Specific DNA-binding by Apicomplexan AP2 transcription factors

Erandi K. De Silva*, Andrew R. Gehrke†, Kellen Olszewski*, Ilsa León*, Jasdave S. Chahal*, Martha L. Bulyk†‡§, and Manuel Llinás*¶

*Department of Molecular Biology and Lewis–Sigler Institute for Integrative Genomics, Princeton University, 246 Carl Icahn Laboratory, Princeton, NJ 08544; †Division of Genetics, Department of Medicine, and ‡Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; and §Harvard/Massachusetts Institute of Technology Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115

Edited by Thomas E. Wellems, National Institutes of Health, Bethesda, MD, and approved April 1, 2008 (received for review March 3, 2008)

Malaria remains one of the most prevalent infectious diseases worldwide, affecting more than half a billion people annually. Despite many years of research, the mechanisms underlying transcriptional regulation in the malaria-causing *Plasmodium* spp., and in Apicomplexan parasites generally, remain poorly understood. In *Plasmodium*, few regulatory elements sufficient to drive gene expression have been characterized, and their cognate DNA-binding proteins remain unknown. This study characterizes the DNA-binding specificities of two members of the recently identified Apicomplexan AP2 (ApiAP2) family of putative transcriptional regulators from *Plasmodium falciparum*. The ApiAP2 proteins contain AP2 domains homologous to the well characterized plant AP2 family of transcriptional regulators, which play key roles in development and environmental stress response pathways. We assayed ApiAP2 protein–DNA interactions using protein-binding microarrays and combined these results with computational predictions of coexpressed target genes to couple these putative *trans* factors to corresponding *cis*-regulatory motifs in *Plasmodium*. Furthermore, we show that protein–DNA sequence specificity is conserved in orthologous proteins between phylogenetically distant Apicomplexan species. The identification of the DNA-binding specificities for ApiAP2 proteins lays the foundation for the exploration of their role as transcriptional regulators during all stages of parasite development. Because of their origin in the plant lineage, ApiAP2 proteins have no homologues in the human host and may prove to be ideal antimalarial targets.

gene expression | malaria | *Plasmodium* | protein-DNA interaction | protein-binding microarray

Malaria is caused by the *Plasmodium* parasite and is responsible for >1.5 million annual deaths worldwide (1). Virtually all clinical symptoms are confined to the 48-hour asexual stage of development, during which the parasite matures and undergoes major morphological changes within the host red blood cell. Although the apparent complexity of parasite development in the human host suggests the existence of a finely tuned transcriptional network, remarkably little is known regarding gene regulation in *Plasmodium* spp. and related parasites of the class *Apicomplexa*. This knowledge void is surprising, given the wealth of recent whole-genome gene expression data analyzing virtually all stages of development (2–5). Bioinformatic analysis of the *Plasmodium falciparum* genome demonstrates a dearth of identifiable specific transcription factors (6, 7). This observation has led to speculation that gene expression in *Plasmodium* spp. is preferentially regulated posttranscriptionally through a combination of mRNA degradation, translational repression, and epigenetic mechanisms (8–10). In accordance with this, the genome reveals a near-complete set of chromatin remodeling machinery and an abundance of proteins containing RNA-binding domains (7). However, there is compelling evidence that the mechanisms of transcriptional regulation in *Plasmodium* spp. are more similar to other eukaryotic systems.

Transcription in *Plasmodium* uses minimal promoter regions that produce monocistronic mRNAs, including both 5′ and 3′ untranslated regions, and which often contain introns (11, 12). Furthermore, the basal eukaryotic transcription factors are conserved, including a canonical TATA-box-binding protein and RNA polymerase II-dependent messenger RNA production (6, 7). In addition, the *Plasmodium* transcriptome follows a dynamic cascade of periodic gene expression initiated upon invasion of the red blood cell (2, 4). In this cascade, most genes are expressed only once in a "just-in-time" fashion, suggesting an important role for stage-specific regulation of gene expression. Despite a number of previous studies suggesting the presence of stage-specific nuclear factors (13–15), the identification of factors that might modulate this expression cascade has remained elusive in *Plasmodium*.

A recent bioinformatic search for DNA-binding domains identified the Apicomplexan AP2 (ApiAP2) family of proteins present not only in *Plasmodium* spp. but also in all Apicomplexan parasites sequenced to date (16). ApiAP2 proteins exhibit weak homology to a family of transcription factors in plants called the AP2/ERF DNA-binding proteins. In the plant *Arabidopsis thaliana*, this family comprises the second largest class of regulators, with >145 members (17). The plant AP2/ERF domain is ≈60 aa in size and can be found either as a single module or in a tandem arrangement (18). In plants, the architecture of these proteins is correlated with their function: single-domain AP2 genes are involved in environmental stress responses from thermotolerance to dehydration and ethylene response, whereas proteins with tandem domains have been implicated in plant development (18).

