Supplementary Methods.

Parasites and DNA Isolation. Parasites were obtained from the Malaria Research and Reagent Resource Repository (MR4, malaria.mr4.org) or additional sources noted (**Fig. 1** and **Supplementary Table 1**). From MR4 the following parasite lines were used: parasite line 3D7 (MRA-151, MR4, ATCC Manassas Virginia); parasite line Dd2 (MRA-156, MR4, ATCC Manassas Virginia); parasite line HB3 (MRA-155, MR4, ATCC Manassas Virginia); parasite line 7G8 (MRA-154, MR4, ATCC Manassas Virginia); parasite line Santa Lucia (MRA-331, MR4, ATCC Manassas Virginia); parasite line V1/S (MRA-176, MR4, ATCC Manassas Virginia); parasite line FCB (MRA-309, MR4, ATCC Manassas Virginia); parasite line D10 (MRA-201, MR4, ATCC Manassas Virginia); parasite line FCC-2 (MRA-733, MR4, ATCC Manassas Virginia); parasite line D6 (MRA-285, MR4, ATCC Manassas Virginia); parasite line FCR3 (MRA-731, MR4, ATCC Manassas Virginia); parasite line RO-33 (MRA-200, MR4, ATCC Manassas Virginia); parasite line 106/1 (MRA-464, MR4, ATCC Manassas Virginia); parasite line K1 (MRA-159, MR4, ATCC Manassas Virginia); parasite line Malayan Camp (MRA-328, MR4, ATCC Manassas Virginia); parasite line ITG-2G2 (MRA-326, MR4, ATCC Manassas Virginia); parasite line FCR8 (MRA-732, MR4, ATCC Manassas Virginia); parasite line W2 (MRA-157, MR4, ATCC Manassas Virginia); parasite line Indochina I (MRA-347, MR4, ATCC Manassas Virginia); parasite line WR87 (MRA-284, MR4, ATCC Manassas Virginia); parasite line T9-94 (MRA-153, MR4, ATCC Manassas Virginia); and parasite line TM93C1088 (MRA-207, MR4, ATCC Manassas Virginia).Patient samples were obtained as part of ongoing studies in Senegal and Malawi described elsewhere[1,2] in accordance with human subject guidelines. Parasites were cultured by standard methods[3] and nucleic acids were obtained using Qiagen genomic-tips (Qiagen, USA). *P. reichenowi* DNA was the generous gift of John Barnwell. Whole genome amplification was performed using Repli-G methods (Qiagen GmbH, Germany). Plasmodipur filters were used to deplete human cells as noted (Euro-Diagnostica, The Netherlands).

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 multi-gene families and included loci on all chromosomes.

SNP Identification. Sequence reads were used for SNP detection from low coverage, PCR, and Dd2 sequence. Reads ends and low quality (PHRED <10) bases were trimmed. Reads less than 100 bases, containing greater than 3% internal N's, 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[4] using BLAT[5] requiring 95% identity, a minimum score of 100, less than 20% gaps, and coverage of at least half of the read. Only the highest scoring alignment for each read was kept and paired reads which aligned more than 10Kb 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[6] to identify collinear blocks between 3D7 and the

released assembly that were then aligned using MLAGAN[7] or ClustalW[8]. HB3 and Dd2 sequences can be found in Genbank under accession numbers AANS01000000 and AASM01000000 respectively.

We did not analyze the highly rearranged and repetitive regions at the ends of chromosomes. 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 5Kb windows across the genome. For each window we computed the percentage of alignments which 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 5Kb windows at the end of each chromosome up until the first point where three consecutive windows each had fewer than 15% failed alignments. In most cases there was an abrupt transition, with all windows showing greater than 70% failures up to the boundary point and less than 10% after. The chromosomal coordinates (boundary) are listed below for each chromosome that mark the region of that chromosome analyzed in this study. The total number of bases for the chromosomal region is also provided.

