nature genetics

A worldwide survey of haplotype variation and linkage disequilibrium in the human genome

Donald F Conrad[1,4], Mattias Jakobsson[2,4], Graham Coop[1,4], Xiaoquan Wen[1], Jeffrey D Wall[3], Noah A Rosenberg[2] & Jonathan K Pritchard[1]

Recent genomic surveys have produced high-resolution haplotype information, but only in a small number of human populations. We report haplotype structure across 12 Mb of DNA sequence in 927 individuals representing 52 populations. The geographic distribution of haplotypes reflects human history, with a loss of haplotype diversity as distance increases from Africa. Although the extent of linkage disequilibrium (LD) varies markedly across populations, considerable sharing of haplotype structure exists, and inferred recombination hotspot locations generally match across groups. The four samples in the International HapMap Project contain the majority of common haplotypes found in most populations: averaging across populations, 83% of common 20-kb haplotypes in a population are also common in the most similar HapMap sample. Consequently, although the portability of tag SNPs based on the HapMap is reduced in low-LD Africans, the HapMap will be helpful for the design of genome-wide association mapping studies in nearly all human populations.

Linkage disequilibrium (LD) is of central importance in diverse aspects of human genetics. Most prominently, patterns of LD and haplotype variation serve as the backdrop for the design of association mapping studies[1,2]. Patterns of LD are also informative about population histories and human migrations[3–5], recent natural selection[2,6,7] and the distribution and evolution of recombination hotspots[8–10].

Recent studies from the International HapMap Project[2] and from Perlegen Sciences[11,12] have created dense genome-wide haplotype maps for populations of European, West African and East Asian descent. These maps provide important resources for the design of SNP-based studies in at least two respects. First, they represent an important source of validated SNPs with dense coverage of the genome. Second, they provide detailed information about haplotype structure that can be used, for example, to select tag SNPs for use in association studies[2].

A practical issue for association studies is whether tag SNPs chosen using the HapMap data will adequately capture patterns of variation in other populations. Several recent studies agree that the HapMap European (CEU) and East Asian data (CHB+JPT) are valuable resources for choosing tag SNPs for additional populations of European or East Asian descent, respectively[13–18]. However, European and East Asian populations have lower levels of differentiation than other continental groups[19], so these results do not necessarily imply that the HapMap will provide good tagging information everywhere in the world. Indeed, recent studies of more divergent populations have provided somewhat conflicting results[20–23].

In this article, we perform a global survey of haplotype variation in 52 worldwide populations, using SNPs spread across 36 genomic regions. Our study is designed to suggest initial answers to several questions. How useful are current SNP databases for studying haplotype variation in diverse human populations? To what extent are patterns of haplotype variation similar—or different—across diverse populations, and what do they imply about human history and patterns of recombination? To what extent do the HapMap populations predict patterns of haplotype diversity found in a worldwide set of populations? Answers to these questions provide a basis for understanding the extent to which tag SNPs derived from the HapMap will be useful for association studies worldwide.

RESULTS

We surveyed SNP variation in 927 unrelated individuals from 52 populations in the Human Genome Diversity Project (HGDP)-Centre d'Etude du Polymorphisme Humain (CEPH) Cell Line Panel[19,24–26]. We designed genotyping assays for 3,024 SNPs spaced across 36 genomic regions, including 32 autosomal regions and four regions on the non-pseudoautosomal X chromosome, covering a total of ∼12 Mb (see Methods). To facilitate inferences about both fine-scale and long-range LD, we designed the study so that each genomic region contained a central high-density 'core' of 60 SNPs at an average spacing of 1.5 kb, as well as additional SNPs at lower density outside the core region, extending 120 kb in each direction at 10-kb spacing. Hence, each region spanned a total of ∼330 kb.

[1]Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA. [2]Department of Human Genetics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109, USA. [3]Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, California 90089, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to N.A.R. (rnoah@umich.edu).

Received 28 June; accepted 22 September; published online 22 October 2006; doi:10.1038/ng1911

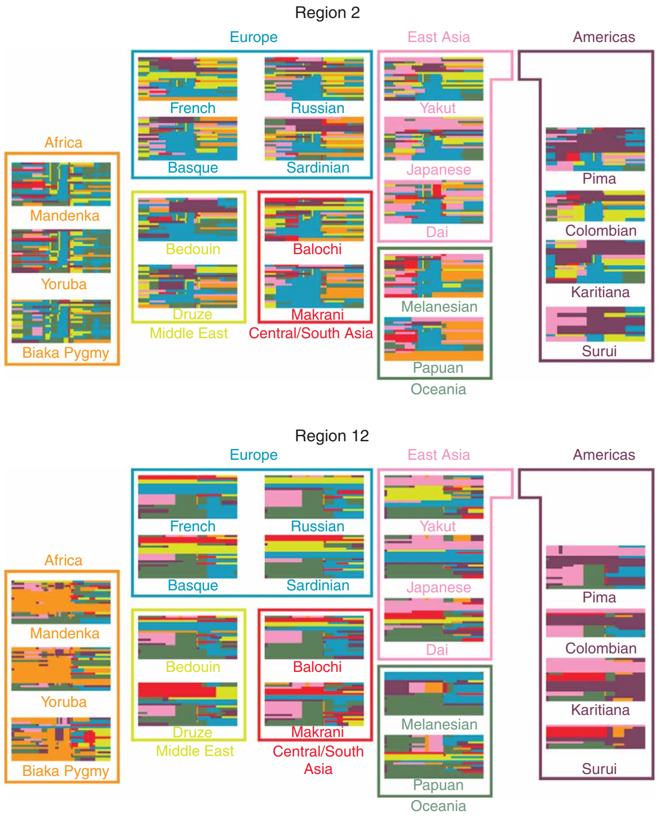

Figure 1 Haplotype structure in diverse populations for two genomic regions of size ~330 kb. For each population, haplotypes are plotted in rows; horizontal position is proportional to physical position in the sequence. Each haplotype is represented as a mosaic of seven 'template' haplotypes, each of which is common in a different part of the world and is colored using the same color as the text for that region (see Methods). Haplotypes are sorted so that within populations, similar haplotypes are drawn close together. A pair of haplotypes is identical across the entire region if and only if both share the same coloring pattern (rare minor alleles not on any template are dropped from the analysis). The same coloring scheme is used for all populations.

To maximize the degree to which the regions were representative of the human genome, we chose regions across the range of local gene densities and meiotic recombination rates (as estimated by the deCODE map[27]) without regard to the genes they contained. For the four X-chromosomal regions, and for 16 of the autosomal regions, we chose SNPs from dbSNP, with priority given to SNPs identified by a discovery effort in a multiethnic panel[12]. The remaining 16 autosomal regions were located on chromosome 21, with all SNPs chosen from a SNP discovery effort that used a uniform multiethnic panel[11].

After excluding SNPs that were monomorphic, that failed genotyping or quality control or that were in Hardy-Weinberg disequilibrium within unstructured populations, our final data set included genotype data for 2,834 SNPs (see Methods). Data quality was extremely high, with an estimated genotype error rate of 2×10^{-4} and a missing data rate of 0.1%. Haplotype phase was estimated using fastPHASE, a method with the attractive feature that its model allows haplotype structure to vary across populations[28]. Extensive testing of phasing accuracy indicates that fastPHASE provides suitable accuracy for our purposes (**Supplementary Note** online).

