nature genetics

A genome-wide map of diversity in *Plasmodium falciparum*

Sarah K Volkman[1,8], Pardis C Sabeti[2,8], David DeCaprio[2], Daniel E Neafsey[2], Stephen F Schaffner[2], Danny A Milner, Jr[1], Johanna P Daily[1], Ousmane Sarr[3], Daouda Ndiaye[3], Omar Ndir[3], Soulyemane Mboup[3], Manoj T Duraisingh[1], Amanda Lukens[1], Alan Derr[2], Nicole Stange-Thomann[2], Skye Waggoner[2], Robert Onofrio[2], Liuda Ziaugra[2], Evan Mauceli[2], Sante Gnerre[2], David B Jaffe[2], Joanne Zainoun[2], Roger C Wiegand[2], Bruce W Birren[2], Daniel L Hartl[4], James E Galagan[2], Eric S Lander[2,5–7] & Dyann F Wirth[1,2]

Genetic variation allows the malaria parasite *Plasmodium falciparum* to overcome chemotherapeutic agents, vaccines and vector control strategies and remain a leading cause of global morbidity and mortality[1]. Here we describe an initial survey of genetic variation across the *P. falciparum* genome. We performed extensive sequencing of 16 geographically diverse parasites and identified 46,937 SNPs, demonstrating rich diversity among *P. falciparum* parasites ($\pi = 1.16 \times 10^{-3}$) and strong correlation with gene function. We identified multiple regions with signatures of selective sweeps in drug-resistant parasites, including a previously unidentified 160-kb region with extremely low polymorphism in pyrimethamine-resistant parasites. We further characterized 54 worldwide isolates by genotyping SNPs across 20 genomic regions. These data begin to define population structure among African, Asian and American groups and illustrate the degree of linkage disequilibrium, which extends over relatively short distances in African parasites but over longer distances in Asian parasites. We provide an initial map of genetic diversity in *P. falciparum* and demonstrate its potential utility in identifying genes subject to recent natural selection and in understanding the population genetics of this parasite.

P. falciparum is a human pathogen that kills 1–2 million people each year—mostly young children in Africa[1]. The parasite uses genetic variability to defeat host immunity and drug treatments. The full genome of a single clone, 3D7, has been sequenced[2], but it provided little insight into genetic variation responsible for phenotypes like drug resistance and virulence. Studies of candidate genes demonstrate that natural selection at known drug resistance loci (*dhfr* and *pfcrt*) produces selective sweeps involving a reduction in genetic diversity around loci under positive natural selection[3,4], and diversifying selection at antigenic determinants (*msp1*, *ama1* and *eba175*) produces highly polymorphic loci[5–7]. However, the genome-wide pattern of genetic diversity is poorly understood. One recent study estimated a frequency of one SNP per 910 bp in coding regions (pairwise nucleotide diversity $\pi = 4.9 \times 10^{-4}$) on chromosome 3 (ref. 8), but studies of candidate genes[9,10], intronic regions[11] and mitochondrial sequences[12,13] have offered highly divergent estimates. Successful vaccines and therapeutics will require better knowledge about genetic variation within and between parasite populations.

To study genome-wide diversity, we first generated high-quality draft genome sequences of two parasite clones, HB3 from Honduras and Dd2 from Indochina, and compared the results to the previously published genome sequence of 3D7, isolated in The Netherlands[14] but of unknown origin (**Fig. 1a** and **Supplementary Table 1** online). The draft sequences were based on approximately sevenfold redundancy of the genome, which is haploid in the human host. Comparison of the three clones uncovered 26,845 SNPs: a frequency of one SNP every 780 bases (**Fig. 1b**), substantially higher than previously reported[8]. The pairwise nucleotide diversity (π) between parasites was 1.29×10^{-3}. We experimentally validated >90% of a subset of candidate SNPs (**Supplementary Methods** online). In addition to SNPs, we noted a high frequency of insertion-deletion (indel) polymorphisms. Comparison of 3D7 and HB3 uncovered 37,039 indels of at least three bases, indicating that indels provide at least as much polymorphism as SNPs in *P. falciparum*.

We next surveyed diversity in *P. falciparum* through light genomic coverage of 12 additional lines (0.25-fold redundancy each) (**Fig. 1a** and **Supplementary Table 1**), identifying another 12,530 SNPs (**Fig. 1b**). Nucleotide diversity between the 12 lines was roughly similar (7.92×10^{-4}) to our estimate from draft-genome sequences.

[1]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA 02115. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA 02142. [3]Faculty of Medicine and Pharmacy, Cheikh Anta Diop University, BP 7325 Dakar, Senegal. [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA 02138. [5]Department of Biology, MIT, Cambridge, Massachusetts, USA 02139. [6]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA 02142. [7]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA 02115. [8]These authors contributed equally to this work. Correspondence should be addressed to D.F.W. (dfwirth@hsph.harvard.edu).

