Science
AAAS

Whole-Genome Patterns of Common DNA Variation in Three Human Populations

David A. Hinds, *et al.*

Science **307**, 1072 (2005);
DOI: 10.1126/science.1105436

3. R. Sharer, D. Grove, Eds., *Regional Perspectives on the Olmec* (Cambridge Univ. Press, New York, 1989).
4. D. Grove, *J. World Prehistory* **11**, 51 (1997).
5. A. Cyphers, in *Olmec to Aztec: Settlement Patterns in the Ancient Gulf Lowlands*, B. Stark, P. Arnold III, Eds. (Univ. of Arizona Press, Tucson, AZ, 1997), pp. 96–114.
6. K. Flannery, J. Marcus, *Early Formative Pottery of the Valley of Oaxaca, Mexico*, Memoirs of the Museum of Anthropology 27 (Univ. of Michigan, Ann Arbor, MI, 1994).
7. J. Blomster, *Etlatongo: Social Complexity, Interaction, and Village Life in the Mixteca Alta of Oaxaca, Mexico* (Wadsworth, Belmont, CA, 2004).
8. J. Blomster, *Ancient Mesoamerica* **9**, 171 (1998).
9. M. Coe, R. Diehl, *In the Land of the Olmec*, Vol. 1: *The Archaeology of San Lorenzo Tenochtitlán* (Univ. of Texas Press, Austin, TX, 1980).
10. K. Flannery, in *Dumbarton Oaks Conference on the Olmec*, E. Benson, Ed. (Dumbarton Oaks, Washington, DC, 1968), pp. 79–117.
11. K. Flannery, J. Marcus, *J. Anthropol. Archaeol.* **19**, 1 (2000).
12. H. Neff, M. Glascock, "Report on Instrumental Neutron Activation Analysis of Olmec Pottery" (Research Reactor Center, Univ. of Missouri, Columbia, MO, 2002).
13. R. Herrera, H. Neff, M. Glascock, J. Elam, *J. Archaeol. Science* **26**, 967 (1999).
14. M. Thieme, in *Procesos de Cambio y Conceptualización del Tiempo: Memoria de la Primera Mesa Redonda de Monte Albán*, N. Robles García, Ed. (Instituto Nacional de Antropología e Historia, Mexico City, 2001), pp. 341–349.
15. J. Lorenzo, *Rev. Mex. Estud. Antropol.* **16**, 49 (1960).
16. M. Kirkby, *The Physical Environment of the Nochixtlán Valley, Oaxaca*, Vanderbilt University Publications in Anthropology 2 (Vanderbilt Univ., Nashville, TN, 1972).
17. M. Glascock, in *Chemical Characterization of Ceramic Pastes in Archaeology*, H. Neff, Ed. (Prehistory Press, Madison, WI, 1992), pp. 11–26.
18. H. Neff, in *Modern Analytical Methods in Art and Archaeology*, E. Ciliberto, G. Spoto, Eds. (Wiley, New York, 2000), pp. 81–134.
19. Materials and methods are available as supporting material on *Science* Online.
20. R. Bishop, H. Neff, in *Archaeological Chemistry IV*, R. Allen, Ed. (American Chemical Society, Washington, DC, 1989), pp. 576–586.
21. P. Joralemon, *A Study of Olmec Iconography*, Studies in Pre-Columbian Art and Archaeology 7 (Dumbarton Oaks, Washington, DC, 1971).
22. J. Clark, M. Pye, in *Olmec Art and Archaeology in Mesoamerica*, J. Clark, M. Pye, Eds. (National Gallery of Art, Washington, DC, 2000), pp. 217–251.

23. We thank the technicians at MURR and the archaeologists in Mexico who provided samples for this study, as well as funding agencies that made their collections possible. The support and permissions of the Instituto Nacional de Antropología e Historia—especially L. Mirambell, M. Serra Puche, J. García-Bárcena, E. López Calzada, and N. Robles Garcia—have been invaluable. Funding for the 80 samples in table S1 came primarily from the Charles J. MacCurdy Endowment, Yale University; the non-Etlatongo samples were provided by M. Winter and M. Coe. The research at the MURR Archaeometry Lab has been supported by NSF (no. SBR-9802366). We are grateful for the support of the Sainsbury Research Unit, University of East Anglia. We thank R. Diehl, A. Joyce, G. Lau, and an anonymous reviewer for comments.

Supporting Online Material
www.sciencemag.org/cgi/content/full/307/5712/1068/DC1
Materials and Methods
Tables S1 to S3
References

16 November 2004; accepted 6 January 2005
10.1126/science.1107599

Whole-Genome Patterns of Common DNA Variation in Three Human Populations

David A. Hinds,[1] Laura L. Stuve,[1] Geoffrey B. Nilsen,[1] Eran Halperin,[2] Eleazar Eskin,[3] Dennis G. Ballinger,[1] Kelly A. Frazer,[1] David R. Cox[1]*

Individual differences in DNA sequence are the genetic basis of human variability. We have characterized whole-genome patterns of common human DNA variation by genotyping 1,586,383 single-nucleotide polymorphisms (SNPs) in 71 Americans of European, African, and Asian ancestry. Our results indicate that these SNPs capture most common genetic variation as a result of linkage disequilibrium, the correlation among common SNP alleles. We observe a strong correlation between extended regions of linkage disequilibrium and functional genomic elements. Our data provide a tool for exploring many questions that remain regarding the causal role of common human DNA variation in complex human traits and for investigating the nature of genetic variation within and between human populations.

Single-nucleotide polymorphisms (SNPs) are the most abundant form of DNA variation in the human genome. It has been estimated that there are ~7 million common SNPs with a minor allele frequency (MAF) of at least 5% across the entire human population (*1*). Most common SNPs are to be found in most major populations, although the frequency of any allele may vary considerably between populations (*2*). An additional 4 million SNPs exist with a MAF between 1 and 5%. In addition, there are innumerable very rare single-base variants, most of which exist in only a single individual.

The relationship between DNA variation and human phenotypic differences (such as height, eye color, and disease susceptibility) is poorly understood. Although there is evidence that both common SNPs and rare variants contribute to the observed variation in complex human traits (*3*, *4*), the relative contribution of common versus rare variants remains to be determined. The structure of genetic variation between populations and its relationship to phenotypic variation is unclear. Similarly, the relative contribution to complex human traits of DNA variants that alter protein structure by amino acid replacement, versus variants that alter the spatial or temporal pattern of gene expression without altering protein structure, is unknown. In some cases, these issues have been studied in limited genomic intervals, but comprehensive genomic analyses have not been possible.

