Evolutionary Bioinformatics

ORIGINAL RESEARCH

# In Silico Prediction of Evolutionarily Conserved GC-Rich Elements Associated with Antigenic Proteins of Plasmodium falciparum

Porkodi Panneerselvam[2,4], Praveen Bawankar[1,5], Surashree Kulkarni[3,6] and Swati Patankar[1]

[1]Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India. [2]Centre for Biotechnology, Anna University, Sardar Patel Road, Guindy, Chennai 600025, India; [3]St. Xavier's College, 5, Mahapalika Marg, Mumbai 400001, India; [4]Present address: Computational and Systems Biology, Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576; [5]Present address: Department of Biochemistry, Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, D-72076 Tuebingen, Germany; [6]Present address: Department of Biological Sciences, University of Wisconsin, Milwaukee, WI 53211, USA.
Corresponding author email: patankar@iitb.ac.in

**Abstract:** The *Plasmodium falciparum* genome being AT-rich, the presence of GC-rich regions suggests functional significance. Evolution imposes selection pressure to retain functionally important coding and regulatory elements. Hence searching for evolutionarily conserved GC-rich, intergenic regions in an AT-rich genome will help in discovering new coding regions and regulatory elements. We have used elevated GC content in intergenic regions coupled with sequence conservation against *P. reichenowi*, which is evolutionarily closely related to *P. falciparum* to identify potential sequences of functional importance. Interestingly, ~30% of the GC-rich, conserved sequences were associated with antigenic proteins encoded by *var* and *rifin* genes. The majority of sequences identified in the 5′ UTR of *var* genes are represented by short expressed sequence tags (ESTs) in cDNA libraries signifying that they are transcribed in the parasite. Additionally, 19 sequences were located in the 3′ UTR of rifins and 4 also have overlapping ESTs. Further analysis showed that several sequences associated with *var* genes have the capacity to encode small peptides. A previous report has shown that upstream peptides can regulate the expression of *var* genes hence we propose that these conserved GC-rich sequences may play roles in regulation of gene expression.

**Keywords:** *Plasmodium*, regulatory elements, comparative genomics, genome bias, antigenic variation

*Evolutionary Bioinformatics* 2011:7 235–255

doi: 10.4137/EBO.S8162

This article is available from http://www.la-press.com.

## Introduction

Regulatory motifs that allow fine-tuning of gene expression are of interest in the malaria parasite *Plasmodium falciparum*. These include promoters, mRNA stability motifs and translation regulatory sequences. Some regulatory motifs also encode non-coding RNAs (ncRNAs) that in turn regulate expression of genes. The importance of regulatory motifs cannot be underestimated in the parasite since mechanisms of regulation of gene expression are still being elucidated in this human pathogen.[1–3] The comparative genomics approach has been successfully employed to identify evolutionarily well-conserved regulatory elements in *C. elegans, S. cerevisiae* and *Homo sapiens*.[4–6] This is based on the rationale that functionally important sequences are often conserved among species. Comparative genomics has also been used in *Plasmodium* species to identify regulatory motifs.[7]

Another feature of the *Plasmodium falciparum* genome that has proved useful in the search for new regulatory elements has been nucleotide bias. *Plasmodium falciparum* has an unusually AT-rich genome,[8] with an average AT content of 80% that increases to 90% in intergenic regions. In such a biased genome, local regions of increased GC content in the non-coding regions appear to correlate with functionally important features. For example, a conserved, GC-rich region found upstream of heat shock protein (*hsp*) genes is a functionally important DNA regulatory element.[9,10] In two reports including one from our group, noncoding RNAs (ncRNAs) were identified in *Plasmodium falciparum* based on searching for conserved GC-rich intergenic regions.[10,11] Similarly, nucleotide compositional contrast has been used to identify ncRNA in the AT rich genome of *Dictyostelium discoideum* and hyperthermophiles.[12–14] This type of screen exploited the fact that most RNA regulatory elements carry out their functions by inter-molecular or intra-molecular base pairing; hence an increase in GC content especially in an AT-rich genome would result in RNAs having more stable secondary structures.[15] Most of these reports also used comparative genomics and evolutionary conservation as a tool to assess functional significance.

The choice of genomes used for comparative genomics is critical. In a bioinformatics screen described previously,[11] since the complete genome of *P. yoelii* was available we chose this species for identifying conserved, GC-rich intergenic regions that were shown to encode ncRNAs. However, with the recent availability of other *Plasmodium* genomes, it is likely that other genomes might be equally useful for comparative genomics. Indeed, *P. yoelii* and *P. falciparum* appear to have diverged >100 million years ago[16] however, *P. falciparum* has been shown to be most closely related to the chimpanzee malaria parasite *P. reichenowi*.[17–19] Apart from housekeeping genes, several ORFs that encode cell surface proteins in *P. falciparum* are conserved between *P. falciparum* and *P. reichenowi*; these include CSP,[20] MSP2[21] and var CSA.[22] In contrast, the *var*, *rifin* and *stevor* multigene families that are involved in antigenic variation in *P. falciparum* are represented by a single multigene family (*yir*) in *P. yoelii* that is most closely related to the *vir* family in *P. vivax*.[23,24] Over the entire genome, *P. yoelii* is most closely related to the other rodent malaria parasites *P. berghei* and *P. chabaudi*.[25]

In this report we ask whether regulatory elements can be identified by a bioinformatics screen using elevated GC content in the *P. falciparum* genome, followed by sequence conservation in other *Plasmodium* species. Due to the large evolutionary distance between *P. falciparum* and *P. yoelii*, we hypothesized that the choice of these two genomes for comparative genomics may not identify regulatory elements associated with immunogenic genes that are specifically expressed in *P. falciparum* and not in *P. yoelii*. Hence for identification of genomic sequences that might be involved in host-specific functions eg, evasion from the immune system or regulation of antigenic variation genes, a primate malaria parasite genome would be more appropriate for the comparative genomics part of any bioinformatics screen.

We show that a large number of GC-rich sequences are conserved in the genomes of *P. falciparum* and the primate parasite *P. reichenowi*. Many of these GC-rich sequences flank genes involved in antigenic variation and some may be transcribed and translated. Several reports in the literature show that short RNAs can regulate transcription[26] and short ORFs can regulate translation of downstream genes.[27,28] Indeed, one of these reports shows that an upstream ORF regulates expression of certain *var* genes.[29] We suggest that the sequences identified in this study may play roles in regulation of antigenic gene expression at the level of transcriptional or translational control.

## Materials and Methods

### GC% filter source data

The genome of *Plasmodium falciparum* 3D7 was downloaded chromosome wise from the online database (http://www.plasmodb.org/). Exon locations of all protein coding genes were also downloaded from the same database. Due to the unavailability of exon location data in the new version—PlasmoDB 5.2, all the data were downloaded from the older version PlasmoDB 4.4.

### GC% C program algorithm

A C program was written which reads large text files of the *Plasmodium falciparum* genome. The program divides the genome into 70 base chunks with the sliding window of 10 bases. It uses exon location data and excludes those chunks which fall within ORFs. The GC% of each chunk was then calculated. An output FASTA file was generated with the sequences of all 70 base chunks with greater than 35% GC according to the sliding window model and lying outside ORFs. If any 70 base chunks with greater than 35% GC were overlapping, these were combined and treated as a single sequence. All such 70 base chunks were associated with their chromosomal locations; note that since overlapping chunks were merged together, some regulatory elements are greater than 70 bases.

### Sequence Conservation Source Data

The genome contigs of *Plasmodium* species *viz. P. yoelii, P. vivax*, *P. reichenowi, P. berghei, P. gallinaceum, P. knowlesi* and *P. chaubadi* were downloaded from PlasmoDB 5.2. The Washington University BLAST version 2.0 (WU-BLAST) downloaded from http://www.blast.wustl.edu/ was employed to analyze sequence conservation. This BLAST version was installed on a Linux machine.

### Shell Script

A shell script was written which took each sequence from the output FASTA file containing sequences having GC content greater than 35% and fed it into the BLAST software. It performs BLAST of all chunks in each of the query files with all the available contigs in the database file. The *E* value cut-off was set as 1e-10.

Positive controls for the above strategy were rRNA, tRNA and the sequences identified with *Plasmodium yoelii* earlier by Upadhyay et al. In short, after running the BLAST analysis of GC-rich sequences using different genomes, we checked whether the 43 annotated tRNAs, 27 annotated rRNAs and 18 ncRNA sequences identified by Upadhyay et al were correctly identified.

## Results and Discussion

### Use of the *P. reichenowi* genome for comparative genomics can identify novel GC-rich conserved sequences

Previous work in our lab had used a bioinformatics strategy to identify GC-rich sequences present in intergenic regions that were conserved between *P. falciparum* and *P. yoelii*. This screen used two cut-offs (35% GC followed by an *E* value cut-off of 1e-10) and identified 18 sequences, many of which were found to be small molecular weight RNAs also known as non-coding RNAs (ncRNAs). These cut-offs were appropriate in searching for ncRNAs since we were able to identify all 43 annotated tRNAs and 27 annotated rRNAs from the *P. falciparum* genome.

We hypothesized that using the same strategy but with different genomes for the comparative genomics part of the screen might give more GC-rich, conserved sequences that are associated with host-specific functions. These sequences might be regulatory DNA sequences, ncRNAs or protein-encoding regions. To ensure that the 35% GC cut-off was appropriate for identifying such regulatory sequences, and particularly to be sure that the probability of finding the GC-rich sequence was greater than chance, we did a simple statistical analysis. The average GC content of the 23 megabase *P. falciparum* genome (19%) was compared to the GC content of the 70 base chunks used in the screen (35%) with a Chi-square test using Minitab software. The *P* value of this test was 0.0003, indicating that the probability of finding a 35% GC-rich sequence of 70 bases in the *P. falciparum* genome, is very low. Hence, any GC-rich sequences identified should be significantly different from the genome in their nucleotide content. We proceeded to test our hypothesis that sequences greater than 35% GC-rich and conserved in other *Plasmodium* species might be regulatory sequences associated with host-specific functions.