Received 28 June; accepted 2 November; published online 10 December 2006; doi:10.1038/ng1930

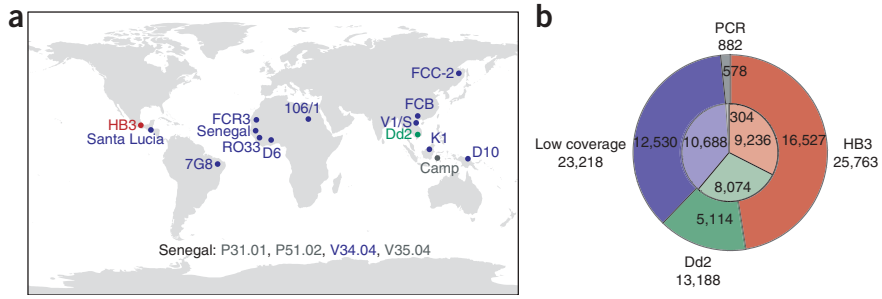

Figure 1 Geographic distribution of parasites and SNPs. (**a**) Sequence data derived from 18 parasites were used for SNP identification, including HB3 (red) and Dd2 (green), for which we obtained the full genome sequence; 12 additional parasites (blue) for which we obtained low-coverage sequence and four additional parasites (gray) that were used with the 12 low-coverage parasites for PCR product sequencing of 20 core regions across the genome (**Supplementary Table 2**). (**b**) SNPs identified from the parasites shown in **a** provide four data sources, including full-genome sequencing of HB3 (red) and Dd2 (green), low-coverage sequencing of 12 additional parasites (blue) and sequencing of 18 kb across 20 core regions (PCR) in 16 parasites (gray). The total number of SNPs identified for each of the four sources (HB3, Dd2, low coverage and PCR) is indicated by source, totaling 46,937 overall. The inner pie chart (light shading) indicates the number of SNPs by source that were found in more than one source ('shared'), identifying a total of 12,188 individual SNP positions that were identified in at least two sources. The outer pie chart (dark shading) indicates the number of SNPs identified only in a single source ('private').

Pairwise comparisons between parasites showed between 15% and 53% correlation in allelic states, with the highest correlations between Asian parasites (**Supplementary Fig. 1** online). The level of correlation was similar for long-term laboratory-adapted cultures and recent isolates from an affected individual, suggesting that culture-adapted parasites in this study are not significantly more or less divergent than recent isolates from affected individuals. Three clones (106/1, FCB and FCR3) were nearly identical, raising questions as to their independent origin and confirming previous observations[4,8,15,16]; we omitted these from further analysis.

We further extended the survey by performing PCR-directed sequencing in 20 genomic regions (**Supplementary Table 2** online) in 16 lines, including the 12 lines above (**Fig. 1** and **Supplementary Table 1**). In each region, we sequenced a core region of 6 kb and 12 regions of 1 kb at specified distances from the core region (a total of 18 kb per region). We chose the regions to represent the gene distribution across the genome while avoiding subtelomeric regions and multigene families. These data facilitated analysis of the allele frequencies and patterns of linkage disequilibrium (LD)[17]. This sequencing rediscovered 304 SNPs seen in the previous data and identified 578 new SNPs (**Fig. 1b**).

Overall, we identified a total of 46,937 SNPs (one every 446 bases; **Fig. 1b**) across the genome (20,950,000 reliably sequenced bases; **Supplementary Methods**), providing a rich data set for analysis of genome-wide diversity. Based on the whole-genome and low-coverage sequence data, pairwise nucleotide diversity (π) was 1.16×10^{-3} across the genome (**Supplementary Table 3** online) but varied threefold among chromosomes ($\pi = 2.41 \times 10^{-3}$ for chromosome 4 and 8.2×10^{-4} for chromosome 14) (**Supplementary Table 3**). Most of this variation was caused by small regions of extremely high diversity associated with antigen-coding genes.

The level of nucleotide diversity correlated strongly with predictions of gene function, as defined by Gene Ontology categories[18]. Gene Ontology categories containing surface molecules involved in cytoadherence and antigenic variation had the highest nucleotide diversity, whereas Gene Ontology categories related to mitochondrial function and electron transport chain showed low diversity (**Fig. 2**)[19]. Individual genes with the greatest diversity were the *var*, *rifin* and *stevor* genes (from $\pi = 1 \times 10^{-3}$ to $\pi = 6.1 \times 10^{-2}$), which encode antigenic variation molecules in the parasite. Conversely, many genes encoding essential metabolic enzymes, such as oxidases, transferases and hydrolases, showed zero nucleotide diversity. These results suggest that nucleotide diversity may help with functional characterization of the >60% of genes in the *P. falciparum* genome whose function remains unknown[20] (**Supplementary Methods**).

Regional variation in nucleotide diversity can be used to identify recent selective sweeps in a parasite population, and several studies have explored this potential[3,4]. By segregating the 12 parasites surveyed according to drug resistance, we searched for regions showing lower diversity in populations under selective pressure than in other populations[21] (**Supplementary Table 1**). Comparing four chloroquine-resistant parasites (CQ^R) to six chloroquine-sensitive (CQ^S) parasites, we found large regions (60–100 kb) of low diversity on chromosomes 5, 7 and 11 (**Fig. 3**). The region on chromosome 7 ($P = 1.8 \times 10^{-4}$) includes the *pfcrt* locus, site of a well-known selective sweep associated with chloroquine resistance[4]; reduced allelic diversity has previously been observed on chromosome 5 (ref. 4). When we compared three pyrimethamine-resistant parasites (PYR^R) to three pyrimethamine-sensitive (PYR^S) parasites, we found two candidate selective sweeps on chromosomes 13 and 14. The well-known selective sweep at the *dhfr*[3,22,23] locus was reflected by a dip in diversity but did not stand out (**Fig. 3**). Notably, a region on chromosome 13 spanning at least 160 kb had a stronger signal than either previously identified selective sweep (**Fig. 3**, $P = 6.4 \times 10^{-5}$). The successful identification of the *pfcrt* sweep illustrates the power of genetic diversity analysis for identifying recent selective sweeps in malaria populations. Further work will be needed to determine the role of selection in other candidate regions.