Genome-wide association studies to identify alleles contributing to complex traits of medical interest are currently performed with subsets of common SNPs, and thus they rely on the expectation that a disease allele is likely to be correlated with an allele of an assayed SNP. Although studies have shown that variants in close physical proximity are often strongly correlated, this correlation structure, or linkage disequilibrium (LD), is complex and varies from one region of the genome to another, as well as between different populations (*5*, *6*). Selection of a maximally informative subset of common SNPs for use in association studies is necessary to provide sufficient power to assess the causal role of common DNA variation in complex human traits. Although a large fraction of all common human SNPs are available in public databases, lack of information concerning SNP allele frequencies and the correlation structure of SNPs within and between human populations has made it difficult to select an optimal subset.

Here we examine the SNP allele frequencies and patterns of LD between 1,586,383 SNPs distributed uniformly across the human genome in unrelated individuals of European, African, and Asian ancestry. Our primary aim was to create a resource for further investigation of the structure of human genetic variation and its relationship to phenotypic differences.

A dense SNP map. To characterize a panel of markers that would be informative in whole-genome association studies, we selected a total of 2,384,494 SNPs likely to be common in individuals of diverse ancestry (*7*). We identified the majority (69%) of the SNPs by performing array-based resequencing

[1]Perlegen Sciences Inc., 2021 Stierlin Court, Mountain View, CA 94043, USA. [2]International Computer Science Institute, Berkeley, CA 94704, USA. [3]Department of Computer Science and Engineering, University of California–San Diego, La Jolla, CA 92093, USA.

*To whom correspondence should be addressed. E-mail: david_cox@perlegen.com

of 24 human DNA samples of diverse ancestry (*5*). These SNPs were supplemented with SNPs chosen from public databases to obtain a more uniform physical distribution across the human genome. Further details of the SNP ascertainment are given in the supporting online material (*7*). We designed 49 high-density oligonucleotide arrays for genotyping these SNPs (*8*, *9*) and roughly 300,000 long-range polymerase chain reaction (PCR) primer pairs covering the selected SNPs, with an average of eight SNPs per individual region being amplified by PCR. The amplicons had an average length of 9 kb and covered ~92% of the available human genome. An average of 6250 amplicons derived from a single individual were pooled and hybridized to a single high-density oligonucleotide array, producing genotypes for ~48,000 SNPs.

We genotyped 71 unrelated individuals from three populations: 24 European Americans, 23 African Americans, and 24 Han Chinese from the Los Angeles area. The 71 individuals genotyped here were not related to the individuals previously used for SNP discovery. DNA samples were selected from the Coriell Cell Repositories' Human Variation Collection, and we relied on Coriell's determinations of sample populations. We complied with all Coriell policies for research use DNA of samples from named populations.

Each SNP was scored with a combination of metrics that had been shown to correlate with genotype quality on our platform, and data for poorly performing SNPs was rejected. These metrics included the call rate; the number of observed genotype clusters; the existence of near-perfect matches for SNP flanking sequences elsewhere in the genome; the presence of other known SNPs in probe-flanking sequences; and consistency with Hardy Weinberg equilibrium. Tests for Hardy Weinberg equilibrium are very effective for identifying some types of genotyping artifacts (*10*); however, because we used these tests for quality control, our genotype data are unsuitable for investigating biologically interesting true deviations from Hardy Weinberg equilibrium. Further details of our genotype quality control are described in the supporting online material (*7*).

A subset of 1,586,383 SNPs was successfully genotyped based on our quality criteria, with two alleles each observed at least once among the 71 individuals. In total, more than 112 million individual genotypes were determined for these SNPs. There were no missing genotypes for 64% of these SNPs, and 92% of these SNPs had less than 5% missing data. The overall frequency of successful genotype calls was 98.6%. SNP assay details and individual genotypes have been deposited in the National Center for Biotechnology Information (NCBI)'s SNP database (dbSNP, build 123, accession nos. ss23145044 to ss24731426). Genotypes for 156,757 SNPs for nine of the European-American individuals that were part of this project had been previously determined by the International HapMap Project, using a variety of genotyping platforms (*11*). Our data for these 1.6 million genotypes is 99.54% concordant with the HapMap project data. The distribution of discordant genotypes is very nonrandom; only 0.3% of the SNPs account for 50% of all the discrepancies, and we estimate that 90% of the SNPs in the complete data set have no incorrect genotypes. Haplotype analyses in particular will generally benefit from this error distribution, because accurate inference of haplotypes requires consistent genotypes across large groups of nearby markers.

The distribution of the 1.6 million high-quality genotyped SNPs (table S1) is similar to that of a previously reported map of 1.42 million SNPs (*12*). More than 95% of the genome is in inter-SNP intervals of less than 50 kb, and roughly two-thirds of the sequenced genome is covered by inter-SNP intervals of 10 kb or less (table S2). The average distance between adjacent SNPs is 1871 base pairs (bp). Although repetitive elements are underrepresented in our collection, we genotyped 269,611 SNPs within repetitive elements where the SNP flanking sequences could be uniquely mapped. There are 735,094 SNPs (46%) in genic regions of the genome, which we define as being within 10 kb of the transcribed intervals for 22,904 protein-coding genes in release 3 of NCBI's build 34 annotations. At least one SNP is present in 78% of all transcripts. When the 10-kb region of DNA upstream and downstream of each transcript is included, 93% of all the protein-coding genes contain at least one SNP. A total of 20,165 SNPs (1.3%) are present in amino acid coding sequences and 9370 of these SNPs are nonsynonymous, leading to an amino acid change (table S3). Although our SNP ascertainment is not random, this subset of SNPs is quite uniformly distributed throughout the human genome with respect to annotated protein-coding genes as well as physical distance.

Common SNPs in three populations. Table 1 illustrates our success in obtaining a set of common SNPs that are informative in human populations of diverse ancestry. Most of the 1,586,383 SNPs with high-quality genotypes are polymorphic in each of the three population samples genotyped in this study. Ninety-four percent of the SNPs (1,483,594 SNPs) have two alleles in the African-American sample; 81% (1,286,277 SNPs) have two alleles in the European-American sample; and 74% (1,168,029 SNPs) have two alleles in the Han Chinese sample. In each population, the majority of the segregating SNPs have a MAF greater than 10%, ranging from 68% of all segregating SNPs in the African-American sample to 57% of all segregating SNPs in the Han Chinese sample. Only 263,029 of the 1,586,383 SNPs (17%) have a MAF of less than 10% in all three of the population samples. The distributions of MAFs we see in the three populations is very similar for the European-American and Han Chinese samples, with a higher frequency of rarer alleles in the African-American sample (fig. S1). Consistent with previous studies (*2*, *13*), we observed the greatest genetic diversity in individuals of African descent. Our SNP ascertainment strategy makes it difficult to make more definitive statements regarding the precise distribution of SNP allele frequencies in different populations.

