A Systematic Map of Genetic Variation in *Plasmodium falciparum*

Claire Kidgell[1], Sarah K. Volkman[2], Johanna Daily[2], Justin O. Borevitz[3¤], David Plouffe[4], Yingyao Zhou[4], Jeffrey R. Johnson[1], Karine G. Le Roch[4], Ousmane Sarr[5], Omar Ndir[5], Souleymane Mboup[5], Serge Batalov[4], Dyann F. Wirth[2], Elizabeth A. Winzeler[1,4*]

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 ≈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

Abbreviations: CLAG, cytoadherence-linked asexual protein; CS, circumsporozoite; CSA, chondroitin sulfate A; HT, host-targeting; MOID, match-only integral distribution; MSP, merozoite surface protein; RESA, ring-infected erythrocyte surface antigen; SFP, single-feature polymorphism; SNP, single-nucleotide polymorphism; SP, signal peptide; TM, transmembrane

* To whom correspondence should be addressed. E-mail: winzeler@scripps.edu

¤ Current address: Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America

Synopsis

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.

solely by meiotic recombination and not through the rapid emergence of novel alleles and haplotypes via mitotic recombination and mutation, which may occur in *P. falciparum.*

Allelic-variation scanning [12], which is a hybridization-based method for discovering and tracking natural variation, has the advantage that no libraries are required, sampling is consistent and is based on an array template, no primers need to be designed, and tens of thousands of genetic markers can be both discovered and typed in as little as one day in any parasite isolate, potentially using only a few milliliters of infected patient blood as starting material. While the nature of this technique dictates that all observed variation will be relative to the sequenced strain represented on the array, reproducibility is high and the method can be used for genotyping and mapping traits in genetic crosses or through the analysis of linkage disequilibrium and, as with full-genome sequencing, the discovery of genes under selection [13]. Most importantly, the method easily discovers genome-amplification events, which generally may not be readily revealed by even laborious full-genome sequencing. The simplicity of this approach could allow full-genome longitudinal studies of parasite variability to be performed on large numbers of samples in response to the introduction of a new drug or vaccine.

Here, we systematically sample genetic variation in the malaria parasite by analyzing the patterns that are derived from hybridizing genomic DNA to an oligonucleotide array covering approximately 50% of the coding regions of the *P. falciparum* genome [14]. We demonstrate a highly nonrandom distribution of variability in terms of functional classes of genes, and we reveal genes that appear to be under increased selection pressure from the host immune system. Understanding patterns of global natural variation in malaria parasites can also assist in the functional characterization of the genes encoded by the parasite's genome. Applying a "guilt by association" principle, we hypothesize that those highly variable, uncharacterized genes that show similar profiles of host selection, expression pattern, and localization to validated immunogens might make promising additional vaccine candidates for further investigation. We also report

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 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,

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

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

CQ, chloroquine; CYC, cycloguanil; PYR, pyrimethamine; QN, quinine; SDX, sulfadoxine.
[a]*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].
[b]Drug-resistance profiles were obtained from [19,55,79].
DOI: 10.1371/journal.ppat.0020057.t001

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, were 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 331,456 and 407,273 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.

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

Strain	PlasmoDB_ID	Gene_ID	Number of Polymorphic Probes Identified	Total Number of Probes	Number of Non-Synonymous Mutations	Number of Synonymous Mutations	Total Number of Probes Re-Sequenced	Number of False-Positive Probes
W2	PF10_0348	810505	5	68	5[a]	2	5	0
	PFE1600w	813129	3	52	3	0	3	1
	PFL0800c	811213	3	22	3	0	3	0
51.02	PF10_0348	810505	7	68	7[a]	2	7	0
	PFE1600w	813129	5	52	5	0	5	0
	PFL0800c	811213	3	22	4[a]	0	3	0

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.
[a]Indicates that more than one nucleotide substitution was observed within a single polymorphic 25-mer feature.
Gene_IDs were derived from Entrez Gene at NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene).
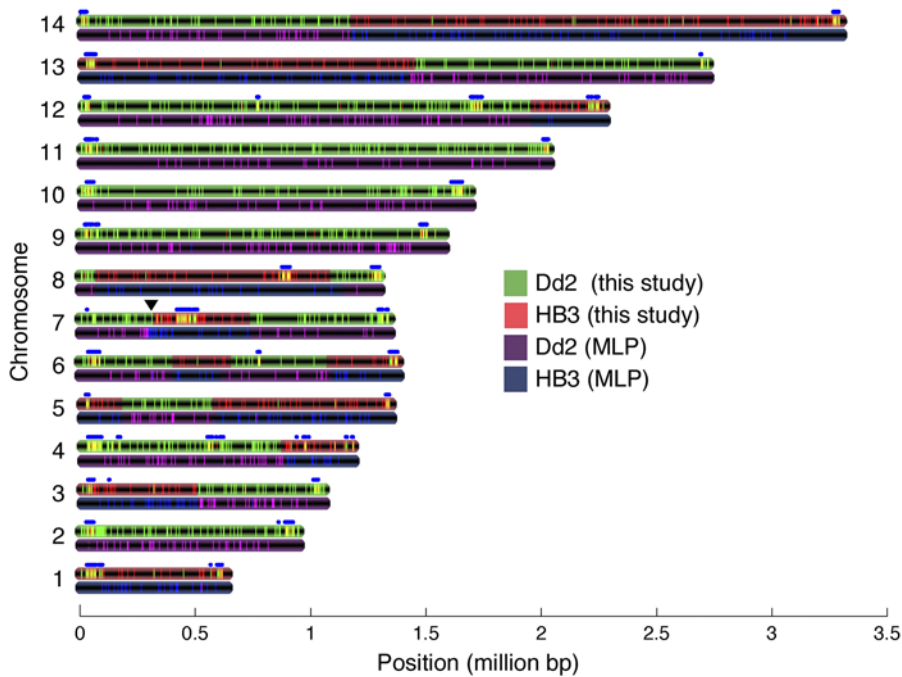DOI: 10.1371/journal.ppat.0020057.t002

