Supplementary Methods for “A worldwide survey of haplotype variation and linkage disequilibrium in the human genome”

This supplement contains expanded versions of the methods for topics not described in full detail in the Methods section of the article. Additional results about phasing, recombination rate estimation, haplotype sharing, and tag SNPs can be found in the **Supplementary Note**.

Contents

1 Study design

The initial study design involved the genotyping of 1048 distinct individuals and four duplicate DNA samples for 3024 single-nucleotide polymorphisms (SNPs). The individuals were drawn from the HGDP-CEPH Human Genome Diversity Cell Line Panel[1,2] (the "HGDP" in many places in this supplement), and the set of 1048 distinct individuals was the same as a previously used collection, the H1048 dataset[3,4].

We designed a modular sampling strategy for selecting SNPs, with the aim of providing a representative view of worldwide patterns of long- and short-range linkage disequilibrium (LD) across the human genome. Each 84-SNP module consists of a "core" region of 60 SNPs spaced at an average of 1.5 kb apart, flanked by two regions of 12 SNPs at 10 kb average spacing. Thus, each module covers an average of 330 kb. Thirty-six modules were selected across the genome: 16 from chromosome 21, 8 from ENCODE regions, 8 from random autosomal regions (not on chromosome 21), and 4 from random regions on the X chromosome.

1.1 Choice of genomic regions

The process for selecting the 36 genomic regions was predicated on a 16-cell block design, in which we would try to sample as evenly as possible from each quartile of the genomic distribution of recombination rate and each quartile of the distribution of gene density. The deCODE genetic map[5] and the "Known Genes" track of the UCSC genome browser (October, 2004) were used to generate estimates of recombination rate and gene density within a grid of 500 kb intervals across the entire genome. All analyses were conducted using the latest public genome assembly (NCBI 35, UCSC hg17). Prior to calculation of recombination rate, we inspected the deCODE data for inconsistencies between the physical and genetic map orders of all mappable markers. In the case of a single-marker conflict we simply discarded the marker; when multiple markers were involved in map inconsistencies within a small region, we discarded the smallest number of makers necessary to create a consistent map. These map inconsistencies led to the removal of 42 markers from the deCODE map. Gene density was estimated by counting the number of records from "Known Genes" with a transcription start site within each 500 kb window.

Because there were many fixed elements of the study design, our goal was to select from eligible regions at random in such a way that the result would be a fairly uniform sample from all cells of our block design. The large number of regions to be selected from chromosome 21 was a particularly influential factor in the selection of regions; the recombination rate across 21 is much higher than the genome average, and the number of usable SNPs in gene-rich regions was low.

The distribution of selected autosomal genomic regions across the sampling grid is shown in Table SM.1. The genomic location and additional properties of each region are in Table SM.2, and the genomic locations are shown pictorially in Figure SM.1. For the X chromosome, we randomly selected one region from each quartile of the deCODE recombination rate.

	Recombination Rate				
Gene density	Quartile 1	Quartile 2	Quartile 3	Quartile 4	Total
Quartile 1	4	6	2	2	14
Quartile 2	0	1	2	4	7
Quartile 3	1	1	1	1	4
Quartile 4	3	1	3	0	7
Total	8	9	8	7	32

Table SM.1: Sampling distribution of the autosomal genomic regions used in this study. Columns are quartiles of genomic recombination rate, and rows are quartiles of genomic gene density.

