A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Despite the ever-accelerating pace of biomedical research, the root causes of common human diseases remain largely unknown, preventative measures are generally inadequate, and available treatments are seldom curative. Family history is one of the strongest risk factors for nearly all diseases—including cardiovascular disease, cancer, diabetes, autoimmunity, psychiatric illnesses and many others—providing the tantalizing but elusive clue that inherited genetic variation has an important role in the pathogenesis of disease. Identifying the causal genes and variants would represent an important step in the path towards improved prevention, diagnosis and treatment of disease.

More than a thousand genes for rare, highly heritable ‘mendelian’ disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease. Common disorders, in contrast, have proven much more challenging to study, as they are thought to be due to the combined effect of many different susceptibility DNA variants interacting with environmental factors.

Studies of common diseases have fallen into two broad categories: family-based linkage studies across the entire genome, and population-based association studies of individual candidate genes. Although there have been notable successes, progress has been slow due to the inherent limitations of the methods; linkage analysis has low power except when a single locus explains a substantial fraction of disease, and association studies of one or a few candidate genes examine only a small fraction of the ‘universe’ of sequence variation in each patient.

A comprehensive search for genetic influences on disease would involve examining all genetic differences in a large number of affected individuals and controls. It may eventually become possible to accomplish this by complete genome resequencing. In the meantime, it is increasingly practical to systematically test common genetic variants for their role in disease; such variants explain much of the genetic diversity in our species, a consequence of the historically small size and shared ancestry of the human population.

Recent experience bears out the hypothesis that common variants have an important role in disease, with a partial list of validated examples including *HLA* (autoimmunity and infection)*1*, *APOE4* (Alzheimer’s disease, lipids)*2*, Factor V*Leiden* (deep vein thrombosis)*3*, *PPARG* (encoding PPARγ; type 2 diabetes)*4,5*, *KCNJ11* (type 2 diabetes)*6*, *PTPN22* (rheumatoid arthritis and type 1 diabetes)*7,8*, insulin (type 1 diabetes)*9*, *CTLA4* (autoimmune thyroid disease, type 1 diabetes)*10*, *NOD2* (inflammatory bowel disease)*11,12*, complement factor H (age-related macular degeneration)*13,14* and *RET* (Hirschsprung disease)*16,17*, among many others.

Systematic studies of common genetic variants are facilitated by the fact that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as linkage disequilibrium (LD); a particular combination of alleles along a chromosome is termed a haplotype.

LD exists because of the shared ancestry of contemporary chromosomes. When a new causal variant arises through mutation—whether a single nucleotide change, insertion/deletion, or structural alteration—it is initially tethered to a unique chromosome on which it occurred, marked by a distinct combination of genetic variants. Recombination and mutation subsequently act to erode this association, but do so slowly (each occurring at an average rate of about 10−8 per base pair (bp) per generation) as compared to the number of generations (typically 103 to 105) since the mutational event.

The correlations between causal mutations and the haplotypes on which they arose have long served as a tool for human genetic research: first finding association to a haplotype, and then subsequently identifying the causal mutation(s) that it carries. This was pioneered in studies of the *HLA* region, extended to identify causal genes for mendelian diseases (for example, cystic fibrosis*18* and diastrophic dysplasia*19*), and most recently for complex disorders such as age-related macular degeneration*12,15*.

Early information documented the existence of LD in the human genome*10,22*; however, these studies were limited (for technical reasons) to a small number of regions with incomplete data, and general patterns were challenging to discern. With the sequencing of the human genome and development of high-throughput genomic methods, it became clear that the human genome generally displays more LD*23* than under simple population genetic models*24*, and that LD is more varied across regions, and more segmentally structured*25−30*, than had previously been supposed. These observations indicated that LD-based methods would generally have great value (because nearby SNPs were typically correlated with many of their neighbours), and also that LD relationships would

*Lists of participants and affiliations appear at the end of the paper.*
need to be empirically determined across the genome by studying polymorphisms at high density in population samples.

The International HapMap Project was launched in October 2002 to create a public, genome-wide database of common human sequence variation, providing information needed as a guide to genetic studies of clinical phenotypes. The project had become practical by the confluence of the following: (1) the availability of the human genome sequence; (2) databases of common SNPs (subsequently enriched by this project) from which genotyping assays could be designed; (3) insights into human LD; (4) development of inexpensive, accurate technologies for high-throughput SNP genotyping; (5) web-based tools for storing and sharing data; and (6) frameworks to address associated ethical and cultural issues. The project follows the data release principles of an international community resource project (http://www.wellcome.ac.uk/doc_WTD003208.html), sharing information rapidly and without restriction on its use.

The HapMap data were generated with the primary aim of guiding the design and analysis of medical genetic studies. In addition, the advent of genome-wide variation resources such as the HapMap opens a new era in population genetics, offering an unprecedented opportunity to investigate the evolutionary forces that have shaped variation in natural populations.

The Phase I HapMap

Phase I of the HapMap Project set as a goal genotyping at least one common SNP every 5 kilobases (kb) across the genome in each of 269 DNA samples. For the sake of practicability, and motivated by the allele frequency distribution of variants in the human genome, a minor allele frequency (MAF) of 0.05 or greater was targeted for study. (For simplicity, in this paper we will use the term ‘common’ to mean a SNP with MAF ≥ 0.05.) The phase has a Phase II, which is attempting genotyping of an additional 4.6 million SNPs in each of the HapMap samples.

To compare the genome-wide resource to a more complete database of common variation—one in which all common SNPs and many rarer ones have been discovered and tested—a representative collection of ten regions, each 500 kb in length, was selected from the ENCODE (Encyclopedia of DNA Elements) Project. Each 500-kb region was sequenced in 48 individuals, and all SNPs in these regions (discovered or in dbSNP) were genotyped in the complete set of 269 DNA samples.

The specific samples examined are: (1) 90 individuals (30 parent–offspring trios) from the Yoruba in Ibadan, Nigeria (abbreviation YRI); (2) 90 individuals (30 trios) from Utah, USA, from the Centre d’Etude du Polymporphisme Humain collection (abbreviation CEU); (3) 45 Han Chinese in Beijing, China (abbreviation CHB); (4) 44 Japanese in Tokyo, Japan (abbreviation JPT).

Because none of the samples was collected to be representative of a larger population such as ‘Yoruba’, ‘Northern and Western European’, ‘Han Chinese’, or ‘Japanese’ (let alone of all populations from ‘Africa’, ‘Europe’, or ‘Asia’), we recommend using a specific local identifier (for example, ‘Yoruba in Ibadan, Nigeria’) to describe the samples initially. Because the CHB and JPT allele frequencies are generally very similar, some analyses below combine these data sets. When doing so, we refer to these ‘analysis panels’ (YRI, CEU, CHB+JPT) to avoid confusing this analytical approach with the concept of a ‘population’.

Important details about the design of the HapMap Project are presented in the Methods, including: (1) organization of the project; (2) selection of DNA samples for study; (3) increasing the number and annotation of SNPs in the public SNP map (dbSNP) from 2.6 million to 9.2 million (Fig. 1); (4) targeted sequencing of the ten ENCODE regions, including evaluations of false-positive and false-negative rates; (5) genotyping for the genome-wide map; (6) intense efforts that monitored and established the high quality of the data; and (7) data coordination and distribution through the project Data Coordination Center (DCC) (http://www.hapmap.org).

Description of the data. The Phase I HapMap contains 1,007,329 SNPs that passed a set of quality control (QC) filters (see Methods) in each of the three analysis panels, and are polymorphic across the 269 samples. SNP genotyping was distributed across centres by chromosomal region, with several technologies employed (Table 1). Each centre followed the same standard rules for SNP selection, quality control data and release; all SNPs were genotyped in the full set of 269 samples. Some centres genotyped more SNPs than required by the rules.

Extensive, blinded quality assessment (QA) exercises documented that these data are highly accurate (99.7%) and complete (99.3%, see

<table>
<thead>
<tr>
<th>ARTICLES</th>
<th>NATURE</th>
<th>Vol 437</th>
<th>27 October 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Genotyping centres</td>
<td>Chromosomes</td>
<td>Technology</td>
</tr>
<tr>
<td>Centre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIKEN</td>
<td>5, 11, 14, 15, 16, 17, 19</td>
<td>Third Wave Invader</td>
<td></td>
</tr>
<tr>
<td>Wellcome Trust Sanger Institute</td>
<td>1, 6, 10, 13, 20</td>
<td>Illumina BeadArray</td>
<td></td>
</tr>
<tr>
<td>McGill University and Génomique Québec Innovation Centre</td>
<td>2, 4p</td>
<td>Illumina BeadArray</td>
<td></td>
</tr>
<tr>
<td>Chinese HapMap Consortium*</td>
<td>3, 8p, 21</td>
<td>Sequenom MassExtend, Illumina BeadArray</td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td>8q, 9, 18q, 22, X</td>
<td>Illumina BeadArray</td>
<td></td>
</tr>
<tr>
<td>Broad Institute of Harvard and MIT</td>
<td>4q, 7q, 18p, Y, mtDNA</td>
<td>Sequenom MassExtend, Illumina BeadArray</td>
<td></td>
</tr>
<tr>
<td>Baylor College of Medicine with ParAllele BioScience</td>
<td>7p</td>
<td>ParAllele MIP</td>
<td></td>
</tr>
<tr>
<td>University of California, San Francisco, with Washington University in St Louis</td>
<td>5Mb (ENCODE) on 2, 4, 7, 8, 9, 12, 18 in CEU</td>
<td>PerkinElmer AcycloPrime-FP</td>
<td></td>
</tr>
<tr>
<td>Perlegen Sciences</td>
<td>High-density oligonucleotide array</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The Chinese HapMap Consortium consists of the Beijing Genomics Institute, the Chinese National Human Genome Center at Beijing, the University of Hong Kong, the Hong Kong University of Science and Technology, the Chinese University of Hong Kong, and the Chinese National Human Genome Center at Shanghai.

Figure 1 | Number of SNPs in dbSNP over time. The cumulative number of non-redundant SNPs (each mapped to a single location in the genome) is shown as a solid line, as well as the number of SNPs validated by genotyping (dotted line) and double-hit status (dashed line). Years are divided into quarters (Q1–Q4).
also Supplementary Table 1). All genotyping centres produced high-quality data (accuracy more than 99% in the blind QA exercises, Supplementary Tables 2 and 3), and missing data were not biased against heterozygotes. The Supplementary Information contains the full details of these efforts.

Although SNP selection was generally agnostic to functional annotation, 11,500 non-synonymous cSNPs (cSNPs in coding regions of genes where the different SNP alleles code for different amino acids in the protein) were successfully typed in Phase I. (An effort was made to prioritize cSNPs in Phase I in choosing SNPs for each 5-kb region; all known non-synonymous cSNPs were attempted as part of Phase II.)

Across the ten ENCODE regions (Table 2), the density of SNPs was approximately tenfold higher as compared to the genome-wide map: 17,944 SNPs across the 5 megabases (Mb) (one per 279 bp).

More than 1.3 million SNP genotyping assays were attempted (Table 3) to generate the Phase I data on more than 1 million SNPs. The 0.3 million SNPs not part of the Phase I data set include 73,652 that passed QC filters but were monomorphic in all 269 samples. The remaining SNPs failed the QC filters in one or more analysis panels mostly because of inadequate completeness, non-mendelian inheritance, deviations from Hardy–Weinberg equilibrium, discrepant genotypes among duplicates, and data transmission discrepancies.

SNPs on the Phase I map are evenly spaced, except on Y and mtDNA. The Phase I data include a successful, common SNP every 5 kb across most of the genome in each analysis panel (Supplementary Fig. 1): only 3.3% of inter-SNP distances are longer than 10 kb, spanning 11.9% of the genome (Fig. 2; see also Supplementary Fig. 2). One exception is the X chromosome (Supplementary Fig. 1), where a much higher proportion of attempted SNPs were rare or monomorphic, and thus the density of common SNPs is lower.