To test this, we initially performed the bioinformatics screen using only chromosome 1 of *P. falciparum*. This screen retained the original parameters of GC threshold and BLAST cutoff (>35% GC rich and BLAST value of e-10), however the BLAST analysis was performed against seven *Plasmodium* species—*P. yoelii*, *P. reichenowi*, *P. berghei*, *P. vivax*, *P. gallinaceum*, *P. knowlesi* and *P. chabaudi*. For all genomes except *P. reichenowi*, no new GC-rich, conserved sequences were identified. Interestingly, eighty-five new sequences could be identified when the screen involved comparison with the chimpanzee parasite, *P. reichenowi*. No new sequences were identified when BLAST was performed against the macaque parasite *P. knowlesi* and the human parasite *P. vivax*. This is consistent with reports that *P. knowlesi* falls in the same phylogenetic group as *P. vivax*.[19] Hence, *P. reichenowi* was chosen as the most appropriate genome to do the comparative analysis for identifying regulatory elements in *P. falciparum.*

## Proximal Intergenic Sequences

The bioinformatics screen was repeated using the entire *P. falciparum* genome to identify GC-rich sequences with a cutoff of 35% GC; these sequences were compared for conservation against the complete *P. reichenowi* genome (BLAST value of e-10) yielding ~1500 conserved GC-rich regions. In order to further prioritize these sequences an additional parameter was applied. This parameter restricted the output to sequences that lie within 500 bases of the start or stop codons of annotated ORFs (termed proximal intergenic regions). The rationale was that a majority of DNA regulatory elements and translational control elements are generally found within 500 bases of the start or stop codons of flanking genes. Hence we decided to sort out sequences that could lie within 5′ or 3′ UTRs of *P. falciparum* genes. Very few *P. falciparum* UTRs have been annotated, nevertheless Watanabe et al conclude from their analysis of a cDNA library that the 5′ UTRs of *P. falciparum* genes are unusually long, averaging 346 bp.[30] Golightly et al report a 3′ UTR of 450 bp in the mRNA of Pgs28, an ookinete protein of the avian parasite *P. gallinaceum*.[31] Hence, we defined all the intergenic sequences within 500 bp of the coding region as 'proximal intergenic sequences'. Those intergenic sequences, which lie greater than 500 bp from the coding sequence, were designated as 'deep intergenic sequences'. Concurrently, Neafesy et al has suggested that conserved CpG dinucleotides enriched in proximal intergenic regions might function as regulatory elements.[32] With these criteria in mind, ~1500 new GC-rich sequences that were identified during the bioinformatics analysis described in this report were pruned down to 151 by screening for proximal intergenic sequences (see Supplementary Table 1).

## Immunogenic Proteins are Conserved in P. falciparum and P. reichenowi

Having shown that 151 sequences that are GC-rich and present in intergenic regions are conserved between *P. falciparum* and *P. reichenowi*, we wished to test our hypothesis that these might be associated with antigenic genes that are found in these two species. As a first step towards this, we tested whether families of antigenic genes found in *P. falciparum* are also present in *P. reichenowi*.

A comparison of the chimpanzee's genetic blueprints with that of the human genome shows that our closest living relatives share 96 percent of our DNA. Humans and chimps originate from a common ancestor, and scientists believe they diverged some six million years ago.[33] Interestingly the human malaria parasite *P. falciparum* diverged from the chimpanzee malaria parasite *P. reichenowi* around 5–7 million years ago[17,34] suggesting that the primate parasites may have diverged at the same period when their hosts diverged.

Several studies have shown that *P. falciparum* is most closely related to *P. reichenowi*.[20,21] This is true not only for housekeeping genes but also for genes that encode proteins involved in host-parasite interactions. These include some of the *var* genes that encode the PfEMP family of proteins important for antigenic variation and evasion of the host immune response. Indeed, Trimnell et al[22] have shown that fragments of the var1CSA and var2CSA genes are conserved between *P. falciparum* and *P. reichenowi* suggesting an ancient origin of some *var* loci. Like *P. falciparum*, *P. reichnowi* is also shown to express key invasion proteins like EBLs and MAEBLs.[35,36] To further test the extent of relatedness of the parasites, an analysis was done for other genes involved in antigenic variation. Antigenic proteins of *P. falciparum* involved in host pathogen interactions were chosen and BLAST analysis of the genes was performed with *P. reichenowi* contigs (PlasmoDB BLAST server—blastn). Two genes were chosen at random from each of the PfEMP, *rifin* and *stevor* families of

**Table 1.** BLAST analysis of antigenic proteins.

| Gene ID and the gene product | No of hits with *P. yoelii* | *E* value of best hit with *P. yoelii* | No of hits with *P. reichenowi* | *E* value of best hit with *P. reichenowi* | Best hit with *P. reichenowi* |
|---|---|---|---|---|---|
| MAL13P1.1 PfEMP1 | 2 | 0.027 | 109 | 2e-87 | Pr_3502696.c000023469.Contig1 |
| PF07_0051 PfEMP1 | 27 | 3e-5 | 107 | 7e-54 | Pr_3502696.c000023041.Contig1 |
| MAL13P1.2 RIFIN | 2 | 0.017 | 194 | e-128 | Pr_3502696.c000027339.Contig1 |
| PFF0850c RIFIN | 0 | – | 37 | 2e-95 | Pr_3502696.c000023726.Contig1 |
| MAL13P1.505 STEVOR | 1 | 0.015 | 35 | e-140 | Pr_3502696.c000023791.Contig1 |
| PFI0045c STEVOR | 4 | 0.014 | 34 | e-136 | Pr_3502696.c000023791.Contig1 |

**Note:** Two members of the PfEMP, rifin and stevor families were chosen arbitrarily from the *P. falciparum* genome and BLAST was performed against the genomes of *Plasmodium yoelii* and *Plasmodium reichenowi*.

antigenic surface proteins and the *P. yoelii* genome was used for comparison. Table 1 shows the results of this analysis.

Except for the *var* gene PF07_0051 there were fewer than 5 matches to the *P. yoelii* genome with the antigenic genes tested. PF07_0051 showed 27 matches with a best *E* value of 3e-5 indicating that this *var* gene may have weak homology to sequences in the *P. yoelii* genome. This is consistent with the data that there have been no genes showing homology to the *var* gene family in reports on *P. yoelii* genome analysis.[8] Instead, the *P. yoelii* genome contains a multigene family (*yir*) that shows homology to the *P. vivax vir* multigene family.[23,24] In contrast, 34–194 matches of the *var*, *rifin* and *stevor* genes were obtained by using BLAST against the *P. reichenowi* genome and these matches gave extremely low *E* values (*E* value $<$ e-140) indicating that the sequences are highly conserved. The high numbers of matches obtained (eg, 194 with a *rifin* gene) indicate that *P. reichenowi* also has three different families of antigenic proteins like *P. falciparum*. Hence the data suggests that the *P. falciparum* genome is more similar to the genome of *P. reichenowi* than *P. yoelii* when antigenic variation genes are analyzed.

## Sequences Proximal to var Genes

Having shown that antigenic variation genes are conserved in *P. falciparum* and *P. reichenowi* and that 151 GC-rich sequences are also conserved in the two genomes, the next question was whether these GC-rich sequences flanked antigenic variation genes.

As mentioned in the previous section, sequestration and rosetting are key determinants of *P. falciparum* pathogenesis and these processes are mediated by the *var* gene family called *Plasmodium falciparum* Erythrocyte Membrane Proteins 1 (PfEMP1). To evade immunity and extend infections, parasites clonally vary the PfEMP1 proteins that are expressed on the surface of the infected red blood cells.[37] Mechanisms of regulation of *var* genes have been a topic of intense research due to the clinical importance of these genes.[38,39] Expression of *var* genes is regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron.[40] Upstream promoters of *var* genes fall into four major sequence classes: upsA, upsB, upsC and upsE[41] of which upsA- upsB- and upsE type *var* genes lie in sub-telomeric regions and upsC-type *var* genes lie in internal clusters. Recent evidence indicates that *var* genes are activated by recruitment of the promoter to a perinuclear site that is permissive for transcription[42] and also that the PfSIR2 regulator plays a role in *var* gene silencing.[43,44] Recent studies indicate that ncRNAs associate with chromatin and thus regulate the expression of *var* gene family.[45] Additionally, an upstream ORF can regulate certain *var* genes.[29]

Interestingly, the BLAST result with *P. reichenowi* showed that 27 of the proximal intergenic GC-rich sequences flank *var* genes (listed in Table 2). All these sequences lie in the 5′ UTR of the flanking *var* genes and most are less than 20 bp away from the predicted ORF of PfEMP1 proteins. The close proximity of the GC-rich sequences to the *var* ORF led us to wonder whether these sequences might be transcribed either as short RNAs or as part of the *var* mRNA transcripts.

A search of PlasmoDB revealed that the Sugano malaria cDNA library[30,46,47] has identified several short transcripts (ESTs AU088275 and AU087013) in the 5′ UTRs of *var* genes. An analysis of the GC-rich sequences that are proximal to *var* genes showed that all except the PfNC4.4var overlap with at least one of the two ESTs AU088275 and AU087013. The two ESTs are transcribed from the same strand as the PfEMP1 mRNA and AU088275 and AU087013 showed alignment with 30 and 16 regions of the *P. falciparum* genome respectively. This bioinformatics study was able to identify 23 out of 30 and 10 out of 16 regions in the case of AU088275 and AU087013 respectively. The GC-rich sequences that were not identified in this study are less conserved compared to *P. reichenowi* and hence did not show up after the BLAST with a cut off of 1e-10. The presence of short transcripts that overlap with the GC-rich sequences

