A Systematic Map of Genetic Variation in *Plasmodium falciparum*

Claire Kidgell¹, Sarah K. Volkman², Johanna Daily², Justin O. Borevitz³*, David Plouffe⁴, Yingyao Zhou⁴, Jeffrey R. Johnson¹, Karine G. Le Roch⁴, Ousmane Sar⁵, Omar Ndir⁵, Soulyemane Mboup⁵, Serge Batalov⁴, Dyann F. Wirth², Elizabeth A. Winzeler¹*¹

¹ Department of Cell Biology, The Scripps Research Institute, La Jolla, California, United States of America, ² Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, United States of America, ³ Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California, United States of America, ⁴ Genomics Institute of the Novartis Research Foundation, San Diego, California, United States of America, ⁵ Faculty of Medicine and Pharmacy, Cheikh Anta Diop University, Dakar, Senegal

Discovering novel genes involved in immune evasion and drug resistance in the human malaria parasite, *Plasmodium falciparum*, is of critical importance to global health. Such knowledge may assist in the development of new effective vaccines and in the appropriate use of antimalarial drugs. By performing a full-genome scan of allelic variability in 14 field and laboratory strains of *P. falciparum*, we comprehensively identified about 500 genes evolving at higher than neutral rates. The majority of the most variable genes have paralogs within the *P. falciparum* genome and may be subject to a different evolutionary clock than those without. The group of 211 variable genes without paralogs contains most known immunogens and a few drug targets, consistent with the idea that the human immune system and drug use is driving parasite evolution. We also reveal gene-amplification events including one surrounding *pfmdr1*, the *P. falciparum* multidrug-resistance gene, and a previously uncharacterized amplification centered around the *P. falciparum* GTP cyclohydrolase gene, the first enzyme in the folate biosynthesis pathway. Although GTP cyclohydrolase is not the known target of any current drugs, downstream members of the pathway are targeted by several widely used antimalarials. We speculate that an amplification of the GTP cyclohydrolase enzyme in the folate biosynthesis pathway may increase flux through this pathway and facilitate parasite resistance to antifolate drugs.


Introduction

Malaria is a significant burden on world health. The most severe form of human malaria is caused by the apicomplexan parasite *Plasmodium falciparum*. Genetic variation in this parasite is central to the pathogenesis of the organism, as allelic variability in different clones is thought to facilitate immune evasion [1]. This in turn has significantly impeded progress towards the development of an effective malaria vaccine [2]. Thus understanding the extent and basis of genetic variation is important for strategies of disease control. In addition, the study of genetic variation in different populations and clones that differ in susceptibility to drugs can reveal genes or genetic regions involved in drug resistance [3], and can thus facilitate the use of more appropriate therapies against *P. falciparum* and can save lives.

Despite a tradition of studying microsatellite-length polymorphisms in *P. falciparum* [4–7], many now feel that the most efficient way to study natural variation and population structure is by shotgun sequencing, single-nucleotide polymorphism (SNP) discovery [8], and genotyping using arrays or a variety of other high-throughput, low-cost approaches that are used in human genetics. These processes all have some limitations. First, sequencing is time-consuming, and even though the cost is falling, shotgun sequencing faces sampling issues and is ill suited to the task of finding genome duplications, deletions, and amplifications, some of which may have profound roles in drug resistance. While polony sequencing [9] or new methodologies [10] may eventually allow the genome of a malaria parasite to be fully sequenced in one day, at a reasonable cost, such sequencing methods are not yet in widespread use, and library construction generally remains a bottleneck in the production of sequence data. Furthermore, *P. falciparum*-genome intergenic regions are up to 90% AT rich and are known to recombine when cloned in other organisms; they created special challenges in the original sequencing of the genome [11]. Thus it is possible that the same low-cost approaches that can be applied to other microbial genomes may not work as well for *P. falciparum*. High-throughput genotyping of a subset of SNP markers discovered through the sequencing of laboratory strains is more efficient than sequencing a whole genome. However, this relies on the assumption that traits such as drug resistance will be transmitted throughout a population...
Variability in the genome of the human malaria parasite, *Plasmodium falciparum*, is key to the parasite’s ability to cause disease and overcome therapeutic interventions such as drugs and vaccines. Elucidating the extent of genetic variation in the malaria parasite will therefore be central to decreasing the malaria disease burden. The authors performed a full-genome scan of variability in different strains of *P. falciparum* and observed a nonrandom distribution of variation. In particular, those genes that are predicted to have roles in evading the host immune response or antimalarial drugs show significantly higher levels of variation. In addition, the authors speculate that a previously unreported genome amplification in the folate biosynthesis pathway correlates with resistance to the antimalarial drug sulfadoxine. Such data enable hypotheses to be made about the function of many of the unknown elements in the parasite’s genome, which may permit the identification of new targets that can be investigated for incorporation into a malaria vaccine and may aid in the understanding of how the parasite withstands drug pressure.

Amplification events in a multidrug-resistance gene and describe a previously uncharacterized amplification in the GTP cyclohydrolase, the first enzyme in the frequently targeted folate biosynthesis pathway.

Results

To investigate genome-wide diversity in the human malaria parasite in an unbiased, comparative fashion, genomic DNA from 14 different cloned *P. falciparum* lines, obtained from four different continents (Table 1), were labeled and hybridized in triplicate to a custom high-density oligonucleotide array [14] specifically designed for studying diversity in *P. falciparum* (see Materials and Methods, Table S1).

A single base change between a 25-mer microarray feature and its genomic target results in a reproducible loss of hybridization signal [12]. As the 25-mer probes used in this study were designed to map to unique locations in the genome, this approach enabled us to identify those genes that carry nucleotide mutations relative to the sequenced strain, 3D7. These single-feature polymorphisms (SFPs) are detected using an F-test on probe intensity values for three hybridizations for the query strain and the reference strain (3D7).

Some of these SFPs correspond to SNPs which result in a change in protein structure, some to silent SNPs [15] and, in a few cases, these SFPs are the result of small insertion/deletion events (“indels”), which are more frequently observed in the non-coding regions [16], or in full deletions. We further required that the signal intensity for the 3D7 DNA be higher than that of the alternate strain since the physical location of gene duplications is unknown. To overcome any issues of variation acquired through long-term in vitro culture of laboratory strains, five of the 14 strains analyzed were isolated recently from Pikine, Senegal. These single clones, as indicated by genetic typing of the merozoite surface protein (MSP) 1 gene (unpublished data), were maintained in vitro culture for only a short period.

Our analysis resulted in the identification of 23,653 SFPs across all strains for the 298,782 25-mer *P. falciparum* probes for the 23-Mb genome (Tables S2 and S3). The seven laboratory-derived strains displayed moderately higher numbers of polymorphic probes (Table S4) compared to the strains isolated from Senegal. However, the laboratory strain D6 (originally isolated from Sierra Leone) showed considerably fewer polymorphisms in comparison to all other strains analyzed.

We had previously validated this method using a small Chromosome 2 microarray [17], and direct nucleotide sequencing of several alleles identified in this current study confirmed our accuracy with a low false-positive rate of <5% (Table 2). We also showed that these SFPs corresponded to heritable changes at specific chromosomal loci by examining their segregation patterns in a haploid progeny strain (C188) which was derived from a previous cross between a chloroquine-sensitive clone (HB3) and a chloroquine-resistant clone (Dd2) [18]. We identified recombination breakpoints that clearly corresponded to those previously published (Figure 1) [19], and the analysis of co-inheritance further confirmed a false-positive rate of less than 5%, although this number is likely to be inflated by the presumed existence of genuine mitotic gene-conversion events. As more genetic crosses between *Plasmodium* species become available,
the rapid and reliable genotyping approach presented here will be an important tool for the mapping of multiple heritable traits, particularly in rodent malaria parasites [20], either individually or through bulk segregant analysis [21].

