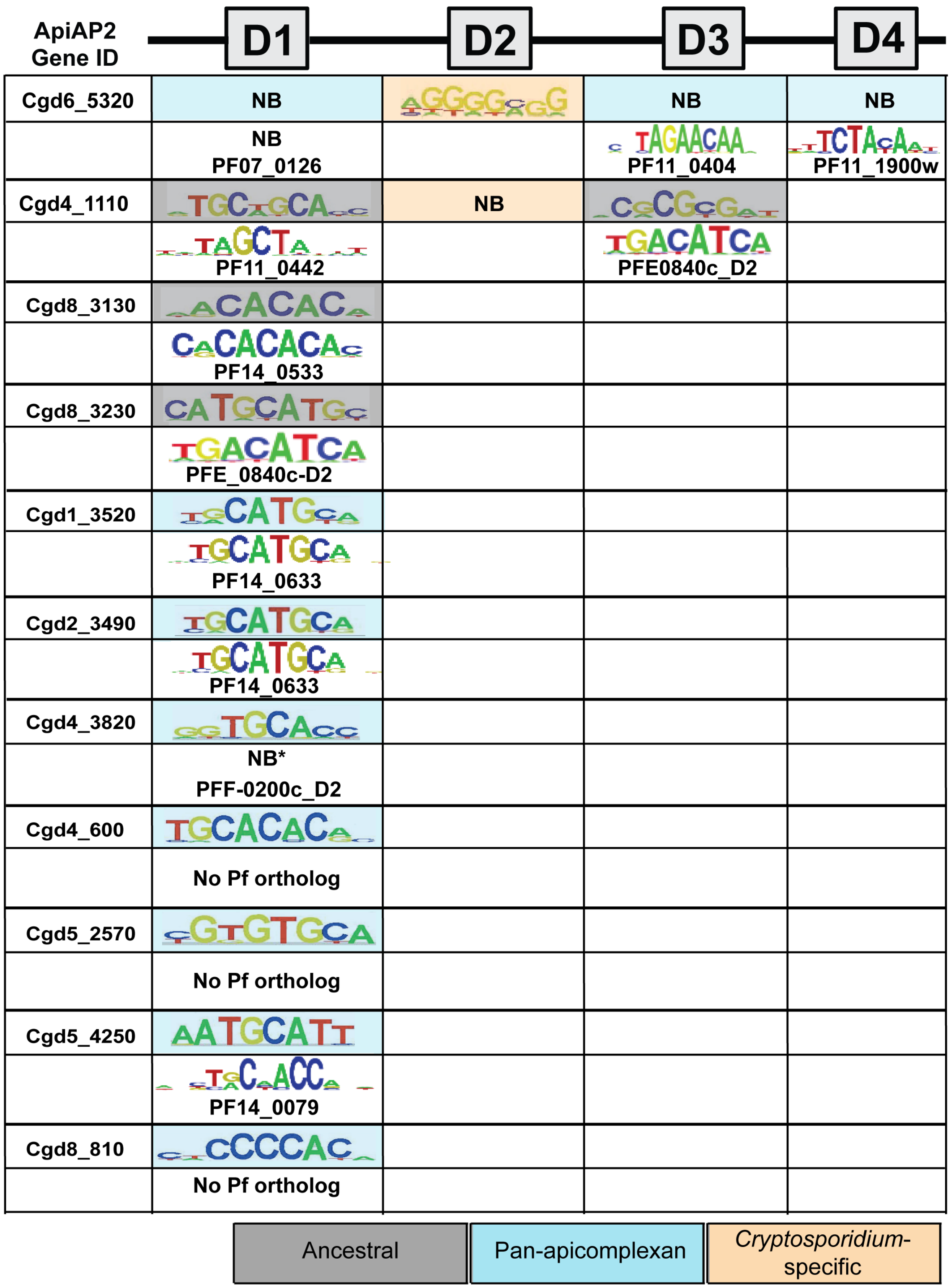


Figure 4.5. ***C. parvum*** **ApiAP2 binding sites compared to** ***P. falciparum*** **ortholog binding sites.**

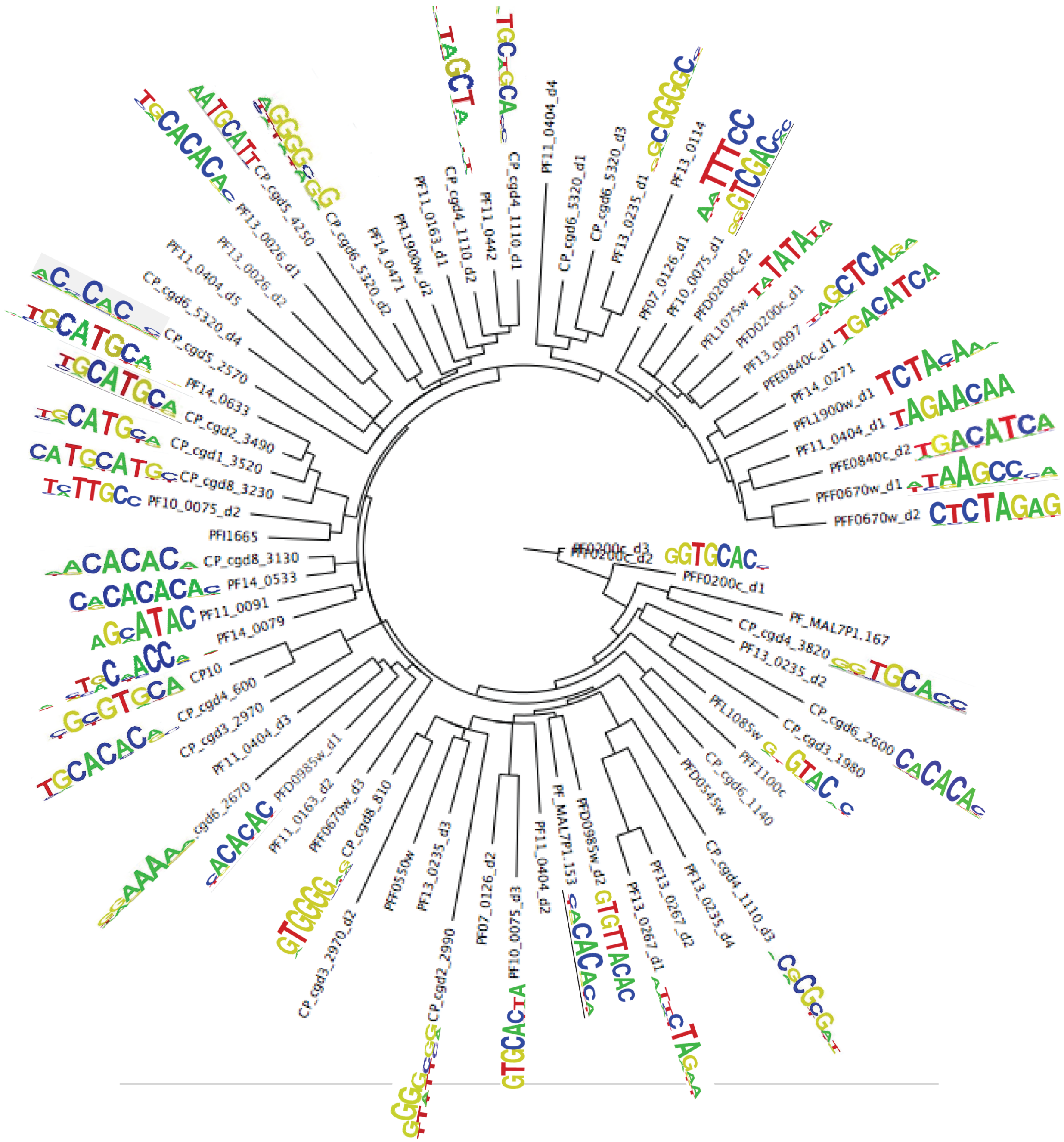


Figure 4.6. Maximum likelihood tree of *P. falciparum* and *C. parvum* ApiAP2 domains and their corresponding DNA binding sites.

