Update TRENDS in Genetics Vol.xx No.xx Month2005

Genome Analysis

Combinatorial gene regulation in *Plasmodium falciparum*

Vera van Noort and Martijn A. Huynen

Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Toernooiveld 1, 3525 ED Nijmegen, The Netherlands

The malaria parasite *Plasmodium falciparum* has a complicated life cycle with large variations in its gene expression pattern, but it contains relatively few specific transcriptional regulators. To elucidate this paradox, we identified regulatory sequences, using an approach that integrates the sequence conservation among species and the correlation in mRNA expression within a species. Our analysis identified several DNA sequence motifs that are associated with mRNA expression, two of which were previously determined experimentally. We found more putative regulatory sequences per gene in *P. falciparum* than in other eukaryotes, such as yeast. We propose that *Plasmodium* uses the few regulatory proteins it has in a combinatorial approach for gene regulation, explaining the relative paucity in regulatory proteins.

Introduction

Half of the population of the world lives in areas where malaria is endemic, causing the death of up to three million people annually. The most lethal form of human malaria is caused by infection with the protozoan parasite *Plasmodium falciparum*. The genome sequence of this eukaryotic organism in addition to mRNA and protein expression data are publicly available; however, the gene-regulatory processes governing the development of the parasite are poorly understood. Proteomics [1,2] and mRNA expression data [3,4] show that *P. falciparum* has major variations in gene expression levels throughout its life cycle. Furthermore, transcription levels are influenced by environmental factors such as temperature and glucose concentration [5,6]. However, the *Plasmodium* genome seems to encode relatively few proteins that are homologous to transcription factors found in other eukaryotes; these transcription factors are expected to contribute to gene-specific transcriptional regulation [7]. How the parasite manages to control the timing of gene expression correctly taking into account the requirements of the cell remains elusive. It has been proposed that histone-modifications or post-transcriptional mechanisms have a larger effect on gene expression than transcriptional regulation in *Plasmodium*. Recently, six genes were shown to contain sequences that might be implicated in translational repression [8].

The target sequences of transcription regulators in *P. falciparum* are largely unknown. Two methods exist to detect *cis*-regulatory elements by bioinformatics approaches. The first method determines shared sequence motifs in upstream regions of genes that have similar expression patterns or similar functions [9,10]. These motifs have, in several cases, been shown to be target sites. The second method is 'phylogenetic footprinting', in which conserved sequences among multiple species in non-coding DNA can indicate regulatory sites [11]. *Cis*-regulatory elements are conserved at a significantly greater level than non-functional DNA among genomes that are as distant as human and mouse genomes [12]. The evolutionary divergance between the rodent parasite *Plasmodium y. yoelii* and the human parasite *P. falciparum* is approximately the same as that between human and mouse [13], leading to the expectation that *cis*-regulatory elements will also be conserved between these two *Plasmodium* species. In a preliminary study, the AlignAce program was used to find *cis*-regulatory elements in *P. falciparum* in upstream regions of genes encoding heat shock proteins [14], leading to the identification of the G-box element. Comparison among different *Plasmodium* species revealed that this element is conserved.

The extreme AT-richness of *Plasmodium* intergenic regions makes it difficult to identify putative regulatory elements by either phylogenetic footprinting or over-representation in functionally related genes. Therefore, we integrated the two approaches to identify these elements (i.e. we used both clusters of co-expressed genes in *P. falciparum* and the evolutionary sequence conservation between *P. y. yoelii* and *P. falciparum*). We found 12 putative regulatory motifs. Based on our results, we hypothesize that *P. falciparum* uses a greater number of transcriptional regulatory sites per gene in a combinatorial fashion compared with other eukaryotes species, such as *Saccharomyces cerevisiae*.