It has been proposed that the analysis of linkage disequilibrium may permit the discovery of loci associated with drug resistance [4,5]. Previous linkage-disequilibrium analysis of 342 polymorphic microsatellite markers has shown that the mutations in the *P. falciparum* chloroquine-resistance transporter gene (*pfcrt*) (MAL7P1.27), which give rise to chloroquine resistance, have arisen on only four independent occasions and that these mutations have spread across the globe through selective sweeps [5]. We also observe that the SFP-derived haplotypes obtained from 4,369 SFPs distinguishing the chloroquine-resistant strains, FCB and Dd2, are nearly identical in the central region of Chromosome 7 and that this region (with the exception of the PfEMP1 cluster) surrounds the *pfcrt* gene (Figure 2). This region is the only region that showed extensive disequilibrium and is accompanied by impressive probability scores. For example, the probability that the exact same haplotype would be observed for FCB and Dd2 by chance over 22 different SFPs in the region between bases 331,436 and 407,275 on Chromosome 7 is 1 in 3.1E–11. Disequilibrium was also noted at the same region in the comparison between FCR3 and Dd2, but was lacking in comparisons between other lines such as HB3 and Dd2. These data suggest that haplotype analysis of DNA derived from un-cloned drug-resistant and drug-sensitive parasites isolated directly from patients could indicate the chromosomal regions associated with drug resistance. In contrast to traditional mapping methods, this approach is not dependent on a set of pre-established markers or specialized primer sets. The entire SFP dataset (Tables S2 and S3), as well as a browsable interface showing the exact location of variation in each gene in the laboratory strains, can be viewed at [http://carrier.gnf.org/publications/SFP](http://carrier.gnf.org/publications/SFP).

Table 1. *P. falciparum* Strains Used in This Analysis

<table>
<thead>
<tr>
<th>Strain</th>
<th>Origin</th>
<th>Description</th>
<th>Drug-Resistance Profile&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D7</td>
<td>The Netherlands&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Laboratory</td>
<td>SDX</td>
<td>Prof. D. Wirth and MR4, USA</td>
<td>[72]</td>
</tr>
<tr>
<td>FCR3</td>
<td>The Gambia</td>
<td>Laboratory</td>
<td>CQ</td>
<td>Prof. I. Sherman</td>
<td>[73]</td>
</tr>
<tr>
<td>FCB</td>
<td>Thailand</td>
<td>Laboratory</td>
<td>CQ, QN, CYC</td>
<td>MR4, USA</td>
<td>[5]</td>
</tr>
<tr>
<td>7G8</td>
<td>Brazil</td>
<td>Laboratory</td>
<td>CQ, PYR, CYC</td>
<td>MR4, USA</td>
<td>[74]</td>
</tr>
<tr>
<td>W2</td>
<td>Thailand</td>
<td>Laboratory</td>
<td>CQ, QN, PYR, SDX</td>
<td>MR4, USA</td>
<td>[75]</td>
</tr>
<tr>
<td>D6</td>
<td>Sierra Leone</td>
<td>Laboratory</td>
<td>—</td>
<td>MR4, USA</td>
<td>[75]</td>
</tr>
<tr>
<td>Dd2</td>
<td>Thailand</td>
<td>Laboratory</td>
<td>CQ, QN, PYR, SDX</td>
<td>MR4, USA</td>
<td>[18]</td>
</tr>
<tr>
<td>HB3</td>
<td>Honduras</td>
<td>Laboratory</td>
<td>PYR</td>
<td>MR4, USA</td>
<td>[76]</td>
</tr>
<tr>
<td>C188</td>
<td>HB3 × Dd2 cross</td>
<td>Laboratory</td>
<td>—</td>
<td>MR4, USA</td>
<td>[19]</td>
</tr>
<tr>
<td>31.01</td>
<td>Pikine, Senegal</td>
<td>Field</td>
<td>—</td>
<td>Dr. J. Daily</td>
<td>n/a</td>
</tr>
<tr>
<td>41.02</td>
<td>Pikine, Senegal</td>
<td>Field</td>
<td>—</td>
<td>Dr. J. Daily</td>
<td>n/a</td>
</tr>
<tr>
<td>48.02</td>
<td>Pikine, Senegal</td>
<td>Field</td>
<td>—</td>
<td>Dr. J. Daily</td>
<td>n/a</td>
</tr>
<tr>
<td>51.02</td>
<td>Pikine, Senegal</td>
<td>Field</td>
<td>—</td>
<td>Dr. J. Daily</td>
<td>n/a</td>
</tr>
</tbody>
</table>

<sup>a</sup>Indicates that more than one nucleotide substitution was observed within a single polymorphic 25-mer feature.

Table 2. Direct Nucleotide Sequencing of Multiple Alleles across Three Loci in the *P. falciparum* Strains W2 and 51.02

<table>
<thead>
<tr>
<th>Strain</th>
<th>PlasmoDB_ID</th>
<th>Gene_ID</th>
<th>Number of Polymorphic Probes Identified</th>
<th>Total Number of Probes</th>
<th>Number of Non-Synonymous Mutations</th>
<th>Number of Synonymous Mutations</th>
<th>Total Number of Probes Re-sequenced</th>
<th>Number of False-Positive Probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2</td>
<td>PF10_0348</td>
<td>810505</td>
<td>5</td>
<td>52</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>PFC1600w</td>
<td>813129</td>
<td>3</td>
<td>52</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>PFL0800c</td>
<td>811213</td>
<td>3</td>
<td>22</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51.02</td>
<td>PF10_0348</td>
<td>810505</td>
<td>7</td>
<td>68</td>
<td>7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>PFC1600w</td>
<td>813129</td>
<td>5</td>
<td>52</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>PFL0800c</td>
<td>811213</td>
<td>3</td>
<td>22</td>
<td>4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Nucleotide sequences derived from independent PCR-generated DNA fragments on both strands, for the three loci in the strains W2 and 51.02, were aligned with the corresponding gene sequences in the published *P. falciparum* 3D7 genome sequence [11]. Sequencing indicated a false-negative rate of ~50%, which was used in converting the SFP rate to the SNP rate. The high false-negative rate is due to the fact that not all probes initially function well, although function is reproducible, and because SNPs outside of the central region of the 25-mer probes sometimes do not produce an SFP. Furthermore, we used stringent parameters in our analysis in order to obtain a very low false-positive rate.

<sup>a</sup>Indicates that more than one nucleotide substitution was observed within a single polymorphic 25-mer feature.

<sup>b</sup>Drug-resistance profiles were obtained from [19,55,79].

DOI: 10.1371/journal.ppat.0020057.t001

DOI: 10.1371/journal.ppat.0020057.t002

CQ, chloroquine; CYC, cycloguanil; PYR, pyrimethamine; QN, quinine; SDX, sulfadoxine.

<sup>a</sup>*P. falciparum* strain 3D7 was derived from *P. falciparum* strain NF54 by limiting dilution [72]. NF54 was derived from a patient living near Schipol airport in the Netherlands, who had never left the Netherlands [77,78].
Classes of Non-Variable Genes

The ratio of SFPs to probes was highly variable depending on a gene’s predicted function, chromosomal position, or life-cycle stage-expression [14]. Although the number of SFPs did not appear biased to any particular chromosome, an uneven distribution of polymorphic features within each chromosome was observed, as previously noted on Chromosome 2 [17]. Non-variable genes (more than ten probes, ≤10% SFP rate) accounted for 82% of the total genes located in the central regions of the chromosome (4,077/4,998 genes). In contrast 7,555 (75%) of the 10,122 probes in subtelomeric regions were polymorphic versus 12,978 (5%) of the remaining 247,910 probes (p ≤ 1.0E−100). Despite this observation, many genes that are located in the central regions of the chromosome show variability on an individual basis.

Low levels of variability were also found in different functional classes. For example, only 49 SFPs were observed for the 2,034 probes mapped to genes having a role in protein biosynthesis for the 13 strains interrogated. Assuming a false-negative SFP discovery rate of 50% and no linkage, this corresponds to one polymorphism for every 7,000 bases in each strain. In contrast, genes with putative roles or associated with immune evasion (vars, rifins, and stevors [22]) displayed a significantly higher estimated SNP rate of at least one polymorphism for every 200 bases (10,746 SFPs/13,018 probes), although this may be an underestimation as one probe may cover multiple SNPs. The fact that the majority of genes were exceptionally well conserved across strains supports the idea that the different strains of *P. falciparum* emerged through a population bottleneck so recently that multiple mutations have not accumulated [23,24].