Chr	*Boundary (kb)*	*Boundary (kb)*	*Bases (kb)*
1	95	575	480
2	65	870	805
3	65	995	930
4	90	1,170	1,080
5	85	1,285	1,200
6	70	1,295	1,225
7	60	1,380	1,320
8	75	1,345	1,270
9	75	1,475	1,400
10	60	1,515	1,455
11	110	1,915	1,805
12	50	2,175	2,125
13	75	2,780	2,705
14	35	3,185	3,150
TOTAL			**20,950**

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 five base 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 could be uniquely aligned to the genome, passed filtering, and satisfied the conditions of the

NQS algorithm. As a final filter, we discarded SNPs when another read from the same sample met the NQS criteria at that position but did not have a sequence difference.

We used two methods to estimate a false positive rate for SNP discovery. First, we used the redundancy in the 8-fold coverage Dd2 reads – 18,126 bases were called as a SNP by at least one read, and 81% of those bases were covered by more than one read. We divided the bases with multiple read coverage into true and false positives. Bases were considered false positives if there was more than one read supporting the reference allele, or if there was only one read supporting the reference allele, but only one read supporting the SNP. Using these criteria, 10.9% of the SNPs identified from individual reads are false positives. Because in the final data set, we threw out all SNPs with any disagreements among the reads, this false positive estimate is accurate for regions with single coverage, but will be an overestimate for regions with higher coverage. Second, we used deep PCR resequencing data. For 186 of the SNPs identified through low coverage sequencing, PCR data was also available for the same sample. PCR data identified 15 of these SNPs (8%) as false positives.

Sample correspondence. We computed the correspondence among parasites using a conditional probability P(a U b | a or b) (**Supplementary Figure 1**). Given two samples *a* and *b*, the correspondence is the conditional probability that a given SNP is present in both *a* and *b* given that it is present in either *a* or *b*. This was computed only for sites that were reliably assayed in both samples.

d_N/d_S. To estimate d_N/d_S, we created two sequences for each gene that contained a SNP. The first sequence was the 3D7 reference sequence. The second sequence was created from the first by substituting in all of the SNPs discovered for that gene, creating a hybrid containing the SNPs from all sequenced parasites. In cases where these substitutions introduced a premature stop codon, the gene was discarded. Nucleotide frequencies were estimated for each codon position from the remaining genes, and these frequencies were used to estimate codon frequencies. We then ran PAML[9] in pairwise mode to estimate d_N and d_S for each gene along with confidence intervals. The expected number of non-synonymous vs. synonymous SNPs for neutral sequence was taken by counting the total number of synonymous and nonsynonymous sites estimated by PAML[9] across all genes for which d_N/d_S was determined (**Supplementary Table 6**).

Sample correspondence. We computed the correspondence among parasites using a conditional probability P(a U b | a or b) (**Supplementary Fig. 1**). Given two samples *a* and *b*, the correspondence is the conditional probability that a given SNP is present in both *a* and *b* given that it is present in either *a* or *b*. This was computed only for sites that were reliably assayed in both samples.

Nucleotide Diversity. π was calculated 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 less than 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, since these have a reasonable chance to share a common founder mutation; grouping the two continents increases statistical power, but at a 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 $\Delta = \frac{\pi_{resistant} - \pi'_{sensitive}}{\pi_{resistant} + \pi'_{sensitive}}$, 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 (a good approximation: 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.

GO Category Analysis. Subsets of genes showing a significant enrichment or deficit of genetic diversity (π) were identified using a 2-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 GO 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.

Genotyping. SNPs were genotyped using a mass spectrometry-based MassArray platform by Sequenom (San Diego, California, United States). SNPs are amplified in multiplex PCR reactions consisting of maximum 24 loci each. Following amplification, the Single Base Extension /SBE/ reaction is performed on the SAP treated PCR product using *iPLEX* enzyme™ and mass-modified terminators™ /Sequenom, SanDiego/. A small volume (~7 nl) of reaction is 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 Inc., 2005). The resulting spectra are analyzed by SpectroTyper v.3.4A software which combines base caller with the clustering algorithm.

PCR- Re-sequencing. 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 were designed to produce amplicons 700 to 900 bp in length.