Haplotype variation

Figure 1 illustrates the phased haplotypes in two of our genomic regions using a new scheme for graphical representation of haplotype structure. The similarity of haplotype structure is greatest for nearby populations, especially for populations from the same continent. The complex and variable mosaics of haplotypes in Africa (and the simpler, less variable haplotypes in Oceania and the Americas) illustrate a steady trend of reduction in haplotype diversity with distance from Africa.

The number of distinct haplotypes per genomic core region declines from Africa along the likely path of human migrations into the Middle East, from the Middle East to Europe and Central and South Asia (referred to as 'Central/South Asia'), from Central/South Asia to East Asia, and from East Asia to Oceania and the Americas (illustrated in the quantitative view of haplotype structure in **Fig. 2**). This result matches the prediction of a serial dilution model[29] in which each successive human migration drew a sample of haplotypes from among those available at the previous location. The model also predicts a larger value for the fraction of haplotypes of location A found in location B than when location A is further from Africa and B is closer. This prediction is a consequence of the fact that the fraction of haplotypes of location A found in location B is the quotient of the number of shared haplotypes (which is the same for comparisons in both directions) and the number of haplotypes in location A, a quantity that is smaller at greater distances from Africa. **Figure 2** shows for each ordered pair of geographic regions the fraction of the haplotypes found in the first region that are also observed in the second. As predicted, for pairs of geographic regions at different numbers of migrational steps from Africa (Africa = 0, Middle East = 1, Europe = Central/South Asia = 2,

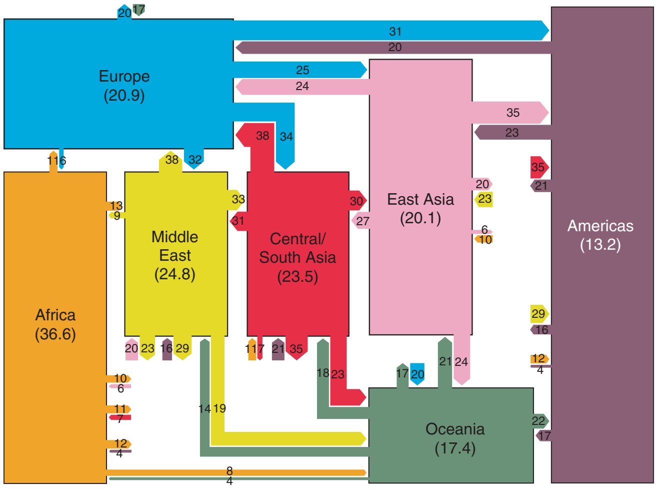

Figure 2 Schematic world map of haplotype diversity. Colored boxes represent regions of the world, positioned geographically. The average number of haplotypes per genomic core region in a sample size of 54 chromosomes is written for each geographic region. Links entering a geographic region indicate the percentages of distinct haplotypes from the geographic region found in other regions (and are drawn proportionately in width). For example, on average 11% of haplotypes observed in Europe in a given part of the genome are found in Africa, whereas 6% of African haplotypes are found in Europe. The links can be viewed as a description of haplotype 'flow': for example, 11% gives a measurement of the proportion of distinct European haplotypes that could have come from Africa (without mutation or recombination), and 6% gives the proportion of African haplotypes that could have come from Europe.

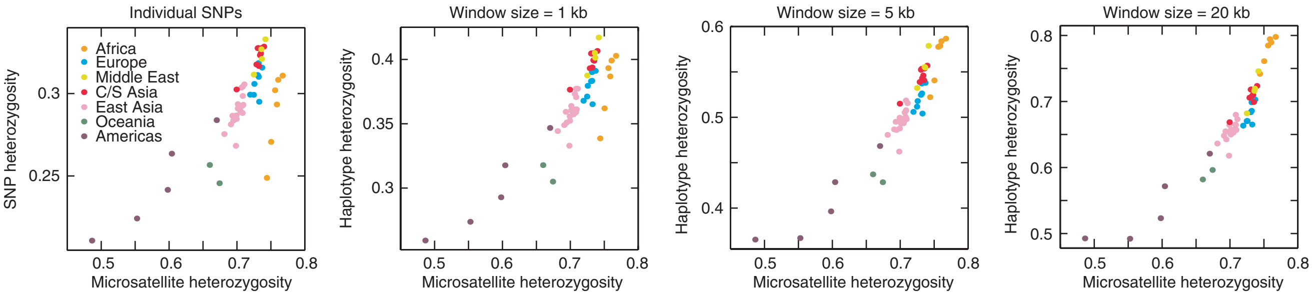

Figure 3 Effect of ascertainment bias on haplotype diversity. The plots show haplotype heterozygosity computed in windows of four sizes, plotted against heterozygosity of microsatellite loci in the same populations[29,41].

East Asia = 3, Oceania = America = 4), the 'flow' of haplotypes is always greater from the region closer to Africa toward the farther region than the flow in the reverse direction.

A potential concern about haplotype diversity computations is that they might be biased by SNP ascertainment procedures[30–33]. Indeed, the average heterozygosity of individual SNPs in our study shows evidence of ascertainment bias. Owing to the inadequate representation of African samples in panels used for SNP discovery, SNP heterozygosity is greatest in the Middle East, Central and South Asia and Europe (**Fig. 3**); in contrast, studies with little or no ascertainment bias have consistently identified African populations as having the greatest genetic diversity[19,29,34–36]. When we consider longer haplotypes, however, African haplotype heterozygosities overtake those of Europe and Asia, so that for haplotypes of length ≥20 kb, there is almost perfect correlation between haplotype heterozygosity and the heterozygosity of microsatellite loci in the same populations (Spearman's $\rho = 0.96$, $P < 10^{-8}$). This effect is likely to be a consequence of the fact that for highly polymorphic markers, the ascertainment scheme has relatively little impact on which markers are ascertained, as the same set of markers will probably be identified with most schemes[33]. Although the ascertainment of individual SNPs, which are not highly polymorphic, may depend heavily on the ascertainment scheme, the same underlying haplotypes, which are highly polymorphic, are likely to be observed regardless of which SNPs are studied in a genomic region. Hence, although ascertainment is known to cause biases in such statistics of individual loci as SNP allele frequencies[30,32,37], it is less likely to bias analyses of long haplotypes.

Recombination rates across populations

To investigate the properties of the underlying recombination process that gives rise to LD, we applied a recently developed method that uses the strength of LD across sites to estimate the historical extent of fine-scale recombination in a genomic region[8]. Estimates are obtained for the population recombination rate $\rho = 4N_er$, where N_e is effective population size and r is the meiotic recombination rate per kb. High levels of LD produce low estimates of ρ, and recombination hotspots lead to localized peaks of high ρ.

Our fine-scale estimates of the recombination landscape for different populations often show marked correspondence in the estimated locations of hotspots (**Fig. 4a**). This result indicates that areas of haplotype breakage are frequently shared across diverse populations; cases in which a hotspot is not detected in some populations despite being observed in others, such as for Yoruba in Region 2, where the estimate of ρ is constant across the entire region, may result from incomplete power of the method to detect hotspots or from the possibility of rapid evolution of hotspot differences across populations[9,10,38].

For each population, averaging over rate variation within genomic regions, estimates of the population recombination rate ρ are strongly correlated with meiotic recombination rate r estimated from Icelandic pedigrees (**Fig. 4b**, $P < 0.01$ for most populations). However, the nature of the relationship between ρ and r varies greatly across populations. Because $\rho = 4N_er$, the slope of a linear regression of ρ on $4r$, denoted $\hat{N}_e(\rho)$, provides an estimate of the effective population size N_e. Hence, for example, the steep slope for Bantu reflects a large N_e, in contrast to the flatter slope for Pima.