Although the small sample sizes in this study preclude any definite conclusion regarding the complete absence of a particular allele in any given population, we observed 291,012 SNPs (18%) that were segregating in only one population sample (''private SNPs''). Most of these private SNPs (75%) were segregating in the African-American sample, although private SNPs were observed for each of the three population samples (Table 1). Although private SNPs tend to have lower MAFs than other SNPs in our collection, a substantial fraction are common: 106,404, or 37%, have MAF > 0.10.

To quantify genetic variation within and between populations, we calculated F_{ST} for each SNP in each pair of populations, as well as combined values across all three popula-

Table 1. SNPs segregating in the three genotyped populations. Percentages are of 1,586,383 genotyped SNPs or of 291,012 private SNPs.

Population	Segregating		MAF > 0.05		MAF > 0.10	
	SNPs	%	SNPs	%	SNPs	%
All SNPs						
African-American	1,483,594	93.5	1,267,594	79.9	1,083,652	68.3
European-American	1,286,277	81.1	1,123,765	70.8	991,046	62.5
Han Chinese	1,168,029	73.6	1,027,109	64.7	910,451	57.4
Private SNPs						
African-American	218,500	75.1	139,536	47.9	88,525	30.4
European-American	44,555	15.3	18,284	6.3	8,062	2.8
Han Chinese	27,957	9.6	15,946	5.5	9,817	3.4

tions (*14*). F_{ST} measures the genetic variance between populations as a fraction of the total genetic variance. Because African Americans are a relatively admixed population with substantial but heterogeneous European genetic contributions (*15*), the F_{ST} estimates for comparisons with this group will be more variable but should generally underestimate the results that would be obtained with a native African sample. The distribution of pairwise F_{ST} is very similar for the African-American versus European-American and European-American versus Han Chinese samples, with more large F_{ST} values between the African-American and Han Chinese samples (fig. S2). These findings are consistent with prior studies (*16*, *17*) showing that most common DNA variation is shared across human populations, with differences in allele frequencies between populations.

Markers with large between-population variance will be useful for admixture mapping studies to identify genetic variants causing phenotypic differences (*18*). Admixture mapping exploits relatively long-range allelic correlations in a recently admixed population to identify functional variants that have different prevalences in the ancestral populations, whether because of genetic drift or local natural selection. The technique requires selection and genotyping of limited numbers of "ancestry-informative markers." Our identification of large numbers of such markers removes one of the major barriers to practical use of this promising but largely untested technique.

Evidence for natural selection between populations. It has been suggested that natural selection distorts the observed distribution of F_{ST} across the human genome and that large F_{ST} values can be used to identify candidate loci likely to have undergone local selection (*13*, *19*). If this is true, then larger F_{ST} values should be found near functional genetic elements. We looked at the distribution of F_{ST} for SNPs that were genic or nongenic, coding or noncoding, and synonymous or nonsynonymous. We performed the analysis within subsets of SNPs grouped by MAF, so that effectively, we looked at the fraction of between-population variance for SNPs with the same total genetic variance (fig. S3). Common SNPs in genic regions do have slightly but significantly higher F_{ST} values than nongenic SNPs with the same MAF [analysis of variance (ANOVA), $P = 1.8 \times 10^{-46}$], and common coding SNPs have slightly higher F_{ST} values than noncoding SNPs in genic regions (ANOVA, $P = 1.1 \times 10^{-4}$). We did not see a significant difference in F_{ST} between synonymous and nonsynonymous coding SNPs, but our sensitivity is limited by the small sample sizes and expected correlations among SNPs within the same transcript. These results are consistent with local selection changing the distribution of F_{ST} near functional sequences. However, because the distributions of F_{ST} among genic and nongenic SNPs are very similar, large F_{ST} values by themselves appear to be very weak evidence of selection.

We performed a similar analysis to see if there is also an association between private SNPs and functional genetic elements. When conditioned on MAF, we saw no difference in frequency of private SNPs among genic and nongenic SNPs or among coding and noncoding SNPs (fig. S4). This indicates that the SNPs responsible for evidence of local selection in the F_{ST} analysis tend not to be private and instead are segregating in multiple populations. Although there are known examples linking population-specific SNP alleles to phenotypic differences (*20*–*22*), our results are more consistent with the conclusion that most functional human genetic variation is not population-specific.

Correlation structure of common SNPs. DNA variants in physical proximity along a chromosome tend to be correlated, and these correlations are known as linkage disequilibrium. LD results from a combination of processes, including mutation, natural selection, and genetic drift. It can initially extend over very long genomic distances but is steadily broken down over time by recombination. The observed structure of LD in any particular genomic interval thus depends on a complex interplay of demographic history, stochastic events, and functional constraints. Several metrics exist for measuring LD between pairs of SNPs; we used r^2, the squared correlation coefficient for a 2 by 2 table of haplotype frequencies (*23*).

We have used a modification of a previously described algorithm to identify bins of common SNPs that are in very strong LD, where each bin has at least one "tag SNP" with an r^2 of at least 0.8 with every other SNP in the bin (*24*). This "greedy" algorithm works by iteratively identifying the largest possible subset with these properties from a list of available SNPs, then removing those SNPs from the list used in the next iteration. By assaying a reduced set of tag SNPs, the genotyping burden of an association study may be substantially reduced while retaining most of the power to discover disease associations of the entire SNP set. Unlike haplotype blocks, which are defined as contiguous groups of SNPs, the SNPs that make up a bin may be interdigitated with SNPs that are part of other bins.

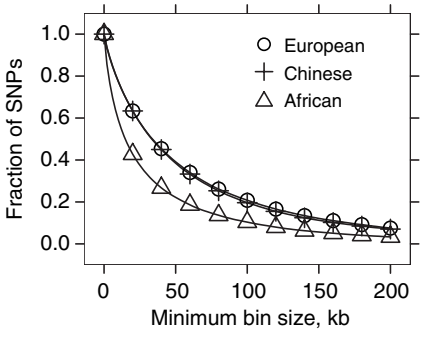


Fig. 1. Size distribution of LD bins. We show, for a given minimum bin size, the fraction of SNPs in bins of that size or larger. The size distributions for the European-American and Han Chinese LD maps are essentially identical.