**Table 2.** Conserved GC rich sequence associated with *var* genes.

| Candidate | Location | PfEMP1 Associated | GC% | Identity | Associated ESTs |
|---|---|---|---|---|---|
| PfNC1.1var | Chr 1: 29631 to 29730 | PFA0005w | 37 | 58/100 | AU088275 |
| PfNC1.2var | Chr 1: 616621 to 616710 | PFA0765c | 38.9 | 33/90 | AU088275 and AU087013 |
| PfNC2.1var | Chr 2: 25101 to 25230 | PFB0010w | 40.8 | 56/130 | AU087013 and AU088275 |
| PfNC2.2var | Chr 2: 923651 to 923750 | PFB1055c | 42 | 58/100 | AU087013 and AU088275 |
| PfNC3.1var | Chr 3: 33511 to 33640 | PFC0005w | 38.5 | 72/130 | AU088275 |
| PfNC3.2var | Chr 3: 1034931 to 1035030 | PFC1120c | 41 | 42/100 | AU088275 |
| PfNC4.1var | Chr 4: 35061 to 35150 | PFD0005w | 46.7 | 32/90 | AU088275 |
| PfNC4.2var | Chr 4: 606841 to 606930 | PFD0635c | 42.2 | 38/90 | AU088275 |
| PfNC4.3var | Chr 4: 970091 to 970160 | PFD1005c | 35 | 34/70 | AU088275 |
| PfNC4.4var | Chr 4: 981221 to 981290 | PFD1015c | 37 | 36/70 | – |
| PfNC4.5var | Chr 4: 1183861 to 1183950 | PFD1245c | 45.6 | 31/90 | AU088275 |
| PfNC6var | Chr 6: 3401 to 3500 | PFF0010w | 42 | 38/100 | AU088275 |
| PfNC7.1var | Chr 7: 30531 to 30670 | MAL7P1.212 | 37.9 | 81/140 | AU088275 and AU087013 |
| PfNC7.2var | Chr 7: 605971 to 606040 | MAL7P1.50 | 37 | 35/70 | AU088275 |
| PfNC7.3var | Chr 7: 614461 to 614570 | PF07_0050 | 40 | 38/110 | AU088275 |
| PfNC7.4var | Chr 7: 644311 to 644440 | MAL7P1.55 | 41.5 | 43/130 | AU087013 |
| PfNC8.1var | Chr 8: 22251 to 22330 | PF08_0142 | 41.3 | 41/80 | AU087013 |
| PfNC8.2var | Chr 8: 441381 to 441450 | PF08_0106 | 38 | 34/70 | AU087013 |
| PfNC8.3var | Chr 8: 1399241 to 1399340 | MAL8P1.220 | 38 | 38/100 | AU088275 and AU087013 |
| PfNC9.1var | Chr 9: 19931 to 20070 | PFI0005w | 40.7 | 92/140 | AU088275 |
| PfNC9.2var | Chr 9: 1503331 to 1503430 | PFI1830c | 37 | 38/100 | AU088275 |
| PfNC10var | Chr 10: 28351 to 28490 | PF10_0001 | 36.4 | 76/100 | AU088275 |
| PfNC11var | Chr 11: 24021 to 24150 | PF11_0007 | 40 | 67/130 | AU088275 |
| PfNC12.1var | Chr 12: 32601 to 32670 | PFL0020w | 37.1 | 41/70 | AU088275 |
| PfNC12.2var | Chr 12: 774191 to 774300 | PFL0935c | 38.2 | 53/110 | AU088275 and AU087013 |
| PfNC12.3var | Chr 12: 1704411 to 1704490 | PFL1955w | 46.3 | 36/80 | AU088275 and AU087013 |
| PfNC12.4var | Chr 12: 2248951 to 2249040 | PFL2665c | 45.6 | 56/90 | AU088275 |

**Note:** All candidates are found in the 5′ UTRs of *var* genes and are within 150 bases of the start codon.

identified in this bioinformatics screen suggests that indeed these sequences are transcribed.

PfNC4.4var was the only sequence with no associated ESTs and this sequence lies 190 bases away from the annotated PfEMP1. A BLAST was performed with the sequence of PfNC4.4.var against the genome of *P. falciparum* and we identified 6 matches that were all proximal to PfEMP1 genes. To test whether any short RNAs are associated with the sequence PfNC4.4var we performed Northern analysis on mixed stage asexual parasites using strand-specific probes. These results indicate that the sequence is not expressed in mixed stage asexual parasites (data not shown); perhaps the expression of this sequence is below the limit of detection by Northern analysis or is stage-specific. Alternatively the sequence may function as a DNA regulatory element rather than as RNA or may be involved in translational control of the flanking *var* gene.

The sequences of the ESTs AU088275 and AU087013 were compared with each other and with the sequence PfNC4.4var using ClustalW (http://www.ebi.ac.uk/clustalw/). The scores obtained show that the ESTs AU088275 and AU087013 are 68% similar to each other at the sequence level while the sequence PfNC4.4var is quite distinct from either of these ESTs showing 25%–32% sequence similarity in the ClustalW analysis. Further analysis of the ESTs showed that AU088275 and AU087013 are in the 5′ UTRs of *var* genes of the upsB or upsBsh subtypes while sequence PfNC4.4var is found in the 5′ UTRs of 7 *var* genes of the upsC subtype.

Having shown that the GC-rich sequences that flank *var* genes are found in short transcripts, we next asked whether these sequences have the capacity to encode proteins, either as upstream ORFs (uORFs) or as N-terminal extensions of the annotated *var* genes. Indeed, a majority of the GC-rich sequences showed the presence of upstream ORFs (uORFs) ranging in size from minimal ORFs (1 amino acid) to 21 amino acids. Several of the uORFs are found in a majority of the GC-rich regions (pentapeptide MYATI found 20 times) and others are found less frequently (MYQNTTKPCMPRYKPRMHDIM found once).

Interestingly, when all the GC-rich sequences that flank *var* genes were aligned with each other, it was noticed that the most conserved sequences (highlighted in grey with asterisks), encoded the uORF pentapeptide MYATI (Fig. 1). In contrast, sequence conservation was poor in the regions surrounding the uORF. This suggests an evolutionary pressure to maintain the uORF encoding sequences indicating these sequences may have functional importance. A sequence alignment between *var*-associated GC-rich sequences of *P. falciparum* and *P. reichenowi* (Fig. 2) shows a significant sequence similarity between PfNC12.4var and the homologous region from *P. reichenowi* and the uORF MYATI is conserved between the two species.

uORFs have been shown to play important roles in translational control. For example, a minimal uORF can regulate translation of certain HIV genes.[48] This minimal ORF (consisting of only a start and a stop codon) overlaps with the start codon of the *vpu* gene and mutating the start and stop codons of this minimal ORF results a reduction of translation of the downstream *env* gene. Upstream AUGs and uORFs in human and rodent genes appear to regulate translation initiation by the ribosome scanning machinery.[27] Finally, and most pertinently for this work, the presence of uORFs has been shown to regulate the expression of the downstream *var* gene.[29] We propose that the uORFs identified in this report flank *var* genes at the 5′ regions and may play similar roles in regulation of *var* gene expression.

## Sequences Proximal to *rifin* Genes

*Rifin* genes constitute the largest multi-gene family in the *P. falciparum* genome with 149 members. Transcription from *rifin* genes is highest at the rings and early trophozoite stages and proteins encoded by these mRNAs are localized to the Maurer's clefts.[49,50] Presence of antibodies against RIFINS in patient sera suggests that these proteins are indeed exposed on the surface of erythrocytes.[51] More recently, the discovery of a PEXEL/VTS transport signal found in proteins exported from the parasite vacuole to the erythrocyte was observed in RIFIN proteins and is consistent with a potential cell surface localization.[52,53] The function of RIFINS is unknown however these proteins may be involved in cytoadherence. Similar to *var* genes, *rifin* genes are also clonally variable although the mechanisms underlying the two processes appear to be different.

A search of the proximal intergenic GC-rich sequences obtained in our screen of the *P. falciparum* genome shows that 19 sequences flank *rifin* genes. The list of sequences is shown in Table 3. All the sequences

```
CLUSTAL 2.1 multiple sequence alignment

PfNC1.1var       ----ACATACATACATACAT---ACACCAA-----------ACCAAACCATGTATGCCACGATATAAACCACGTAC------CACGTATGACATAATGTAG--TCA----TGAATAACC-
PfNC12.2var      ----ACATACATACATACAT---ACCCCTACCAA--ACACCACCAAACCATGTATGCCACGATATAAACCACGTATGT-ATGCATGTATGACATAATGTAG---------TGCACGGAC-
PfNC4.5var       ----ACATACATACAATCAC---CCCACACCACACCACACCACCAAACCATGTATGCCACGATATAAACCACGTATACCACGTATGCATGACATAATGTAG--TGC---ACGGACAACC-
PfNC10var        ----ACATACATACAATCAC---CCCACACCACACCACACCACCAAACCATGTATGCCACGATATAAACCACG---------TATGCATGACATAATGTAG--TCCGAAACAATAAAAC-
PfNC7.4var       ----ACATACG--CAATACG---CCACCACCACCGCCCACA--CGAACCATGCATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAG--TGG----TGGTGTT---
PfNC9.1var       ----ACATACG--CAATACG---CCACCACCACCGCCCACA--CTTACCATGTATGCCACGATATAAACCACGTATGT-ATGCATGTATGACATCATGTAG--TGG----TGGAGTTAAC
PfNC7.1var       ----ACATACG--CAATACG---CCACCGCCACCGCCAACA--CAAACCATGTATGCCACGATATAAACCACGTATG-----TATGTATGACATAATGTAG--TCG----GGAAGAAGAA
PfNC11var        ----ACATACG--CAATACA---CCACCACCACCGCCCACA--CGAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAG--TGC----ACCAATAACG
PfNC4.1var       ----ACATACG--CAATACG---CCACCACCACACCACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATCATGTAG--TCG----TGAACAA---
PfNC4.2var       ----ACATACAT-ACACCCA---CGTACGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAG--TGCAC---CAATAACG-
PfNC6var         ----ACATACAT-ACCCCCA---CGTACGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAG--TGCAC---CAATAACCA
PfNC3.2var       ----ACATATAT-ACCCCCA---CGTACGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----TATG----ACATAATGTAG--TGCACGAACGATAAAC-
PfNC9.2var       -CATACATACAT-ATATACA---CGTATGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAG--TGCACGAAAGATAAAC-
PfNC7.3var       ----ACATACAT-ACCCCCA---CGTACGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAGTCTGGAAGAAGAAGAATAC
PfNC8.3var       ----ACATATAT-ACCCCCA---CGTACGTACCAAAACACCACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATG----ACATAATGTAGTCATGAATAA-------CC
PfNC3.1var       ----ACATACAATCACCCCA---CACCACACCACACCACC----AAACCATGTATGCCACGATATAAACCACGTAT---------GTATGACATCATGTTG--TCG--------GTACA-
PfNC12.4var      -----CATACAATCACCCCA---CACCACACCACACCACCTACCAAACCATGTATGCCACGATATAAACCACGTATG-----CATGTATGACATCATGTTG--TCG--------GTACA-
PfNC2.1var       ACATACATACAATCACCCCA---CACCACACCACACCACC----AAACCATGTATGCCACGATATAAACCACGTATGT-ATGCATGTATGACATCATGTTG--TCG--------CAACC-
PfNC12.3var      -ACCCCTACCAAACACCTAC---CACTCCACCGCCCACAC----GAACCATGTATGCCACGATATAAACCACGTATG-----TATGTATGACATCATGTTG--TCG--------CAACC-
PfNC2.2var       ----ACATACATACCCCCACGTACGTACCAAAACACCACC----AAACCATGTATGCCACGATATAAACCACGTATG-----CATGTATGACATCATGTTG--TCG--------CAACC-
                      *::. .  .    .    .                     :****** ********************            *    ****.****:*
                                                                  M Y A T I

PfNC1.1var       ---AAAATGGTG 98
PfNC12.2var      ---AAAATGGTG 109
PfNC4.5var       ---ACAATGGCG 116
PfNC10var        ---AAAATGGCG 110
PfNC7.4var       ---AAAATGGCG 100
PfNC9.1var       --AAAAATGGGG 112
PfNC7.1var       TAAAAAATGGCG 110
PfNC11var        ---AAAATGGAG 103
PfNC4.1var       -----AATGGTG 100
PfNC4.2var       ---AAAATGGCG 106
PfNC6var         ---AAAATGGCG 107
PfNC3.2var       ---AAAATGGGT 109
PfNC9.2var       ---AAAATGGCT 112
PfNC7.3var       ---AAAATGGCG 112
PfNC8.3var       ---AAAATGGTG 105
PfNC3.1var       ------ATGGTT 95
PfNC12.4var      ------ATGGGG 102
PfNC2.1var       ------ATGGCG 107
PfNC12.3var      ------ATGGCG 102
PfNC2.2var       ------ATGGGG 102
                       ****
```