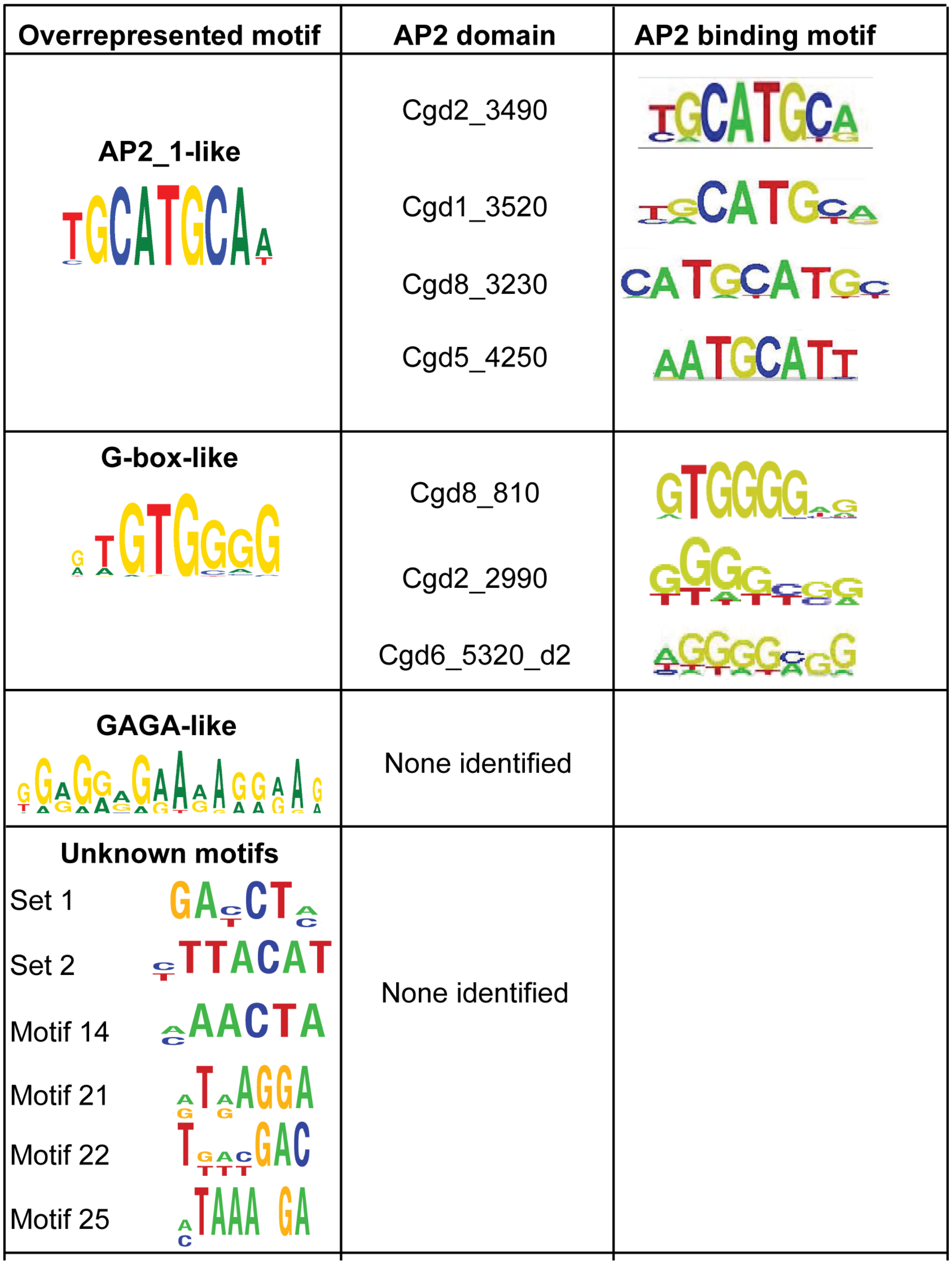


Figure 4.7. Overrepresented *C. parvum* motifs bound by ApiAP2 domains.

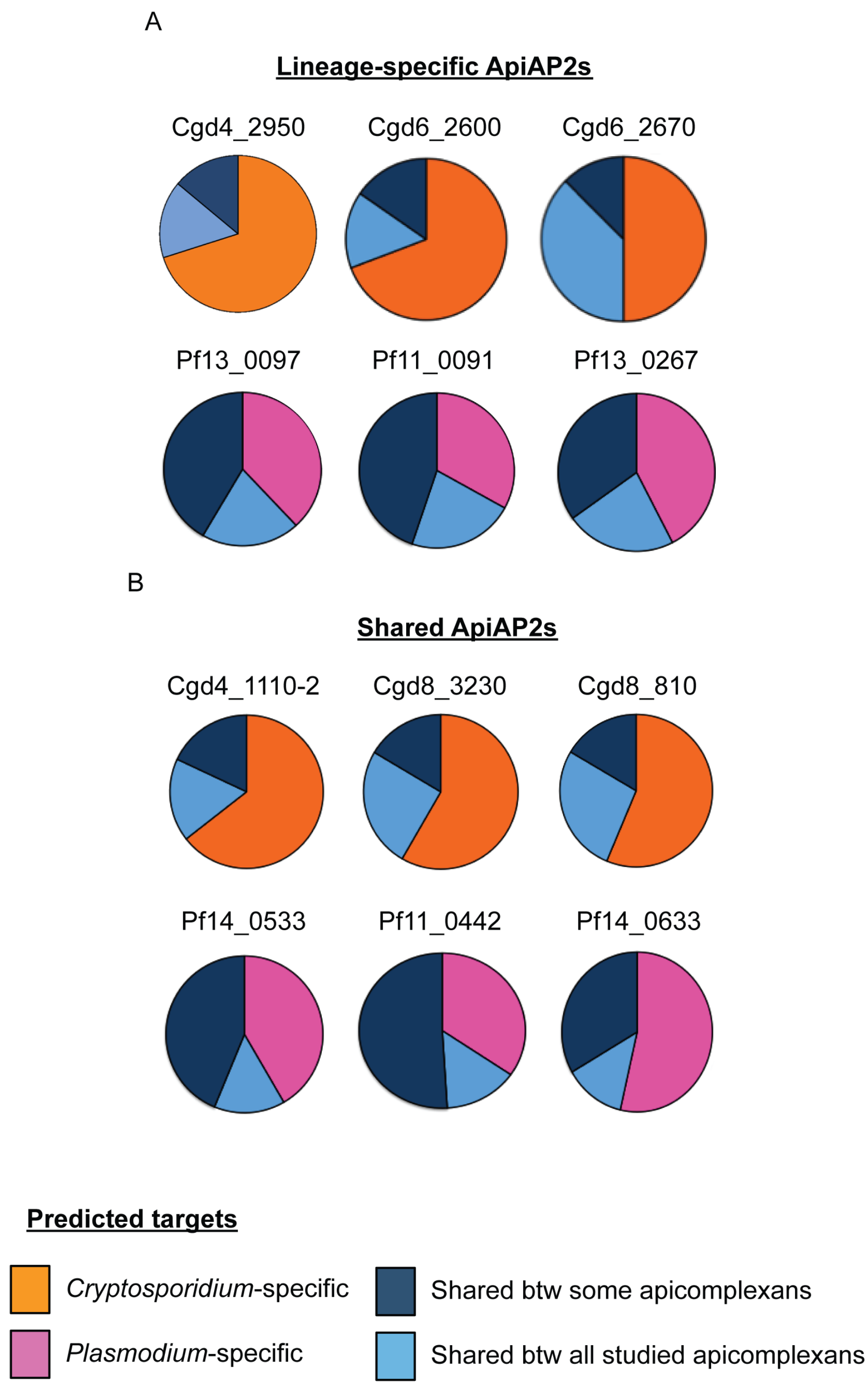
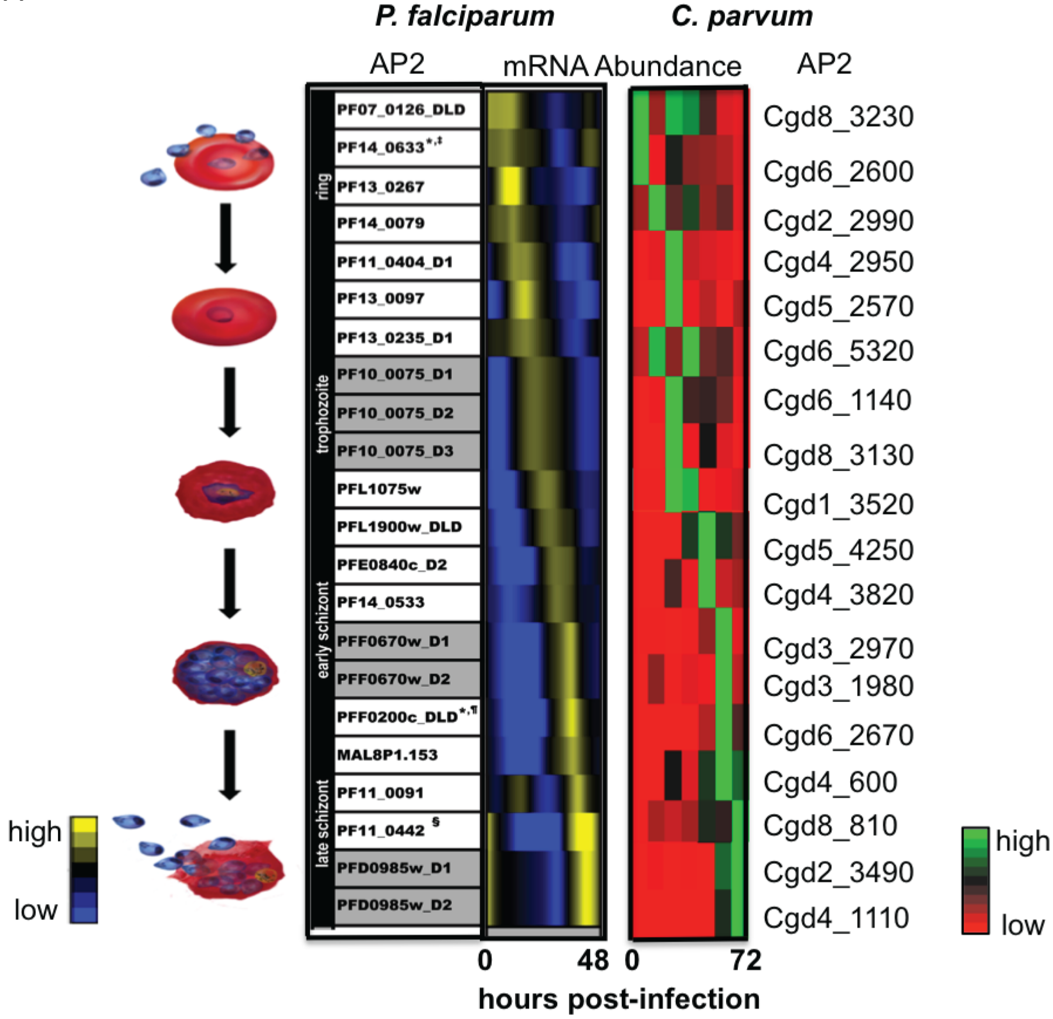


Figure 4.8. Evolutionary classification of predicted target genes for select lineage-specific and shared ApiAP2s.