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

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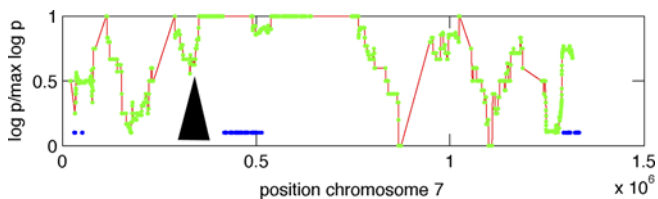 each chromosome was observed, as previously noted on Chromosome 2 [17]. Non-variable genes (more than ten probes, <10% SFP rate) accounted for 82% of the total genes located in the central regions of the chromosome (4,077/4,998 genes). In contrast 7,555 (≈75%) of the 10,122 probes in subtelomeric regions were polymorphic versus 12,978 (≈5%) of the remaining 247,910 probes ($p \leq 1.0E{-}100$). 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 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 ($n = 580$), while for expressed variable genes (SFP rate >10%, more than nine probes) it was 0.70 ($n = 99$) (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 ≈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 *stevor* genes ($n \approx 290$), 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, *pfcrt (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 ortholog/paralog 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 *stevor* 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 *vars*, *stevors*, and *rifins* 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 (SICA*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]. Immunoepidemiologic 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 polymorphism data available for approximately 70 of the annotated *var* genes, 17/70 genes showed significantly less variability (less than ten SFPs) than the remaining 53 genes, such as the previously characterized *var2CSA* gene *PFL0030c* [37]. Eleven out of 17 of these "conserved" *var* genes are annotated as truncated coding sequences, and one as a "pseudogene". The inactivation of these coding sequences over time may be a cause or consequence of their specific role in malaria pathogenesis.

Immunogens

Since certain blood-stage antigens of the malaria parasite, such as MSP-1 *(PFI1475w)*, are highly expressed when the parasite invades the erythrocyte, they are likely to be under selection pressure from the host, and consequently are believed to display elevated rates of nucleotide substitution compared to the genome as a whole [40]. Such variable genes are thought to be the principal targets of protective immunity. To test the hypothesis that such genes would be under positive selection, we compared the SFP rate of probable immunogens to housekeeping genes. The SFP-to-probe ratio was significantly higher for characterized parasite immunogens in clinical development (CSP *[PFC0210c]*; TRAP *[PF13__0201]*; LSA1 *[PF10__0356]*; LSA3 *[PFB0915w]*; MSP-1 *[PFI1475w]*; MSP-2 *[PFB0300c]*; MSP-3 *[PF10__0345]*; AMA1 *[PF11__0344]*; SERA *[PFB0340c]*; GLURP *[PF10__0344]*; RESA *[PFA0110w]*; and EXP1 *[PF11__0224]* [2] (220 SFPs/882 probes), compared with either housekeeping genes (such as those with roles in protein biosynthesis [49 SFPs/2,034 probes; $p = 4.62E–76$]) or compared with genes encoding proteins that are localized to the mitochondrion (114 SFPs/4,185 probes; $p = 3.24E–94$). Seven of these 11 characterized immunogens encode genes that are amongst the top 20 most highly variable, non-deleted annotated genes (Table 3); a $\approx$17-fold enrichment in the group of highly variable genes is shown relative to the genome as a whole. Between the 14 *P. falciparum* strains analyzed, strains Dd2 (Thailand), FCR3 (the

Gambia), and FCB (Thailand) display the highest levels of polymorphisms in characterized parasite immunogens, while strains D6 (Sierra Leone) and 7G8 (Brazil) showed low levels of variation at these specific loci (Table S9). Overall, the laboratory strains show greater variation at these 11 loci compared to the Senegal strains, particularly at the MSP-1 locus *(PFI1475w)* (Table S9).

A conserved pentameric sequence (Pexel motif) [41] and an 11-amino acid signal (host-targeting [HT] motif) [42], which are frequently found in parasite proteins that are targeted to the red blood cell, have been identified, and such proteins have been proposed as candidates for proteins interacting with the host immune system. Excluding the highly variable *var*, *rifin*, and *stevor* genes, we observed a large enrichment for polymorphisms in genes for those proteins bearing a Pexel motif (609 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 *(PFI1475w)* 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* [43], 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 sites in a 424-amino acid protein [45]. In contrast, introns of housekeeping genes show low SNP rates (one SNP/27,000 bases) [23]. The list of highly variable genes (Table 3) includes the gene that is known to play a role in modulating antimalarial drug resistance, the *P. falciparum* chloroquine-resistance transporter *(pfcrt)* (MAL7P1.27) (four SFPs/36 probes; 11% SFP rate) [46]. The well-known drug-resistance gene dihydrofolate reductase *(dhfr)* *(PFD0830w)* [47], also shows an increased rate of variability (five SFPs/54 probes; 9% SFP rate) compared to the neutral rate (4%) [24].

As many antimalarials are derived from natural products that are likely to have off-target effects and to contribute to parasite toxicity, other enzymes that display variation (Table 3) may represent additional off-site targets. Such candidates potentially include heme A farnesyltransferase *(PFE0970w)* (seven SFPs/52 probes; 13 % SFP rate) or the heme-binding protein *(PFI0885w)* (three SFPs/17 probes; 17% SFP rate), which could be targeted by antimalarials that bind to heme

Table 3. Highly Variable Annotated Genes (n = 66)