We next sought to better characterize malaria population genetics. For this purpose, we genotyped 372 SNPs across the 20 genomic regions described above in 54 parasite samples (26 African, 24 Asian, 4 American) from culture-adapted lines or from single-infection samples from infected individuals. We also genotyped a subset of these SNPs in 12 additional isolates (ten African, two Asian) (**Supplementary Table 1**) to provide better resolution for phylogenetic analysis. These data were able to separate parasite populations into continental groups (**Supplementary Fig. 2** online), consistent with previous observations[4,8]. Finally, we successfully genotyped several SNPs in *P. reichenowi*, the most closely related species to *P. falciparum*, providing an outgroup to determine the probable ancestral state of *P. falciparum* alleles.

The allele frequency spectrum in *P. falciparum* is dominated by rare, derived (non-ancestral) SNPs (**Supplementary Fig. 3** online). Of the SNPs genotyped, 47% have a minor allele frequency (MAF) <5%, and 35% have a derived allele frequency (DAF) <5%. This is a large fraction, given our low-coverage SNP ascertainment, which favors

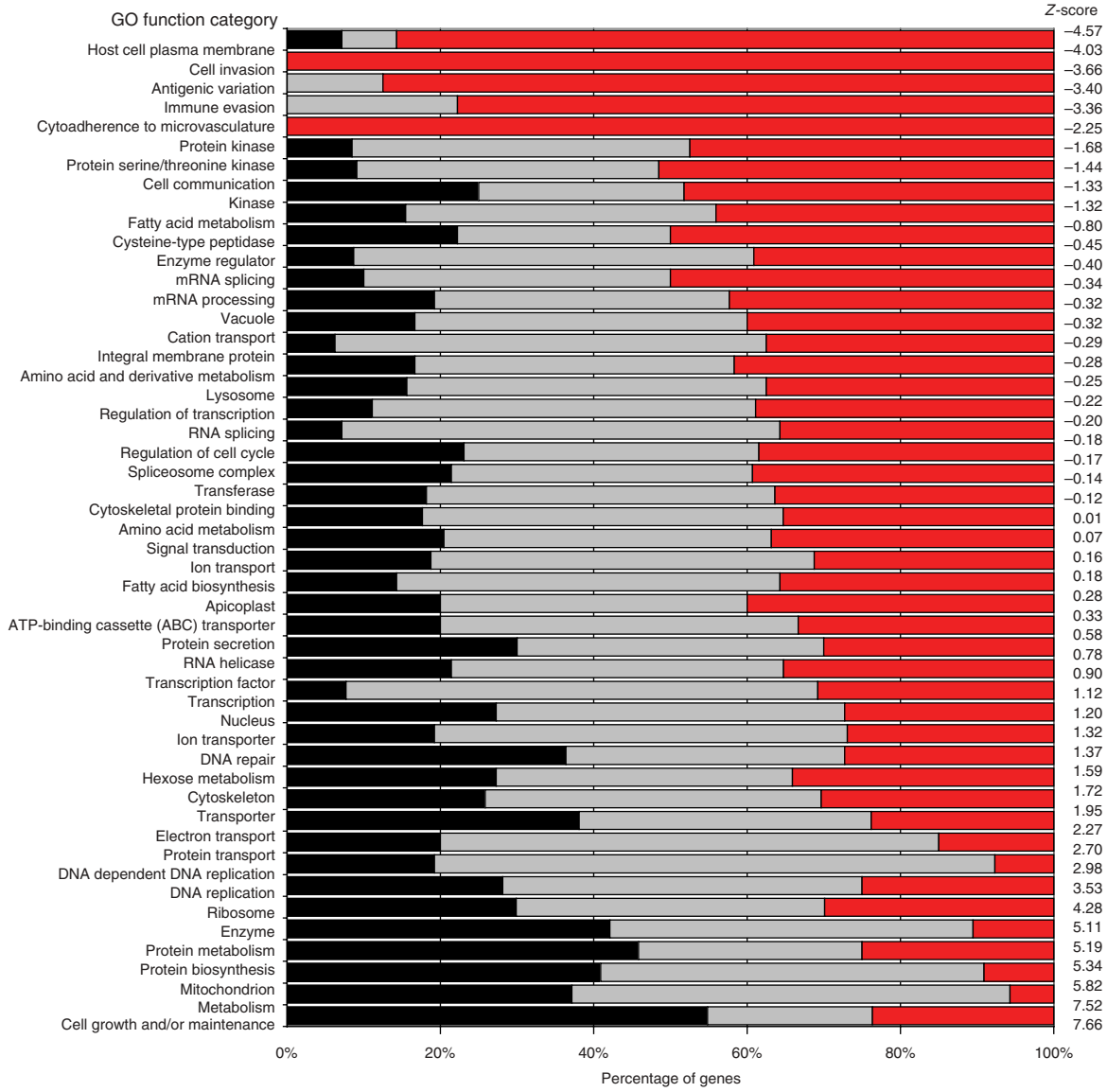

Figure 2 Genetic diversity within Gene Ontology (GO) functional categories. Genes were classified by GO functional categories[18] and stratified by level of nucleotide diversity (π), as estimated by *Z* scores (see Methods). Select categories (the categories with the five highest and five lowest *Z* scores along with categories in between that differ by incremental *Z* scores) are shown with the proportion of genes having high π values ($\pi > 8.36 \times 10^{-4}$) (red), genes with low π values ($\pi < 8.36 \times 10^{-4}$) (gray) and genes with $\pi = 0$ (black). The majority of genes in GO categories for molecules found at the cell membrane have high levels of nucleotide diversity, whereas most of the genes classified into GO categories for functionally conserved molecules lack nucleotide diversity (**Supplementary Methods**).

common alleles. Purifying selection provides some bias toward low-MAF SNPs (nonsynonymous coding SNPs are 15% of our total), but demographic effects, including population expansion, probably contribute as well. This predominance of rare variants suggests that most SNPs have yet to be identified and that further sequencing will be required for a thorough description of polymorphism in this organism.