Integrating evolutionary conservation with expression correlation

Our method for putative regulatory-element detection looks simultaneously for motifs that correlate with mRNA expression profiles and have evolutionary sequence conservation. An overview of the method is given in Figure 1, details can be found in the supplementary material online. First, we calculated similarity in expression of gene pairs based on two expression data sets [3,4] by multiplying the gene–gene correlations from the individual data sets. Then we

Corresponding author: van Noort, V. (V.vanNoort@cmbi.ru.nl).

www.sciencedirect.com 0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2005.12.002

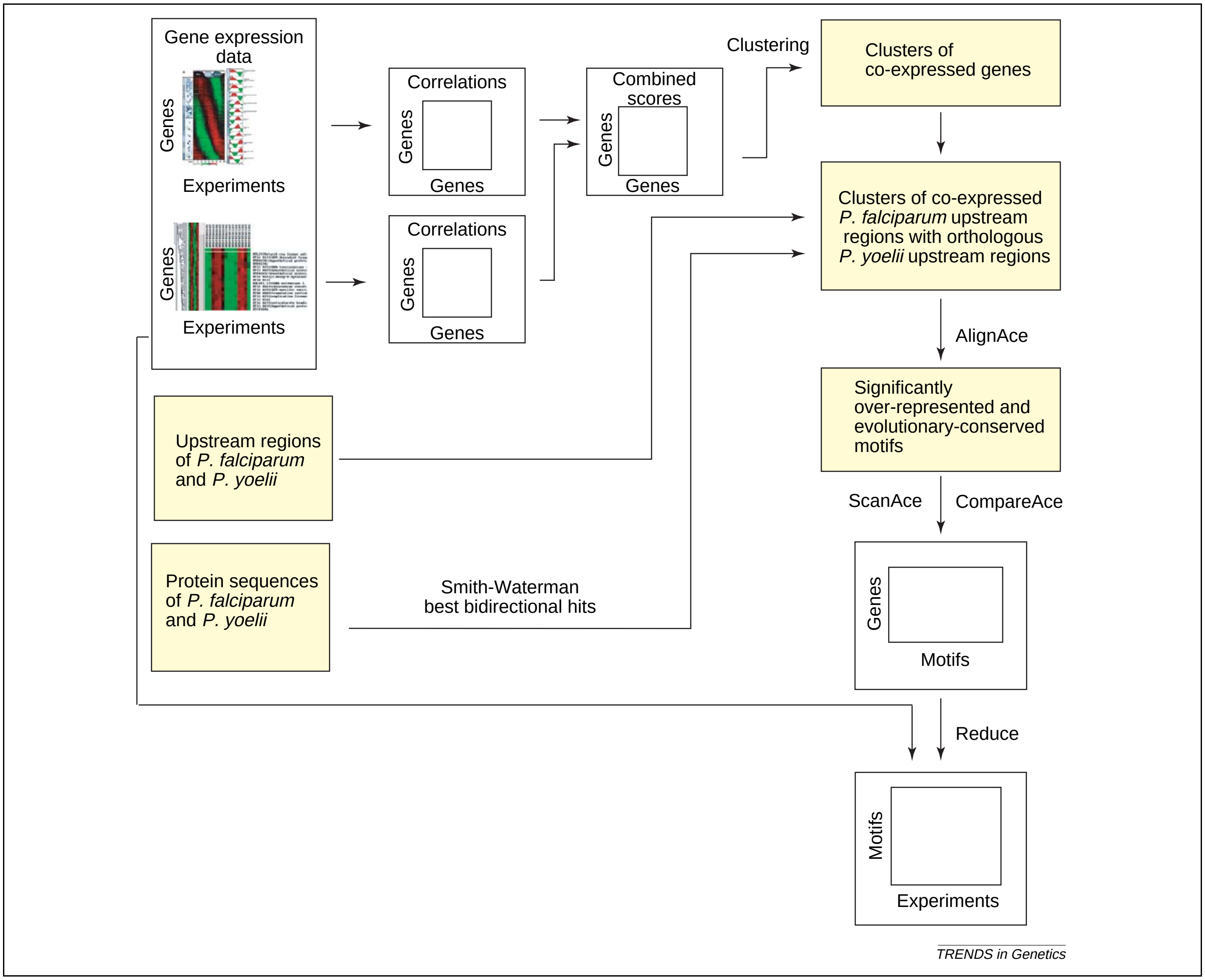
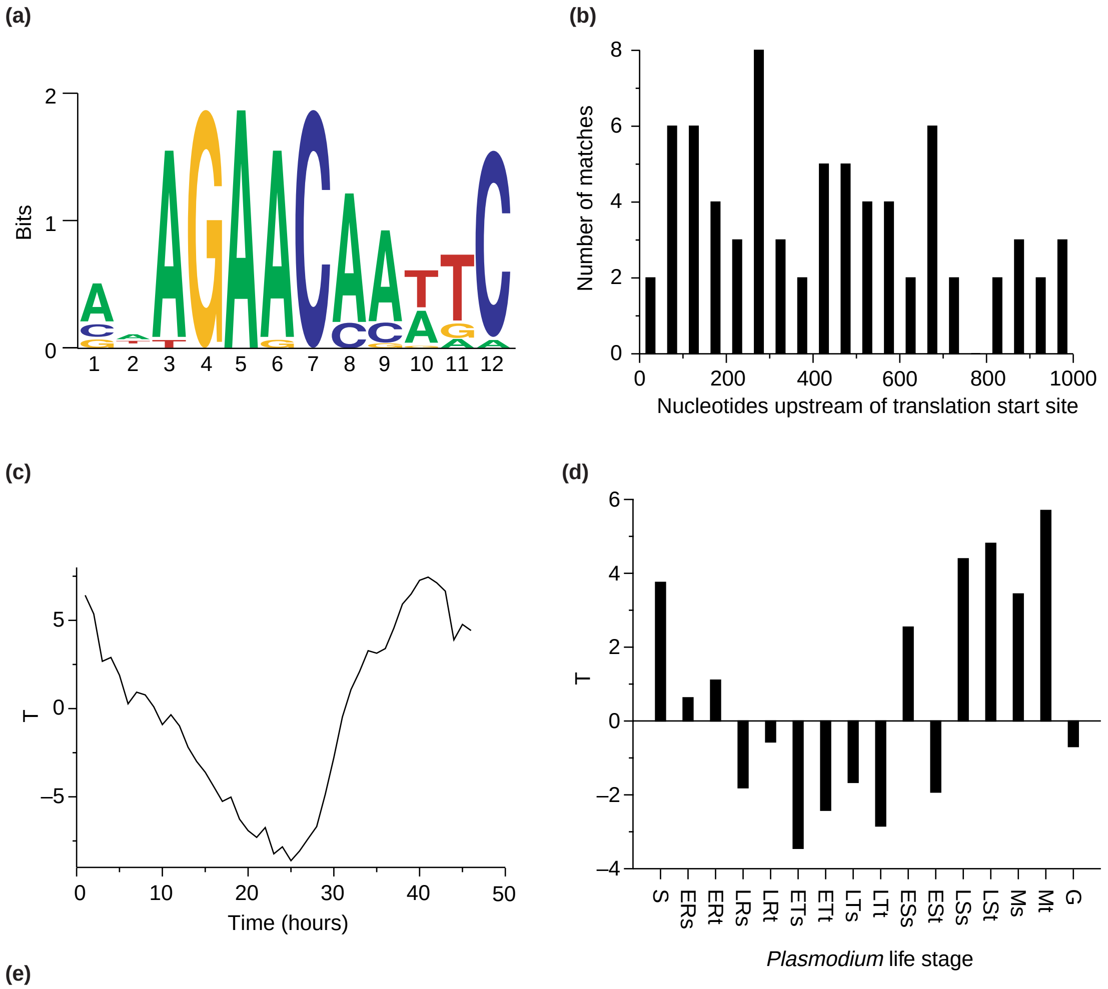