Two intentional exceptions to the regular spacing of SNPs on the physical map were the mitochondrial chromosome (mtDNA), which does not undergo recombination, and the non-recombining portion of chromosome Y. On the basis of the 168 successful, polymorphic SNPs, each HapMap sample fell into one of 15 (of the 18 known) mtDNA haplogroups‡‡ (Table 4). A total of 84 SNPs that characterize the unique branches of the reference Y genealogical tree§§–§§ were genotyped on the HapMap samples. These SNPs assigned each Y chromosome to 8 (of the 18 major) Y haplogroups previously described (Table 4).

Highly accurate phasing of long-range chromosomal haplotypes. Despite having collected data in diploid individuals, the inclusion of parent–offspring trios and the use of computational methods made it possible to determine long-range phased haplotypes of extremely high quality for each individual. These computational algorithms take advantage of the observation that because of LD, relatively few of the large number of possible haplotypes consistent with the genotype data actually occur in population samples.

The project compared a variety of algorithms for phasing haplotypes from unrelated individuals and trios††, and applied the algorithm that proved most accurate (an updated version of PHASE§§)

Table 2: ENCODE project regions and genotyping

<table>
<thead>
<tr>
<th>Region name</th>
<th>Chromosome band</th>
<th>Genomic interval (NCBI) (base numbers)</th>
<th>Gene density (%)</th>
<th>Conservation score (%)</th>
<th>Pedigree-based recombination rate (cM/Mb)</th>
<th>Population-based recombination rate (cM/Mb)</th>
<th>G+C content#</th>
<th>Available SNPs</th>
<th>Successfully genotyped SNPs‡</th>
<th>Sequencing centre/genotyping centre(§)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENr12</td>
<td>2p16.3</td>
<td>51,633,239–52,133,238</td>
<td>0</td>
<td>3.8</td>
<td>0.8</td>
<td>0.9</td>
<td>0.35</td>
<td>1,570</td>
<td>1,762</td>
<td>3,332</td>
<td>2,275</td>
</tr>
<tr>
<td>ENr13</td>
<td>2q37.1</td>
<td>234,778,639–235,278,638</td>
<td>4.6</td>
<td>1.3</td>
<td>2.2</td>
<td>2.5</td>
<td>0.43</td>
<td>1,736</td>
<td>1,259</td>
<td>2,995</td>
<td>1,910</td>
</tr>
<tr>
<td>ENr13</td>
<td>4q26</td>
<td>118,705,475–119,205,474</td>
<td>0</td>
<td>3.9</td>
<td>0.6</td>
<td>0.9</td>
<td>0.35</td>
<td>1,444</td>
<td>2,053</td>
<td>3,497</td>
<td>2,201</td>
</tr>
<tr>
<td>ENm010</td>
<td>7p15.2</td>
<td>26,699,793–27,999,792</td>
<td>5.0</td>
<td>22.0</td>
<td>0.9</td>
<td>0.9</td>
<td>0.44</td>
<td>1,220</td>
<td>1,795</td>
<td>3,015</td>
<td>1,271</td>
</tr>
<tr>
<td>ENm013*</td>
<td>7q21.13</td>
<td>89,395,718–89,859,717</td>
<td>5.5</td>
<td>4.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.38</td>
<td>1,394</td>
<td>1,917</td>
<td>3,311</td>
<td>1,807</td>
</tr>
<tr>
<td>ENm014*</td>
<td>7q31.33</td>
<td>126,135,436–126,652,577</td>
<td>2.9</td>
<td>11.2</td>
<td>0.4</td>
<td>0.9</td>
<td>0.39</td>
<td>1,320</td>
<td>1,664</td>
<td>2,984</td>
<td>1,966</td>
</tr>
<tr>
<td>ENr321</td>
<td>8q24.11</td>
<td>118,769,628–119,269,627</td>
<td>3.2</td>
<td>11.4</td>
<td>0.6</td>
<td>1.1</td>
<td>0.41</td>
<td>1,430</td>
<td>1,508</td>
<td>2,938</td>
<td>1,758</td>
</tr>
<tr>
<td>ENr232</td>
<td>9q34.11</td>
<td>127,061,347–127,561,346</td>
<td>5.9</td>
<td>8.3</td>
<td>2.7</td>
<td>2.6</td>
<td>0.52</td>
<td>1,444</td>
<td>1,523</td>
<td>2,967</td>
<td>1,324</td>
</tr>
<tr>
<td>ENr232</td>
<td>12q12</td>
<td>38,626,477–39,126,476</td>
<td>3.1</td>
<td>1.7</td>
<td>0.3</td>
<td>0.8</td>
<td>0.36</td>
<td>1,877</td>
<td>1,379</td>
<td>3,256</td>
<td>1,792</td>
</tr>
<tr>
<td>ENr235</td>
<td>18q12.1</td>
<td>23,717,221–24,217,220</td>
<td>0.9</td>
<td>7.4</td>
<td>1.2</td>
<td>0.9</td>
<td>0.37</td>
<td>1,330</td>
<td>1,459</td>
<td>2,789</td>
<td>1,640</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14,765</td>
<td>16,319</td>
<td>31,084</td>
<td>17,944</td>
</tr>
</tbody>
</table>

McGill-GGIC, McGill University and Genome Quebec Innovation Centre.

*These regions were truncated to 500 kb for resequencing.
††Sequence build 34 coordinates.
§Gene density is defined as the percentage of bases covered either by Ensembl genes or human mRNA best BLAT alignments in the UCSC Genome Browser database.
∥Non-exonic conservation with mouse sequence was measured by taking 125 base non-overlapping sub-windows inside the 500,000 base windows. Sub-windows with less than 75% of their bases in a mouse alignment were discarded. Of the remaining sub-windows, those with at least 80% base identity were used to calculate the conservation score. The mouse alignments in regions corresponding to the following were discarded: Ensembl genes, all GenBank mRNA Blast alignments, FGenesh++ gene predictions, Twinscan gene predictions, spliced EST alignments, and repeats.
¶Recombination rate based on estimates from LDhat§.§
#G+C content calculated from the sequence of the stated coordinates from sequence build 34.
‡SNPs in dbSNP build 121 at the time the ENCODE resequencing began and SNPs added to dbSNP in builds 122–125 independent of the resequencing.
‡‡SNPs successfully genotyped in all analysis panels (YRI, CEU, CHB + JPT).
§Perlegen genotyped a subset of SNPs in the CEU samples.

© 2005 Nature Publishing Group
separately to each analysis panel. (Phased haplotypes are available for
download at the Project website.) We estimate that ‘switch’ errors—
where a segment of the maternal haplotype is incorrectly joined to the
paternal—occur extraordinarily rarely in the trio samples (every 8 Mb
in CEU; 3.6 Mb in YRI). The switch rate is higher in the CHB+JPT
samples (one per 0.34 Mb) due to the lack of information from
parent–offspring trios, but even for the unrelated individuals, statis-
tical reconstruction of haplotypes is remarkably accurate.

**Table 3 | HapMap Phase I genotyping success measures**

<table>
<thead>
<tr>
<th>SNP categories</th>
<th>YRI</th>
<th>CEU</th>
<th>CHB + JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assays submitted</td>
<td>1,273,716</td>
<td>1,302,849</td>
<td>1,273,703</td>
</tr>
<tr>
<td>Passed QC filters</td>
<td>1,123,296 (88%)</td>
<td>1,157,650 (89%)</td>
<td>1,134,726 (89%)</td>
</tr>
<tr>
<td>Did not pass QC filters*</td>
<td>150,420 (12%)</td>
<td>145,199 (11%)</td>
<td>138,977 (11%)</td>
</tr>
<tr>
<td>&gt; 2% missing data</td>
<td>98,116 (65%)</td>
<td>107,626 (74%)</td>
<td>93,710 (67%)</td>
</tr>
<tr>
<td>&gt; 1 duplicate inconsistent</td>
<td>7,575 (5%)</td>
<td>6,254 (4%)</td>
<td>10,725 (8%)</td>
</tr>
<tr>
<td>&gt; 1 mendelian error</td>
<td>22,815 (15%)</td>
<td>13,600 (9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>&lt; 0.001 Hardy-Weinberg P-value</td>
<td>12,052 (8%)</td>
<td>9,721 (7%)</td>
<td>16,176 (12%)</td>
</tr>
<tr>
<td>Other failures†</td>
<td>23,478 (16%)</td>
<td>17,692 (12%)</td>
<td>23,722 (17%)</td>
</tr>
<tr>
<td>Non-redundant (unique) SNPs</td>
<td>1,076,392</td>
<td>1,104,980</td>
<td>1,087,305</td>
</tr>
<tr>
<td>Monomorphic</td>
<td>156,290 (15%)</td>
<td>234,482 (21%)</td>
<td>268,325 (25%)</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>920,102 (85%)</td>
<td>870,498 (79%)</td>
<td>818,980 (75%)</td>
</tr>
</tbody>
</table>

| All analysis panels                   |              |              |               |
| Unique QC-passed SNPs                 | 1,156,772    |              |               |
| Passed in one analysis panel          | 52,204 (5%)  |              |               |
| Passed in two analysis panels         | 97,231 (8%)  |              |               |
| Passed in three analysis panels       | 1,007,337 (87%) |              |               |
| Monomorphic across three analysis panels | 75,997      |              |               |
| Polymorphic in all three analysis panels | 682,397    |              |               |
| MAF ≤ 0.05 in at least one of three analysis panels | 877,351    |              |               |

*Out of 95 samples in CEU, YRI: 94 samples in CHB + JPT.
† ‘Other failures’ includes SNPs with discrepancies during the data transmission process. Some SNPs failed in more than one way, so these percentages add up to more than 100%.

**Figure 2 | Distribution of inter-SNP distances.** The distributions are shown for each analysis panel for the HapMappable genome (defined in the Methods), for all common SNPs (with MAF ≥ 0.05).
estimated from the ENCODE data, where deep sequencing reduces bias due to SNP ascertainment. Consistent with previous studies, most SNPs observed in the ENCODE regions are rare: 46% had MAF < 0.05, and 9% were seen in only a single individual (Fig. 4). Although most varying sites in the population are rare, most heterozygous sites within any individual are due to common SNPs. Specifically, in the ENCODE data, 90% of heterozygous sites in each individual were due to common variants (Fig. 4). With ever-deeper sequencing of DNA samples the number of rare variants will rise linearly, but the vast majority of heterozygous sites in each person will be explained by a limited set of common SNPs now contained (or captured through LD) in existing databases (Fig. 5).

Consistent with previous descriptions, the CEU, CHB and JPT samples show fewer low frequency alleles when compared to the YRI samples (Fig. 5), a pattern thought to be due to bottlenecks in the history of the non-YRI populations.

In contrast to the ENCODE data, the distribution of allele frequencies for the genome-wide data is flat (Fig. 5), with much more similarity in the distributions observed in the three analysis panels. These patterns are well explained by the inherent and intentional bias in the rules used for SNP selection: we prioritized using validated SNPs in order to focus resources on common (rather than rare or false positive) candidate SNPs from the public databases. For a fuller discussion of ascertainment issues, including a shift in frequencies over time and an excess of high-frequency derived alleles due to inclusion of chimpanzee data in determination of double-hit status, see the Supplementary Information (Supplementary Fig. 3).

SNP allele frequencies across population samples. Of the 1.007 million SNPs successfully genotyped and polymorphic across the three analysis panels, only a subset were polymorphic in any given panel: 85% in YRI, 79% in CEU, and 75% in CHB+JPT. The joint distribution of frequencies across populations is presented in Fig. 6 (for the ENCODE data) and Supplementary Fig. 4 (for the genome-wide map). We note the similarity of allele frequencies in the CHB and JPT samples, which motivates analysing them jointly as a single analysis panel in the remainder of this report.