We examined LD across the malaria genome by analyzing allelic correlations in the 20 genomic regions described above (**Supplementary Table 2**), using the traditional statistic D′ (ref. 17). One of the 20 regions showed an unusual pattern of LD, and we studied it separately (see the following paragraph). LD extended for extremely short physical distances (~1.5 kb) in African parasites but substantially longer in Asian parasites (~16 kb)[8] (**Fig. 4**). (The extent of LD was measured as the distance at which D′ fell to 0.5 (ref. 24).) Given the high recombination rate in *P. falciparum* (~17 kb per cM; ref. 25), these physical distances correspond to ~0.1 cM and ~1 cM, respectively. Understanding of the extent and structure of LD is central to designing association studies to identify genes responsible for traits, as illustrated by recent analyses in human[26], mouse[27] and dog[28]. The longer physical extent of LD in Asian parasites[8] suggests that fewer markers will be needed for association studies in those parasites than in African parasites[26].

One region, on chromosome 7, showed significant LD between SNPs over much larger distances (**Fig. 4b**), regardless of the samples' geographic origin. **Figure 4c** shows a SNP in this region with a high-frequency derived allele (allele frequency of 39% in all samples) residing on a long haplotype extending across the analyzed region

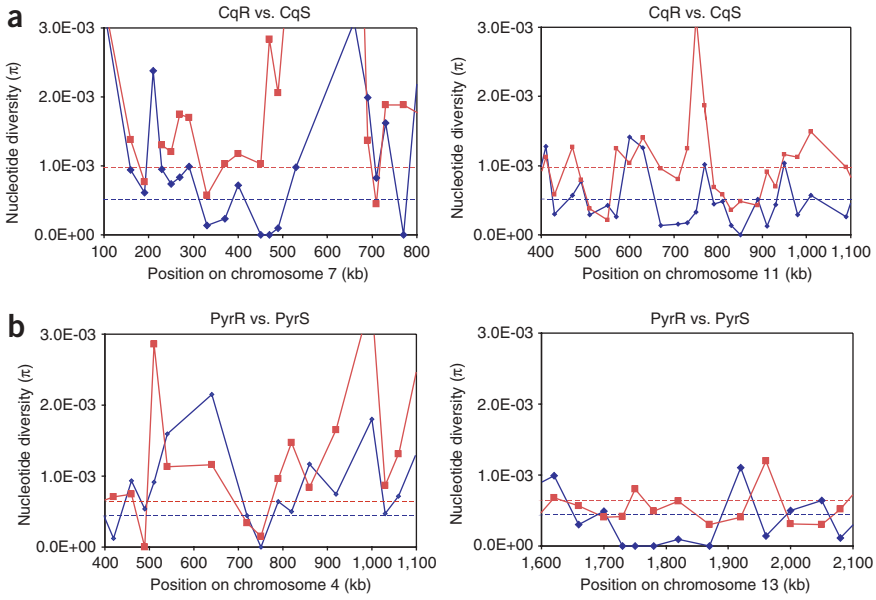

Figure 3 Nucleotide diversity identifies potential selective sweeps. Nucleotide diversity (π) was determined for 20-kb bins across the genome in all examined parasites (**Supplementary Fig. 4**). (**a**) Nucleotide diversity for parasites segregated by chloroquine sensitivity (CQ^S, red) and chloroquine resistance (CQ^R, blue) across chromosome 7 (centered on *pfcrt*, MAL7P1.27) and across chromosome 12. (**b**) Nucleotide diversity for parasites segregated by pyrimethamine sensitivity (PYR^S, red) and pyrimethamine resistance (PYR^R, blue) across chromosome 4 (centered on *dhfr*, PFD0830w) and across chromosome 13. Dashed lines represent genome-wide average for each category.

(>64 kb). This pattern (an allele that has risen to high frequency so rapidly that long-range allelic associations have not yet been broken down by recombination) is an expected signature of positive natural selection and has been extensively studied in humans[29]. Although the genome-wide distribution of long-range associations in *P. falciparum* will need to be thoroughly characterized, the background absence of long-range LD suggests that this approach will be a powerful tool to uncover recent natural selection in malaria parasites.

Finally, we experimented with whole-genome amplification (WGA) using samples purified by several approaches (**Supplementary Table 4** online). In the rest of this study, we used culture-adapted lines or DNA carefully extracted from patient samples, but these approaches are not suitable for large-scale studies, which require efficient protocols for genotyping of field samples. We found excellent success rates and accuracy with WGA of filtered samples from affected individuals as well as with WGA of *P. reichenowi* DNA.

The results presented here provide an initial picture of genetic diversity in *P. falciparum*, both genome-wide and worldwide. There is clearly substantial variation between parasites, with a pairwise nucleotide diversity of 1.16×10^{-3} from SNPs alone and at least an equal frequency of indel polymorphisms. There are detectable differences between continental populations of parasites, both in allele frequencies and in the extent of LD.