Table 2 summarizes bin characteristics across the genome, excluding the Y chromosome, for each of the three population samples. We focused on common SNPs with MAF > 10% in this analysis, because estimates of LD for variants with lower MAF are unreliable unless large numbers of individuals are genotyped (*23*). Although most LD bins contained just one SNP, these isolated SNPs were a small proportion of all SNPs, and most SNPs were tightly correlated with multiple other SNPs. In the European-American data, 52.3% of 293,677 bins contained one SNP; however, these constituted only 15.5% of the 991,185 common SNPs. A substantial portion of all SNPs qualified as tag SNPs by having a high r^2 value with every other bin member, indicating that the bins are generally quite densely connected. For the African-American sample, there were substantially fewer bins made up of large numbers of SNPs extending over large distances (Fig. 1). It should be kept in mind that the LD structure we observed is based on an analysis of only 25% of all common SNPs in the genome. Although the sizes of longer intervals of LD should be relatively robust to our incomplete ascertainment, the proportion of all common SNPs in high LD with other SNPs may be substantially underestimated from our data.

LD and functional elements. We observed a strong relationship between extended intervals of LD and functional genomic fea-

Table 2. LD bin statistics in three populations. Bins were classified by the number of SNPs they contained.

Size*	Bins	% Bins	kb†	SNPs	% SNPs
		African-American			
1	362,465	67.4	0.0	362,465	33.5
2 to 4	131,737	24.5	12.4	337,877	31.2
5 to 9	32,081	6.0	37.2	202,512	18.7
≥10	11,530	2.1	78.4	180,556	16.7
Total	537,813			1,083,410	
		European-American			
1	153,511	52.3	0.0	153,511	15.5
2 to 4	84,890	28.9	14.6	226,172	22.8
5 to 9	33,745	11.5	37.3	218,491	22.0
≥10	21,531	7.3	89.5	393,011	39.7
Total	293,677			991,185	
		Han Chinese			
1	129,759	50.8	0.0	129,759	14.3
2 to 4	74,232	29.1	13.2	198,422	21.8
5 to 9	30,569	12.0	34.8	198,429	21.8
≥10	20,708	8.1	83.7	383,580	42.1
Total	255,268			910,190	

*The number of SNPs per LD bin. †Average distance spanned by the SNPs in each LD bin, in kb.

tures (Table 3). Large bins were significantly overpopulated with genic versus nongenic SNPs (trend test, $P \approx 0$), and in genic regions, coding SNPs were significantly enriched over noncoding SNPs (trend test, $P = 1.9 \times 10^{-26}$). Large bins were also overrepresented for nonsynonymous versus synonymous SNPs (trend test, $P = 5.3 \times 10^{-4}$). This result is consistent with the hypothesis of an association between selection and some regions of extended LD (*25*, *26*) and suggests that some genomic regions of extended LD may play a particularly important role in determining the genetic basis of human phenotypic differences.

We identified five bins of more than 200 SNPs each and 17 genomic intervals containing bins that span more than 1000 kb in one or more populations (tables S4 and S5). Several of these large bins spanned similarly large genes. The bin with the most SNPs was on chromosome 17 in the European-American map and had an unusual pattern of variation, with two previously reported haplotypes extending across 518 SNPs and spanning a distance of 800 kb (*27*). The rarer haplotype had a frequency of 25% in the European-American sample and a 9% frequency in the African-American sample and was absent in the Han Chinese sample. This bin includes the gene for microtubule-associated protein tau, mutations of which are associated with a variety of neurodegenerative disorders; a gene coding for a protease similar to presenilins, mutations of which result in Alzheimer's disease; and the gene for corticotropin-releasing hormone receptor, which mediates immune, endocrine, autonomic, and behavioral responses to stress (*27*–*29*).

Large-scale patterns of LD. The distribution of SNPs and LD across the entire human genome is shown in Fig. 2 and can be examined in more detail online. The top track illustrates the relative uniformity of coverage of the analyzed SNPs apart from intervals of centromeric and telomeric heterochromatin. The middle track shows the fraction of common SNPs that are in high LD with at least one other SNP. In most regions, we observed a high level of redundancy for the European-American and Han Chinese samples and somewhat less redundancy in the African-American sample. The bottom track shows the fraction of common SNPs observed to be in relatively large LD bins in each population. This track shows substantial structure on a scale of megabases. Although there is generally good agreement between populations, there are also intervals where there is substantial divergence.

Our whole-genome analysis reveals that the large-scale structure of LD across the genome is correlated with large-scale differences in recombination rates, consistent with previous findings for a single chromosome (*30*). In particular, regions of very strong LD are mostly located in regions of low recombination (fig. S5). This correlation of large-scale LD structure with recombination rate and the finding that regions of extended LD show evidence of selection provide strong support for the hypothesis that the LD structure of the human genome has functional significance and is not simply a byproduct of random genetic drift and population demographics.

SNP subsets capture most common variation. As only a fraction of all common SNPs in human populations have been characterized to date, association studies based on available subsets of SNPs rely on the expectation that an undiscovered, disease-associated variant is likely to be correlated with an allele of an assayed SNP. The statistical power to detect an unassayed, disease-associated allele indirectly with a correlated allele of an assayed SNP is related to r^2. Specifically, the power to detect an association indirectly in N individuals is equivalent to the power to detect it directly in Nr^2 individuals (*31*). The actual power to detect a particular causal variant depends on that variant's mode of action and penetrance as well as details of the study design. Thus, r^2 can only be used to answer the narrower question of what is the sample size penalty, in an otherwise appropriately designed study, for not directly assaying a causal variant.

To determine our ability to detect unassayed, disease-associated variants with this SNP collection, we took advantage of the fact that the European-American and African-American individuals genotyped in this study were also sequenced across selected genes by the SeattleSNPs Program for Genomic Applications (PGA) (*32*). For these individuals, this data provides an essentially complete assessment of genetic variation in the sequenced regions, allowing us to estimate the fraction of all variation contained in our SNP set. In addition, the data allows us to determine the coverage of our genotyped SNPs for the sites we did not directly assay.