PlasmoDB_ID	Gene_ID	Description	Number of SFPs	Total Number of Probes	Life-Cycle Stage-Expression Le Roch et al. [14]	Bozdech et al. [25]	Pexel Motif [41]	HT Motif [42]	SignalP	TM Domain
PF10_0343	810500	S-antigen	10	11	Schizont	Late schizont	No	No	Yes	0
PFB0100c	812616	Knob-associated histidine-rich protein	56	70	Ring/trophozoite	Late ring	Yes	Yes	No	0
PF14_0010	811601	Glycophorin-binding protein–related antigen	14	19	Trophozoite	0	Yes	Yes	No	1
PFB0095c	812615	Erythrocyte membrane protein 3	60	76	Ring/trophozoite	0	Yes	Yes	Yes	1
PFI1710w	813622	Cytoadherence-linked protein	38	52	Schizont/gametocyte	0	No	No	No	0
PFI1800w	813640	Enzyme, putative	22	31	Low expression	0	No	No	Yes	0
PFD1205w	812411	Predicted integral membrane protein	12	17	Merozoite	0	No	No	Yes	2
PFI1475w	813575	MSP-1, precursor	106	162	Schizont	Late schizont	No	No	Yes	1
PFC0110w	814340	CLAG	21	42	Schizont	Late schizont	No	No	Yes	0
PF10_0345	810502	MSP-3	14	31	Schizont	Late schizont	No	No	Yes	0
PFL2570w	811564	Acyl-coa ligase antigen	22	52	Schizont/gametocyte	0	No	No	Yes	0
PFC0120w	814342	CLAG	23	57	Schizont	Late schizont	No	No	Yes	0
MAL13P1.176	813730	*P. falciparum* reticulocyte-binding protein 2	19	53	Schizont	Late schizont	No	No	No	1
PFL1155w	811283	GTP cyclohydrolase I[a]	11	33	Trophozoite/gametocyte	Late trophozoite	No	No	No	0
PF13_0201	814170	Sporozoite surface protein 2	19	61	Sporozoite	0	No	No	Yes	0
PFB0300c	812660	MSP-2 precursor	12	40	Schizont	Very late schizont	No	No	No	0
PF10_0356	810513	Liver stage antigen, putative	16	55	Low expression	0	No	No	No	1
PF11_0344	810891	Apical membrane antigen 1 precursor	18	65	Schizont	Very late schizont	No	No	Yes	1
PF13_0115	814086	Frameshifted ebl1, pseudogene	62	239	Low expression	0	No	No	No	0
MAL7P1.159	2655035	Antioxidant protein, putative	3	12	Low expression	0	No	No	No	0
PF10_0346	810503	MSP-6	6	24	Schizont	Late schizont	No	No	No	0
PF10_0051	810209	ADP/ATP carrier protein, putative	7	32	Constitutive	0	No	No	No	5
PF10_0374	810531	Gene 11–1 protein precursor	97	470	Sporozoite/gametocyte	0	No	No	No	1
PFC0210c	814364	CS protein	5	25	Sporozoite	0	No	No	No	1
PF08_0071	2655270	Fe-superoxide dismutase	4	20	Trophozoite	0	No	No	No	0
PF07_0128	2654998	Erythrocyte-binding antigen	25	127	Ring	0	No	No	No	0
PF11_0161	810708	Falcipain-2 precursor, putative	3	16	Ring/trophozoite	Late ring	No	No	No	1
PFC0050c	814328	Long chain fatty acid ligase, putative	15	85	Constitutive	0	No	No	Yes	0
PFI0885w	813457	Heme-binding protein, putative	3	17	Ring/trophozoite	Very late schizont	No	No	No	1
PFC0310c	814391	ATP-dependent CLP protease, putative	4	23	Trophozoite	Very late schizont	No	No	Yes	0
PF13_0197	814166	MSP 7	5	29	Schizont	Late schizont	No	No	Yes	0
PFE1350c	813086	Ubiquitin-conjugating enzyme, putative	3	18	Constitutive	0	No	No	No	0
MAL7P1.86	2655049	Transcription initiation factor iie, alpha subunit	5	30	Constitutive	0	No	No	No	0
MAL13P1.337	813904	Skp1 family protein, putative	2	12	Trophozoite	Very late schizont	No	No	No	0
PFB0310c	812662	MSP-4	4	26	Sporozoite/schizont	Very late schizont	No	No	No	2
PFC0870w	814510	Elongation factor 1 (EF-1), putative	2	13	Constitutive	0	No	No	No	0
PFB0935w	812791	CLAG-2	18	121	Schizont	Late schizont	No	No	Yes	0
PF07_0046	2655094	50S ribosomal protein L1, putative	4	28	Ring/trophozoite	0	Yes	No	No	0
PFE1430c	813100	Cyclophilin, putative	3	22	Constitutive	0	No	No	No	0
PFE0970w	813009	Cytochrome c oxidase assembly protein (heme A: farnesyltransferase), putative	7	52	Trophozoite	0	No	No	No	4
PFC0255c	814375	Ubiquitin-conjugating enzyme E2, putative	2	15	Constitutive	0	No	No	No	0
MAL7P1.26	2655198	O-sialoglycoprotein endopeptidase, putative	2	15	Constitutive	Late ring	No	No	No	0
PFB0687c	812738	RING zinc finger protein, putative	3	23	Schizont	0	No	No	No	0
PF11_0509	811041	RESA, putative	12	95	Ring	0	Yes	No	No	0
PF07_0071	2655115	Queuine tRNA-ribosyltransferase; putative	7	56	Ring/early trophozoite	0	No	No	No	0
PFL2000w	811452	Mitochondrial carrier protein, putative	3	24	Constitutive	0	No	No	No	4

