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 replicates 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.

23. E. Lander, L. Kruglyak, *Nature Genet.* **11**, 241 (1995).
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.
25. L. Zheng *et al.*, *Science* **276**, 425 (1997).
26. S. F. Traoré, O. Niare, K. D. Vernick, data not shown.
27. J. Flint, R. Mott, *Nature Rev. Genet.* **2**, 437 (2001).
28. C. J. Talbot *et al.*, *Nature Genet.* **21**, 305 (1999).
29. C. L. Peichel *et al.*, *Nature* **414**, 901 (2001).
30. E. B. Holub, *Nature Rev. Genet.* **2**, 516 (2001).
31. J. Bergelson, M. Kreitman, E. A. Stahl, D. Tian, *Science* **292**, 2281 (2001).
32. K. W. Deitsch, E. R. Moxon, T. E. Wellems, *Microbiol. Mol. Biol. Rev.* **61**, 281 (1997).
33. H. M. Ferguson, A. F. Read, *Proc. R. Soc. London Ser. B* **269**, 1217 (2002).
34. G. Pringle, *Trans. R. Soc. Trop. Med. Hyg.* **60**, 626 (1966).

35. We acknowledge the support of O. Doumbo and the Equipe Parasitologique (FMPOS, Mali) and the organizational assistance of S. Karambe, R. Sakai, and R. Gwadz. Supported by a grant from the National Institute of Allergy and Infectious Disease/NIH to K.D.V., a DOE/Sloan Foundation Fellowship in Computational Molecular Biology to K.M., DFG Sonderforscungbereich 544 to F.C.K., a Ph.D. Fellowship from the University of Heidelberg to J.V., and a James S. McDonnell Centennial Fellowship to L.K.

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5591/213/DC1
Materials and Methods
Figs. S1 and S2
Table S1
Data Sets

30 April 2002; accepted 9 September 2002

Excess Polymorphisms in Genes for Membrane Proteins in *Plasmodium falciparum*

Sarah K. Volkman,[1] Daniel L. Hartl,[2] Dyann F. Wirth,[1] Kaare M. Nielsen,[2,3] Mehee Choi,[2] Serge Batalov,[4] Yingyao Zhou,[4] David Plouffe,[4] Karine G. Le Roch,[4] Ruben Abagyan,[4] Elizabeth A. Winzeler[4,5]*

The detection of single-nucleotide polymorphisms in pathogenic microorganisms has normally been carried out by trial and error. Here we show that DNA hybridization with high-density oligonucleotide arrays provides rapid and convenient detection of single-nucleotide polymorphisms in *Plasmodium falciparum*, despite its exceptionally high adenine-thymine (AT) content (82%). A disproportionate number of polymorphisms are found in genes encoding proteins associated with the cell membrane. These genes are targets for only 22% of the oligonucleotide probes but account for 69% of the polymorphisms. Genetic variation is also enriched in subtelomeric regions, which account for 22% of the chromosome but 76% of the polymorphisms.

The complete genomic sequence of *P. falciparum* has been determined and is in the final stages of assembly and annotation (*1*–*3*). The great challenge now is how best to use the genome sequence for public health and clinical applications. One approach is to identify single-nucleotide polymorphisms (SNPs) in order to pinpoint the origin and map the spread of contagious diseases, to identify and track new mutations that confer resistance to drugs or vaccine-induced immunity, and potentially to identify candidate genes for novel therapeutic or immunological intervention. Although SNP detection on a genome-wide scale is technically difficult, we reasoned that it might be feasible to detect SNPs in *P. falciparum* by means of high-density oligonucleotide arrays, even though such arrays were originally developed for gene-expression studies.

Owing to the relatively short probe sequences used for oligonucleotide arrays, the strength of hybridization between a probe and its target sequence depends largely on perfect complementarity between the probe and the target. An SNP or a small deletion or insertion in the target will reduce the hybridization signal (*4*). Because the exact genomic position of each probe is known, the location of variant sequences can be found by comparing the intensity of oligonucleotide hybridization between genomic DNA from an unknown strain and that from the 3D7 reference strain of *P. falciparum*, from whose genomic sequence the oligonucleotides were designed (*5*). There are two major technical obstacles to the use of oligonucleotide arrays for *P. falciparum*. First, the AT content of the genome is unusually high even in coding regions, and in some genes the AT content of the third positions of codons is nearly as high as the genetic code allows. Second, preparations of DNA from *P. falciparum* may also contain quantities of human DNA.

The feasibility of the approach was tested with the complete sequence of chromosome 2, which contains 210 annotated genes (*6*). Using an oligonucleotide selector algorithm, we chose a unique 25-nucleotide (nt) sequence to match protein-coding sequences at intervals of ~200 nucleotides (yielding between 2 and 117 probes per gene, depending on size). The probes were designed to have similar melting temperatures and to avoid runs of As and Ts but otherwise to be as different from each other as possible. A total of 4167 single-stranded probes were designed (*7*), manufactured by means of photosensitive DNA synthesis technology (*8*), and positioned by Affymetrix onto a prototype array that also included 395,833 probes for ~80,000 different cDNAs from human tissues.