A simple measure of population differentiation is Wright’s $F_{ST}$ which measures the fraction of total genetic variation due to between-population differences$^{45}$. Across the autosomes, $F_{ST}$ estimated from the full set of Phase I data is 0.12, with CEU and CHB+JPT showing the lowest level of differentiation ($F_{ST} = 0.07$), and YRI and CHB+JPT the highest ($F_{ST} = 0.12$). These values are slightly higher than previous reports$^{46}$, but differences in the types of variants (SNPs versus microsatellites) and the samples studied make comparisons difficult.

As expected, we observed very few fixed differences (that is, cases in which alternate alleles are seen exclusively in different analysis panels). Across the 1 million SNPs genotyped, only 11 have fixed differences between CEU and YRI, 21 between CEU and CHB+JPT, and 5 between YRI and CHB+JPT, for the autosomes.

The extent of differentiation is similar across the autosomes, but higher on the X chromosome ($F_{ST} = 0.21$). Interestingly, 123 SNPs on the X chromosome were completely differentiated between YRI and CHB+JPT, but only two between CEU and YRI and one between CEU and CHB+JPT. This seems to be largely due to a single region near the centromere, possibly indicating a history of natural

---

**Table 4 | mtDNA and Y chromosome haplogroups**

<table>
<thead>
<tr>
<th>mtDNA haplogroup</th>
<th>DNA sample*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YRI (60)</td>
</tr>
<tr>
<td>L1</td>
<td>0.22</td>
</tr>
<tr>
<td>L2</td>
<td>0.35</td>
</tr>
<tr>
<td>L3</td>
<td>0.43</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
</tr>
<tr>
<td>M/E</td>
<td>-</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
</tr>
<tr>
<td>V</td>
<td>-</td>
</tr>
<tr>
<td>J</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
</tr>
<tr>
<td>U</td>
<td>-</td>
</tr>
<tr>
<td>W</td>
<td>-</td>
</tr>
</tbody>
</table>

*Number of chromosomes sampled is given in parentheses.

---

**Figure 3 | Allele frequency and completeness of dbSNP for the ENCODE regions. a-c.** The fraction of SNPs in dbSNP, or with a proxy in dbSNP, are shown as a function of minor allele frequency for each analysis panel (a, YRI; b, CEU; c, CHB+JPT). Singletons refer to heterozygotes observed in a single individual, and are broken out from other SNPs with MAF < 0.05. Because all ENCODE SNPs have been deposited in dbSNP, for this figure we define a SNP as “in dbSNP” if it would be in dbSNP build 125 independent of the HapMap ENCODE resequencing project. All remaining SNPs (not in dbSNP) were discovered only by ENCODE resequencing; they are categorized by their correlation ($r$) to those in dbSNP. Note that the number of SNPs in each frequency bin differs among analysis panels, because not all SNPs are polymorphic in all analysis panels.
Figure 4 | Minor allele frequency distribution of SNPs in the ENCODE data, and their contribution to heterozygosity. This figure shows the polymorphic SNPs from the HapMap ENCODE regions according to minor allele frequency (blue), with the lowest minor allele frequency bin (<0.05) separated into singletons (SNPs heterozygous in one individual only, shown in grey) and SNPs with more than one heterozygous individual. For this analysis, MAF is averaged across the analysis panels. The sum of the contribution of each MAF bin to the overall heterozygosity of the ENCODE regions is also shown (orange).

Properties of LD in the human genome

Traditionally, descriptions of LD have focused on measures calculated between pairs of SNPs, averaged as a function of physical distance. Examples of such analyses for the HapMap data are presented in Supplementary Fig. 6. After adjusting for known confounders such as sample size, allele frequency distribution, marker density, and length of sampled regions, these data are highly similar to previously published surveys.

Because LD varies markedly on scales of 1–100 kb, and is often discontinuous rather than declining smoothly with distance, averages obscure important aspects of LD structure. A fuller exploration of the fine-scale structure of LD offers both insight into the causes of LD and understanding of its application to disease research.

LD patterns are simple in the absence of recombination. The most natural path to understanding LD structure is first to consider the simplest case in which there is no recombination (or gene conversion), and then to add recombination to the model. (For simplicity we ignore genotyping error and recurrent mutation in this discussion, both of which seem to be rare in these data.)

In the absence of recombination, diversity arises solely through mutation. Because each SNP arose on a particular branch of the genealogical tree relating the chromosomes in the current populations, multiple haplotypes are observed. SNPs that arose on the same branch of the genealogy are perfectly correlated in the sample, whereas SNPs that occurred on different branches have imperfect correlations, or no correlation at all.

We illustrate these concepts using empirical genotype data from 36 adjacent SNPs in an ENCODE region (ENr131.2q37), selected because no obligate recombination events were detectable among them in CEU (Fig. 7). (We note that the lack of obligate recombination events in a small sample does not guarantee that no recombinants have occurred, but it provides a good approximation for illustration.)

In principle, 36 such SNPs could give rise to $2^{36}$ different haplotypes. Even with no recombination, gene conversion or recurrent mutation, up to 37 different haplotypes could be formed. Despite this great potential diversity, only seven haplotypes are observed (five seen more than once) among the 120 parental CEU chromosomes studied, reflecting shared ancestry since their most recent common ancestor among apparently unrelated individuals.

In such a setting, it is easy to interpret the two most common pairwise measures of LD: $D^*$ and $r^2$. (See the Supplementary Information for fuller definitions of these measures.) $D^*$ is defined to be 1 in the absence of obligate recombination, declining only due to recombination or recurrent mutation. In contrast, $r^2$ is simply

Figure 5 | Allele frequency distributions for autosomal SNPs. For each analysis panel we plotted (bars) the MAF distribution of all the Phase I SNPs with a frequency greater than zero. The solid line shows the MAF distribution for the ENCODE SNPs, and the dashed line shows the MAF distribution expected for the standard neutral population model with constant population size and random mating without ascertainment bias.

selection at this locus (see below; M. L. Freedman et al., personal communication).

Haplotype sharing across populations. We next examined the extent to which haplotypes are shared across populations. We used a hidden Markov model in which each haplotype is modelled in turn as an imperfect mosaic of other haplotypes (see Supplementary Information). In essence, the method infers probabilistically which other haplotype in the sample is the closest relative (nearest neighbour) at each position along the chromosome.

Unsurprisingly, the nearest neighbour most often is from the same analysis panel, but about 10% of haplotypes were found most closely to match a haplotype in another panel (Supplementary Fig. 5). All individuals have at least some segments over which the nearest neighbour is in a different analysis panel. These results indicate that although analysis panels are characterized both by different haplotype frequencies and, to some extent, different combinations of alleles, both common and rare haplotypes are often shared across populations.
The squared correlation coefficient between the two SNPs. Thus, $r^2$ is 1 when two SNPs arose on the same branch of the genealogy and remain undisrupted by recombination, but has a value less than 1 when SNPs arose on different branches, or if an initially strong correlation has been disrupted by crossing over.

In this region, $D' = 1$ for all marker pairs, as there is no evidence of historical recombination. In contrast, and despite great simplicity of haplotype structure, $r^2$ values display a complex pattern, varying from 0.0003 to 1.0, with no relationship to physical distance. This makes sense, however, because without recombination, correlations among SNPs depend on the historical order in which they arose, not the physical order of SNPs on the chromosome.

Most importantly, the seeming complexity of $r^2$ values can be deconvolved; in a sample made up of only seven different SNP configurations exist in this region, with all but two chromosomes matching five common haplotypes, which can be distinguished from each other by typing a specific set of four SNPs. That is, only a small minority of sites need be examined to capture fully the information in this region.

**Variation in local recombination rates is a major determinant of LD.** Recombination in the ancestors of the current population has typically disrupted the simple picture presented above. In the human genome, as in yeast44, mouse45 and other genomes, recombination rates typically vary dramatically on a fine scale, with hotspots of recombination explaining much crossing over in each region46. The generality of this model has recently been demonstrated through computational methods that allow estimation of recombination rates (including hotspots and coldspots) from genotype data46,47.

The availability of nearly complete information about common DNA variation in the ENCODE regions allowed a more precise estimation of recombination rates across large regions than in any previous study. We estimated recombination rates and identified recombination hotspots in the ENCODE data, using methods previously described46 (see Supplementary Information for details). Hotspots are short regions (typically spanning about 2 kb) over which recombination rates rise dramatically over local background rates.

Whereas the average recombination rate over 500 kb across the human genome is about 0.5 cM48, the estimated recombination rate across the 500-kb ENCODE regions varied nearly tenfold, from a minimum of 0.19 cM (ENm013.7q21.13) to a maximum of 1.25 cM (EN232.9q34.11). Even this tenfold variation obscures much more dramatic variation over a finer scale: 88 hotspots of recombination were identified (Fig. 8; see also Supplementary Fig. 7)—that is, one per 57 kb—with hotspots detected in each of the ten regions (from 4 in 12q12 to 14 in 2q37.1). Across the 5 Mb, we estimate that about 80% of all recombination has taken place in about 15% of the sequence (Fig. 9, see also refs 46, 49).

**A block-like structure of human LD.** With most human recombination occurring in recombination hotspots, the breakdown of LD is often discontinuous. A ‘block-like’ structure of LD is visually apparent in Fig. 8 and Supplementary Fig. 7: segments of consistently high $D'$ that break down where high recombination rates, recombination hotspots and obligate recombination events40 all cluster.

When haplotype blocks are more formally defined in the ENCODE data (using a method based on a composite of local $D'$

---

**Figure 6** Comparison of allele frequencies in the ENCODE data for all pairs of analysis panels and between the CHB and JPT sample sets. For each polymorphic SNP in a sample we identified the minor allele across all panels (a–d) and then calculated the frequency of this allele in each analysis panel/sample set. The colour in each bin represents the number of SNPs that display each given set of allele frequencies. The purple regions show that very few SNPs are common in one panel but rare in another. The red regions show that there are many SNPs that have similar low frequencies in each pair of analysis panels/sample sets.
values, or another based on the four gamete test, most of the sequence falls into long segments of strong LD that contain many SNPs and yet display limited haplotype diversity (Table 5).

Specifically, addressing concerns that blocks might be an artefact of low marker density, in these nearly complete data most of the sequence falls into blocks of four or more SNPs (67% in YRI to 87% in CEU) and the average sizes of such blocks are similar to initial estimates. Although the average block spans many SNPs (30–70), the average number of common haplotypes in each block ranged only from 4.0 (CHB + JPT) to 5.6 (YRI), with nearly all haplotypes in each block matching one of these few common haplotypes. These results confirm the generality of inferences drawn from disease-mapping studies and genomic surveys with smaller sample sizes and less complete data.

Long-range haplotypes and local patterns of recombination. Although haplotypes often break at recombination hotspots (and block boundaries), this tendency is not invariant. We identified all unique haplotypes with frequency more than 0.05 across the 269 individuals in the phased data, and compared them to the fine-scale recombination map. Figure 10 shows a region of chromosome 19 over which many such haplotypes break at identified recombination hotspots, but others continue. Thus, the tendency towards localization of recombination sites does not imply that all haplotypes break at each recombination site.

Some regions display remarkably extended haplotype structure based on a lack of recombination (Supplementary Fig. 8a, b). Most striking, if unsurprising, are centromeric regions, which lack recombination: haplotypes defined by more than 100 SNPs span several megabases across the centromeres. The X chromosome has multiple regions with very extensive haplotypes, whereas other chromosomes typically have a few such domains.

Most global measures of LD become more consistent when measured in genetic rather than physical distance. For example, when plotted against physical distance, the extent of pairwise LD
varies by chromosome; when plotted against average recombination rate on each chromosome (estimated from pedigree-based genetic maps) these differences largely disappear (Supplementary Fig. 6). Similarly, the distribution of haplotype length across chromosomes is less variable when measured in genetic rather than physical distance. For example, the median length of haplotypes is 54.4 kb on chromosome 1 compared to 34.8 kb on chromosome 21. When measured in genetic distance, however, haplotype length is much more similar: 0.104 cM on chromosome 1 compared to 0.111 cM on chromosome 21 (Supplementary Fig. 9).