There are 26 members of the *P. falciparum* ApiAP2 protein family, all currently annotated as conserved hypothetical proteins (19). Subsets of *Plasmodium* ApiAP2 proteins are expressed throughout the four stages of the intraerythrocytic development cycle (IDC): the ring, trophozoite, early schizont, and late schizont stages (2, 16). As in plants, the predicted AP2 domains in *Plasmodium* are ≈60 aa in size and can be found as both single and tandem domain architectures, although there is an additional architecture containing three AP2 domains in a single protein. Within the *Plasmodium* spp., there is virtually 100% identity between orthologous AP2 domains, and often homologues of lower sequence similarity can be identified in distant *Apicomplexa* (e.g., *Cryptosporidium* vs. *Plasmodium*) (16).

Author contributions: E.K.D.S. and M.L. designed research; E.K.D.S., A.R.G., K.O., I.L., and J.S.C. performed research; M.L.B. contributed new reagents/analytic tools; E.K.D.S., A.R.G., K.O., and M.L. analyzed data; and E.K.D.S., A.R.G., K.O., and M.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¶To whom correspondence should be addressed. E-mail: mllinas@princeton.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0801993105/DCSupplemental.

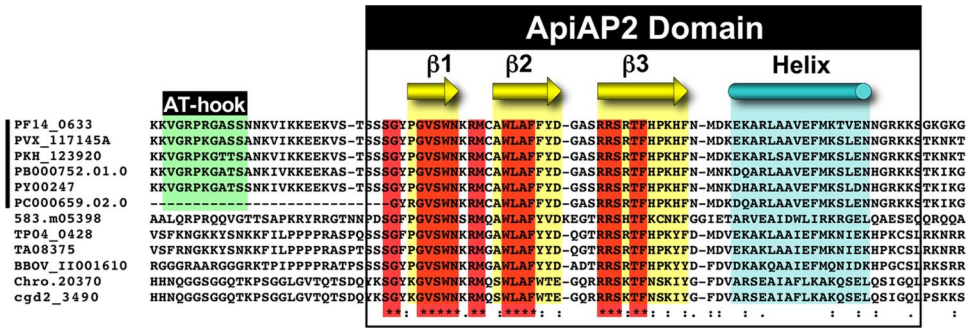

Fig. 1. Alignment of the AP2 domain from PF14_0633 (amino acids 63–123) to orthologues in five additional *Plasmodium* spp. and six Apicomplexan species. The AP2 domain (boxed) is highly conserved across all species. Conservation of residues is most significant in the three β-strands (shaded yellow) of the AP2 domain and is less significant in the α-helix (shaded blue). The AT-hook domain (shaded green) is found upstream of the AP2 domain in *Plasmodium* spp. (vertical black line). Absolutely conserved residues likely to be involved in DNA binding are highlighted in red. Secondary structure predictions were made by using Jnet (39). PF, *P. falciparum*; PVX, *Plasmodium vivax*; PKH, *Plasmodium knowlesi*; PB, *Plasmodium berghei*; PY, *Plasmodium yoelii*; PC, *Plasmodium chabaudi*; 583, *T. gondii*; TP, *Theileria parvum*; TA, *Theileria annulata*; BBOV, *Babesia bovis*; Chro, *Cryptosporidium hominis*; cgd2, *Cryptosporidium parvum*.

The AP2 domains in *P. falciparum* are located within the context of much larger proteins ranging in size from 200 to >4,000 aa, although there is generally very low homology in regions outside of the AP2 domains themselves.

In this study, we demonstrate the DNA-binding specificities of two ApiAP2 proteins representing different classes of AP2 domain architectures from *P. falciparum* using protein-binding microarrays (PBMs) (20, 21). We report that these AP2 domains have a high specificity for unique DNA sequence motifs found in the upstream regions of distinct sets of genes that are coregulated during asexual development. We also show that despite extensive sequence divergence between ApiAP2 proteins from distantly related Apicomplexan species (*P. falciparum* and *Cryptosporidium parvum*), the DNA-binding specificities of orthologous pairs of AP2 domains are fundamentally conserved, although their downstream targets are not. This demonstrates a previously undescribed interaction between *Plasmodium trans* factors and their putative target sequences. These results, along with computational predictions of genome-wide motif enrichment, allow us to begin constructing a network of regulatory interactions in *Plasmodium*.

Results

ApiAP2 Proteins of Differing Domain Architecture. We selected two different architectures of *P. falciparum* AP2 domains that resemble the single and tandem domain plant architectures. The first gene, *pf14_0633*, encodes an 813-aa protein, is maximally expressed during the ring stage of development (2) and contains a single 60-aa AP2 domain and an adjacent AT-hook DNA-binding domain (16). PF14_0633 not only has orthologues in all other sequenced *Plasmodium* genomes but also in all of the other sequenced Apicomplexan genomes (Fig. 1). Although the AT-hook is conserved only in *Plasmodium* spp., residues within the AP2 domain are well conserved in all *Apicomplexa*.