Annotated
start codon

**Figure 1.** Sequence alignment of proximal upstream regions of upsB var genes.
**Notes:** The box shows that conserved GC-rich sequences contain the putative upstream ORF MYATI. Grey highlights show-conserved sequences, indicating that sequences flanking the putative uORF are less conserved than the regions encoding the uORF. The annotated start codon is highlighted in grey.

except for one (PfNC10.1rif) lie in the 3′ UTR of *rifin* genes and are 1 to 500 bases away from the stop codon of the *rifin* open reading frame. PfNC10.1rif is located in the 5′ UTR of *rifin* gene PF10_0002w. Four of the GC-rich regions that flank *rifins* are associated with short ESTs (BI816203 and BQ577081) and all the ESTs are transcribed from the same strand as the *rifin* gene. There is a paucity of information regarding regulation of *rifin* gene expression. A recent study has mapped promoter elements that are required for expression of one *rifin* gene (PF11_0009) that is highly expressed in 3D7 parasites.[54] The promoter elements include two repressor regions that are bound by nuclear proteins expressed at different stages of the parasite life cycle. While 5′ flanking sequences are essential for transcriptional regulation, it is tempting to speculate that events in the 3′ UTRs of *rifin* genes, particularly the GC-rich sequences discovered in this study may play roles in gene regulation.

```
PfNC12.4var                        ACCAAACCATGTATGCCACGATATAAACCACGTATG----CATGTATGA
Pr_3502696.c000023441.Contig1      ACCAAACCATGTATGCCACGATATAAACCACGTATGTATGCATGTATAA
                                   ************************************    ******* *

                                            M   Y   A   T   I

PfNC12.4var                        CATCATGTTGTCGG
Pr_3502696.c000023441.Contig1      CATCATGCTGTCGG
                                   ******* ******
```

**Figure 2.** BLAST result of PfNC12.4var against the *Plasmodium reichenowi* genome.
**Note:** The regions of conservation are shown with stars and the uORF is highlighted in grey.

**Table 3.** Conserved GC rich regions associated with *rifin* genes.

| S.no | Candidate | Associated RIFIN | GC% | Identity | Associated ESTs |
|---|---|---|---|---|---|
| PfNC1.1rif | Chr 1: 62341 to 62410 | PFA0045c | 35% | 36/70 | – |
| PfNC1.2rif | Chr 1: 81921 to 81990 | PFA0080c | 38% | 55/70 | – |
| PfNC2rif | Chr 2: 32951 to 33020 | PFB0015c | 35% | 36/70 | – |
| PfNC3rif | Chr 3: 1025611 to 1025680 | PFC1115w | 35% | 34/70 | BI816203 |
| PfNC4rif | Chr 4: 67831 to 67940 | PFD0025w | 37.3% | 78/110 | – |
| PfNC6rif | Chr 6: 1352101 to 1352170 | PFF1575w | 37% | 50/70 | – |
| PfNC7.1rif | Chr 7: 45441 to 45540 | MAL7P1.215 | 35% | 79/100 | – |
| PfNC7.2rif | Chr 7: 55261 to 55330 | MAL7P1.217 | 37% | 45/70 | BQ577081 |
| PfNC7.3rif | Chr 7: 1454751 to 1454820 | PF07_0134 | 37% | 47/70 | – |
| PfNC9.1rif | Chr 9: 42361 to 42460 | PFI0025c | 34% | 80/100 | – |
| PfNC9.2rif | Chr 9: 1479191 to 1479290 | PFI1810w | 35% | 81/100 | – |
| PfNC10.1rif | Chr 10: 39021 to 39090 | PF10_0002w | 35.7% | 89/100 | – |
| PfNC10.2rif | Chr 10: 47981 to 48050 | PF10_0005 | 35.7% | 88/100 | – |
| PfNC10.3rif | Chr 10: 1623881 to 1623950 | PF10_0398 | 35.7% | 98/100 | – |
| PfNC12.1rif | Chr 12: 43711 to 43790 | PFL0025c | 33.8% | 92/100 | BQ577081 |
| PfNC12.2rif | Chr 12: 2239401 to 2239480 | PFL2660w | 35% | 87/100 | – |
| PfNC13.1rif | Chr 13: 30631 to 30700 | MAL13P1.2 | 37.1% | 94/100 | BQ577081 |
| PfNC13.2rif | Chr 13: 53591 to 53670 | PF13_0006 | 40% | 95/100 | – |

**Note:** All candidates are found in the 3′ UTRs of *rifin* genes.

## Conclusion

In conclusion, this report shows that a bioinformatics strategy involving a search for GC-rich intergenic regions that are conserved between *P. falciparum* and *P. reichenowi* can be used to uncover conserved GC-rich sequences proximal to antigenic variation genes. These sequences are transcribed and may also encode short upstream ORFs. It will be of interest to test the functional importance of these sequences in regulation of antigenic variation and clinical disease.

## Acknowledgements

SP thanks the Department of Biotechnology, Government of India for funding this project. PP was supported by funding from the Indian Institute of Technology (IIT) Bombay during her summer internship. PB received a fellowship from the Council for Scientific and Industrial Research (CSIR).

## Disclosures

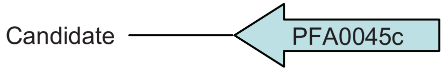
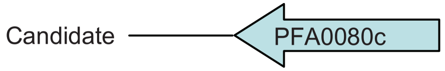
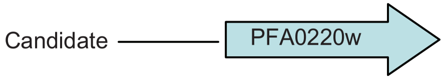
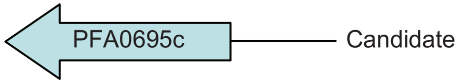
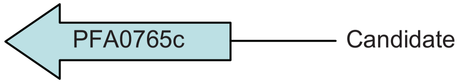
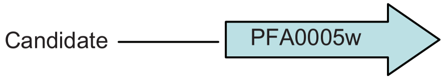
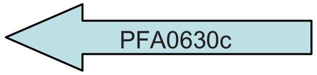
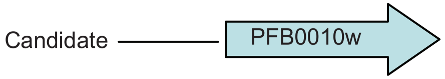
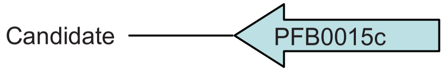
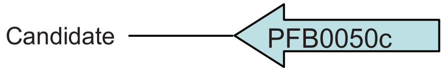
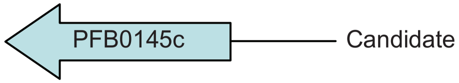
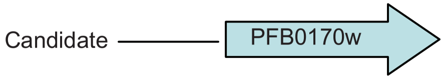
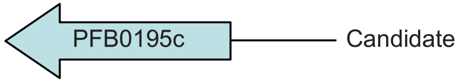
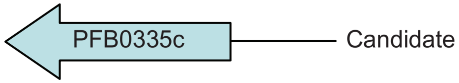
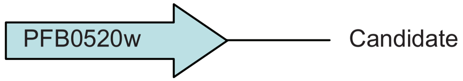
Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

## References

1. Gomez C, Esther RM, Calixto-Galvez M, Medel O, Rodriguez MA. *J Biomed Biotechnol*. 2010:726045.
2. Coleman BI, Duraisingh MT. *Cell Microbiol*. 2008;10:1935–46.
3. Deitsch K, Duraisingh M, Dzikowski R, et al. *Am J Trop Med Hyg*. 2007;77:201–8.
4. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. *Nature*. 2003;423:241–54.
5. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. *Genome Res*. 2004;14:451–8.
6. McCutcheon JP, Eddy SR. *Nucleic Acids Research*. 2003;31:4119–28.
7. Wu J, Sieglaff DH, Gervin J, Xie XS. *Bioinformatics*. 2008;24:1843–9.
8. Gardner MJ, Hall N, Fung E, et al. *Nature*. 2002;419:498–511.
9. Militello KT, Dodge M, Bethke L, Wirth DF. *Mol Biochem Parasitol*. 2004;134:75–88.
10. Mourier T, Pain A, Barrell B, Griffiths-Jones S. *Rna*. 2005;11:119–22.
11. Upadhyay R, Bawankar P, Malhotra D, Patankar S. *Mol Biochem Parasitol*. 2005;144:149–58.
12. Klein RJ, Misulovin Z, Eddy SR. *Proc Natl Acad Sci U S A*. 2002;99:7542–7.
13. Schattner P. *Nucleic Acids Research*. 2002;30:2076–82.
14. Larsson P, Hinas A, Ardell DH, Kirsebom LA, Virtanen A, Soderbom F. *Genome Res*. 2008;18:888–99.
15. Klein RJ, Misulovin Z, Eddy SR. *Proc Natl Acad Sci U S A* 2002;99:7542–7.
16. Roy SW, Hartl DL. *Genome Res*. 2006;16:750–6.
17. Escalante AA, Ayala FJ. *Proc Natl Acad Sci U S A*. 1994;91:11373–7.
18. Jeffares DC, Pain A, Berry A, et al. *Nat Genet*. 2007;39:120–5.
19. Qari SH, Shi YP, Pieniazek NJ, Collins WE, Lal AA. *Mol Phylogenet Evol*. 1996;6:157–65.
20. Lal AA, Goldman IF. *J Biol Chem*. 1991;266:6686–9.
21. Polley SD, Weedall GD, Thomas AW, Golightly LM, Conway D. *J Mol Biochem Parasitol*. 2005;142:25–31.