Using previously published microarray data from strain 3D7, we found some correlation between the time of mRNA expression [14] and variability, with genes that are expressed during the parasite’s brief extra-cellular phases (merozoite and sporozoite phases) showing almost twice the rate of

---

**Figure 1.** Mapping the Recombination Breakpoints across the 14 *P. falciparum* Chromosomes in the Chloroquine-Sensitive F1 Progeny Strain C188

A total of 14,289 SFPs were identified over three hybridizations by comparing HB3, Dd2, or the recombinant line C188 to 3D7. SFPs were recorded as HB3 (green bars) in the recombinant line if the probe was scored as polymorphic in both C188 and HB3 but not Dd2, or were recorded as Dd2 (red bars) if scored as polymorphic in Dd2 and C188 [19] but not HB3. Regions with a large number of markers scored as both HB3 and Dd2 are shown in yellow and are associated with the known locations of *var*, *rifin*, and *stevor* genes (shown as blue dots). Markers determined by microsatellite length mapping (MLP) are coloured purple and blue [7].

DOI: 10.1371/journal.ppat.0020057.g001

**Figure 2.** Linkage Disequilibrium surrounding *pfcrt* (MAL7P1.27) in FCB and Dd2 on Chromosome 7

Scores were generated by calculating the probability of observing the same genotype by chance over a moving 40-kbp window (with the probability of observing the same genotype for any one SFP by chance placed at 0.33). The plot shows the ratio between the probability and the maximum possible probability for regions with at least four SFPs, with 1 indicating the best possible score. The position of antigenic variation clusters (*vars*, *stevors*, or *rifins*), are shown in blue and are marked. SFPs mapping to these genes were excluded from the calculations because our data indicate that mitotic recombination may be occurring in these genes. *pfcrt*, which is located between bases 307,926 and 311,020 on Chromosome 7, is shown as a black triangle. The trough at *pfcrt* is likely to be due to the strong selective pressure on *pfcrt*, and is consistent with the observation that FCB and Dd2 have different alleles of *pfcrt* even though published SNP data also shows disequilibrium in surrounding regions [5]. Data for other chromosomes are shown in Figure S1.

DOI: 10.1371/journal.ppat.0020057.g002
variation (644 SFPs/8,368 probes for merozoite-expressed genes [$p = 2.33E-46$] and 1,043/5,396 for sporozoite-expressed genes [$p \leq 1.0E-100$]) as genes expressed during the late- trophozoite phase when DNA replication is occurring (373 SFPs/11,751 probes) (Table S5). The diversity observed within the genes expressed during these two extra-cellular life-cycle stages (merozoite and sporozoite stages) may result from an increased selection pressure from the host compared to those parasite stages that reside within the erythrocyte. Although we hypothesized that non-variable genes would show more reproducible expression profiles across strains and across various microarray-expression platforms, this was not the case. The average correlation coefficient for strains HB3 [25] and 3D7 [14] across expression profiles for expressed asexual erythrocytic stage genes, with SFP rates of less than 2% (more than nine probes), was 0.75 ($n = 99$) (http://www.plasmodb.org) suggesting that expression patterns are generally not affected by variation. The greatest predictor of expression correlation across the strains and platforms was relative transcript abundance, with genes exhibiting low expression levels in the erythrocyte and high expression levels in the gametocyte or sporozoite stages ($n = 394$) showing the poorest correlation (0.28).

It has been suggested that intronic sequences are under neutral selection, and thus we calculated their SFP rates [26]. We observed a total of 309 SFPs/5,686 intronic probes, corresponding to a polymorphism for every $\approx 3,000$ bases. However, 76/309 intronic SFPs are found within multi-gene families, while only 239/5,686 intronic probes are located in variable multi-gene families described below, suggesting that intronic sequences are not subject to neutrality when located in multi-gene families ($p = 1.16E-32$). This contrasts with the results of Castillo-Davis et al., who showed that intronic nucleotide substitution was not accelerated in duplicated genes in their comparison of the P. yoelii yoelii and P. falciparum genomes [26]. This discrepancy may be due to the fact that many members of P. falciparum multi-gene families do not have recognizable orthologs in the P. yoelii yoelii genome [27] and were excluded from the analysis. When intron sequences from multi-gene families are omitted, these findings suggest that members of housekeeping families, such as protein biosynthesis (49 SFPs/2,034 probes; SFP rate 2%), are evolving at less than the neutral rate (233 SFPs/5,447 probes; SFP rate 4%) ($p = 0.000095$).

### Types of Variable Genes

Approximately 10% ($n = 457$) of the genes in the malaria genome showed SFP rates (number of SFPs/total number of probes) of greater than 10% (with a minimum of ten probes)—more than twice the rate of neutrality—and thus could be classified as highly variable. These genes could be broken down into three broad categories, including members of multi-gene families, highly expressed immunogens (which are predicted to be under selection from the host immune system), and enzymes that may play a role in the parasite’s response to drug pressure. Excluding the highly variable var, rifin, and stever genes ($n = 290$), known to be involved in antigenic variation, we identified a total of 234 variable genes, across all strains, with 49% (114/234 genes) containing a signal peptide (SP) or at least one transmembrane (TM) domain, suggesting that parasite-mediated exposure of proteins through externalization is important to the immune response. Of these 234 variable genes, 66 were annotated (Table 3) whereas 168 were classed as uncharacterized (Table S6), similar to that which was found in an analysis of SNPs on Chromosome 3 [8]. Within the individual strains, 128 variable genes were identified in the laboratory strains and 86 in the Senegal isolates (Tables S7 and S8). Interestingly, many of the best characterized P. falciparum genes (circumsporozoite [CS] protein, sporozoite surface protein 2, MSPs, erythrocyte-binding antigens) and genes associated with drug resistance, such as the P. falciparum chloroquine-resistance transporter, pfert (MAL7P1.27), reside within the highly variable, annotated genes that we identified (Table 3).

### Multi-Gene Families

We hypothesized that the presence of a paralog would predispose a gene to a loss of the 3D7 haplotype [28,29]. Gene-conversion events, which result in small rearrangements between different alleles or paralogs of the same gene, are the result of the cell’s repair of a double-strand break (reviewed in [30]). In yeast, introducing a paralog can increase the rate of gene conversion [31]. In haploid, asexually replicating eukaryotic organisms such as P. falciparum or Saccharomyces cerevisiae, a double-strand break would likely result in cell death if the break occurred in a gene without a paralog that could be used in the repair process. However, if a paralogous gene were available for repair, the event would likely not be lethal, and evidence of the repair event would remain in the genome as a novel allele. It is also feasible that the presence of a paralog promotes higher rates of non-homologous recombination [29]. Indeed, one of the strongest predictors of whether an uncharacterized gene would show excessive variability was whether or not it was assigned an orthologparalog group [32] (11,862 SFPs/23,781 probes for genes having a paralog versus 8,671 SFPs/234,251 probes for genes lacking a paralog ($p \leq 1.0E-100$). In fact, members of multi-gene families showed a 20-fold enrichment in the group of highly variable genes, relative to the genome as a whole, and a 5-fold increase (1,462 SFPs/11,048 probes; $p \leq 1.0E-100$) after members of the var, rifin, and stever multi-gene families were excluded (10,746 SFPs/13,018 probes). While it may be possible that hybridization-based methods cannot be used to examine variability in multi-gene families, sequencing of PF10_0348, a member of the MSP family, and PFE1600w, a member of the ring-infected erythrocyte surface antigen (RESA) family, showed almost 100% fidelity in all polymorphisms scored (Table 2).