The second ApiAP2 gene examined, *pff0200c*, shows maximal expression in late-stage parasites and encodes a 1,979-aa protein possessing two AP2 domains in tandem, linked by a conserved 17-aa sequence. In *Plasmodium* spp., the amino acid sequence identity across the orthologous tandem AP2 domains of PFF0200c approaches 95% [supporting information (SI) Fig. S1]. In contrast, the individual AP2 domains of PFF0200c share only 35% identity with each other. In plants, it has been shown that the two tandem AP2 domains of AINTEGUMENTA in *A. thaliana*, which share 43% identity, bind two different DNA motifs (22). The functional relevance of this sequence divergence between tandem domains in *P. falciparum* remains unknown.

***P. falciparum* AP2 Domains Bind Specific DNA Motifs.** To elucidate whether isolated AP2 domains from *P. falciparum* bind DNA, and if so, to determine the specificity of binding, we assayed purified AP2 domains using PBMs. PBMs are a methodology used to determine the specificity of protein–DNA interactions and have been extensively used to characterize transcription factors from yeast to human (20). The array is not organism-specific but contains all possible 10-mer DNA sequences spread across 44,000 double-stranded DNA 60-mers, providing extensive negative and positive specificity controls across the array (20). As a proof of principle, we first measured the binding of a 63-aa *A. thaliana* ERF1 AP2 domain (residues 144–206) and recovered the expected GCC-box motif reported in the literature (data not shown) (23).

Using PBMs, we obtained distinct and highly specific DNA sequence motifs for the single AP2 domain from PF14_0633 and for the double AP2 domain from PFF0200c. As is common for transcription factor-binding sites, these motifs are palindromic, with PF14_0633 binding the TGCATGCA consensus sequence and PFF0200c-binding GTGCAC (Fig. 2, Dataset S1). The motifs for the *P. falciparum* AP2 domains were noticeably more AT-rich than the canonical GCC-box motifs that are bound by plant AP2 domains. This is consistent with predictions that the regulatory motifs would be more AT-rich than other eukaryotic *cis*-acting motifs, given that the AT-content in the intergenic region of *P. falciparum* approaches 90% (16, 19).

Of the 26 ApiAP2 proteins in *P. falciparum*, all are conserved in the other sequenced *Plasmodium* spp., whereas only a small subset span all Apicomplexan genomes. The AP2 domain from the *C. parvum* gene *cgd2_3490* has 47% identity (68% similarity) to the plasmodial PF14_0633 (Fig. 1). To determine whether the conserved residues between evolutionarily distant orthologues were sufficient to confer similar DNA-binding specificity, we tested the cgd2_3490 AP2 domain by PBM. Remarkably, our results show that the *C. parvum* AP2 domain and its PF14_0633 *Plasmodium* orthologues have highly similar DNA-binding specificities (Fig. 2). This unexpected result demonstrates that the Apicomplexan parasites have conserved not only the AP2 DNA-binding domain architecture, but also the sequence specificity.

Although the AP2 region for PF14_0633 is conserved between *Apicomplexa*, the AT-hook DNA-binding motif is found only in the *Plasmodium* spp. (Fig. 1). We tested the possible contribution of this additional domain to DNA-binding using a GST-fusion protein containing both the AT-hook and the AP2 domain from PF14_0633 and found no change in the DNA-binding motif recognized (data not shown). This suggests that DNA binding by the AP2 domain is sufficient for specific

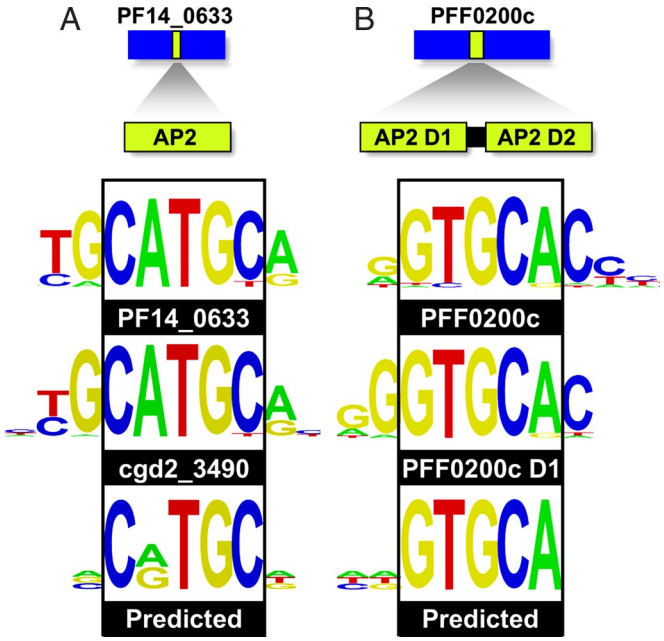

Fig. 2. DNA motifs specifically bound by AP2 domains predicted using PBMs and computational analysis (FIRE algorithm). (*A*) The core nucleotides (boxed) in the motif specifically bound by the *P. falciparum* AP2 domain of PF14_0633 are highly similar to those bound by its *C. parvum* orthologue cgd2_3490 (*Top* and *Middle*). The motifs determined from the PBM are very similar to motifs predicted using the FIRE algorithm (*Bottom*, Predicted) (24). (*B*) The PBM-derived motif bound by the tandem AP2 domains of PFF0200c (*Top*) is highly similar to the motif bound by the first domain alone (*Middle*). Domain 2 of PFF0200c did not bind a specific DNA motif (data not shown). Both PBM-derived motifs for PFF0200c match the computationally predicted motif (*Bottom*).