A high degree of variation, together with a high recombination rate, points to a powerful strategy for mapping traits in *P. falciparum* that have arisen recently owing to strong positive natural selection, such as drug resistance. Such selective events should leave a distinctive signature involving large regions that are depleted of polymorphism in the descendant parasites. The magnitude of the signal will depend on the strength of selection, which governs how fast the advantageous

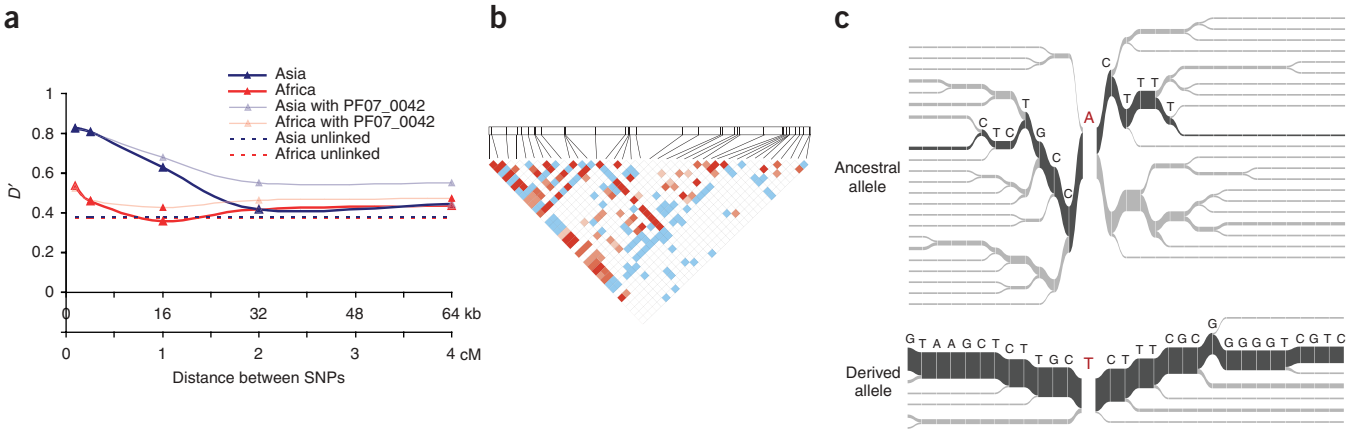

Figure 4 Patterns of linkage disequilibrium at 20 loci distributed across the *P. falciparum* genome in Asian and African parasites. (**a**) The relationship between the standard LD measure (D') and both physical (kb) and genetic (cM) distance for 20 loci distributed across the *P. falciparum* genome is shown using 54 samples with high-quality genotypes in Africa (red) and Asia (blue). (**b**) The hypothetical gene PF07_0042 has an unusual pattern of LD. Although there is little LD across the region in general, there are individual SNPs that maintain long-range LD with other SNPs in the region. LD calculations were carried out using Haploview (**Supplementary Methods**). White: $D' < 1$ with LOD < 2; pink: $D' < 1$ with LOD ≥ 2; blue: $D' = 1$ with LOD < 2; bright red, $D' = 1$ with LOD ≥ 2. (**c**) A haplotype bifurcation diagram[29] visualizes the long-range associations between the 13th SNP in the PF07_0042 region. The SNP is an A→T mutation; T is the nonancestral (derived) allele based on comparison with *P. reichenowi*. Although the long-range associations between the ancestral A allele have been whittled away by recombination, the T allele maintains long-range associations with other SNPs, suggesting it arose recently and that not enough time has passed for recombination to substantially break down these associations.

trait rises in frequency. For strong selection, the affected region may be >100 kb and already detectable using the current SNP catalog[21]. For weaker selection coefficients (for example, 1%–5%), the selective sweep may be tens of kilobases long and will require the use of much denser SNPs maps.

Partial selective sweeps can also be identified through alleles that show long-range LD against a background of short LD. This LD signal is useful for detecting recently selected alleles, but the signal will persist for only a short time. To study other genetic variation, association studies of alleles with phenotypes will probably be key. For such studies, the short span of LD, especially in African parasites, suggests that a very dense map of variants will be needed to detect most associations. Instead of dense genotyping, complete genome sequencing of each sample might be preferable; this may soon become feasible for this small genome (24 Mb).

Our initial survey underscores both the importance and the feasibility of creating a comprehensive map of genetic diversity in *P. falciparum*. Such a map will make it possible to identify specific loci under natural selection, to find genes mediating drug resistance and virulence and to trace the past spread of malaria. Two other groups, in parallel, have undertaken complementary genome-wide studies of genetic diversity: one group focused its analysis on coding regions[20], identifying potential vaccine targets, and the other undertook an evolutionary comparison of a recent Ghanaian clinical isolate with *P. reichenowi*[30], the chimpanzee parasite. Thus, powerful new genomic technologies now present opportunities to advance the understanding of the biology and evolution of this major pathogen, which is responsible for much human suffering.

METHODS

Parasites and DNA isolation. Parasites were obtained from the Malaria Research and Reagent Resource Repository (MR4) or additional sources as noted (**Fig. 1** and **Supplementary Table 1**). We used the following parasite lines from the MR4 repository of the American Type Culture Collection (ATCC): 3D7 (MRA-151); Dd2 (MRA-156); HB3 (MRA-155); 7G8 (MRA-154); Santa Lucia (MRA-331); V1/S (MRA-176); FCB (MRA-309); D10 (MRA-201); FCC-2 (MRA-733); D6 (MRA-285); FCR3 (MRA-731); RO-33 (MRA-200); 106/1 (MRA-464); K1 (MRA-159); Malayan Camp (MRA-328); ITG-2G2 (MRA-326); FCR8 (MRA-732); W2 (MRA-157); Indochina I (MRA-347); WR87 (MRA-284); T9-94 (MRA-153); and TM93C1088 (MRA-207). Patient samples were obtained as part of ongoing studies in Senegal and Malawi described elsewhere (**Supplementary Methods**) in accordance with human subject guidelines; written informed consent was obtained from all participants. Parasites were cultured by standard methods (**Supplementary Methods**) and nucleic acids were obtained using Qiagen genomic tips. Whole-genome amplification was performed using Repli-G methods (Qiagen). Plasmodipur filters were used to deplete human cells as noted (Euro-Diagnostica).

