Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum

Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*

Tobias Mourier,[1,2,7] Celine Carret,[3,7] Sue Kyes,[4,7] Zoe Christodoulou,[4] Paul P. Gardner,[5] Daniel C. Jeffares,[3] Robert Pinches,[4] Bart Barrell,[3] Matt Berriman,[3] Sam Griffiths-Jones,[6] Alasdair Ivens,[3] Chris Newbold,[4] and Arnab Pain[3,8]

[1]*Ancient DNA and Evolution Group, Department of Biology, University of Copenhagen, Copenhagen DK-2100, Denmark;* [2]*Department of Experimental Medical Science, Lund University, Lund 22184, Sweden;* [3]*Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom;* [4]*Molecular Parasitology Group, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom;* [5]*Department of Molecular Biology, University of Copenhagen, Copenhagen DK-2200, Denmark;* [6]*Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, United Kingdom*

We undertook a genome-wide search for novel noncoding RNAs (ncRNA) in the malaria parasite *Plasmodium falciparum*. We used the RNAz program to predict structures in the noncoding regions of the *P. falciparum* 3D7 genome that were conserved with at least one of seven other *Plasmodium* spp. genome sequences. By using Northern blot analysis for 76 high-scoring predictions and microarray analysis for the majority of candidates, we have verified the expression of 33 novel ncRNA transcripts including four members of a ncRNA family in the asexual blood stage. These transcripts represent novel structured ncRNAs in *P. falciparum* and are not represented in any RNA databases. We provide supporting evidence for purifying selection acting on the experimentally verified ncRNAs by comparing the nucleotide substitutions in the predicted ncRNA candidate structures in *P. falciparum* with the closely related chimp malaria parasite *P. reichenowi*. The high confirmation rate within a single parasite life cycle stage suggests that many more of the predictions may be expressed in other stages of the organism's life cycle.

[Supplemental material is available online at www.genome.org.]

While the tight differential regulation of gene expression in the human malaria *Plasmodium falciparum* is well documented (Bozdech et al. 2003; Le Roch et al. 2004), analysis of the published genome sequence indicates a paucity of recognizable transcription factors and a lack of components for classical microRNA (miRNA)-mediated gene regulation (Gardner et al. 2002; Aravind et al. 2003; Coulson et al. 2004; Rathjen et al. 2006). This, together with the evidence for the epigenetic control of expression of the virulence-associated *var* gene family, post-transcriptional gene regulation in the sexual stages of rodent malaria parasites, and widespread antisense transcription in the asexual blood stages of *P. falciparum* (Gunasekera et al. 2004; Duraisingh et al. 2005; Hall et al. 2005; Mair et al. 2006; Voss et al. 2006), suggests important roles for additional mechanisms of gene regulation.

RNA is a structural component of complexes controlling a variety of core cellular processes such as translation and transcription (Storz 2002; Storz et al. 2005). A large proportion of the eukaryotic transcriptome is likely to consist of noncoding RNAs (ncRNAs), once regarded primarily as transcriptional noise (The ENCODE Project Consortium 2007; Washietl et al. 2007). Recent studies indicate that a large number of these noncoding structured RNAs play critical roles in regulating gene expression at multiple levels in diverse organisms (Eddy 2001; Washietl et al. 2005a; Mattick and Makunin 2006; Pedersen et al. 2006; Backofen et al. 2007; Prasanth and Spector 2007). In recent years, more than 800 ncRNA genes have been described in mammals, a subset of which are alternatively spliced and show tissue-specific expression or developmental regulation (Cheng et al. 2005; Jongeneel et al. 2005; Washietl et al. 2005a; Pedersen et al. 2006; Ravasi et al. 2006), and a few have been implicated in disease phenotypes or developmental disorders (Szymanski and Barciszewski 2006). More recently, the ENCODE pilot project has identified a large number of non-protein-coding transcripts, with many overlapping protein-coding loci and a large number of mappings to genomic regions that were previously thought to be transcriptionally inactive (The ENCODE Project Consortium 2007). A significant proportion of these noncoding transcripts map to structured ncRNA candidates, predicted in the ENCODE region (Washietl et al. 2007) and thus shedding further light on the widespread occurrence of structured RNAs and their potential roles in shaping the functional landscape of the human genome. However, in the *P. falciparum* genome, only a minimal set of tRNAs and a small set of ribosomal and spliceosomal RNAs have been identified by sequence annotation (Gardner et al. 2002; Upadhyay et al. 2005). An in silico comparison based on the identification of homologous GC rich regions across several *Plasmodium* species sequences identified a total of 18 different potential RNA types present in all, of which nine had homology with known noncoding RNAs (Upadhyay et al. 2005). Transcripts were detected in asexual parasites' RNA for six of the 18 candidates, but only one of these six was a novel RNA. Recently, a more elaborate study, using comparative genomics and RNA analysis, has identified several components of the structured RNAs of known function (i.e., spliceosomal and small nucleolar RNAs, telomerase RNA) and a small number of structured RNAs of unknown function in *P. falciparum* (Chakrabarti et al. 2007).

[7]These authors contributed equally to this work.
[8]Corresponding author.
E-mail ap2@sanger.ac.uk; fax 44-01223-494919.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6836108. Freely available online through the *Genome Research* Open Access option.

Here we undertake a broader, systematic in silico screen to identify novel conserved non-protein-coding RNA structures across multiple *Plasmodium* genomes.

Results

Prediction of structured RNAs

We used two criteria of phylogenetic conservation and RNA structure prediction to identify candidate ncRNAs in the *P. falciparum* genome. Figure 1 shows a flowchart of the prediction and analysis procedure. To identify non-protein-coding conserved regions we compared all nonexonic regions of the *P. falciparum* clone 3D7 reference genome with conserved genomic sequence from seven other *Plasmodium* species: *P. berghei*, *P. chabaudi*, *P. gallinaceum*, *P. knowlesi*, *P. reichenowi*, *P. vivax*, and *P. yoelii* (http://www.sanger.ac.uk/Projects/; http://www.tigr.org) using WUBLASTN (http://blast.wustl.edu). We then employed the comparative RNA gene-finding tool RNAz (Washietl et al. 2005b) on all conserved non-protein-coding alignments, predicting 855 RNA structures. After filtering out 187 predictions—sequences with similarity to protein-coding genes or with low complexity—a total of 668 secondary structures remained that were conserved between 3D7 (*P. falciparum*) and at least one other *Plasmodium* species. Sixty-four RNA predictions overlapped or displayed similarity to known RNA genes such as tRNA, rRNA, and snRNA (subsequently referred to as "known" RNAs), leaving 604 potentially novel RNA structures ("candidate" RNA predictions) (Fig. 1; Supplemental Fig. S1; Supplemental Table S1). Because RNA structures may exert their functions when cotranscribed with mRNAs or transcribed independently (for reviews, see Mignone et al. 2002; Pedersen et al. 2006), our analysis included all conserved RNA structures regardless of their distance from protein-coding genes. A graphical display of the features (RNAz score, GC-content, distance to nearest CDS and size of prediction) of candidate RNA predictions is shown in Supplemental Figure S2.

Some of the candidate sequences were highly similar and could be grouped into 40 clusters of between two and 27 sequences. Often the members within each cluster are located at a similar distance from a member of a specific gene family (most often *rif* and *var* genes). However, it appears that the similar distances are a result of highly conserved sequences flanking the genes (which may stem from duplication events leading to paralogous expansion of these gene families), rather than a feature maintained specifically by selection (Supplemental Table S2).

As expected, the repertoire of previously known ncRNA genes (tRNAs, rRNAs, and snRNAs) is well conserved across all *Plasmodium* species (Fig. 2). By contrast, the novel candidate RNA predictions display much narrower distributions, with the vast majority shared between *P. falciparum* and the closely related *P. reichenowi* and *P. gallinaceum*. Interestingly, a small number of the candidate RNA predictions display a host-associated phylogenetic distribution (Fig. 2), such as those exclusively present in parasites that infect primates. The predicted novel candidate RNAs appear unique to *Plasmodium* species since none of the candidates are identifiable in primary sequence from available sequenced genomes of other related apicomplexan parasites: *Toxoplasma gondii*, *Theileria parva*, *Theileria annulata*, *Cryptosporidium parvum*, and *Eimeria tenella* or from any noncoding RNAs predicted in the Rfam database (http://www.sanger.ac.uk/Software/Rfam/) (Griffiths-Jones et al. 2005).

