# Comparative Genomics of Transcriptional Control in the Human Malaria Parasite *Plasmodium falciparum*

Richard M.R. Coulson, Neil Hall and Christos A. Ouzounis

*Genome Res.* 2004 14: 1548-1554
Access the most recent version at doi:10.1101/gr.2218604

Letter

# Comparative Genomics of Transcriptional Control in the Human Malaria Parasite *Plasmodium falciparum*

Richard M.R. Coulson,[1,3] Neil Hall,[2] and Christos A. Ouzounis[1]

[1]*Computational Genomics Group, The European Bioinformatics Institute, European Molecular Biology Laboratory Cambridge Outstation, Cambridge CB10 1SD, United Kingdom;* [2]*The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*

**The life cycle of the parasite *Plasmodium falciparum*, responsible for the most deadly form of human malaria, requires specialized protein expression for survival in the mammalian host and insect vector. To identify components of processes controlling gene expression during its life cycle, the malarial genome—along with seven crown eukaryote group genomes—was queried with a reference set of transcription-associated proteins (TAPs). Following clustering on the basis of sequence similarity of the TAPs with their homologs, and together with hidden Markov model profile searches, 156 *P. falciparum* TAPs were identified. This represents about a third of the number of TAPs usually found in the genome of a free-living eukaryote. Furthermore, the *P. falciparum* genome appears to contain a low number of sequences, which are highly conserved and abundant within the kingdoms of free-living eukaryotes, that contribute to gene-specific transcriptional regulation. However, in comparison with these other eukaryotic genomes, the CCCH-type zinc finger (common in proteins modulating mRNA decay and translation rates) was found to be the most abundant in the *P. falciparum* genome. This observation, together with the paucity of malarial transcriptional regulators identified, suggests *Plasmodium* protein levels are primarily determined by posttranscriptional mechanisms.**

The most lethal form of human malaria is caused by infection with the parasite *Plasmodium falciparum*, which is transmitted by the mosquito *Anopheles gambiae*. Primarily, as a consequence of the unique biology of *P. falciparum*, efforts to develop novel treatments for malaria are greatly hindered. The organism is able to regulate its pattern of gene expression to generate a sequence of forms that exploit the highly diverse environments in which it resides (Bannister and Mitchell 2003). In the invertebrate vector, mammalian-infectious forms (sporozoites) are found in salivary glands, whereas in the vertebrate host, both intracellular erythrocytic (asexual trophozoite and sexual gametocyte stages) and extracellular (merozoites—the invasive stage of erythocytes) forms are observed in the bloodstream.

Transcriptional regulation has been shown to be involved with controlling gene expression in the various *P. falciparum* life cycle forms. Firstly, the parasites differentially express structurally distinct sets of rRNA genes in a stage-specific manner (Waters 1994). This may alter the properties of the ribosomes and, through altered translation rates, modify patterns of cell growth and development. Secondly, in situ *var* gene switching is mediated at the level of transcriptional inititiation (Scherf et al. 1998); this switching mechanism results in only one *var* gene being actively transcribed in a single parasite, whereas the remaining copies remain silenced, and plays an important role in *Plasmodium* survival and virulence. Finally, two *P. falciparum* promoters have been identifed whose activation marks the developmental switches executed during the sexual differentiation process (Dechering et al. 1999).

Mechanisms controlling transcriptional activation are fundamentally different between prokaryotes and eukaryotes (Struhl 1999), with transcription-associated protein (TAP) families showing very little sharing across the three domains of life (Kyrpides and Ouzounis 1999; Coulson et al. 2001). Additionally, in eukaryotes, the transcription initiation complex is highly conserved, whereas transcriptional regulator families are primarily taxon specific (Coulson and Ouzounis 2003). *P. falciparum* proteomics data (Florens et al. 2002; Lasonder et al. 2002) show a considerable level of regulation of gene expression taking place throughout its life cycle; only 6% of the proteins identified were present in all four of the parasite stages examined, with 49% of the sporozoite proteins unique to this stage (Florens et al. 2002). To investigate which aspects of the *Plasmodium* transcriptional process have diverged from those in crown eukaryote group species, the genome sequence of *P. falciparum* clone 3D7 (Gardner et al. 2002) was profiled with reference sets of TAPs and transcriptional regulator domains.

**[3]Corresponding author.**
**E-MAIL coulson@ebi.ac.uk; FAX 44-1223-494468.**
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2218604. Article published online ahead of print in July 2004.