Chromosome	Start (bp)	End (bp)	Region number	Number of genes	cM/Mb
Random autosomal regions					
1	3000000	3500000	1	4	1.11
2	234750000	235250000	28	2	2.49
3	55500000	56000000	2	3	1.9
9	12500000	13000000	3	3	1.37
9	127000000	127500000	34	19	1.19
12	71000000	71500000	4	0	1.40
18	19500000	20000000	5	22	1.37
22	25500000	26000000	6	0	4.77
ENCODE regions					
2	51570356	52070355	27	0	0.84
4	118604259	119104258	29	0	0.49
7	26730761	27230760	30	14	0.75
7	89428340	89928339	31	12	0.19
7	126174898	126672039	32	1	0.72
8	118882221	119382220	33	0	0.41
12	38626477	39126476	35	3	0.16
18	23719232	24219731	36	1	0.93
Chromosome 21 regions					
21	14500000	15000000	7	8	3.57
21	16500000	17000000	8	0	3.02
21	21500000	22000000	9	0	0.87
21	22500000	23000000	10	0	0.69
21	23500000	24000000	11	0	1.63
21	24500000	25000000	12	1	1.26
21	27000000	27500000	13	3	2.74
21	28500000	29000000	14	1	0.54
21	29000000	29500000	15	13	0
21	29500000	30000000	16	5	0.50
21	30000000	30500000	17	3	0.88
21	30500000	31000000	18	22	1.34
21	36000000	36500000	19	7	1.02
21	38500000	39000000	20	9	1.44
21	40000000	40500000	21	3	3.59
21	44000000	44500000	22	22	0
X chromosome regions					
X	8000000	8500000	23	2	1.59
X	32000000	32500000	24	2	2.10
X	78500000	80000000	25	0	0
X	125500000	126000000	26	0	0.64

Table SM.2: Summary of the 36 genomic regions used in the study. Recombination rates in cM/Mb are obtained from the deCODE map.

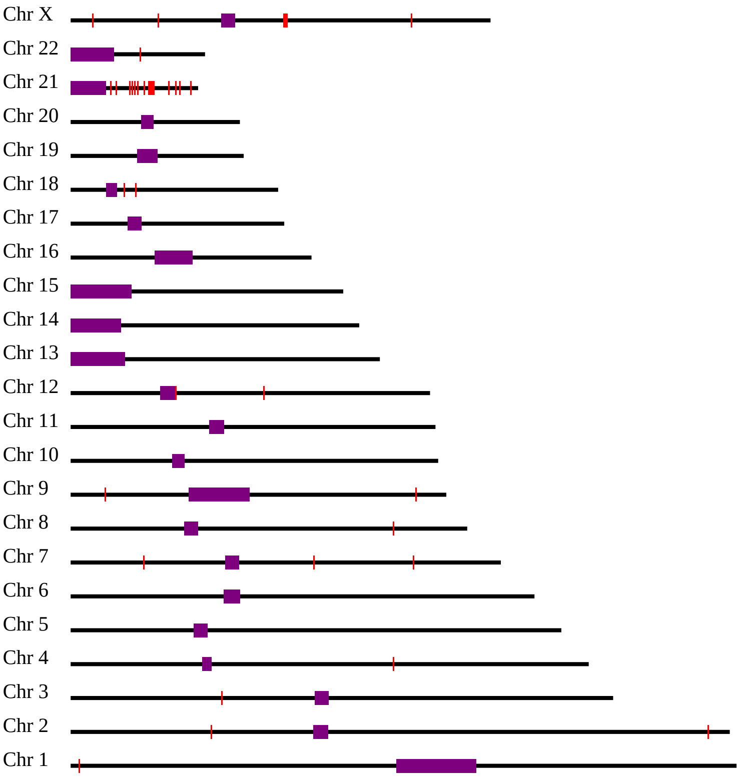


Figure SM.1: Genomic locations of 36 regions. In this schematic diagram, the physical positions of the 36 regions selected for genotyping are depicted with a vertical red line. Centromeres are depicted in purple. In order to improve visibility, regions are not drawn to scale (some regions are drawn flush against each other). All positions are based on the NCBI 35 assembly.