Table 3. Continued

PlasmoDB_ID	Gene_ID	Description	Number of SFPs	Total Number of Probes	Life-Cycle Stage–Expression		Pexel Motif [41]	HT Motif [42]	SignalP	TM Domain
					Le Roch et al. [14]	Bozdech et al. [25]				
PFE0070w	812870	Interspersed repeat antigen, putative	9	76	Ring	Very late schizont	No	No	No	0
PFI0955w	813471	Sugar transporter, putative	4	34	Sporozoite	0	No	No	No	11
PFI1030c	813486	Ubiquitin conjugating enzyme E2, putative	2	17	Constitutive	0	No	No	No	0
PF11_0165	810712	Falcipain 2 precursor	2	17	Ring/trophozoite	Late ring	No	No	No	1
PFE0845c	812984	60S ribosomal subunit protein L8, putative	3	26	Ring/trophozoite	Late ring	No	No	No	0
PFE1420w	813098	F-actin capping protein, alpha subunit, putative	3	26	Trophozoite/gametocyte	0	No	No	No	0
PFL1920c	811436	Hydroxyethylthiazole kinase, putative	3	26	Constitutive	0	No	No	No	0
PF10_0025	810183	PF70 protein	5	44	Ring	Very late schizont	Yes	Yes	No	1
MAL7P1.27	2655199	Chloroquine-resistance transporter, putative	4	36	Ring/trophozoite	Late ring	No	No	No	10
PFI1670c	813614	Vacuolar ATP synthase subunit E, putative	2	18	Constitutive	0	No	No	No	0
PF11_0224	**810771**	**CS-related antigen**	**2**	**18**	**Ring/trophozoite**	**Very late schizont**	**No**	**No**	**Yes**	**1**
PF11_0461	811004	Rab6	2	18	Schizont	0	No	No	No	0
PFB0340c	**812668**	**Cysteine protease, putative**	**11**	**100**	**Schizont/gametocyte**	**Late schizont**	**No**	**No**	**No**	**0**
PF11_0257	810804	Ethanolamine kinase, putative	4	37	Schizont/gametocyte	Late schizont	No	No	No	0
MAL7P1.38	2655211	Regulator of chromosome condensation protein	7	65	Constitutive	0	No	No	No	0
PF10_0225	810382	Orotidine-monophosphate-decarboxylase, putative	3	28	Constitutive	0	No	No	No	0
PFB0355c	812671	Cysteine protease, putative	10	95	Schizont	0	No	No	No	0
PFD1155w	812436	Erythrocyte-binding antigen, putative	17	163	Schizont	Late schizont	No	No	Yes	1
MAL6P1.296	2654838	Nucleolar GTP-binding protein 1, putative	6	58	Constitutive	0	No	No	No	0
PF08_0102	2655301	Asparagine-rich antigen Pfa55-14	8	78	Gametocyte	0	No	No	No	0

Those genes considered "absent" in one strain, based on the MOID algorithm, are underlined. 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 Richie 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).
[a]Potential duplication.

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

Strain	Duplicated Region	Function	Representative Genes in Duplicated Region
FCB	*PFL1130c..PFL1165w*	Folate biosynthesis	GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24
HB3	*PFL1125w.. PFL1315w*		Proliferating cell nuclear antigen (PCNA), cytochrome c heme lyase, putative, GTP cyclohydrolase I
7G8	*PFL1130c..PFL1160c*		GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24
3D7	*PFL1155w*		GTP cyclohydrolase I
W2	*PFL1145w..PFL1160c*		GTP cyclohydrolase I, uncharacterized protein, ribosomal protein L24
Dd2	*PFL1145w..PFL1155w*		GTP cyclohydrolase I, cAMP-dependent protein kinase subunit, ribosomal protein L24, DNA GyrAse subunit, phospholipid-transporter ATPase
FCB	*PFD0090c..PFD0120w*	Surface proteins	Reticulocyte-binding protein, uncharacterized proteins
HB3	*PF11_0504..PF11_0513*		RESA and RESA-2
W2	*PFD0100c..PFD0115c; MAL7P1.52; PFL1365w.. PFL1385c*		Reticulocyte-binding protein, uncharacterized proteins; RIF pseudogene; NIMA-related protein kinase (Pfnek-1), 101 kd malaria antigen
D6	*PF07_0135*		Rifin
31.01	*PF10_0006..PF10_0013*		Hypothetical proteins, *pfemp1*
FCB	*PFE1065w..PFE1155c*	Antimalarial drug resistance	S-adenosylmethionine-dependent methyltransferase, multidrug-resistance protein, *pfmdr1*
FCR3	*PFE0940c.. PFE1175w*		Adenosylhomocysteinase(S-adenosyl-L-homocysteine hydrolase), S-adenosylmethionine-dependent methyltransferase, *pfmdr1*, cytochrome c oxidase assembly protein (heme A: farnesyltransferase), putative, vacuolar ATP synthetase, putative
Dd2	*PFE1095w.. PFE1160w*		S-adenosylmethionine-dependent methyltransferase, putative, *pfmdr1*
FCR3	*PFL0115w..PFL0130c*	Cyclophilins	Cyclophilin, uncharacterized proteins
W2	*MAL7P1.137.. PF07_0106*	Others	Mago nashi protein homolog, kelch protein, exonuclease i

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.t004

derivatives, such as chloroquine and an ADP/ATP carrier protein *(PF10__0051)* (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 *(PFD0110w)* 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 antimalarials (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)* *(PFD0830w)* [47] or dihydropteroate synthase *(dhps)* *(PF08__0095)* 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 *(PF10__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
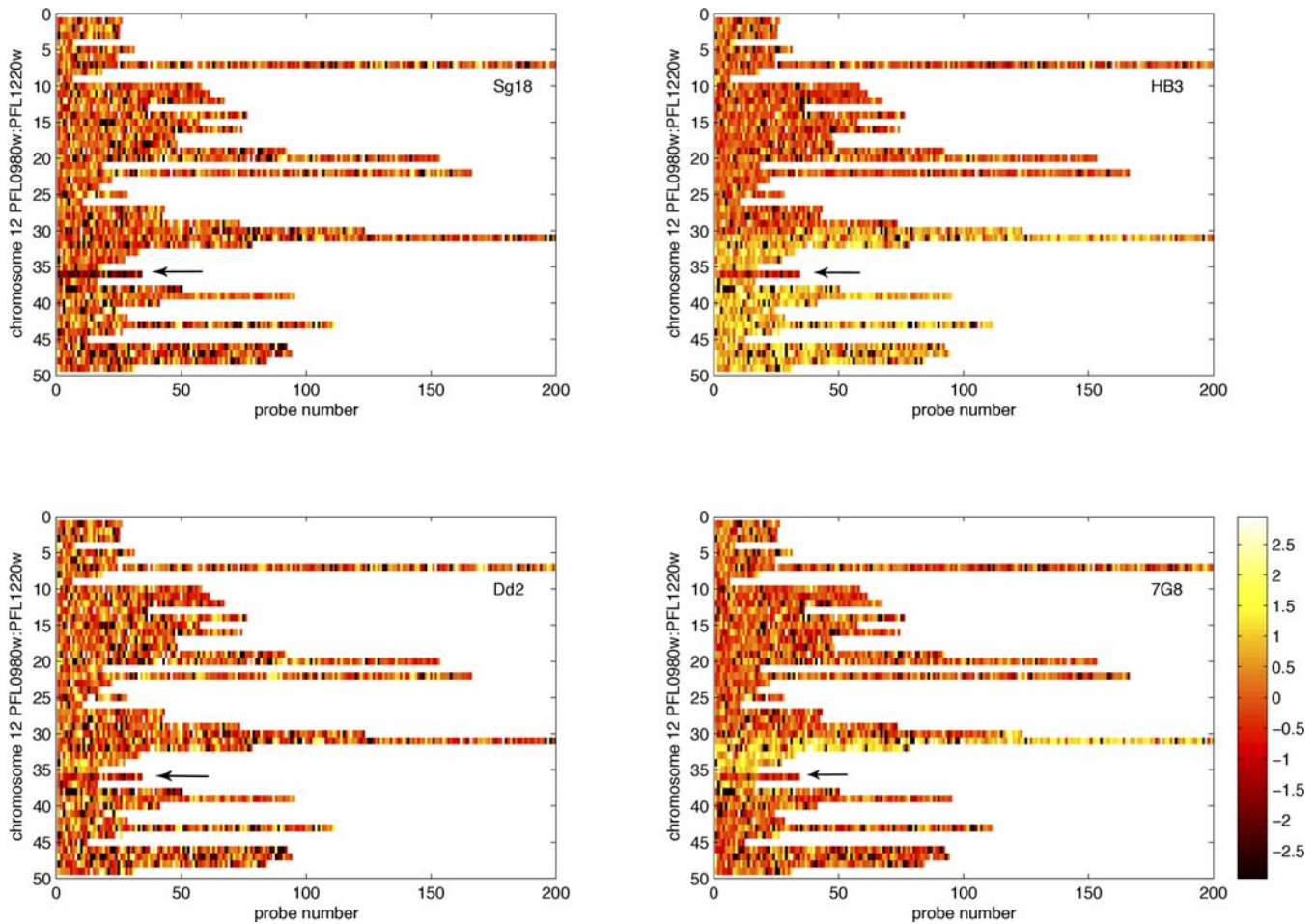
Figure 3. The Heat Map Shows the Base 2 Logarithm of the Ratio of the Normalized, Background Subtracted Probe Signal for the Listed Isolate (18.02, Dd2, 7G8 or HB3), Relative to the Sequenced Isolate, 3D7