A



B

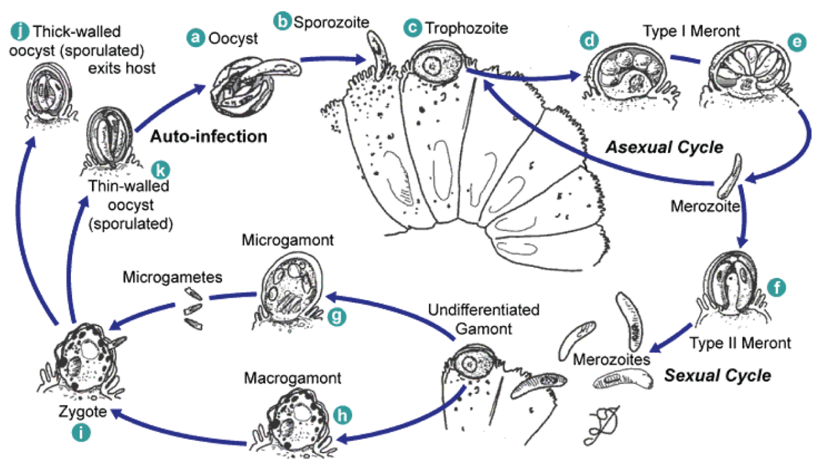


Figure 4.9. Cascade of ApiAP2 protein expression across the *P. falciparum* and *C. parvum* life cycles.

Table 4.1. Distribution and quantification of AP2 proteins and domains across chromalveolates and algal endosymbionts.

Eukaryota			Species	Total AP2 proteins	Total AP2 domains
Chromalveolates	Apicomplexans		*P. falciparum*	30	51
			P. knowlesi...	30	50
			P. vivax………	33	52
			P. yoelli………	26	43
			T. parva……..	22	27
			T. annulata…	24	31
			B. bovis………	22	29
			N. caninum…	61	89
			T. gondii…….	64	93
			C. muris……..	17	24
			C. parvum…..	18	23
			P. marinus…..	2-13*	3-14*
	Dinoflagellates		*K. brevis…….*	6**	11**
			A. tamarense	4**	5**
	Ciliates		*T. thermophila*	6	6
			P. tetraurelia	0	0
	Stramenopiles		*T. pseudonana*	15	19
			P. tricornutum	2	2
Rhodophytes			*C. merolae…..*	0	0
			P. purpurea…	0	0
			P. yezoensis..	0	0
Viridiplantae	Chlorophytes		*C. reinhardtii..*	13	17
			Micromonas sp. RCC299	17	27
			Total	412-423	604-615

Table 4.2. Domain counts by evolutionary group.

Domain classification	**Number of domains in group**
Present in *P. marinus* and all or most apicomplexans	1 to 5
Present in all or most apicomplexans	11 to 13
Plasmodium and piroplasms	2 to 5
Plasmodium and coccidians	7 to 11
Coccidians and piroplasms	1
Plasmodium-specific	18 to 29
Piroplasm-specific	8 to 9
Theileria-specific	2 to 6
Cryptosporidium-specific	7 to 15
Coccidian-specific	44 to 69

Table 4.3. ***C. parvum*** **ApiAP2 secondary motifs.**

ApiAP2		Secondary motif	Relationship to Primary Motif	Enrichment Score
Cgd1_3520			Core change	0.4586
Cgd4_1110	D1		Similar	0.4949
	D3		Similar	0.4730
			Similar	0.4800
Cgd4_2950			Rev com	0.4915
Cgd4_600			Rev com	0.4974
Cgd5_2570			Rev com	0.4940
Cgd6_2600			Rev com	0.4874
Cgd6_2670			Rev com	0.4822
Cgd6_5320 (D2)			Rev com	0.4618
Cgd8_3130			Rev com	0.4719
Cgd8_3230			Truncation (?)	0.4960
Cgd8_810			Rev com	0.4931

CHAPTER 5

Discussion and future directions

Malaria diagnostics

We have shown that malaria diagnostic assays developed to conserved, repetitive sequences in several human-infecting malaria parasite genomes are species-specific and more sensitive than existing molecular diagnostics (Chapter 2). We have developed multiplexable assays to detect both *P. falciparum* and *P. vivax* in a single step. Though we haven't tried a three-species multiplex with *P. knowlesi*, we designed the *P. knowlesi* assay with the intent that it could be multiplexed with the existing *P. falciparum* and *P. vivax* assays as well (with consideration to size of the target, primer melting temperature, and possible cross-reactivity between primer sets). Using our data-mining methodology, I have developed assays for the detection of *P. ovale* as well that I haven't discussed in this dissertation; these assays appear very species-specific, but we do not have enough clinical isolates to confirm sensitivity and specificity. These assays have been designed with multiplex goals in mind as well.

The eventual goal of our work is to have a single multiplex assay that can detect all five human-infecting malaria parasites in the same reaction—a savings of ~83% over the existing molecular diagnostic gold standard, which requires reagents and expendables for six tubes and multiple rounds of PCR. Though we have put much consideration into developing assays with the potential for multiplex, conditions can be much more difficult to optimize than those for standard PCR, and the reaction can be much more sensitive to fluctuations in the concentrations of various reagents, as well as varying concentrations of templates [1]. Problems with spurious amplification products, uneven or no amplification of some target sequences, and reproducibility are common. Considerable parameter optimization and possibly new target design will likely be needed to develop a truly viable multiplex that will detect multiple species of malaria parasite.

Apicomplexan transcriptional regulation

We have presented the first comprehensive study of a major transcription factor family in *Cryptosporidium parvum*, the ApiAP2s, and present evidence that *C. parvum* may not be as reliant on ApiAP2 regulation as previous research has indicated for other apicomplexans. Many *C. parvum* ApiAP2s bind redundant motifs, and the majority of *C. parvum* ApiAP2 domains bind only one motif (Chapter 4). Thus *C. parvum* ApiAP2 regulation does not appear to be as multi-faceted as proposed in *P. falciparum* [2]. The presence of more non-ApiAP2 transcription factors in the *C. parvum* genome may explain the decreased diversity of ApiAP2 binding.

The E2F motif is the most abundant motif in the upstream regions of the *C. parvum* genome, being found upstream of 161 of 200 predicted co-regulated gene clusters (Chapter 3). E2Fs are notably absent in *Plasmodium* and other apicomplexans [3]. It is possible that the two predicted E2F transcription factors are responsible for a disproportionate amount of transcriptional regulation, such that *C. parvum* is less reliant on ApiAP2s. The apparent redundancy in *C. parvum* ApiAP2 binding motifs may also be important to stage-specific transcriptional regulation, as ApiAP2s binding the same or similar motifs are expressed at various points across the lifecycle.

A next logical step to expand on the work presented in this dissertation would be to experimentally examine the reliance *C. parvum* has on E2F transcription factors. Do E2Fs actually bind the overrepresented E2F-like motifs in the *C. parvum* genome? What about the ApiAP2s, what proportion of putative target genes do they bind? Though we have no genetic tools in *C. parvum*, ChIP-seq with an antibody to E2F would elucidate the proportion of motifs that are bound. I have already created several GST-tagged ApiAP2 constructs that could potentially be used for ChIP-seq analyses.

If E2Fs are the primary regulators in the *C. parvum* genome, how did this come to be? No other sequenced apicomplexans have E2F transcription factors [4], nor is the E2F motif overrepresented in the upstream regions of any other sequenced apicomplexan genome [5]. E2F evolutionary history has been more extensively studied than ApiAP2 evolutionary history. E2Fs are present in almost all studied eukaryotes to date, including animals, amoebazoans, plants, and basal eukaryotes such as *Trichomonas* and *Giardia*, while being absent in nearly all fungal lineages [4]. This broad phyletic distribution suggests that E2Fs are ancient transcription factors, and they are part of a group of at least seven DNA-binding domains that can be traced to the last eukaryotic common ancestor [4]. Thus the presence of the E2F domain in *C. parvum* most likely represents the ancestral state, and there were subsequent losses in all the other apicomplexan lineages. How an acquired factor like the ApiAP2 domain came to be so integral to apicomplexan transcriptional regulation is still mysterious, though it can be guessed at. Acquisition of genes via horizontal transfer, recruitment of DNA-binding domains from the transferred genes, lineage-specific expansion of the acquired gene and subsequent repurposing of the acquired DNA-binding domains as transcription factors has been a relatively common pathway taken for the establishment of a number of eukaryotic lineage-specific transcription factor families, including the AP2 and WRKY families in plants, the WRKY-related Rcsp1 and Af2p families in yeast, among several others [4,6].

However the ApiAP2 domain rose to prominence in the Apicomplexa, its role in *Cryptosporidium* transcriptional regulation is less certain. The conservation of the ApiAP2 expression cascade between *Cryptosporidium* and *Plasmodium* (Chapter 4) and the binding of a number of these factors to the second-most overrepresented motif in the genome suggests they are very important regulators. Here we have provided a global overview of *C. parvum* transcriptional regulation. This work can be used to further dissect *C. parvum* transcriptional regulation, with focus on individual factors and their putative targets. Further comparisons can be made between these data and those being gathered across the Apicomplexa to further elucidate the evolutionary history of transcriptional regulation in these parasites.

REFERENCES

1. Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: A practical approach. Journal of Clinical Laboratory Analysis 16: 47-51.

2. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. Plos Pathogens 6.

3. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. (2004) Comparative analysis of Apicomplexa and genomic diversity in eukaryotes. Genome Research 14: 1686-1695.

4. Iyer LM, Anantharaman V, Wolf MY, Aravind L (2008) Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. International Journal for Parasitology 38: 1-31.

5. Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M, et al. (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. Genome Res 13: 1787-1799.

6. Babu MM, Iyer LM, Balaji S, Aravind L (2006) The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Research 34: 6505-6520.

CHAPTER 6

Appendices

6.1 A new single-step PCR assay for the detection of the zoonotic malaria parasite *Plasmodium knowlesi*

Naomi W. Lucchi, Mitra Poorak, Jenna Oberstaller, Jeremy DeBarry, Ganesh Srinivasamoorthy, Ira Goldman, Maniphet Xayavong, Alexandre J. da Silva, David. S. Peterson, John W. Barnwell, Jessica C. Kissinger, Venkatachalam Udhayakumar. 2012. *PLoS One*. 7(2): e31848. doi:10.1371/journal.pone.0031848. Reprinted here with permission of publisher.

ABSTRACT

Background

Recent studies in Southeast Asia have demonstrated substantial zoonotic transmission of *Plasmodium knowlesi* to humans. Microscopically, *P. knowlesi* exhibits several stage-dependent morphological similarities to *P. malariae* and *P. falciparum*. These similarities often lead to misdiagnosis of *P. knowlesi* as either *P. malariae* or *P. falciparum* and PCR-based molecular diagnostic tests are required to accurately detect *P. knowlesi* in humans. The most commonly used PCR test has been found to give false positive results, especially with a proportion of *P. vivax* isolates. To address the need for more sensitive and specific diagnostic tests for the accurate diagnosis of *P. knowlesi*, we report development of a new single-step PCR assay that uses novel genomic targets to accurately detect this infection.

Methodology and Significant Findings

We have developed a bioinformatics approach to search the available malaria parasite genomes for the identification of suitable DNA sequences relevant for molecular diagnostic tests. Using this approach, we have identified multi-copy DNA sequences distributed in the *P. knowlesi* genome. We designed and tested several novel primers specific to new target sequences in a single-tube, non-nested PCR assay and identified one set of primers that accurately detects *P. knowlesi*. We show that this primer set has 100% specificity for the detection of *P. knowlesi* using three different strains (Nuri, H, and Hackeri), and one human case of malaria caused by *P. knowlesi*. This test did not show cross reactivity with any of the four human malaria parasite species including 11 different strains of *P. vivax* as well as 5 additional species of simian malaria parasites.

Conclusions

The new PCR assay based on novel *P. knowlesi* genomic sequence targets was able to accurately detect *P. knowlesi*. Additional laboratory and field-based testing of this assay will be necessary to further validate its utility for clinical diagnosis of *P. knowlesi*.

INTRODUCTION

Until recently, only four *Plasmodium* species, *P. falciparum, P. vivax, P. malariae* and *P. ovale*, were thought to contribute to human malaria infections. However, recent studies in Southeast Asia have shown zoonotic transmission of *P. knowlesi* to humans [1-15]. *P. knowlesi* is a parasite species that readily infects Old World monkeys, reviewed in [16]. The natural hosts of this simian malaria parasite are the long-tailed (*Macaca facsicularis*) and pig-tailed (*M. nemestrina*) macaque monkeys and langurs (*Presbytis* sp.) [17-19] that are distributed throughout much of Southeast Asia. The transmission of *P. knowlesi* is closely related to its vector species in the *Anopheles leucophyrus* group, which are forest-dwelling mosquitoes found in forest canopies or on forest fringes [8-10]. Indeed, many of the reported human *P. knowlesi* cases were found either near forests or as imported cases from individuals known to have visited the forests [20-22]. To date, no human-to-human transmission has been documented and chloroquine is effective in treating these infections [3]. *P. knowlesi* has a 24-hour asexual life cycle [23], the shortest observed, thus far, for human-infecting parasites. This short cycle can lead to rapid increases in parasitemia and can lead to severe disease including fatalities as reported in recent studies [1,2]. Given these observations, human infections with *P. knowlesi* require immediate and appropriate treatment, which in turn depends upon a prompt and accurate diagnosis.

Microscopically, *P. knowlesi* exhibits stage-dependent morphological similarities to *P. malariae* and *P. falciparum* [3,24]. These similarities have contributed to misdiagnosis of *P. knowlesi* as *P. malariae* [1,3] or *P. falciparum*. For example, a study in the Kapit Division of Malaysian Borneo, found that 58% of previously diagnosed *P. malariae* cases were actually *P. knowlesi* infections [3]. In this study by Singh et al., [3] a nested PCR-based diagnostic test for the detection of *P. knowlesi* 18S ribosomal RNA genes was developed and has been used in numerous subsequent studies [1,7-9,25-27]. However, this test was recently noted to cross-react with *P. vivax* leading to potential false positive results for a small proportion of human clinical *P. vivax* samples [26]. This observation was confirmed by results from our laboratory, in which cross reactivity with *P. vivax* and other simian *Plasmodium* species (*P. cynomolgi*, *P. inui, P. coatneyi,* and *P. hylobati*) was observed (Figure 1). These findings have raised some concern about the actual extent of the reported *P. knowlesi* cases [28,29], although *P. knowlesi* DNA from some of the diagnosed cases was sequenced in order to confirm the presence of this parasite [7-9]. Therefore, development of an improved molecular diagnostic test is critical not only for the proper diagnosis of human infections, but also for estimating the true burden of *P. knowlesi* infection in human populations.

Imwong *et al*. recently reported a nested PCR assay with 100% specificity for detecting *P. knowlesi* [26]. In addition, a loop mediated isothermal amplification (LAMP) method designed to detect the beta tubulin gene of *P. knowlesi* [30] and two real-time PCR assays [31,32] have been reported to be highly specific for the detection of *P. knowlesi.* We recently reported on the use of a bioinformatics approach to mine available genome data and identify suitable DNA sequences that are highly specific to a given species of malaria parasite [33]. Using this approach, we have identified highly-specific, multi-copy sequences from the *P. knowlesi* genome and designed novel primers that can be used in a single-tube, non-nested PCR diagnostic test. We have identified one set of primers that has high specificity (100%) for the detection of *P. knowlesi* at a low level of parasitemia (1 parasite per uL).

METHODS

Plasmodium parasites and clinical samples

Different *Plasmodium* species available in our laboratories were utilized to test the specificity of the novel *P. knowlesi* primers: *P. falciparum* (3D7 clone), *P. vivax* (South Vietnam IV), *P. malariae* (Uganda I), *P. ovale* (Nigeria I), and 11 other *P. vivax* strains (Ong, Thai III, India VII, Honduras I, Salvador II, Panama I, Chesson, Vietnam IV, Pakchong, Mauritania I and Indonesia XIX). Three *P. knowlesi* isolates (Nuri, H, and Hackeri) and 5 simian malaria parasites (*P. simiovale, P. inui, P. cynomolgi, P. hylobati* and *P. coatneyi*) available in the CDC laboratory collection were included. In addition, DNA from 52 clinical samples, previously diagnosed using a nested PCR method [34] (14 *P. falciparum*, 9 *P. vivax*, 1 *P. malariae*, 12 *P. ovale, 2 P. falciparum/P. malariae, 1 P. vivax/P. ovale, 2 P. falciparum/P. ovale* mixed infections, 1 *P. knowlesi* and 10 malaria negative samples), were tested in a blinded manner. The *P. knowlesi* sample was acquired from a traveler who returned infected after a trip to the Philippines in 2008, representing the first recognized case of imported simian malaria in several decades in the United States [35]. These clinical samples were obtained from the CDC molecular diagnostic parasitology reference laboratory (Dr. A. da Silva).

DNA extraction

The QIAamp DNA Mini Kit (QIAGEN, Valencia, CA-(Qiagen method) was used to isolate DNA from the different *Plasmodium* parasites following the manufacturer's protocol. The DNA was aliquoted and stored at −20°C until used in the experiments.

Novel *P. knowlesi* target validation

Assembled genome sequence data for *P. knowlesi* was obtained from PlasmoDB (http://plasmodb.org/plasmo/; release 5.5). The sequence candidates were selected as previously described [33]. Briefly, genome sequence data were mined for repetitive content. The identified repetitive sequences were screened for a number of properties that would negate their utility as PCR targets, such as tandem repeats and human or artificial (vector) sequence similarity. Repeats passing these screens were evaluated for species-specificity. The copy number of candidate targets satisfying a length requirement of 300 bp was determined, and targets with greater than 5 copies/genome were further considered as potential diagnostic targets. Primers were designed manually to the candidate targets and screened for GC-content, melting temperature, secondary structure, and primer-dimer forming potential. Primer pairs were optimized by means of gradient PCR using *P. knowlesi* DNA (strain H) to determine the optimum annealing temperature, primer concentration (concentrations from 0.25µM to 1.0µM were tested) and $MgCl_2$ concentrations (2.0mM - 4.0mM were tested). Primers were then tested for *P. knowlesi* specificity and sensitivity.

Specificity assay

Primers that passed the initial validation tests were further tested for specificity using 11 *P. vivax* strains and different simian *Plasmodium* species and strains. DNAs from 52 clinical samples were tested blindly.