Although variation in var, stevor, and rifin was expected, we identified other groups of consistently variable genes, such as one which contains 13 members (ortholog group 762970 [32]), almost all of which show extensive variation (in terms of SFPs) across the strains that we have examined (315 SFPs/1,051 probes). This gene group displays some homology to the P. knowlesi schizont-infected cell-agglutination variant antigen (SICAv-var) multi-gene family. All members ($n = 13$) of this gene family have previously been designated as the surf genes (surface-associated interspersed genes) [33], which encode a protein family, SURFINs, present on the surface of the infected erythrocyte and merozoite. The high degree of variability observed within these surface antigens implies that members of the surf gene family are likely to be evolving at higher rates [34] relative to the neutral polymorphism rate.
(30% versus 4%), which may also indicate a putative role for them in antigenic variation.

Other variable gene families include a recently identified multi-gene family called the Maurer’s cleft proteins, which are trafficked to the surface of infected red cells [35] (ortholog group 765506 [32]) and showed variation at 29 of the 48 probes specific for this gene family. Some of the variable gene families may be related to human physiology: of the 12 genes that show an over-expression of mRNA transcripts in vivo compared to in vitro [36], nine are members of multi-gene families that show variation. One of these genes (PF14_0752), encodes an uncharacterized protein that is member of a 32-member multi-gene family (ortholog group 774486 [32]). A total of 91 SFPs/319 probes are variable within this ortholog group. Generally, a proportion of these multi-gene families show variation of mRNA expression of family members between different P. falciparum strains, suggesting that selective transcriptional silencing may further regulate their expression.

As previously noted, relatively conserved var genes have also been observed and are associated with binding to the host receptor, chondroitin sulfate A (CSA) [37–39]. Immunonepidemiologic evidence has suggested that this particular var gene variant is conserved between isolates and that it is implicated in malaria infection during pregnancy [37]. Of the 48 probes specific for this gene family, some of the uncharacterized genes that show similar patterns of variability and gene expression to known antigens, such as MSP-1 (PF11475w) or the CS protein (PFC0210c), also interact with the host immune system. Indeed, volunteers protected with an irradiated sporozoite vaccine all exhibited an immune response to the product of PFL0800c [49], a gene that shows a life cycle-dependent expression pattern that is similar to the well-characterized P. falciparum pre-erythrocytic stage antigen, the CS protein (PFC0210c), and which we demonstrate exhibits a comparable degree of variation (four SFPs/22 probes versus five SFPs/25 probes).

Drug Resistance

While single point mutations in drug targets and receptors, which may not be identified by our analysis, can result in drug resistance, evidence suggests in some cases that multiple mutations are associated with resistance [44]. This may be because resistance mutations impose a fitness disadvantage and additional, compensatory mutations are required for resistance or to maintain a protein’s 3D structure. It is known that the P. falciparum chloroquine-resistance transporter (pfcrt) (MAL7P1.27) shows a high rate of substitution with at least 15 different known haplotypes and 15 non-synonymous polymorphisms in genes for those proteins bearing a Pexel motif (309 SFPs/4,401 probes versus 9,178 SFPs/240,613 probes; p = 2.63E−157) or a HT motif (381 SFPs/2,983 probes versus 9,406 SFPs/242,031 probes; p = 7.75E−88) versus those that do not, although not all proteins bearing this signal show variation.

Based upon these observations, it is our hypothesis that many of the uncharacterized genes that show similar patterns of variability and gene expression to known antigens, such as MSP-1 (PF11475w) or the CS protein (PFC0210c), also interact with the host immune system. Indeed, volunteers protected with an irradiated sporozoite vaccine all exhibited an immune response to the product of PFL0800c [49], a gene that shows a life cycle-dependent expression pattern that is similar to the well-characterized P. falciparum pre-erythrocytic stage antigen, the CS protein (PFC0210c), and which we demonstrate exhibits a comparable degree of variation (four SFPs/22 probes versus five SFPs/25 probes).
<table>
<thead>
<tr>
<th>PlasmoDB_ID</th>
<th>Gene_ID</th>
<th>Description</th>
<th>Number of SFPs</th>
<th>Total Number of Probes</th>
<th>Life-Cycle Stage–Expression</th>
<th>Pexel Motif</th>
<th>HT Motif</th>
<th>SignalP</th>
<th>TM Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF10_0343</td>
<td>810500</td>
<td>S-antigen</td>
<td>10</td>
<td>11</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>PF0100c</td>
<td>812616</td>
<td>Knob-associated histidine-rich protein</td>
<td>56</td>
<td>70</td>
<td>Ring/trophozoite</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF14_0010</td>
<td>811601</td>
<td>Glycophorin-binding protein-related antigen</td>
<td>14</td>
<td>19</td>
<td>Trophozoite</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PF0095c</td>
<td>812615</td>
<td>Erythrocyte membrane protein 3</td>
<td>60</td>
<td>76</td>
<td>Ring/trophozoite</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF1170w</td>
<td>813622</td>
<td>Cytoadherence-linked protein</td>
<td>38</td>
<td>52</td>
<td>Schizont/gametocyte</td>
<td></td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF1100w</td>
<td>813640</td>
<td>Enzyme, putative</td>
<td>22</td>
<td>31</td>
<td>Low expression</td>
<td></td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF13205w</td>
<td>812411</td>
<td>Predicted integral membrane protein</td>
<td>12</td>
<td>17</td>
<td>Merozoite</td>
<td></td>
<td>No</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>PFI1475w</td>
<td>813575</td>
<td>MSP-1, precursor</td>
<td>106</td>
<td>162</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>PFC0110w</td>
<td>814340</td>
<td>CLAG 21</td>
<td>21</td>
<td>42</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>PF10_0345</td>
<td>810502</td>
<td>MSP-3</td>
<td>14</td>
<td>31</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>PF12570w</td>
<td>811564</td>
<td>Acylcoa ligase antigen</td>
<td>22</td>
<td>52</td>
<td>Schizont/gametocyte</td>
<td></td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PFI1475w</td>
<td>813575</td>
<td>MSP-1, precursor</td>
<td>106</td>
<td>162</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>PF10_0345</td>
<td>810502</td>
<td>MSP-3</td>
<td>14</td>
<td>31</td>
<td>Schizont</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>PFI0100c</td>
<td>812616</td>
<td>Knob-associated histidine-rich protein</td>
<td>56</td>
<td>70</td>
<td>Ring/trophozoite</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PF14_0010</td>
<td>811601</td>
<td>Glycophorin-binding protein-related antigen</td>
<td>14</td>
<td>19</td>
<td>Trophozoite</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PF0095c</td>
<td>812615</td>
<td>Erythrocyte membrane protein 3</td>
<td>60</td>
<td>76</td>
<td>Ring/trophozoite</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PF1170w</td>
<td>813622</td>
<td>Cytoadherence-linked protein</td>
<td>38</td>
<td>52</td>
<td>Schizont/gametocyte</td>
<td></td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF1100w</td>
<td>813640</td>
<td>Enzyme, putative</td>
<td>22</td>
<td>31</td>
<td>Low expression</td>
<td></td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>PF13205w</td>
<td>812411</td>
<td>Predicted integral membrane protein</td>
<td>12</td>
<td>17</td>
<td>Merozoite</td>
<td></td>
<td>No</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>PlasmoDB_ID</td>
<td>Gene_ID</td>
<td>Description</td>
<td>Number of SFPs</td>
<td>Total Number of Probes</td>
<td>Life-Cycle Stage—Expression</td>
<td>Pexel Motif</td>
<td>HT Motif</td>
<td>SignalP</td>
<td>TM Domain</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>--------------------------------------------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>------------------------------</td>
<td>-------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>PF0070w</td>
<td>812870</td>
<td>Interspersed repeat antigen, putative</td>
<td>9</td>
<td>76</td>
<td>Ring</td>
<td>Very late schizont</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF1055w</td>
<td>813471</td>
<td>Sugar transporter, putative</td>
<td>4</td>
<td>34</td>
<td>Sporozoite</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF1030c</td>
<td>813486</td>
<td>Ubiquitin conjugating enzyme E2, putative</td>
<td>2</td>
<td>17</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF11_0165</td>
<td>810712</td>
<td>Falcipain 2 precursor</td>
<td>2</td>
<td>17</td>
<td>Ring/trypomastigote</td>
<td>Late ring</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF00845c</td>
<td>812984</td>
<td>60S ribosomal subunit protein L8, putative</td>
<td>3</td>
<td>26</td>
<td>Ring/trypomastigote</td>
<td>Late ring</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF1420w</td>
<td>813098</td>
<td>F-actin capping protein, alpha subunit, putative</td>
<td>3</td>
<td>26</td>
<td>Trophozoite/gametocyte</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF11920c</td>
<td>811436</td>
<td>Hydroxethylthiazole kinase, putative</td>
<td>3</td>
<td>26</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF10_0025</td>
<td>810183</td>
<td>PF70 protein</td>
<td>5</td>
<td>44</td>
<td>Ring</td>
<td>Very late schizont</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>MAL/P1.27</td>
<td>2653199</td>
<td>Chloroquine-resistance transporter, putative</td>
<td>4</td>
<td>36</td>
<td>Ring/trypomastigote</td>
<td>Late ring</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF1670c</td>
<td>813614</td>
<td>Vacuolar ATP synthase subunit E, putative</td>
<td>2</td>
<td>18</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF11_0224</td>
<td>810771</td>
<td>CS-related antigen</td>
<td>2</td>
<td>18</td>
<td>Ring/trypomastigote</td>
<td>Very late schizont</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PF11_0461</td>
<td>811004</td>
<td>Rab6</td>
<td>2</td>
<td>18</td>
<td>Schizont</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF20340c</td>
<td>812668</td>
<td>Cysteine protease, putative</td>
<td>11</td>
<td>100</td>
<td>Schizont/gametocyte</td>
<td>Late schizont</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF11_0257</td>
<td>810804</td>
<td>Ethanolamine kinase, putative</td>
<td>4</td>
<td>37</td>
<td>Schizont/gametocyte</td>
<td>Late schizont</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MAL/P1.38</td>
<td>2655211</td>
<td>Regulator of chromosome condensation protein</td>
<td>7</td>
<td>65</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF10_0225</td>
<td>810382</td>
<td>Orotidine-monophosphate-decarboxylase, putative</td>
<td>3</td>
<td>28</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF00355c</td>
<td>812671</td>
<td>Cysteine protease, putative</td>
<td>10</td>
<td>95</td>
<td>Schizont</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF011155w</td>
<td>812436</td>
<td>Erythrocite-binding antigen, putative</td>
<td>17</td>
<td>163</td>
<td>Schizont</td>
<td>Late schizont</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>MAL/P1.296</td>
<td>2654838</td>
<td>Nucleolar GTP-binding protein 1, putative</td>
<td>6</td>
<td>58</td>
<td>Constitutive</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PF08_0102</td>
<td>2655301</td>
<td>Asparagine-rich antigen Pfa55–14</td>
<td>8</td>
<td>78</td>
<td>Gametocyte</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Those genes considered “absent” in one strain, based on the MOID algorithm, are undefined. Candidate vaccine immunogens [2] are shown in bold. TM domain refers to the number of TM domains as predicted by the TMHMM2 algorithm. The candidate vaccine immunogens listed are derived from Ritchie and Saul (2002) [2] and include CSP, TRAP, LSA1, EXP1, LSA3, MSP-1, MSP-2, MSP-3, AMA1, SERA, GLURP, and RESA. The life-cycle expression data are derived from Le Roch et al. [14] and from Bozdech et al. 2003 [25]; 0 indicates a lower than 2-fold induction of expression at ≥50th percentile or no information. Gene IDs were derived from Entrez Gene at NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene).