1.2 Choice of SNPs

For all genomic regions selected, we screened candidate SNPs on several criteria. All known tri-allelic SNPs were removed, as were SNPs whose flanking sequence in dbSNP mapped to more than one location in the genome. For the regions on chromosome 21, we considered only SNPs typed by Patil et al.[6] in their study of 20 total chromosomes of multiethnic origin. Use of these SNPs provides a relatively clear understanding of the SNP ascertainment process, as well as empirical knowledge of the phase of some haplotypes from three continental groups. For the remaining regions (ENCODE, X, and random autosomal regions), we considered two types of SNPs in dbSNP for inclusion in our study. All SNPs designated as “two-hit” SNPs were included, as were all SNPs genotyped by Perlegen Sciences in their latest large data release of 1.5 million genotypes from 71 individuals, which debuted with dbSNP build 123[7].

Quality control: A total of 26,749 candidate SNPs (for all 36 genomic regions) were submitted to Illumina for quality control. Of these SNPs, Illumina gave 18,020 of them a “high-quality” score (greater than or equal to 0.6). An additional 2334 SNPs were given a quality score in the range 0.4-0.59, and were considered usable. The remaining SNPs were unusable. Among the SNPs included finally selected for genotyping, $\sim 95\%$ had a quality score of at least 0.6, while the other $\sim 5\%$ had scores in the 0.4-0.59 range.

Prioritization: For each region not on chromosome 21, the choice of SNPs (of acceptable quality and at appropriate spacing within the region) was prioritized in the following way: (i) if available, a Perlegen SNP[7] was chosen; (ii) if no Perlegen SNPs were available, a HapMap Phase I SNP was chosen; (iii) if no Perlegen or HapMap Phase I SNPs were available, a dbSNP “two-hit” SNP was chosen.

1.3 Summary of SNP design

The final set of 3024 SNPs queued for genotyping consisted of 2433 Perlegen SNPs, with the number of Perlegen SNPs per genomic region ranging from 31 to 84. The average distance between "core" SNPs was 1499 bp, and the average spacing among flanking SNPs was about 10,100 bp. The average Illumina quality control score was 0.83. After genotyping was complete, we discovered that some SNPs from three regions (numbers 30, 31, and 32) had inadvertently been mapped to an alternate chromosome 7 genome assembly (CRA_TCAGchr7.v2). For each of these regions, there are two clusters of SNPs with roughly the same properties as intended in our original design.

2 Data

SNP genotypes were obtained for 1039 of the 1048 distinct individuals and three of the four duplicates. As a result of sample failures, the remaining individuals did not produce any genotype data. Of the 3024 SNPs, poor-quality data were obtained for 115 of them, leaving 2909 SNPs for further study.

2.1 Individuals

Of the 1039 distinct individuals, 22 produced a large amount of missing data due to a failure of a group of assays that accounted for approximately half of the SNPs. For these individuals, at least 1444 of the 2909 high-quality SNPs had missing data. These 22 individuals were excluded from consideration, leaving a collection of 1017 individuals, all of whom had relatively small amounts of missing data (all individuals had $\leq$ 93 SNPs missing among the 2909 high-quality SNPs). This set of 1017 individuals included 927 individuals from the H952 dataset[4], a collection of individuals not likely to contain any first- or second-degree relationships. Except where specified, all subsequent analyses utilized this collection of 927 unrelated individuals — 610 males and 317 females. The sample sizes in Africa, Europe, the Middle East, Central/South Asia, East Asia, Oceania and the Americas were 103, 149, 158, 199, 229, 27 and 62 individuals, respectively.

Apparent heterozygotes among males for X-chromosomal loci were recoded as missing data. With one exception, all individuals (among the 1039 genotyped) reported to be male had at most six SNPs heterozygous on the X chromosome. The only exception was that Mozabite #1263 was heterozygous for 34 SNPs, including 32 SNPs in genomic region 23, suggesting that this individual was duplicated for the entire region. All individuals previously reported to be female had at least 20 SNPs heterozygous on the X chromosome, with eight exceptions among native Americans, who were expected *a priori* to be more homozygous than other populations. Thus, the reported sex information was assumed to be accurate.