The estimate $\hat{N}_e(\rho)$ based on the slope of the regression line between ρ and $4r$ provides a summary of the extent of LD, with larger $\hat{N}_e(\rho)$ corresponding to smaller amounts of LD. Both the estimate $\hat{N}_e(\rho)$ and microsatellite-based heterozygosity are expected

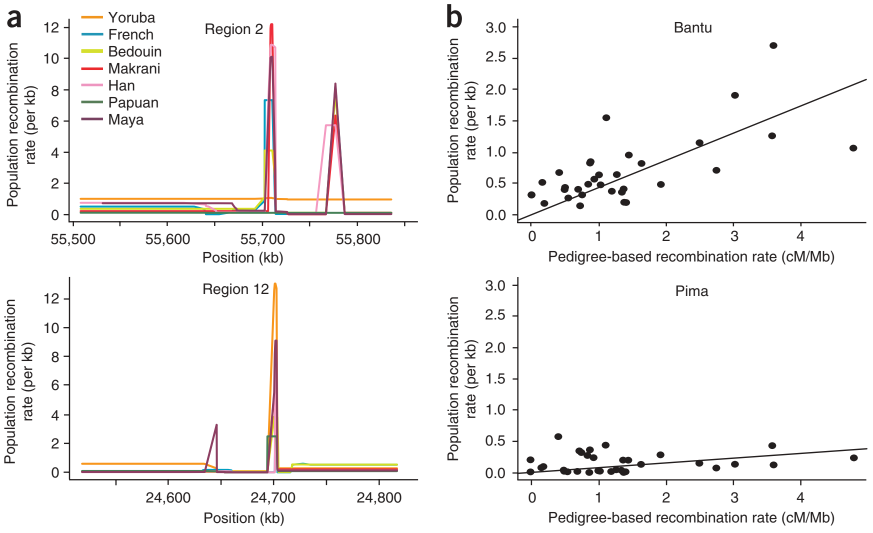

Figure 4 Population recombination rates (ρ) across genomic regions. (**a**) Fine-scale estimates of variation in ρ across the same two genomic regions shown in **Figure 1**. (**b**) Scatter plot of the relationship between ρ and the meiotic recombination rates for each autosomal region, as estimated from Icelandic pedigrees[27] (data shown for autosomal regions only).

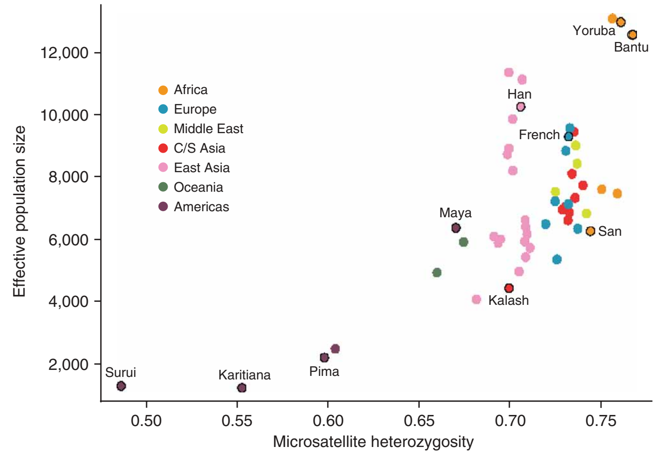

Figure 5 Relationship between $\hat{N}_e(\rho)$ and microsatellite heterozygosity for individual populations.

to increase with N_e, so that under simple demographic models, these two quantities should be highly correlated (illustrated in **Fig. 5**, which plots the relationship between the extent of LD and microsatellite heterozygosity). In addition, LD is extremely strong (low $\hat{N}_e(\rho)$) and diversity very low in the American populations and, to a lesser extent, in the two populations from Oceania (**Fig. 5**). Notably, populations similar to those in the HapMap sample (Yoruba, French and Han), all of which are representatives of large cosmopolitan groups, have among the lowest LD of any of the populations in our data set. We see no systematic difference between the chromosome 21 regions and other autosomal regions, despite the use of different SNP ascertainment protocols (**Supplementary Note**).

Several populations have unusual levels of LD relative to their neighbors. By far the highest $\hat{N}_e(\rho)$ among the Native Americans is seen in the Maya, perhaps owing to the presence of European chromosomes in these samples[19] that could show evidence of recombination. We have observed previously that the Kalash population from Pakistan clusters distinctly from other Eurasian populations[19], and, in fact, relative to other Eurasians, the Kalash show reduced $\hat{N}_e(\rho)$ and heterozygosity, consistent with a long-term history of isolation. Among Africans, the San, a hunter-gatherer group, have the lowest $\hat{N}_e(\rho)$, although their microsatellite diversity is not substantially reduced. Because demographic events such as bottlenecks can have a larger effect on LD than on genetic diversity[4], this result may suggest a past bottleneck in the San.

Worldwide portability of the HapMap

An important motivation for studying LD is to inform the design of association mapping studies. The HapMap study provides dense, genome-wide information on LD in only a few populations, and it is of interest to ask what fraction of worldwide haplotype variation is captured by its samples. To address this issue, we considered windows of fixed size within the core regions, and for each population and each window position, we computed the fraction of common haplotypes (frequency >10%) that were also common in the most similar HapMap population (**Fig. 6**). For example, averaging across populations, 83% of common haplotypes are also common in the most similar HapMap population (at a window size of 20 kb).

Comparing HapMap Asians (CHB+JPT) to our new data, we find that as expected, haplotype sharing is highest with our Han and Japanese samples. Similarly, our Yoruba sample is the population that has the highest sharing with the HapMap YRI. Although there is no direct analog of the HapMap CEU population in our data set, sharing with HapMap CEU is high with several of our European populations