PCR assays

All PCR tests were completed on a BioRad iCycler (BioRad, Hercules, CA). Nested PCR for *P. knowlesi* was performed with primers and cycling conditions as described before [3]. The confirmatory nested PCR used to test the 52 clinical samples was as previously described [34]. The PCR amplified material was analyzed using gel electrophoresis (2% agarose gel) to visualize the bands of appropriate size. Amplification of *P. knowlesi* using the novel primers was performed in a 25µl reaction containing 1 X Taq Buffer (containing 10mM Tris-HCl, 50mM KCl, 1.5mM $MgCl_2$) , 200µM each dNTP, 1.25 units of Taq DNA Polymerase (all from New England Biolabs, Ipswich MA, USA), 250nM each oligonucleotide primer, and 1µl of DNA template. The sequences of the final oligonucleotide primer set (Pkr140-5) selected for *P. knowlesi* detection are shown in Table 1. Reactions were performed under the following cycling parameters: initial denaturation at 95°C for 2 minutes, and then 35 cycles of 95°C for 30 seconds, 57°C for 30 seconds, and 72°C for 45 seconds, followed by final extension at 72°C for 5 minutes. Ten mL of PCR products were visualized by gel electrophoresis on a 2% agarose gel.

Limits of detection of the PCR amplification using the new primers

The analytical sensitivity of the assay was determined using a well-quantified *P. knowlesi* H strain sample obtained from an infected monkey. The WHO recommended protocol for the preparation of standards for use in the quality control of rapid diagnostic tests (http://www.wpro.who.int/sites/rdt/using_rdts/qa/lot_testing.htm) was used to prepare the parasite standard for this study. The *P. knowlesi* parasites were at either the ring or early trophozoite stages of development when the sample was utilized. The percent parasitemia of the infected monkey was determined by three expert microscopists by counting the number of infected erythrocytes in 10,000 erythrocytes. The total number of erythrocytes per microliter was determined through use of a coulter counter and the number of parasites/mL was then determined from the total number of RBCs/mL. The resulting parasitemia was determined to be 225,000 parasites/mL. This standard sample was then diluted from the initial parasitemia to 100,000 parasites/mL using uninfected blood and then serial diluted ten-fold to 1p/µL using a 250mL volume. DNA was extracted from each dilution point using 200mL of sample. These diluted samples were used to test the limits of detection of the previously described primers [3] and the novel Pkr140-5 primer set described here.

RESULTS

Primer Design

Four genomic sequence targets passed the *in silico* tests and were selected for validation. A total of 14 primers were designed to these targets and empirically tested in conventional PCR amplification assays using *P. knowlesi*-H DNA sample. Of the 14 primers designed, three sets (Pkr140-3, Pkr140-4 and Pkr140-5), all of which recognize the Pkr140 repeat sequence, were selected for further evaluation as they correctly amplified *P. knowlesi* as evidenced by clean, intense, single bands of the expected size. The Pkr140 sequence exists in 7 copies in the available *P. knowlesi* genome sequence.

Tests for assay specificity

The three Pkr140 primer sets were tested for species-specificity initially using DNA from *P. falciparum, P. vivax, P. ovale* and *P. malariae.* No cross-reactivity was observed with these species (Figure 2 and data not shown) as there was no amplification of these DNA. Second, the primers were tested for specificity against 5 different simian malaria parasites (*P. simiovale, P. inui, P. cynomolgi, P. hylobati* and *P. coatneyi*) and strains thereof. Primer set Pkr140-3 produced non-specific bands with *P. inui, P. cynomolgi,* and *P. hylobati* (Figure 3A) and primer set Pkr140-4 with *P. cynomolgi* (Figure 3B). These two primer sets were not evaluated further. Primer set Pkr140-5 (Table 1) detected only the three *P. knowlesi* isolates (H, Nuri, and Hackeri) used in this study (Figure 3C) and did not amplify DNA from any of the eleven *P. vivax* isolates tested (Figure 4).

Further test for specificity using clinical samples

Forty two DNA specimens extracted from clinical samples previously confirmed by PCR as positive for malaria (14 *P. falciparum*, 9 *P. vivax*, 1 *P. malariae*, 12 *P. ovale, 2 P.*

falciparum/P. malariae, 1 P. vivax/P. ovale, 2 P. falciparum/P. ovale mixed infections, 1 *P. knowlesi*) and 10 malaria negative clinical samples were further used to test the specificity of primer set Pkr140-5 in a blinded manner. This primer set correctly identified the single sample with known *P. knowlesi* infection [35] and did not show any cross-reactivity with any of the other samples.

Limits of detection of primer set Pkr140-5

Using known quantities of *P. knowlesi*-H DNA, both the previously published *P. knowlesi* primers and the novel Pkr140-5 primer set were able to detect up to 1 parasite of *P. knowlesi* /mL of blood with the novel primer set showing better resolution than the previously published primer set (Figure 5).

Characteristics of Pkr140

The Pkr140 sequence repeats are present in 7 copies distributed across 6 chromosomes (Figure 6). Six of the copies have an average size of 424 bps. The seventh copy (closest to the end of chromosome 5) is truncated (only 42 bps) and is not amplified by primer set Pkr140-5. We previously identified repetitive sequence targets in *P. falciparum and P. vivax* that were distributed to subtelomeric regions or to contigs thought to belong to subtelomeric regions [33]. In contrast, the Pkr140 sequences are found both near chromosome ends and interior regions. The Pkr140 sequences do not appear to be protein-encoding and they have not been annotated as serving any particular function. Moreover, searches of PlasmoDB (http://plasmodb.org) did not reveal any possible function for these sequences. Interestingly, 5 of the 7 Pkr140 repeat sequences (including the truncated copy) are located near genes that encode the *SICAvar* antigen, a member of one of the main variant gene families in *P. knowlesi* [36,37]

DISCUSSION

In this study, we report a new PCR assay based on novel genomic target sequences for *P. knowlesi* detection. We have previously reported on the use of a bioinformatics method to mine parasite genome sequences in search of species-specific and multi-copy sequences that can be used to design diagnostic PCR primers for malaria detection [33]. Using this genome-mining approach, 14 primer sets were designed and tested for their utility for *P. knowlesi* detection in a non-nested PCR assay. Three sets of primers were found to amplify *P. knowlesi* consistently. However, two of these sets produced non-specific bands with some simian malaria parasites and were not tested further as our goal was to identify primers that specifically amplify *P. knowlesi*. We identified primer set Pkr140-5 as specific for the detection of *P. knowlesi* as it did not detect any other human malaria parasites nor any of the five simian malaria species tested, including the closely related species *P. inui and P. cynomolgi.*

Previously identified diagnostic targets in *P. falciparum* and *P. vivax* [33] were distributed at chromosome ends or unassembled contigs belonging to chromosome ends. Subtelomeric regions in these species have been shown (to varying degrees) to be enriched for species-specific and multi-copy genes [38] and genes involved in antigen variation [39,40]. The *P. knowlesi* genome organization differs from *P. falciparum* and *P. vivax* with genes involved in antigenic variation distributed across chromosomes and not concentrated at their ends. Given this difference in genome organization, and the proximity of the identified Pkr140 targets to *SICAvar* genes, it is perhaps not surprising that the targets are also distributed across both chromosome ends and interiors. Based on the results from three *Plasmodium* species, regions near multi-gene families are potentially rich areas for the mining of diagnostic targets.

Our data, reported here, further confirms a previous report of cross reactivity between 18S ribosomal RNA gene primers [3] and *P. vivax* parasites. In addition, our results demonstrate that the 18S ribosomal RNA gene primers also cross-react with at least four simian malaria parasites (*P. inui, P. hylobati, P. cynomolgi,* and *P. coatneyi*). The difficulty of *P. knowlesi* diagnosis with the 18S ribosomal RNA gene-based PCR assay was also recently highlighted in a study in which 2 samples determined to be positive for *P. knowlesi* could not be confirmed by DNA sequencing analysis [41]. The primer set described here showed 100% specificity and no cross reactivity observed with any of the non- *P. knowlesi* samples tested. In addition, this primer set showed a limit of detection of 1 parasite/mL which was shown to be comparable to the limits of detection of the previously described nested PCR test [3]. This is promising as the primer set can be used for the detection of low parasite levels without the need to perform a nested PCR. A limitation of the current study is the fact that only one clinical *P. knowlesi* sample was available for use to test the novel primer sets; however, three *P. knowlesi* strains obtained from monkeys were included to validate the specificity. Given the fact that the occurrence of human *P. knowlesi* is a pretty novel phenomenon that is rather confined mainly in Southeast Asia, it was not immediately possible to evaluate a large number of *P. knowlesi* samples. Therefore, further validation of these primers in regions known to have *P. knowlesi* transmission will be required to test their utility for *P. knowlesi* diagnosis. However, the lack of a large sample size does not negate the fact that these primers are indeed specific and sensitive to detect *P. knowlesi*.

Molecular tools for *P. knowlesi* detection have been reported including nested PCR assays, two real-time PCR assays and a loop mediated isothermal amplification (LAMP) assay [30-32,42]. The PCR test described here does not require nested amplification, simplifying the performance of the reaction and saving on costs. The LAMP assay holds potential for use in regions with limited or fewer resources, as it does not necessitate the use of expensive thermal cyclers. The real-time PCR assays' use is limited to settings with real-time PCR capabilities such as reference laboratories. It remains to be determined if these different assays vary in their sensitivity and specificity to diagnose *P. knowlesi* infection in field/clinical settings.

Human *P. knowlesi* infections have been mostly reported in Southeast Asia [1-15]. Recently, several imported cases in other parts of the world have also been reported [20-22,43] including the United States[35]. The novel non-nested PCR assay described in this study is a suitable alternative for the accurate diagnosis of *P. knowlesi* by PCR in most laboratories. However, additional laboratory and field-based testing of this assay will be necessary to validate its utility for clinical diagnosis of *P. knowlesi.*

ACKNOWLEDGEMENTS

We thank Allyson Byrd for her help in generating the Circos map of *P. knowlesi* targets.