The exception is again the X chromosome, which has more extensive haplotype structure after accounting for recombination rate (median haplotype length = 0.135 cM). Multiple factors could explain different patterns on the X chromosome: lower SNP density, smaller sample size, restriction of recombination to females and lower effective population size.

A view of LD focused on the putative causal SNP
Although genealogy and recombination provide insight into why nearby SNPs are often correlated, it is the redundancies among SNPs that are of central importance for the design and analysis of association studies. A truly comprehensive genetic association study must consider all putative causal alleles and test each for its potential role in disease. If a causal variant is not directly tested in the disease sample, its effect can nonetheless be indirectly tested if it is correlated with a SNP or haplotype that has been directly tested.

Figure 8 | Comparison of linkage disequilibrium and recombination for two ENCODE regions. For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB+JPT analysis panels are shown: white, $D' < 1$ and LOD $< 2$; blue, $D' = 1$ and LOD $< 2$; pink, $D' < 1$ and LOD $\geq 2$; red, $D' = 1$ and LOD $\geq 2$. Below each of these plots is shown the intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles.
correlation with one or more others. When two variants are perfectly correlated, testing one is exactly equivalent to testing the other; we refer to such collections of SNPs (with pairwise $r^2 = 1.0$ in the HapMap samples) as ‘perfect proxy sets’.

Considering only common SNPs (the target of study for the HapMap Project) in CEU in the ENCODE data, one in five SNPs has 20 or more perfect proxies, and three in five have five or more. In contrast, one in five has no perfect proxies. As expected, perfect proxy sets are smaller in YRI, with twice as many SNPs (two in five) having no perfect proxy, and a quarter as many (5%) having 20 or more (Figs 11 and 12). These patterns are largely consistent across the range of frequencies studied by the project, with a trend towards fewer proxies at MAF < 0.10 (Fig. 11). Put another way, the average common SNP in ENCODE is perfectly redundant with three other SNPs in the YRI samples, and nine to ten other SNPs in the other sample sets (Fig. 13).

Of course, to be detected through LD in an association study, correlation need not be complete between the genotyped SNP and the causal variant. For example, under a multiplicative disease model and a single-locus $\chi^2$ test, the sample size required to detect association to an allele scales as $1/r^2$. That is, if the causal SNP has an $r^2 = 0.5$ to one tested in the disease study, full power can be maintained if the sample size is doubled.

The number of SNPs showing such substantial but incomplete correlation is much larger. For example, using a looser threshold for declaring correlation ($r^2 \geq 0.5$), the average number of proxies found for a common SNP in CHB+JPT is 43, and the average in YRI is 16 (Fig. 12). These partial correlations can be exploited through haplotype analysis to increase power to detect putative causal alleles, as discussed below.

**Evaluating performance of the Phase I map.** To estimate the proportion of all common SNPs captured by the Phase I map, we
evaluated redundancy among SNPs on the genome-wide map, and performed simulations based on the more complete ENCODE data. The two methods give highly similar answers, and indicate that Phase I should provide excellent power for CEU, CHB and JPT, and substantial power for YRI. Phase II, moreover, will provide nearly complete power for all three analysis panels.

**Redundancies among SNPs in Phase I HapMap.** Redundancy offers one measure that Phase I has sampled densely in comparison to the underlying scale of correlation. Specifically, 50% (YRI) to 75% (CHB+JPT, CEU) of all SNPs on the Phase I map are highly correlated ($r^2 \geq 0.8$) to one or more others on the map (Fig. 13; see also Supplementary Fig. 10). Over 90% of all SNPs on the map have highly statistically significant correlation to one or more neighbours. These partial correlations can be combined to form haplotypes that are even better proxies for a SNP of interest.

**Modelling Phase I HapMap from complete ENCODE data.** A second approach to evaluating the completeness of the Phase I data involves thinning the more complete ENCODE data to match Phase I for allele frequency and SNP density. Simulated Phase I HapMaps were used to evaluate coverage in relation to the full set of common SNPs (Table 6), and provided nearly identical estimates to those above: 45% (YRI) to 74% (CHB+JPT, CEU) of all common SNPs are predicted to have a proxy with $r^2 \geq 0.8$ to a SNP included in the Phase I HapMap (Supplementary Fig. 11).

Statistical power in association studies may be more closely approximated by the average (maximal) correlation value between a SNP and its best proxy on the map, rather than by the proportion exceeding an arbitrary (and stringent) threshold. The average values for maximal $r^2$ to a nearby SNP range from 0.67 (YRI) to 0.85 (CEU and CHB+JPT).

**Modelling Phase II HapMap from complete ENCODE data.** A similar procedure was used to generate simulated Phase II HapMaps from ENCODE data (Table 6). Phase II is predicted to capture the majority of common variation in YRI: 81% of all common SNPs should have a near perfect proxy ($r^2 \geq 0.8$) to a SNP on the map, with the mean maximal $r^2$ value of 0.90. Unsurprisingly, the CEU, CHB and JPT samples, already well served by Phase I, are nearly perfectly captured: 94% of all common sites have a proxy on the map with $r^2 \geq 0.8$, with an average maximal $r^2$ value of 0.97.

These analyses indicate that the Phase I and Phase II HapMap resources should provide excellent coverage for common variation in these population samples.

**Selection of tag SNPs for association studies**

A major impetus for developing the HapMap was to guide the design and prioritization of SNP genotyping assays for disease association studies. We refer to the set of SNPs genotyped in a disease study as tags. A given set of tags can be analysed for association with a
phenotype using a variety of statistical methods which we term tests, based either on the genotypes of single SNPs or combinations of multiple SNPs.

The shared goal of all tag selection methods is to exploit redundancy among SNPs, maximizing efficiency in the laboratory while minimizing loss of information\(^{24,27}\). This literature is extensive and varied, despite its youth. Some methods require that a single SNP serve as a proxy for other, untyped variants, whereas other methods allow combinations of alleles (haplotypes) to serve as proxies; some make explicit use of LD blocks whereas others are agnostic to such descriptions. Although it is not practical to implement all such methods at the project website, the HapMap genotypes are freely available and investigators can apply their method of choice to the data. To assist users, both a single-marker tagging method and a more efficient multimarker method have been implemented at http://www.hapmap.org.

**Tagging using a simple pairwise method.** To illustrate general principles of tagging, we first applied a simple and widely used pairwise algorithm\(^{32,34}\). SNPs are selected for genotyping until all common SNPs are highly correlated (\(r^2 \geq 0.8\)) to one or more members of the tag set.

Starting from the substantially complete ENCODE data, the density of common SNPs can be reduced by 75–90% with essentially no loss of information (Fig. 14). That is, the genotyping burden can be reduced from one common SNP every 500 bp to one SNP every 2 kb (YRI) to 5 kb (CEU and CHB+JPT). Because LD often extends for long distances, studies of short gene segments tend to underestimate the redundancy across the genome\(^{35}\).

**Although tags selected based on LD offer the greatest improvements in efficiency and information capture, even randomly chosen subsets of SNPs offer considerable efficiencies (Fig. 14).**

The data also reveal a rule of diminishing returns: a small set of highly informative tags captures a large fraction of all variation, with additional tags each capturing only one or a few proxies. For example, in CHB+JPT the most informative 1% of all SNPs (one per 50 kb) is able to proxy (at \(r^2 \geq 0.8\)) for 40% of all common SNPs, whereas a substantial proportion of SNPs have no proxies at all.

These observations are encouraging with respect to genome-wide association studies. A set of SNPs typed every 5–10 kb across the genome (within the range of current technology) can capture nearly all common variation in the genome in the CEU and CHB+JPT samples, with more SNPs required in the YRI samples.

**Tagging from the genome-wide map.** Whereas analysis of the complete ENCODE data set reveals the maximal efficiency likely to be possible with this tag selection strategy, analysis of the Phase I map illuminates the extent to which the current resource can be used for near-term studies. Specifically, using the same pairwise tagging approach above, 260,000 (CHB+JPT) to 474,000 (YRI) SNPs are required to capture all common SNPs in the Phase I data set (Table 7). That is, being incomplete and thus less redundant, the Phase I data are much less compressible by tag SNP selection than are the ENCODE data. Nevertheless, even at this level a half to a third of all SNPs can be selected as proxies for the remainder (and, by inference, the bulk of other common SNPs in the genome).

**Increasing the efficiency of tag SNPs.** Although the pairwise method

---

**Table 6** | Coverage of simulated Phase I and Phase II HapMap to capture all common SNPs in the ten ENCODE regions

<table>
<thead>
<tr>
<th>Analysis panel</th>
<th>Per cent maximum (r^2 \geq 0.8)</th>
<th>Mean maximum (r^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I HapMap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YRI</td>
<td>45</td>
<td>0.67</td>
</tr>
<tr>
<td>CEU</td>
<td>74</td>
<td>0.85</td>
</tr>
<tr>
<td>CHB+JPT</td>
<td>72</td>
<td>0.83</td>
</tr>
<tr>
<td>Phase II HapMap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YRI</td>
<td>81</td>
<td>0.90</td>
</tr>
<tr>
<td>CEU</td>
<td>94</td>
<td>0.97</td>
</tr>
<tr>
<td>CHB+JPT</td>
<td>94</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Simulated Phase I HapMaps were generated from the phased ENCODE data (release 16ct) by randomly picking SNPs that appear in dbSNP build 121 (excluding ‘non-rs’ SNPs in release 16a) for every 5-kb bin until a common SNP was picked (allowing up to three attempts per bin). The Phase II HapMap was simulated by picking SNPs at random to achieve an overall density of 1 SNP per 1 kb. These numbers are averages over 20 independent iterations for all ENCODE regions in all three analysis panels.

---

**Figure 14** | Tag SNP information capture. The proportion of common SNPs captured with \(r^2 \geq 0.8\) as a function of the average tag SNP spacing is shown for the phased ENCODE data, plotted (left to right) for tag SNPs prioritized by Tagger (multimarker and pairwise) and for tag SNPs picked at random. Results were averaged over all the ENCODE regions.
is simple, complete and straightforward, efficiency can be improved with a number of simple changes. First, relaxing the threshold on $r^2$ for tag SNP selection substantially reduces the number of tag SNPs selected, with only a modest decrease in the correlations among SNPs (Table 7). For example, reducing the $r^2$ threshold from 0.8 to 0.5 decreases the number of tag SNPs selected from the HapMap by 39% in CHB+JPT (260,000 to 159,000) and 32% in YRI (474,000 to 325,000). The average $r^2$ value between tags and other (unselected) SNPs falls much less dramatically than the number of tags selected, increasing efficiency. Whether such a loss of power is justified by the disproportionate reduction in work is a choice each investigator will need to make.

A second enhancement exploits multimarker haplotypes. Many investigators have discussed using multiple SNPs (in haplotypes and regression models) to serve as proxies for untyped sites\(^{35-38}\), which may reduce the number of tags required and increase the power of analyses performed. Figure 14 illustrates the point with one such method\(^3^{9}\), showing that a multimarker method allows greater coverage for a fixed set of markers (or, alternatively, fewer markers to achieve the same coverage). Although a full consideration of this issue is beyond the scope of this paper, the availability of these and other data should allow the comparison and application of such methods.

A third approach to increasing efficiency is to prioritize tags based on the number of other SNPs captured. Whereas 260,000 SNPs are required to provide $r^2 \geq 0.8$ for all SNPs in the Phase I HapMap (CHB+JPT), the best 10,000 such SNPs (4%) capture 22% of all common variable sites with $r^2 \geq 0.8$ (Table 8). Such prioritization can be applied using different weights for SNPs based on genomic annotation (for example, non-synonymous coding SNPs, SNPs in conserved non-coding sequence, and candidate genes of biological interest).