Population structure and labeling errors: We used *structure*[8] to search for potential labeling errors. Our previous work[3,9] showed that individuals in the HGDP can be clustered into groups that reflect geographic origin. To check that the clustering patterns were similar to those found previously using microsatellites, we repeated this analysis using the new SNP data. The analysis is not strictly valid with the present data, because the clustering algorithm assumes linkage equilibrium among all markers[8]. However it might be expected that with 36 independent regions there is enough independence in the data that the overall clustering results would be driven more by population structure than by spurious clustering due to LD between markers.

As can be seen in Figure SM.2, our results broadly recapitulate those from the original analysis of microsatellites[9]. To aid the comparison, the full sample of 1017 individuals with nearly complete SNP data and the same *structure* settings were used for the SNP analysis as in the original microsatellite analysis, including the use of the *structure* admixture model with correlated allele frequencies. One clear difference is that the data seem more "noisy" with the SNPs, perhaps because the LD among markers diminishes the total amount of information. Secondly, one cluster at $K = 6$ is spread across a number of Eurasian populations, rather than separating out the Kalash population, as was seen for microsatellites.

Perhaps most importantly, there are no cases of individuals whose cluster membership is clearly discrepant between the microsatellite and SNP datasets (a mislabeled Biaka Pygmy and a mislabeled Japanese individual in the microsatellite analysis were not among the individuals genotyped with SNPs). This concordance argues against the presence of serious labeling errors in the SNP data.

2.2 Populations

The two Bantu groups, from Kenya and southern Africa, were combined for data analysis, so that the data we analyzed consisted of 52 populations. Populations were classified by geographic region in the same manner as in previous work with the same sample of individuals[3,9].