We evaluated data for 16,601 sequence variants identified in 152 genes, of which 2465 were part of our SNP set. The concordance between our genotype data and the PGA data for these 2465 SNPs was 99.2%. Our SNP set contained ~24% of all SNPs with a MAF ≥ 10% for these 152 genes in the African-American and European-American samples. SNPs with low MAF are underrepresented in our data compared to the PGA data, because our SNPs were typically discovered with sequence data from fewer distinct chromosomes. These rarer variants account for relatively small fractions of the total nucleotide diversity. In the PGA data for the European Americans, 45% of SNPs have MAF < 10% but account for only 15% of nucleotide diversity; for the African Americans, 58% of SNPs have MAF < 10% and account for 23% of nucleotide diversity.

Table 4 shows the average r^2 and the fraction of r^2 values exceeding thresholds, for

Table 3. Distribution of genic, synonymous, and nonsynonymous coding SNPs spanned by bins of extended LD in any of the three population samples. Genic SNPs are defined as within 10 kb of a protein-coding gene annotation.

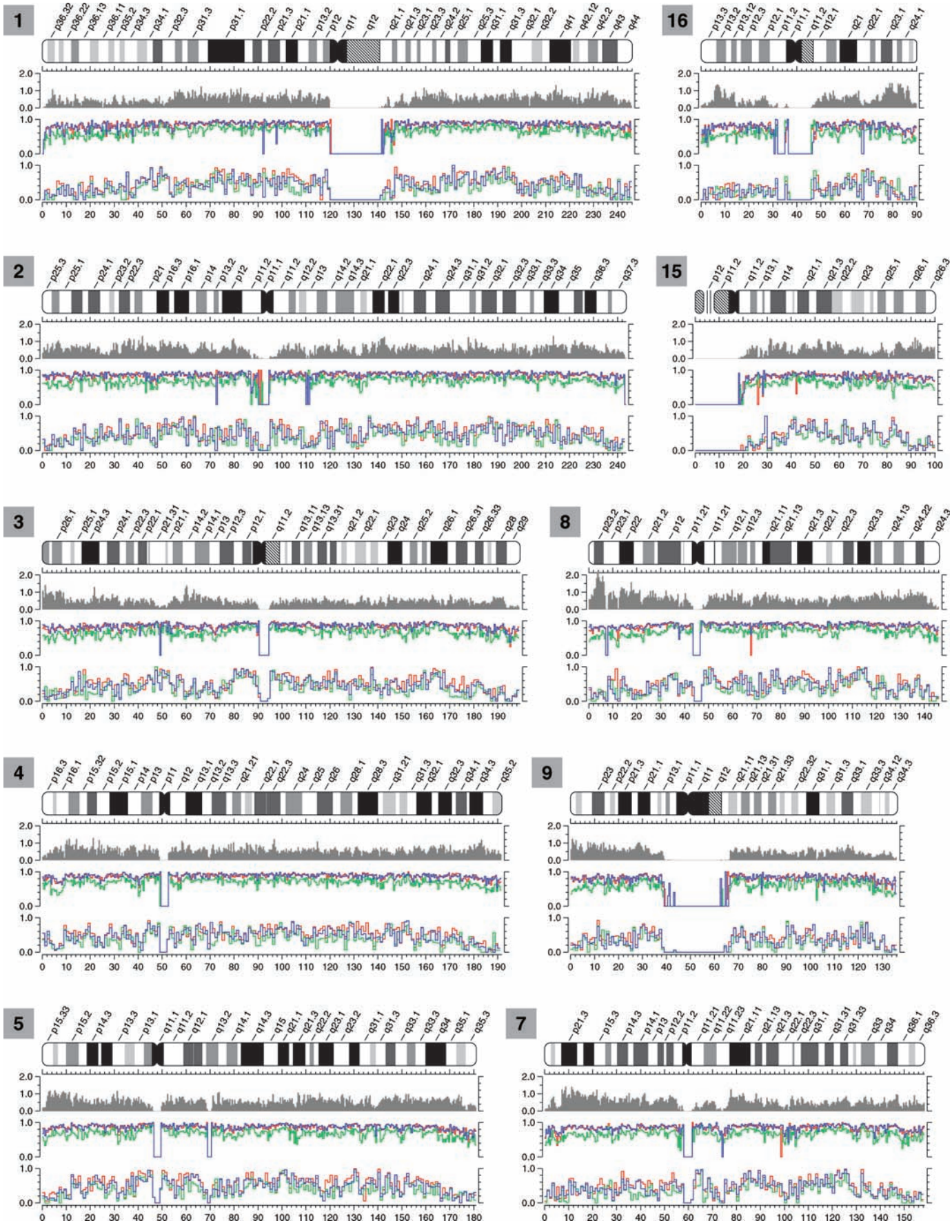
Longest spanning LD bin (kb)	SNPs	Genic		Synonymous		Nonsynonymous	
		SNPs	%	SNPs	%	SNPs	%
<500	1,536,094	707,950	46.1	10,330	0.67	8,898	0.58
500 to 1000	42,432	22,189	52.3	347	0.82	302	0.71
≥1000	7,857	4,955	63.1	120	1.52	171	2.17