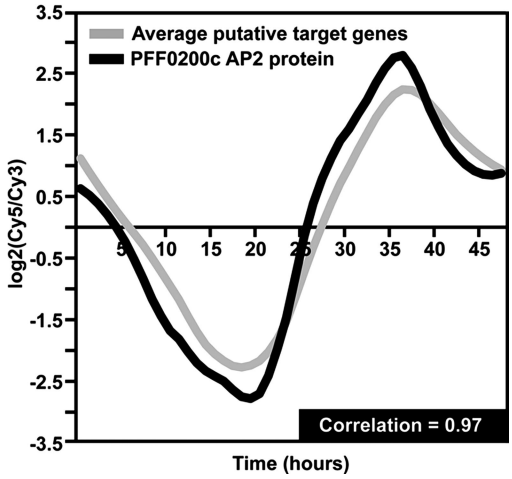

Fig. 3. Blood-stage gene expression profile of the PFF0200c ApiAP2 protein compared with the averaged expression profiles of putative FIRE-predicted target genes. The 48-hour gene expression profiles are positively correlated with a Pearson correlation coefficient of 0.97.

binding, although it is possible that the AT-hook region may increase affinity through nonspecific interactions with the DNA.

In plants, it has been demonstrated that proteins with tandem AP2 domain architectures absolutely require both domains for specific DNA binding (22). To explore the relative contribution of each AP2 domain in the PFF0200c tandem AP2 architecture, we dissected the protein and tested each domain separately using PBMs. Surprisingly, domain 1 of PFF0200c was sufficient for specific DNA binding and bound the identical GTGCAC motif as the full-length tandem double domain of PFF0200c (Fig. 2). The isolated PFF0200c domain 2 did not demonstrate any specific protein–DNA interaction. These results suggest that, in contrast to plants, the second AP2 domain in PFF0200c may not contribute to the protein's DNA-binding specificity.

Prediction and Validation of AP2 Target Genes. Ultimately, we are interested in the genes that may be regulated by ApiAP2 proteins in *Plasmodium*. The two binding sites identified biochemically in this study are highly similar to two motifs predicted independently by Elemento *et al.* (24) using the Finding Informative Regulatory Elements (FIRE) algorithm (Fig. 2). The FIRE algorithm compiles a list of candidate target genes associated with each predicted motif (Dataset S2). These target genes share two characteristics: (*i*) at least one instance of the considered motif in their promoter regions and (*ii*) peak mRNA abundance levels within a particular phase of the IDC transcriptome that is significantly enriched in genes whose promoters contain the motif (2). We compared the expression profile of PFF0200c to the mean profile of its highest confidence target set (66 predicted target genes, Dataset S2) containing the GTGCA motif (Fig. 3). The highly positive Pearson correlation (0.97) between PFF0200c, and its putative targets suggests that it functions to activate this set of genes (Fig. 3). We also note that the induction of gene expression (20–35 h postinvasion) for PFF0200c precedes the average expression profile of the putative target genes by 1–2 h. The mean profile from the top cluster (34 genes) containing the PF14_0633 consensus motif, however, is not strongly correlated with the PF14_0633 expression profile (Pearson correlation = 0.25) (Fig. S2, Dataset S2) and may suggest that other auxiliary regulatory factors are involved for expression timing. The striking and independent convergence of these computational motif predictions with our biochemical DNA-binding specificity results strongly suggests that these motifs are of significant biological importance.

From the FIRE-predicted gene sets, we selected candidate genes to be tested for specific DNA–protein interactions by EMSA. A radiolabeled 40-bp sequence (Table S1), found upstream of the putative target gene *pfi0540w* and containing the TGCATGCA motif, could be specifically shifted using the purified AP2 domain from PF14_0633 (Fig. 4). This interaction could be competed by an unlabeled oligonucleotide of identical sequence, but a 50-fold excess of a related, mutant oligonucleotide (AT to GC change in the motif) did not disrupt binding. We also confirmed the PFF0200c-binding interaction using an oligonucleotide sequence from the upstream region of a candidate target gene *mal7p1.119* containing the GTGCAC DNA motif (Fig. 4). These results validate our PBM data and extend this analysis by demonstrating that ApiAP2 domains can bind specifically to sequences present upstream of putative target genes in the *Plasmodium* genome.

Analysis of Putative Regulons. The FIRE algorithm predicts a total of 194 putative targets (Dataset S2) containing the GTGCAC motif associated with PFF0200c, of which 115 (59.3%) are annotated as hypothetical. Gene Ontology (GO) analysis reveals significant enrichment for genes involved in protein modification (P value = 9.85e-5), particularly protein phosphorylation (P = 3.83e-4) and cysteine peptidase activity (P = 1.30e-3), and in genes associated with the rhoptry organelle (P = 4.12e-5) and apical complex (P = 6.81e-6) or involved in the invasion machinery (P = 1.30e-3). These targets include the rhoptry-associated proteins RAP1, RAP2, and RAP3; the merozoite surface proteins MSP1, MSP7, and MSP9; and the cytoadherence-linked asexual proteins CLAG2, CLAG3.1, CLAG3.2, and CLAG9. This suggests that PFF0200c regulates late-stage genes involved in the critical process of preparing the parasite for host cell rupture and reinvasion.