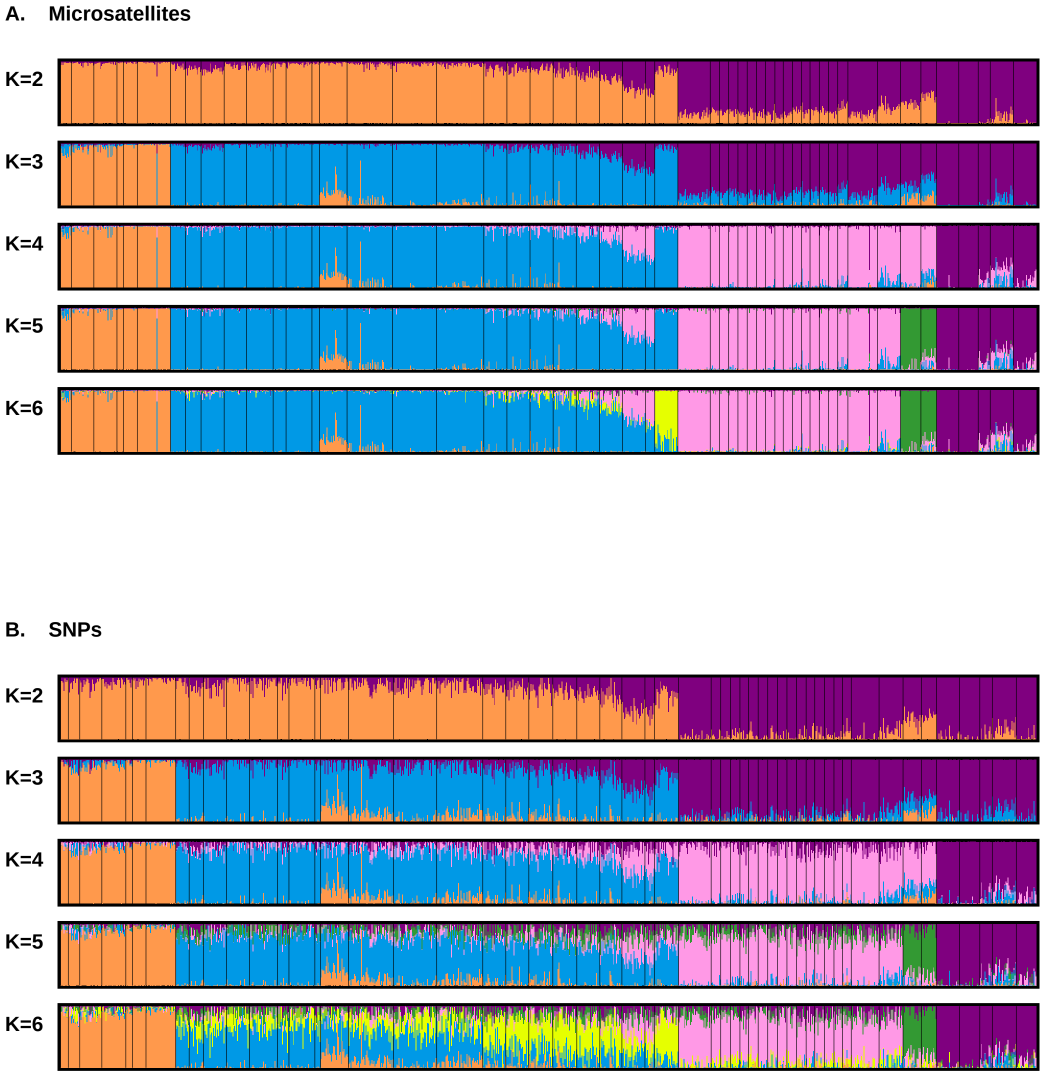


Figure SM.2: Comparison of *structure* results based on the same samples, using different genetic markers. **(A)** Microsatellites, **(B)** SNPs. For both microsatellites and SNPs, the panels run from $K = 2$ clusters (top) to $K = 6$ clusters (bottom). Each cluster is represented by a different color, and each individual by a thin vertical line. An individual's proportion of membership in each cluster is indicated by the proportion of the line length that is drawn in each color. The results in part **A** are taken directly from Figure 1 of Rosenberg et al.[9].

2.3 SNPs

Monomorphic SNPs: Of the 2909 SNPs with high-quality data, 50 were found to be monomorphic in the sample of 1017 individuals with low levels of missing data, and were excluded from consideration.

SNPs with missing data: Among the SNPs polymorphic in the set of 1017 individuals, 10 SNPs with at least 10% missing data were excluded. For autosomal SNPs, the fraction of missing data was calculated as the total fraction of individuals whose genotypes were missing, whereas for X-chromosomal loci, it was equal to $(2f' + m')/(2f + m)$, where f, m, f', and m' respectively denote the number of females considered (360), the number of males considered (657), the number of females with missing genotypes, and the number of males with missing genotypes.

SNPs were identified for which one or more populations had a sample size fewer than 5 alleles in the sample of 927 unrelated individuals. One additional polymorphic SNP with this property was then excluded.

SNPs not in Hardy-Weinberg equilibrium: From the set of 927 unrelated individuals, three population groupings with relatively low levels of population structure in previous work[9] were constructed: Europeans excluding the Adygei, Basque, and Sardinian populations (83 individuals); sub-Saharan Africans excluding the Biaka Pygmy, Mbuti Pygmy and San populations (62 individuals); and East Asians excluding the Cambodian, Japanese, Lahu, and Yakut populations (159 individuals).

A chi-squared test of the null hypothesis of Hardy-Weinberg equilibrium was performed in each of these population groups, taking into account the Yates continuity correction[10]. For X-chromosomal SNPs, males were ignored in the tabulation of allele frequencies and in the hypothesis test. SNPs that had the following pair of properties were then discarded: (1) both alleles had at least five copies in at least two of the three population groups; (2) the chi-squared test statistic was greater than four in at least two of the population groups for which both alleles had at least five copies. Using these criteria, SNPs for which the minor allele was very rare in at least two of the three population groups were assumed to be in Hardy-Weinberg equilibrium.

