are linked, multiple comparisons at the same locus are correlated, and not all DNA markers are equally informative. Instead of a Bonferroni correction, we used an empirical permutation test. We created replications of the observed sample, in which genotype and phenotype distributions were preserved, but any linkage between genotype and phenotype was removed by randomly reassigning the observed phenotypic values. We then asked how often any replicates that conformed to the null hypothesis (the independence of genotype and phenotype) produced \( P \) values in excess of the \( P \) value observed for the real data. This test preserved both the observed marker informativeness and the overall phenotype distribution.

24. Linkage of the resistance allele to two distinct marker alleles, 2 and 4, could be explained by recombination between resistance and marker alleles or by multiple entry of the resistance allele into the family. An attempt to distinguish by reconstructing flanking markers, although not conclusive, appeared more consistent with the recombination hypothesis.

26. S. F. Taore, O. Niare, K. D. Vernick, data not shown.
sequence averaging 25-nt probe sequence. These differences can account for the reduction in hybridization strength of the hybridization signal across isolates. Our criterion for a candidate SNP was the finding that all three hybridizations with DNA from a given isolate showed significantly reduced signal intensity as compared with three control hybridizations with DNA from the reference strain 3D7. Relative to 3D7, 320 putative SNPs were identified in W2, 107 in D6, 230 in 7G8, and 324 in HB3. Altogether, 585 of the 4167 probes showed a difference in at least one isolate.

To validate the SNP detection, we determined the genomic sequence for each of 17 variable, nontelomeric probes in each of the isolates (table S1). In all 17 cases, an SNP or small deletion was identified within the 25-nt probe sequence. These differences can account for the reduction in hybridization signal intensity on the array. Around a variable probe in each of five genes, flanking sequences averaging ~400 base pairs (bp) were also determined in each of five to seven additional isolates; four of these flanking regions contained additional SNPs (fig. S2).

We also determined genomic sequences across a set of nonvariable probes that, as expected, failed to reveal SNPs. Nevertheless, oligonucleotide hybridization is more prone to false-negatives than to false-positives, because SNPs near the extreme ends of the oligonucleotide may remain undetected.

The distribution of variable probes by gene function is highly nonrandom (Fig. 2). The greatest variation was found in proteins associated with the cell membrane, which accounted for 22% of all probes but 69% of all variable probes ($P \approx 0$). Among these membrane-associated proteins are several well-characterized vaccine targets, including the transmission-blocking target antigen PfS230, several PfEMP1-related molecules (erythrocyte membrane proteins), merozoite surface proteins 2 and 4 (MSP2 and MSP4), and the ring-infected erythrocyte surface antigen RESA-H3. Proteins associated with general cellular processes were also significantly more variable than expected, accounting for 1.8% of all probes but 5.5% of variable probes ($P \approx 10^{-10}$). However, in this category the variability was found predominantly in two open reading frames (PFB0085c and PFB0090c), both of which show similarity to the bacterial chaperone DnaJ and also to P. falciparum RESA antigens. Hence, the most variable proteins are molecules with a high likelihood of interacting with the host immune system. About 20% of the variable probes are in coding sequenc-
es for hypothetical proteins, some of which may also be associated with the cell membrane. The remaining categories of protein are much less variable (Fig. 2), accounting for 33% of all probes but only 6% of the variable probes ($P = 0$).

The spatial distribution of genetic variation across chromosome 2 is also highly nonrandom (Fig. 3). The 700 kb in the central region of the chromosome is the least variable. Across this region there are 170 genes queried by 3383 probes, of which 142 probes in 69 genes were variable. Some genes are exceptionally polymorphic—in for example, the MSP2 gene, in which 10 of 17 probes were variable. On the other hand, the central region of the chromosome also included 101 genes with no variable probes. Most of the variation was located in the subtelomeric regions within 100 kb of the chromosome ends, accounting for only 22% of the total chromosome length; these regions contained 443 of the 585 variable probes (76%).