**Tag transferability across populations.** The most complete set of tags would be those based on all 269 samples; however, many studies may be performed in individuals more closely related to one particular HapMap population, and efficiency may be gained by selecting tags only from that population sample. (Selecting tags in a HapMap population sample that is known to be more distantly related than is another, for example, using CEU to pick tags for a study of Japanese, seems inefficient.)

An important question is how tags selected in one or more analysis panels will transfer to disease studies performed in these or other populations. Our data do not address this question directly, although the known similarity of allele and haplotype frequencies across populations within continents\(^{41}\) is encouraging. More data are clearly needed, however.

**Tag selection based on initial genotyping.** Whereas the discussions above assume *de novo* selection of SNPs, many investigators will have already performed initial studies, and wish to design follow-on experiments. The HapMap data can be used to highlight SNPs that might potentially explain a positive association signal, or those that were poorly captured (and thus still need to be tested) after a negative scan. In cases where multiple SNPs are both associated with the trait and with each other, the HapMap data can be queried to identify whether samples from any other analysis panel show a breakdown of LD in that region, and thus the possibility of narrowing the span over which the causal variant may reside.

**Applications to the analysis of association data**

Beyond guiding selection of tag SNPs, HapMap data can inform the subsequent analysis and interpretation in disease association studies.

**Analysis of an existing genotype data set.** The HapMap can be used to inform association testing, regardless of how tags were selected. Specifically, as long as the SNPs genotyped in a disease study have also been typed in the HapMap samples, it is possible to identify which SNPs are well captured by the genotyped SNPs (either singly, or in haplotype combinations), and which are not\(^{35}\).

This is of particular importance for genome-wide association studies performed using array-based, standardized genotyping reagents, which do not allow investigators to choose their own sets of tag SNPs. The Affymetrix 120K SNP array data included in Phase I of the HapMap provides a simple example: in CEU 48% of HapMap SNPs have substantial pairwise correlation ($r^2 \geq 0.5$) to one or more of the 120K SNPs on the array. An additional 13%, however, are not correlated to a single SNP, but are to a specific haplotype of two members of the 120K panel. By identifying such haplotype predictors in the HapMap, and testing them (in addition to the single SNPs) in a disease study, it is likely that power will be increased (I. Pe'er *et al.*, manuscript in preparation).

**Evaluating statistical significance and interpreting results.** An important challenge in genome-wide association testing is to develop statistical procedures that minimize false positives without greatly sacrificing true positives. The challenge is amplified by the correlated nature of polymorphism data, which makes simple frequentist approaches that assume independence (such as Bonferroni correction) highly conservative. To illustrate this point, we used the ENCODE data to estimate the ‘effective number of independent tests’ (the statistical burden of testing all common (MAF $\geq 0.05$) variation) across large genomic regions. Specifically, we re-sampled from the phased ENCODE chromosomes to create mock case-control panels in which all common SNPs were observed, but there was not a causal allele. The resulting $\chi^2$ distribution for association indicates that complete testing of common variation in each 500-kb region is equivalent to performing about 150 independent statistical tests (in CEU and CHB+JPT) and about 350 tests (in YRI). Although it will probably be desirable to perform such empirical estimates of significance within each disease study, these results illustrate how Bonferroni correction overestimates the statistical penalty of performing many correlated tests.

**Study of less common alleles.** We have focused primarily on the hypothesis that a single, common causal allele exists, and needs to be tested for association to disease. Of course, in many cases the causal allele(s) will be less common, and might be missed by such an approach.

It is possible to perform additional haplotype tests, beyond those

---

**Table 7** | Number of selected tag SNPs to capture all observed common SNPs in the Phase I HapMap

<table>
<thead>
<tr>
<th>Tag SNP set size</th>
<th>YRI captured (%)</th>
<th>CEU captured (%)</th>
<th>CHB+JPT captured (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>12.3</td>
<td>20.4</td>
<td>21.9</td>
</tr>
<tr>
<td>20,000</td>
<td>19.1</td>
<td>30.9</td>
<td>33.2</td>
</tr>
<tr>
<td>50,000</td>
<td>32.7</td>
<td>50.4</td>
<td>53.6</td>
</tr>
<tr>
<td>100,000</td>
<td>47.2</td>
<td>68.5</td>
<td>72.2</td>
</tr>
<tr>
<td>250,000</td>
<td>70.1</td>
<td>94.1</td>
<td>98.5</td>
</tr>
</tbody>
</table>

As in Table 7, tag SNPs were picked to capture common SNPs in HapMap release 16c1 using Haploview, selecting SNPs in order of the fraction of sites captured. Common SNPs were captured by fixed-size sets of pairwise tags at $r^2 \geq 0.8$.  

---

© 2005 Nature Publishing Group

---

**Table 8** | Proportion of common SNPs in Phase I captured by sets of tag SNPs

<table>
<thead>
<tr>
<th>Tag SNP set size</th>
<th>YRI captured (%)</th>
<th>CEU captured (%)</th>
<th>CHB+JPT captured (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>12.3</td>
<td>20.4</td>
<td>21.9</td>
</tr>
<tr>
<td>20,000</td>
<td>19.1</td>
<td>30.9</td>
<td>33.2</td>
</tr>
<tr>
<td>50,000</td>
<td>32.7</td>
<td>50.4</td>
<td>53.6</td>
</tr>
<tr>
<td>100,000</td>
<td>47.2</td>
<td>68.5</td>
<td>72.2</td>
</tr>
<tr>
<td>250,000</td>
<td>70.1</td>
<td>94.1</td>
<td>98.5</td>
</tr>
</tbody>
</table>
that capture known polymorphisms, in the hope of capturing less common or unreported alleles\textsuperscript{39}. Such haplotype analysis has a long history and proven value in mendelian genetics; the causal mutation is generally rare and unexamined during initial genotyping, but is frequently recognized by its presence on a long, unique haplotype of common alleles\textsuperscript{34,39–42}.

Admixture mapping. Although not designed specifically to enable admixture mapping\textsuperscript{39}, the HapMap has helped lay the groundwork for this approach. Admixture mapping requires a map of SNPs that are highly differentiated in frequency across population groups. By typing many SNPs in samples from multiple geographical regions, the data have helped to identify such SNPs for the design of genome-wide admixture mapping panels\textsuperscript{54,65} and can be further used to identify candidate SNPs with large allele frequency differences for follow-up of positive admixture scan results\textsuperscript{66}.

**Loss of heterozygosity in tumours.** Loss of heterozygosity (LOH) in tumour tissue can be a powerful indicator of the location of tumour suppressor genes, and genome-wide, fine-scale LOH analysis has been empowered by genome-wide SNP arrays\textsuperscript{67}. Germline DNA is not always available from the same subjects, however, and even if available, typing of germline DNA doubles project costs. In lower density scans for LOH (with markers far apart relative to the scale of LD), long runs of homozygosity in tumours are nearly always indicative of LOH. However, at higher densities runs of homozygosity can be due to haplotype homozygosity in the inherited germline DNA, rather than LOH.

The HapMap data can help minimize this difficulty; previous probabilities for homozygosity based on known frequencies of haplotypes in the HapMap data can be used to distinguish homozygosity due to haplotype sharing rather than LOH\textsuperscript{68}.

**Identifying structural variants in HapMap data**

Structural variations—segments where DNA is deleted, duplicated, or rearranged—are common\textsuperscript{44,69} and have an important role in diseases\textsuperscript{71–73}. The HapMap can provide some insight into structural variation because, in many cases, structural variants reveal themselves through signatures in SNP genotype data. In particular, polymorphic deletions are important to discover, because loss of genetic material is of obvious functional relevance, and results in aberrant patterns of SNP genotypes. These include apparent non-mendelian inheritance of SNP alleles, null genotypes and deviations from Hardy–Weinberg equilibrium. However, such SNPs are routinely discarded as technical failures of genotyping.

Thus, we scanned the unfiltered Phase I HapMap data using an approach developed and validated to identify polymorphic deletions from clusters of SNPs with aberrant genotype patterns (calibrated across the multiple centres and genotyping platforms\textsuperscript{74}). In total, 541 candidate deletion polymorphisms were identified, of which 150 were common enough to be observed as homozygotes.

The properties of these candidate deletions, including experimental validation of 90 candidates, are described in ref. 74. Validated

---

**Figure 15 | Length of LD spans.** We fitted a simple model for the decay of linkage disequilibrium\textsuperscript{189} to windows of 1 million bases distributed throughout the genome. The results of model fitting are summarized for the CHB+JPT analysis panel, by plotting the fitted $r^2$ value for SNPs separated by 30 kb. The overall pattern of variation was very similar in the other analysis panels\textsuperscript{44} (see Supplementary Information).
polymorphisms include 10 that remove coding exons of genes, such that in many cases individuals are homozygous null for the encoded transcript. Analysis of confirmed deletions often shows strong LD with nearby SNPs, indicating that LD-based approaches can be useful for detecting disease associations due to structural (as well as SNP) variants.

Polygenic inversions may also be reflected in the HapMap data as long regions where multiple SNPs are perfectly correlated: because recombination between an inverted and non-inverted copy is lethal, the inverted and non-inverted copies of the region evolve independently. A striking example corresponds to the known inversion polymorphism on chromosome 17, present in 20% of the CEU chromosomes, that has been associated with fertility and total recombination rate in females among Icelanders. Long LD may also arise, however, due to a low recombination rate or certain forms of natural selection, as discussed below.

Insights into recombination and natural selection

In addition to its intended function as a resource for disease studies, the HapMap data provide clues about the biology of recombination and history of natural selection.

A genome-wide map of recombination rates at a fine scale. On the basis of the HapMap data, we created a fine-scale genetic map spanning the human genome (Supplementary Fig. 12), including 21,617 identified recombination hotspots (one per 122 kb).

Both the number and intensity of hotspots contribute to overall variation in recombination rate. For example, we selected 25 regions of 5 Mb as having the highest (>2.75 cM Mb\(^{-1}\)) and lowest (<0.5 cM Mb\(^{-1}\)) rates of recombination in the deCODE (pedigree-based) genetic map. We detected recombination hotspots in all regions, even the low. But in the high cM Mb\(^{-1}\) regions hotspots are more closely spaced (one per 84 kb) and have a higher average intensity (0.124 cM) as compared to the low cM Mb\(^{-1}\) regions (one every 208 kb, and 0.051 cM, respectively).

Estimates of recombination rates and identified hotspots are robust to the specific markers and samples studied. Specifically, we compared these results to a similar analysis of the data of ref. 77 with about 1.6 million SNPs genotyped in 71 individuals. We find nearly complete correlation in rate estimates at a coarse scale (5 Mb) between these two surveys (\(r^2 = 0.99\)) and to the pedigree map (\(r^2 = 0.95\)). Very substantial correlation is found at finer scales: \(r^2 = 0.8\) at 50 kb and \(r^2 = 0.59\) at 5 kb. Moreover, of the 21,617 hotspots identified using the HapMap data, 78% (16,923) were also identified using the data of ref. 77.

The ability to detect events depends on marker density, with the larger number of SNPs studied by ref. 77 increasing power to detect hotspots, and presumably precision of rate estimates. There are, however, substantial genomic regions where the HapMap data have a higher SNP density. For example, more hotspots are detected on chromosomes 9 and 19 from the HapMap data. We expect that Phase II of HapMap will provide a genome-wide recombination map of substantially greater precision than either ref. 77, or Phase I, at fine scales.

Little is yet known about the molecular determinants of recombination hotspots. In an analysis of the data of ref. 77, another study (ref. 76) found significant evidence for an excess of the THE1A/B retrotransposon-like elements within recombination hotspots, and more strikingly for a sixfold increase of a particular motif (CCTCCGCT) within copies of the element in hotspots, compared to copies of the element outside hotspots. In analysing the HapMap data, we confirmed these findings (Supplementary Fig. 13). Furthermore, THE1B elements with the motif are particularly enriched within 1.5 kb of the centre of the hotspots compared to flanking sequence (\(P < 10^{-16}\)).