Selection of core regions. Twenty core regions were selected across the genome. These regions included several genes of interest, such as drug resistance loci and known antigens. For the remaining regions, we included hypothetical genes and genes with expected housekeeping roles in the parasite. We avoided subtelomeric regions and multigene families and included loci on all chromosomes (**Supplementary Table 2**).

SNP identification. Sequence reads were used for SNP detection from low-coverage, PCR-derived and Dd2 sequence. Read-end and low-quality (PHRED score < 10) bases were trimmed. Reads shorter than 100 bases, containing >3% internal Ns or containing a mononucleotide repeat covering greater than 80% of the read were discarded. Reads were aligned to the PlasmoDB version 5 of the 3D7 genome (**Supplementary Methods**) using BLAT (**Supplementary Methods**) requiring 95% identity, a minimum score of 100, <20% gaps and coverage of at least half of the read. Only the highest-scoring alignment for each read was kept, and paired reads that aligned more than 10 kb apart or in the wrong orientation were discarded. For the PCR reads, we discarded any reads that aligned outside of their known primer locations. For HB3, we used PatternHunter (**Supplementary Methods**) to identify blocks collinear between 3D7 and the released assembly; we then aligned these using MLAGAN (**Supplementary Methods**) or ClustalW (**Supplementary Methods**).

We did not analyze the highly rearranged and repetitive regions at the ends of chromosomes (**Supplementary Methods**). We determined the bounds for these regions by excluding the end regions where we saw a significant drop in alignment quality. Specifically, we examined the Dd2 and low-coverage read alignments in 5-kb windows across the genome. For each window, we computed the percentage of alignments that failed our quality checks, either because the alignment contained over 20% gaps or because paired reads did not align together. The bounds were set by discarding all 5-kb windows at the end of each chromosome up until the first point where three consecutive windows each had <15% failed alignments. In most cases there was an abrupt transition, with all windows showing >70% failures up to the boundary point and <10% after.

The Neighborhood Quality Score (NQS) algorithm was used to distinguish real polymorphisms from sequence errors. This algorithm uses the PHRED quality scores at the position of the mismatch as well as those at the neighboring bases to select SNPs. We required the SNP to have a quality score of 20 and the 5-bp neighborhood to have a score of 15. We allowed one mismatch and no indels in the neighborhood. Because of stringent requirements to identify a SNP, only 42% of the low-coverage sequence uniquely aligned to the genome, passed filtering and satisfied the conditions of the NQS algorithm. As a final filter, we discarded a SNP when another read from the same sample met the NQS criteria at that position but did not have a sequence difference.

Nucleotide diversity. We calculated π for all aligned HB3, Dd2 and low-coverage reads described previously. For each site with a good call from at least two of the parasites being compared, a count of the two alleles was made, and the mean number of differences per pairwise comparison calculated. Mean π within a bin was calculated by averaging over sites, weighting each by $\sum_{i=1}^{n-1} \frac{1}{i}$, where n is the number of aligned parasites. Bins with a weighted coverage of <30% were discarded.

Selective sweeps. Parasites were divided into groups based on known drug susceptibility (**Supplementary Table 1**). African and Asian resistant parasites (excluding the three nearly identical parasites) were grouped together, as these have a reasonable chance to share a common founder mutation; grouping the two continents increases statistical power but at the cost of reducing our ability to identify sweeps with different founder mutations in Asia and Africa. The groups were CQ^R (Dd2, Senegal P34.04, V1/S and K1), CQ^S (3D7, HB3, D6, FCC-2, D10 and Santa Lucia), PYR^R (Dd2, V1/S and K1) and PYR^S (3D7, D6 and D10). π was calculated within each group in 20-kb bins. Because π was systematically lower in resistant parasites, $\pi(CQ^R)$ and $\pi(PYR^R)$ were scaled to have the same mean as found for the sensitive groups. Regions around *pfcrt* and *dhfr* previously reported to have experienced selective sweeps were omitted in determining the scaling factor.

To identify swept regions, the statistic $\Delta = \frac{\pi_{\text{resistant}} - \pi'_{\text{sensitive}}}{\pi_{\text{resistant}} + \pi'_{\text{sensitive}}}$, where π' is the scaled diversity, was calculated for each bin with sufficient coverage. No shape was assumed for the distribution of π under neutral evolution. Instead, candidate loci were identified by clustering of extreme values ($\pi > 0.6$) from the empirical distribution. The *pfcrt* and *dhfr* regions were omitted in determining this distribution. Values in adjacent bins were assumed to be uncorrelated (which is a good approximation, as r^2 was 0.02 for the PYR groups and 0.004 for the CQ groups). Local significance was determined by calculating the probability of finding the observed number of consecutive high-π bins in a 5-bin window. Genome-wide significance was calculated as the probability of finding that many consecutive high-π bins among our informative bins across the entire genome.

Gene Ontology category analysis. Subsets of genes showing a significant enrichment or deficit of genetic diversity (π) were identified using a two-tailed Mann-Whitney U test, with a Bonferroni correction applied for multiple testing. Select categories were ranked from high to low genetic diversity, and

individual members of those Gene Ontology categories were classified as having high ($\pi > 8.36 \times 10^{-4}$); low ($\pi < 8.36 \times 10^{-4}$) or no ($\pi = 0$) genetic diversity, with $\pi = 8.36 \times 10^{-4}$ representing the mean of the distribution (**Supplementary Methods**).