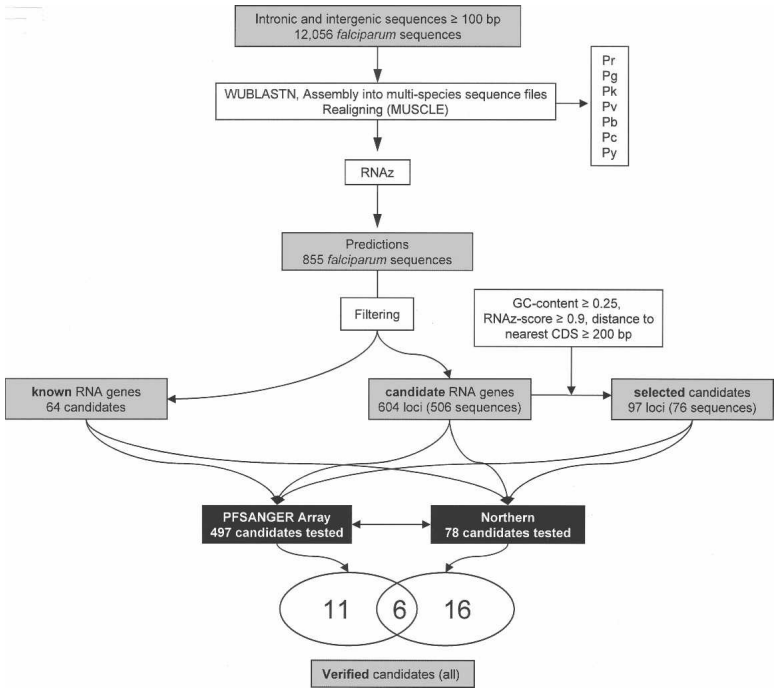
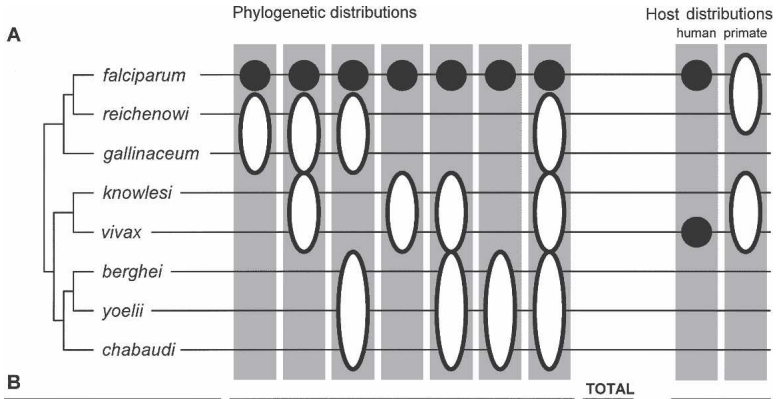


Figure 1. Flowchart of the prediction procedure, resulting in three classes of predictions; non-RNA predictions (non), known RNA predictions (known), and putative novel RNA predictions (candidates). Known RNA genes were identified from genome annotation and similarity to previously identified RNA genes (see in Methods). A set of candidates (selected) was chosen for Northern blotting analysis using the following criteria: (*1*) G+C content of the alignment >25%, (*2*) RNAz score ≥0.9, and (*3*) distance to nearest annotated protein coding gene >200 bases. Analyses and verification experiments are indicated by white boxes with black borders. Microarray indicates microarray analysis using the PFSANGER Affymetrix array; Northern, Northern blotting. (Pr indicates *P. reichenowi*; Pg, *P. gallinaceum*; Pk, *P. knowlesi*; Pv, *P. vivax*; Pb, *P. berghei*; Py, *P. yoelii*; Pc, *P. chabaudi*.)

Assessment of performance

To assess the sensitivity of our computational predictions, we examined how many previously annotated ribosomal, transfer, and spliceosomal RNA genes were predicted by our methods. The annotation of the *P. falciparum* 3D7 genome (version 2002.10.03, available at http://www.PlasmoDB.org/) contained 49 RNA genes with known genomic coordinates. Although this set of RNA genes is incomplete compared with the previously described repertoire (Gardner et al. 2002), these genes were used for our estimation of sensitivity. Of the 49 annotated RNA genes, 41 (83.7%) overlap with our identified regions of sequence similarity between *P. falciparum* and at least one other *Plasmodium* genome, and are thus present in the alignments on which RNA gene prediction is performed. Of these, RNAz correctly predicts 33 of these 41 sequences to be RNA genes, corresponding to ~80%



B

							TOTAL		
Known RNA predictions	1	1	1	0	2	1	59	64	0	0
Candidate RNA predictions	517	29	16	5	5	3	29	604	1	32
Selected candidates	83	1	3	1	1	0	8	97	0	2
Verified candidates	23	0	1	0	1	1	7	33	0	0

Figure 2. Phylogenetic distribution of predicted RNA structures. (*A*) A schematic *Plasmodium* phylogeny, based on the method of Escalante and Ayala (1994), is shown *top left*. Phylogenetic distribution combinations are marked in gray boxes. Black circles denote candidates found in the corresponding species. Ellipses indicate that a candidate is found in at least one of the species covered. Two phylogenetic distributions ("human" and "primate") are functionally denoted by the host of the *Plasmodium* species. (*B*) The table indicates the number in each prediction class belonging to a given distribution.

sensitivity with RNAz in the aligned regions. The overall sensitivity is 67% (33/49).

Ribosomal RNAs are the poorest detected known RNA genes, although there did not seem to be a strong systematic bias as all types of rRNAs were predicted at least once. A list of the 49 annotated RNA genes in *P. falciparum* 3D7 genome is provided in Supplementary Material (Supplemental Table S3).

To estimate the specificity of our predictions, we adopted the approach of Washietl et al. (2005a) and shuffled the alignments (Washietl and Hofacker 2004). These randomized data suggested a false discovery rate of 35.6% (RNAz *P*-value of 0.5) and 23.3% (RNAz *P*-value of 0.9). This rate is higher than that reported for mammalian RNA genes (28.8% and 19.2%, respectively) (Washietl et al. 2005a), as is the number of predictions per kilo base pair of shuffled sequence (mammals [$P > 0.5$]: 0.32 predictions/kbp [$P > 0.9$]: 0.08. *Plasmodium* [$P > 0.5$]:0.66 [$P > 0.9$]: 0.27). These numbers underline the high level of "noise" in the *Plasmodium* data, possibly reflecting either the alignment or prediction difficulties imposed by the extreme A+T composition of *P. falciparum* genome.

Experimental verification

The life cycle of *Plasmodium* is complex but can be broadly divided into life cycle stages that appear in the invertebrate host (i.e., the mosquito stages) and the vertebrate host (i.e., the preerythrocytic and the erythrocytic stages). Clinical symptoms of malaria are manifested during the erythrocytic (blood) stages of the life cycle, and only erythrocytic forms of *Plasmodium falciparum* can be cultured conveniently in the laboratory. We examined the transcription of the ncRNA predictions in two points in the asexual life cycle of the vertebrate host (young ring forms and mature pigmented trophozoite forms) by combining expression data obtained from a genome-tiling-like microarray and Northern blots (Fig. 3). We used the PFSANGER array, a 2.32 million feature 25-mer oligonucleotide Affymetrix array, which has features (probes) representing over 90% of the *P. falciparum* 3D7 genome. Probes for 641 of the 855 initial RNAz predictions (75%), and for 443 of the 604 candidate predictions (73%) were represented on the PFSANGER microarrays (Fig. 3; Supplemental Fig. S1). RNAs from ring and pigmented trophozoite stages were hybridized to the PFSANGER arrays. Applying a log2 cutoff of 1 (see Methods; Supplemental Fig. S3), we verified the expression of 17 novel RNA predictions. Of these, six are specifically expressed in the ring stage and six in the trophozoite stage, and a further five are expressed in both stages (Supplemental Table S4).

Recently, ncRNA predictions in *Saccharomyces cerevisiae* have been compared with array expression data, resulting in expression verification for 16.8% of the RNA predictions (Steigele et al. 2007). As 443 of our candidate RNA predictions in *Plasmodium* were presented on the PFSANGER array (see in Supplemental Fig. S1), the 17 novel RNAs correspond to a little less than 4%. However, the extensive transcription of the *S. cerevisiae* genome (David et al. 2006) and the limited coverage of *Plasmodium* life cycles in our approach make these numbers difficult to compare directly.

The PFSANGER array hybridizations uncovered some intriguing diversity of ncRNA expression phenomena in *P. falciparum*. Several candidate RNAs exhibited intraerythrocytic stage-dependent differential expression. For example, candidate 3370 is expressed at both ring and trophozoite stage, but with relatively higher expression during the trophozoite stage, (log2 ratio = 5.3, $P = 0.012$) (Supplemental Table S4). Interestingly, some of the ncRNA candidate loci appear to be transcribed from both strands, such as candidate 1537, which maps downstream of the *cyclin 4* gene (PF13_0022, which is expressed in neither rings nor trophozoites) (Supplemental Fig. S4A). Candidate 1678 is detected on the microarray (Supplemental Fig. S4B), and is located upstream of a putative *sir2* homolog (PF13_0152), which is expressed at both stages (data not shown). Candidate 2814 is predicted within the intron of the putative 40S ribosomal protein S23 (PFC0290w) and is expressed in trophozoites only (log2 ratio = 1.7, $P = 0.12$) (Supplemental Table S4), although the protein coding mRNA for PFC0290w is expressed in both stages. Finally, candidates 3217, 3967, and 1320 are closely related members of a larger novel ncRNA family, and all are expressed in both ring and trophozoite stages (Fig. 4; Supplemental Table S4). This novel ncRNA family was first identified by us as a family of 15 conserved GC-rich sequence elements during our initial analysis of the *P. falciparum* genome (Hall et al. 2002). Members of this family are present only in *P. falciparum* and *P. reichenowi* and are exclusively located near the chromosome-internal clusters of antigenically variant *var* genes (Hall et al. 2002).