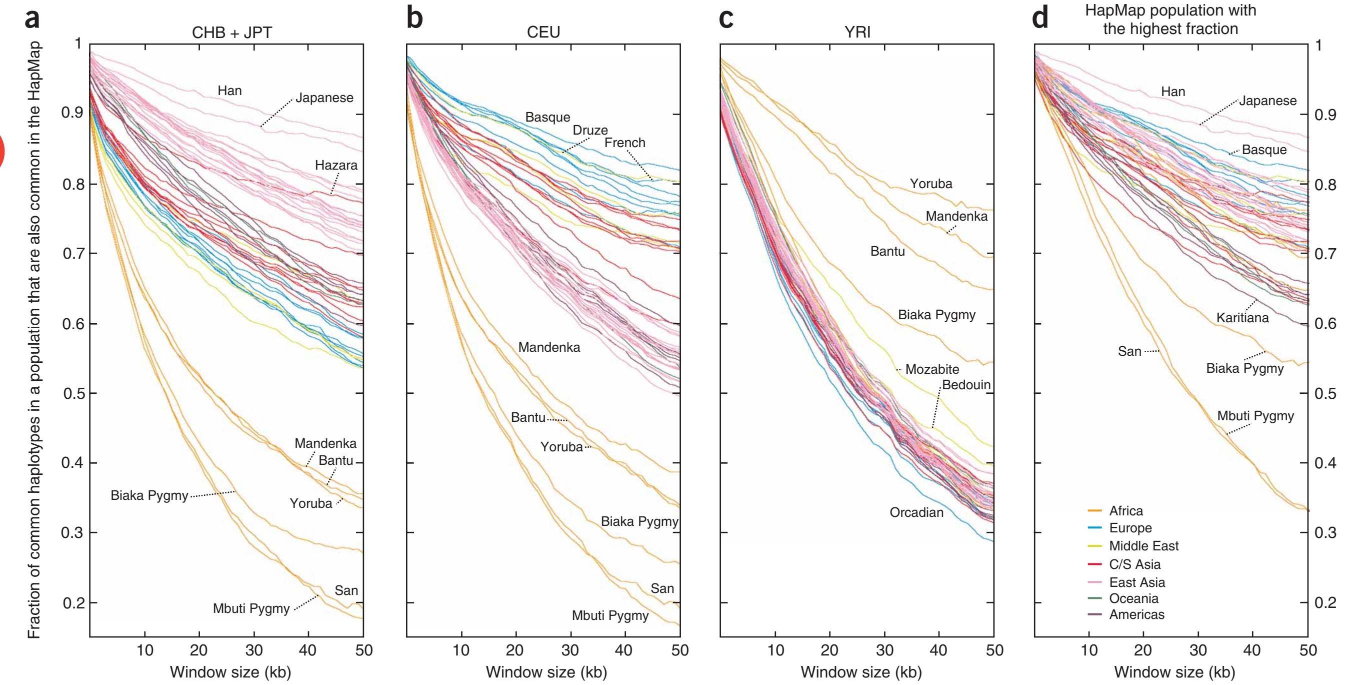

Figure 6 The fraction of common haplotypes in individual populations that are also common in the HapMap. For each plot, we used haplotypes based on the SNPs that overlap between Phase II of the HapMap and our autosomal core regions and averaged over all windows of a given length. Each curve shows the fraction of the common haplotypes of a population (those with >10% frequency) that are also common in a HapMap sample. The HapMap samples are taken as (**a**) CHB+JPT, (**b**) CEU, (**c**) YRI and (**d**) the maximum for each population of the values in **a**, **b** and **c**, taken pointwise.

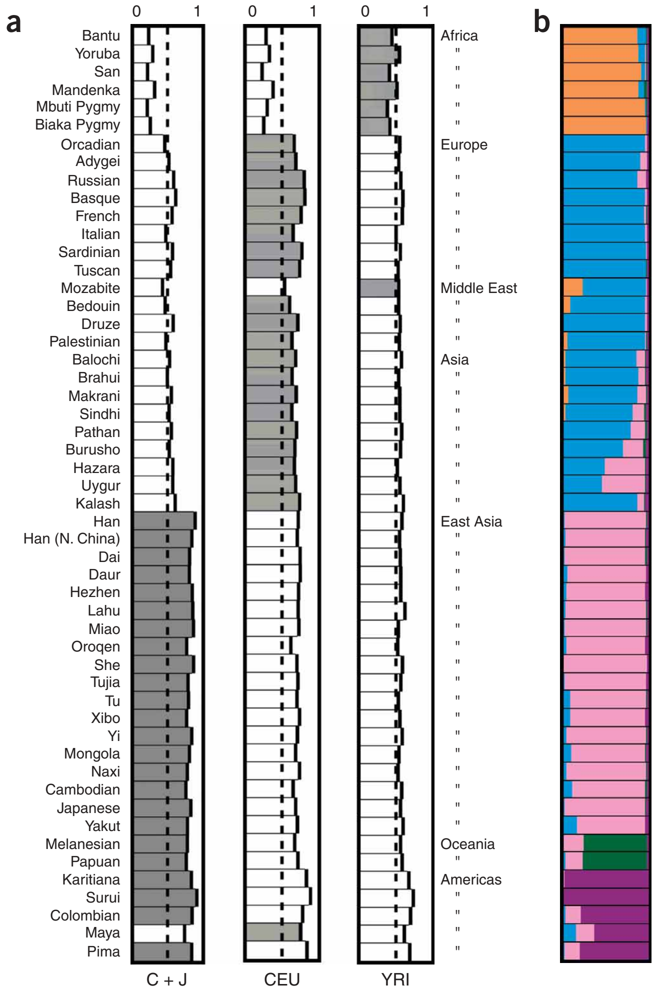

Figure 7 Portability of tag SNPs chosen using the HapMap. (**a**) These columns show for each of 52 populations the proportion of polymorphic non-tag SNPs that have $r^2 > 0.85$ with at least one tag SNP. Tag SNPs were chosen separately from each of the three HapMap groups (CHB+JPT, CEU, YRI; shown separately in each column) to maximize the number of SNPs tagged within those groups, subject to an average tag SNP density of 1 per 7.8 kb (corresponding to a genome-wide screen of about 400,000 SNPs). For each population, the gray bar indicates which tag SNP set is best. Vertical dashed lines indicate 50% tag portability. (**b**) Estimated worldwide population structure based on microsatellite data from the individuals in our study. For each population, the horizontal bar is split into colored segments with lengths proportional to the estimated membership of the population in each of five clusters identified in previous analysis[41] by the program STRUCTURE[47].

and is highest for French and Basques. The patterns of haplotype sharing do not depend noticeably on window size and are also robust to the definition of 'common' and to variation in sample size (**Supplementary Note**). Note that owing to sampling variation, none of the populations reach 100% sharing with any of the HapMap samples.

Overall, we found that sharing with the HapMap is high in European and East Asian populations, a result consistent with previous work[13–15,17], but that sharing for most other populations was lower. Most notably, for larger window sizes, the common haplotypes in African hunter-gatherer groups differed markedly from those found in any of the HapMap populations. Other populations that have lower sharing with the HapMap are those that are geographically distant from all of the HapMap populations, including populations from Oceania (Papuan, Melanesian) and the Americas (Colombian, Karitiana, Surui), and genetically distinctive populations from Central Asia (Kalash and Uygur) and North Africa (Mozabite).

A related matter is whether tag SNP sets designed using HapMap data will perform well in association studies in diverse populations[14,23]. To investigate this issue, we designed three tag SNP panels based on the HapMap samples (CHB+JPT, CEU and YRI). As the use of more complex multimarker approaches does not tend to change the qualitative order across populations of results concerning tag SNPs[39], we followed a popular approach for choosing tag SNPs based on pairwise r^2 (ref. 40). For each HapMap sample, we identified the set of 333 tag SNPs that maximized the total fraction of SNPs tagged at $r^2 > 0.85$ in that HapMap sample. As all three tag SNP panels used the same density of markers (corresponding to a genome-wide panel of ~400,000 markers), the performances of all three panels are directly comparable.

Our results show considerably improved tagging for tag SNP panels based on the 'nearest' HapMap population in comparison with those based on either of the other two HapMap samples (**Fig. 7**). The groupings that describe the HapMap population with the best tag SNP performance closely follow clusters of human population structure estimated from microsatellites genotyped in the same individuals[19,41]: sub-Saharan Africans are best tagged by YRI, populations from Europe, the Middle East and Central and South Asia by CEU and populations from East Asia, Oceania and the Americas by CHB+JPT. Of the 52 populations, only the Maya and Mozabite populations do not follow this trend. Maya are tagged best by CEU, perhaps reflecting partial European admixture, and Mozabites, who have a noticeable proportion of sub-Saharan ancestry, are best tagged by YRI[19].