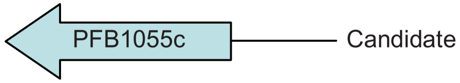
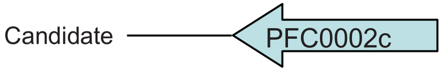
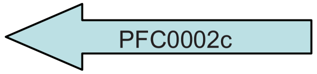
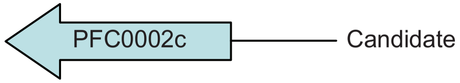
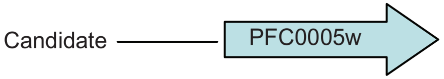
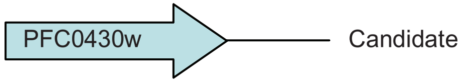
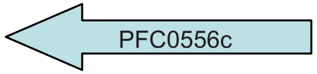
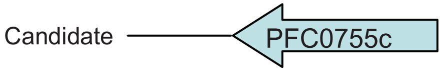
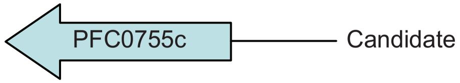
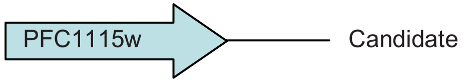
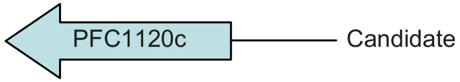
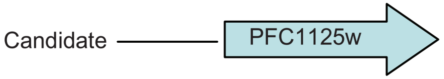
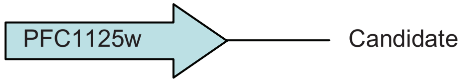
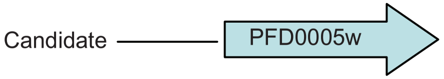
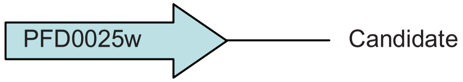
22. Trimnell AR, Kraemer SM, Mukherjee S, et al. *Mol Biochem Parasitol*. 2006;148:169–80.
23. Carlton JM, Angiuoli SV, Suh BB, et al. *Nature*. 2002;419:512–9.
24. Janssen CS, Barrett MP, Turner CM, Phillips RS. *Proc Biol Sci*. 2002;269:431–6.
25. Perkins SL, Sarkar IN, Carter R. *Infect Genet Evol*. 2007;7:74–83.
26. Mattick JS. *PLoS Genet*. 2009;5:e1000459.
27. Iacono M, Mignone F, Pesole G. *Gene*. 2005;349:97–105.
28. Calvo SE, Pagliarini DJ, Mootha VK. *Proc Natl Acad Sci U S A*. 2009;106:7507–12.
29. Amulic B, Salanti A, Lavstsen T, Nielsen MA, Deitsch, KW. *PLoS Pathog*. 2009;5:e1000256.
30. Watanabe J, Sasaki M, Suzuki Y, Sugano S. *Gene*. 2002;291:105–13.
31. Golightly LM, Mbacham W, Daily J, Wirth DF. *Mol Biochem Parasitol*. 2000;105:61–70.
32. Neafsey DE, Hartl DL, Berriman M. *Mol Biol Evol*. 2005;22;1621–6.
33. Goodman M. *Am J Hum Genet*. 1999;64:31–9.
34. Hughes AL, Verra F. *Molecular Phylogenetics and Evolution*. 2010;57:135–43.
35. Rayner JC, Huber CS, Barnwell JW. *Molecular and Biochemical Parasitology*. 2004;138:243–7.
36. Verra F, Polley SD, Thomas AW, Conway DJ. *Parassitologia*. 2006;48:567–72.
37. Kraemer SM, Smith JD. *Curr Opin Microbiol*. 2006;9:374–80.
38. Frank M, Deitsch K. *Int J Parasitol*. 2006;36:975–85.
39. Scherf A. *Cell*. 2006;124:251–3.
40. Calderwood MS, Gannoun-Zaki L, Wellems TE, Deitsch KW. *J Biol Chem*. 2003;278:34125–32.
41. Kyes SA, Kraemer SM, Smith JD. *Eukaryot Cell*. 2007;6:1511–20.
42. Ralph SA, Scheidig-Benatar C, Scherf A. *Proc Natl Acad Sci U S A*. 2005;102:5414–9.
43. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, et al. *Cell*. 2005;121:25–36.
44. Duraisingh MT, Voss TS, Marty AJ, et al. *Cell*. 2005;121:13–24.
45. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW. *Rna*. 2009;15:116–27.
46. Watanabe J, Suzuki Y, Sasaki M, Sugano S. *Nucleic Acids Res*. 2004;32:D334–8.
47. Watanabe J, Wakaguri H, Sasaki M, Suzuki Y, Sugano S. *Nucleic Acids Res*. 2007;35:D431–8.
48. Krummheuer J, Johnson AT, Hauber I, et al. *Virology*. 2007;363:261–71.
49. Craig A, Scherf A. *Mol Biochem Parasitol*. 2001;115:129–43.
50. Sherman IW, Eda S, Winograd E. *Microbes Infect*. 2003;5:897–909.
51. Fernandez V, Hommel M, Chen Q, Hagblom P, Wahlgren M. *J Exp Med*. 1999;190:1393–404.
52. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF. *Science*. 2004;306:1930–3.
53. Marti M, Baum J, Rug M, Tilley L, Cowman AF. *J Cell Biol*. 2005;171:587–92.
54. Tham WH, Payne PD, Brown GV, Rogerson SJ. *Int J Parasitol*. 2007;37:605–15.

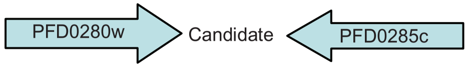
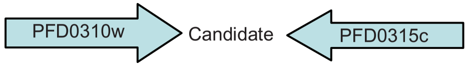
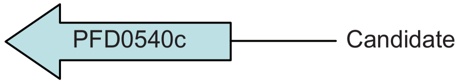
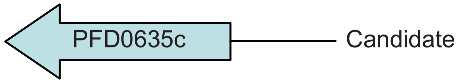
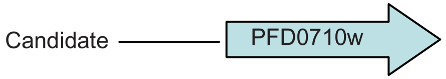
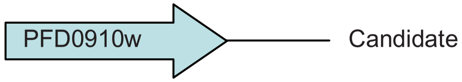
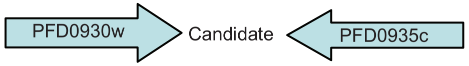
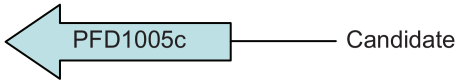
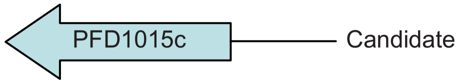
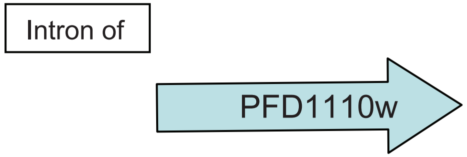
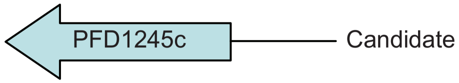
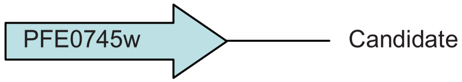
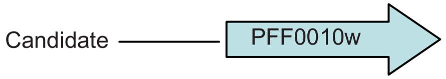
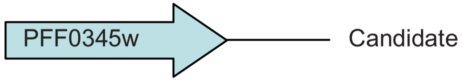
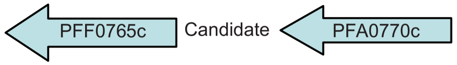
**Table S1.** List of 151 GC-rich sequences proximal to annotated genes identified in *P. falciparum*.

| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 1 | Chr 1: 62341 to 62410 |  | RIFIN | 10 | 35% and 36/70 |
| 2 | Chr 1: 81921 to 81990 |  | RIFIN | 8 | 38% and 55/70 |
| 3 | Chr 1: 197781 to 197850 |  | Ubiquitin carboxyl terminal hydrolase | 5 | 35% and 62/70 |
| 4 | Chr 1: 556971 to 557100 |  | PfEMP1 | 320 | 35% and 126/130 |
| 5 | Chr 1: 616621 to 616710 |  | PfEMP1 | 8 | 38.9% and 33/90 |
| 6 | Chr 1: 29631 to 29730 |  | PfEMP1 | 3 | 37% and 58/100 |
| 7 | Chr 1: 503731 to 503800 | Intron of  | Hypothetical protein | | 35% and 40/70 |
| 8 | Chr 2: 25101 to 25230 |  | PfEMP1 | 2 | 40.8% and 56/130 |
| 9 | Chr 2: 32951 to 33020 |  | RIFIN | 10 | 35% and 36/70 |
| 10 | Chr 2: 54331 to 54410 |  | STEVOR isoform gam beta | 8 | 43.8% and 73/80 |
| 11 | Chr 2: 147571 to 147640 |  | Hypothetical protein | 7 | 37% and 58/70 |
| 12 | Chr 2: 165911 to 165990 |  | Hypothetical protein | 154 | 35% and 57/80 |
| 13 | Chr 2: 197361 to 197430 |  | Hypothetical protein | 445 | 35% and 66/70 |
| 14 | Chr 2: 301801 to 301970 |  | Cysteine protease putative | 8 | 40.6% and 162/170 |
| 15 | Chr 2: 473501 to 473600 |  | Protein kinase putative | 10 | 37% and 93/100 |

(*Continued*)