*Potential duplication.*

DOI: 10.1371/journal.ppat.0020057.t003
derivatives, such as chloroquine and an ADP/ATP carrier protein (PF10_0005) (seven SFPs/32 probes; 21% SFP rate), which could modulate resistance to drugs targeting mitochondrial function, such as Atovaquone. A few non-enzymes that are not paralogs and do not contain a TM or SP domain were also variable. These non-SP/TM–containing proteins include the 50S (PF07_0046) and 60S (PFE0845c) ribosomal proteins. As many antibiotics target the ribosomal subunits, this variation may represent increased selection from indiscriminate antibiotic usage in the human host.

**Gene Amplifications**

Notably absent from the list of highly variable genes (Table 3) was the *P. falciparum* multidrug-resistance gene, *pfmdr11* (PFE1150w), which has been implicated in mefloquine resistance in both human [48] and rodent malaria infections [49]. This was not unexpected since it is amplifications of the *pfmdr1* gene that are associated with resistance, and we required all SFPs scored to show a higher signal in 3D7 relative to the alternate strain. Our analysis of hybridization ratios (see Materials and Methods) confirms the known *pfmdr1* amplification in the mefloquine-resistant strain Dd2 [50], and shows independent amplification events in strains FCR3 and FCB (Table 4). We are also able to detect further gene amplifications in other regions. For example, gene duplication is noted in the strain W2 encompassing the *Pfrh1* gene (PFD0110c) on Chromosome 4 [51] and in the center of Chromosome 7. Our data identified an increased hybridization intensity surrounding the GTP cyclohydrolase gene (PFL1155w) on Chromosome 12 in seven out of the 14 strains analyzed (Table S10). GTP-cyclohydrolase encodes the first enzyme in the folate biosynthesis pathway. The folate pathway is the target of many antifolates (Fansidar, WR99210, sulfadoxine, and pyrimethamine), chemotherapy agents (Methotrexate), and antibiotics (sulfa drugs). The mechanism of resistance to antifolates in vivo has been attributed to mutations within either the dihydrofolate reductase (*dhfr*) (PFD0830c) [47] or dihydropteroate synthase (*dhps*) (PFE0845c) genes [52], but mutations in the *dhps* gene may not always correlate with sulfadoxine resistance [53]. Although the regions surrounding the GTP cyclohydrolase gene (PFL1155w) appeared to be amplified, the median signal for all probes to GTP cyclohydrolase itself was actually significantly lower in the majority of strains relative to 3D7, suggesting an undiscovered amplification in this reference-sequenced strain (Figure 3).

Real-time PCR analysis of the GTP-cyclohydrolase gene (PFL1155w), along with a β-tubulin control gene (PFE10_0084), confirmed the independent amplification of this gene in 3D7 and in the *P. falciparum* strains W2, HB3, FCB, D6, 7G8, and Dd2 but not in FCR3 or any of the five strains isolated from Senegal (Table S11). Interestingly, it has been shown that the expression of bacterial GTP cyclohydrolase-1 in transgenic *Arabidopsis* results in a 1,250-fold and 2- to 4-fold enhancement in the levels of pterins and folates, respectively, suggesting that this is the rate-limiting enzyme in plants [54]. The presence of the amplification in most laboratory strains may indicate the importance of de novo folate biosynthesis in culture. However the absence of this amplification in the Senegal strains may also correlate with a lack of selection pressure in these recently acquired isolates, since antifolate drug regimes replaced chloroquine in Pikine, Senegal, only