Gene expression data
Genes
Experiments
Genes
Experiments
Correlations
Genes
Genes
Correlations
Genes
Genes
Combined scores
Genes
Genes
Clustering
Clusters of co-expressed genes
Clusters of co-expressed *P. falciparum* upstream regions with orthologous *P. yoelii* upstream regions
AlignAce
Significantly over-represented and evolutionary-conserved motifs
ScanAce
CompareAce
Genes
Motifs
Reduce
Motifs
Experiments
Upstream regions of *P. falciparum* and *P. yoelii*
Protein sequences of *P. falciparum* and *P. yoelii*
Smith-Waterman best bidirectional hits
TRENDS in Genetics

Figure 1. *Cis*-regulatory motif detection. First, the correlations between gene pairs were calculated on the basis of two mRNA expression data sets [3,4]. Genes were then clustered, and the clusters that contained at least 20 genes were considered for further analysis. The squares indicate the calculated data (correlation and scoring matrices). Next, the regions that were 1-kb upstream from the co-expressed *Plasmodium falciparum* genes and their orthologous genes in *Plasmodium yoelii* were selected. The clusters of upstream regions were subsequently used as input for the AlignAce program [9,26], which finds over-represented motifs by a Gibbs-sampling algorithm. The upstream regions of all *P. falciparum* genes were scanned for the presence of over-represented motifs, resulting in a scoring matrix of 5334 genes by 79 motifs. To obtain motifs that correlated with the expression data, we used the multivariate regression approach with forward motif selection used in Reduce [15]. We included motifs until $P<$ 0.01 of the most significant motif. Time courses (T-values) were calculated for all significant motifs.