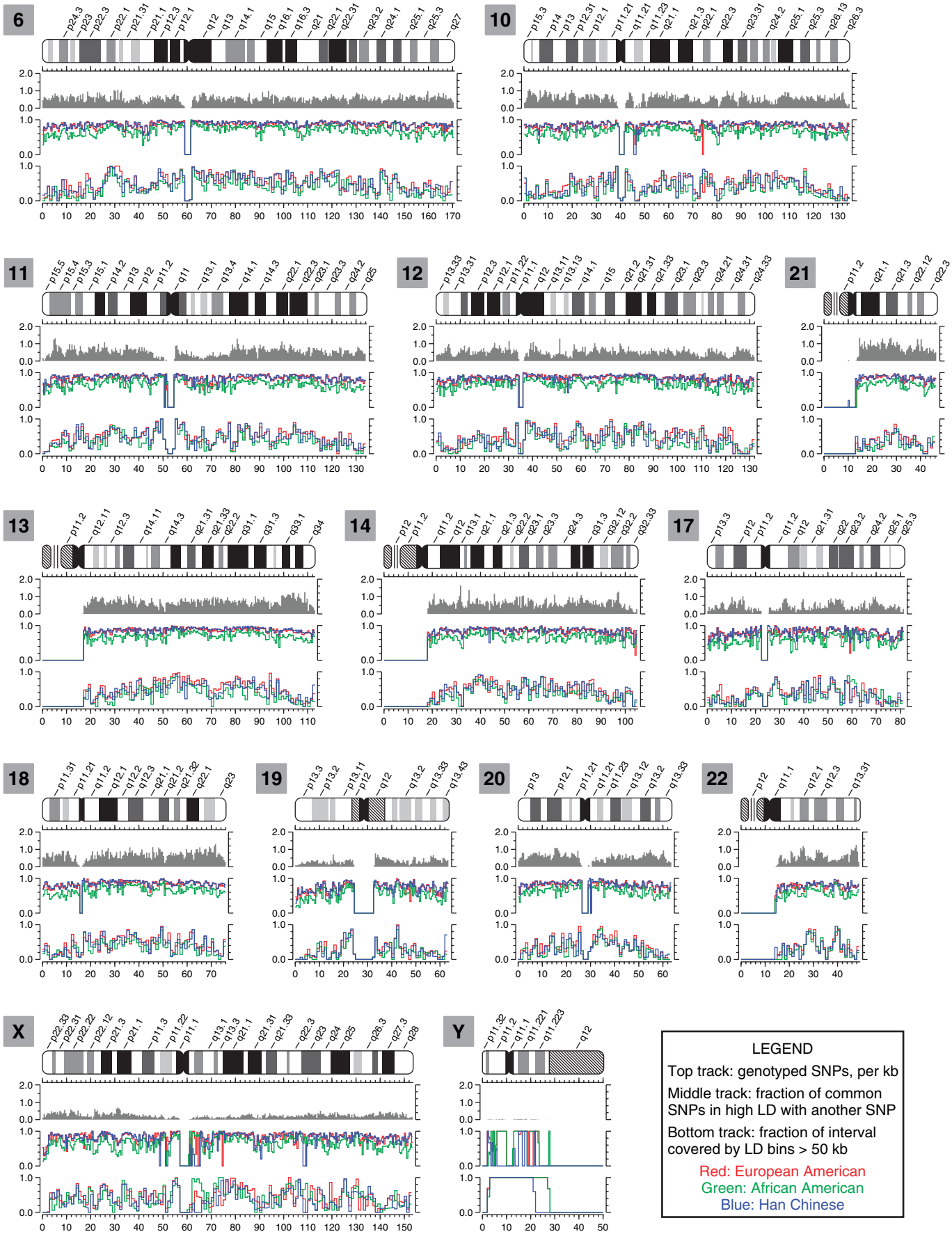
Table 4. LD statistics for common SNPs genotyped in this study, with common variants identified by complete resequencing in 152 genes.

Subset*	Yield (%)†	r^2‡	$r^2 > 0.5$ (%)§	$r^2 > 0.8$ (%)	$r^2 = 1.0$ (%)
African-American					
All	23.3	0.715	70.9	53.7	41.5
Tag	12.3	0.698	70.1	51.9	33.2
European-American					
All	25.0	0.841	86.5	72.6	62.4
Tag	8.1	0.810	85.6	69.7	44.8

*SNPs from the current study: either all common SNPs or a minimal tagging subset. †Percentage of all SeattleSNPs PGA variants that were in the selected set. ‡Across all PGA variants, the mean maximum r^2 with a selected SNP in the same locus. §Percentages of PGA variants having an r^2 greater than the specified threshold with any selected SNP in the same locus.

Fig. 2. Distribution of SNP positions and LD structure across the genome. For each chromosome, the top track shows SNP density per kb, with a window size of 500 kb. The middle track shows, for each population, the fraction of common SNPs with MAF > 10% that are in high LD ($r^2 > 0.8$) with at least one other common SNP, with a window size of 500 kb. The bottom track shows, for each population, the fraction of common SNPs that are in an LD bin extending over at least 50 kb, with a window size of 1000 kb.





any PGA SNP with the most-correlated SNP in the same region that was included in our SNP set. These results indicate that, with the stringent threshold of $r^2 > 0.8$, our SNP set ascertains ~73% of common variation in the European-American sample and ~54% of common variation in the African-American sample. These values are similar to those previously predicted if 2.7 million SNPs from public databases were developed into genotyping assays (*17*). This analysis sets a very conservative lower bound on coverage, because it treats SNPs below the threshold of $r^2 = 0.8$ as completely uncovered and does not reward coverage that exceeds the threshold. Using a less stringent threshold of $r^2 > 0.5$, coverage would improve to 86% in the European-American sample and 71% in the African-American sample. The skewed distribution of r^2 toward high values is apparent in the mean values of 0.84 for the European-American sample and 0.72 for the African-American sample. These numbers are especially impressive considering that we did not genotype 75% of all the common SNPs in these intervals.

Selection of one tag SNP from each LD bin for the three population samples yielded 296,313 of the 991,398 SNPs segregating in the European-American sample (30%); 256,766 of the 909,824 SNPs segregating in the Han Chinese sample (28%); and 540,533 of the 1,083,638 SNPs segregating in the African-American sample (50%). When tag SNPs from European Americans and African Americans were used to assess common variation in the PGA data, for MAF ≥ 10%, the amount of all common variation ascertained was reduced very little compared to that ascertained with the complete sets of common SNPs (Table 4). These tag SNP numbers are smaller than have previously been predicted with a similar selection strategy (*24*); however, we did not attempt to achieve 100% coverage as in that work. Although choosing subsets of SNPs based on bin relationships reduces the genotyping burden for a comprehensive whole-genome scan to some degree in all populations, these data indicate that even taking advantage of such tag SNP selection, a comprehensive whole-genome association study requires genotyping each individual for at least several hundreds of thousands of SNPs.

Haplotype block structure. LD maps and haplotype maps represent somewhat different aspects of the local structure of genetic variation. The genetic architecture of a particular phenotype will determine which representation is most powerful for the identification of functional variants (*33*). In parallel with our LD analysis, we used the HAP program (*34*) to infer haplotypes from our diploid genotype data. We partitioned these reconstructed haplotypes into blocks with limited diversity, separately for each of the three population samples. These blocks were defined as sets of SNPs for which at least 80% of the inferred haplotypes could be grouped into common patterns with population frequencies of at least 5%.

Table 5 summarizes the structure of the three resulting haplotype maps for the whole genome, excluding the Y chromosome. The haplotype map statistics across the three populations appear qualitatively similar to the LD maps, with substantially more blocks in the map derived from the African-American sample than in the maps from the European-American and Han Chinese samples. The numbers of SNPs required to represent frequencies of common haplotype patterns were similar to the numbers of tag SNPs identified in the LD maps. Substantial fractions of LD bins of two or more SNPs crossed haplotype block boundaries, ranging from 33% in the Han Chinese map to 48% in the African-American map.