## RESULTS

The occurence of genes within the *P. falciparum* genome encoding proteins associated with transcriptional regulation was examined using a sensitive sequence-searching method that enables the detection of sequences homologous to multiple alignments of protein domains. This search utilized 51 profile-hidden Markov models (HMMs) of transcriptional regulators (obtained via TransFac, see Methods). Additionally, the *Schizosaccharomyces pombe* (Wood et al. 2002), *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000), *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), *A. gambiae* (Holt et al. 2002), *Drosophila melanogaster* (Adams et al. 2000), and *Homo sapiens* (Lander et al. 2001; Venter et al. 2001) genomes, were searched for matches to these HMMs (Table 1). Only 17 of the HMMs had significant matches to sequences in the malarial genome (corresponding to 71 protein hits in total, denoted by purple characters in Figure 1, and 69 unique proteins), suggesting that *Plasmodium* has few transcriptional regulator families with homology to those observed in fungi, plants, and animals (represented by the color-coded strips in Fig. 1). This contrasts with the *A. gambiae* genome—another species in which transcriptional control is poorly understood—where, of the 42 HMMs that match fruit fly sequences, 40 also match mosquito sequences, albeit at different relative abundances (Fig. 1).


14:1548–1554 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

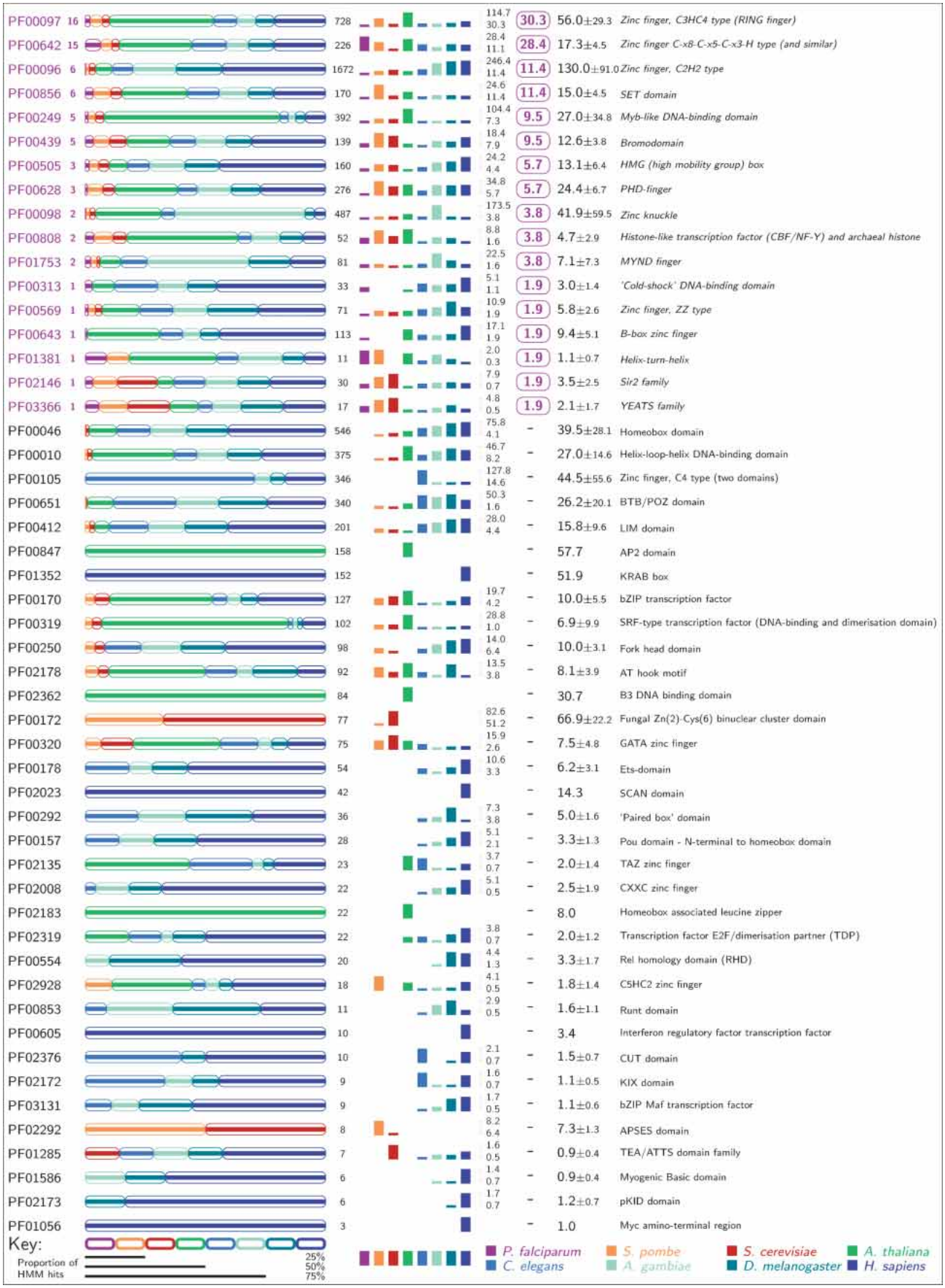


**Figure 1** (Legend on next page)

**Table 1. HMMER Search of Eight Eukaryotic Genomes Using HMMs Linked to TRANSFAC**

| Species | Size[a] | Matches | % Genome |
|---|---|---|---|
| *P. falciparum* | 5273 | 69 | 1.3 |
| *S. cerevisiae* | 6292 | 249 | 4.0 |
| *S. pombe* | 4882 | 205 | 4.2 |
| *C. elegans* | 19,099 | 936 | 4.9 |
| *D. melanogaster* | 13,707 | 865 | 6.3 |
| *A. thaliana* | 27,387 | 1807 | 6.6 |
| *H. sapiens* | 29,303 | 1987 | 6.8 |
| *A. gambiae* | 15,101 | 1045 | 6.9 |