Genotyping. SNPs were genotyped using a mass spectrometry-based MassArray platform by Sequenom. SNPs were amplified in multiplex PCRs consisting of a maximum of 24 loci each. After amplification, the single-base extension (SBE) reaction was performed on the shrimp alkaline phosphatase (SAP)-treated PCR product using iPLEX enzyme and mass-modified terminators (Sequenom). A small volume ($\sim$7 nl) of reaction was then loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl of matrix (3-hydroxypicolinic acid). SpectroCHIPs are analyzed in automated mode by a MassArray Compact system with a solid-phase laser mass spectrometer (Bruker Daltonics). The resulting spectra are analyzed by SpectroTyper v.3.4A software, which combines base calling with the clustering algorithm.

PCR resequencing. Sixteen *P. falciparum* lines, described above, were sequenced for 470 regions across 20 genomic loci. They were amplified using a standard PCR protocol, processed for Sanger-style sequencing using ABI BigDye Terminator chemistry and detected on ABI 3730 capillary machines. M13-tailed PCR primers (**Supplementary Methods** online) were designed to produce amplicons 700–900 bp in length. Universal M13 primers were used for the sequencing amplification. Each 10-µl PCR reaction was composed of 5 µl mixed forward and reverse PCR primers (final concentration, 0.5 µM) (IDT), 2 µl genomic DNA (5 ng/µl), 0.04 µl Taq polymerase (Qiagen, HotStarTaq), 1.0 µl 10× buffer (Qiagen), 0.4 µl 25 mM $MgCl_2$ (Qiagen), 0.08 µl 100-mM dNTPs (25 mM each) (ABI) and 1.48 µl UltraPure DNase/RNase-free water (Invitrogen). Quality control of the PCR reactions was performed by gel electrophoresis (2%, 96-well Invitrogen E-gels). Excess PCR primers and dNTPs were eliminated by incubation with SAP and exonuclease I. Each SAP-exonuclease reaction was composed of 0.45 µl SAP (1 unit (U)/µl, Amersham), 0.30 µl exonuclease I (20 U/µl, Fermentas) and 2.25 µl UltraPure DNase/RNase-free water (Invitrogen). Reactions were then sent for cycle sequencing. Sequence bases were called with 3XX caller from ABI. SNP detection was performed automatically with the SNP Compare Analysis Suite (developed at the Broad Institute).

LD analysis. We carried out LD analysis on the 372 SNPs across 20 genomic regions in 22 malaria samples from Africa and 22 samples from Asia. These samples were from culture-adapted lines or genomic DNA from samples from affected individuals with a single infection. For our analysis of LD, we used only a subset of 372 SNPs in each population that had at least four copies of the minor allele within a continental group (frequency of 18% or greater); that corresponded to 108 SNPs in Africa and 86 SNPs in Asia. LD was examined using two standard measures: the pairwise-marker statistic D' (**Supplementary Methods**) and r^2. For each genomic region, pairwise LD was visualized and presented using the HaploView program (**Supplementary Methods**). We evaluated the correlation between LD and distance[24] by binning pairwise markers at varying distances (1.5 kb, 4 kb, 16 kb, 32 kb and 64 kb). We identified an unusual pattern of LD in one region on chromosome 7 and analyzed this region separately. We compared these values to the average background correlation that occurs between unlinked markers (on different chromosomes) in this small sample set. We performed similar analysis for data from 56 malaria isolates (28 African and 28 Asian) from a recent study of chromosome 3 (ref. 8). Similarly, we calculated D' and r^2 for SNPs of $>$18% frequency and give the background correlation for likely unlinked SNPs ($>$200 kb apart) (**Supplementary Methods**).

Extended haplotypes. We visualized the decay of the extended ancestral chromosome (haplotype) on which an allele arose using the program Bifurcator (**Supplementary Methods**). The root of each diagram is an allele, identified by an open square. The diagram is bidirectional, portraying both centromere-proximal and centromere-distal LD. Moving in one direction, each marker is an opportunity for a node; the diagram either divides or does not divide based on the presence of either one or both alleles for each adjacent marker. Thus, the breakdown of LD away from the allele of interest is portrayed at progressively longer distances. The thickness of the lines corresponds to the number of samples with the indicated long-distance haplotype.

Accession codes. GenBank: HB3, AANS01000000; Dd2, AASM01000000.

URLs. Malaria Research and Reagent Resource Repository (MR4): http://www.malaria.mr4.org. The HB3 genome assembly can be accessed at http://www.broad.mit.edu/annotation/microbes/plasmodium_falciparum_hb3/. Additional information is available on The Broad Institute of MIT and Harvard website (http://www.broad.mit.edu/mpg/pubs/).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all members of the sample collection team in Senegal (A. Ahouidi, L. Ndiaye, O. Ly, Y. Diedhiou, T. Sene, A. Mbaye and D. Diop). We thank T. Taylor, K. Seydel, J. Montgomery, E. Dembo, M. Molyneux and S. Rogerson for help collecting the samples from Malawi. We also thank all the members of the Broad Sequencing Platform and M. Koehrsen, D. Richter and O. Shamovsky for sequence analysis help. Our thanks to M. Goyette and T. Rachupka for help with Sequenom genotyping. Thanks to N. Mahesh and G. Ramirez for help with sample preparation and parasite cultures and to J. Mu and X. Su for typing samples. We thank MR4 for providing us with malaria parasites contributed by the following depositors: W.E. Collins (MRA-347); D.E. Kyle (MRA-154, MRA-157, MRA-159, MRA-176, MRA-207, MRA-284, MRA-285); L.H. Miller and D. Baruch (MRA-331, MRA-326, MRA-328); W. Trager (MRA-731, MRA-732, MRA-733); D. Walliker (MRA-151, MRA-153, MRA-200); T.E. Wellems (MRA-155, MRA-156, MRA-309, MRA-464); and Y. Wu (MRA-201). Special thanks to J. Barnwell for providing *P. reichenowi* DNA. Our thanks to PlasmoDB (http://www.plasmodb.org) for access to the 3D7 genome sequence. The authors are supported by the US National Institutes of Health, SPARC funding of The Broad Institute of MIT and Harvard, the Burroughs-Wellcome Fund, The Bill and Melinda Gates Foundation, the NIAID Microbial Sequencing Center, the Ellison Medical Foundation, Fogarty International and the Exxon Mobil Foundation. P.C.S. is funded by the Damon Runyon Cancer Fellowship and the L'Oreal for Women in Science Award.