Correlations of LD with genomic features. Variation in recombination rate is important, in large part, because of its impact on LD. We thus examined genome-wide LD for correlation to recombination rates, sequence composition and gene features.

We confirmed previous observations that LD is generally low near telomeres, elevated near centromeres, and correlated with chromosome length (Fig. 15; see also Supplementary Figs 8b and 14). These patterns are due to recombination rate variation as discussed above. We also confirmed previously described relationships between LD and G+C content (Fig. 16), sequence polymorphism and repeat composition (Fig. 16).

We observe, for the first time, that LD tracks with both the density and functional classification of genes. We examined quartiles of the genome based on extent of LD, and looked for correlations to gene density. Surprisingly, we find that both the top and bottom quartiles of the genome have greater gene density as compared to the middle quartiles (6.7 as compared to 6.1 genes per Mb), as well as percentage of bases in codons (1.24% as compared to 1.08%). We have no explanation for this observation.

Although the majority of gene classes are equally divided between these two extreme quartiles of the genome, some classes of genes show a marked skew in their distribution. Genes involved in immune responses and neurophysiological processes are more often located in regions of low LD, whereas genes involved in DNA and RNA metabolism, response to DNA damage and the cell cycle are preferentially located in regions of strong linkage disequilibrium. It is intriguing to speculate that the extent of LD (and sequence diversity) might track with gene function due to natural selection, with increased diversity being favoured in genes involved in interface with the environment such as the immune response, and disadvantage for core cell biological processes such as DNA repair and packaging.

Natural selection. The preceding observation highlights the hypothesis that signatures of natural selection are present in the HapMap data. The availability of genome-wide variation data makes it

---

**Table 9 | High-differentiation non-synonymous SNPs**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (base number)</th>
<th>Gene*</th>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54,772,833</td>
<td>THEA</td>
<td>rs1702003</td>
</tr>
<tr>
<td>1</td>
<td>156,000,000</td>
<td>FY</td>
<td>rs12075</td>
</tr>
<tr>
<td>1</td>
<td>244,000,000</td>
<td>Q8NGY8_human</td>
<td>rs7555046</td>
</tr>
<tr>
<td>2</td>
<td>3,184,917</td>
<td>COLEC1</td>
<td>rs756783</td>
</tr>
<tr>
<td>2</td>
<td>73,563,622</td>
<td>ALMS1</td>
<td>rs3813227</td>
</tr>
<tr>
<td>2</td>
<td>73,589,553</td>
<td>ALMS1</td>
<td>rs654837</td>
</tr>
<tr>
<td>2</td>
<td>73,591,645</td>
<td>ALMS1</td>
<td>rs6724782</td>
</tr>
<tr>
<td>2</td>
<td>73,593,163</td>
<td>ALMS1</td>
<td>rs654839</td>
</tr>
<tr>
<td>2</td>
<td>73,629,222</td>
<td>ALMS1</td>
<td>rs2056648</td>
</tr>
<tr>
<td>2</td>
<td>73,629,311</td>
<td>ALMS1</td>
<td>rs1093972</td>
</tr>
<tr>
<td>2</td>
<td>109,000,000</td>
<td>EDAR</td>
<td>rs3827760</td>
</tr>
<tr>
<td>3</td>
<td>182,000,000</td>
<td>FXR1</td>
<td>rs13499</td>
</tr>
<tr>
<td>3</td>
<td>185,000,000</td>
<td>MCEF2L2</td>
<td>rs7639705</td>
</tr>
<tr>
<td>4</td>
<td>41,844,599</td>
<td>SLC30A9</td>
<td>rs1047626</td>
</tr>
<tr>
<td>4</td>
<td>46,567,077</td>
<td>ENSG000000172895.1</td>
<td>rs5825</td>
</tr>
<tr>
<td>8</td>
<td>101,000,000</td>
<td>ADH1B</td>
<td>rs1229984</td>
</tr>
<tr>
<td>8</td>
<td>10,517,787</td>
<td>RPL1</td>
<td>rs6601495</td>
</tr>
<tr>
<td>8</td>
<td>146,000,000</td>
<td>SLC39A4</td>
<td>rs1871534</td>
</tr>
<tr>
<td>10</td>
<td>50,402,145</td>
<td>ERCC6</td>
<td>rs4253047</td>
</tr>
<tr>
<td>11</td>
<td>46,701,579</td>
<td>F2</td>
<td>rs5896</td>
</tr>
<tr>
<td>15</td>
<td>46,213,776</td>
<td>SLC2A4</td>
<td>rs1426654</td>
</tr>
<tr>
<td>15</td>
<td>61,724,262</td>
<td>HERC1</td>
<td>rs7162473</td>
</tr>
<tr>
<td>16</td>
<td>30,996,126</td>
<td>ZNF646</td>
<td>rs749670</td>
</tr>
<tr>
<td>16</td>
<td>46,815,699</td>
<td>ABC1T</td>
<td>rs7822931</td>
</tr>
<tr>
<td>19</td>
<td>26,322,430</td>
<td>RNFL1</td>
<td>rs7225888</td>
</tr>
<tr>
<td>19</td>
<td>26,399,303</td>
<td>ENSG00000184253.2</td>
<td>rs6505228</td>
</tr>
<tr>
<td>19</td>
<td>66,022,323</td>
<td>RTTN</td>
<td>rs3911730</td>
</tr>
<tr>
<td>19</td>
<td>5,782,891</td>
<td>FUT6</td>
<td>rs364637</td>
</tr>
<tr>
<td>19</td>
<td>47,723,209</td>
<td>CEACAM1</td>
<td>rs810904</td>
</tr>
<tr>
<td>22</td>
<td>18,164,095</td>
<td>GBTL1</td>
<td>rs2073770</td>
</tr>
<tr>
<td>22</td>
<td>65,608,007</td>
<td>EDA2R</td>
<td>rs1385609</td>
</tr>
</tbody>
</table>

*Where no standard gene abbreviation exists, the ENSEMBL gene ID has been given.

**Figures**

16.322.430 RNF195 rs7225888

18.164.095 GBTL1 rs2073770

65.608.007 EDA2R rs1385609

---

© 2005 Nature Publishing Group

1313
possible to scan the genome for such signatures to discover genes that
were subject to selection during human evolution\(^6\); the HapMap
data also provide a genome-wide empirical distribution against
which previous claims of selection can be evaluated (rather than
relying solely on theoretical computer simulations).

Natural selection influences patterns of genetic variation in
various ways, such as through the removal of deleterious mutations,
the fixation of advantageous variants, and the maintenance of multiple
alleles through balancing selection. Each form of selection may have
occurred uniformly across the world (and thus be represented in all
human populations) or have been geographically localized (and thus
differ among populations).

Nearly all methods for recognizing natural selection rely on the
collection of complete sequence data. The HapMap Project’s focus on
common variation—and the process of SNP selection that achieved a
preponderance of high-frequency alleles (Fig. 5)—thus prevents their
straightforward application. Adjusting for the effect of SNP choice is
complex, moreover, because SNP choice varied over time as dbSNP
evolved, and was implemented locally at each centre.

For these reasons, we focus here on two types of analysis. First, we
examined the distributions of signatures of selection across the
genome. Although the absolute value of these measures is difficult
to interpret (owing to SNP ascertainment), the most extreme cases in
a genome-wide distribution are important candidates to evaluate for
selection. Second, we compared across functional categories, because
SNP choice was largely agnostic to such features, and thus systematic
differences may be a sign of selection.

The outcomes of these analyses confirm a number of previous
hypotheses about selection and identify new loci as candidates for
selection. Evidence for selective sweeps in particular genomic regions.
First we consider population differentiation, generally accepted as a clue
to past selection in one of the populations. The HapMap data reveal
926 SNPs with allele frequencies that differ across the analysis panels

---

**Table 10 | Candidate loci in which selection occurred**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (base number) at centre</th>
<th>Genes in region</th>
<th>Population</th>
<th>Haplotype frequency</th>
<th>Empirical P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>137,224,699</td>
<td>LCT</td>
<td>CEU</td>
<td>0.65</td>
<td>1.25 × 10⁻⁹</td>
</tr>
<tr>
<td>5</td>
<td>22,296,347</td>
<td>CDH12, PMCH1</td>
<td>YRI</td>
<td>0.25</td>
<td>5.77 × 10⁻⁸</td>
</tr>
<tr>
<td>7</td>
<td>79,904,387</td>
<td>CD36</td>
<td>YRI</td>
<td>0.24</td>
<td>2.72 × 10⁻⁶</td>
</tr>
<tr>
<td>7</td>
<td>73,747,934</td>
<td>PM32L5, WBSCR16</td>
<td>CEU</td>
<td>0.76</td>
<td>3.37 × 10⁻⁶</td>
</tr>
<tr>
<td>12</td>
<td>109,892,896</td>
<td>CUTL2</td>
<td>CEU</td>
<td>0.36</td>
<td>7.95 × 10⁻⁹</td>
</tr>
<tr>
<td>15</td>
<td>78,558,508</td>
<td>ARNT2</td>
<td>YRI</td>
<td>0.32</td>
<td>6.92 × 10⁻⁷</td>
</tr>
<tr>
<td>16</td>
<td>75,661,011</td>
<td>Desert</td>
<td>YRI</td>
<td>0.46</td>
<td>5.01 × 10⁻⁷</td>
</tr>
<tr>
<td>17</td>
<td>3,945,580</td>
<td>ITGAE, GSG2, HSA277841, CAMKK1, P2RX1</td>
<td>YRI</td>
<td>0.70</td>
<td>9.26 × 10⁻⁷</td>
</tr>
<tr>
<td>18</td>
<td>24,502,756</td>
<td>Desert</td>
<td>CEU</td>
<td>0.57</td>
<td>2.23 × 10⁻⁷</td>
</tr>
<tr>
<td>22</td>
<td>32,459,471</td>
<td>LARGE</td>
<td>YRI</td>
<td>0.36</td>
<td>7.82 × 10⁻⁹</td>
</tr>
<tr>
<td>X</td>
<td>20,171,291</td>
<td>Desert</td>
<td>YRI</td>
<td>0.33</td>
<td>5.02 × 10⁻⁹</td>
</tr>
<tr>
<td>X</td>
<td>64,323,320</td>
<td>HEPH</td>
<td>YRI</td>
<td>0.55</td>
<td>3.02 × 10⁻⁸</td>
</tr>
<tr>
<td>X</td>
<td>42,763,073</td>
<td>MAOB</td>
<td>CEU</td>
<td>0.53</td>
<td>4.21 × 10⁻⁹</td>
</tr>
<tr>
<td>X</td>
<td>34,399,948</td>
<td>Desert</td>
<td>CEU</td>
<td>0.57</td>
<td>8.85 × 10⁻⁸</td>
</tr>
</tbody>
</table>
in a manner as extreme as the well-accepted example of selection at
the Duffy (FY) locus (Supplementary Fig. 8c). Of these 926 SNPs, 32
are non-synonymous coding SNPs and many others occur in tran-
scribed regions, making them strong candidates for functional
polymorphisms that have experienced geographically restricted
selection pressures (see Table 9 and Supplementary Information
for details). In particular, the ALMS1 gene on chromosome 2 has
six amino acid polymorphisms that show very strong population
differentiation.

Another signature of an allele having risen to fixation through
selection is that all other diversity in the region is eliminated (known
as a selective sweep). We identified extreme outliers in the joint
distribution of heterozygosity (as assessed from shotgun sequencing
SNP discovery projects) and either population differentiation or
skewing of allele frequency towards rare alleles in each analysis
panel (Supplementary Fig. 15). We identified 19 such genomic
regions (13 on autosomes, 6 on the X chromosome) as candidates
for future study (Supplementary Table 4); these include candidates
for population-specific sweeps and sweeps in the ancestral
population. Encouragingly, this analysis includes among its top-
scoring results the LCT gene, which influences the ability to digest
dairy products90 and has been shown to be subject to past natural
selection91.