Columns: Size, no. protein sequence entries for a genome; Matches, no. sequences matching the 51 HMMs; % Genome, percentage of genome matching the HMMs.
[a]Total no. sequences searched: 121,044

The proportion of malaria sequences, of the eight eukaryotic genomes searched, that matched a particular HMM usually represented the lowest number of sequences matched (Fig. 1). Hence, only 1.3% of the malaria genome matched the HMMs, compared with an average of 5.7% for the other seven genomes (Table 1). This low level of abundance of known transcriptional regulators in the *P. falciparum* genome becomes more pronounced when the number of matches are normalized against genome size, represented by the number of protein sequence entries (see bar graphs in Fig. 1).

However, of the four HMMs that have the highest number of matches to sequences in the *P. falciparum* genome (the top four entries in Fig. 1), the sequences that match three of them (**PF00097**, **PF00096**, **PF00856**) have the lowest relative abundance compared with the other genomes. This is most striking with the matches to the HMM of the C2H2-type zinc finger (**PF00096**, Fig. 1); on average, in the crown eukaryote group genomes, of every 10,000 genes, 130 contain this type of motif. This contrasts with *P. falciparum* genome, in which only 11 genes per 10,000 encode a C2H2-type zinc finger. The HMM of the CCCH-type zinc finger is an exception to this pattern (**PF00642**); this has an average of 17 genes per 10,000 encoding this domain in the crown eukaryote group genomes, whereas in *P. falciparum*, it exhibits the highest relative abundance with nearly double the number of genes (28 vs. 17 on average).

No examples of highly amplified gene families encoding transcriptional regulators were observed in the malarial genome, examples in the other genomes being the Myb family in plants (Riechmann et al. 2000), the C4-type zinc finger in worms (The *C. elegans* Sequencing Consortium 1998), and the zinc knuckle family in mosquitoes (Zdobnov et al. 2002; Fig. 1). The absence of gene duplications is not confined to transcriptional regulators, but also extends to the rest of the genome, notable exceptions being the *var* and *rif* gene families containing 59 and 149, respectively, highly polymorphic members (Gardner et al. 2002). Thus, the HMM searches indicate that large, paralogous protein families encoding transcription factors, commonly observed in crown group eukaryotes, are virtually absent from *P. falciparum*.

To identify proteins associated with different aspects of the *P. falciparum* transcriptional process from those covered by the HMMs, its genome and those of the seven crown eukaryote group species were queried with 17,054 TAPs. The TAPs and their corresponding homologs were clustered on the basis of sequence similarity using the TRIBE-MCL algorithm (Enright et al. 2002). After a process of manual annotation by inspection of SWISS-PROT function annotation records, 84 clusters that contained at least one *P. falciparum* gene and at least one TAP were identified, and these comprised 95 *P. falciparum* genes (Fig. 2). Of these genes, 87 were uniquely detected by this procedure, with eight also identified by the HMM searches. Hence, 61 of the 69 sequences were only identified by the HMM searches, resulting in a total of 156 malarial TAPs.

Querying the Gene Ontology (GO; Ashburner et al. 2000) annotations in PlasmoDB (Kissinger et al. 2002) with the word 'transcription' identified 45 sequences involved with transcriptional control, all of which were identifed by the above HMM searches and TRIBE-MCL clustering. Thus, this combined strategy using two complementary sequence-searching methods has facilitated the retrieval of proteins from two databases with different contents and annotation levels, and involved in all aspects of the transcriptional process. The presence of 156 genes in *P. falciparum* with homology to TAPs suggests only 3% of the proteins encoded by the malarial genome participate in the transcriptional process, compared with an average of 10% in the crown eukaryote group genomes (Coulson and Ouzounis 2003).

The HMM searches (Table 1) showed that there are relatively few proteins in malaria with homology to known transcriptional activators. To investigate whether the *P. falciparum* basal transcriptional apparatus exhibits a similar degree of divergence, TRIBE-MCL protein families containing the subunits of RNA polymerase II and the preinitiation complex were identified (Fig. 2). Homologs of all 12 subunits of RNA polymerase II, TATA-box binding protein (TBP), and TFIIB were observed. Consistent with the phylogenetic distribution of basal transcription factors across crown eukaryote group kingdoms (Coulson and Ouzounis 2003), homologs of TFIIE$\alpha$ and the TFIIH components XPB, XPD, p52, p44, p34, and MAT1 were also found. No homolog of cyclin H, a component of the TFIIH complex, was identified. However, a sequence homologous to human cyclin K, which is associated with RNA polymerase II (Edwards et al. 1998), is present in the *P. falciparum* (**PF13_0022**, Fig. 2). These results imply that the basic mechanism of transcription initiation in malaria is highly similar to that observed in the crown eukaryote group organisms.