REFERENCES

[1] Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, et al. (2008) *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. Clin Infect Dis 46: 165-171.

[2] Cox-Singh J, Hiu J, Lucas SB, Divis PC, Zulkarnaen M, et al. (2010) Severe malaria - a case of fatal *Plasmodium knowlesi* infection with post-mortem findings: a case report. Malar J 9: 10.

[3] Singh B, Kim Sung L, Matusop A, Radhakrishnan A, Shamsul SS, et al. (2004) A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363: 1017-1024.

[4] Jongwutiwes S, Buppan P, Kosuvin R, Seethamchai S, Pattanawong U, et al. (2011) *Plasmodium knowlesi* Malaria in humans and macaques, Thailand. Emerg Infect Dis 17: 1799-1806.

[5] Jongwutiwes S, Putaporntip C, Iwasaki T, Sata T, Kanbara H (2004) Naturally acquired *Plasmodium knowlesi* malaria in human, Thailand. Emerg Infect Dis 10: 2211-2213.

[6] Putaporntip C, Hongsrimuang T, Seethamchai S, Kobasa T, Limkittikul K, et al. (2009) Differential prevalence of *Plasmodium* infections and cryptic *Plasmodium knowlesi* malaria in humans in Thailand. J Infect Dis 199: 1143-1150.

[7] Van den Eede P, Van HN, Van Overmeir C, Vythilingam I, Duc TN, et al. (2009) Human *Plasmodium knowlesi* infections in young children in central Vietnam. Malar J 8: 249.

[8] Lee KS, Divis PC, Zakaria SK, Matusop A, Julin RA, et al. (2011) *Plasmodium knowlesi*: Reservoir Hosts and Tracking the Emergence in Humans and Macaques. PLoS Pathog 7: e1002015.

[9] Tan CH, Vythilingam I, Matusop A, Chan ST, Singh B (2008) Bionomics of *Anopheles latens* in Kapit, Sarawak, Malaysian Borneo in relation to the transmission of zoonotic simian malaria parasite *Plasmodium knowlesi*. Malar J 7: 52.

[10] Vythilingam I, Tan CH, Asmad M, Chan ST, Lee KS, et al. (2006) Natural transmission of *Plasmodium knowlesi* to humans by Anopheles latens in Sarawak, Malaysia. Trans R Soc Trop Med Hyg 100: 1087-1088.

[11] Osman MM, Nour BY, Sedig MF, De Bes L, Babikir AM, et al. (2010) Informed decision-making before changing to RDT: a comparison of microscopy, rapid diagnostic test and molecular techniques for the diagnosis and identification of malaria parasites in Kassala, eastern Sudan. Trop Med Int Health 15: 1442-1448.

[12] Khim N, Siv S, Kim S, Mueller T, Fleischmann E, et al. (2011) *Plasmodium knowlesi* infection in humans, Cambodia, 2007-2010. Emerg Infect Dis 17: 1900-1902.

[13] Luchavez J, Espino F, Curameng P, Espina R, Bell D, et al. (2008) Human Infections with *Plasmodium knowlesi*, the Philippines. Emerg Infect Dis 14: 811-813.

[14] Vythilingam I, Noorazian YM, Huat TC, Jiram AI, Yusri YM, et al. (2008) *Plasmodium knowlesi* in humans, macaques and mosquitoes in peninsular Malaysia. Parasit Vectors 1: 26.

[15] Ng OT, Ooi EE, Lee CC, Lee PJ, Ng LC, et al. (2008) Naturally acquired human *Plasmodium knowlesi* infection, Singapore. Emerg Infect Dis 14: 814-816.

[16] Collins WE (2012) *Plasmodium knowlesi*: A Malaria Parasite of Monkeys and Humans. Annu Rev Entomol 57: 107-121.

[17] Eyles DE, Laing AB, Warren M, Sandosham AA (1962) Malaria parasites of Malayan leaf monkeys of the genus *Presbytis*. Med JMalaya 17: 85-86.

[18] Eyles DE, Laing AB, Dobrovolny CG (1962) The malaria parasites of the pig-tailed macaque, *Macaca nemestrina nemestrina* (Linnaeus), in Malaya. Ind J Malariol 16: 285-298.

[19] Coatneyi GR, Collins WE, Warren M, Contacos PG (1971) The Primate Malarias. Bethesda: U.S. National Institute of Allergy and Infectious Diseases. pp. 381.

[20] Berry A, Iriart X, Wilhelm N, Valentin A, Cassaing S, et al. (2011) Imported *Plasmodium knowlesi* Malaria in a French Tourist Returning from Thailand. Am J Trop Med Hyg 84: 535-538.

[21] Hoosen A, Shaw MT (2011) *Plasmodium knowlesi* in a traveller returning to New Zealand. Travel Med Infect Dis.

[22] Ta TT, Salas A, Ali-Tammam M, Martinez Mdel C, Lanza M, et al. (2010) First case of detection of *Plasmodium knowlesi* in Spain by Real Time PCR in a traveller from Southeast Asia. Malar J 9: 219.

[23] Chin W, Contacos PG, Collins WE, Jeter MH, Alpert E (1968) Experimental mosquito-transmission of *Plasmodium knowlesi* to man and monkey. Am J Trop Med Hyg 17: 355-358.

[24] Lee KS, Cox-Singh J, Singh B (2009) Morphological features and differential counts of *Plasmodium knowlesi* parasites in naturally acquired human infections. Malar J 8: 73.

[25] Lee KS, Cox-Singh J, Brooke G, Matusop A, Singh B (2009) *Plasmodium knowlesi* from archival blood films: further evidence that human infections are widely distributed and not newly emergent in Malaysian Borneo. Int J Parasitol 39: 1125-1128.

[26] Imwong M, Tanomsing N, Pukrittayakamee S, Day NP, White NJ, et al. (2009) Spurious amplification of a *Plasmodium vivax* small-subunit RNA gene by use of primers currently used to detect *P. knowlesi*. J Clin Microbiol 47: 4173-4175.

[27] Daneshvar C, Davis TM, Cox-Singh J, Rafa'ee MZ, Zakaria SK, et al. (2009) Clinical and laboratory features of human *Plasmodium knowlesi* infection. Clin Infect Dis 49: 852-860.

[28] Van den Eede P, Vythilingam I, Ngo DT, Nguyen VH, Le XH, et al. (2010) *Plasmodium knowlesi* malaria in Vietnam: some clarifications. Malar J 9: 20.

[29] Cox-Singh J (2009) Knowlesi malaria in Vietnam. Malar J 8: 269.

[30] Iseki H, Kawai S, Takahashi N, Hirai M, Tanabe K, et al. (2010) Evaluation of a loop-mediated isothermal amplification method as a tool for diagnosis of infection by the zoonotic simian malaria parasite *Plasmodium knowlesi*. J Clin Microbiol 48: 2509-2514.

[31] Divis PC, Shokoples SE, Singh B, Yanow SK (2010) A TaqMan real-time PCR assay for the detection and quantitation of *Plasmodium knowlesi*. Malar J 9: 344.

[32] Babady NE, Sloan LM, Rosenblatt JE, Pritt BS (2009) Detection of *Plasmodium knowlesi* by real-time polymerase chain reaction. Am J Trop Med Hyg 81: 516-518.

[33] Demas A, Oberstaller J, Debarry J, Lucchi NW, Srinivasamoorthy G, et al. (2011) Applied genomics: Data mining reveals species-specific malaria diagnostic targets more sensitive than 18S rRNA. J Clin Microbiol 49: 2411-2418.

[34] Johnston SP, Pieniazek NJ, Xayavong MV, Slemenda SB, Wilkins PP, et al. (2006) PCR as a confirmatory technique for laboratory diagnosis of malaria. J Clin Microbiol 44: 1087-1089.

[35] (2009) Simian malaria in a U.S. traveler--New York, 2008. MMWR Morb Mortal Wkly Rep 58: 229-232.

[36] Lapp SA, Korir CC, Galinski MR (2009) Redefining the expressed prototype SICAvar gene involved in *Plasmodium knowlesi* antigenic variation. Malar J 8: 181.

[37] al-Khedery B, Barnwell JW, Galinski MR (1999) Antigenic variation in malaria: a 3 genomic alteration associated with the expression of a *P. knowlesi* variant antigen. Mol Cell 3: 131-141.

[38] Debarry JD, Kissinger JC (2011) Jumbled Genomes: Missing Apicomplexan Synteny. Mol Biol Evol 28: 2855-2811.

[39] Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498-511.

[40] Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455: 757-763.

[41] Sulistyaningsih E, Fitri LE, Loscher T, Berens-Riha N (2010) Diagnostic difficulties with *Plasmodium knowlesi* infection in humans. Emerg Infect Dis 16: 1033-1034.

[42] Putaporntip C, Buppan P, Jongwutiwes S (2011) Improved performance with saliva and urine as alternative DNA sources for malaria diagnosis by mitochondrial DNA-based PCR assays. Clin Microbiol Infect 17: 1484-1491.

[43] Ong CW, Lee SY, Koh WH, Ooi EE, Tambyah PA (2009) Monkey malaria in humans: a diagnostic dilemma with conflicting laboratory data. Am J Trop Med Hyg 80: 927-928.

[44] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639-1645.

Figure and table legends

Figure 6.1. 18S ribosomal RNA gene based *P. knowlesi* primer cross-reacts with *P. vivax* and other simian-infecting malaria parasite species.