To evaluate the robustness and specificity of the *P. falciparum* probes and to rule out artifacts due to contaminating human DNA, we prepared genomic DNA from different parasite isolates that had been cultured in human erythrocytes, then labeled the DNA and hybridized it to the oligonucleotide array. After hybridization, the integrated intensity (an estimate of the copy number) was determined for each of the 210 *P. falciparum* genes probed on the array (Fig. 1). The mean of the integrated 3D7 intensity for probes at the chromosome ends was higher than for probes within the central region of the chromosome, presumably because of duplication of some of these probe sequences. Even though oligonucleotide probes are thought to differ substantially in their hybridization properties, the integrated signals were generally consistent, varying by not much more than a factor of 2 (fig. S1). In contrast, for the W2 isolate, the integrated intensity across a region on the left arm of the chromosome was 30- to 50-fold lower than the mean integrated intensity for genes in the central region of the

[1]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA. [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [3]Department of Pharmacy, University of Tromso, Tromso N-9037, Norway. [4]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA. [5]Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

*To whom correspondence should be addressed. E-mail: winzeler@gnf.org

chromosome (fig. S1). This finding is consistent with the observation that W2 carries a deletion in the region around the *PfEMP3* gene (*9*), and our hybridization data define the extent of the deletion as including genes PFB0070w through PFB0100c. We found virtually no cross-hybridization to the human oligonucleotides with *P. falciparum* DNA extracted from erythrocyte cultures. Nor was there significant cross-hybridization to the *P. falciparum* oligonucleotides with human DNA from individuals not infected with the parasite (fig. S1).

Using oligonucleotide hybridization as an SNP detector, we tested four isolates of *P. falciparum* from the geographically diverse areas of Honduras (HB3), Southeast Asia (W2), Sierra Leone (D6), and Brazil (7G8). (The reference strain 3D7 was isolated in the Netherlands.) DNA was isolated from each parasite culture and hybridized to the array in three independent experiments. Probes detecting putative SNPs are expected to vary in 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, nonsubtelomeric 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 ($P \approx 10^{-6}$). 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-

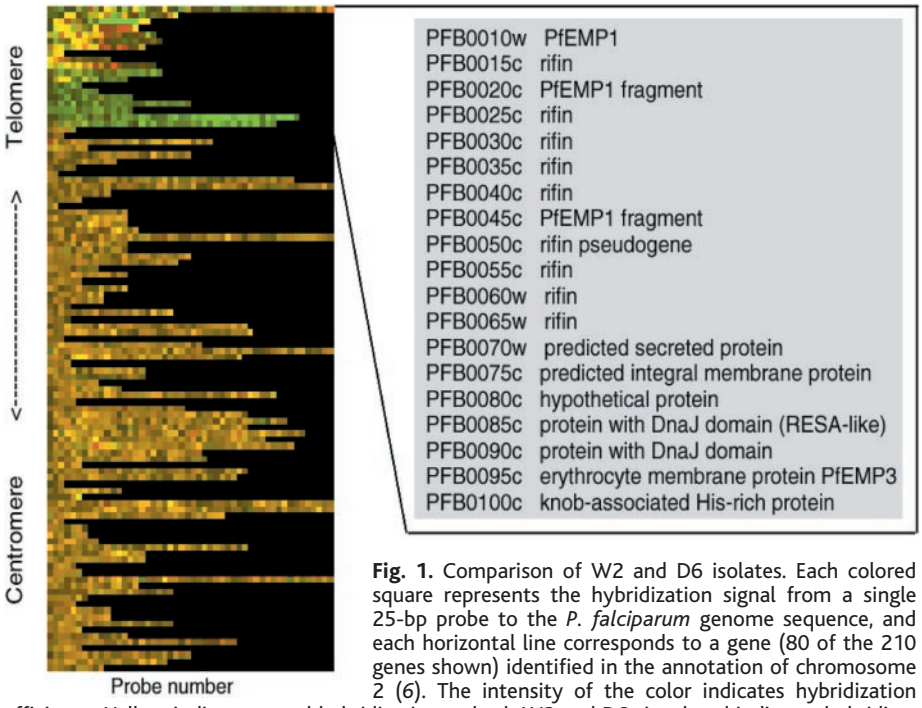

Fig. 1. Comparison of W2 and D6 isolates. Each colored square represents the hybridization signal from a single 25-bp probe to the *P. falciparum* genome sequence, and each horizontal line corresponds to a gene (80 of the 210 genes shown) identified in the annotation of chromosome 2 (*6*). The intensity of the color indicates hybridization efficiency. Yellow indicates equal hybridization to both W2 and D6 signal, red indicates hybridization to D6 but not to W2, and green indicates hybridization to W2 but not to D6.