The clustering/annotation procedure identified *Plasmodium* homologs of proteins present in two complexes, which in mammalian and yeast cells, regulate transcription from various promoters (Fig. 2), (1) three sequences having homology to the subunits that comprise the heterotrimeric transcription factor CCAAT box-binding factor (CBF or NF-Y; Maity and de Crombrugghe 1998), and (2), sequences with significant homology to the *S. cerevisiae* proteins CCR4, CAF1, and five NOT proteins

**Figure 1** Phylogenetic distribution of genes matching TRANSFAC-linked HMMs. The first column shows the Pfam accession numbers of the HMMs used in the genome searches (Table 1). If the HMM matches any malaria sequences, it is shown in purple, along with the number of *P. falciparum* sequences matching it (also in purple). The length of the strips in the next column (color-coded by species, see the key), indicates the proportion of sequences in a particular species, of the eight genomes that are matched by the HMM. The total number of sequences matching the HMM is indicated on the *right* of these strips. The bar graphs show the number of sequences that match the HMM, normalized by genome size and expressed as number of sequences per 10,000 genes. The two numbers *above* each other, and immediately to the *right* of the bar graphs, show the maximum and minimum normalized number of matches observed when the genomes were searched with the particular HMM. In the following column, if an HMM matches *P. falciparum* sequences, the normalized number of matching malaria sequences is displayed in the purple rounded box. HMMs matching only crown eukaryote group sequences are indicated by dashes. The penultimate column shows the average, normalized number of matches to the HMM in the seven crown eukaryote group genomes, along with the standard deviation. The final column shows the Pfam description of the HMM — this is in italics if the HMM matches malarial sequences.