We selected 76 high-scoring candidate sequences (subsequently referred to as "selected" candidates) for Northern blot hybridization, using criteria shown in Figure 1 and Supplemental

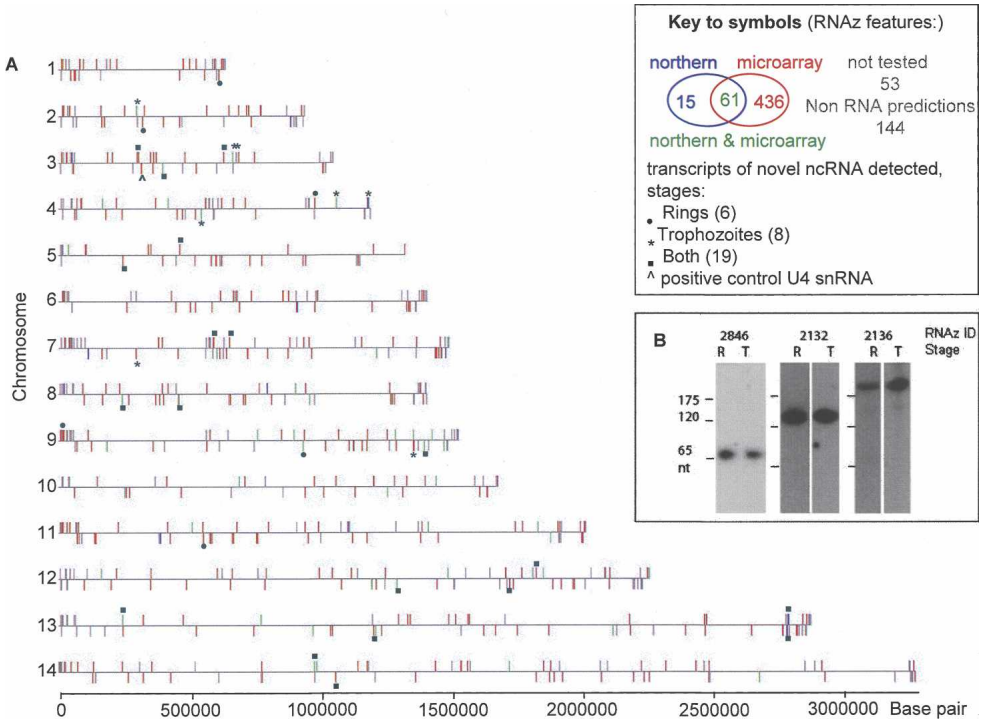


Figure 3. Chromosomal mapping of predicted RNA structures. (*A*) Chromosomal location of the initial 855 RNAz features predicted on *P. falciparum* genome. Each colored bar represents a feature, plotted along the 14 chromosomes as indicated. Keys to the colors and symbols are shown in the *inset* box. (*B*) Northern blots showing ring (*R*) and pigmented trophozoite (*T*) stage total RNA, hybridized with reverse-complement oligonucleotide probes to predicted transcripts, for three of the 78 RNAz candidates. Size references indicated in nucleotides (as derived in Supplemental Fig. S5). Candidates 2846 and 2132 are novel predicted snoRNAs, with transcription confirmed by microarray. Candidate 2136 is unique to the Northern blot screen, as no PFSANGER microarray probe covered this locus.

Figure S1. Northern blots with asexual stage RNAs detected 20 candidates, although five candidates were detected only at low transcript levels (Fig. 3; Supplemental Fig. S5). For two of the 76 selected candidates tested, signals were detected initially only on the microarray, possibly reflecting the relative inefficiency of Northern blot oligonucleotide probes. These were designed for detecting only one of the two possible transcript orientations, were relatively short, and thus relied on exactly predicting the feature boundaries and strand. Supporting this explanation, expression was confirmed for these two (candidates 1537 and 1678), by Northern blot hybridization with riboprobes representing the whole feature, on each strand. This same procedure detected two additional lower-scoring candidates (not on the initial selected list, but microarray-positive candidates 2814 and 3971; see Methods) (Supplemental Figs. S1, S5; Supplemental Table S5). A total of 22, out of 78 candidates tested, were positive by Northern blot. The majority are present in both ring and trophozoite stage RNA. Transcribed candidate 4323 represents one of the 15 member internal *var*-adjacent ncRNA family.

Overall, we confirmed expression of 33 novel candidates (subsequently referred to as "verified" candidates), six by both Northern and microarray, 16 by Northern alone, and 11 by microarray alone (Fig. 1). Predicted structures of verified candidates are shown in Supplemental Figure S6.

Among the 20 selected candidates with positive Northerns, only four corresponded to microarray positives, leaving 16 apparently discrepant "Northern only" positives. The discrepancy between the two methods has several explanations: (1) The PFSANGER microarray design does not represent all the ncRNA candidate sequences tested by Northern blot. Of the 16 discrepant candidates, 10 have incomplete microarray probe coverage (for details, see Supplemental Table S1). (2) Six of the 16 selected candidates detected only by Northern are not explained by probe design limitations, but these might be due to differences in RNA handling. In the microarray analysis, we measured the relative abundance of ncRNA transcripts in size-fractionated in vitro polyadenylated RNA samples, while on the Northern blots, we detected ncRNA transcripts in total RNA. Thus a relative inefficiency of in vitro polyadenylation and subsequent labeling, inefficient size separation, or loss of specific short RNAs in the purification process could lead to inefficient detection on the microarray.

Likelihood of protein coding potential of verified RNAs

We tested open reading frames (ORFs, defined as the longest sequence devoid of stop codons) in all verified RNA predictions and publicly available ESTs matching candidate predictions for protein-coding potential. We identified potential protein-coding ORFs of 19–48 codons in length for verified RNA predictions, and 30–95 codons for ESTs matching RNA predictions. Both these groups have ORF lengths with means significantly shorter than protein coding exons (data not shown).

In an effort to test further whether our candidates could be unannotated exons, we developed a Markov model for probabilistically generating (and scoring) coding and noncoding like *P. falciparum* sequences. Rather than using existing software such as CRITICA (Badger and Olsen 1999), this approach allowed us to account for any peculiar genomic features of *Plasmodium*. When