AUTHOR CONTRIBUTIONS

S.K.V. designed experiments, prepared samples, carried out Sequenom genotyping and PCR resequencing, analyzed genotyping data and wrote the paper. P.C.S. designed experiments; carried out Sequenom genotyping and PCR resequencing; analyzed data for linkage disequilibrium, allele frequency and long-range haplotypes and wrote the paper. D.D.C. analyzed sequence data, prepared sequence data for subsequent analysis and defined polymorphisms. D.E.N. analyzed data for allele frequency and nucleotide diversity by GO category. S.F.S. analyzed data for nucleotide diversity and selective sweeps. D.A.M., J.P.D., O.S., D.N., O.N., S.M. and M.T.D. helped with sample collection. J.P.D. and M.T.D. helped with culture adaptation. D.A.M. and A.L. helped with parasite cultures. A.D. helped with GO function analysis. N.S.-T. and J.Z. prepared libraries for sequencing. S.W. and R.O. helped with PCR resequencing. L.Z. helped with Sequenom genotyping. E.M., S.G. and D.B.J. created the genome assemblies for HB3 and Dd2. R.C.W. coordinated project flow. B.W.B. supervised sequencing. D.L.H. consulted on population genetic analysis. J.E.G. supervised and advised on sequence analysis. E.S.L. designed experiments, consulted on project outcomes and wrote the paper. D.F.W. designed experiments, coordinated all efforts, supervised project at all levels, consulted on project outcomes and wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. World Health Organization. WHO expert committee on malaria. *World Health Organ. Tech. Rep. Ser.* **892**, 1–74 (2000).
2. Gardner, M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
3. Nair, S. *et al.* A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Mol. Biol. Evol.* **20**, 1526–1536 (2003).
4. Wootton, J.C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418**, 320–323 (2002).

5. Polley, S.D., Chokejindachai, W. & Conway, D.J. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics* **165**, 555–561 (2003).
6. Baum, J., Thomas, A.W. & Conway, D.J. Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and P. vivax. *Genetics* **163**, 1327–1336 (2003).
7. Ferreira, M.U., Ribeiro, W.L., Tonon, A.P., Kawamoto, F. & Rich, S.M. Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of *Plasmodium falciparum*. *Gene* **304**, 65–75 (2003).
8. Mu, J. *et al.* Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol.* **3**, e335 (2005).
9. Rich, S.M., Licht, M.C., Hudson, R.R. & Ayala, F.J. Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA* **95**, 4425–4430 (1998).
10. Hughes, A.L. & Verra, F. Extensive polymorphism and ancient origin of *Plasmodium falciparum*. *Trends Parasitol.* **18**, 348–351 (2002).
11. Volkman, S.K. *et al.* Recent origin of *Plasmodium falciparum* from a single progenitor. *Science* **293**, 482–484 (2001).
12. Conway, D.J. *et al.* Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. *Mol. Biochem. Parasitol.* **111**, 163–171 (2000).
13. Joy, D.A. *et al.* Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300**, 318–321 (2003).
14. Delemarre, B.J. & van der Kaay, H.J. Tropical malaria contracted the natural way in the Netherlands. *Ned. Tijdschr. Geneeskd.* **123**, 1981–1982 (1979).
15. Robson, K.J., Hall, J.R., Davies, L.C., Crisanti, A., Hill, A.V. & Wellems, T.E. Polymorphism of the TRAP gene of *Plasmodium falciparum*. *Proc. Biol. Sci.* **242**, 205–216 (1990).
16. Kaneko, O., Soubes, S.C. & Miller, L.H. *Plasmodium falciparum*: invasion of Aotus monkey red blood cells and adaptation to Aotus monkeys. *Exp. Parasitol.* **93**, 116–119 (1999).
17. Lewontin, R.C. On measures of gametic disequilibrium. *Genetics* **120**, 849–852 (1988).
18. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
19. Volkman, S.K. *et al.* Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298**, 216–218 (2002).
20. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat. Genet.* advance online publication 10 December 2006 (doi:10.1038/ng1924).
21. Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
22. Roper, C. *et al.* Antifolate antimalarial resistance in southeast Africa: a population-based analysis. *Lancet* **361**, 1174–1181 (2003).
23. Roper, C. *et al.* Intercontinental spread of pyrimethamine-resistant malaria. *Science* **305**, 1124 (2004).
24. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
25. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286**, 1351–1353 (1999).
26. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
27. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
28. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
29. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
30. Jeffares, D.C. *et al.* Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat. Genet.* advance online publication 10 December 2006 (doi:10.1038/ng1931).