The bin structure for SNPs in the region of the *CFTR* gene on chromosome 7 (Fig. 3) demonstrates some of the differences between the LD bin and haplotype block maps and further illustrates that there can be substantial population differences in local map structure. In this interval, the European-American and African-American LD maps have similar complexity, with multiple overlapping bins, but the Han Chinese map is dominated by two disjoint bins of highly correlated SNPs. Conversely, a break point near the 116,790-kb position is shared in the African-American and Han Chinese LD maps but is bridged by multiple LD groupings in the European-American map. All three haplotype maps share this break point. However, the African-American map contains many more distinct haplotype blocks than the maps for the other two population samples.

Common genetic variation and human health. Our focus on common genetic variation has several motivations. Common variants account for a larger share of human nucleotide diversity than rare variants and are more experimentally tractable. For the same allelic effect, a common variant represents a larger fraction of phenotypic variance and population attributable risk than a rare one, so common variants are more valuable from the perspective of diagnostics and intervention. Finally, detecting and characterizing effects of rare variants requires very large sample sizes to obtain statistically meaningful numbers of individuals carrying

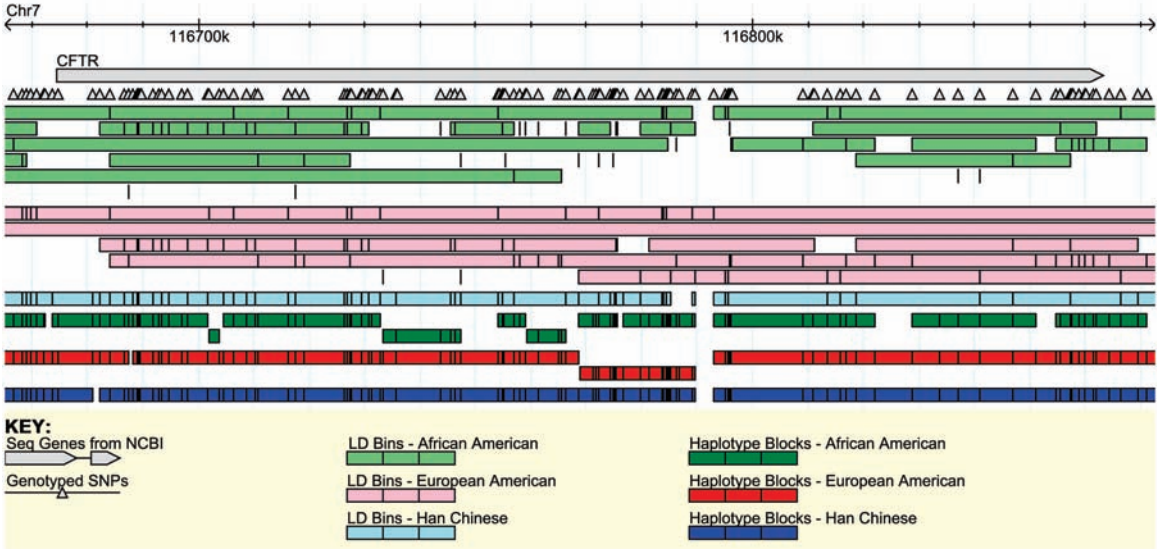


Fig. 3. Extended LD bin and haplotype block structure around the *CFTR* gene. LD bins, where each bin has at least one SNP with $r^2 > 0.8$ with every other SNP, are depicted as light horizontal bars, with the positions of constituent SNPs indicated by vertical tick marks as well as the extreme ends of the bars. Isolated SNPs are indicated by plain tick marks. Haplotype blocks, within which at least 80% of observed haplotypes could be grouped into common patterns with frequencies of at least 5%, are depicted as dark horizontal bars. Unlike haplotype blocks that are by design sequential and nonoverlapping, SNPs in one LD bin can be interdigitated with SNPs in multiple other overlapping bins.

a rare allele. There is no doubt that rare variants play a role in the etiology of common disease, but pursuit of common variants is more tractable with available technologies.

Common human diseases, such as cardiovascular disease and psychiatric illness, are caused by the interplay of multiple genetic and environmental factors. The bounded nature of the human genome and the availability of the complete human genome sequence have resulted in extensive efforts to define the genetic basis of a wide variety of complex human traits. One approach for identifying such genetic risk factors is the case-control association study, in which a group of individuals with disease is found to have an increased frequency of a particular genetic variant compared to a group of control individuals. A number of genetic risk factors for common disease have been identified by such association studies (*3*, *4*, *35*, *36*). These studies suggest that many different genes distributed throughout the human genome contribute to the total genetic variability of a particular complex trait, with any single gene accounting for no more than a few percent of the overall variability of the trait (*37*). Case-control study designs that include on the order of 1000 individuals can provide adequate power to identify genes accounting for only a few percent of the overall genetic variability of a complex trait, even using the very stringent significance levels required when testing large numbers of common DNA variants (*37*). Using such study designs in conjunction with the detailed description of common human DNA variation presented here, it may be possible to identify a set of major genetic risk factors contributing to the variability in a complex disease and/or treatment response. Although knowledge of a single genetic risk factor can seldom be used to predict the treatment outcome of a common disease, knowledge of a large fraction of all the major genetic risk factors contributing to a treatment response or common disease could have immediate utility, allowing existing treatment options to be matched to individual patients without requiring additional knowledge of the mechanisms by which the genetic differences lead to different outcomes.

In our analyses, we selected representations of the data, including pairwise LD as well as a haplotype-based approach, that we felt would be most useful for an initial characterization of this resource. We focused attention on pairwise LD analyses because they provide a particularly simple framework for evaluating coverage and information content of different SNP collections. The optimal representation of genetic variation data remains an area of active research. Although we have determined example haplotype maps of the human genome in these three populations, the most appropriate representation of the data depends substantially on the specific questions to be answered. There will be many maps of human genetic variation, each tailored for specific uses.