---

**Table 4. Description of the Genes Amplified in the *P. falciparum* Strains Analyzed**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Duplicated Region</th>
<th>Function</th>
<th>Representative Genes in Duplicated Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCB</td>
<td>PFL1130c..PFL1165w</td>
<td>Folate biosynthesis</td>
<td>GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24</td>
</tr>
<tr>
<td>HB3</td>
<td>PFL1125w..PFL1135w</td>
<td>GTP cyclohydrolase I</td>
<td>Proliferating cell nuclear antigen (PCNA), cytochrome c heme lyase, putative, GTP cyclohydrolase I</td>
</tr>
<tr>
<td>7GB</td>
<td>PFL1130c..PFL1160c</td>
<td>GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24</td>
<td></td>
</tr>
<tr>
<td>3D7</td>
<td>PFL1155w</td>
<td>GTP cyclohydrolase I</td>
<td>GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24</td>
</tr>
<tr>
<td>W2</td>
<td>PFL1145w..PFL1160c</td>
<td>GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24</td>
<td></td>
</tr>
<tr>
<td>Dd2</td>
<td>PFL1145w..PFL1155w</td>
<td>GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24</td>
<td></td>
</tr>
<tr>
<td>FCB</td>
<td>PFD0090c..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>HB3</td>
<td>PFD0090c..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>W2</td>
<td>PFD0100c..PFD0115c; MAL7P1.52; PFL1365w..PFL1385c</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins; Rif pseudogene; NIMA-related protein kinase (Pfnek-1), 101 kd malaria antigen</td>
</tr>
<tr>
<td>D6</td>
<td>PFD0090c..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>31.01</td>
<td>PF10_0006..PF10_0013</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>FCB</td>
<td>PFD0115w..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>FCR3</td>
<td>PFD0115w..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>Dd2</td>
<td>PFD0115w..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>FCR3</td>
<td>PFD0115w..PFD0120w</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
<tr>
<td>W2</td>
<td>MAL7P1.137..PF07_0106</td>
<td>Surface proteins</td>
<td>Reticulocyte-binding protein, uncharacterized proteins</td>
</tr>
</tbody>
</table>

No gene amplifications were observed in the *P. falciparum* strains 18.02, 41.02, 48.02, and 51.02 isolated from Senegal. DOI: 10.1371/journal.ppat.0020057.004
after the isolation of the strains used in this analysis. From the literature, it appears that this gene amplification correlates with the sulfadoxine-resistance profile previously published for particular strains analyzed [55–57]. We propose that partial sulfadoxine and/or pyrimethamine resistance may be linked to duplication of GTP cyclohydrolase (PFL1155w), which may permit an increased metabolic flux through the de novo pathway, although further investigation of a larger number of strains is required to confirm this association. Additional analysis of this duplication event from recently in vitro–adapted isolates in other geographic areas, as well as the continued analysis for the appearance of GTP-cyclohydrolase duplication in the Senegal strains over time (both in vivo and in vitro), will also be required.

**Gene Deletions**

Previous studies have shown that arrays can be used to identify potential deletions using a comparative genome-hybridization approach that examines the ratios between the hybridization signals for two strains [25,58]. However, such a method makes it difficult to distinguish between a deletion and those cases where a gene is simply highly variable in one strain. The problem is compounded if only one or two probes per gene are used [25] versus the scores of probes used in these analyses. Here we use an algorithm, based on the match-only integral distribution (MOID) algorithm that returns a “present” call based on whether a probe-set distribution for a gene is similar to a series of background control probes [59] on the array. These control probes, which are not expected to hybridize to human or *Plasmodium falciparum* sequences, are predicted to have little signal associated with them and to show the same patterns as “deleted” genes. Analysis of background sequences indicated that they only had a 2% chance of being misclassified as “present”, and thus provides an estimate of the number of likely false positives.

In applying this algorithm, every gene with at least six probes (n = 4,954) was considered present for 3D7 indicating a low false-negative rate as well, although gene-duplication events in strains other than 3D7 could mask the deletion of a very closely related gene (Table S12). While one might predict that highly variable genes might not be called “present,” the
deletion analysis showed that hybridization to variable genes, even for the most variable var, rifin, and stevor genes, is generally higher than the background level (Figure 4; Table S9). For example, PF10_0348 bears 23 SFPs/68 probes and is considered to be highly variable in our analysis. The MOID algorithm–derived E value for this gene ranged from 124 to 269 units with ''present'' p-values in the range of 1.0E\(^{-17}\) to 1.0E\(^{-47}\). In contrast, genes within the previously characterized 42-kbp deletion in P. falciparum strain W2 showed different patterns [17,60]. This deletion is likely to be associated with a loss of cytoadherence in parasite lines maintained in in vitro culture. The signal for PfEMP3 (PFBO095c), a gene in this region, ranged from 200–792 units in strains where it was considered present and from 18.1 to 4.6 units in strains where it was considered deleted. The p-values were in the range of 1.0E\(^{-222}\) for strains where it was considered present and from 1.0E\(^{-3.6}\) to 1.0E\(^{-1}\) where it was considered deleted. These differences are remarkable and allow a high degree of confidence in calling deletions.

We also identify deletions (Table S13) in several genes on the right arm of Chromosome 9 in strains FCR3 and FCB, including those encoding an enzyme (PF11800w), an uncharacterized protein (PF11785w), and the cytoadherence-linked asexual protein (CLAG, PF11710w) [61]. The Maurer’s cleft proteins [35], which co-localize with a 130-KDa membrane-associated protein found in both the Maurer’s clefts and knobs, but only in a cytoadherent P. falciparum strain [62] (thus indicating a potential role in cytoadherence) were also often deleted. Despite the confidence that our method offers, difficulties in distinguishing a full deletion from a large partial deletion or substitution do exist, particularly for large genes such as MSP-1 (PFII1475w). The MSP-1 gene, to which there are 162 probes, was considered “absent” in one strain (HB3), although visual inspection of the gene showed a residual signal at the 3 prime end of the gene, consistent with the known 3 prime conservation of this well-studied immunogen. Because most other deleted genes were located in chromosomal clusters and were themselves adjacent to other deleted genes, which was not true for the S-antigen (PF10_0343) and MSP-1 (PFII1475w), our confidence in calling these genes “absent” is decreased. Efforts to PCR-amplify the entire 10-kbp MSP-1 region using primers outside of the variable regions were not successful.

**Discussion**

Two recent studies examined selection in the P. falciparum genome: the first measured codon volatility [63], and the second identified orthologous gene sequences under selection through comparative sequencing of rodent parasite genomes [64]. A lack of correlation was observed between each gene’s average SFP value, as determined by this study,
and the corresponding p-value assigned by codon volatility (r = 0.06, Pearson’s correlation) or the dN/dS ratio (r = 0.01, Pearson’s correlation). However, because of the fundamental differences in the datasets and experimental approaches, we did not expect to observe a correlation. While the codon-volatility data measurement should be comparable, various groups have suggested that codon volatility makes unfounded assumptions about underlying mutational processes and is potentially not a robust approach to detect selection [65–67].

Secondly, Hall et al. [64] examined speciation events that occurred over a much larger time scale, while our data consider events that likely occurred more recently, poten-tially within the past 50 y. Furthermore, only 20% of the highly variable P. falciparum genes have detectable orthologs in the rodent malaria species, and only 13 out of the 77 genes showing a dN/dS ratio of >1 between P. berghei and P. chabaudi show an orthologous gene sequence present in P. falciparum (versus 70% of all genes). However, integrating the rodent dataset with our variability data may allow better prediction of potential drug-resistance genes. Characterized P. falciparum enzymes such as dhfr, orotidine monophosphate dehydrogenase, and heme A farnesyltransferase show very low dN/dS ratios but high rates of variability in our assay, suggesting roles in drug resistance. These data may also be useful to systems biology research.

Traditional gene-by-gene approaches to understanding gene function are difficult and time-consuming in P. falciparum. Systematic genome-wide analyses such as those described here can assist in the credentialing of genes to identify the most promising drug and vaccine targets. As with expression patterns, protein–protein interaction maps, or global-localization studies, the SFP rate offers an insight as to the potential function of uncharacterized genes; the SFP rate may indicate that a gene is likely to be involved in antigenic variation, that it is interacting with the host’s immune system, or that it is evolving under drug pressure. Probes that tile through the P. falciparum genome can now be placed on a single microarray, which should permit the identification, at a higher resolution, of genes that are changing in response to the introduction of new drug treatment regimes, such as new artemisinin-based combination therapies [68]. Even if a single gene cannot be identified as being differentially under selection when comparing treated and untreated parasite populations, hybridization haplotypes associated with drug resistance [69,70] will likely be found. These data will also assist in the interpretation of traditional mapping data, because it is likely that SNPs in genes that are highly variable, or members of multi-gene families, will be less reliable than those that lie in genes that are largely monomorphic.