All probes on the array were selected to be unique within 3D7. Each horizontal bar represents a single gene on the right arm of Chromosome 12. Dark probes, with low signal intensities, do not match any region in the genome of the listed isolate, while light probes, displaying high signal intensities, such as those near the GTP-cyclohydrolase gene *(PFL1155w)* in Dd2, HB3, and 7G8 contain matches to multiple regions of the genome. The signal intensity and the consecutive nature of the pattern suggest that strains 7G8, HB3, and Dd2 contain more than one copy of the GTP-cyclohydrolase gene *PFL1155w*. An arrow represents the location of the GTP-cyclohydrolase*PFL1155w* gene on Chromosome 12.
DOI: 10.1371/journal.ppat.0020057.g003

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-cyclo-hydrolase 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 *P. 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
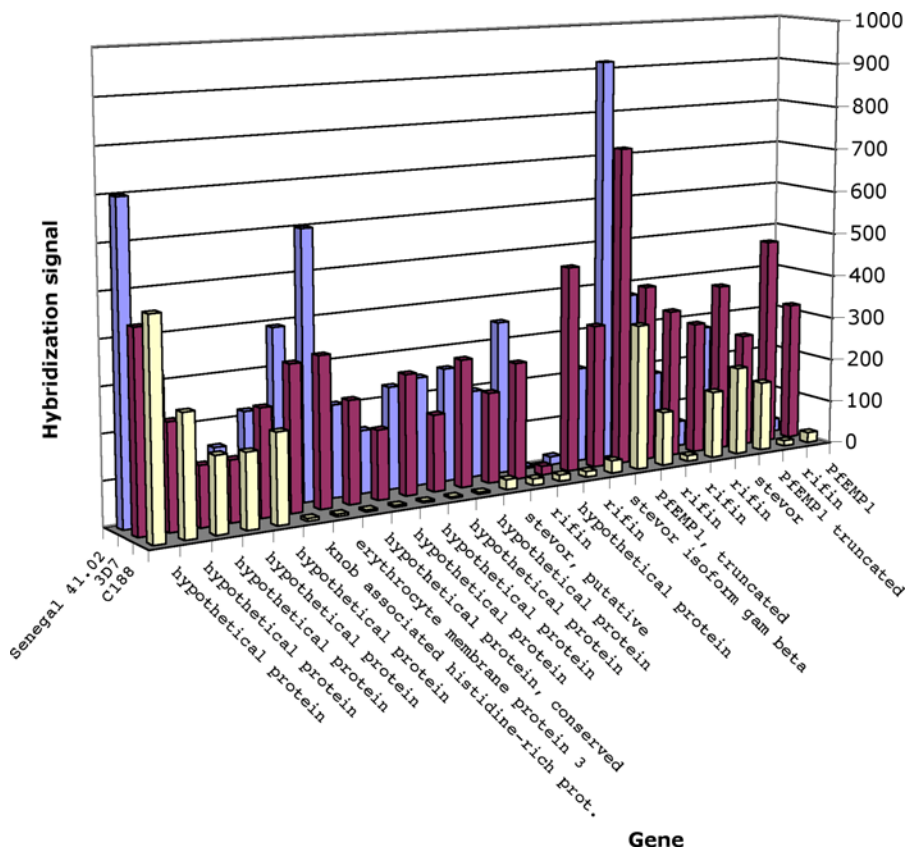
Figure 4. Example of a Deleted Region

Hybridization levels were calculated with the MOID algorithm as previously described [59]. The figure shows the signal for 26 genes beginning with *PFB0010w* and ending with *PFB0120w* for three strains (41.02, 3D7, and C188). The background level (zero) was determined by analyzing the signal of non-*P. falciparum* genes on the array. Data for all genes and all strains are available from Table S12.