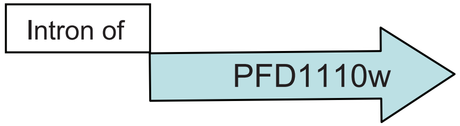
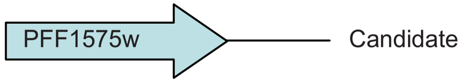
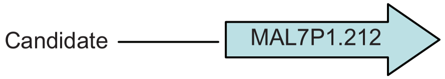
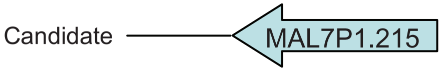
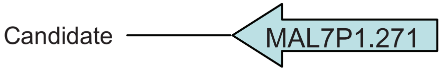
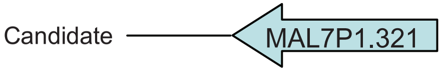
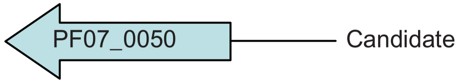
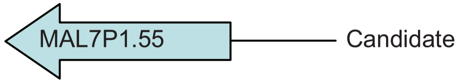
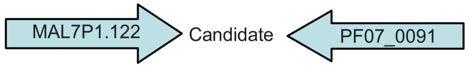
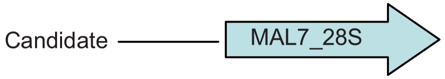
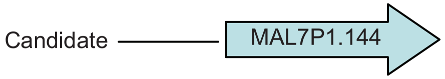
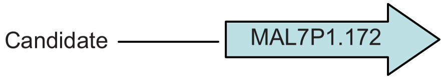
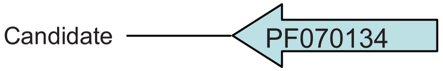
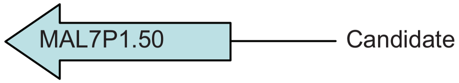
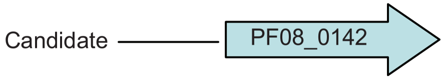
**Table S1.** (*Continued*)

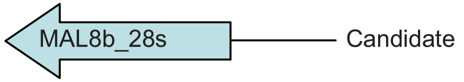
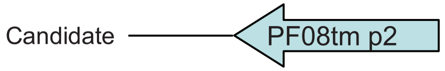
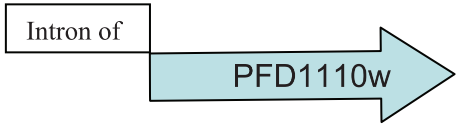
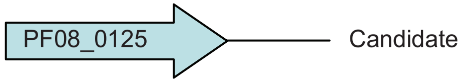
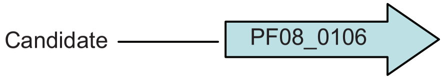
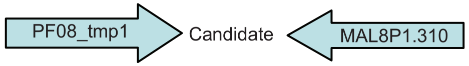
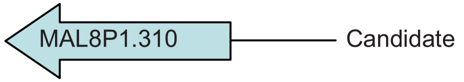
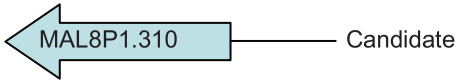
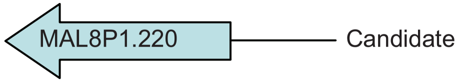
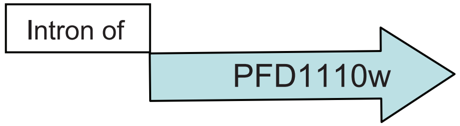
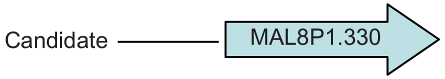
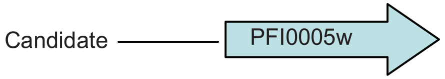
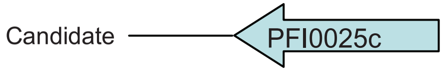
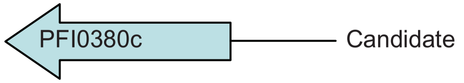
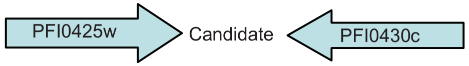
| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 16 | Chr 2: 923651 to 923750 |  | PfEMP1 | 3 | 42% and 58/100 |
| 17 | Chr 3: 8031 to 8160 |  | Hypothetical protein | 234 | 36.2% and 93/130 |
| 18 | Chr 3: 8461 to 8540 | Intron of  | Hypothetical protein |  | 35% and 62/80 |
| 19 | Chr 3: 10961 to 11050 |  | Hypothetical protein | 216 | 34.4% and 74/90 |
| 20 | Chr 3: 33511 to 33640 |  | PfEMP1 | 1 | 38.5% and 72/130 |
| 21 | Chr 3: 443031 to 443120 |  | Hypothetical protein | 197 | 34.4% and 80/90 |
| 22 | Chr 3: 540901 to 540980 | Intron of  | Hypothetical protein |  | 35% and 40/50 |
| 23 | Chr 3: 686651 to 686720 |  | Protein kinase putative | 107 | 38% and 31/70 |
| 24 | Chr 3: 691491 to 691560 |  | Protein kinase putative | 3 | 35% and 66/70 |
| 25 | Chr 3: 1025611 to 1025680 |  | Rifin | 12 | 35% and 34/70 |
| 26 | Chr 3: 1034931 to 1035030 |  | Var gene | 7 | 41% and 42/100 |
| 27 | Chr 3: 1046441 to 1046510 |  | Hypothetical protein | 351 | 35% and 35/70 |
| 28 | Chr 3: 1051191 to 1051280 |  | Hypothetical protein | 213 | 34.4% and 65/90 |
| 29 | Chr 4: 35061 to 35150 |  | PfEMP1 | 3 | 46.7% and 32/90 |
| 30 | Chr 4: 67831 to 67940 |  | RIFIN | 445 | 37.3% and 78/110 |

| | | | | | |
|---|---|---|---|---|---|
| 31 | Chr 4: 311851 to 311920 |  | Hypothetical protein and lysine decarboxylase | 36 and 298 | 35% and 42/70 |
| 32 | Chr 4: 336301 to 336410 |  | Sexual stage specific precursor and hypothetical protein | 135 and 5 | 39.1% and 106/110 |
| 33 | Chr 4: 500431 to 500520 |  | Hypothetical protein | 159 | 35.6% and 65/90 |
| 34 | Chr 4: 606841 to 606930 |  | PfEMP1 | 9 | 42.2% and 38/90 |
| 35 | Chr 4: 667311 to 667410 |  | GTP binding protein | 197 | 34% and 95/100 |
| 36 | Chr 4: 851901 to 851970 |  | Hypothetical protein | 404 | 35% and 54/70 |
| 37 | Chr 4: 862591 to 862660 |  | CGI141 protein homolog, and hypothetical protein | 322 and 218 | 35% and 56/70 |
| 38 | Chr 4: 970091 to 970160 |  | PfEMP 1 | 41 | 35% and 34/70 |
| 39 | Chr 4: 981221 to 981290 |  | PfEMP1 | 191 | 37% and 36/70 |
| 40 | Chr 4: 1064251 to 1064380 |  | Hypothetical protein | | 37.7% and 80/130 |
| 41 | Chr 4: 1183861 to 1183950 |  | PfEMP1 | 13 | 45.6% and 31/90 |
| 43 | Chr 5: 619951 to 620030 |  | Hypothetical protein | 8 | 38.8% and 56/80 |
| 44 | Chr 6: 3401 to 3500 |  | PfEMP1 | 3 | 42% and 38/100 |
| 45 | Chr 6: 296831 to 296920 |  | Translation initiation factor IF2 | 38 | 34.4% and 77/90 |
| 46 | Chr 6: 661791 to 661880 |  | Hypothetical proteins | 6 and 551 | 31.1% and 55/90 |

(*Continued*)

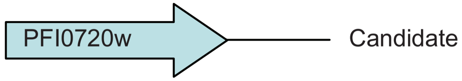
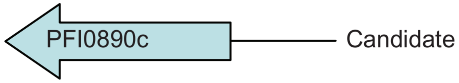
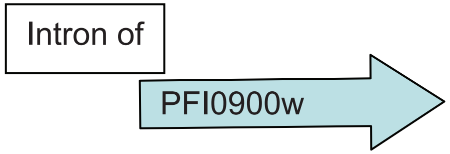
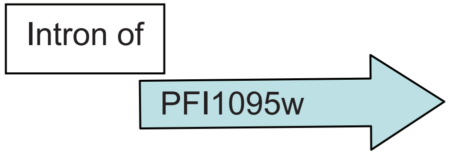
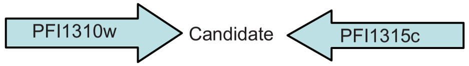
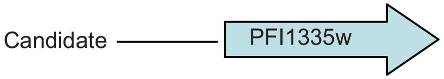
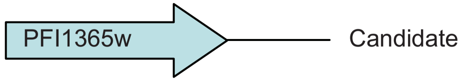
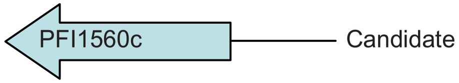
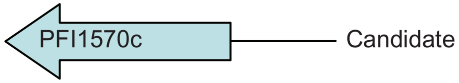
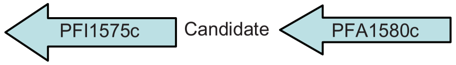
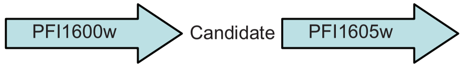
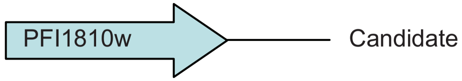
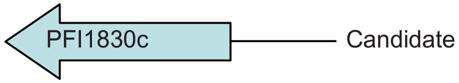
**Table S1.** (*Continued*)