clustered *Plasmodium* genes based on the combined co-expression scores. Finally, these *P. falciparum* gene clusters were combined with their *P. y. yoelii* orthologs to find conserved motifs in the upstream regions using the AlignAce program. Note that we did not make alignments of upstream regions but instead used the Gibbs sampler from AlignAce [9] to find motifs that correlate simultaneously with expression and with evolutionary conservation.

Correlation of motifs with expression data

Using the different co-expression clusters, we found 79 over-represented motifs. The upstream regions of all *P. falciparum* genes were scanned for presence of the motifs. For each motif cluster, we took the greatest score for each gene resulting in a scoring matrix of 5334 genes by 79 motifs. Among these motifs we expected that some are functional and, therefore, correlate with specific expression patterns, whereas others might just be over-represented in the whole genome or occur because of other biases, specifically in the AT-rich genome of *P. falciparum*. Therefore, we used a second algorithm, Reduce, to calculate the correlation (r) between the motif scores and expression levels [15], to obtain potentially functional motifs. This method ensures that if similar motifs correlate with the same gene expression levels, we will only identify the motif that has the strongest correlation with these gene expression levels. We found 12 putative regulatory motifs [i.e. motifs that significantly correlated ($P<0.01$) with mRNA expression]. Note that out of 79 over-represented motifs, only 12 correlated with mRNA expression. Indeed by just detecting over-representation, we identified motifs that were the result of biases in the genome.



(e)

```
PFB0275w    agtgttaagaAGAGAACAATTCttaatttgct
PFB0665w    aaatttttgtATAGAACCATTCtaggtttgat
PFC1045c    taaaaaaaaaAAAGAACCATGCtaccttaata
PFD0110w    accataaaatATAGAACAATGCcgagaataag
PFD0395c    aaaaaaaaaaAAAGAGCCATGCataattcttt
PF10_0100   tataaaacagAAAGAACAAATCttttttttta
PF10_0145   tatttggtaaAAAGAACAACGCaggaagataa
PF10_0352   aaataaataaAAAGAACACTTCttatttatgt
PF11_0089   aatgaaatttATAGAACACGTCctataatgaa
PFL1170w    aatagcaagcCTAGAACCACGCaaatcaattt
PFL2460w    cctactatatGTAGAACACATCtacatataac
MAL13P1.315 aataatatggAAAGAACACTTCacctaattat
                      * ******* **
```

TRENDS in Genetics

Figure 2. The late-schizont motif. **(a)** Logo of the late-schizont regulatory motif. The height per position represents the information content and the height of the letters indicates the frequency. **(b)** Positions of non-overlapping matches counted in 50-nt bins. Most matches of the late-schizont motif are 250–300 nt upstream of the translation start site (position preference $P<0.05$). **(c)** Time course (T) for the late-schizont motif in the Bozdech data set. **(d)** Time course for the late-schizont motif in the LeRoch data set. **(e)** The alignment of upstream regions of genes expressed in late schizonts that contain the motif. The asterisks indicate conserved positions. The region matching the motif is in capital letters. Abbreviations: ER, early ring; ES, early schizont; ET, early trophozoite; G, Gametocyte; LR, late ring; LS, late schizont; LT, late trophozoite; M, Merozoite; S, sporozoite; s, sorbitol-synchronized parasite; t, temperature-synchronized parasite.

Previously determined sequence motifs

The sequence logos of each of these motifs and correlations with expression data can be found in Table 1 (supplementary material online). There is little experimental data available about expression-related sequence motifs in *P. falciparum*. Two of the motifs we identified had been experimentally determined previously. Motif 5 contains consensus sequence TGTATATATG and is correlated with upregulation in schizonts in mRNA expression data sets and in gametocytes in the LeRoch data set. Motif 5 is similar to and present in the same genes as the sequence TGTAT(G/A)TG, which was found to regulate *var* genes in an experimental study [16]. We also observed a poly(dAdT) repeat (motif 11), which correlated with upregulation in gametocytes, this motif was recently found to regulate calmodulin activity [17].