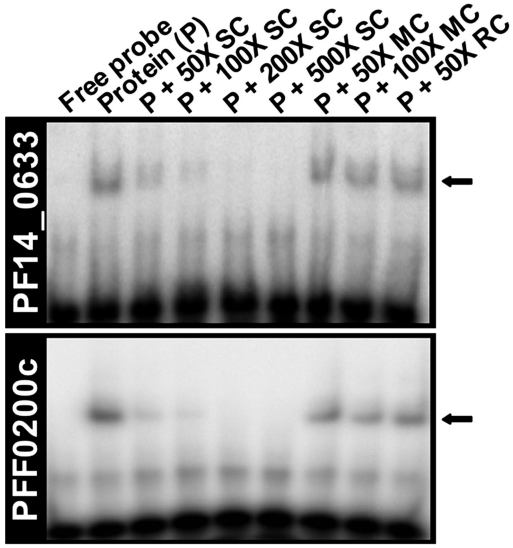

Fig. 4. EMSA of AP2 domains. The AP2 domain from PF14_0633 (*Upper*) and the tandem AP2 domains from PFF0200c (*Lower*) were used to shift 40 bp of a radioactively labeled DNA probe derived from the upstream region of a computationally predicted target genes (PFI0540w and MAL7P1.119, respectively) (lane 2). Competition experiments show that an unlabeled specific competitor (SC) can deplete the labeled shifted band (lanes 3–6). An unlabeled mutant competitor (MC, lanes 7 and 8) or a random competitor (RC, lane 9) cannot deplete the shift.

The extremely high conservation between the PF14_0633- and cgd2_3490-binding sites led us to compare their putative regulons to determine whether these proteins regulate similar gene sets. Of the 5,460 *P. falciparum* genes, 31.5% (1,722) have homologues in *C. parvum*. Surprisingly, only 26 of the 127 FIRE-predicted targets of the PF14_0633 motif (20.4%, $P = 3.43e{-}3$) are conserved between these organisms, suggesting that, whereas the sequence specificity of these two AP2 domains is absolutely conserved, a significant rewiring of the transcriptional network has occurred since these species diverged. In contrast, there is no significant deviation from the background frequency homology for the 194 FIRE-predicted targets of PFF0200c (37.6%, $P = 0.972$), as is expected because of the lack of a PFF0200c homologue in *C. parvum*.

We used ScanACE to examine the 2.0-kbp 5′ upstream regions from every gene in the *P. falciparum* and *C. parvum* genomes for the occurrence of the PF14_0633 and cgd2_3490 AP2 DNA motif, respectively. The 1,003 *C. parvum* genes containing at least one instance of the cgd2_3490 DNA binding motif are most significantly enriched for transmembrane proteins ($P = 5.40e{-}9$), whereas the 775 putative PF14_0633 targets are most enriched for cytoadherence to the microvasculature ($P = 2.51e{-}12$). There are no common enriched annotations between the two sets, further indicating that they target very different regulons. Interestingly, we found members of the upsB and upsC *var* gene subfamilies in the PF14_0633 target set. The *var* genes in *P. falciparum* encode PfEMP1 (erythrocyte membrane protein) one of the major surface antigens involved in cytoadherence and host immune evasion (25). Sequence alignments of the upsB upstream regions revealed an almost perfectly conserved instance of the consensus PF14_0633 motif CATGCA between 1,478 and 1,352 bp of the ATG and another instance between 1,218 and 1,093 bp upstream, which corresponds to the SPE1 site identified by Voss *et al.* (15). These data suggest that PF14_0633 may play a role in *var* gene regulation. In addition, the potential target set also includes PFF0200c, which contains an exact match to the sequence TGCATGCA 1,746 bp upstream from the ATG start. This raises the intriguing possibility that PFF0200c is itself regulated by PF14_0633, in what may be one link in an ApiAP2 regulatory cascade (Fig. S3).

Discussion

The majority of the *P. falciparum* ApiAP2 family of proteins are expressed throughout asexual development and represent the first candidate set of transcriptional regulators in this parasite (16). Our analysis of two *Plasmodium* asexual-stage ApiAP2 proteins establishes that these proteins bind highly specific DNA sequences that are enriched in the upstream regions of subsets of *Plasmodium* genes. We have further demonstrated that two representative architectures (single and tandem AP2 domains) specifically interact with unique DNA sequence motifs, although it remains unclear why up to three AP2 domains are present in plasmodial ApiAP2 proteins. We also examined the significance of the conservation of AP2 domains across *Apicomplexa* by characterizing an orthologous AP2 domain from *C. parvum*. The conservation of DNA-binding specificities likely occurs through the highly conserved predicted β-strand residues (Fig. 1, Fig. S1). However, we provide computational evidence that the putative target genes are not conserved between distant *Apicomplexa*, implying that the regulons in different organismal classes are likely to be highly diversified. These results further caution against the inference of transcriptional regulatory networks based solely on homology, as was recently demonstrated for the yeast transcription factor Ste12, which targets different gene sets in three yeast species (26).