DOI: 10.1371/journal.ppat.0020057.g004

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 1.0*E*–17 to 1.0*E*–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 (PFB0095c),* 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 1.0*E*–222 for strains where it was considered present and from 1.0*E*–3.6 to 1.0*E*–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 *(PFI1800w),* an uncharacterized protein *(PFI1785w),* and the cytoadherence-linked asexual protein (CLAG, *PFI1710w)* [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 *(PFI1475w).* 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 *(PFI1475w),* 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, potentially 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* have 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 produce as 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 5,179 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 are perfect reverse complements of another. Because forward and reverse probes 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 extraction. RNA 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 1 and end-labeled with biotin [12]. Hybridization was subsequently carried out to 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 streptavadin R-phycoerythrin 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* for all hybridizations was determined by 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* (usually resulting in a 5% to 10% change). 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 strain 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 −4. A paired *t*-test was performed on the *R* values for the remaining probes in the probe-set using $p \leq 0.01$. Next, we identified genes with a median *R* of greater than 0.7 for the remaining probes when 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 surrounding 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, but such deletions were not held in common by all. Only two genes were shared 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 gametocytogenesis. 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 cyclohydrolase (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, this 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 "present" if it is required to have both an 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).

Mapping inheritable markers. SFPs were identified in the strains Dd2, HB3, and C188, relative to 3D7 as described above. The inheritance of markers was determined by creating a matrix containing the SFP data for the two parental strains (Dd2 and HB3) and the progeny strain C188. The data were scored accordingly: if a marker is present in Dd2 and in the recombinant, but not in HB3, score as Dd2, but if a marker is in HB3 and the recombinant, but not Dd2, score as HB3. These data were then plotted relative to chromosome position, and subsequently potential recombination breakpoints could be clearly defined.

Analysis of gene-expression data. A tab-separated text file containing the correlation coefficient of expressed asexual erythrocytic genes across the *P. falciparum* strains 3D7, HB3, and Dd2, derived from the Affymetrix high-density oligonucleotide and glass-slide oligonucleotide microarray-expression platforms was obtained from PlasmoDB (http://www.plasmodb.org; Chris Stoeckert, personal communication). The average correlation coefficient of highly variable, expressed erythrocytic genes between strains 3D7 (Affymetrix platform, Winzeler laboratory) and strain HB3 (glass-slide platform, DeRisi laboratory) were derived.

Real-time PCR analysis. GTP-cyclohydrolase *(PFL1155w),* the uncharacterized protein *(PFL1145w),* ribosomal protein L24 *(PFL1150c),* and *pfmdr1 (PFE1150w)* copy number were assessed by

real-time PCR amplification using the BioRad MyiQ apparatus (Hercules, California, United States) in the presence of SYBR-green. The following oligonucleotide primers were designed to specifically amplify each gene: PFL1155w-F 5′-AAACACCATCTTTTACCTTTT GAA-3′; PFL1155w-R 5′-AGCATCGTGCTCTTTAACTCC-3′; pfmdr1-F 5′-CAAGCGGAGTTTTTGCATTT-3′; pfmdr1-R 5′-TTGAGCGCTTTGACTGAATC-3′; β-tubulin-F 5′-TCGTCAACTTCCTTTGTGGA-3′; β-tubulin-R 5′-TCCCATTCC CACGTTTACAT-3′; PFL1145w-F 5′-GGGTATGCCCCTGGAAG TAT-3′; PFL1145w-R 5′-TCCACATGCTCCACCAATAA-3′; PFL1150c-F 5′-ACGCAAGAAATGCCTATTCA-3′; and PFL1150c-R 5′-CTTTGATGGGAGGCCCTATT-3′. Parallel amplification reactions were carried out in 96-well plates in 25 μl containing 0.5 μm of each forward and reverse primer, 1 μl of template DNA, and 12.5 μl of iQSYBR green PCR master mix (50 mM KCl, 20 mM Tris-HCl, 0.2 mM each dNTP [dATP, dCTP, dGTP, and dTTP], 0.6U iTaq DNA polymerase, 3 mM MgCl$_2$, SYBR Green I, 10 nM fluorescein) (BioRad). Forty cycles were performed (95 °C 10 min, 95 °C 10 s, 55 °C 30 s, 60 °C 30 s) (Table S11).

Nucleotide sequencing. Oligonucleotide primers were designed to specifically amplify polymorphic regions of the *PF10_0348, PFL0800c,* and *PFE1600w* coding sequences (Table S14). Independent PCR-generated DNA fragments from both strands were sequenced using dRhodamine-labeled terminators. Nucleotide sequences were edited using the 4Peaks DNA analysis program v1.6 (http://www.mekentosj.com/4peaks).

Statistical analysis. The Fisher's exact test (http://www.matforsk.no/ola/fisher.htm) was used to determine the statistical relationship between each of the categorical variables in this analysis.

All the *P. falciparum* 3D7 genome data and gene annotations used for analysis in the work described in this paper are available for download at http://www.plasmodb.org. The gene-identification numbers (Gene_ID) were derived from Entrez Gene at NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene).

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 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 indicates 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. A: Chromosomes 1–7, B: 8–14.

Found at DOI: 10.1371/journal.ppat.0020057.sg001 (2.4 MB PDF).

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

As expected, a tight 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 classed 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.sg002 (22 KB TIF).

Table S1. Probe File (38Mb) for the SCRmalaria 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 any probes from the *Plasmodium yoelii* sequence. Affymetrix standard 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.st001 (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 23,653 SFPs.

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

Table S3. SFP Data by Gene

This tab-separated text 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 known roles 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 ($n = 34$)

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 ($p = 0.45$) genes between the two groups of isolates, but variation in multi-gene families was significant ($p = 6.37E{-}58$).

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 ($n = 25$)

The notably higher variation observed in the Senegal strains at the glycophorin-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 Richie and Saul [2].

Found at DOI: 10.1371/journal.ppat.0020057.st009 (56 KB PDF).

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 *(PFL0650c)*, also located on Chromosome 12. The ratio and copy number of the *P. falciparum* multidrug-resistance protein, *pfmdr1* *(PFE1150w)* 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 *PFE1150w)*, 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).