Within geographic regions, there was generally little reduction of the proportion of variation tagged (PVT) in transferring HapMap tag SNPs to a different population. For example, in East Asia, PVT for our Japanese and Han Chinese samples was comparable to that of other East Asian populations[14,16,18]. The Yoruba were tagged by the HapMap YRI sample slightly better than were other African populations, though the difference is modest, perhaps because the Yoruba have low LD even among Africans (**Figs. 4** and **8**).

The portability of tags to a population was affected much more by its level of LD than by its proximity to the donor HapMap population (**Fig. 8**). Proximity predicted which HapMap sample provided the highest PVT, as 48 of 52 populations were best tagged by the HapMap sample to which they had the lowest F_{ST} genetic distance (the exceptions being Hazara, Maya, Mozabite and Uygur). Additionally, in a linear regression controlling for r^2 decay distance, we found a significant contribution to PVT for F_{ST} distance to the nearest HapMap sample ($P < 10^{-4}$). However, the Spearman correlation of PVT with r^2 decay distance (0.72) was considerably greater in magnitude than the corresponding correlation of PVT with F_{ST} (–0.16). Thus, portability of HapMap tag SNPs might be high for high-LD populations that are genetically distant from the HapMap populations and low for low-LD populations whose genetic similarity to one of the HapMap groups is greater. Owing to their low levels of LD, African populations are the most difficult populations to tag (even though an African population is included in the HapMap), and Native Americans, with their high levels of LD, are among the easiest, despite the fact that they have relatively low haplotype sharing with the HapMap populations.

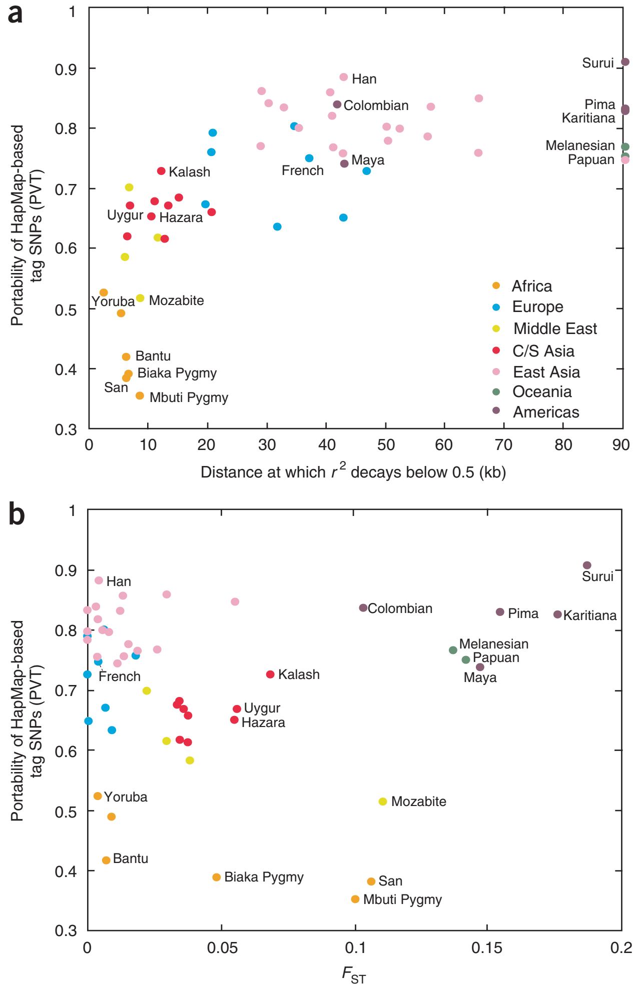

Figure 8 The determinants of portability of HapMap tag SNPs. (**a**) The relationship between tag portability and the distance at which the r^2 measure of linkage disequilibrium decays below 0.5. (**b**) The relationship between tag portability and F_{ST} genetic distance to the HapMap population that produces the highest tag portability. For each population, tag portability is computed as the maximum of the three PVT values in **Figure 7a**. Spearman correlation coefficient equals 0.72 between tag portability and r^2-decay distance and equals –0.16 between tag portability and F_{ST}.

By considering additional schemes for tag SNPs, it is potentially possible to obtain more effective tagging of some of the populations that are genetically most distant from HapMap groups, such as African hunter-gatherers, Native Americans and populations genetically intermediate between HapMap samples. For example, when we use four additional tag SNP panels, relying on one of the three pairs of HapMap populations or on the combination of all three groups, tag portability improves for several of the genetically intermediate populations (**Supplementary Fig. 1** online). Hazara, Kalash and Uygur show improved tagging when tag SNPs are chosen based on the combination of the HapMap CEU, CHB and JPT samples, whereas Bedouin and Mozabite have higher PVT when using the combination of CEU and YRI to choose tag SNPs.

One problem that affects tag portability is that HapMap tag SNPs are frequently monomorphic in these most distant populations. For example, it might have been expected that tagging would be essentially perfect in the American populations, where LD is so extensive. However, we found that 47% of the tag SNPs designed based on East Asians have minor allele frequency (MAF) <0.1 in Surui, compared with 13% in our Han sample. A similar problem occurs in Africa, where 47% of HapMap YRI tag SNPs have MAF <0.1 in San, compared with 16% in our sample of Yoruba. Some improvement in tag portability can be obtained by careful consideration of SNP ascertainment procedures, which affect proportions of monomorphic SNPs and low-frequency alleles. For example, tag portability based on the HapMap CEU data is noticeably higher for SNPs in our data that were ascertained in the multiethnic panel of ref. 11 than it is for other SNPs (**Supplementary Fig. 2** online), as such multiethnically ascertained tag SNPs are likely to have generally higher minor allele frequencies in most populations. Indeed, for SNPs with average MAF >0.2 across populations, these multiethnically ascertained SNPs have higher frequencies than all remaining SNPs (Wilcoxon two-sample $P = 0.004$). At the same time, tag portability based on YRI is lower for SNPs from the multiethnic panel (**Supplementary Fig. 3** online): owing to the higher level of genetic variation in Africa, tag SNPs chosen from such a panel (in comparison with those from, for example, a European-ascertained panel) are likely to include many polymorphisms private or nearly private to Africa that will perform poorly as tags elsewhere. This claim is supported by the fact that the fraction of SNPs with a higher MAF (MAF > 0.2) among pooled Africans and lower MAF (MAF ≤ 0.1) among pooled non-Africans is considerably larger (11.1%) for the SNPs from ref. 11 than for the remaining SNPs (5.2%).

DISCUSSION

Our study provides the first data on global haplotype variation across multiple megabases of sequence in multiple genomic regions. We find that although SNP ascertainment distorts patterns of variation at the single-SNP level, patterns of haplotype diversity at larger scales show greatly reduced evidence of bias. Global patterns of haplotype variation accord well with a population model in which genetic variation passed through a serial dilution process as humans spread progressively further from our African source. In the future, autosomal haplotype data will surely provide an important tool for unraveling the history of human migrations.

A related application of haplotype data is identifying genomic regions that have been targets of natural selection[2,6,7]. Such tests, which attempt to identify unusually long common haplotypes likely to have reached high frequencies as a result of recent natural selection, have reduced power in populations that are strongly bottlenecked[7], owing to the property that common haplotypes are often extremely long even in neutral regions. Hence, these tests will unfortunately be less powerful in the populations that have extreme long-range LD (especially in Oceania and the Americas).