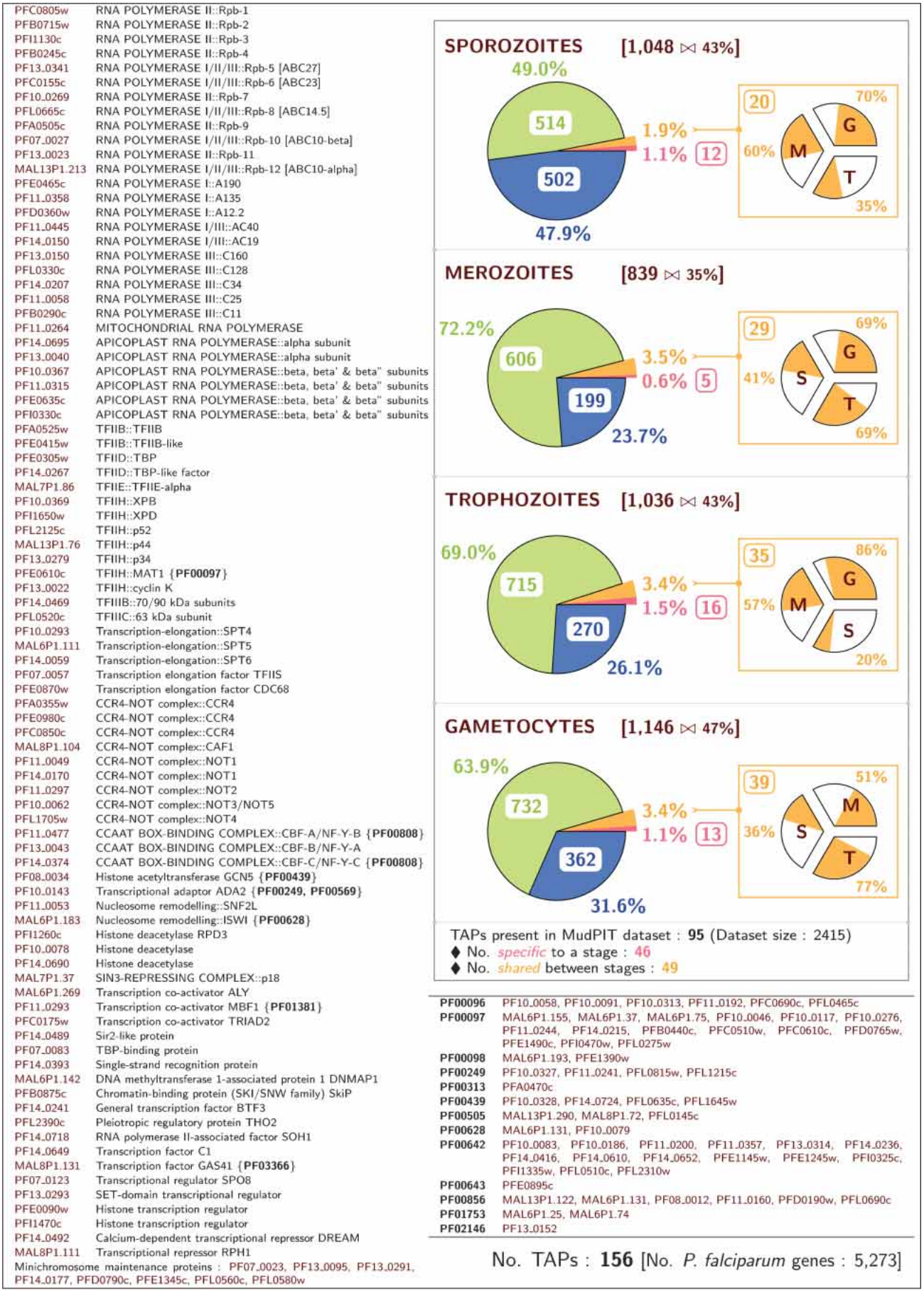


**Figure 2** (Legend on next page)

(NOT1–NOT5), which form the core of the large, CCR4-NOT transcriptional coactivator complexes (Bai et al. 1999). Sequences with homology to the histone acetylase (HAT) GCN5 and the adaptor ADA2, both components of yeast complexes that acetylate histones H3 and H2B in nucleosomes (Grant et al. 1997), are also present in *P. falciparum* (Fig. 2). Furthermore, three genes were found to encode histone deacetylases (HDACs), one of which appears to be a homolog of the HDAC RPD3. Interestingly, a malarial sequence similar to the human protein SAP18, which is a member of the SIN3-RPD3 complex associated with transcriptional silencing (Kadosh and Struhl 1998), was also identified. Furthermore, two proteins homologous to the catalytic subunits of chromatin remodeling complexes (Narlikar et al. 2002; SNF2L, Ji and Arnot 1997; ISWI, Fig. 2) are present in the malarial genome. These similarities in the transcriptional process between malaria and model eukaryotes are further enforced with the existence of a *Plasmodium* homolog of the transcriptional coactivator multiprotein bridging factor 1 (MBF1, Fig. 2), whose function is to facilate communication between regulatory factors and the basal transcription machinery (Takemaru et al. 1998).

High-throughput proteomics identified 2415 proteins that constitute the proteomes of the sporozoite, merozoite, trophozoite, and gametocyte stages of *P. faciparum* (Florens et al. 2002). A total of 95 of the 156 malarial TAPs are observed in these proteomes (Fig. 2, top, right). About half of the malarial TAPs are specific to one of the four stages, with sporozoite TAPs showing the highest level of stage specificity (12 of 32, corresponding to 38%) and the merozoites showing the lowest (5 of 34, corresponding to 15%). A similar pattern is observed with the non-TAP sequences in the proteomes (Fig. 2; Florens et al. 2002). The number of non-stage-specific TAPs ranges from 20 to 39 (Fig. 2, orange boxes), with sporozoites containing the lowest proportion of such proteins; the level of these insect stage TAPs ranges from 20% to 41% in the blood forms (or an average of 32%). In contrast, the proportion range of non-stage-specific TAPs is 51%–60% for merozoites, 35%–77% for trophozoites, and 69%–86% for gametocytes (or an average of 64% for mammalian form TAPs). The largest proportion of non-stage-specific TAPs expressed in trophozoites and gametocytes (86% and 77%, respectively), are present in both of these intracellular erythrocytic stages, that is, 30 TAPs (of 35 and 39, respectively) are common to these forms. These data show TAP expression exhibits a high degree of stage specificity, although a greater similarity in expression patterns is observed in stages residing in identical environments. Hence, the variation in TAP levels across the life cycle is consistent with them functioning in controlling gene expression.

## DISCUSSION

The HMM searches (Fig. 1), combined with the clustering and annotation procedure (Fig. 2), indicates that the genome of *P. falciparum* encodes about a third of the number of proteins associated with the transcriptional process than do the genomes of free-living eukaryotes, even though enzymes are found in equivalent abundances in malaria and yeast (Gardner et al. 2002). The CCCH-type zinc finger functions in regulating mRNA stability and localization, and translation (Lai et al. 2000), and is most prevalent in the *Plasmodium* genome when its abundance within a genome is normalized against the size of the genome (**PF00642**, Fig. 1). The CCR4-NOT coactivator possesses cytoplasmic mRNA deadenylase activity (Tucker et al. 2001), and sequences homologous to the core components of this complex, including homologs of CCR4—which serves as the catalytic subunit (Swanson et al. 2003), are present in the parasite (Fig. 2). Deadenylation modifies the rates of translation initiation and mRNA decay, and as the *P. falciparum* genome encodes nearly twice the number of proteins containing CCCH zinc fingers compared with the average number observed in the crown eukaryote group genomes (when genome size is accounted for), this provides evidence for the proposal that protein levels during the malaria life cycle are controlled through posttranscriptional mechanisms (Wirth 2002).