Acknowledgments

References

1. Biggs BA, Gooze L, Wycherley K, Wollish W, Southwell B, et al. (1991) Antigenic variation in *Plasmodium falciparum*. Proc Natl Acad Sci U S A 88: 9171–9174.
2. Richie TL, Saul A (2002) Progress and challenges for malaria vaccines. Nature 415: 694–701.
3. Wellems TE, Walker-Jonah A, Panton LJ (1991) Genetic mapping of the chloroquine-resistance locus on *Plasmodium falciparum* Chromosome 7. Proc Natl Acad Sci U S A 88: 3382–3386.
4. Roper C, Pearce R, Nair S, Sharp B, Nosten F, et al. (2004) Intercontinental spread of pyrimethamine-resistant malaria. Science 305: 1124.
5. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, et al. (2002) Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature 418: 320–323.
6. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, et al. (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Mol Biol Evol 17: 1467–1482.
7. Su X, Ferdig MT, Huang Y, Huynh CQ, Liu A, et al. (1999) A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. Science 286: 1351–1353.
8. Mu J, Duan J, Makova KD, Joy DA, Huynh CQ, et al. (2002) Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. Nature 418: 323–326.
9. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728–1732.
10. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
11. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511.
12. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, et al. (1998) Direct allelic variation scanning of the yeast genome. Science 281: 1194–1197.
13. Kidgell C, Winzeler EA (2005) Elucidating genetic diversity with oligonucleotide arrays. Chromosome Res 13: 225–235.
14. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503–1508.
15. Hughes AL, Friedman R (2005) Amino acid sequence constraint and gene expression pattern across the life history in the malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol 142: 170–176.

16. Anderson TJ, Su XZ, Roddam A, Day KP (2000) Complex mutations in a high proportion of microsatellite loci from the protozoan parasite *Plasmodium falciparum*. Mol Ecol 9: 1599–1608.

17. Volkman SK, Hartl DL, Wirth DF, Nielsen KM, Choi M, et al. (2002) Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. Science 298: 216–218.

18. Wellems TE, Panton LJ, Gluzman IY, do Rosario VE, Gwadz RW, et al. (1990) Chloroquine resistance not linked to mdr-like genes in a *Plasmodium falciparum* cross. Nature 345: 253–255.

19. Su X, Kirkman LA, Fujioka H, Wellems TE (1997) Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant *P. falciparum* in Southeast Asia and Africa. Cell 91: 593–603.

20. Martinelli A, Cheesman S, Hunt P, Culleton R, Raza A, et al. (2005) A genetic approach to the de novo identification of targets of strain-specific immunity in malaria parasites. Proc Natl Acad Sci U S A 102: 814–819.

21. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513–523.

22. Kyes S, Horrocks P, Newbold C (2001) Antigenic variation at the infected red cell surface in malaria. Annu Rev Microbiol 55: 673–707.

23. Volkman SK, Barry AE, Lyons EJ, Nielsen KM, Thomas SM, et al. (2001) Recent origin of *Plasmodium falciparum* from a single progenitor. Science 293: 482–484.

24. Rich SM, Licht MC, Hudson RR, Ayala FJ (1998) Malaria's Eve: Evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc Natl Acad Sci U S A 95: 4425–4430.

25. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1: E5.

26. Castillo-Davis CI, Bedford TB, Hartl DL (2004) Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. Mol Biol Evol 21: 1422–1427.

27. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature 419: 512–519.

28. Nielsen KM, Kasper J, Choi M, Bedford T, Kristiansen K, et al. (2003) Gene conversion as a source of nucleotide diversity in *Plasmodium falciparum*. Mol Biol Evol 20: 726–734.

29. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, et al. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. Nature 407: 1018–1022.

30. Paques F, Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev 63: 349–404.

31. Fairhead C, Dujon B (1993) Consequences of unique double-stranded breaks in yeast chromosomes: Death or homozygosis. Mol Gen Genet 240: 170–178.

32. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

33. Winter G, Kawai S, Haeggstrom M, Kaneko O, von Euler A, et al. (2005) SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. J Exp Med 201: 1853–1863.

34. Hughes MK, Hughes AL (1995) Natural selection on *Plasmodium* surface proteins. Mol Biochem Parasitol 71: 99–113.

35. Sam-Yellowe TY, Florens L, Johnson JR, Wang T, Drazba JA, et al. (2004) A *Plasmodium* gene family encoding Maurer's cleft membrane proteins: Structural properties and expression profiling. Genome Res 14: 1052–1059.

36. Daily JP, Le Roch KG, Sarr O, Ndiaye D, Lukens A, et al. (2005) In vivo transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. J Infect Dis 191: 1196–1203.

37. Rowe JA, Kyes SA, Rogerson SJ, Babiker HA, Raza A (2002) Identification of a conserved *Plasmodium falciparum* var gene implicated in malaria in pregnancy. J Infect Dis 185: 1207–1211.

38. Kyes SA, Christodoulou Z, Raza A, Horrocks P, Pinches R, et al. (2003) A well-conserved *Plasmodium falciparum* var gene shows an unusual stage-specific transcript pattern. Mol Microbiol 48: 1339–1348.

39. Elliott SR, Duffy MF, Byrne TJ, Beeson JG, Mann EJ, et al. (2005) Cross-reactive surface epitopes on chondroitin sulfate A-adherent P*lasmodium falciparum*-infected erythrocytes are associated with transcription of var2csa. Infect Immun 73: 2848–2856.

40. Conway DJ, Cavanagh DR, Tanabe K, Roper C, Mikes ZS, et al. (2000) A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. Nat Med 6: 689–692.

41. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306: 1930–1933.

42. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, et al. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306: 1934–1937.

43. Doolan DL, Southwood S, Freilich DA, Sidney J, Graber NL, et al. (2003) Identification of *Plasmodium falciparum* antigens by antigenic analysis of genomic and proteomic data. Proc Natl Acad Sci U S A 100: 9952–9957.

44. White NJ (2004) Antimalarial drug resistance. J Clin Invest 113: 1084–1092.

45. Bray PG, Martin RE, Tilley L, Ward SA, Kirk K, et al. (2005) Defining the role of PfCRT in *Plasmodium falciparum* chloroquine resistance. Mol Microbiol 56: 323–333.

46. Sidhu AB, Verdier-Pinard D, Fidock DA (2002) Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by pfcrt mutations. Science 298: 210–213.