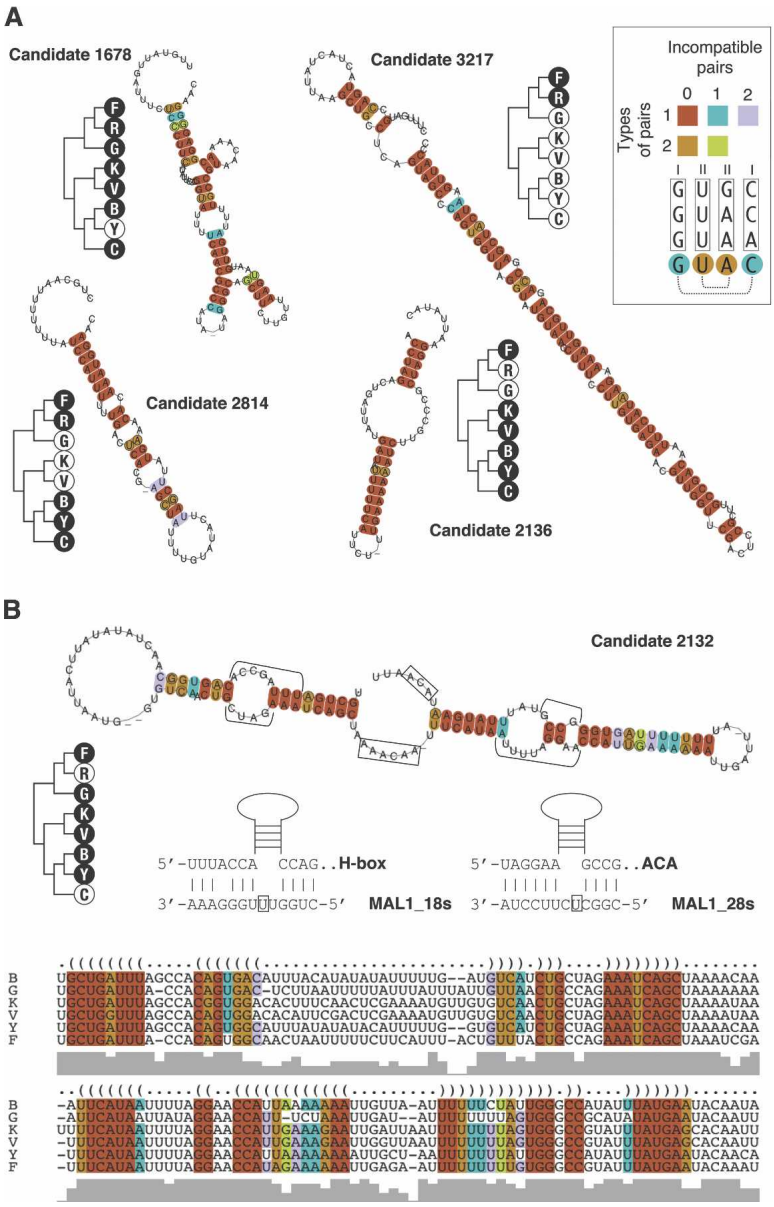


Figure 4. Examples of predicted structures for verified candidate RNAs. (*A*) Examples of structures of experimentally verified RNA candidates. Consensus structures are shown as described by Washietl et al. (2005a). Variable positions in stems are shown in circles. The number of different base pairs (in the alignment) supporting a given structure are indicated by colors (*insert*). An example is provided: Four columns from a hypothetical alignment involved in two base-pairings. Columns I contain one type of (correct) pairs (G:C), and one incompatible pair (G:A). Columns II contain two types of pairs, and no incompatible pairs. For each candidate, the phylogenetic distribution is indicated on a tree structure. Dark dots denote presence in a particular species (F indicates *P. falciparum*; R, *P. reichenowi*; G, *P. gallinaceum*; K, *P. knowlesi*; V, *P. vivax*; B, *P. berghei*; Y, *P. yoelii*; C, *P. chabaudi*). Candidate 3217 corresponds to the previously reported GC-rich repeats associated with *var* gene clusters (Hall et al. 2002). (*B*) Prediction 2132 is a H/ACA box snoRNA. H and ACA motifs are highlighted in boxes (the *P. falciparum* sequence has a degenerate H box) on the predicted structure; putative guide regions complementary to 18S and 28S rRNA are indicated by parentheses. Predicted complementarity to ribosomal RNAs is shown with target uridines (U673 and U3414 respectively) in boxes. The alignment of the six relevant *Plasmodium* species sequences is provided (b indicates *berghei*; g, *gallinaceum*; k, *knowlesi*; v, *vivax*; y, *yoelii*; f, *falciparum*).

testing a coding model against two models based on intergenic and intronic sequence, respectively, the putative protein-coding sequences from the above-mentioned ESTs scored similarly to intron sequences, suggesting that they were unlikely to encode *P. falciparum* proteins (Fig. 5A,B). Scores for hypothetical protein-coding ORFs from verified RNA predictions were intermediate between coding and intronic sequences (Fig. 5A,B). Thus although the Markov model approach suggests that the verified RNAs are not protein coding, it does not completely rule out any protein-coding potential of the verified RNAs.

Interestingly, sequences from known RNA genes score higher using the coding model than using the intergenic and the intronic model (Fig. 5A,B), signifying that RNA genes can be discriminated from noncoding DNA in terms of composition. However, the verified RNA sequences score poorly using a model based on known RNAs (Fig. 5C), indicating that our novel *Plasmodium*-specific RNA genes differ from the previously known RNA genes in this respect.

We also conducted a simple compositional codon usage analysis. From the composition within *Plasmodium* codons (GenomeAtlas) (Hallin and Ussery 2004), we deduced five expectations of *Plasmodium* protein coding potential: (1) occurrence of A at first codon position (A1) is higher than at codon position 3 (A3); (2) occurrence of T at first codon position is higher than at codon position 3; (3) occurrence of G at first codon position is higher than the occurrence of C; (4) occurrence of A at first codon position is higher than the occurrence of T; and (5) occurrence of G or C at first codon position (GC1) is higher than at codon position 3 (GC3). We compared the number of ORFs from verified RNA fulfilling these expectations with a sample of all 41 *P. falciparum* annotated CDS shorter than 200 amino acids (Table 1). With the exception of expectation 1 (and to a lesser extent expectation 5), the vast majority of short annotated CDS fulfils the expectations for base composition within codons. In contrast, only around half of the ORFs in verified RNAs fulfils the expectations, as would be expected if composition was not restricted by a periodicity of three.

As a final test of the protein coding potential of these ncRNAs, we analyzed the ratio of tyrosine codons, based on the observation of a high ratio between the synonymous codons UAU and UAC (ratio: 8.27) (Nakamura et al. 2000). The UAU:UAC ratio in verified RNAs does not match the observed ratios of known protein-coding genes (Table 1), but it correlates well with the observed ratio A:C in the verified RNAs (ratio: 3.26). Thus, the overall

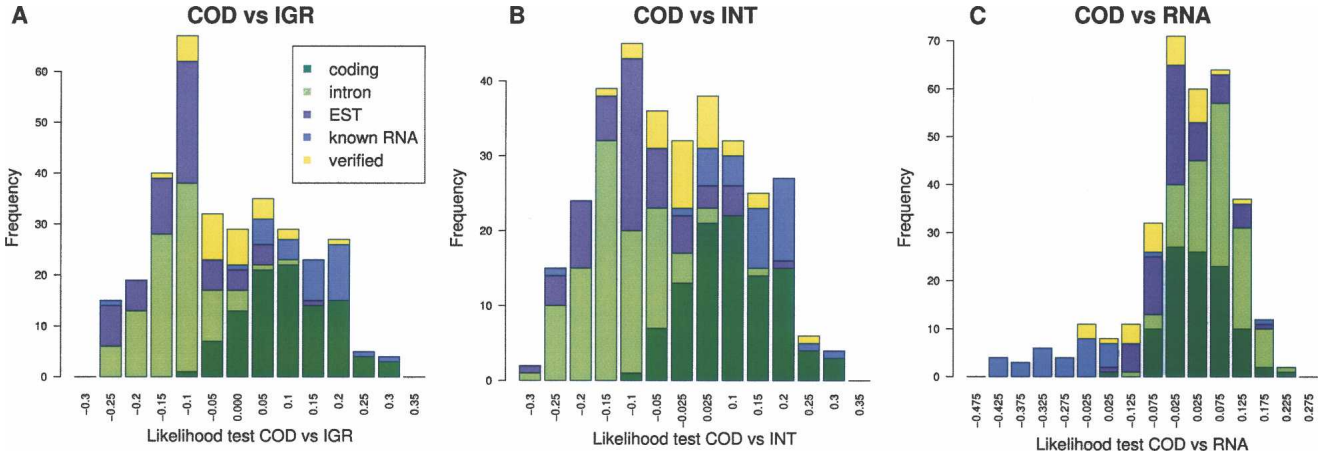


Figure 5. Distributions of log-likelihood ratios. Histograms of the distributions of log-likelihood ratios of the coding model (COD) versus the three dinucleotide-based background models (IGR indicates intergenic; INT, intronic; RNA, RNA genes). Coding indicates set of annotated protein coding codons; intron, set of intron sequences; EST, set of longest ORFs from ESTs with matches to predicted RNAs; known RNA, set of sequences from known tRNAs, rRNAs, and snRNAs; verified, set of longest ORFs from predicted RNAs with verified expression.

composition—rather than codon bias—determines the ratio between UAU and UAC triplets in the verified RNAs.

In summary, the Markov model approach showed that the ORFs of ESTs matching predicted RNAs are unlikely to be protein-coding but did not allow us to rule out the possibility that any ORFs in verified RNAs are protein-coding. However, based on our codon usage analysis, we argue that the absence of stop codons in these sequences is not a result of selection maintaining protein-coding capacity.

Nucleotide substitutions

To determine whether the predicted RNA genes are subject to purifying selection, we examined the density of nucleotide substitutions compared with other regions of noncoding DNA (intergenic regions and introns). Nucleotide substitution and polymorphism data (Jeffares et al. 2007) were obtained from alignments of sequence reads from the chimpanzee parasite *P. reichenowi* and two recently sequenced *P. falciparum* strains, a Ghanaian clinical isolate and the IT strain (a cultured laboratory strain) to the *P. falciparum* strain 3D7 (Gardner et al. 2002). Few polymorphisms were observed between clone 3D7 and the two other *falciparum* strains, and all analysis was restricted to *P. falciparum*/*P. reichenowi* substitutions (Fig. 6A). The *P. falciparum*–*P. reichenowi* sequence divergence in known RNAs was significantly less than noncoding DNA ($P = 1.3 \times 10^{-4}$, using a hypergeometric distribution, see Methods), indicating strong functional constraints on nucleotide substitutions. Divergence in selected and verified candidates was also significantly less than noncoding DNA ($P = 0.0075$ and $P = 0.019$, respectively), indicating that these regions are also subject to selective constraint. However, the divergence across all candidate RNA predictions was not significantly different from noncoding DNA ($P = 0.53$) (Fig. 6; Supplemental Table S6).