Public data availability. We have implemented an instance of the Generic Genome Browser (*38*) at http://genome.perlegen.com for viewing the SNP, LD, and haplotype data reported here; this data will also be available from *Science* upon request. More detailed haplotype analysis results are available at http://research.calit2.net/hap/wgha/ and through dbSNP. The data reported here represent a massive increase in the available number of SNPs characterized in multiple populations. For comparison, although the public SNP database, dbSNP build 122, contained map positions for more than 8.1 million human SNPs, frequencies were available for only 797,000 of these SNPs, mostly in just one population, and genotypes were available for only 210,000 SNPs. Our data also complement the results of the International HapMap Project (*11*), by providing data for many more SNPs across fewer individuals.

This work enables detailed analyses of the structure of human genetic variation on a whole-genome scale. We examined genetic variation in individuals from three populations with substantially different histories and describe general features of variation within and between populations. Because these samples do not capture the full genetic diversity of the populations from which they were selected, our data are not suitable for answering many questions about the detailed genetic structure of human populations (*39*). However, the public availability of these data will enable a wide variety of additional analyses to be carried out by scientists investigating the structure of human genetic variation as well as the genetic basis of human phenotypic differences.

Table 5. Haplotype block partition results for the three populations.

Population	Blocks	Average size, kb*	Required SNPs†
African-American	235,663	8.8	570,886
European-American	109,913	20.7	275,960
Han Chinese	89,994	25.2	220,809

*Average distance spanned by segregating sites in each block. †Minimum number of SNPs required to distinguish common haplotype patterns with frequencies of 5% or higher.

References and Notes

1. L. Kruglyak, D. A. Nickerson, *Nature Genet.* **27**, 234 (2001).
2. C. Romualdi *et al.*, *Genome Res.* **12**, 602 (2002).
3. J. P. Hugot *et al.*, *Nature* **411**, 599 (2001).
4. A. D. Roses, *Neurogenetics* **1**, 3 (1997).
5. N. Patil *et al.*, *Science* **294**, 1719 (2001).
6. S. B. Gabriel *et al.*, *Science* **296**, 2225 (2002).
7. Materials and methods are available as supporting material on *Science* Online.
8. D. A. Hinds *et al.*, *Am. J. Hum. Genet.* **74**, 317 (2004).
9. D. A. Hinds *et al.*, *Hum. Genom.* **1**, 421 (2004).
10. L. Hosking *et al.*, *Eur. J. Hum. Genet.* **12**, 395 (2004).
11. International HapMap Consortium, *Nature* **426**, 789 (2003), available at www.hapmap.org.
12. International SNP Map Working Group, *Nature* **409**, 928 (2001).
13. A. M. Bowcock *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 839 (1991).
14. B. S. Weir, C. C. Cockerham, *Evolution* **38**, 1358 (1984).
15. E. J. Parra *et al.*, *Am. J. Hum. Genet.* **63**, 1839 (1998).
16. N. A. Rosenberg *et al.*, *Science* **298**, 2381 (2002).
17. C. S. Carlson *et al.*, *Nature Genet.* **33**, 518 (2003).
18. N. Patterson *et al.*, *Am. J. Hum. Genet.* **74**, 979 (2004).
19. J. M. Akey, G. Zhang, K. Zhang, L. Jin, M. D. Shriver, *Genome Res.* **12**, 1805 (2002).
20. Q. Xiao, H. Weiner, D. W. Crabb, *J. Clin. Invest.* **98**, 2027 (1996).
21. P. Duggal *et al.*, *Genes Immun.* **4**, 245 (2003).
22. K. Nakayama *et al.*, *J. Hum. Genet.* **47**, 92 (2002).
23. B. Devlin, N. Risch, *Genomics* **29**, 311 (1995).
24. C. S. Carlson *et al.*, *Am. J. Hum. Genet.* **74**, 106 (2004).
25. G. A. Huttley, M. W. Smith, M. Carrington, S. J. O'Brien, *Genetics* **152**, 1711 (1999).
26. P. C. Sabeti *et al.*, *Nature* **419**, 832 (2002).
27. A. M. Pittman *et al.*, *Hum. Mol. Genet.* **13**, 1267 (2004).
28. G. Van Gassen, W. Annaert, *Neuroscientist* **9**, 117 (2003).
29. E. R. De Kloet, *Ann. N.Y. Acad. Sci.* **1018**, 1 (2004).
30. E. Dawson *et al.*, *Nature* **418**, 544 (2002).
31. J. K. Pritchard, M. Przeworski, *Am. J. Hum. Genet.* **69**, 1 (2001).
32. SeattleSNPs, National Heart, Lung, and Blood Institute Program for Genomic Applications, University of Washington–Fred Hutchinson Cancer Research Center, Seattle, WA, available at http://pga.gs.washington.edu.
33. J. S. Bader, *Pharmacogenomics* **2**, 11 (2001).
34. E. Halperin, E. Eskin, *Bioinformatics* **20**, 1842 (2004).
35. D. Altshuler *et al.*, *Nature Genet.* **26**, 76 (2000).
36. L. A. Pennacchio *et al.*, *Science* **294**, 169 (2001).
37. N. J. Risch, *Nature* **405**, 847 (2000).
38. L. D. Stein *et al.*, *Genome Res.* **12**, 1599 (2002).
39. D. Serre, S. Pääbo, *Genome Res.* **14**, 1679 (2004).

40. We thank B. Margus and S. Fodor for many helpful discussions, S. Ptak for comments on the manuscript, and the following individuals for expert technical assistance: high throughput genotyping, C. Chen, P. Chu, D. Dalija, J. Doshi, P. Jain, A. Johnson, L. Kamigaki, J. Karbowski, C. Kautzer, V. Mendoza, M. Morenzoni, B. Nguyen, C. Owyang, N. Patil, K. Perry, R. Patel, C. Pethiyagoda, T. L. Pham, C. Sanders, A. Sparks, R. Stokowski, D. Telman, R. Vergara, P. Vu, and P.-H. Wang; bioinformatics design, W. Barrett, H. Huang, M. Jen, X. Li, B. Mooney, and S. Pitts; data analysis, A. Berno, K. Konvicka, A. Ollmann, K. Pant, and J. Sheehan; laboratory information management, R. Gupta, E. Jacobs, C. Radu, and P. Starink; engineering and instrumentation, R. Hartlage, M. Norris, G. Park, and A. Yee; computer systems and operations, T. Fleury, R. Galvez, R. Gordon, P. Hickey, C. LaPlante, J. Nordhal, T. Ogi, and J. VandenHengel. E.E. is supported by the California Institute for Telecommunications and Information Technology.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5712/1072/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S5
References and Notes

20 September 2004; accepted 14 January 2005
10.1126/science.1105436