These data also shed light on the time of P. falciparum speciation. If one considers just the average number of SNPs [8] or SFPs, as in this study, the P. falciparum species appears ancient, but after excluding variation in genes that are likely to be evolving by mitotic recombination, the species looks considerably younger. While much additional work will be clearly required to confirm associations between mutations and resistance to individual drugs, such connections, if supported, will allow the design and implementation of more effective therapies. Ultimately, some of these genes involved in drug resistance, particularly those whose copy number is critical to parasite survival, may become the targets of novel drugs to treat malaria.

Materials and Methods

P. falciparum Affymetrix full-genome array. Although useful for gene-expression analysis, the malaria full-genome oligonucleotide array [14] was specifically designed for studying diversity in P. falciparum. The design was optimized so as to obtain many nonoverlapping probes in coding regions as possible. The array contains 327,989 specific single-stranded 25-mer probes, of which 281,552 map to the exons of 5,170 genes that comprise ~55% of the 23-Mbp genome. After excluding overlap between probes and repetitive regions (~18% of the coding sequence), on average more than 53% of the exon sequences are covered by at least one probe. Probe coverage is significantly lower in the intragenic regions because of the frequency of repetitive sequences and the high AT content (up to 90%). Of the 327,989 single-stranded probes, 29,207 probe complements of another SFP and reverse complements do not always exhibit identical hybridization behavior, the forward probe might be capable of revealing an SFP while its complement might not. Thus, reverse-complement probes were excluded from the SFP tally only if both the forward and the reverse complement detected an SFP. Reverse complements were excluded from the number of probes per genes listed in Tables 1–4.

DNA methods. P. falciparum parasites were cultured in leukocyte-depleted human O+ erythrocytes as previously described [71], and genomic DNA was isolated by standard phenol–chloroform extrac-tion. DNA contamination was removed by treatment with RNase at 37 °C for 1 h.

Genomic DNA (15 μg) from each P. falciparum strain (Table 1) and 2.5 ng each of Bio B, Bio C, Bio D, and Cre control plasmids were fragmented by treatment with DNase I and end-labeled with biotin [12]. Hybridization was subsequently carried out on the P. falciparum full-genome array [14] at 45 °C overnight, followed by a standard wash protocol and antibody staining (Affymetrix, Santa Clara, California, United States) and detection with streptavidin R-phycocerythrin conjugate and normal goat IgG. Arrays were scanned with an emission wavelength of 560 nm at 3-μm resolution using a confocal scanner (Affymetrix), and the signal intensity for each feature on the array was determined using the 70th percentile method in Microarray Suite 5 (Affymetrix). Three replicate hybridizations for each strain were carried out.

Identification of SFPs. SFPs were identified as described in Winzeler et al. [12]. Grids were aligned to the scanned images by the known feature dimensions of the array. The hybridization intensities for each of the elements in the grid were determined by the 75th percentile method in the Affymetrix GeneChip software package. For each hybridization, an adjusted array-hybridization intensity value (I) was determined as the mean of the log signals of all features that showed minimal variation across all hybridizations. Then, for each feature on the array, a linear regression of the logarithm on I was calculated using the least squares method, first under the null hypothesis that the reference (P. falciparum strain 3D7) and test P. falciparum strains had the same response and then under the alternative hypothesis that the reference (3D7) strain had a greater signal than the test strain. The models were compared with the F-test, and the same signal model was rejected in favor of a marker with 99% confidence.

Gene amplifications and deletions. Gene amplifications were identified by comparing an array hybridized with DNA for each strain to an array that had been hybridized with 3D7 DNA. Comparisons were made between arrays that appeared to be overall closest to one another in overall signal. As a first step, background was subtracted from probe signals, and overall array signals were normalized as previously described for the two arrays to be compared [14]. Background values were determined by analyzing nonhuman, non-P. falciparum control sequences on the array, also as previously described [14]. This step places the overall signal intensities of arrays in the same range as one another, but does not take into account the fact that strains that are different from the sequenced strain, 3D7, are expected to give an overall lower intensities because they contain SFPs. Thus, we computed an additional normalization factor, N, using the ratio of the overall signal intensities of the 3D7 array to the 3D7 array for each feature. This was calculated by computing the log2 of the median signal ratio between the test strain and 3D7 for all probes to all genes that contained no SFPs in any of the strains analyzed (~1,400 genes). In addition, probes with negative values in 3D7 or the test strain, or which showed log2 ratios of more than 4, or less than −4, were discarded.

The median value for all probe-sets was chosen as N. To identify duplications, all 3D7 probe values were multiplied by N, determined independently for each strain. Then the log2 of the ratio (R) of the
test signal relative to N-normalized 3D7 was calculated for each probe in each probe-set. In order to eliminate probes complementary to SFPs or which could behave as outliers, for each probe-set, probes were discarded if R was greater than 4 or less than -1. A paired t-test was performed on the R values for the remaining probes in the probe-set using Fisher’s r-to-z. Next, we identified genes with a median number of greater than 0.7 for the remaining probes. The number of remaining probes was greater than six and the t-test indicated a likely difference between the intensities of the probe signal for 3D7 relative to test strain. The t-test was added in order to distinguish between a probe-set being amplified and a probe-set being polymorphic at only several probes. The value of 0.7 (roughly a 1.5-fold increase in signal) was determined empirically, and is slightly less than the expected 2-fold change because array probe signals are not expected to be perfectly linear with respect to copy number, and because the ratios for amplified genes are expected to show a distribution of values. A comparison of the HB3 and 3D7 ratios (Figure S2) shows a tight distribution around zero, as expected, with a secondary peak around 0.8. Genes with values of greater than 0.7 in the histogram are found exclusively in two chromosomal blocks; one surrounding GTP cyclohydrolase (PFL1155w) on Chromosome 12 and one on Chromosome 11 at the PF10_0374 locus. However, while this method worked well for our data, it may not work as well for saturated arrays or low-intensity arrays. Potential duplications were visually examined to verify that the regions showed contiguous increases in signals. Short amplifications of small genes may have been missed.

To identify potential amplifications in 3D7, we compared the list of genes showing a 1.5-fold or greater change in 3D7 relative to each of the Senegal isolates (n = 5) as described above. This returned some genes that were likely to be highly polymorphic in individual Senegal strains. Subsequent analysis of such probes was not held in common by all the Senegal strains examined. These two genes were for GTP cyclohydrolase (PFL1155w) and P. falciparum 11-1 protein PF10_0374, the gene11-1 product, which is highly expressed during gameteogenesis. Examination of the ratios for these genes (Table S7) is consistent with a 3D7 amplification (generally from 1.5- to 4-fold changes) rather than a Senegal deletion where ratios show a 20-fold difference in signal. Quantitative real-time PCR analysis further confirmed a probable amplification in 3D7 for GTP cyclohydrase (Table S8).

SSS; SFPs were identified as follows. The custom-designed Affymetrix malaria full-genome array consists of 2,397 genes that were likely to be highly polymorphic in individual Senegal strains. Intensities from these probes represent the level of cross-hybridization for a deleted gene. A new probe-to-gene map was generated for the gene11-1 product, which is highly expressed during gameteogenesis. Examination of the ratios for these genes (Table S7) is consistent with a 3D7 amplification (generally from 1.5- to 4-fold changes) rather than a Senegal deletion where ratios show a 20-fold difference in signal. Quantitative real-time PCR analysis further confirmed a probable amplification in 3D7 for GTP cyclohydrase (Table S8).

In contrast, gene deletions were identified as follows. The custom-designed Affymetrix malaria full-genome array consists of 2,397 probes for 100 viral genes that serve as background controls [14]. Intensities from these probes represent the level of cross-hybridization for a deleted gene. A new probe-to-gene map was generated to include both sense and antisense probes, and the MOID algorithm [59] was applied to assign “present” and “absent” calls to each gene. Based on all the background control data collected in this study, analysis, similar to that described in Le Roch et al. [14], shows that a deleted gene has only a 2% chance to be misclassified as “absent” if it is required to have a minimum intensity level of E > 10 and a Kolmogorov–Smirnov test of log10P of less than –0.5. Excluding genes with less than six probes and the highly variable var, rifin, and stevor genes, we found a total of 33 genes being called “absent” in at least two out of the three hybridizations for each strain (Table S9).