Our results extend recent work showing that the HapMap is a valuable resource for designing association studies in diverse human populations. We find that most populations have extensive haplotype sharing with at least one HapMap population. Some exceptions exist, however, especially in Africa; as our sample includes only six African populations, considerable scope exists for improved understanding of the extent of LD patterns and tag SNP portability in Africa.

We observe that there is a clear advantage to developing distinct sets of tag SNP panels based on each of the three HapMap groups rather than having a single panel to be used worldwide. For populations where tag portability is lower, several strategies, including use of SNPs ascertained in a more optimal manner or consideration of additional tag SNP designs, may potentially be used to provide improvements. However, for the populations we have studied, neither ascertainment nor the use of additional tag SNP panels affects the fact that tag portability is determined to a greater extent by levels of LD than by proximity to the nearest HapMap population (**Supplementary Figs. 4** and **5** online). Thus, although such strategies (as well as inclusion of additional populations in studies of similar magnitude to the HapMap study) will improve tag portability in low-LD populations, these populations will likely continue to require an increased density of tag SNPs to achieve the same proportion of variation tagged as can be obtained with fewer SNPs in higher-LD populations.

METHODS

Samples. We genotyped 1,048 distinct individuals and four duplicate DNA samples for 3,024 SNPs. The individuals were drawn from the HGDP-CEPH Human Genome Diversity Cell Line Panel[24,25], and the set of 1,048 distinct individuals was the same as a previously used collection, the H1048 data set[26,41]. We removed 22 individuals with large amounts of missing data and additional individuals who were first- or second-degree relatives of others in the data set[41], obtaining a final data set of 927 individuals. The smallest sample size from any major geographic region (regions defined as Africa, Europe, the Middle East, Central and South Asia, East Asia, Oceania and the Americas[19]) is 27 (from Oceania); hence, some analyses that are sensitive to variable sample sizes use subsamples of 54 chromosomes from each continent. Further details on many aspects of the methods are available in **Supplementary Methods**.

Choice of genomic regions and SNPs. The 36 genomic regions include 16 regions scattered across the autosomes, 16 from chromosome 21 and four from the non-pseudoautosomal X chromosome. The first set of 16 autosomal regions included eight that fell within regions studied in detail by the ENCODE Project[2].

For both sets of 16 regions, we aimed to include one region to represent each category in a 4 × 4 table, each of whose cells corresponds to a quartile of the genomic distribution of recombination rate (based on the deCODE map[27]) and a quartile of the distribution of gene density (from the UCSC genome browser). We sampled the four X-chromosomal regions to represent each recombination rate quartile. For the non-ENCODE autosomal regions, we filled out the rest of the 4 × 4 table as well as possible given the properties of the ENCODE loci. Apart from the eight predetermined ENCODE regions, all regions were sampled at random subject to these criteria (**Supplementary Methods**).

Each region was studied using 84 SNPs, including a 'core' of 60 SNPs spaced at an average of 1.5 kb apart, flanked by two sets of 12 SNPs at 10 kb average spacing. Thus, each region covered ~330 kb. For the chromosome 21 regions, all SNPs were selected from those discovered in a uniform screen of 20 chromosomes of multiethnic origin[11]. For the other regions, SNPs were chosen in the following order of priority: (i) Perlegen SNPs (most of which were discovered in a multiethnic panel[11]); (ii) HapMap Phase I SNPs[2]; (iii) dbSNP 'two-hit' SNPs.

Genotype data. Genotyping was performed using the Illumina BeadLab 1000 platform. Of the initial 3,024 SNPs, 190 were excluded: 115 failed genotyping, 50 were monomorphic, 20 SNPs failed Hardy-Weinberg checks in unstructured populations and 11 had >10% missing data (six of these also failed Hardy-Weinberg). These exclusions yielded a final cleaned data set of 2,834 non-monomorphic SNPs typed in 927 individuals. The missing data rate in the cleaned data was 0.1%. The genotype error rate in the cleaned data was estimated at $\sim 2 \times 10^{-4}$ based on comparisons of three duplicate samples and on mendelian error checks between related individuals in the full sample. Details of the SNPs used are available in **Supplementary Table 1** online.

For some analyses, we also make use of the Phase II HapMap data in the regions studied here. For those analyses, we only use SNPs from our data set that are also in the HapMap data set; there are 2,078 such SNPs.

Haplotype phasing. Haplotype phasing was performed using fastPHASE v. 0.9 (ref. 28). A pre-publication release of the program was supplied by P. Scheet. We chose to use fastPHASE for several reasons. The related program PHASE was found to have the best performance in a recent comparison of phasing methods[42]. fastPHASE achieves similar accuracy to PHASE but at much greater speed in large samples such as ours[42]. fastPHASE has the added benefit of allowing separate parameters for each population, a feature that is attractive for our data set and that leads to improved performance (**Supplementary Note**). We also used fastPHASE to impute the 0.1% of missing genotypes so that the analyzed data set does not contain any missing values (except where otherwise specified).

In order to phase the data, there were a number of choices that had to be made, including how to label and group the population samples and the number of haplotype clusters to assume. Our main approach was to perform a series of fastPHASE runs in which 10% of the genotype data were hidden at random. We computed the error rates in the genotypes imputed by fastPHASE and then chose parameter combinations that minimized the overall error rate. This is essentially the approach suggested by ref. 28. We also checked our results using haplotype reconstructions from the set of parent-offspring pairs available in our full data. Finally, for HapMap data in our regions, we compared values of the LD statistic r^2 estimated from data phased by fastPHASE (parents only) to the same data phased using trio information, and we found almost perfect agreement (**Supplementary Note**).

Haplotype visualization. In order to visualize the haplotypes in each genomic region, we developed the following algorithm. Our method was conceptually motivated in part by the model developed in ref. 28, but it differs in being less model-based.

We start by identifying, for each of seven major geographic regions, the single most common haplotype spanning a genomic region. These seven haplotypes will be called the 'template' haplotypes. (The assignment of populations to seven geographic regions is the same as that used in ref. 19). Occasionally, the most common haplotype is identical for two or more geographic regions. In that case, we take as one of the templates the second-place haplotype that is most frequent within its region. Each template is assigned a distinct color. Next, we color each observed haplotype as a mosaic of the seven templates, requiring exact matches between the observed haplotype and the template that is being copied. Roughly speaking, the coloring minimizes the number of switches between templates (see **Supplementary Methods** for details). Rare alleles not found on any template were dropped from the analysis in the version shown in **Figure 1**.

Finally, for each population shown in **Figure 1**, 20 haplotypes were sampled without replacement for plotting. Surui and Colombians have <20 total haplotypes, so for these populations, all haplotypes are shown. For clarity, the plotted chromosomes are sorted by the coloring in the center of the region.

Estimation of recombination rates. The reversible jump Markov Chain Monte Carlo method LDhat v2.0 (ref. 8) was used to estimate maps of the population-genetic recombination rate ρ from the SNP data for each genomic region. The program requires a choice of value for a smoothing parameter that determines the penalty for introducing a new zone with distinct recombination rate; following ref. 8, this quantity was set to 20. As the LDhat method uses unphased genotype data rather than phased data, this analysis used the unphased genotypes and did not use the data version with missing data imputed. For all results, the mean value of the recombination rate was obtained over 10^6 iterations of the MCMC (with a thinning interval of 2,000), following a burn-in of 10^5 iterations.