This constitutes a global test, and the underlying assumption of a random distribution of substitutions in the genome is unlikely to be fulfilled. To further assess the significance of the observed substitution densities at a local level, we took each RNA prediction and changed the coordinates, so that the real predictions would be compared against structurally identical—yet completely hypothetical—predictions (i.e., with identical structures, enforced regardless of primary sequence, referred to as relocated predictions) residing a fixed distance away from the real predictions. This was done for 250, 500, 750, 1000, 1500, 2000, and 3000 base pairs both upstream and downstream. In agreement with the global test, we found that the divergence in the known RNA predictions indeed appear to be lower than the corresponding relocations (Fig. 6C). A weaker tendency is observed in the selected candidates, whereas candidate predictions overall show no difference, and the verified candidates display no consistent pattern. The latter is most likely due to the small sample size.

The divergence was significantly lower in regions predicted to be involved in base-pairing (Fig. 6B) for all candidate predictions ($P = 5.3 \times 10^{-6}$), selected candidates ($P = 0.0048$), and known RNA predictions ($P = 3.7 \times 10^{-10}$), although not for the "verified" candidates ($P = 0.052$) (Supplemental Table S6). Further, the difference in substitution density between stem and loop is consistently more pronounced in all classes of predictions compared with their corresponding relocations (Fig. 6D; Supplemental Table S6). In addition, we obtained evidence for mutually compensatory substitutions (12 substitutions out of 233) in the predicted base-paired (stem) regions in the "verified candidates" between *P. falciparum* and *P. reichenowi*, changing primary se-

Table 1. Compositional expectations for Plasmodium ORFs and Tyrosine codon usage

	Annotated CDS <200 amino acids	Verified RNA candidates	Codon usage database[a]
Compositional expectations			
1 A1>A3	24 (59%)	15 (52%)	
2 T1>T3	37 (90%)	13 (45%)	
3 G1>C1	34 (83%)	13 (45%)	
4 A1>T1	36 (88%)	14 (48%)	
5 GC1>GC3	29 (71%)	12 (41%)	
Tyrosine codon usage			
UAU	113	82	153001
UAC	17	23	18498
Ratio	6.65	3.57	8.27

[a]Data for all *P. falciparum* CDS as obtained from Nakamura et al. (2000).

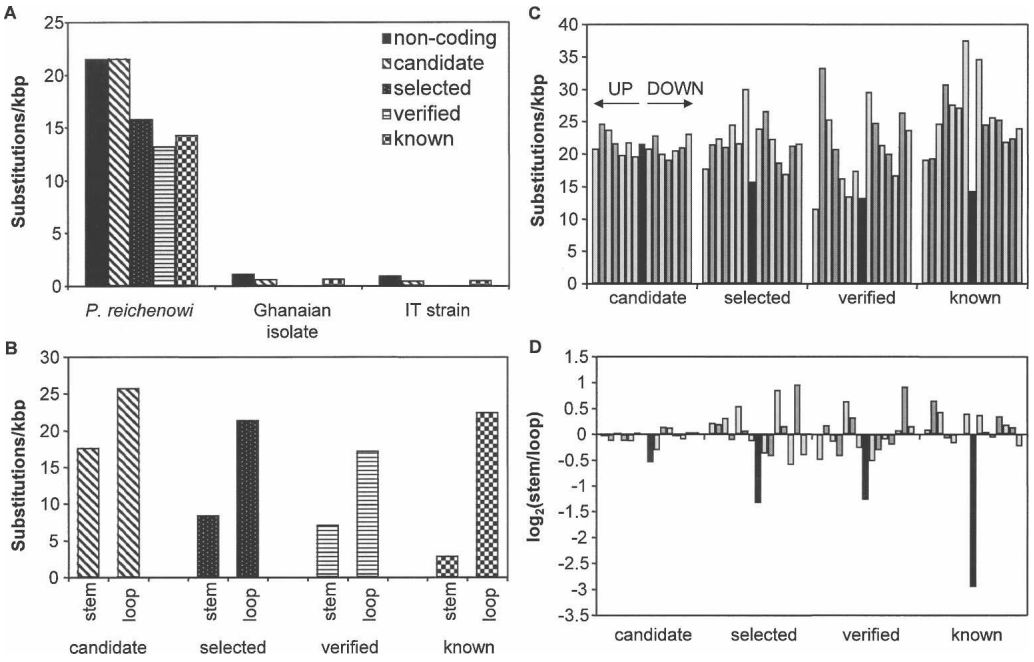


Figure 6. Substitution densities. (*A*) Density of substitutions and polymorphisms per Kbp for *P. falciparum* noncoding DNA within the different classes of predictions. Densities are shown for three different alignments; *P. falciparum* Clone 3D7 against *P.f.* Ghanaian clinical isolate, *P.f.* IT strain and *P. reichenowi*. (*B*) Each base within each locus was classified as either stem (predicted to form base pairs) or loop (predicted not to form base pairs). Prediction classes as in *A*. (*C*) Substitution densities by prediction class (as in *A*; black bars). Each class of prediction is preceded by virtual predictions translocated to a position of 3000, 2000, 1500, 1000, 750, 500, and 250 bp upstream (based on genomic coordinates, regardless of orientation of prediction), respectively, and succeeded by virtual predictions translocated to similar distances downstream (gray bars). (*D*) Log(2) ratios of substitution density between STEM and LOOP positions (as defined in *B*). Black bars denote real candidates, gray bars translocations as described in *C*.

quence while maintaining secondary structure (Supplemental Fig. S7). None of the four substitutions in stem regions of known tRNA genes are compensatory (data not shown).

Our procedure for constructing the multiple alignments resulted, in some cases, in alignments with long trailing sequences only found in a minority of genomes that are not believed to be part of the RNA gene. This could potentially affect the above analyses of substitution densities. We therefore repeated all the above analyses ignoring positions in the alignments with gaps in the consensus sequence from RNAz. This did not essentially alter the observed pattern (Supplemental Fig. S8; Supplemental Table S7).

In summary, although the obtained *P*-values are not particularly strong, we have been able to detect a consistent signal of divergence level differences in RNA predictions. Considering the relative sparseness of substitutions, the fact that functional RNA sequences are expected to be under purifying selection, and the uncertainty of structure prediction, the finding of differences in divergence levels between stem and loop sequences is remarkable. We propose that divergence data may serve as a useful indicator in RNA gene validation, but that the size of the current *Plasmodium* data sets imposes significant restrictions on any practical use in prediction.

Function of novel RNAs

It is difficult to identify biological roles for the majority of these transcripts on a genome scale. Determining the function of novel ncRNAs will require detailed experimentation. However, sequence and structure analysis, conservation patterns, and known motifs of the novel RNA genes enable us to identify potential classes. The sequence and structure conservation of candidate 2132 is consistent with a novel H/ACA box small nucleolar RNA (snoRNA) (Fig. 4). H/ACA snoRNAs guide the essential pseudouridylation of ribosomal (and other) RNAs, although the absence of RNA modification maps in *Plasmodium* makes target prediction difficult. By manual inspection, the most consistent of the predicted targets are U673 of 18S rRNA and U3414 of 28S rRNA (Fig. 4; Schattner et al. 2005). Candidate 2846 is a predicted C/D box snoRNA, and candidates 1335 and 1537 may represent fragments of C/D box snoRNA predictions. Candidate 1335 shows a low scoring match to a putative human C/D box snoRNA. Our screen has further identified three of the four selenocysteine insertion sequence (SECIS) elements known in *P. falciparum* (Mourier et al. 2005; Lobanov et al. 2006).

We tested all predicted candidates for hairpin structures with miRNA precursor potential using RNAmicro (Hertel and Stadler 2006). Five candidates were classified as miRNAs with high probability (Supplemental Fig. S9). However, processing of ncRNA precursors to smaller (22–23 bp) miRNA species has not been shown to occur in *Plasmodium* (Rathjen et al. 2006), so detailed experimental analysis is required to test if these hairpin structures are in fact processed to miRNAs. As conventional miRNA processing machinery is not detected in *Plasmodium*, processing would have to be carried out by host machinery, or by a novel mechanism with parasite-encoded proteins.

The primary sequence and predicted secondary structures of the verified RNAs are not homologous with functionally annotated RNAs (e.g., those in the Rfam database). We suggest that initial functional studies are led by contextual information from the surrounding genomic regions.

Discussion

In this study, we present a systematic in silico screen for conserved noncoding RNA structures by comparing the *P. falciparum* genome with seven other *Plasmodium* species, resulting in 604 novel candidate ncRNA structures. By combining microarray and Northern data, we provide evidence for expression of 33 novel ncRNA candidates in the asexual blood stages of *P. falciparum*, a subset of which show stage specificity during the asexual life cycle. Few of the expressed novel candidates have primary sequence similarity to previously characterized loci. Given the stage-specific expression of transcripts and/or proteins in *P. falciparum* (Bozdech et al. 2003; Le Roch et al. 2004), many predicted but as-yet-unverified candidates may be expressed elsewhere in the parasite's complex life cycle or may be below the level of detection of Northern blots or microarray hybridizations.