### Supporting Information

**Figure S1.** Linkage Disequilibrium Analysis

Shown is the linkage disequilibrium analysis of SFP-derived haplotypes obtained from 4,369 SFPs distinguishing the chloroquine resistant strains FCB and Dd2, across the 14 P. falciparum chromosomes. Scores were generated by calculating the probability of observing the same genotype by chance over a moving 40 kb window (with the probability of observing the same genotype for any one SFP by chance placed at 0.33). The plot shows the ratio between the probability and the maximum possible probability for regions with at least four SFPs with 1 indicating the best possible score. The position of all markers, variation clusters (var, rifins, and stevor genes), are shown in blue and are marked. SFPs mapping to these genes were excluded from the calculations because our data indicates that mitotic recombination may be occurring in these genes. pfet, which is located between bases 507,926 and 511,029 on chromosome 7 is shown as a black triangle. A: Chromosomes 1–7, B: 8–14.

Found at DOI: 10.1371/journal.ppat.0020057.s001 (2.4 MB PDF).

**Figure S2.** A Ratio Histogram of the HB3/3D7 Signal Intensities

As expected, a distribution around zero is observed, with a secondary peak around 0.8, for 4,300 genes with more than six probes, excluding the var, rifin, and stevor genes. Genes with log2 ratios of greater than 0.7, and classified as gene amplifications in HB3, are found exclusively in two chromosomal blocks: one on Chromosome 12 surrounding the GTP-cyclohydrolase gene (PFL1155w) and one on Chromosome 11, encompassing the region from PF11_0504 to PF11_0513.

Found at DOI: 10.1371/journal.ppat.0020057.s002 (22 KB TIF).

**Table S1.** Probe File (38Mb) for the SCRNamida Array

This file contains data for 327,989 P. falciparum-specific 25 mer probes selected for use in this analysis. Probes located in both the coding and non-coding sequences from both strands are included. Non-specific and un-mapped probes have been removed. This file does not contain probe sets from the 25 mer probes for Pfmdr1 specific control probes and probes from human and mouse are marked as controls in the “Reporter Usage” column. Generic mismatches are marked as “background” in the “Reporter Usage” column.

Found at DOI: 10.1371/journal.ppat.0020057.s001 (6.2 MB ZIP).
Table S2. SFP Data by Strain
This tab-separated file shows the strain hybridization pattern for every *P. falciparum* isolate for the 23,053 SFPs. Found at DOI: 10.1371/journal.ppat.0020057.st002 (3.2 MB XLS).

Table S3. SFP Data by Gene
This tab-separated file shows SFP data compiled by gene. Found at DOI: 10.1371/journal.ppat.0020057.st003 (257 KB TXT).

Table S4. Polymorphism Variation across the *P. falciparum* Strains Analyzed
Found at DOI: 10.1371/journal.ppat.0020057.st004 (55 KB PDF).

Table S5. The Number of Polymorphic Probes per Cluster of Genes Expressed at Each Stage of the *P. falciparum* Life Cycle
Clusters 1–15, along with representative genes for each cluster, are further described in Le Roch et al. [14]. Essentially, genes were grouped into clusters, based on the expression correlation throughout the *P. falciparum* life cycle, by a robust k-means program. While many of the genes belonging to each cluster encode proteins with unknown function, those genes with defined cellular roles permitted a life-cycle stage to be attributed to each of the 15 clusters defined. Virtually all known sporozoite-specific genes, such as the CS protein, reside within cluster 1, and most genes that have been identified in invasion, such as the MSPs, are found in cluster 15, validating the cluster assignments. Found at DOI: 10.1371/journal.ppat.0020057.st005 (57 KB PDF).

Table S6. Highly Variable, Uncharacterized Genes (n = 168)
Genes that are likely to be deleted are underlined. Found at DOI: 10.1371/journal.ppat.0020057.st006 (73 KB PDF).

Table S7. A List of the Highly Variable, Annotated Genes Observed in the Laboratory Strains (n = 54)
Aside from those deleted genes (underlined), no distinct differences were observed between the classes of highly variable genes within the laboratory strains compared to the Senegal strains (Table 5). Indeed, no significant differences were observed in the variation within immunogens (\(p = 0.87\)) and protein biosynthesis (\(p = 0.45\)) between the two groups of isolates, but variation in multi-gene families was significant (\(p = 0.37\)–58). Found at DOI: 10.1371/journal.ppat.0020057.st007 (66 KB PDF).

Table S8. A List of the Highly Variable, Annotated Genes in the *P. falciparum* Strains from Senegal (n = 235)
The notably higher variation observed in the Senegal strains at the glycosphosphoinositol-binding protein-related antigen (PF14_0010) locus is due to a deletion of this gene (underlined) in strain 51.02. Found at DOI: 10.1371/journal.ppat.0020057.st008 (64 KB PDF).

Table S9. Variation in Characterized Parasite Immunogens in Clinical Development
Candidate parasite immunogens were derived from Richier and Saul [2].

References

Acknowledgments
Author contributions. CK, SKV, SB, DFW, and EAW conceived and designed the experiments. CK and DP performed the experiments. CK, SKV, JD, JOB, YZ, JRJ, SB, DFW, and EAW analyzed the data. CK, SKV, JD, JOB, YZ, JRJ. KGRL, OS, ON, SM, SB, DFW, and EAW contributed reagents/materials/analysis tools. CK and EAW wrote the paper.

Funding. E. A. W is supported by a New Scholars Award from the Ellison Foundation and a 2004 Keck Distinguished Young Scholar in Medical Research award from the W. M. Keck Foundation and by funding by NIH Grant NIH5RO3AI054687-02. We would also like to acknowledge the Malaria Research and Reference Reagent Resource Center (MR4) and GCRC (funded by NIH grant MOIR00835) for providing biological reagents.

Competing interests. The authors have declared that no competing interests exist.

Found at DOI: 10.1371/journal.ppat.0020057.st002 (3.2 MB XLS).

Table S10. Predicted Copy Number for Amplified Genes
The log2 (Ratio) and copy number, for all *P. falciparum* Strains, of the GTP-Cyclohydrolase I (PFL1155w), the two flanking genes (PFL1150c and PFL1160w), as well as an uncharacterized control gene (PFL1150w), also located on Chromosome 12. The ratio and copy number of the *P. falciparum* multidrug-resistance protein, *pfmdr1* (PFL1150w) and the *P. falciparum* 11-1 protein (PF10_0374) are also represented. A ratio histogram of the log2 (Ratio) for the HB3/3D7 signal intensity (Figure S1) verifies the thresholds used to identify gene amplifications. The GTP-cyclohydrolase copy numbers were extrapolated based on three copies in 3D7, and gene *pf11-1* was extrapolated based on two copies. Found at DOI: 10.1371/journal.ppat.0020057.st010 (56 KB PDF).

Table S11. Threshold Cycle Values Derived from the Real-Time PCR Amplification of the *P. falciparum* Genes (PFL1155w, PFL1145w, PFL1150c, and PFL1150w), along with a β-Tubulin Control Gene (PF10_0084)
A parallel reaction was carried out for each primer pair, and the mean threshold cycle (Ct) value and standard deviation are presented. Found at DOI: 10.1371/journal.ppat.0020057.st011 (58 KB PDF).

Table S12. Gene Deletions Results of the MOID Analysis Showing Genes Detected as “Present”
Found at DOI: 10.1371/journal.ppat.0020057.st012 (8.9 MB XLS).

Table S13. Description of the Genes Deleted in the *P. falciparum* Strains Analyzed; Deletions Are Reported Only for Genes with ≥ Six Probes
No deletions in strain D6 were observed. Found at DOI: 10.1371/journal.ppat.0020057.st013 (75 KB PDF).

Table S14. Oligonucleotide Primer Sequences
Found at DOI: 10.1371/journal.ppat.0020057.st014 (54 KB PDF).