Our results also demonstrate that the AP2 domain from *Plasmodium* is sufficient for specific DNA binding and lead us to question the functional roles of other regions in the ApiAP2 proteins. Outside of the AP2 domains, sequence homology between orthologous ApiAP2 proteins deteriorates, making the identification of other relevant domains such as transcriptional activation domains difficult. The 26 ApiAP2 proteins in *P. falciparum* contain no other known functional domains in the Pfam repository, with the exception of PF14_0633, which contains an AT-hook. In light of this, we examined all *P. falciparum* ApiAP2 proteins for the presence of additional informative sequences. We were unable to detect any motifs for apicoplast targeting, mitochondrial transit, endoplasmic reticulum trafficking, transmembrane domains, or host cell surface targeting by the PEXEL/VTS. However, we are able to identify classical lysine- and arginine-rich nuclear localization signals (Table S2) in the majority of the ApiAP2 proteins using PredictNLS (27), supporting the conclusion that this protein family consists of transcription factors.

Although the AP2/ERF proteins are established as transcriptional regulators in plants, there is currently little understanding of how these proteins interact with the basal transcriptional machinery. We hypothesize that ApiAP2 proteins may mediate specific protein–protein interactions. A recent *Plasmodium*-based global yeast two-hybrid study suggests that ApiAP2 proteins interact with each other and with chromatin remodeling factors, including the *Plasmodium* histone acetyltransferase GCN5 (28). Binding to chromatin remodeling factors may serve to recruit these complexes to specific chromosomal locations and facilitate interaction with the core transcription machinery. One report in plants has suggested that the *A. thaliana* AP2 protein, CBF1, interacts with GCN5 and two transcriptional adaptor proteins (ADA2 and ADA3) (29). Alternatively, protein–protein interactions may occur directly through an AP2 domain, as has been seen for the plant proteins DORNRÖSCHEN and DORNRÖSCHEN-LIKE (30). As we gain further knowledge regarding the DNA-binding specificity of all AP2 domains in *Plasmodium* spp., it will become equally important to biochemically identify interacting partners to define the full network of transcriptional regulation.

A similar motif to the GTGCAC motif bound by PFF0200c has been identified upstream of the upsB-type *var* genes (SPE2) in *P. falciparum* and was found repeated 5–18 times 565–1,035 bp

upstream of the transcription start site (−1,167 bp from ATG) (15). This motif was shown to be specifically bound by a nuclear factor expressed in late-stage parasites and was further found to induce late-stage expression from an otherwise minimal promoter (31). This has been interpreted to suggest a late-stage *var* gene silencing role for this nuclear factor. Both the binding activity and reporter gene induction, however, also coincide with the maximal expression of PFF0200c and its predicted target gene set (Fig. 3). If PFF0200c is the nuclear factor identified (15), it may serve dual roles as a silencer of *var* genes and as an activator of late-stage invasion genes. However, it is also possible that there are two different DNA-binding proteins that recognize the GTGCAC motif.

Interestingly, the TGCATGCA sequence bound by the AP2 domains of PF14_0633 and *C. parvum* cgd2_3490 is the most highly conserved and overrepresented motif in the upstream regions of all sequenced Apicomplexan genomes (32). This motif is seven times overrepresented in *P. falciparum*, and 11×, 18×, and 18× overrepresented in *Toxoplasma gondii*, *Cryptosporidium parvum*, and *Eimeria tenella*, respectively. Why this motif is so ubiquitous remains unexplored, although it is well established that the presence of a motif does not necessarily imply that it is functional. Future chromatin immunoprecipitation experiments will help to decipher the precise targets for PF14_0633 and PFF0200c.

We note that the FIRE algorithm reveals significant enrichment for 21 distinct motifs from all phases of the intraerythrocytic developmental cycle (24). This number is similar to the number of ApiAP2 proteins, which are actively transcribed during the asexual stages, supporting the idea put forth by Balaji *et al.* (16) that gene expression during the IDC could be controlled by a cascade of ApiAP2 proteins, each regulating a specific phase of development (16) (Fig. S3). Furthermore, we find that the FIRE-predicted DNA motifs also occur in the upstream regions of the majority of ApiAP2 genes, suggesting they may regulate one another. In fact, the upstream region of PFF0200c contains five copies of its own GTGCA motif (−240 to −839 bp from the ATG) and a copy of the TGCATGCA motif (−1,746 bp from the ATG) recognized by PF14_0633.

This study links specific DNA-binding proteins to sequences found in putative regulatory regions of Apicomplexan genomes. The existence of a possible network of ApiAP2 DNA-binding proteins contributes an important piece to the puzzle of gene regulation in *Plasmodium.* Although it remains to be determined how ApiAP2 proteins function as transcriptional regulators, the DNA-binding sequence specificity of these proteins, their conservation across *Apicomplexa*, and the highly coherent expression patterns of their predicted downstream targets suggests an essential role in regulating parasite development. Additionally, several *P. falciparum* ApiAP2 proteins show no detectable expression during the IDC and may prove to be functionally important in the mosquito or exoerythrocytic liver stages.