The apparent dominance of post-transcriptional over transcriptional mechanisms as a means of controlling gene expression is not restricted to malaria; no developmental regulation of RNA polymerase II transcription has been reported in trypanosomatids (Clayton 2002). Even though the environments in which *Leishmania*, *Plasmodium*, and *Trypanosoma* reside are highly diverse from each other, all their life cycles progress in a highly "deterministic" manner. Additionally, the range of environments encountered by these organisms is likely to be far more limited than those encountered by crown group eukaryotes. Transcriptional control may enable more elaborate responses in crown group eukaryotes to these wider variety of stimuli; however, parasitic eukaryotes may not be exposed to such stimuli and, therefore, do not require the same level of sophistication in their gene control mechanisms. Furthermore, posttranscriptional mechanisms may elicit more rapid changes in patterns of gene expression; the environment of the parasite can alter very quickly, for example, when it enters the host while the vector takes a bloodmeal. One consequence of posttranscriptional control being the major feature controlling gene expression would be relatively little constraint on nucleotide composition in introns and intergenic regions. Hence, there would be few sites in the DNA that bind regulatory proteins, and this could allow, as in *P. falciparum*, the overall (A+T) composition to rise to ~90% in such regions (Gardner et al. 2002).

*Plasmodium* sequences encoding all 12 subunits of RNA polymerase II, TFIIB, TFIIEα, and seven of the eight conserved polypeptides of TFIIH were identified. This high level of conservation is also observed with proteins functioning in other aspects of the transcriptional process; sequences homologous to the general transcription-elongation factors SPT4, SPT5, SPT6 (Winston 2001), and TFIIS (Maniatis and Reed 2002) were also observed. The presence of these components of the general transcription machinery in the parasite is expected from previous analyses (Coulson and Ouzounis 2003), but strikingly, the only component of TFIID found to be homologous to malarial sequences is

**Figure 2** Malarial TAP homologs and their expression patterns. The *left* column lists the 95 *P. falciparum* sequences identified by the TRIBE-MCL clustering and annotation procedure. Sequence identifiers are shown in brown. Next to each sequence identifier is shown a functional annotation of the protein. If the annotation contains a double colon (::), then the complex of which the sequence is a component is listed to the *left* of the double colon. Annotations containing {PFnnnnn} indicate the Pfam accession number(s) of the HMM(s) that match(es) the sequence. (*Bottom*) The *right* column (between the two parallel lines) shows the Pfam accession number of a HMM and the identifiers of the malarial sequences that match the HMM. (*Top*) The right-hand column represents as pie charts, the proportion of the malarial TAPs expressed in the sporozoite (S), merozoite (M), trophozoite (T), and gametocyte (G) stages. The expression profiles were obtained from the multidimensional protein identification technology (MudPIT) dataset. The blue and green sectors of the piecharts indicate the number of proteins not associated with transcription (also shown as percentage of total number expressed in the stage), whose expression is either specific to or unrestricted to a stage, respectively. The number of TAPs that are only expressed in the stage is indicated in the red box, and the number in the rounded orange box indicates the number that are also expressed in other stages. The percentage of these TAPs, of the total number of proteins expressed in the stage, is also indicated. In the square orange box, each of the three sectors of the pie chart shows the percentage of the nonstage-specific TAPs expressed in the other three stages.

TBP. This contrasts greatly with crown group eukaryotes, in which TFIID exhibits strong conservation between the species, even though transcription cofactor complexes are diversified extensively (Levine and Tijan 2003). No subunits of TFIIA, which along with TAFs, exhibit high promoter selectively (Aoyagi and Wassarman 2000; Green 2000), were detected. This again suggests that gene-specific transcriptional activation plays less of a role in controlling gene expression in *Plasmodium* than it does in free-living eukaryotes. However, TAP family taxon specificity appears to correlate with evolutionary distance and not cellular complexity (Coulson and Ouzounis 2003), contrasting greatly with other fundamental cellular processes such as metabolism; half of *Escherichia coli* metabolic enzymes have homologs in all domains of life (Peregrin-Alvarez et al. 2003). Thus, the parasite could contain equivalent abundances of transcriptional activators to those in crown group eukaryotes, but these are highly diverged from their functional analogs in free-living eukaryotes.

Genes encoding the catalytic subunits of enzyme complexes that remodel nucleosomes (ISWI and SNF2L) and covalently modify histones (HATs and HDACs) are present in *P. falciparum*. In conjunction with the presence of parasite sequences homologous to subunits that comprise general transcription factors, this suggests that the basic mechanisms activating and repressing genes in malaria are similar to those observed in crown eukaryote group species. These similarities in transcriptional control are emphasized by studies showing that sequences regulating reporter gene expression exhibit, by virtue of assembly into chromatin, appropriate patterns of silencing and temporal transcription (Deitsch et al. 2001; Horrocks et al. 2002). Furthermore, it appears that promoter and terminator sequences immediately adjacent to malaria genes contain the necessary information for the correct timing and expression level of a gene (Horrocks et al. 1998). This is similar to the organization of transcription units typically found in other unicellular eukaryotes, where upstream activator sequences and silencer elements are spaced within ~100–200 bp of the core promoter (Levine and Tijan 2003).