We estimated ρ/kb for each population and genomic region by taking the mean map length for each genomic region and each population and dividing by the total length in kb of the region in that population. The average population-genetic recombination rate per kb in region *reg* and population *pop* is denoted $\rho_{pop,reg}$. Genomic region 1 is distal to the first deCODE marker on chromosome 1p, so the pedigree-based recombination rate is unreliable. This genomic region was excluded from the analysis. The effective size of a population N_{pop} was estimated from the population-genetic recombination rates by a model that allowed for an error in the pedigree-based estimate of the recombination rate (the most extreme example of which is the excluded genomic region 1). In this model, we assume

$$\rho_{pop,reg} = 4N_{pop}(d_{reg}+b_{reg}) + \varepsilon_{pop,reg} \tag{1}$$

where d_{reg} is deCODE's pedigree-based estimate of recombination rate per kb for the region, and b_{reg} is the 'error' in the pedigree rate estimate for a region when used at a local scale. To constrain this model, we required that the sum across regions of the values of b_{reg} be 0. The model was fitted to minimize $\sum_{pop,reg} \varepsilon^2_{pop,reg}$ by a hill-climbing algorithm, where the sum ranges over 52 × 31 (52 populations and 31 autosomal regions, excluding region 1) population-region combinations. Further analyses regarding the robustness of the results and the use of different estimation methods can be found in the **Supplementary Note**.

Definition of haplotypes for computations of haplotype summary statistics. We computed haplotype summary statistics based on haplotypes within genomic windows of a specified length *w*. For these analyses, the entire window was required to lie within our genomic 'core' regions. For each SNP, we defined a haplotype locus that extended from the position of the SNP (a) along the chromosome to the SNP position plus the window size ($a + w$). The haplotype of a particular phased chromosome was then specified by the set of allele states at all SNPs located between a and $a + w$ (including position a but excluding position $a + w$). Haplotypes were considered to be identical if and only if they had the same genotype for all SNPs with position in $[a,a + w)$. For each value of w, except for the ϕ statistic (defined below), the summary statistics presented are means over all haplotype loci with the given window size. The ϕ statistic was computed by averaging across haplotype loci within each of the genomic core regions and was then averaged across regions. The computations of ϕ also differed in that estimates involving the HapMap excluded from consideration SNPs not among the 2,078 in our data set that were contained in the HapMap.

Numbers of distinct haplotypes and private haplotypes. Because sample sizes differed across geographic regions, this computation used the rarefaction approach[43,44] to obtain sample-size corrected estimates of the numbers of distinct haplotypes and private haplotypes. For a given summary statistic, if a sample has size N, the idea of this approach is to choose a value for $g \le N$ and to compute the mean value of the statistic expected across all possible subsamples of size g from the original sample of size N. Because the smallest sample size among the seven geographic regions equaled 54 chromosomes (Oceania), we used $g = 54$ for all computations.

Following the notation in ref. 44, for a given haplotype-locus, let N_j represent the number of haplotypes sampled in the jth geographic region, let N_{ij} represent the number of haplotypes of type i in the jth geographic region and let m represent the total number of distinct haplotypes observed ($N_j = \sum_{i=1}^{m} N_{ij}$). Let P_{ijg} be the probability of observing at least one haplotype of type i in a sample of size g haplotypes from geographic region j, and let Q_{ijg} be the probability of not observing any haplotypes of type i in a sample of size g from region j. For region j, the expected number of distinct haplotypes that will be observed in a sample of size g, or $\alpha_g^{(j)}$, equals[43]

$$\alpha_g^{(j)} = \sum_{i=1}^{m} P_{ijg} \tag{2}$$

where $P_{ijg} = 1 - Q_{ijg}$ and

$$Q_{ijg} = \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}}$$

Let J equal the number of geographic regions under consideration ($J = 7$ in our case). The number of distinct haplotypes private to region j expected in a sample that contains g haplotypes from each of the J geographic regions, or $\pi_g^{(j)}$, equals[44]

$$\pi_g^{(j)} = \sum_{i=1}^{m}\left(P_{ijg} \prod_{\substack{j'=1 \\ j' \neq j}}^{J} Q_{ij'g} \right) \tag{3}$$

Pairwise haplotype sharing of geographic regions. To compute the fraction of distinct haplotypes that are shared between two geographic regions, say regions j and j', for each haplotype-locus, we first computed the numbers of distinct haplotypes and the numbers of private haplotypes for the particular pair of geographic regions. This computation used the rarefaction approach with $g = 54$ and $J = 2$, and in this analysis, private haplotypes for population j refer to those not found in j'.

The expected number of distinct haplotypes that will be found in a sample of size g from geographic region j that will also be found in a sample of size g from region j' is then equal to the difference between the expected number of distinct haplotypes in region j and the expected number of private haplotypes in region j, or $S_g^{(j)} = \alpha_g^{(j)} - \pi_g^{(j)}$. Thus, we can view the following statistic as an estimator of the fraction of distinct haplotypes observed in a sample of size g from region j that will also be observed in a sample of size g from region j':

$$z_g^{(j)} = \frac{\alpha_g^{(j)} - \pi_g^{(j)}}{\alpha_g^{(j)}} \tag{4}$$

Fraction of common haplotypes that are also common in the HapMap. Suppose that haplotypes with frequency greater than a cutoff value c are considered to be common haplotypes. Making a slight modification to Equation (2), the expected number of distinct common haplotypes at a given haplotype locus that will be observed in a sample of size g from population j, or $\alpha_{g,c}^{(j)}$, equals

$$\alpha_{g,c}^{(j)} = \sum_{i=1}^{m} P_{ijg}\chi_{\{f_{ij}>c\}} \tag{5}$$

where the indicator function $\chi_{\{f_{ij}>c\}} = 1$ if the frequency f_{ij} of haplotype i in population j exceeds the cutoff c, and equals 0 otherwise. The expected number of distinct common haplotypes $\gamma_{g,c,j'}^{(j)}$ in a sample of size g from population j that also have the property of being common in population j' is given by the following expression:

$$\gamma_{g,c,j'}^{(j)} = \sum_{i=1}^{m} P_{ijg}\chi_{\{f_{ij}>c\}}\chi_{\{f_{ij'}>c\}} \tag{6}$$

Therefore, for the ordered pair of populations (j,j'), for a sample of size g from population j, we used the following expression to determine the average fraction of common haplotypes in population j that were also common in population j':

$$\phi_{g,c,j'}^{(j)} = \frac{\gamma_{g,c,j'}^{(j)}}{\alpha_{g,c}^{(j)}} \tag{7}$$

Heterozygosity. For a given haplotype locus, the heterozygosity of the haplotypes of a particular population, H_j, was computed as follows[45]:

$$H_j = \frac{N_j}{N_j - 1}\left[1 - \sum_{i=1}^{m}\left(\frac{N_{ij}}{N_j}\right)^2\right] \tag{8}$$

In this equation, N_j equals the number of haplotypes in population j, and N_{ij} is the number of haplotypes of type i in population j. Microsatellite heterozygosity for each population was computed as the average across 783 loci[41], pooling the two Bantu groups into a single population.