Primer	***Sequence***
M13-tailed Forward Primer	GTAAAACGACGGCCAGT
M13-tailed Reverse Primer	CAGGAAACAGCTATGACC

Universal M13 primers were used for the sequencing amplification. Each 10ul PCR reaction is comprised of the following: 5ul F+R Mixed PCR primer (0.5uM) (IDT), 2ul genomic DNA (5ng/ul), 0.04ul Taq polymerase (Qiagen, hotstart), 1.0ul 10X Buffer (Qiagen), 0.4ul 25mM $MgCl_2$ (Qiagen), 0.08ul 100mM dNTPs (25mM each) (ABI), 1.48ul Ultra-pure DNase/RNase-free water (Invitrogen). QC of the PCR reactions was performed by gel electrophoresis (2%, 96-well E-gels from Invitrogen). Excess PCR primer and dNTPs are eliminated by incubation with Shrimp Alkaline Phosphatase and Exonuclease I. Sap/Exo mix per reaction is comprised of the following: 0.45ul SAP (1U/ul, Amersham), 0.30ul Exo I (20U/ul, Fermentas), 2.25ul Ultra-pure DNase/RNase-free water (Invitrogen). Reactions are then sent for cycle sequencing. Sequence bases are called with 3XX caller from ABI. SNP detection was performed automatically with the SNP Compare analysis suite (developed at the Broad).

Phylogeny. A phylogenet ic tree was constructed using the neighbor-joining algorithm[10] in ClustalW[8], using corrected pairwise distances. One thousand bootstrap replicates were performed, and nodes exhibiting less than 50% support were collapsed. A circular representation of the tree was derived using Mesquite (mesquiteproject.org). Identical tree topologies and comparable bootstrap support were observed using parsimony analysis in Mega3.1[11]. An identical tree topology was also observed using a maximum likelihood approach in PAUP* 4.0b10[12]. F_{ST} was calculated using DNAsp 4.0[13].

Linkage Disequilibrium. 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 patient samples with a single infection. For our analysis of LD we used the only a subset of 372 SNPs in each population which had at least 4 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, pairwise marker D` statistic[14] and r^2. For each genomic region pairwise LD was visualized and presented using the HaploView program[15]. We evaluated the correlation between LD and distance[16] 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 for 56 malaria isolates (28 African and 28 Asian) from a recent study of chromosome 3[17]. We similarly calculate D′ and r^2 for SNPs of greater than 18% frequency and present background correlation for likely unlinked SNPs (> 200kb apart).

Extended haplotypes. We visualize the decay of the extended ancestral chromosome (haplotype) on which an allele arose using the program Bifurcator[18]. The root of each diagram is an allele, identified by an open square. The diagram is bi-directional, 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 not based on whether both or only one allele for each adjacent marker is present. 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.

URLs. Additional information is available on The Broad Institute of MIT and Harvard website: http://www.broad.mit.edu/mpg/pubs/.

REFERENCES

1. Thomas, S.M. et al. In vitro chloroquine susceptibility and PCR analysis of pfcrt and pfmdr1 polymorphisms in Plasmodium falciparum isolates from Senegal. *Am J Trop Med Hyg* **66**, 474-80 (2002).
2. Montgomery, J. et al. Genetic Analysis of Circulating and Sequestered Populations of Plasmodium falciparum in Fatal Pediatric Malaria. *J Infect Dis* **194**, 115-22 (2006).
3. Trager, W. & Jensen, J.B. Human malaria parasites in continuous culture. *Science* **193**, 673-5 (1976).
4. Bahl, A. et al. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* **31**, 212-5 (2003).
5. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
6. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440-5 (2002).
7. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721-31 (2003).
8. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
9. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6 (1997).
10. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25 (1987).
11. Kumar, S., Tamura, K., and Nei, M. MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics* **5**, 150-163 (1993).
12. Swofford, D.L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* (Sinauer Associates, Sunderland, Massachusetts, 2003).

13. Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-7 (2003).
14. Lewontin, R.C. The Interaction of Selection and Linkage. Ii. Optimum Models. *Genetics* **50**, 757-82 (1964).
15. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
16. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).
17. Mu, J. et al. Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol* **3**, e335 (2005).
18. Sabeti, P.C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-7 (2002).