Eighteen noncoding RNA genes have previously been reported in the *P. falciparum* genome (Upadhyay et al. 2005); 11 of these candidates are present in our alignments, and six are found among our RNA predictions (Supplemental Table S8). Only one of those novel candidates was verified as transcribed by Upadhyay et al. (2005), and our microarray analysis confirms expression of this candidate (our candidate 3370). This comparison indicates the difficulty posed by identifying noncoding RNA genes in *P. falciparum*. The combined Northern and microarray analyses presented here extend the repertoire of novel ncRNA transcripts to a total of 33, transcribed within asexual stages of the life cycle. Twenty-nine of our candidates overlap structured ncRNAs (three previously known snRNAs, four of our "candidate" RNAs, and 22 of our "selected" RNAs) described recently by Chakrabarti et al. (2007). Interestingly, this study has provided independent evidence for the expression of the chromosome-internal, *var* gene-associated, novel structured ncRNA family in *P. falciparum* described previously.

Plasmodium genomes vary significantly in their GC-content (from 20% in *P. falciparum* to 42% in *P. vivax*). Thus, ncRNA candidates that are conserved across the three major taxonomic clades of *Plasmodium* (as shown in Fig. 2) should be regarded as strong candidates. If we assume that all the candidates that are conserved outside *P. falciparum* and the closely related *P. reichenowi* are true ncRNAs, and add the number of candidates only found in *P. falciparum*/*reichenowi* that are either verified or show compensatory mutations, this gives a conservative estimate of 120 high-confidence predictions of novel noncoding structural RNAs in *Plasmodium* (Supplemental Table S9). Those ncRNA candidates that show a narrow phylogenetic distribution are presumably involved in species-specific functions, but we have no evidence at present as to what any of these functions may be.

Genome-wide informatics approaches are extremely useful for the prediction of novel ncRNA transcripts but, in general, are poor at predicting function. Determining the function of the ncRNAs identified in this study will require a thorough analysis of carefully chosen candidates and will rely on the fact that they confer a readily assessable phenotype. This is now a high priority but, owing to the inefficiency of transfection in *P. falciparum*, might take some years to complete. This is the first study in the *Plasmodium* field that aims to identify conserved ncRNA structures by comparing eight *Plasmodium* genomes, using the RNAz program. RNA gene finding algorithms are still in their infancy and differ in their sensitivities and specificities (Washietl et al. 2007). It is expected therefore that that many more ncRNA species still remain undetected in the *P. falciparum* genome. Nevertheless, our study highlights the abundance of these novel RNA structures in *P. falciparum* and provides a framework for functional studies to elucidate cellular roles for these ncRNAs in *Plasmodium* biology. Since the conventional RNAi pathway is absent from *Plasmodium*, ncRNAs may play a major role in gene regulation, using novel pathways or mechanisms unique to *Plasmodium* parasites.

Methods

The annotation of *P. falciparum* predicted ncRNA candidates described in this manuscript is available at GeneDB (http://www.geneDB.org/). The complete nucleotide sequences of all the 855 predicted structured ncRNA candidates have been provided in Supplemental Table S11. The microarray data for the ncRNA candidates described in this study have been submitted to Array Express under the accession number E-SGRP-8.

Genomic data sources

Genomes of *Plasmodium falciparum* (version 2002.10.03) (Gardner et al. 2002), *P. berghei* (version 092304), *P. chabaudi* (version 070104), *P. gallinaceum* (version 070104), *P. knowlesi* (version 031104), *P. reichenowi* (version 031104), *P. vivax* (version 22-Nov-2004), and *P. yoelii* (version 2002.09.10) (Carlton et al. 2002) were downloaded from GeneDB (http://www.genedb.org/) and PlasmoDB (http://www.plasmodb.org/). During the course of this work, higher-coverage genomic sequence became available for *P. berghei*, *P. chabaudi*, and *P. knowlesi*. These sequences were used for the phylogenetic distribution shown in Figure 2 only.

Sequence similarity searches

All intergenic and intronic sequences with a length of at least 100 bp were retrieved for *P. falciparum*. This resulted in 12,056 sequences with a total length of 10,848,102 bp, comprising 47% of the genome. These sequences were searched for similarity against the other *Plasmodium* genomes using WU-BLASTN (http://blast.wustl.edu/) with dust and seg low-complexity filters. Only hits with a length ≥50 bp and with at least 70% identity were further processed.

BLAST-matches were mapped back at the *P. falciparum* genome and merged using the following procedure. If a given sequence from *P. falciparum* had a match to genome X and this overlapped another match to genome Y, a *P. falciparum* sequence spanning the combined length of the two matches was retrieved along with the sequence of the two matches (and so forth for all compared genomes). These sequences were then realigned using MUSCLE (Edgar 2004; Gardner et al. 2005).

RNAz (Washietl et al. 2005b) was run on the resulting 4845 alignments and their reverse complements. Alignments with a length exceeding 400 bp were scanned using a window of 400 bp in steps of 50 bp. As the number of sequences in an alignment is restricted to six by RNAz (Washietl et al. 2005b), all alignments containing seven or more sequences were reduced by the following rule: First, if a *P. reichenowi* sequence is present it is removed. Second, if *P. reichenowi* is absent or an additional sequence needs to be removed, the *P. vivax* sequence is removed. Removed sequences were not used for further sequence analysis, although their presences were recorded for phylogenetic distribution analysis.

By using RepeatMasker (*RepeatMasker Open-3.0*. 1996–2004; http://www.repeatmasker.org/), we filtered out all predictions in which ≥30% of the *P. falciparum* sequence was masked as low-

complexity sequence. All predictions with similarity to either protein-coding genes in *P. falciparum* or any non-RNA gene sequence deposited in GenBank were removed from the data set. All 187 predictions removed using the above criteria are referred to as "non-RNA predictions" and were discarded from phylogenetic and sequence divergence analysis.

From the genome annotation of *P. falciparum*, all RNA genes were extracted together with the *P. falciparum* entries in Rfam (Griffiths-Jones et al. 2005); this constituted the set of annotated RNA genes. The remaining predictions were tested for similarity to these annotated RNA genes. tRNAscan-SE (Lowe and Eddy 1997) predicted 20 additional tRNAs. Searches with the covariance model at Rfam (Griffiths-Jones et al. 2005) detected an additional single SSU rRNA. Finally, a previously published thermoregulated ncRNA (GenBank accession: AY496275), a selenocysteine tRNA, and SECIS element (Griffiths-Jones et al. 2005; Mourier et al. 2005) were detected in the set of predictions. The above predictions are referred to as "known" RNA predictions.

Functional assignment of novel RNA predictions

Structural similarity to snoRNAs was tested using SnoGPS (Lowe and Eddy 1999; Schattner et al. 2004) and Snoreport (http://www.tbi.univie.ac.at/~jana/software/SnoReport.html).

Markov models and ORFs

For each sequence, the longest run devoid of stop codons was recorded. For verified RNA predictions, this resulted in ORF lengths between 19 and 48 codons. For all three reading frames, each verified RNA sequence contains on average 12.6 stop codons (range 2–23). Due to the uncertainty of orientation of both candidate predictions and ESTs, the longest ORF was determined from both orientations of the ESTs matching candidate predictions (ORF lengths between 30 and 95 codons).

For the coding model, we assume the input predicted ORF sequences are in the correct frame and the lengths are a factor of three. We compute the conditional probability of the sequence given the protein coding model (COD) using the following expression:

$$P(seq|COD) := \prod_{i=1}^{N/3} p^1(x_{3i-2})p^2(y_{3i-1})p^3(z_{3i}),$$

where N is the sequence length, x_j, y_k, and z_l are the nucleotides at positions j, k, and l and $p^m(x_j)$ is the probability of observing nucleotide x in codon position m. This probability is estimated from the nucleotide frequencies by Hallin and Ussery (2004).

In a similar fashion, we compute the conditional probability of the sequence given the dinucleotide distribution background models (α) using the following first-order model:

$$P(seq|\alpha) := p(x_1)\prod_{i=2}^{N} p(x_i|y_{i-1}),$$

where $p(x_i|y_{i-1})$ is the probability of observing nucleotide x at position i given that the nucleotide at position $i-1$ is y. Similar to the COD model, the probability is estimated from the observed dinucleotide distributions. We tested the sequences using a number of different dinucleotide models, including intronic (α = INT), intergenic (α = IGR), and structural RNA (α = RNA).