Long haplotypes as candidates for natural selection. Selective
sweeps that fail to fix in the population, as well as balancing selection,
lead to haplotypes that are relatively high in frequency and long in
duration. In the HLA region (which is widely believed to have been
influenced by balancing selection) multiple haplotypes of 500 SNPs
that extend more than 1 cM in length are observed with a frequency
in the HapMap samples of more than 1%. We identified other such
occurrences of long haplotypes across the genome (Supplementary
Fig. 8 and Supplementary Tables 5 and 6).

An approach to long haplotypes designed specifically to identify
regions having undergone partial selective sweeps is the long range
haplotype (LRH) test92,93, which compares the length of each haplo-
type to that of others at the locus, matched across the genome based
on frequency. Previously identified outliers to the genome-wide
distribution for the LRH test (Fig. 16) that have been identified as
candidates for selection include the LCT gene in the CEU sample
(empirical P-value = $1.3 \times 10^{-5}$), which was an outlier for the
heterozygosity/allele frequency test above, and the HBB gene
(empirical P-value = $1.39 \times 10^{-5}$) in the YRI sample. However,
most of the strongest signals in the LRH test (Table 10) were not
previously hypothesized as undergoing selection.

These four tests overlap only partially in the hypotheses they
address—heterozygosity, for example, is sensitive to older sweeps,
whereas the haplotype tests are most powerful for partial sweeps—but
encouragingly some candidate regions are found by more than
one test. In particular, six regions are identified both by long
haplotypes and by low heterozygosity, and three regions (LCT on
chromosome 2, and two regions on the X chromosome at 20 and
65 Mb) are identified by three different tests.

Confirming purifying selection at conserved non-coding
elements. Finally, we used the HapMap data to test an important
hypothesis from comparative genomics. Genomic sequencing has
shown that about 5% of the human sequence is highly conserved
across species, yet less than half of this sequence spans known
functional elements such as exons94. It is widely assumed that
conserved non-genic sequences lack diversity because of selective
constraint (that is, purifying selection), but such regions may simply
be coldspots for mutation, and thus be of little value as candidates for
functional study.

Analysis of allele frequencies helps to resolve this uncertainty.
Functional constraint, but not a low mutation rate, results in a
downward skew in allele frequencies for conserved sequences as
compared to neutral sequences95,96. We find that conserved non-
genic sequences display a greater skew towards rare alleles than do
intergenic regions, as predicted under purifying selection. This skew
is less extreme than that observed for exons (Supplementary Fig. 16),
reflecting either stronger purifying selection or the prioritization of
coding SNPs for genotyping by the HapMap centres regardless of
validation status. This novel evidence for ongoing constraint shows
that conserved non-genic sequences are not mutational coldspots,
and thus remain of high interest for functional study.

Conclusions

The International HapMap Project set out to create a resource that
would accelerate the identification of genetic factors that influence
medical traits. Analyses reported here confirm the generality of
hotspots of recombination, long segments of strong LD, and
limited haplotype diversity. Most important is the extensive
redundancy among nearby SNPs, providing (1) the potential to
extract extensive information about genomic variation without
complete resequencing, and (2) efficiencies through selection of
tag SNPs and optimized association analyses. Beyond the bio-
medical context, these data have made it possible to identify deletion
variants in the genome, explore the nature of fine-scale recombi-
nation and identify regions that may have been subject to natural
selection.

The HapMap Project (along with a previous genome-wide assess-
ment of LD97) is a natural extension of the Human Genome Project.
Where the reference sequence constructed by the Human Genome
Project is informative about the vast majority of bases that are
invariant across individuals, the HapMap focuses on DNA sequence
differences among individuals. Our understanding of SNP variation
and LD around common variants in the sampled populations is
reasonably complete; the current picture is unlikely to change with
additional data. In other aspects—such as the fine details of local
correlation among SNPs, rarer alleles, structural variants, and inter-
population differences—these resources are only a first step on the
path towards a complete characterization of genetic variation of
the human population. Planned extensions of the Phase I map
include Phase II of HapMap, with genotyping of another 4.6 million SNPs
attempted in the HapMap samples, and detailed genotyping of the
HapMap ENCODE regions in additional members of each HapMap
population sampled, as well as in samples from additional popu-
lations. These results should guide understanding of the robustness
and transferability of LD inferences and tag SNPs selected from the
current set of HapMap samples.

An important application of the HapMap data is to help make
possible comprehensive, genome-wide association studies. There are
now laboratory tools that make it practical to undertake such studies,
and initial results are encouraging98. Given the low prior probability
of causality for each SNP in the genome, however, rigorous standards
of statistical significance will be needed to avoid a flood of
false-positive results. Multiple replications in large samples provide
the most straightforward path to identifying robust and broadly
relevant associations. Given the potential for confusion if associa-
tions of uncertain validity are widely reported (and a persistent
tendency towards genetic determinism in public discourse), we urge
conservatism and restraint in the public dissemination and interper-
anation of such studies, especially if non-medical phenotypes are
explored. It is time to create mechanisms by which all results of
association studies, positive and negative, are reported and discussed
without bias.

The success of the HapMap will be measured in terms of the
gene discoveries enabled, and improved knowledge of disease
aetiology. Specifically, identifying which genes and pathways are
causal in humans has the potential to provide a new and solid
foundation for biomedical research. This is equally true whether the
variants that lead to the discovery of those genes are themselves rare
or common, or of large or small effect. The impact on diagnostics
and targeted prevention, however, will depend on how predictive each
given allele may be. Where genetic mechanisms underlie treatment
responses, both more efficient trials and individualized preventive and treatment strategies may become practical.

Success identifying alleles conferring susceptibility or resistance to common diseases will also provide a deeper understanding of the architecture of disease; how many genes are involved in each case, whether and how alleles interact with one another and with environmental exposures to shape clinical phenotypes. In this regard, it will be important to invest heavily in the discovery and characterization of relevant lifestyle factors, environmental exposures, detailed characterization of clinical phenotypes, and the ability to obtain such information in longitudinal studies of adequate size. Where environmental and behavioural factors vary across studies, replication will be hard to come by (as will clinical utility) unless we can learn to capture these variables with the same precision and completeness as genotypic variation. Technological innovation and international collaboration in these realms will probably be required (as they have been in the Human Genome Project and the HapMap) to advance the shared goal of understanding, and ultimately preventing, common human diseases.

METHODS

The project was undertaken by investigators from Japan, the United Kingdom, Canada, China, Nigeria and the United States, and from multiple disciplines: sample collection, sequencing and genotyping, bioinformatics, population genetics, statistics, and the ethical, legal, and social implications of genetic research. The Supplementary Information contains information about project participants and organization.

Choice of DNA samples. Any choice of DNA samples represents a compromise: a single population offers simplicity, but cannot be representative, whereas grid-sampling is representative of the current worldwide population, but is neither practical nor captures historical genetic diversity. The project chose to include DNA samples based on well-known patterns of allele frequencies across populations6, reflecting historical genetic diversity.5,12

For practical reasons, the project focused on SNPs present at a minor allele frequency (MAF) ≥ 0.05 in each analysis panel, and thus studied a sufficient number of individuals to provide good power for this frequency range13. Cell lines and DNA are available at the Coriell Institute for Medical Research (http://locus.umdnj.edu/nigms/products/hapmap.html).

Community engagement was employed to explain the project, and to learn how the project was viewed, in the communities where samples were collected.12,13

Papers describing the community engagement processes are being prepared.

One JPT sample was replaced for technical reasons, but not in time for inclusion in this report. We surveyed cryptic relatedness among the study participants, and identified a small number of pairs with unexpectedly high allele sharing (Supplementary Information). As the total level of sharing is not great, and as a subset of analyses performed without these individuals were unchanged, we include these individuals in the data and analyses presented here.

Genome-wide SNP discovery. The project required a dense map of SNPs, ideally containing information about validation and frequency of each candidate SNP. When the project started, the public SNP database (dbSNP) contained 2.6 million candidate SNPs, few of which were annotated with the required information.

To generate more SNPs and obtain validation information, shotgun sequencing of DNA from whole-genome libraries and flow-sorted chromosomes was performed11, augmented by analysis of sequence traces produced by Applied Biosystems10,11, and information on 1.6 million SNPs genotyped by Perlegen Sciences12, including 425,000 not in dbSNP when released (Supplementary Table 7). The HapMap Project contributed about 6 million new SNPs to dbSNP.

At the time of writing (October 2005) dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) contains 9.2 million candidate human SNPs, of which 3.6 million have been validated by both alleles having been seen two or more times during discovery (‘double-hit’ SNPs), and 2.4 million have genotype data (Fig. 1).

Comprehensive study of common variation across 5 Mb of DNA. To study patterns of genetic variation as comprehensively as possible, we selected ten 500-kb regions from the ENCODE Project13. These ten regions were chosen in aggregate to approximate the genome-wide average for G+C content, recombination rate, percentage of sequence conserved relative to mouse sequence, and gene density (Table 2).

In each such region additional sequencing and genotyping were performed to obtain a much more complete inventory of common variation. Specifically, bidirectional PCR-based sequencing was performed across each 500-kb region in 48 individuals (16 YRI, 16 CEU, 8 CHB, 8 JPT). Although the intent was for these same DNA samples to be included in Phase I, eight Yoruba and one Han Chinese sample used in sequencing were not among the 269 samples genotyped. (The nine samples are available from Coriell.)

All variants found by sequencing, and any others in dbSNP (build 121) not found by sequencing, were genotyped in all 269 HapMap samples. If the first attempt at genotyping was unsuccessful, a second platform was tried for each SNP.

False-positive and false-negative rates in PCR-based SNP discovery. The false-positive rate of SNP discovery by PCR-based resequencing was estimated at 7–11% (for the two sequencing centres), based on genotyping of each candidate SNP in the same samples used for discovery.

The false-negative rate of SNP discovery by PCR resequencing was estimated at 6%, using as the denominator a set of SNPs previously in dbSNP and confirmed by genotyping in the specific individuals sequenced. The false-negative rate was considerably higher, however, for singletons (SNPs seen only as a single heterozygote): 15% of singletons covered by high-quality sequence data were not detected by the trace analysis, and another 25% were missed due to a failure to obtain a high-quality sequence over the relevant base in the one heterozygous individual (D. J. Richter et al., personal communication).

False-positive and false-negative rates in dbSNP. The false positive rate (candidate SNPs that cannot be confirmed as variable sites) estimated for dbSNP was 17%. This represents an upper bound, because dbSNP entries that are not polymorphic in the 269 HapMap samples could be rare variants, or polymorphic in other samples. We note that as the catalogue of dbSNP gets deeper, the rate at which candidate SNPs are monomorphic in any given sample is observed to rise (Supplementary Table 8). This is expected because the number of rare SNPs and false positives scales with depth of sequencing, whereas the number of true common variants will plateau.

SNP genotyping for the genome-wide map. Genotyping assays were designed from dbSNP, with priority given to SNPs validated by previous genotyping data or both alleles having been seen more than once in discovery. Data from the Chimpanzee Genome Sequencing Project14 were used in SNP validation if they confirmed the ancestral status of a human allele seen only once in discovery (Supplementary Information). Non-synonymous coding SNPs were also prioritized for genotyping. Two whole-genome, array-based genotyping reagents were used efficiently to increase SNP density: 40,000 SNPs from Illumina, and 120,000 SNPs from Affymetrix10.