The presence of malarial sequences, whose function is to alter chromatin structure, suggests a potential mechanism of gene control in *P. falciparum*; the chromatin environment defines specific regions of the chromosomes as either transcriptionally repressed (i.e., the genes are silenced) or activate/accessible and facilitating constituitive (unregulated) transcription. Interchangeability between these two states could be accomplished through histone modifications, with the combination of modifications (a "histone code"; Jenuwein and Allis 2001) reflecting the life-cycle stage of the parasite, and so enabling stage-specific transcriptional repression or activation. There appears to be highly coordinated expresssion in *P. falciparum* of genes involved in a common process, and chromosomal clusters of coexpressed proteins are observed (Florens et al. 2002). Additionally, the highest proportion of apicoplast-targeted genes are found on chromosome 5 (Hall et al. 2002), and this uneven distribution involves nonorthologous genes. These types of gene organization are consistent with the existence of a mode of developmental gene regulation in which the covalent modification of histones primarily determine gene activity. Molecular evidence indicates that erythrocytic merozoites are not identical to those derived from the sporozoite invasion of hepatocytes (Preiser et al. 2002). As a response to residing in different cell types, histone modifications present in hepatic merozoites may be dissimilar to those in erythrocytic merozoites, and the resulting differences in chromatin structure leading to subtle differences in patterns of gene activation and silencing.

# METHODS

## Extraction of TAP Reference Set and Pfam Database Entries

The extraction of the TAP reference set was performed by retrieval of database entries that contained 'transcription' in their keyword or gene description fields, as previously described (Coulson and Ouzounis 2003). Sequences present in both the TRANSFAC (Wingender et al. 1996) and the SWISS-PROT (Bairoch and Apweiler 2000) databases were linked to their corresponding Protein Families (Pfam) Database (Bateman et al. 2002) annotations by the Sequence Retrieval System (SRS), version 7.0 and the Icarus language (Etzold and Argos 1993). This procedure obtained 51 profile-hidden Markov models (HMMs) of protein domains known to be involved with eukaryotic transcriptional control. The HMMs were stored, along with the genome and TAP sequences, in a MySQL relational database system.

## HMM Searches

*P. falciparum* sequences were filtered for compositional bias using CAST (Promponas et al. 2000), with a threshold of 20, before being queried. Sequences were then searched with the HMMs using HMMER version 2.2 (Eddy 1998), and any that matched an HMM with a score greater than the lowest score for sequences included in the Pfam full alignment of the family were considered as hits. A total of 71 hits were observed with the filtered *P. falciparum* sequences and 89 for unfiltered sequences.

## Sequence Comparison and Clustering

Sequence comparisons were performed using BLAST version 2.0 (Altschul et al. 1997) and sequence similarities with an *E*-value $\leq 10^{-6}$ were considered as significant. Eukaryotic homologs and TAPs from the reference set were clustered using TRIBE-MCL (Enright et al. 2002) at an inflation value of 2.0, as described previously (Coulson and Ouzounis 2003). This parameter value was chosen to minimize cluster granularity and ensure maximum coverage of the corresponding protein families.

All results are available at: http://www.ebi.ac.uk/research/cgg/projects/transcription/plasmodium.

# ACKNOWLEDGMENTS

R.M.R.C. and C.A.O. thank the Medical Research Council for supporting this work through a Special Training Fellowship in Bioinformatics to R.M.R.C.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

# REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aoyagi, N. and Wassarman, D.A. 2000. Genes encoding *Drosophila melanogaster* RNA polymerase II general transcription factors: Diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation. *J. Cell. Biol.* **150:** F45–F50.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bai, Y., Salvadore, C., Chiang, Y.C., Collart, M.A., Liu, H.Y., and Denis, C.L. 1999. The CCR4 and CAF1 proteins of the CCR4-NOT complex are physically and functionally separated from NOT2, NOT4, and NOT5. *Mol. Cell. Biol.* **19:** 6642–6651.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.
Bannister, L. and Mitchell, G. 2003. The ins, outs and roundabouts of malaria. *Trends Parasitol.* **19:** 209–213.
Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.
The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.
Clayton, C.E. 2002. Life without transcriptional control? From fly to man and back again. *EMBO J.* **21:** 1881–1888.
Coulson, R.M. and Ouzounis, C.A. 2003. The phylogenetic diversity of eukaryotic transcription. *Nucleic Acids Res.* **31:** 653–660.
Coulson, R.M., Enright, A.J., and Ouzounis, C.A. 2001. Transcription-associated protein families are primarily taxon-specific. *Bioinformatics* **17:** 95–97.
Dechering, K.J., Kaan, A.M., Mbacham, W., Wirth, D.F., Eling, W., Konings, R.N., and Stunnenberg, H.G. 1999. Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell. Biol.* **19:** 967–978.
Deitsch, K.W., Calderwood, M.S., and Wellems, T.E. 2001. Malaria. Cooperative silencing elements in var genes. *Nature* **412:** 875–876.
Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Edwards, M.C., Wong, C., and Elledge, S.J. 1998. Human cyclin K, a novel RNA polymerase II-associated cyclin possessing both carboxy-terminal domain kinase and Cdk-activating kinase activity. *Mol. Cell. Biol.* **18:** 4291–4300.
Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30:** 1575–1584.
Etzold, T. and Argos, P. 1993. SRS—An indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.* **9:** 49–57.
Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., et al. 2002. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419:** 520–526.
Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498–511.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.
Grant, P.A., Duggan, L., Cote, J., Roberts, S.M., Brownell, J.E., Candau, R., Ohba, R., Owen-Hughes, T., Allis, C.D., Winston, F., et al. 1997. Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: Characterization of an Ada complex and the SAGA (Spt/Ada) complex. *Genes & Dev.* **11:** 1640–1650.
Green, M.R. 2000. TBP-associated factors (TAFIIs): Multiple, selective transcriptional mediators in common complexes. *Trends Biochem. Sci.* **25:** 59–63.
Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., et al. 2002. Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419:** 527–531.
Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129–149.
Horrocks, P., Dechering, K., and Lanzer, M. 1998. Control of gene expression in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **95:** 171–181.
Horrocks, P., Pinches, R., Kriek, N., and Newbold, C. 2002. Stage-specific promoter activity from stably maintained episomes in *Plasmodium falciparum*. *Int. J. Parasitol.* **32:** 1203–1206.
Jenuwein, T. and Allis, C.D. 2001. Translating the histone code. *Science* **293:** 1074–1080.
Ji, D.D. and Arnot, D.E. 1997. A *Plasmodium falciparum* homologue of the ATPase subunit of a multi-protein complex involved in chromatin remodelling for transcription. *Mol. Biochem. Parasitol.* **88:** 151–162.
Kadosh, D. and Struhl, K. 1998. Targeted recruitment of the Sin3-Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin in vivo. *Mol. Cell. Biol.* **18:** 5121–5127.
Kissinger, J.C., Brunk, B.P., Crabtree, J., Fraunholz, M.J., Gajria, B., Milgram, A.J., Pearson, D.S., Schug, J., Bahl, A., Diskin, S.J., et al. 2002. The *Plasmodium* genome database. *Nature* **419:** 490–492.
Kyrpides, N.C. and Ouzounis, C.A. 1999. Transcription in archaea. *Proc. Natl. Acad. Sci.* **96:** 8545–8550.
Lai, W.S., Carballo, E., Thorn, J.M., Kennington, E.A., and Blackshear, P.J. 2000. Interactions of CCCH zinc finger proteins with mRNA. Binding of tristetraprolin-related zinc finger proteins to Au-rich elements and destabilization of mRNA. *J. Biol. Chem.* **275:** 17827–17837.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G., et al. 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419:** 537–542.
Levine, M. and Tijan, R. 2003. Transcription regulation and animal diversity. *Nature* **424:** 147–151.
Maity, S.N. and de Crombrugghe, B. 1998. Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends Biochem. Sci.* **23:** 174–178.
Maniatis, T. and Reed, R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506.
Narlikar, G.J., Fan, H.Y., and Kingston, R.E. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108:** 475–487.
Peregrin-Alvarez, J.M., Tsoka, S., and Ouzounis, C.A. 2003. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13:** 422–427.
Preiser, P.R., Khan, S., Costa, F.T., Jarra, W., Belnoue, E., Ogun, S., Holder, A.A., Voza, T., Landau, I., Snounou, G., et al. 2002. Stage-specific transcription of distinct repertoires of a multigene family during *Plasmodium* life cycle. *Science* **295:** 342–345.
Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., and Ouzounis, C.A. 2000. CAST: An iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics* **16:** 915–922.
Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., et al. 2000. *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290:** 2105–2110.
Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J., and Lanzer, M. 1998. Antigenic variation in malaria: In situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.* **17:** 5418–5426.
Struhl, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98:** 1–4.
Swanson, M.J., Qiu, H., Sumibcay, L., Krueger, A., Kim, S.J., Natarajan, K., Yoon, S., and Hinnebusch, A.G. 2003. A multiplicity of coactivators is required by Gcn4p at individual promoters in vivo. *Mol. Cell. Biol.* **23:** 2800–2820.
Takemaru, K., Harashima, S., Ueda, H., and Hirose, S. 1998. Yeast coactivator MBF1 mediates GCN4-dependent transcriptional activation. *Mol. Cell. Biol.* **18:** 4971–4976.
Tucker, M., Valencia-Sanchez, M.A., Staples, R.R., Chen, J., Denis, C.L., and Parker, R. 2001. The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in *Saccharomyces cerevisiae*. *Cell* **104:** 377–386.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Waters, A.P. 1994. The ribosomal RNA genes of *Plasmodium*. *Adv. Parasitol.* **34:** 33–79.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24:** 238–241.
Winston, F. 2001. Control of eukaryotic transcription elongation. *Genome Biol.* **2:** reviews1006.1001–1006.1003.
Wirth, D.F. 2002. Biological revelations. *Nature* **419:** 495–496.
Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415:** 871–880.
Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., et al. 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298:** 149–159.

*Received November 28, 2003; accepted in revised form April 28, 2004.*