Published 18S ribosomal RNA gene based *P. knowlesi* primers [3] were used to test 5 different simian-infecting malaria parasite species (*P. simiovale, P. inui, P. cynomolgi, P. hylobati and P. coatneyi*) including 3 different *P. knowlesi* isolates (A) and 11 *P. vivax* strains (B). A no template control (NTC) was included. Cross reactivity was observed with some of the simian malaria parasites and some *P. vivax* strains.

Figure 6.2. Primer Pkr140-5 tested with the 4 human-infecting malaria parasite species.

To test the specificity of the novel primers, DNA from the four additional human-infecting *Plasmodium* parasites were tested. *P. knowlesi* (H strain) was used as a positive control (expected size = 200bp). A no template control (NTC) was also included.

Figure 6.3. Specificity of the *P. knowlesi* primers tested using simian-infecting malaria parasite species.

To test the specificity of the *P. knowlesi* primers, 5 different simian-infecting malaria parasite species (*P. simiovale, P. inui, P. cynomolgi, P. hylobati and P. coatneyi*) including 3 different *P. knowlesi* isolates were tested. The no template control (NTC) was included as a negative control. A; primer set Pkr140-3 (expected size = 230bp), B; primer set Pkr140-4 (expected size = 280bp) and C; primer set Pkr140-5 (expected size = 200bp). Circles indicate non-specific amplification.

Figure 6.4. Primer set Pkr140-5 does not cross react with *P. vivax*.

Multiple *P. vivax* strains were tested using primer set Pkr140-5 (expected size 200bp) in order to test the primers' specificity. NTC = No template control.

Figure 6.5. Limits of detection of primer set Pkr140-5

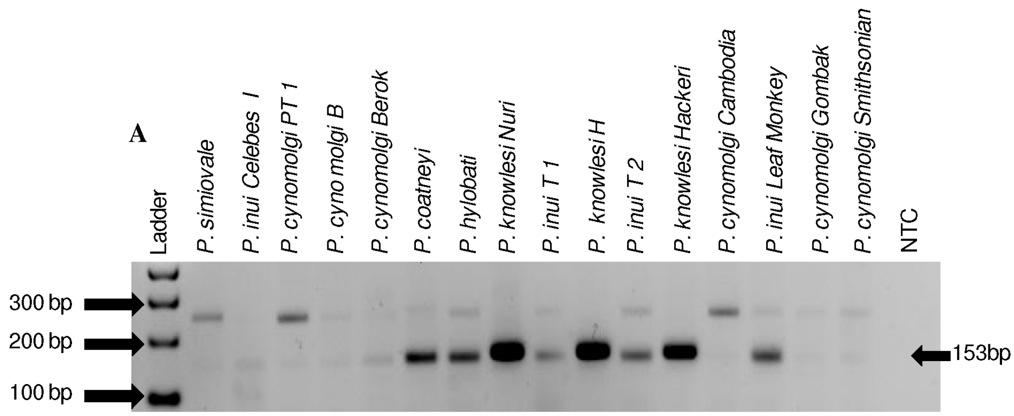
The analytical sensitivity of primer set Pk140-5 (A) and the primers from a published study [3] (B) were determined using a well-quantitated *P. knowlesi* DNA standard. The blood sample was serially diluted ten-fold with a starting parasitemia of 100,000p/µl to 1p/µl. The expected base pair sizes for the two primers are included. Three different experiments are shown.

Figure 6.6. Spatial distribution of Pkr140 sequence targets across the *P. knowlesi* genome.

The circle represents chromosomes. Each chromosome is labeled with the 2-letter genus and species abbreviation for *P. knowlesi* and the chromosome number. Tick marks indicate 1 mb of sequence. Lines inside the circle indicate the location of Pkr140 copies and are not to scale. Circos 0.51 (http://mkweb.bcgsc.ca/circos/) was used to generate this map [44].

Table 6.1. Sequence of the novel Pkr140-5 primer set. The primers were designed to target Pkr140, which is present in 7 copies distributed across 6 different chromosomes in the available genome sequence.

A



B

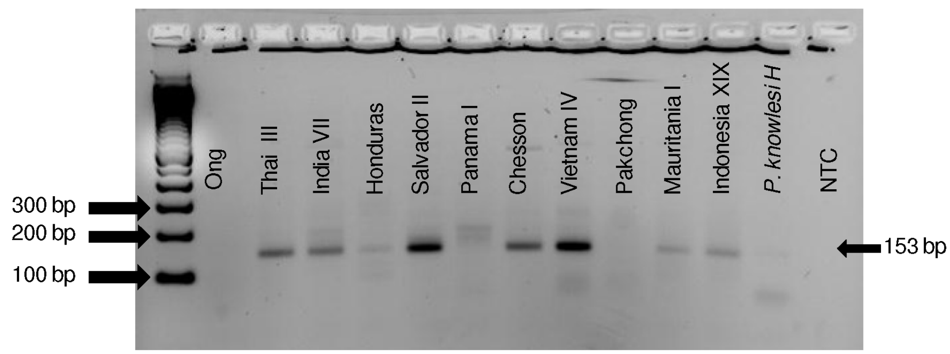


Figure 6.1. 18S ribosomal RNA gene based *P. knowlesi* primer cross-reacts with *P. vivax* and other simian-infecting malaria parasite species.

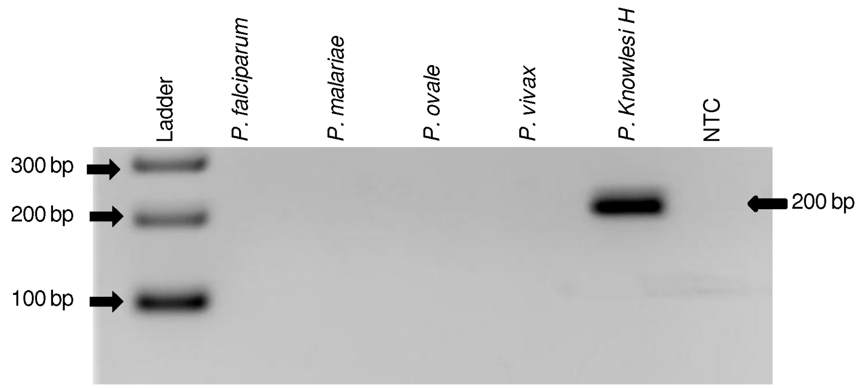


Figure 6.2. Primer Pkr140-5 tested with the 5 human-infecting malaria parasite species.

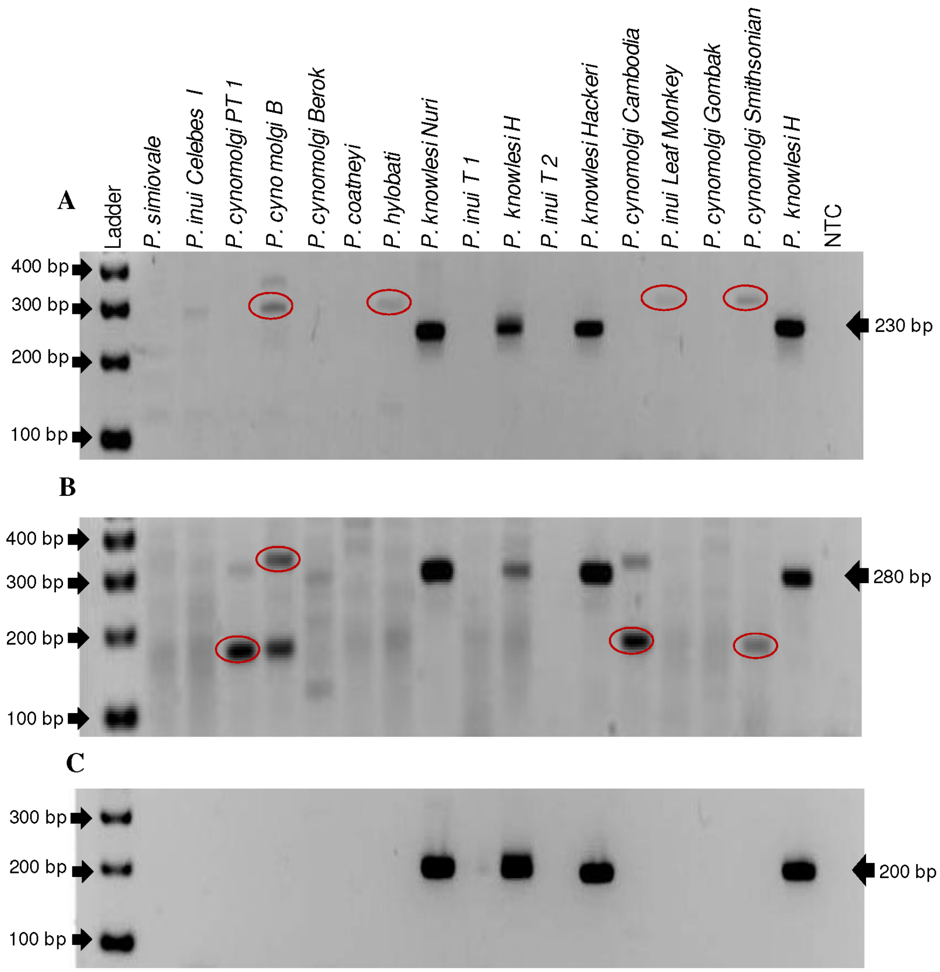


Figure 6.3. Specificity of the *P. knowlesi* primers tested using simian-infecting malaria parasite species.