Materials and Methods

PBMs. N-terminal GST fusion proteins were made by using the pGEX4-T1 vector (GE Healthcare) and AP2 domains from PF14_0633 (residues 63–123), AT-hook plus AP2 domain from PF14_0633 (residues 37–123), cgd2_3490 (residues 340–399), the two tandem domains from PFF0200c (residues 177–312) and its dissected domains. Domain boundaries were defined as (16), PCR-amplified and cloned into BamHI and XhoI restriction sites in pGEX4-T1. Proteins were expressed in *Escherichia coli* BL21 (RIL Codon PLUS) (Stratagene) cells at 25°C and purified by using Uniflow Glutathione Resin (Clontech). Proteins were eluted in 10 mM reduced glutathione, 50 mM Tris·HCl, pH 8.0. Purity was verified by silver stain SDS/PAGE and Western blot analysis with an anti-GST antibody (Invitrogen).

A minimum of two PBM experiments were performed as described (20, 21) for each protein tested. Briefly, this methodology utilizes a custom 60-mer single-stranded Agilent DNA microarray with ≈44,000 features covering all possible 10-mers. Primer extension from a universal 24-mer region on all 60-mers generates a double-stranded DNA microarray platform. Purified proteins were diluted to a final concentration of 100–500 nM in PBS, 2% (wt/vol) milk, 51.3 ng/μl salmon testes DNA (Sigma), 0.2 μg/μl BSA (New England Biolabs), and incubated for 1 h at 20°C. After washing, specific DNA–protein interactions were visualized by using a GSI Lumonics ScanArray 5000 scanner to detect fluorescence from an Alexa488-conjugated anti-GST antibody. After data normalization as described (20, 21), enrichment scores were calculated for all 8-mers, and the ''Seed-and-Wobble'' algorithm (20) was used to construct position weight matrices (Dataset S1) that represent the PBM-derived DNA-binding specificities, which we represent using the Web-based tool enoLOGOS (33).

Computational Analysis. Data from the FIRE analysis of the *P. falciparum* transcriptome (24) was downloaded from http://tavazoielab.princeton.edu/FIRE. Genes with highly conserved upstream regions, including the *var*, *rifin*, and *stevor* subfamilies, are filtered to contain only one randomly selected member per cluster by the FIRE algorithm, not to bias the sequence analysis. For our analyses, we defined the putative targets of a given motif as those genes, which (*i*) possess at least one instance of the motif within the sequence region 1 kb upstream from their ATG start and (*ii*) fall into a phase bin that is significantly enriched for that motif.

Peptides involved in targeting to various organelles and the host cell surface and putative transmembrane domain were analyzed by using tools at PlasmoDB.org (34). GO analyses were performed in GOLEM (35) with the most recent *P. falciparum* annotation file found on the Gene Ontology web site (1/21/2008). The reported *P* values were corrected for multiple hypothesis testing by using a false discovery rate (FDR) of 0.05. Because of the lack of an official GO annotation for *C. parvum*, functional enrichment analyses were performed by using the DAVID Functional Analysis Tool with an FDR threshold of 0.05 (36). To determine orthologous gene pairs, conservation analysis was performed by reciprocal best BLAST by using a lenient E-value cutoff of 1 on the *P. falciparum* and *C. parvum* predicted protein sequences [PlasmoDB 5.4 (34) and CryptoDB 3.7 release (37)]. Deviation from the background level conservation within a subset of genes was scored for statistical significance using the hypergeometric distribution. To derive expanded target sets for the PF14_0633 and cgd2_3490 motifs, we converted the PBM-derived nucleotide frequency matrices into ScanACE files (38) and searched the 2 kb upstream regions of every predicted ORF in the *P. falciparum* or *C. parvum* genomes. We used a significance threshold of 2 standard deviations below the mean score in the alignment file.

EMSA. Forty-base pair DNA oligonucleotides containing the PBM-derived motifs were taken from the 5′ upstream regions of predicted target genes PFI0540w and MAL7P1.119 (Table S1). ^{32}P radiolabeled double-stranded DNA probes (10 fmol) were incubated with purified GST-tagged AP2 domains from either PF14_0633 or PFF0200c in binding buffer (50 mM KCl, 5 mM $MgCl_2$, 1 mM EDTA, 50 ng/μl poly dI/dC) for 1 h at 25°C. Competition experiments included 50–500× excess unlabeled probes. Binding reactions were separated on 6% nondenaturing acrylamide gels run in 0.5× TBE buffer.