If we assume that hybridization with 25-nt oligomers can reliably detect SNPs anywhere within the middle 15 nucleotides, then the 3383 probes assay 50,745 bp of coding sequence and detect 142 SNPs. The frequency of SNPs is therefore about one in 350 bp, which is significantly greater than the 1 in 1400 observed in introns ($P = 0.007$) (10). Assuming that about 30% of these are synonymous (11), the estimated frequency of synonymous SNPs across chromosome 2 is $\sim 1$ per 1.2 kb. This estimate is comparable to that reported for coding sequences in chromosome 3 when the latter is corrected for SNPs in regions of DNA with low sequence complexity (11).

The SNPs in chromosome 2 are not randomly distributed, but instead cluster in 19 highly polymorphic genes in which the proportion of variable probes in each gene exceeds 10%. These 19 genes encode known antigens, predicted membrane-associated proteins, or hypothetical proteins and account for 36% of all the SNPs detected on chromosome 2. It seems likely that the high level of polymorphism in many of these genes is maintained by some form of selection. Estimates of the age of the most recent common ancestor of P. falciparum that include the most highly polymorphic genes may therefore be biased toward a more ancient common ancestry (11, 12), because the theoretical basis of the estimation assumes that the SNPs are selectively neutral.

The proportion of variable probes in the remaining 151 genes in the central region of chromosome 2 is less than 0.2% and not significantly different from that observed in introns ($P = 0.09$). We find no evidence for extensive regions of exceptionally high polymorphism or of exceptionally low polymorphism, which might be expected if the chromosome had recently experienced one or more selective sweeps that reduced genetic variation locally.

The oligonucleotide-hybridization approach, validated here for chromosome 2, provides an experimental platform for systematic genomewide studies of reference isolates. Assessing the nature and extent of genetic variation across the genome of P. falciparum has potential implications for control strategies including the identification of new targets for drug or vaccine development (13). In chromosome 2, most of the variation is concentrated in the subtelomeric 100 kb at each end, regions that are known to be rich in repetitive sequences and prone to gene conversion and unequal crossing-over (14, 15).

In the central region of the chromosome, genetic variation is much reduced compared with the subtelomeric regions. The functional categories of polymorphic genes are highly nonrandom, with the most frequent polymorphisms being in known antigenic determinants and proteins associated with the cell membrane. Discounting hypothetical proteins and those of unknown function, membrane-associated proteins are queried by less than 40% of all probes but account for more than 85% of all detected polymorphism. A number of hypothetical proteins are also highly polymorphic, suggesting that these genes may be under genetic selection pressures similar to those experienced by antigenic and membrane-protein genes. These could represent genes that have important functions in parasite viability or virulence and that warrant further functional characterization.

References and Notes
1. Sequence data for P. falciparum chromosomes 1, 3 through 9, and 13 can be obtained from The Sanger Institute (www.sanger.ac.uk/Projects/P_falciparum/).
2. Sequence data for P. falciparum chromosome 12 can be obtained from the Stanford Genome Technology Center (www-sequence.stanford.edu/group/malaria/).
3. Preliminary sequence data for P. falciparum chromosomes 10, 11, and 14 can be obtained from The Institute for Genomic Research (www.tigr.org).
7. Supplementary material are available on Science Online.
16. Supported by NIH grants GM61335 (S.L.H. and D.F.W.); Burroughs Wellcome Fund New Initiatives in Malaria Research Award: Ellison Medical Foundation, Program in Career Development, Research and Training in Global Infectious Diseases (D.L.H. and D.F.W.); ExxonMobil Program on Malaria in Africa (D.F.W.); and a new scholars award from the Ellison Medical Foundation (E.A.W.). We wish to thank the scientists and funding agencies comprising the International Malaria Genome Sequencing Project for making sequence data from the genome of P. falciparum (3D7) public prior to publication of the completed sequence. The Sanger Centre (UK) provided sequence for chromosomes 1, 3 through 9, and 13, with financial support from the Wellcome Trust. A consortium composed of The Institute for Genome Research, along with the Naval Medical Research Center (USA), sequenced chromosomes 2, 10, 11, and 14, with support from NIAID/NIH, the Burroughs Wellcome Fund, and the U.S. Department of Defense. The Stanford Genome Technology Center (USA) sequenced chromosome 12, with support from the Burroughs Wellcome Fund.

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5951/216/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Table S1
References and Notes
2 July 2002; accepted 9 August 2002