Two known regulatory sequences were not present in our results. The first is an experimentally determined sequence associated with sexual and early mosquito stages: the recognition site for the PAF-1 transcription factor [18]. The second motif that we did not retrieve is the CCAAT box, although this is to be expected because the *Plasmodium* genome encodes the complete CCAAT-box-binding complex [7]. Additional mRNA expression data will be needed for the identification of these sequence motifs.

Newly discovered motifs

In addition to motif 5, we found two more T(G/A) repeat motifs (motifs 3 and 4) that also correlate with mRNA expression in schizonts and gametocytes. We found other important motifs including oligo(dA)oligo(dT) repeat (motif 1), correlating with gene upregulation in the ring stage and a poly(dG)poly(dA) motif (motif 2), correlating with gene upregulation in the trophozoite stage. Figure 2 shows the late-schizont motif, a newly discovered, putative regulatory motif that correlates with upregulation of 72 genes in late schizonts (Figure 2c,d). The late-schizont motif occurs preferentially ($P<0.05$) between 250 and 300 nucleotides upstream of the translation start site (Figure 2b). Binding of transcriptional regulators in yeast occurs predominantly at ~180 bases upstream of the start codon [19]. Thus, we expect that the late-schizont motif also constitutes a transcription-factor-binding site. Similar to the late-schizont motif, all motifs have a significant position preference relative to the translation start site. The same motifs that are correlated with expression in the Bozdech data set also correlate with expression in the LeRoch data set. Both data sets describe asexual intra-erythrocytic blood stages. The LeRoch data set describes three additional parasite life stages, enabling the discovery of additional motifs.

Functional significance of discovered motifs

In the absence of large amounts of experimental data on transcription-factor-binding sites in *P. falciparum,* we used other data and randomizations to asses the functional significance of the discovered motifs.

First, we examined the robustness of our results by applying other strategies to find *cis*-regulatory motifs. Rather than examining co-expressed genes, we analyzed upstream regions of genes with similar functions (occurring in one pathway as defined by the Kyoto encyclopedia of genes and genomes (KEGG) data base [20]), we combined them with the upstream regions of their orthologs in *P. y. yoelii* and used those as input sets for AlignAce to detect over-represented motifs. These motifs are clustered and correlated with expression patterns. Logos [21] of motifs with the greatest correlations with expression are shown in Table 2 (supplementary material online). We found similar motifs with this approach, confirming the functional relevance of the discovered motifs. A few forms of the T(G/A) repeat (motif 15, 17, 18 and 27), the two different AT-rich repeats [poly(dAdT), motif 28; oligo(dA)oligo(dT) repeat, motif 13] and the poly(dG)poly(dA) motif (motif 14) seem to regulate a functionally coherent set of genes. However, the variation that can be explained by these motifs is less than that explained by motifs in clusters of co-expressed genes (Figure 3a).

Second, we found that simpler methods do not give better or even as good results as our approach. We performed an exhaustive search for oligomers, up to 7 nt in length, in the upstream regions of *P. falciparum* and examined their predictive value. These oligomers gave a lower correlation with the expression data than the over-represented motifs; for example, the top scoring oligomer, GACCGC, only has a maximum r^2 value of 0.0125, whereas our top scoring motif (motif 3) has a score of 0.108.

Third, to examine the value of including the upstream regions from the second species in the study, we repeated the analysis without *P. y. yoelii*, by first determining over-represented motifs in upstream regions of *P. falciparum* and then examining their correlation with gene expression. This resulted in motifs that were not correlated with the expression data (Figure 3a), underscoring the value of combining sequence conservation with co-expression data in determining regulatory elements.

Finally, we verified that the statistical significance of the correlations of motifs with mRNA expression corresponds to a real signal in the upstream regions of genes and not to other biases in intergenic regions of *P. falciparum* by randomizing our data. All expression profiles were randomly reassigned to the genes, effectively detaching upstream regions of genes from the gene expression data. Next, we re-clustered the genes according to their new expression profiles and repeated the motif discovery procedure. This resulted in sequence motifs that had little correlation with the (still randomized) expression profiles (Figure 3a), indicating that in the *Plasmodium* genome there are DNA sequences in the upstream regions of genes that correlate with mRNA expression. Whether these DNA sequences are transcription-factor-binding sites will have to be solved experimentally. It is possible that the sequence elements are related to mRNA stability or chromosome accessibility.