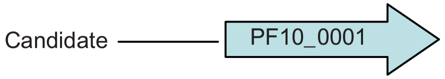
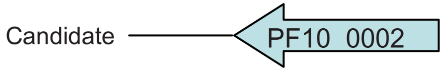
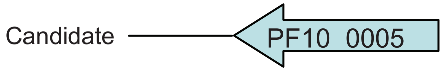
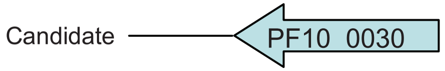
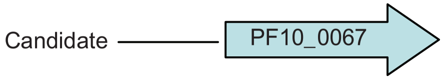
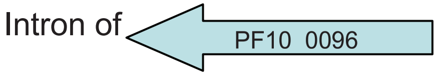
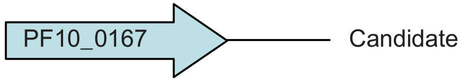
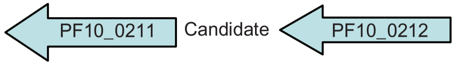
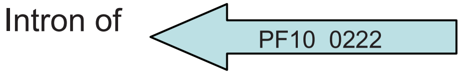
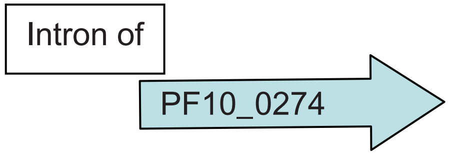
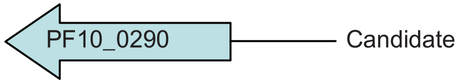
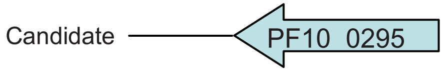
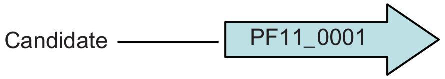
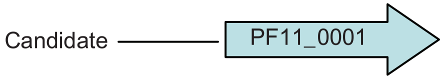
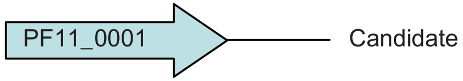
| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 47 | Chr 6: 672491 to 672560 |  | Hypothetical protein |  | 35% and 64/70 |
| 48 | Chr 6: 1352101 to 1352170 |  | RIFIN | 80 | 37% and 50/70 |
| 49 | Chr 7: 30531 to 30670 |  | PfEMP1 | 3 | 37.9% and 81/140 |
| 50 | Chr 7: 45441 to 45540 |  | RIFIN | 447 | 35% and 79/100 |
| 51 | Chr 7: 55261 to 55330 |  | RIFIN | 1 | 37% and 45/70 |
| 52 | Chr 7: 98671 to 98740 |  | Hypothetical protein | 215 | 37% and 70/70 |
| 53 | Chr 7: 614461 to 614570 |  | PfEMP1 | 3 | 40% and 38/110 |
| 54 | Chr 7: 644311 to 644440 |  | PfEMP1 | 8 | 41.5% and 43/130 |
| 55 | Chr 7: 1012111 to 1012190 |  | Conserved GTP binding protein and cell cycle control protein cwf15 homologue | 390 and 108 | 32.5% and 44/80 |
| 56 | Chr 7: 1145161 to 1145250 |  | MAL7_28Sa | 7 | 47.8% and 64/90 |
| 57 | Chr 7: 1155801 to 1155890 |  | Hypothetical protein | 6 | 36.7% and 87/90 |
| 58 | Chr 7: 1393981 to 1394050 |  | Hypothetical protein | 287 | 35% and 70/70 |
| 59 | Chr 7: 1454751 to 1454820 |  | RIFIN | 10 | 37% and 47/70 |
| 60 | Chr 7: 605971 to 606040 |  | PfEMP1 | 71 | 37% and 35/70 |
| 61 | Chr 8: 22251 to 22330 |  | PfEMP1 | 39 | 41.3% and 41/80 |

| | | | | | |
|---|---|---|---|---|---|
| 62 | Chr 8: 98871 to 99000 |  | MAL8b_28s rRNA | 176 | 36.2% and 117/130 |
| 63 | Chr 8: 99751 to 99850 |  | PF08_tmp2 | 164 | 36% and 98/100 |
| 64 | Chr 8: 187771 to 187870 |  | Hypothetical protein | | 31% and 93/100 |
| 65 | Chr 8: 233921 to 233990 |  | Tubulin gamma chain | 45 | 35% and 66/70 |
| 66 | Chr 8: 441381 to 441450 |  | PfEMP1 | 80 | 38% and 34/70 |
| 67 | Chr 8: 1289051 to 1289130 |  | PF08_tmp1 r RNA and putative senescence associated protein | 225 and 129 | 50% and 53/80 |
| 68 | Chr 8: 1289821 to 1289940 |  | Senescence associated protein | 163 | 40% and 116/120 |
| 69 | Chr 8: 1290151 to 1290250 |  | Senescence associated protein | 493 | 37% and 96/100 |
| 70 | Chr 8: 1399241 to 1399340 |  | PfEMP1 | 7 | 38% and 38/100 |
| 71 | Chr 8: 1410561 to 1410640 |  | Hypothetical protein | | 35% and 56/80 |
| 72 | Chr 8: 1412661 to 1412770 |  | Hypothetical protein | 10 | 38.2% and 52/110 |
| 73 | Chr 9: 19931 to 20070 |  | PfEMP1 | 10 | 40.7% and 92/140 |
| 74 | Chr 9: 42361 to 42460 |  | RIFIN | 449 | 34% and 80/100 |
| 75 | Chr 9: 369131 to 369220 |  | Formylmethionine deformylase | 349 | 31.1% and 55/90 |
| 76 | Chr 9: 406351 to 406460 |  | Transporter protein and hypothetical protein | 74 and 54 | 33.6% and 108/110 |

(*Continued*)

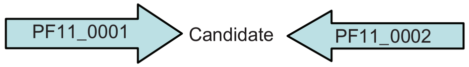
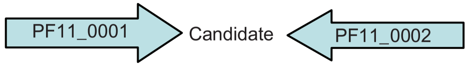
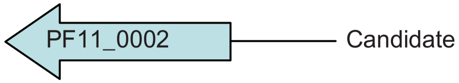
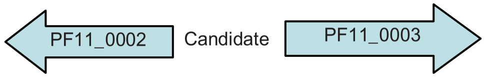
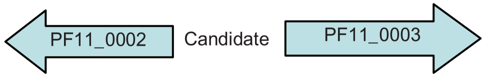
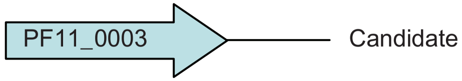
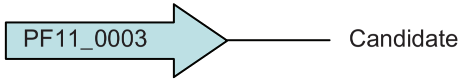
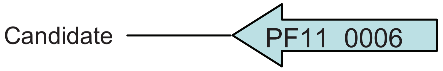
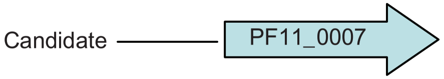
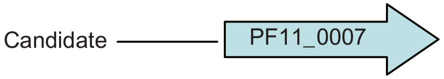
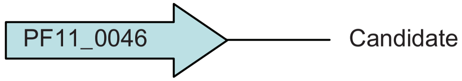
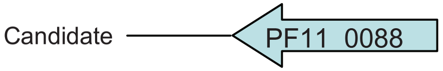
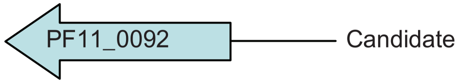
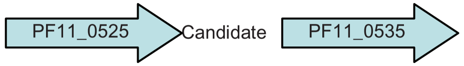
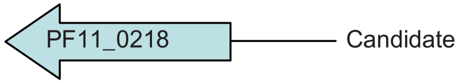
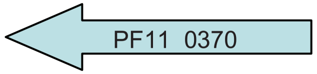
**Table S1.** (*Continued*)

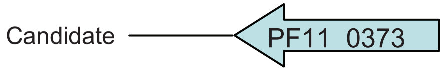
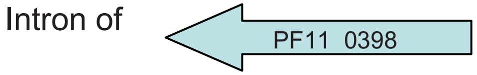
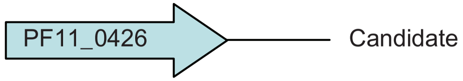
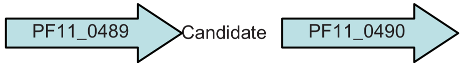
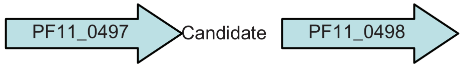
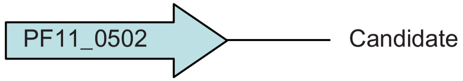
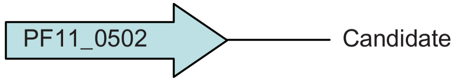
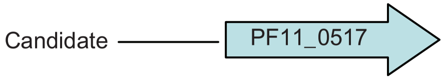
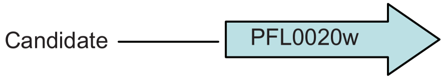
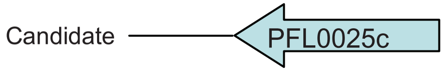
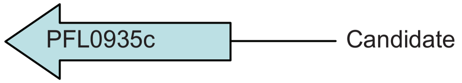
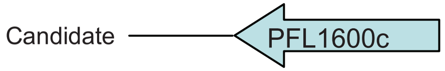
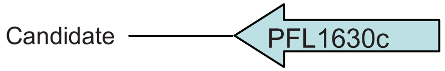
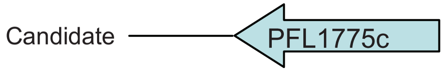
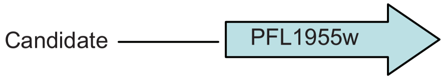
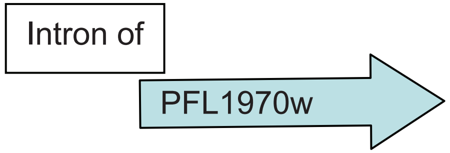
| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 77 | Chr 9: 632261 to 632330 |  | Hypothetical protein | 63 | 35% and 58/70 |
| 78 | Chr 9: 749991 to 750080 |  | Large ribosomal subunit protein L3, prokaryotic (50S)like | 115 | 32.2% and 88/90 |
| 79 | Chr 9: 757341 to 757430 |  | Hypothetical protein | | 33.3% and 50/90 |
| 80 | Chr 9: 907861 to 907930 |  | Hypothetical protein | | 37% and 64/90 |
| 81 | Chr 9: 1092431 to 1092510 |  | NAD synthase and Hypothetical protein | 11 and 324 | 32.5% and 75/80 |
| 82 | Chr 9: 1107301 to 1107400 |  | Hypothetical protein | 2 | 44% and 100/100 |
| 83 | Chr 9: 1130251 to 1130370 |  | Cytochrome c oxidase subunit, | 321 | 32.5% and 103/120 |
| 84 | Chr 9: 1283801 to 1283870 |  | Hypothetical protein | 359 | 37% and 70/70 |
| 85 | Chr 9: 1291101 to 1291170 |  | Hypothetical protein | 84 | 35% and 54/70 |
| 86 | Chr 9: 1293991 to 1294060 |  | Peptide release factor and DHHC type zinc finger protein | 353 and 199 | 37% and 62/70 |
| 87 | Chr 9: 1314241 to 1314350 |  | mRNA processing protein and Hypothetical protein | 526 and 192 | 35.5% and 101/110 |
| 88 | Chr 9: 1479191 to 1479290 |  | RIFIN | 467 | 35% and 81/100 |
| 89 | Chr 9: 1503331 to 1503430 |  | PfEMP1 | 7 | 37% and 38/100 |