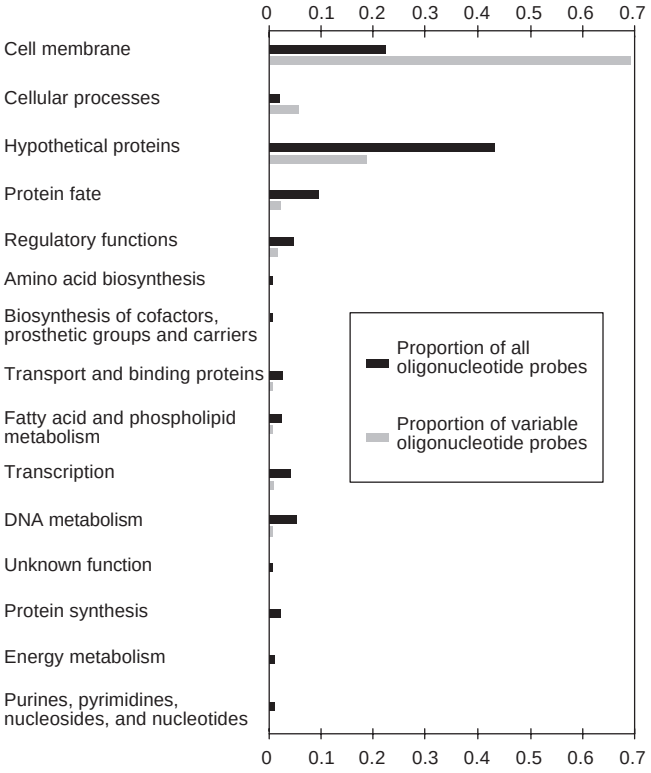

Fig. 2. Proportion of polymorphic probes per functional classification. The proportion of all oligonucleotide probes is represented by black bars, and the proportion of variable oligonucleotide probes is represented by gray bars.

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 ($P \approx 0$).

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—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 ~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 polymorphisms. 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.

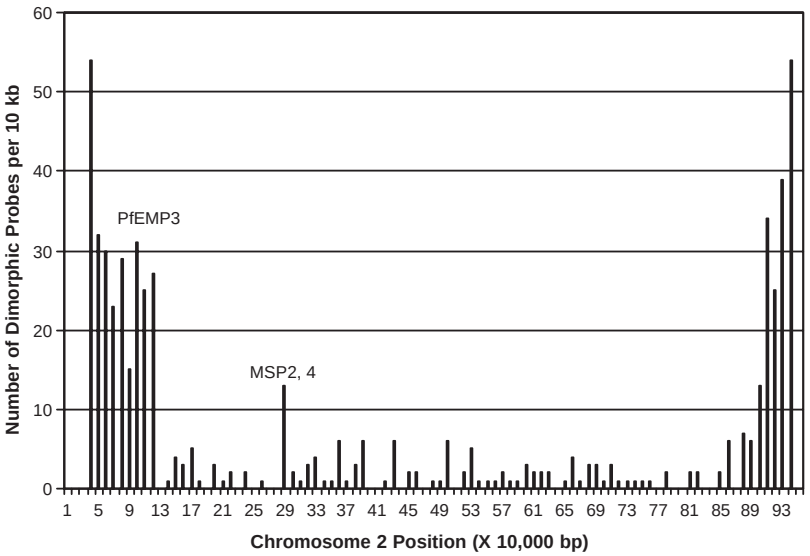

Fig. 3. The distribution of all the variation on chromosome 2, with each position bin representing 10 kb of sequence.

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).
4. M. Chee *et al.*, *Science* **274**, 610 (1996).
5. E. A. Winzeler *et al.*, *Science* **285**, 901 (1999).
6. M. J. Gardner *et al.*, *Science* **282**, 1126 (1998).
7. Supplementary material are available on *Science* Online.
8. A. C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994).
9. L. M. Corcoran, J. K. Thompson, D. Walliker, D. J. Kemp, *Cell* **53**, 807 (1988).
10. S. K. Volkman *et al.*, *Science* **293**, 482 (2001).
11. J. Mu *et al.*, *Nature* **418**, 323 (2002).
12. A. L. Hughes, F. Verra, *Proc. R. Soc. London Ser. B* **268**, 1855 (2001).
13. A. G. Clark, *Nature* **418**, 283 (2002).
14. H. M. Taylor, S. A. Kyes, C. I. Newbold, *Mol. Biochem. Parasitol.* **110**, 391 (2000).
15. L. H. Freitas-Junior *et al.*, *Nature* **407**, 1018 (2000).

16. Supported by NIH grant GM61351 (S.K.V., D.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/298/5591/216/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Table S1
References and Notes

2 July 2002; accepted 9 August 2002