Combinatorial gene regulation

To elucidate the paradox of the small number of regulators encoded in the *Plasmodium* genome, we counted the number of regulatory elements (motif clusters that correlate significantly with expression) per gene for yeast and *Plasmodium* (Figure 3b). Most *Plasmodium* genes have four or five different regulatory elements in their upstream region. This contrasts strongly with the situation in yeast, where most genes are regulated by only one or two regulators (Figure 3b). Chromatin immunoprecipitation (ChIP-on-chip) data from *Plasmodium* are not yet available, therefore, we chose to compare the distribution of regulators per gene with the results of a similar computational method in yeast (Figure 3b). ChIP-on-chip data also show that the vast majority of yeast genes have only one regulator binding to their promoter region [22]. However, in *Plasmodium* it seems that fewer regulators are used in different combinations of five elements per promoter to obtain the same level of diversity in expression profiles. A simple calculation shows that with ten regulatory proteins and five elements per gene,

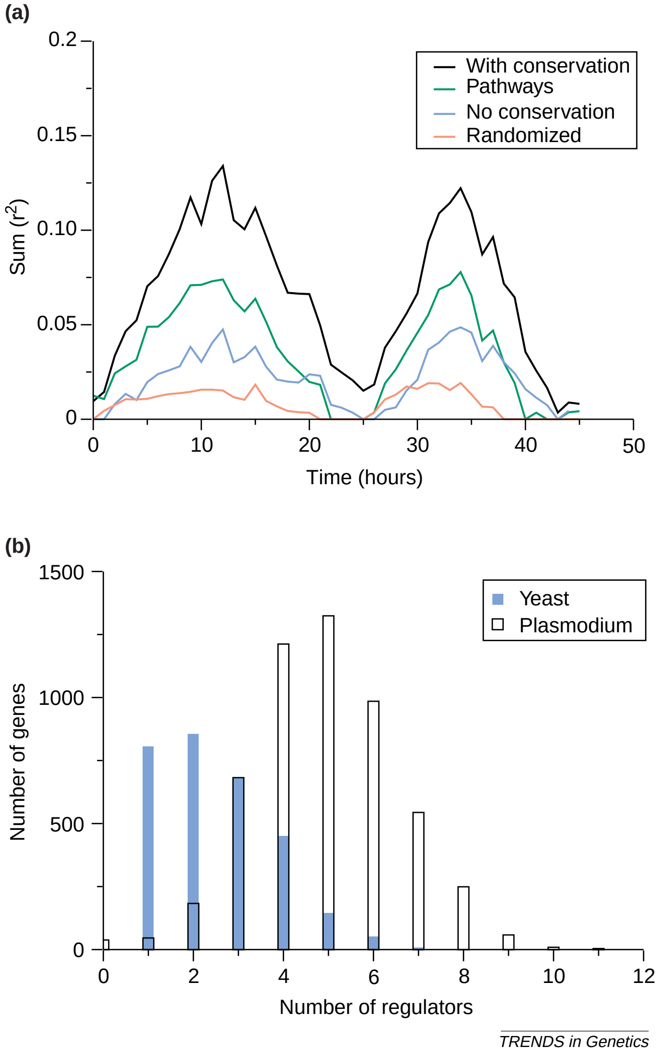


Figure 3. Regulatory motifs in *Plasmodium falciparum.* **(a)** The variation in gene expression that can be explained by motifs. For each experiment (shown on the *x*-axis), we calculated the total amount of variation (r^2) in expression levels in the Bozdech data set that can be explained by motifs identified using four different methods [motifs found with the procedure as depicted in Figure 1 (black); motifs found without conservation (blue); motifs found using KEGG pathways (green); motifs found with upstream regions and expression profiles randomized relative to each other (red)]. The greatest level of variation that can be explained by motif scores was obtained by the regulatory-element-detection method that integrates both mRNA expression and evolutionary sequence conservation. **(b)** The number of regulators per gene. In yeast, the number of unique oligomers correlating with expression was obtained from Ref. [15] and counted in 1000-bp regions upstream of yeast genes that do not overlap with coding regions (blue). For *Plasmodium*, the number of different motifs that correlated significantly with expression was counted in 1000-bp regions upstream of *Plasmodium falciparum* overlapping genes (white).