To monitor progress, the genome was partitioned into 5-kb bins, with genotyping continuing through iterative rounds until a set of predetermined ‘stopping rules’ was satisfied in each analysis panel. (1) Minor allele frequency: in each analysis panel a common SNP (MAF ≥ 0.05) was obtained in each 5-kb bin. (2)Spacing: the distance between adjacent SNPs was 2–8 kb, with at least 9 SNPs across 50 kb. (3)‘HapMappable’ genome: with available technologies it is challenging to study centromeres, telomeres, gaps in genome sequence, and segmental duplications. The project identified such regions15 (Supplementary Table 9), spanning 4.4% of the finished human genome sequence, in which only a single attempt to develop a genotyping assay was required. (4) Three strikes, you’re out: if the above rules were not satisfied after three attempts to develop an assay in a given 5-kb region, or if all available SNPs in dbSNP had been tried, genotyping was considered complete for Phase I. Two attempts were considered sufficient if one attempt was of a SNP previously shown to have MAF ≥ 0.05 in the appropriate population sample in a previous genome-wide survey. (5) Quality control: ongoing and standardized quality control (QC) filters and three rounds of quality assessment (QA) were used to ensure and document the high quality of the genotype data.

QC filters were systematically performed, with each SNP tested for completeness (>80%), consistency across five duplicate genotypes (± 1 discrepancy), mendelian inheritance in 60 trios (± 1 discrepancy in each of YRI and CEU), and Hardy–Weinberg equilibrium (P > 0.001)16. SNPs in the Phase I data set passed all the QC filters in all the analysis panels and were polymorphic in the HapMap samples. Failing SNPs were released (with a special flag), as they can help to identify polymorphisms under primers, insertions/deletions, paralogous loci and natural selection.

Three QA exercises were carried out. First, a calibration exercise to ‘benchmark’ each platform and laboratory protocol. Second, a mid-project evaluation of each genotyping centre. Third, a blind analysis of a random sample of the complete Phase I data set. A number of SNPs were genotyped more than once during the project, or by other investigators, providing additional information about data quality. See the Supplementary Information for full information about the QA exercises.

An exhaustive approach was taken to mtDNA. Alignment of more than 1,000 publicly available mtDNA sequences of African (n = 87), European (n = 928) and Asian (n = 238) geographical origin18 was used to identify 210 common
variants (MAP ≥ 0.05 in at least one continental region) that were attempted in the samples.

Data release. Data deposited at the Data Coordination Center and released at http://www.hapmap.org, a Japanese mirror site http://hapmap.jst.go.jp/ and dbSNP include ascertainment status of each SNP at the time of selection, primer sequences, protocols for genotyping, genotypes for each sample, allele frequencies, and, for SNPs that failed QC filters, a code indicating the mode(s) of failure.

Initially, because of concern that third parties might seek patents on HapMap data, users were required to agree to a web-based ‘‘click-wrap license’’, assenting that they would not prevent others from using the data (http://www.hapmap.org/cgi-perl/registration). In December 2004 this license was dropped, and all data were released without restriction into the public domain.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: J. Beck, C. Beiswanger, D. Coppock, A. Leach, J. Mintzer and L. Toji (Coriell Institute for Medical Research) for helpful advice; P. D. (donnelly@stats.ox.ac.uk). ARTICLES NATURE|Vol 437|27 October 2005

1318 © 2005 Nature Publishing Group

Author Contributions David Altshuler, Lisa D. Brooks, Aravinda Chakravarti, Francis S. Collins, Mark J. Daly and Peter Donnelly are members of the writing group responsible for this manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A. (altshuler@molphg.mgh.harvard.edu) or P.D. (donnelly@stats.ox.ac.uk).
The International HapMap Consortium (Participants are listed by institution, alphabetized, and then alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Genotyping centres: Baylor College of Medicine and ParAllele BioScience Richard A. Gibbs (Principal Investigator),1 John W. Belmont,1 Andrew Boudreau,2 Suzanne M. Leal,1 Paul Hardenbol,2 Shiran Pasternak,2 David A. Wheeler,3 Thomas D. Willis,2 Fuli Yu,2 Beijing Genomics Institute Huanming Yang (Principal Investigator),3 Changqing Zeng (Principal Investigator);4 Yang Gao,4 Haoran Hu,4 Weitao Hu,4 Chaohua Li,5 Wei Lin,5 Siqi Liu,5 Hao Pan,6 Xiaoli Tang,6 Jian Wang,6 Wei Wang,6 Jun Yu,6 Bo Zhang,6 Qingrun Zhang,6 Hongbin Zhao,6 Hui Zhao,6 Jun Zhou,7 Broad Institute of Harvard and Massachusetts Institute of Technology Stacey B. Gabriel (Project Leader),8 Rachel Barry,9 Brendan Blumensienthal,10 Amy Camargo,10 Matthew Defelice,10 Maura Faggert,10 Mary Goyette,7 Supriya Gupta,11 Jamie Moore,7 Huy Nguyen,12 Robert C. Onofrio,7 Melissa Parkin,9 Jessica Roy,11 Erich Stahl,10 Ellen Winchester,11 Liuda Ziaugra,9 David Altschuler (Principal Investigator)13; Chinese National Human Genome Center at Beijing Yan Shen (Principal Investigator),14 Zhihui Yao;14 Chinese National Human Genome Center at Shanghai Wei Huang (Principal Investigator),15 Xin Chu,15 Yungang He,15 Li Jin,15 Yangfan Liu,15 Yayun Shen,15 Weixui Sun,15 Haifeng Wang,15 Yi Wang,15 Ying Wang,15 Xiaoyang Xiong,15 Liang Xu,15 Chinese University of Hong Kong Mary M. Y. Wayne (Principal Investigator),16 Stephen K. W. Tsui,17 Hong Kong University of Science and Technology Hong Xue (Principal Investigator),18 J. Tze-Fei Wong8; Illumina21,22 Illumina21,22 Laura M. Galver (Project Leader),19 Jia-Bing Fan9,19 Arnold R. Oliphant19,22 Mark S. Chee (Principal Investigator)19,22; McGill University and Quebec genomics Centre Alexandre Montpetit (Project Leader)19,22 Fanny Chagnon19,22,23 Vincent Ferretti19,22,23 Martin Leboeuf19,22,23 Jean-François Olivier19,22,23 Michael S. Phillips19,22,23 Stéphanie Roux19,22,23 Clémentine Sallée19,22,23 Andrei Verner19,22,23 Thomas J. Hudson (Principal Investigator)19,22,23; Perlegen Sciences Kelly A. Fraser (Principal Investigator),15 Dennis G. Ballinger,15 David R. Cox,15 David A. Hinds33; Laura L. Stuve15; University of California at San Francisco and Washington University Pui-Yan Kwok (Principal Investigator)6,16 Dongmei Cai15, Daniel C. Koboldt15, Raymond D. Miller,15 Ludmila Pawlickowska15, Patricia Taillon-Miller15, Ming Xiao15; University of Hong Kong Lap-Chee Tsui (Principal Investigator)15, William Mak15, Pak C. Sham15, You Qiang Song14, Paul K. H. Tam15; University of Tokyo and RIKEN Yusuke Nakamura (Principal Investigator)19,20,21 Takahisa Kawaguchi20, Takyu Kitamoto20, Takashi Morizono20, Atsushi Nagashima20, Yozo Ohnishi20; Akihiko Sekine20, Toshikazu Totsuka20, Tatsuo Katsuya20; Wellcome Trust Sanger Institute Panos Deloukas (Project Leader)24, Christine P. Bird24, Marcos Delgado24, Emmanouil T. Dermitzakis24, Rhian Gwilliam24, Sarah Hunt24, Jonathan Morrison24, Don Powell24, Barbara E. Stranger24, Pamela Whittaker24, David R. Bentley (Principal Investigator)24

Analysis groups: Broad Institute of Harvard and Massachusetts Institute of Technology Mark J. Daly (Project Leader)4,5, Paul I. W. de Bakker4,5, Jeff Barrett4,5, Ben Fry16, Julian Maller4,5, Steve McCarroll4,5, Nick Patterson4,5, Itsik Pe’er4,5, Shaun Purcell4, Thomas J. Richter, Pardis Sabeti24, Richa Saxena4,6, Stephen F. Schaffner24, Patrick Varilly4, David Altschuler (Principal Investigator)4,5; Cold Spring Harbor Laboratory Lincoln D. Stein (Principal Investigator)24, Lalitha Krishnan24, Albert Vernon Smith24, Gudmundur A. Thorsson24; Johns Hopkins University School of Medicine Aravinda Chakravarti (Principal Investigator)24, Peter E. Chen24, David J. Cutler24, Carl S. Kashuk24, Shin Lin24; University of Michigan Goncalo R. Abecasis (Principal Investigator)24, WeiHua Guan24, Heather M. Munro24, Zhaohui Steve Qin24, Daryl J. Thomas24; University of Oxford Gideon McVean (Project Leader)26, Leonardo Bottolo27, Susana Eyheramendy27, Colin Freeman27, Jonathan Marchini27, Simon Myers27, Chris Spencer27, Matthew Stephens28, Peter Donnelly (Principal Investigator)27; University of Oxford, Wellcome Trust Centre for Human Genetics Lon R. Cardon (Principal Investigator)29, Geraldine Clarke29, David M. Evans29, Andrew P. Morris29, Bruce S. Weir29, RIKEN Tatsukiko Tsunoda29; US National Institutes of Health James C. Mullikin31, Stephen T. Sherry32, Michael Feolo32

Community engagement and sample collection groups: Beijing Normal University and Beijing Genomics Institute Houcan Zhang34, Changqing Zeng; Hui Zhao;35 Health Sciences University of Hokkaido, Eubios Ethics Institute, and Shinsu University Ichiro Matsuda (Principal Investigator)34, Yoshimitsu Fukushima34, Darryl R. Macer34, Eiko Suda34; Howard University and University of Ibadan Charles N. Rotimi (Principal Investigator)35, Clement A. Adebamowo28, Ile Ajayi35, Toyin Aniagwu35, Patricia A. Marshall36, Chibuzor Nkwodimah35, Charmaine D. M. Royal35; University of Utah Mark F. Leppert (Principal Investigator)41, Missy Dixon41, Andy Peiffer41

Ethical, legal and social issues: Chinese Academy of Social Sciences Renzong Qiu38; Genetic Interest Group Alastair Kent39; Kyoto University Kazuto Kato41; Nagasaki University Norio Nikawa41; University of Ibadan School of Medicine Isaac F. Adeowo38; University of Montréal Bartha M. Knoppers41; University of Oklahoma Morris W. Foster41; Vanderbilt University Ellen Wright Clayton41; Wellcome Trust Jessica Watkin48

SNP discovery: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)1, John W. Belmont1, Donna Mumy1, Lynne Nazareth1, Erica Sodergren1, George M. Weinstock1, David A. Wheeler1, Imtiaz Yakub; Broad Institute of Harvard and Massachusetts Institute of Technology Stacey B. Gabriel (Project Leader)1, Robert C. Onofrio1, Daniel J. Richter1, Liuda Ziaugra,9 Bruce W. Birren,1 Mark J. Daly40,115, David Altschuler (Principal Investigator)1, Washington University Richard K. Wilson (Principal Investigator)41, Lucinda L. Fulton41; Wellcome Trust Sanger Institute Jane Rogers (Principal Investigator)27, John Burton11, Nigel P. Carter11, Christopher M. Cree11, Mark Griffiths11, Matthew C. Jones11, Kirsten Mclay11, Robert W. Plumb11, Mark T. Ross11, Sarah K. Sims11, David W. Willey21


Affiliations for participants: 1 Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, Texas 77030, USA. 2 ParAllele Bioscience, Inc., 7300 Shoreline Court, South San Francisco, California 94080, USA. 3 Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China. 4 The Broad Institute of Harvard and Massachusetts Institute of Technology, 1 Kendall Square, Cambridge, Massachusetts 02139, USA. 5 Massachusetts General Hospital and Harvard Medical School, Simchess Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. 6 Chinese National Human Genome Center at Beijing.