Sequence divergence analysis

Divergence and polymorphism data were taken from Jeffares et al. (2007). For each prediction category (candidate, selected candidates, verified candidates, and known predictions), we calculated the probability of obtaining the observed number of substitutions (x) (or fewer) in the category sequences from the number of base pairs in the category sequences (n) based on a total population consisting of the combined number of substitutions in the category sequences and the noncoding DNA (k), and of the combined number of base pairs in the category sequences and the noncoding DNA (N). The hypergeometric probability of obtaining x substitutions in n base pairs are calculated as

$$h(x|N,n,k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}.$$

Analysis of *P. falciparum* EST sequences that map to RNA predictions

All *P. falciparum* ESTs from dbEST at NCBI (http://www.ncbi.nlm.nih.gov/dbEST) were downloaded and searched for similarity to the RNA predictions. ESTs with matches of at least 50 bp and with minimum 95% identity to candidate RNA predictions were retrieved. Genomic mapping were performed for these ESTs using sim4 (Florea et al. 1998). All mappings successfully aligning at least 90% of the EST sequence were considered. If an EST mapped to a candidate RNA prediction and also to a protein-coding gene (minimum 50 bp and 90% ID), candidate RNA prediction was considered part of the protein-coding transcript. Such structural RNA predictions were then considered as cotranscribed with a protein-coding gene and discarded from further analysis.

Parasites

3D7 parasites were cultured according to the method of Trager and Jensen (2005) and were synchronized with sorbitol (Lambros and Vanderberg 1979). Cells were harvested from cultures at 5%–10% parasitemia, at 10–15 h post-invasion (rings) and 30 h post-invasion (pigmented trophozoites).

RNA extraction and size-fractionation

For microarrays and Northern samples tested by riboprobe, total RNA was isolated after thorough saponin-lysis of parasite infected red blood cells (RBCs), to remove contaminating RBCs. RBCs have been demonstrated to carry large amounts of globin RNA, which affects human microarray analysis (Feezor et al. 2004). Additionally, we and others have found that uninfected RBCs contain abundant human miRNAs (Rathjen et al. 2006; our observations [unpubl.]). Ring stage infected RBCs (6 mL) or pigmented trophozoite-infected RBCs (1 mL) were washed in 1× PBS (0.01 M Na-phosphate, 0.0027 M KCl, 0.138 M NaCl at pH 7.4). Cells were incubated in 5 pellet volumes 0.01% saponin/1× PBS at room temperature, releasing parasites from RBCs and lysing the bulk of uninfected RBCs. Cells were centrifuged at 12,000 rpm in a microfuge, 2 min for rings (1 min for trophozoites). This was repeated several times until RBC lysis was complete. After all supernatant was removed and 1 mL TRIzol was added per original 5 mL infected RBCs; this was processed for total RNA as the method of Kyes et al. (2000). Briefly, 0.2 mL $CHCl_3$ was added per 1 mL TRIzol, followed by a 13,200 rpm spin, 25 min at 4°C. The aqueous layer (0.6 mL) was precipitated in 0.5 mL of isopropanol for 2 h on ice, vortexed, and then centrifuged 30 min at 4°C, 13,200 rpm. The pellets were resuspended in 0.2 mL of DEPC-water and checked for quality and concentration by agarose gel analysis (Kyes et al. 2000) and by measuring OD_{260}.

To size-fractionate RNA samples for microarrays, RNA was electrophoresed on gels with a 4.5% acrylamide/7 M urea/1× TBE stacker layered on a 15% acrylamide/7 M urea/1× TBE running gel (1× TBE: 0.089 M Tris, 0.089 M boric acid, 2 mM EDTA). Twenty-five micrograms of heat-denatured RNA (in formamide) was loaded per gel, and four gels were run for each biological replicate (two replicates each of rings and trophozoites, 100 µg total RNA for each replicate). Gels were divided into "long" RNA (longer than 500 nucleotides) and "short" RNA (smaller than 500 nucleotides) by cutting at a predetermined distance from the wells (this was measured by running Ambion Century markers on a mock gel, and each gel was run under exactly the same conditions). RNA was eluted in 0.3 M sodium acetate (pH 5.2), overnight, and then precipitated in 4 volumes ethanol (overnight at −20°C). Short RNA was coprecipitated in the presence of carrier, ~24 µg glycogen (Roche). Samples were resuspended in RNase-free water. The short RNA fraction was polyA-tailed using Ambion polyA polymerase (final 1 unit in a 50 µL reaction) and 1 mM rATP, 60 min at 37°C. The RNA was precipitated and processed as above for target amplification and labeling. For total RNA microarray samples, 20 µg total RNA (from saponin-lysed parasites) was treated with Turbo DNase, as per manufacturer's instructions (Ambion) and then ethanol precipitated. For Northern blot samples tested only by oligonucleotide hybridization, total RNA was prepared from ring and pigmented trophozoite stages as described by Kyes et al. (2000), with no saponin-lysis step.

Expression analysis using PFSANGER Affymetrix genome-wide array

To check the overall expression of any of the predicted RNAz features, a microarray approach was chosen. Affymetrix PFSANGER arrays are high-density 8-µm custom 25-mer oligonucleotide arrays, whose tiling-like design was based on the *P. falciparum* genomic sequence released in January 2005 (http://www.genedb.org). The arrays are PM-only (perfect match) and comprise 2.44 million *Plasmodium* probes, of which 2.32 million are unique and specific to *P. falciparum*. Due to specificity/isothermal constraints during design, the probes are distributed nonrandomly throughout the genome. Both strands are represented, with 1.7 M probes in noncoding regions of the genome, and 0.6 M within exons.

Poly-A tailed short RNA (shorter than 500 nt; ~15 µg) and unmodified long RNA (longer than 500 nt; 10 µg) from both ring and pigmented trophozoite stages was reverse transcribed and biotin-labeled as cRNA, using the GeneChip IVT Labeling kit as recommended by Affymetrix. Hybridizations were carried out for 16 h at 45°C with constant rotation at 60 rpm. Following hybridization, the solutions were removed and the arrays washed and stained on a fluidics station (Affymetrix FS 450). Gene arrays were then scanned at an emission wavelength of 570 nm at 1.56 µm pixel-resolution using a confocal scanner (Affymetrix GeneChip Scanner 3000 7G).

After scanning, the hybridization intensity for each 25-mer feature was computed using Affymetrix GCOS v1.3 software, and the CEL files were transferred into the R/Bioconductor environment (http://www.bioconductor.org/, http://www.r-project.org/) (Gentleman et al. 2004) for downstream analyses. Probesets (groups of probes mapping a RNAz predicted locus) for 641 of the initial 855 predicted candidates were identified. A chip definition file (CDF) for the RNAz predicted loci was generated in-house for these analyses (affy, makecdfenv, altcdfenv packages) (Gautier et al. 2004). Arrays were background adjusted and quantile normalized using the robust multiarray averaging algorithm (RMA) (Irizarry et al. 2003). Differential expression between conditions (contrasts e.g., "short" rings vs. "long" rings) was estimated by the application of linear models, with Bayesian correction, via the "limma" Bioconductor package. Log2 ratios ("coefficients" or *M*-values) were generated and corrected for false discovery rate using the Bonferroni-Hochberg method (Smyth and Speed 2003; Smyth et al. 2005) with associated measures of statistical significance, for each locus in each contrast. The physical chromosomal views of the results were achieved using Genespring (Silicon Genetics).

Northern blots and hybridization

Total RNA extracted from 3D7 ring or trophozoite parasites (not saponin-lysed), 1 µg per lane, was denatured at 65°C in formamide for 2 min, then size-fractionated on 15% acrylamide, 7 M urea 1× TBE (0.089 M Tris, 0.089 M boric acid, 2 mM EDTA) gels. After ethidium bromide staining the gel, RNA was transferred to Hybond N+ by electroblotting in 0.5× TBE, 50 v for 15 min (for oligonucleotide hybridized blots) or in 7.5 mM sodium hydroxide by capillary transfer overnight (for riboprobe blots). The blot was UV-cross-linked (Stratalinker), and UV-visible ethidium bromide bands were marked directly on the filter as size references. Blots were then cut between lanes into strips for hybridization. Strips were prehybridized for ~1 h at 37°C, in hybridization buffer (7% sodium dodecyl sulfate/ 5% dextran sulfate/ 0.25 M sodium phosphate buffer at pH 7.2/0.25 M sodium chloride/1 mM EDTA).

For oligonucleotide probes, reverse complement oligonucleotides were designed for the 76 selected candidates from the RNAz predictions, as well as for a predicted U4 snRNA (PFC0358w, positive control; candidate 2830) and low scoring region candidate 2962 (negative control) (Supplemental Table S10). For each oligonucleotide, 25–30 ng was end-labeled in a 20 µL reaction, with 1 µL 6000 Ci/mmol, 10 mCi/mL γ-^{32}P-ATP (GE Healthcare/Amersham Biosciences), using 5 units T4 polynucleotide kinase in 1× reaction buffer A supplied by manufacturer (Fermentas), for 45 min at 37°C. Two prehybridized Northern blot strips (ring and trophozoite) were placed into 3 mL of fresh hybridization buffer with heat-denatured probe and allowed to incubate overnight at 37°C. Blots were washed in 3× SSC (0.45 M sodium chloride, 45 mM sodium citrate at pH 7)/ 0.1% SDS, twice, at 37°C. The washed membranes were exposed to autoradiographic film (1 h to several days). Sizes of hybridized bands were estimated from the molecular weight markers.