Tag SNP analysis. For analysis of tag SNP portability, we used overlapping SNPs with the HapMap Phase II data (release 19) for 29 regions (X-chromosomal regions 23–26, and regions 30–32 with gaps were excluded). Of the SNPs typed in the current study, 2,078 are present in the phase II HapMap. The HapMap data were phased with the same protocol used to phase the HGDP-CEPH genotypes; phasing and analysis were performed together with the parental genotypes only in the case of the CEU and YRI samples. CHB and JPT samples were combined into one 90-sample population for phasing and all subsequent analyses. All SNPs were used to assist the phasing, but the tag SNP analysis described below was limited to our high-density 'core' regions. The number of core SNPs present in the HapMap ranges from 27 to 58 per region, of a total possible 60. For each HapMap population separately (CEU, YRI, CHB+JPT), we selected 333 LD-based tag SNPs with the goal of maximizing the total number of SNPs that have $r^2 > 0.85$ with at least one tag SNP; our algorithm was roughly as described by ref. 40.

The central aim of these analyses was to measure the amount of variation indirectly assayed in one population (the 'target') by typing genetic markers selected in another (the 'donor'). We define a simple metric called the PVT (proportion of variation tagged) as our measure of tag portability:

$$PVT = \frac{\sum_{r=1}^{n} t_r - s_r}{\sum_{r=1}^{n} p_r - s_r}$$

where the number of tag SNPs within genomic region r that are polymorphic in the target population is denoted s_r, the number of SNPs 'tagged' (which includes tag SNPs) is t_r, the total number of polymorphic SNPs within region r is p_r and the total number of genomic regions is $n = 29$. Because sample sizes vary across populations, it was important to control for the effect of sample size in our analyses. A linear relationship between PVT and sample size (in the relevant range) was observed in simulations based on subsampling from large populations (**Supplementary Note**). Hence, all PVT scores were adjusted to the mean sample size across HGDP-CEPH populations (36 chromosomes) using the following procedure. For populations with more than 36 chromosomes, we corrected the PVT score empirically by resampling 36 chromosomes from the population 30 times and averaging PVT scores across these subsamples. For populations with fewer than 36 chromosomes, we used a regression adjustment to 'bring them up' to 36 (**Supplementary Note**).

PVT was measured for each population using each of the HapMap samples, as well as using combinations of the HapMap samples. The analysis was also performed separately using only the chromosome 21 SNPs overlapping with the HapMap and only the overlapping SNPs not on chromosome 21.

Determinants of tag SNP portability. To compare the influence of levels of LD and relatedness of the most similar HapMap population on portability of tag SNPs, we analyzed the relationships between tag portability as measured by PVT and each of these two variables. For each population, tag portability was computed as the maximum of the three PVT estimates based on the three HapMap samples (**Fig. 7a**). F_{ST} was computed between each population and the HapMap sample that produced the highest portability. This computation was performed for each SNP in a core region (excluding regions 23–26 and 30–32) among those SNPs overlapping with the HapMap data, using equation 5.3 of ref. 46.

The distance at which the LD statistic r^2 decayed below a specified cutoff c was obtained by considering all pairs of SNPs in the same core region among SNPs overlapping with the HapMap data (excluding regions 23–26 and 30–32), subject to both SNPs having MAF $\geq m$ in the population under consideration. Given c and a percentage p, increasing the distance d continuously, all points were located where the percentage of SNP pairs separated by distance $\leq d$ that had $r^2 > c$ crossed from being greater than p to less than p. The desired distance d^* was obtained at the largest of these crossing points as the smallest distance d^* so that $<p$% of SNP pairs separated by distance $\leq d^*$ had $r^2 > c$. To correct for the influence of sample size, for each population, r^2 for each SNP was obtained as the average across 1,000 resamples (without replacement) of 12 chromosomes. In populations where r^2 did not decay below c at the distances we investigated, d^* was set to the approximate average length of the core regions, or 90 kb. The computations shown in **Figure 8** used $m = 0.2$, $c = 0.5$ and $p = 50\%$.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. DeYoung and the Southern California Genotyping Consortium for genotyping, P. Scheet for providing prepublication access to fastPHASE, C. Shtir for assistance and M. Przeworski for comments. This work was supported by grants from the Burroughs Wellcome Fund (J.K.P., N.A.R.), the Sloan Foundation (J.D.W., J.K.P., N.A.R.), the Packard Foundation (J.K.P.) and the US National Science Foundation (J.D.W., N.A.R.).

AUTHOR CONTRIBUTIONS

J.D.W., N.A.R., and J.K.P. conceived and jointly supervised the study, D.F.C. performed the SNP design, N.A.R. and J.K.P. cleaned the data, and J.K.P. performed the phasing. All authors analyzed the data, with the following primary contributions: haplotype visualization, X.W.; haplotype diversity statistics and haplotype sharing with the HapMap, M.J.; recombination rate estimation, G.C.; tag SNP portability, D.F.C.; determinants of portability, M.J. and D.F.C. The supplementary information was written by D.F.C, G.C., M.J., N.A.R. and J.K.P., and the paper was written primarily by J.K.P. and N.A.R., with assistance from all other authors.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Zondervan, K.T. & Cardon, L.R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89–100 (2004).
2. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1319 (2005).
3. Tishkoff, S.A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
4. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
5. Plagnol, V. & Wall, J.D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, 972–979 (2006).
6. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
7. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
8. McVean, G.A.T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
9. Ptak, S.E. *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.* **37**, 429–434 (2005).
10. Fearnhead, P. & Smith, N.G. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.* **77**, 781–794 (2005).
11. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
12. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
13. De Bakker, P.I.W., Graham, R.R., Altshuler, D., Henderson, B.E. & Haiman, C.A. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pac. Symp. Biocomput.* **11**, 478–486 (2006).
14. Huang, W. *et al.* Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc. Natl. Acad. Sci. USA* **103**, 1418–1421 (2006).
15. Montpetit, A. *et al.* An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet.* **2**, e27 (2006).
16. Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.* **38**, 556–560 (2006).
17. Willer, C.J. *et al.* Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol.* **30**, 180–190 (2006).
18. Yoo, Y.K. *et al.* Fine-scale map of Encyclopedia of DNA Elements regions in the Korean population. *Genetics* **174**, 491–497 (2006).
19. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
20. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).
21. Sawyer, S.L. *et al.* Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.* **13**, 677–686 (2005).
22. Bonnen, P.E. *et al.* Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat. Genet.* **38**, 214–217 (2006).
23. Gonzalez-Neira, A. *et al.* The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* **16**, 323–330 (2006).
24. Cann, H.M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
25. Cavalli-Sforza, L.L. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).
26. Rosenberg, N.A. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet. published online* 29 March 2006 (doi:10.1111/j.1469-1809.2006.00285.x).
27. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).

28. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
29. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* **102**, 15942–15947 (2005).
30. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
31. Mountain, J.L. & Cavalli-Sforza, L.L. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc. Natl. Acad. Sci. USA* **91**, 6515–6519 (1994).
32. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genomics* **1**, 218–224 (2004).
33. Rogers, A.R. & Jorde, L.B. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet.* **58**, 1033–1041 (1996).
34. Bowcock, A.M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
35. Crawford, D.C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**, 610–622 (2004).
36. Stephens, J.C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
37. Nielsen, R., Hubisz, M.J. & Clark, A.G. Reconstructing the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168**, 2373–2382 (2004).
38. Winckler, W. *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111 (2005).
39. De Bakker, P.I.W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
40. Carlson, C.S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2005).
41. Rosenberg, N.A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, 660–671 (2005).
42. Marchini, J. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
43. Hurlbert, S.H. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577–586 (1971).
44. Kalinowski, S.T. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conserv. Genet.* **5**, 539–543 (2004).
45. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
46. Weir, B.S. *Genetic Data Analysis II* (Sinauer, Sunderland, Massachusetts, 1996).
47. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).