$\binom{10}{5} = 252$ different combinations can be made. If promoter sequences contain only one regulatory element, then 252 instead of ten regulatory proteins would be needed to obtain the same number of expression profiles. If they contain two elements per gene, then 23 regulatory proteins would be needed.

The most abundant combination of regulatory elements is $1+4+8+10+11$, a combination of elements some of which have opposite effects on gene expression. For example, motifs 1 and 4 have opposite effects in all experiments (Table 1 in the supplementary material online) and 775 genes have a combination of these motifs. This leads to the hypothesis that *Plasmodium* uses combinatorial effects of gene regulators, exploiting the possibilities of the relatively few regulators that it possesses [7]. Instead of using one regulatory protein for each expression profile, different combinations of regulators are employed to obtain a variety of expression profiles. Experimental results support this hypothesis. First, studies of the promoter of *GBP130* showed that this gene was most likely to be regulated by multiple, possibly different nuclear factors [23]. Furthermore, a study of *var* genes showed that silencing occurs through the cooperative action of multiple sequence elements [24]. Finally, a study of the Polδ promoter revealed regions that have both positive and negative effects on gene expression [25].

Concluding remarks

We have identified DNA motifs in the upstream regions of *Plasmodium falciparum* genes that significantly correlate with mRNA expression. To find these motifs it was necessary to integrate phylogenetic footprinting techniques with the over-representation of motifs in co-expressed genes. The results provide an explanation for the paradox between the large variation in *P. falciparum* gene expression and the reported paucity of specific transcription factors. It can be explained by a combinatorial mode of gene regulation, in which every gene is regulated by multiple factors. Our results suggest that this is the general mode of transcriptional regulation in *Plasmodium;* that is combinations of regulatory motifs contribute to overall promoter activity.

Acknowledgements

We thank Harmen Bussemaker for careful reading of the article and Chris Ponting for useful discussions. This work was supported in part by a grant from the Netherlands Organization for Scientific Research (NWO).

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.12.002

References

1 Florens, L. *et al*. (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419, 520–526
2 Lasonder, E. *et al*. (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419, 537–542
3 Bozdech, Z. *et al*. (2003) The transcriptome of the intra-erythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* 1, E5
4 Le Roch, K.G. *et al*. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301, 1503–1508
5 Fang, J. and McCutchan, T.F. (2002) Thermoregulation in a parasite's life cycle. *Nature* 418, 742
6 Fang, J. *et al*. (2004) The effects of glucose concentration on the reciprocal regulation of rRNA promoters in *Plasmodium falciparum*. *J. Biol. Chem.* 279, 720–725
7 Coulson, R.M. *et al*. (2004) Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res.* 14, 1548–1554
8 Hall, N. *et al*. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic and proteomic analyses. *Science* 307, 82–86
9 Roth, F.P. *et al*. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16, 939–945
10 van Helden, J. *et al*. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827–842
11 Cliften, P. *et al*. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301, 71–76
12 Liu, Y. *et al*. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* 14, 451–458
13 Carlton, J.M. *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii* yoelii. *Nature* 419, 512–519
14 Militello, K.T. *et al*. (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* 134, 75–88
15 Bussemaker, H.J. *et al*. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171
16 Calderwood, M.S. *et al*. (2003) *Plasmodium falciparum var* genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. *J. Biol. Chem.* 278, 34125–34132
17 Polson, H.E. and Blackman, M.J. (2005) A role for poly(dA)poly(dT) tracts in directing activity of the *Plasmodium falciparum* calmodulin gene promoter. *Mol. Biochem. Parasitol.* 141, 179–189
18 Dechering, K.J. *et al*. (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
19 Harbison, C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104
20 Ogata, H. *et al*. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34
21 Crooks, G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190
22 Lee, T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
23 Horrocks, P. and Lanzer, M. (1999) Mutational analysis identifies a five base pair *cis*-acting sequence essential for GBP130 promoter activity in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* 99, 77–87
24 Deitsch, K.W. *et al*. (2001) Malaria. Cooperative silencing elements *in var* genes. *Nature* 412, 875–876
25 Porter, M.E. (2002) Positive and negative effects of deletions and mutations within the 5′ flanking sequences of *Plasmodium falciparum* DNA polymerase delta. *Mol. Biochem. Parasitol.* 122, 9–19
26 Hughes, J.D. *et al*. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214