For riboprobes, PCR products were generated for the length of each of five predicted ncRNAs. Lig n' Scribe T7 RNA polymerase adapters were added for transcription in each direction (as per manufacturer's instructions, Ambion), and labeled transcripts were prepared using 10 µCi α-^{32}P-UTP (3000 Ci/mmol; GE Healthcare), Maxiscript kit (as per manufacturer's instructions, Ambion, except cold nucleotide concentrations were 7.5 mM each, and reactions were incubated for 1 h, chased with cold UTP, and then incubated a further hour). Northern blots containing the same ring- and trophozoite-stage RNA samples used in the microarray analysis were hybridized with probes heat denatured at 65°C, hybridization at 50°C, and washes at 50°C in 1× SSC/ 0.1% SDS.

Acknowledgments

We thank Carol Churcher, Mike Quail, David Harris, and the pathogen sequencing teams. as well as acknowledge the support of the Wellcome Trust Sanger Institute core sequencing and informatics groups. We also thank Thomas Keane and Ulrike

Böhme for helpful advice with this manuscript. We thank Jane Carlton and staff members of The Institute for Genome Research and the PlasmoDB team (http://www.PlasmoDB.org/) for making the *P. vivax* genome sequence data available ahead of publication. This work was supported by the Wellcome Trust. T.M. was supported by a grant from the Lundbeck Foundation. P.P.G. is supported by The Carlsberg Foundation

References

Aravind, L., Iyer, L.M., Wellems, T.E., and Miller, L.H. 2003. Plasmodium biology: Genomic gleanings. *Cell* **115:** 771–785.

Backofen, R., Bernhart, S.H., Flamm, C., Fried, C., Fritzsch, G., Hackermuller, J., Hertel, J., Hofacker, I.L., Missal, K., Mosig, A., et al. 2007. RNAs everywhere: Genome-wide annotation of structured RNAs. *J. Exp. Zool. B Mol. Dev. Evol.* **308:** 1–25.

Badger, J.H. and Olsen, G.J. 1999. CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16:** 512–524.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1:** E5. doi: 10.1371/journal.pbio.0000005.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419:** 512–519.

Chakrabarti, K., Pearson, M., Grate, L., Sterne-Weiler, T., Deans, J., Donohue, J.P., and Ares Jr., M. 2007. Structural RNAs of known and unkbown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA* **13:** 1–17.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Coulson, R.M., Hall, N., and Ouzounis, C.A. 2004. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res.* **14:** 1548–1554.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci.* **103:** 5320–5325.

Duraisingh, M.T., Voss, T.S., Marty, A.J., Duffy, M.F., Good, R.T., Thompson, J.K., Freitas-Junior, L.H., Scherf, A., Crabb, B.S., and Cowman, A.F. 2005. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell* **121:** 13–24.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797. doi: 10.1093/nar/gkh340.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Escalante, A.A. and Ayala, F.J. 1994. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl. Acad. Sci.* **91:** 11373–11377.

Feezor, R.J., Baker, H.V., Mindrinos, M., Hayden, D., Tannahill, C.L., Brownstein, B.H., Fay, A., MacMillan, S., Laramie, J., Xiao, W., et al. 2004. Whole blood and leukocyte RNA isolation for gene expression analyses. *Physiol. Genomics* **19:** 247–254.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

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.

Gardner, P.P., Wilm, A., and Washietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* **33:** 2433–2439. doi: 10.1093/nar/gki541.

Gautier, L., Moller, M., Friis-Hansen, L., and Knudsen, S. 2004. Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics* **5:** 111.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5:** R80. doi: 10.1186/gb-2004-5-10-r80.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. 2005. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33:** D121–D124.

Gunasekera, A.M., Patankar, S., Schug, J., Eisen, G., Kissinger, J., Roos, D., and Wirth, D.F. 2004. Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* **136:** 35–42.

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.

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., et al. 2005. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307:** 82–86.

Hallin, P.F. and Ussery, D.W. 2004. CBS Genome Atlas Database: A dynamic storage for bioinformatic results and sequence data. *Bioinformatics* **20:** 3682–3686.

Hertel, J. and Stadler, P.F. 2006. Hairpins in a Haystack: Recognizing microRNA precursors in comparative genomics data. *Bioinformatics* **22:** e197–e202.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249–264.

Jeffares, D.C., Pain, A., Berry, A., Cox, A.V., Stalker, J., Ingle, C.E., Thomas, A., Quail, M.A., Siebenthall, K., Uhlemann, A.C., et al. 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat. Genet.* **39:** 120–125.

Jongeneel, C.V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C.D., Khrebtukova, I., Kuznetsov, D., Stevenson, B.J., Strausberg, R.L., Simpson, A.J., et al. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* **15:** 1007–1014.

Kyes, S., Pinches, R., and Newbold, C. 2000. A simple RNA analysis method shows var and rif multigene family expression patterns in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **105:** 311–315.

Lambros, C. and Vanderberg, J.P. 1979. Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *J. Parasitol.* **65:** 418–420.

Le Roch, K.G., Johnson, J.R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S.F., Williamson, K.C., Holder, A.A., Carucci, D.J., et al. 2004. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* **14:** 2308–2318.

Lobanov, A.V., Delgado, C., Rahlfs, S., Novoselov, S.V., Kryukov, G.V., Gromer, S., Hatfield, D.L., Becker, K., and Gladyshev, V.N. 2006. The *Plasmodium* selenoproteome. *Nucleic Acids Res.* **34:** 496–505. doi: 10.1093/nar/gkj450.

Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964. doi: 10.1093/nar/25.5.955.

Lowe, T.M. and Eddy, S.R. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.

Mair, G.R., Braks, J.A., Garver, L.S., Wiegant, J.C., Hall, N., Dirks, R.W., Khan, S.M., Dimopoulos, G., Janse, C.J., and Waters, A.P. 2006. Regulation of sexual development of *Plasmodium* by translational repression. *Science* **313:** 667–669.

Mattick, J.S. and Makunin, I.V. 2006. Non-coding RNA. *Hum. Mol. Genet.* **115:** R17–R29.

Mignone, F., Gissi, C., Liuni, S., and Pesole, G. 2002. Untranslated regions of mRNAs. *Genome Biol.* **3:** doi: 10.1186/gb-2002-3-3-reviews0004.

Mourier, T., Pain, A., Barrell, B., and Griffiths-Jones, S. 2005. A selenocysteine tRNA and SECIS element in *Plasmodium falciparum*. *RNA* **11:** 119–122.

Nakamura, Y., Gojobori, T., and Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.* **28:** 292. doi: 10.1093/nar/28.1.292.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** e33. doi: 10.1371/journal.pcbi.0020033.

Prasanth, K.V. and Spector, D.L. 2007. Eukaryotic regulatory RNAs: An answer to the "genome complexity" conundrum. *Genes & Dev.* **21:** 11–42.

Rathjen, T., Nicol, C., McConkey, G., and Dalmay, T. 2006. Analysis of short RNAs in the malaria parasite and its red blood cell host. *FEBS Lett.* **580:** 5185–5188.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi,

R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16:** 11–19.
Schattner, P., Decatur, W.A., Davis, C.A., Ares Jr., M., Fournier, M.J., and Lowe, T.M. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: Analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* **32:** 4281–4296. doi: 10.1093/nar/gkh768.
Schattner, P., Brooks, A.N., and Lowe, T.M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33:** W686–W689.
Smyth, G.K. and Speed, T. 2003. Normalization of cDNA microarray data. *Methods* **31:** 265–273.
Smyth, G.K., Michaud, J., and Scott, H.S. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21:** 2067–2075.
Steigele, S., Huber, W., Stocsits, C., Stadler, P.F., and Nieselt, K. 2007. Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biol.* **5:** 25.
Storz, G. 2002. An expanding universe of noncoding RNAs. *Science* **296:** 1260–1263.
Storz, G., Altuvia, S., and Wassarman, K.M. 2005. An abundance of RNA regulators. *Annu. Rev. Biochem.* **74:** 199–217.
Szymanski, M. and Barciszewski, J. 2006. RNA regulation in mammals. *Ann. N. Y. Acad. Sci.* **1067:** 461–468.
Trager, W. and Jensen, J.B. 2005. Human malaria parasites in continuous culture. 1976. *J. Parasitol.* **91:** 484–486.
Upadhyay, R., Bawankar, P., Malhotra, D., and Patankar, S. 2005. A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **144:** 149–158.
Voss, T.S., Healer, J., Marty, A.J., Duffy, M.F., Thompson, J.K., Beeson, J.G., Reeder, J.C., Crabb, B.S., and Cowman, A.F. 2006. A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. *Nature* **439:** 1004–1008.
Washietl, S. and Hofacker, I.L. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342:** 19–30.
Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. 2005a. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23:** 1383–1390.
Washietl, S., Hofacker, I.L., and Stadler, P.F. 2005b. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci.* **102:** 2454–2459.
Washietl, S., Pedersen, J.S., Korbel, J.O., Stocsits, C., Gruber, A.R., Hackermuller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **17:** 852–864.

Received June 23, 2007; accepted in revised form October 23, 2007.