| | | | | | |
|---|---|---|---|---|---|
| 90 | Chr 10: 28351 to 28490 |  | PfEMP1 | 1 | 36.4% and 76/100 |
| 91 | Chr 10: 39021 to 39090 |  | RIFIN | 17 | 35.7% and 89/100 |
| 92 | Chr 10: 47981 to 48050 |  | RIFIN | 3 | 35.7% and 88/100 |
| 93 | Chr 10: 125441 to 125510 |  | Hypothetical protein | 430 | 37.1% and 57/70 |
| 94 | Chr 10: 274111 to 274200 |  | Hypothetical protein | 182 | 37.8% and 87/90 |
| 95 | Chr 10: 401521 to 401590 |  | Hypothetical protein | | 37.1% and 54/70 |
| 96 | Chr 10: 694111 to 694200 |  | Hypothetical protein | 10 | 35.6% and 50/90 |
| 97 | Chr 10: 886941 to 887040 |  | Hypothetical proteins | 4 and 347 | 34% and 98/100 |
| 98 | Chr 10: 960765 to 960854 |  | Hypothetical protein | | 32.2% and 88/90 |
| 99 | Chr 10: 1162615 to 1162694 |  | Hypothetical protein | | 35% and 79/80 |
| 100 | Chr 10: 1211885 to 1211964 |  | Hypothetical protein | 80 | 36.3% and 77/80 |
| 101 | Chr 10: 1231655 to 1231784 |  | Hypothetical protein | 19 | 33.8% and 100/130 |
| 102 | Chr 10: 1623881 to 1623950 | | RIFIN | 22 | 35.7% and 98/100 |
| 103 | Chr 11: 5321 to 5460 |  | Hypothetical protein | 388 | 37.1% and 115/140 |
| 104 | Chr 11: 5631 to 5730 |  | Hypothetical protein | 118 | 34% and 66/100 |
| 105 | Chr 11: 6141 to 6220 |  | Hypothetical protein | 51 | 35% and 65/80 |

(*Continued*)

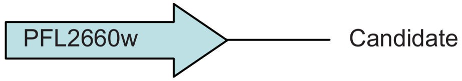
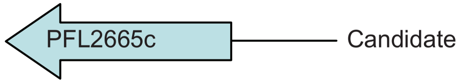
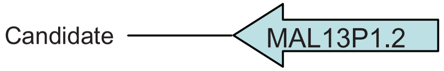
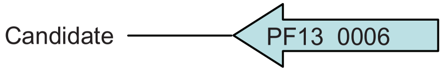
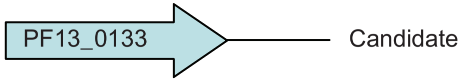
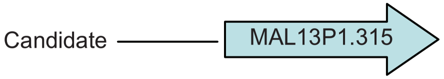
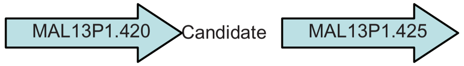
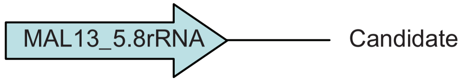
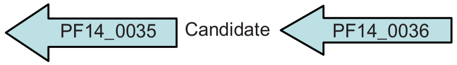
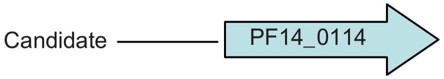
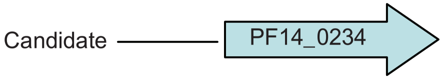
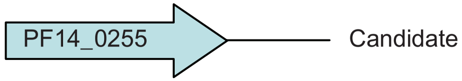
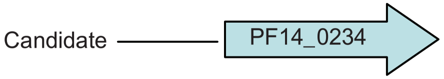
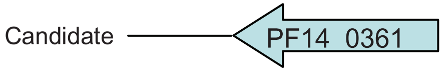
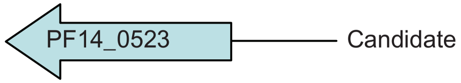
**Table S1.** (*Continued*)

| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 106 | Chr 11: 6311 to 6420 |  | Hypothetical protein | 221 and 401 | 35.5% and 73/110 |
| 107 | Chr 11: 6581 to 6760 |  | Hypothetical proteins | 491 and 61 | 34.4% and 149/180 |
| 108 | Chr 11: 6901 to 7110 |  | Hypothetical protein | 71 | 36.2% and 145/210 |
| 109 | Chr 11: 7271 to 7420 |  | Hypothetical protein | 232 and 340 | 41.3% and 106/150 |
| 110 | Chr 11: 7501 to 7600 |  | Hypothetical protein | 462 and 180 | 33% and 57/100 |
| 111 | Chr 11: 7941 to 8210 |  | Hypothetical protein | 61 | 40% and 119/270 |
| 112 | Chr 11: 8411 to 8720 |  | Hypothetical protein | 262 | 34.2% and 251/310 |
| 113 | Chr 11: 18451 to 18530 |  | Hypothetical protein | 17 | 35% and 69/80 |
| 114 | Chr 11: 24021 to 24150 |  | PfEMP1 | 10 | 49% and 67/120 |
| 115 | Chr 11: 150597 to 150706 |  | Hypothetical protein | 103 | 36.4% and 107/110 |
| 116 | Chr 11: 151007 to 151126 |  | Hypothetical protein | 105 | 35% and 118/120 |
| 117 | Chr 11: 317947 to 318026 |  | Hypothetical protein | 99 | 32.5% and 80/80 |
| 118 | Chr 11: 347207 to 347276 |  | Hypothetical protein | 240 | 37.1% and 69/70 |
| 119 | Chr 11: 569106 to 569175 |  | Hypothetical protein | 56 and 220 | 35.7% and 69/70 |
| 120 | Chr 11: 796606 to 796695 |  | Hypothetical protein | 173 | 35.6% and 90/90 |
| 121 | Chr 11: 1395814 to 1395883 | Intron of  | Hypothetical protein | | 36.8% and 66/70 |

| | | | | | |
|---|---|---|---|---|---|
| 122 | Chr 11: 1417434 to 1417503 |  | Hypothetical protein | 453 | 35.7% and 66/70 |
| 123 | Chr 11: 1527984 to 1528073 |  | Hypothetical protein | | 35.6% and 81/90 |
| 124 | Chr 11: 1663894 to 1664013 |  | Hypothetical protein | 216 | 30.8% and 117/120 |
| 125 | Chr 11: 1918214 to 1918333 |  | Hypothetical protein | 217 and 74 | 33.3% and 112/120 |
| 126 | Chr 11: 1927134 to 1927213 |  | Hypothetical protein | 5 and 255 | 36.3% and 44/80 |
| 127 | Chr 11: 1929634 to 1929743 |  | Hypothetical protein | 58 | 38.2% and 78/110 |
| 128 | Chr 11: 1929974 to 1930053 |  | Hypothetical protein | 282 | 36.3% and 58/80 |
| 129 | Chr 11: 2010214 to 2010293 |  | RIFIN | 4 | 40% and 97/100 |
| 130 | Chr 12: 32601 to 32670 |  | PfEMP1 | 33 | 37.1% and 41/70 |
| 131 | Chr 12: 43711 to 43790 |  | RIFIN | 10 | 33.8% and 92/100 |
| 132 | Chr 12: 774191 to 774300 |  | PfEMP1 | 1 | 38.2% and 53/110 |
| 133 | Chr 12: 1360341 to 1360430 |  | Hypothetical protein | 190 | 34.4% and 54/90 |
| 134 | Chr 12: 1404951 to 1405020 |  | Hypothetical protein | 149 | 37.1% and 67/70 |
| 135 | Chr 12: 1529361 to 1529430 |  | Hypothetical protein | 379 | 37.1% and 70/70 |
| 136 | Chr 12: 1704411 to 1704490 |  | PfEMP1 | 10 | 46.3% and 36/80 |
| 137 | Chr 12: 1739561 to 1739630 |  | PfEMP1 | | 37.5% and 35/70 |

(*Continued*)

**Table S1.** (*Continued*)

| S.no | Name of the candidate | Proximal gene and orientation | Proximal gene | Distance from the proximal gene | GC% and identity |
|---|---|---|---|---|---|
| 138 | Chr 12: 2239401 to 2239480 |  | RIFIN | 1 | 35% and 87/100 |
| 139 | Chr 12: 2248951 to 2249040 |  | PfEMP1 | 6 | 45.6% and 56/90 |
| 140 | Chr 13: 30631 to 30700 |  | RIFIN | 8 | 37.1% and 94/100 |
| 141 | Chr 13: 53591 to 53670 |  | RIFIN | 500 | 40% and 95/100 |
| 142 | Chr 13: 977431 to 977520 |  | Aspartyl (acid) protease, putative | 55 | 38.9% and 82/90 |
| 143 | Chr 13: 2517471 to 2517550 |  | Hypothetical protein | 123 | 33.8% and 79/80 |
| 144 | Chr 13: 2791651 to 2791760 |  | Hypothetical protein conserved | 227 and 76 | 30.9% and 102/110 |
| 145 | Chr 13: 2799331 to 2799400 |  | MAL13_5.8SrRNA rRNA | 311 | 35.7% and 57/70 |
| 146 | Chr 14: 141091 to 141160 |  | Hypothetical protein and acid phosphatase | 206 and 250 | 37.1% and 62/70 |
| 147 | Chr 14: 472871 to 472940 |  | GTP-binding protein, putative | 2 | 35.7% and 59/70 |
| 148 | Chr 14: 989170 to 989279 |  | DNA directed DNA polymerase | 82 | 35.5% and 103/110 |
| 149 | Chr 14: 1086373 to 1086442 |  | Hypothetical protein | 380 | 35.7% and 69/70 |
| 150 | Chr 14: 1213632 to 1213711 |  | Hypothetical protein | 177 | 35% and 80/80 |
| 151 | Chr 14: 1540333 to 1540412 |  | Translocation protein sec62, putative | 170 | 35% and 77/80 |
| 152 | Chr 14: 2247864 to 2247933 |  | Protein phosphatase 2C, putative | 349 | 35.7% and 69/70 |