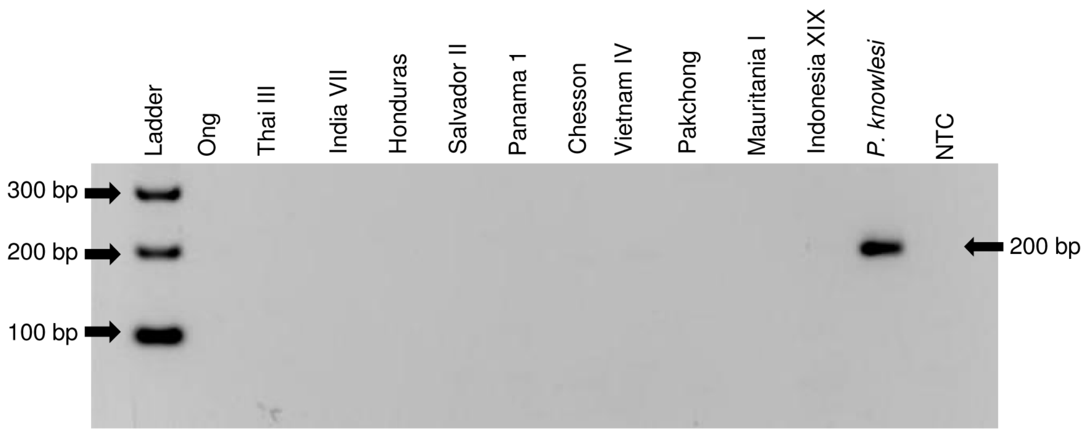


Figure 6.4. Primer set Pkr140-5 does not cross react with *P. vivax.*

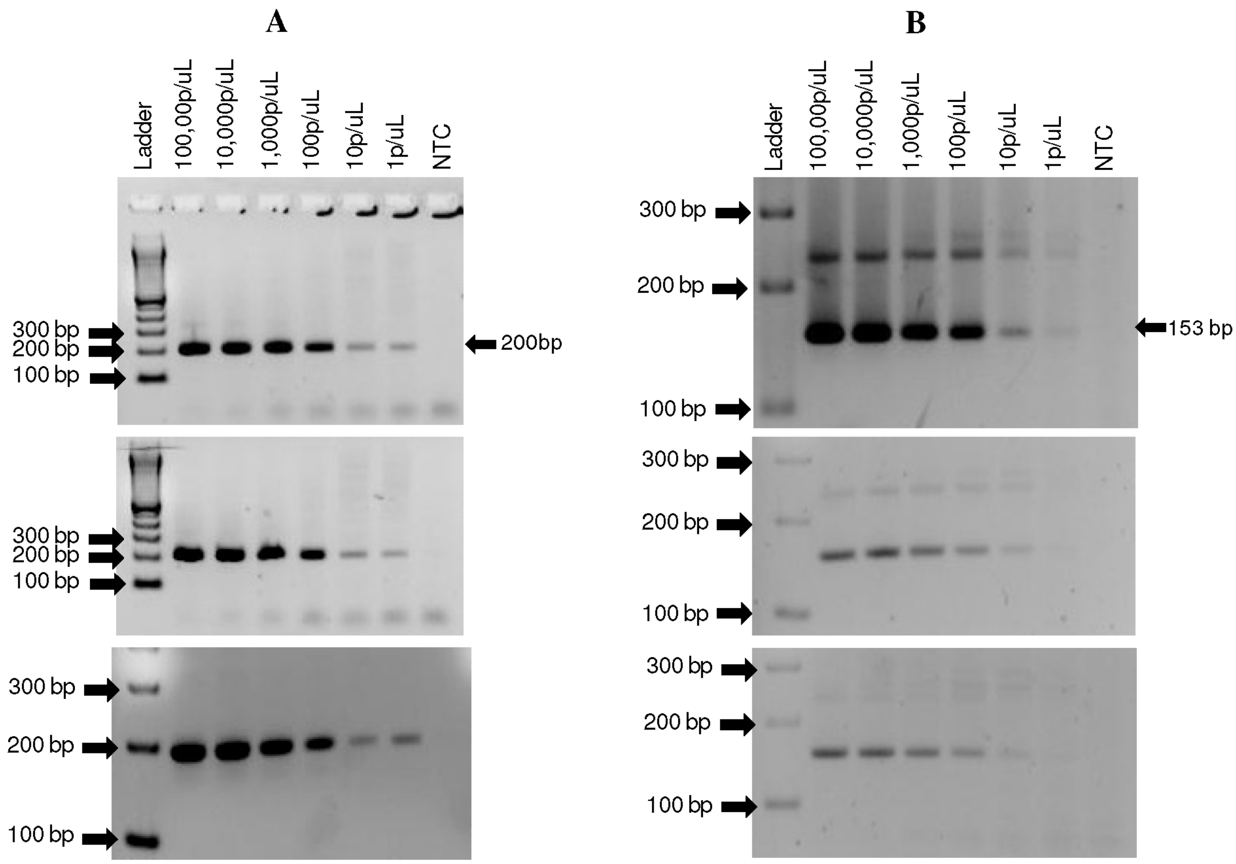


Figure 6.5. Limits of detection of primer set Pkr140-5.

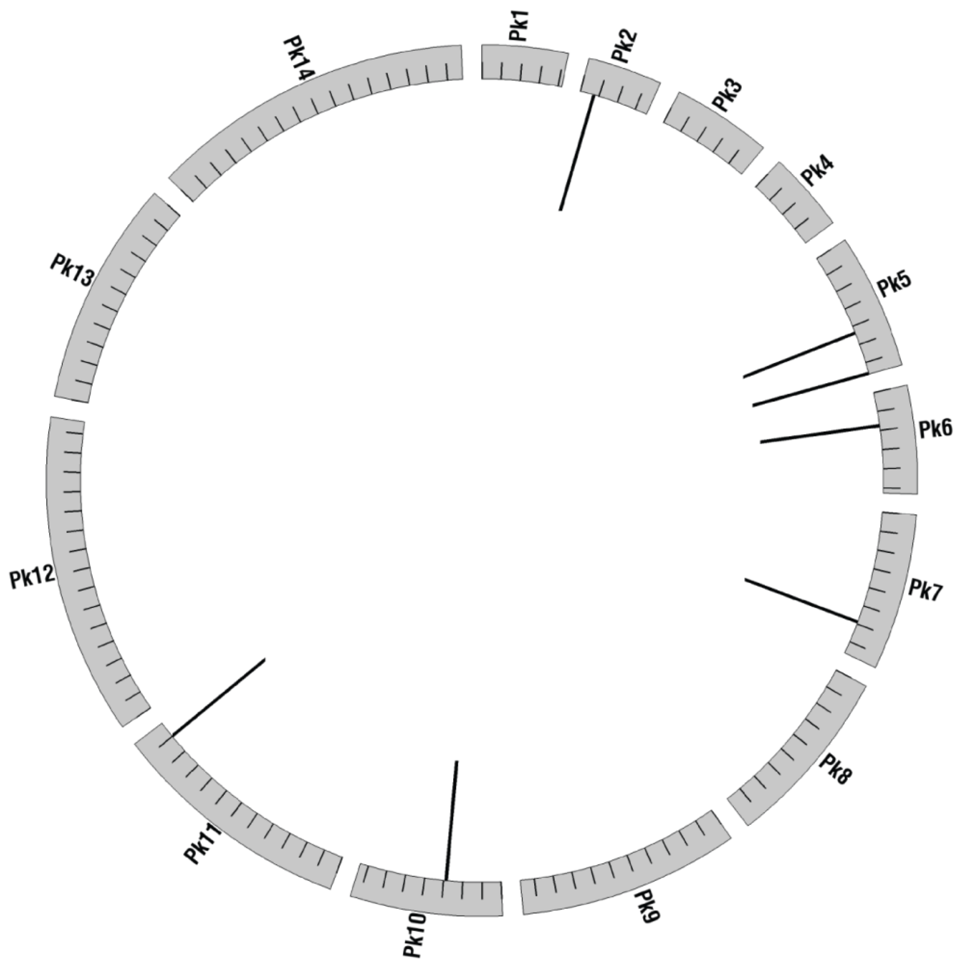


Figure 6.6. Spatial distribution of Pkr140 sequence targets across the *P. knowlesi* genome.

Table 6.1. Sequence of the novel Pkr140-5 primer set.

Primer	Sequence
Forward	5′- CAGAGATCCGTTCTCATGATTTCCATGG -3′
Reverse	5′- CTRAACACCTCATGTCGTGGTAG-3′

6.2 Author contributions

Chapter 2

JD JO and JCK designed the data mining experiments, JO GS and JD performed the bioinformatics experiments. JD, JO, DSP and JCK analyzed the bioinformatics data. AD with supervision from NL designed primer sequences. AD and NL performed the experiments with supervision from VU. DS, AMK, LV, AAE, SPK and JWB provided samples, participated in discussions and reviewed the paper. JO, JD, AD, NL, GS, JCK and VU wrote the paper.

Chapter 3

MM and CL generated the *C. parvum* qPCR data. SJJ clustered the resulting expression profiles, ran MEME on all clusters and evaluated GO term enrichment. JO built and purified AP2 domain constructs, evaluated similarity of overrepresented motifs, analyzed motif expression patterns, and analyzed upstream regions of functionally related groups of genes. JO and SJJ wrote the paper, with editing by JCK.

Chapter 4

Experiments conceived by JO, JCK and ML. YP and AS performed the protein-binding microarrays. JO performed all other bench/bioinformatics work and analyses. JO wrote the paper, with editing by JCK.

Chapter 6

JO, JD, and GS performed the datamining experiments and analyses. NL, MP, IG, MX, AJD, DSP and JB designed primers, performed bench experiments, provided samples or otherwise provided discussion. NL, JO, JD, JCK, and VU wrote the paper.