47. Peterson DS, Walliker D, Wellems TE (1988) Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in *falciparum* malaria. Proc Natl Acad Sci U S A 85: 9114–9118.

48. Cowman AF, Galatis D, Thompson JK (1994) Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. Proc Natl Acad Sci U S A 91: 1143–1147.

49. Cravo PV, Carlton JM, Hunt P, Bisoni L, Padua RA, et al. (2003) Genetics of mefloquine resistance in the rodent malaria parasite *Plasmodium chabaudi*. Antimicrob Agents Chemother 47: 709–718.

50. Jeffress M, Fields S (2005) Identification of putative *Plasmodium falciparum* mefloquine resistance genes. Mol Biochem Parasitol 139: 133–139.

51. Triglia T, Duraisingh MT, Good RT, Cowman AF (2005) Reticulocyte-binding protein homologue 1 is required for sialic acid-dependent invasion into human erythrocytes by *Plasmodium falciparum*. Mol Microbiol 55: 162–174.

52. Triglia T, Cowman AF (1994) Primary structure and expression of the dihydropteroate synthetase gene of *Plasmodium falciparum*. Proc Natl Acad Sci U S A 91: 7149–7153.

53. Hayton K, Ranford-Cartwright LC, Walliker D (2002) Sulfadoxine-pyrimethamine resistance in the rodent malaria parasite *Plasmodium chabaudi*. Antimicrob Agents Chemother 46: 2482–2489.

54. Hossain T, Rosenberg I, Selhub J, Kishore G, Beachy R, et al. (2004) Enhancement of folates in plants through metabolic engineering. Proc Natl Acad Sci U S A 101: 5158–5163.

55. Rathod PK, McErlean T, Lee PC (1997) Variations in frequencies of drug resistance in *Plasmodium falciparum*. Proc Natl Acad Sci U S A 94: 9389–9393.

56. Wang P, Brobey RK, Horii T, Sims PF, Hyde JE (1999) Utilization of exogenous folate in the human malaria parasite *Plasmodium falciparum* and its critical role in antifolate drug synergy. Mol Microbiol 32: 1254–1262.

57. Wang P, Read M, Sims PF, Hyde JE (1997) Sulfadoxine resistance in the human malaria parasite *Plasmodium falciparum* is determined by mutations in dihydropteroate synthetase and an additional factor associated with folate utilization. Mol Microbiol 23: 979–986.

58. Carret CK, Horrocks P, Konfortov B, Winzeler E, Qureshi M, et al. (2005) Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. Mol Biochem Parasitol 144: 177–186.

59. Zhou Y, Abagyan R (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. BMC Bioinformatics 3: 3.

60. Corcoran LM, Forsyth KP, Bianco AE, Brown GV, Kemp DJ (1986) Chromosome size polymorphisms in *Plasmodium falciparum* can involve deletions and are frequent in natural parasite populations. Cell 44: 87–95.

61. Holt DC, Gardiner DL, Thomas EA, Mayo M, Bourke PF, et al. (1999) The cytoadherence linked asexual gene family of *Plasmodium falciparum*: Are there roles other than cytoadherence? Int J Parasitol 29: 939–944.

62. Sam-Yellowe TY, Fujioka H, Aikawa M, Hall T, Drazba JA (2001) A *Plasmodium falciparum* protein located in Maurer's clefts underneath knobs and protein localization in association with Rhop-3 and SERA in the intracellular network of infected erythrocytes. Parasitol Res 87: 173–185.

63. Plotkin JB, Dushoff J, Fraser HB (2004) Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature 428: 942–945.

64. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, et al. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82–86.

65. Hahn MW, Mezey JG, Begun DJ, Gillespie JH, Kern AD, et al. (2005) Evolutionary genomics: Codon bias and selection on single genomes. Nature 433: E5–E6.

66. Chen Y, Emerson JJ, Martin TM (2005) Evolutionary genomics: Codon volatility does not detect selection. Nature 433: E6–E7.

67. Nielsen R, Hubisz MJ (2005) Evolutionary genomics: Detecting selection needs comparative data. Nature 433: E6.

68. Jambou R, Legrand E, Niang M, Khim N, Lim P, et al. (2005) Resistance of *Plasmodium falciparum* field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. Lancet 366: 1960–1963.

69. Pearce R, Malisa A, Kachur SP, Barnes K, Sharp B, et al. (2005) Reduced variation around drug-resistant dhfr alleles in African *Plasmodium falciparum*. Mol Biol Evol 22: 1834–1844.

70. Mu J, Awadalla P, Duan J, McGee KM, Joy DA, et al. (2005) Recombination hotspots and population structure in *Plasmodium falciparum*. PLoS Biol 3: e335.

71. Trager W, Jenson JB (1978) Cultivation of malarial parasites. Nature 273: 621–622.

72. Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, et al. (1987) Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 236: 1661–1666.

73. Jensen JB, Trager W (1978) *Plasmodium falciparum* in culture: Establishment of additional strains. Am J Trop Med Hyg 27: 743–746.

74. Burkot TR, Williams JL, Schneider I (1984) Infectivity to mosquitoes of *Plasmodium falciparum* clones grown in vitro from the same isolate. Trans R Soc Trop Med Hyg 78: 339–341.

75. Oduola AM, Weatherly NF, Bowdre JH, Desjardins RE (1988) *Plasmodium falciparum:* Cloning by single-erythrocyte micromanipulation and heterogeneity in vitro. Exp Parasitol 66: 86–95.

76. Bhasin VK, Trager W (1984) Gametocyte-forming and non-gametocyte-forming clones of *Plasmodium falciparum*. Am J Trop Med Hyg 33: 534–537.

77. Delemarre BJ, van der Kaay HJ (1979) Tropical malaria contracted the natural way in the Netherlands. Ned Tijdschr Geneeskd 123: 1981–1982.

78. Ponnudurai T, Leeuwenberg AD, Meuwissen JH (1981) Chloroquine sensitivity of isolates of *Plasmodium falciparum* adapted to in vitro culture. Trop Geogr Med 33: 50–54.

79. Fidock DA, Rosenthal PJ, Croft SL, Brun R, Nwaka S (2004) Antimalarial drug discovery: Efficacy models for compound screening. Nat Rev Drug Discov 3: 509–520.