ACKNOWLEDGMENTS. We thank J. Kissinger (University of Georgia, Athens) for the *C. parvum* genomic DNA, B. Krizek (University of South Carolina, Columbia) for the *A. thaliana* genomic DNA, and E. Bush for cloning the *C. parvum* cgd2_3490 AP2 domain. We also thank A. Caudy, O. Elemento, J. Forman, M. Szpara, S. Tavazoie, and members of the Llinás lab at Princeton University; M. Berger at Brigham and Women's Hospital and Harvard Medical School; and A. Vaidya at Drexel University College of Medicine for valuable discussion and critical reading of the manuscript. This work was mainly supported by the Arnold and Mabel Beckman Foundation (M.L.), National Institutes of Health Grant P50 GM071508 (to M.L.), and in part by National Institutes of Health/National Human Genome Research Institute Grant R01 HG003985 (to M.L.B.). K.O. is funded by a National Science Foundation Graduate Research Fellowship.

1. Sachs J, Malaney P (2002) The economic and social burden of malaria. *Nature* 415:680–685.
2. Bozdech Z, *et al.* (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum. PLoS Biol* 1:E5.
3. Hall N, *et al.* (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* 307:82–86.
4. Le Roch KG, *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301:1503–1508.
5. Tarun AS, *et al.* (2008) A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc Natl Acad Sci USA* 105:305–310.
6. Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S (2005) Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: Conserved features and differences relative to other eukaryotes. *BMC Genomics* 6:100.
7. Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res* 14:1548–1554.

8. Hakimi MA, Deitsch KW (2007) Epigenetics in Apicomplexa: Control of gene expression during cell cycle progression, differentiation and antigenic variation. *Curr Opin Microbiol* 10:357–362.
9. Mair GR, *et al.* (2006) Regulation of sexual development of Plasmodium by translational repression. *Science* 313:667–669.
10. Shock JL, Fischer KF, DeRisi JL (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intraerythrocytic development cycle. *Genome Biol* 8:R134.
11. Horrocks P, Dechering K, Lanzer M (1998) Control of gene expression in Plasmodium falciparum. *Mol Biochem Parasitol* 95:171–181.
12. Lanzer M, Wertheimer SP, de Bruin D, Ravetch JV (1993*) Plasmodium*: Control of gene expression in malaria parasites. *Exp Parasitol* 77:121–128.
13. Gissot M, Briquet S, Refour P, Boschet C, Vaquero C (2005) PfMyb1, a Plasmodium falciparum transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. *J Mol Biol* 346:29–42.
14. Lanzer M, de Bruin D, Ravetch JV (1992) A sequence element associated with the Plasmodium falciparum KAHRP gene is the site of developmentally regulated protein-DNA interactions. *Nucleic Acids Res* 20:3051–3056.
15. Voss TS, Kaestli M, Vogel D, Bopp S, Beck HP (2003) Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters. *Mol Microbiol* 48:1593–1607.
16. 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.
17. Riechmann JL, *et al.* (2000) *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* 290:2105–2110.
18. Riechmann JL, Meyerowitz EM (1998) The AP2/EREBP family of plant transcription factors. *Biol Chem* 379:633–646.
19. Gardner MJ, *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419:498–511.
20. Berger MF, *et al*. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 11:1429–1435.
21. Mukherjee S, *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 36:1331–1339.
22. Krizek BA (2003) AINTEGUMENTA utilizes a mode of DNA recognition distinct from that used by proteins containing a single AP2 domain. *Nucleic Acids Res* 31:1859–1868.
23. Ohme-Takagi M, Shinshi H (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell* 7:173–182.
24. 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.
25. Frank M, Deitsch K (2006) Activation, silencing and mutually exclusive expression within the var gene family of Plasmodium falciparum. *Int J Parasitol* 36:975–985.
26. Borneman AR, *et al.* (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317:815–819.
27. Cokol M, Nair R, Rost B (2000) Finding nuclear localization signals. *EMBO Rep* 1:411–415.
28. LaCount DJ, *et al.* (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438:103–107.
29. Stockinger EJ, Mao Y, Regier MK, Triezenberg SJ, Thomashow MF (2001) Transcriptional adaptor and histone acetyltransferase proteins in Arabidopsis and their interactions with CBF1, a transcriptional activator involved in cold-regulated gene expression. *Nucleic Acids Res* 29:1524–1533.
30. Chandler JW, Cole M, Flier A, Grewe B, Werr W (2007) The AP2 transcription factors DORNROSCHEN and DORNROSCHEN-LIKE redundantly control Arabidopsis embryo patterning via interaction with PHAVOLUTA. *Development* 134:1653–1662.
31. Voss TS, *et al*. (2007) Alterations in local chromatin environment are involved in silencing and activation of subtelomeric var genes in Plasmodium falciparum. *Mol Microbiol* 66:139–150.
32. Ling KH, *et al.* (2007) Sequencing and analysis of chromosome 1 of Eimeria tenella reveals a unique segmental organization. *Genome Res* 17:311–319.
33. Workman CT, *et al.* (2005) enoLOGOS: A versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* 33:W389–92.
34. Bahl A, *et al.* (2003) PlasmoDB: The Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* 31:212–215.
35. Sealfon RS, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG (2006) GOLEM: An interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics* 7:443.
36. Dennis G, Jr, *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:P3.
37. Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC (2004) CryptoDB: The Cryptosporidium genome resource. *Nucleic Acids Res* 32:D329–31.
38. 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.
39. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40:502–511.