With these criteria for violation of Hardy-Weinberg equilibrium, 20 SNPs were identified. Of these SNPs, it is noteworthy that six of them were among those that had previously been excluded due to missing data. Given that only 11 SNPs were discarded on account of missing data, for the six loci that were both in Hardy-Weinberg disequilibrium and had substantial missing data, it is likely that systematic errors in the SNP assays were responsible for both problems. If substantial missing data and Hardy-Weinberg violations were independent phenomena, then the expected number of polymorphic SNPs with both properties would equal $(20)(11)/2859 \approx 0.08$, considerably less than the number observed, namely 6.

Summary of excluded SNPs: In summary, from our initial design of 3024 SNPs, the number we retained for analysis equaled 2834, or 93.7% of the original SNPs. The discarded SNPs included 115 with data of sufficiently low quality that no genotypes were reported, 50 monomorphic SNPs, 11 polymorphic SNPs with high levels of missing data, and 14 polymorphic SNPs with low levels of missing data but with Hardy-Weinberg disequilibrium. The final set of SNPs for analysis included 2540 autosomal and 294 X-chromosomal SNPs.

2.4 Missing data rate

Of the $2(927)(2540) = 4,709,160$ autosomal genotypes possible in the sample of 927 distinct individuals, the number of missing genotypes was 4494. Of the $(610 + 2 \times 317)(294) = 365,736$ X-chromosomal genotypes possible, the number missing was 518. Combining all 2834 SNPs, the missing data rate was 0.099%.

2.5 Genotyping error rate

Three duplicate samples were genotyped: Biaka samples #452 and #1087 (male), Han samples #813 and #1022 (female), and Melanesian samples #657 and #826 (female). Considering all SNPs for which both individuals in a duplicate pair were genotyped (and excluding X-chromosomal SNPs in male duplicates), only one discrepancy was observed among 8171 genotypes, a rate of 1.22×10^{-4} discrepancies per genotype.

Mendelian error checks were performed by considering 66 parent/offspring pairs and 17 trios included among the 1039 individuals for which (at least partial) SNP data were obtained[4]. The fraction of SNP genotypes with Mendelian incompatibilities was approximately 2×10^{-4} in both of two separate estimates, one using the parent/offspring pairs and one using the trios.

3 Haplotype phasing

Haplotype phasing was performed using fastPHASE v. 0.9[11]. 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[12]. Additionally, fastPHASE produces haplotype inferences that are nearly as accurate as those produced by PHASE, despite a much smaller computation time[11]. In our study, the computational speed is an issue in view of the large number of individuals (PHASE is quadratic in the number of individuals). Finally, fastPHASE has the added benefit of allowing separate parameters for each population, a feature that is attractive for our worldwide dataset.

3.1 Phasing strategy

In order to phase this dataset, there were a number of choices that needed to be made, including how to label and group the population samples, and the choice of K, the number of haplotype clusters to assume. Several other parameters relating to the details of the phasing also needed to be specified. 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 Scheet and Stephens[11].

These testing runs used parameters $H = 500$, $T = 20$, and $C = 25$ (the fastPHASE documentation provides a full description of these parameters), and either $K = 10$ or $K = 20$ clusters. These parameter choices were found to be sensible during a larger set of preliminary runs. When analyzing the full dataset, we found that using 20 clusters was roughly optimal, giving slightly improved performance compared to a choice of 10 clusters. However, the smaller number of clusters was better with smaller samples.

A novelty of fastPHASE is that it models haplotypes as being shared across populations, but allows haplotype frequencies and cluster jump rates to vary across populations. We found that in comparison with alternative schemes, grouping haplotypes by geographic region in the phasing produced the best results by most measures (see **Supplementary Note**). Consequently, the analyses in the main text of the paper rely on this approach.

3.2 Phasing performance

Phasing performance was assessed in three different ways. First, we masked 10% of the genotypes and then used fastPHASE to impute the missing data. The error rate in the entire sample, as inferred from the fraction of missing genotypes correctly imputed, is only 4.4%.

Next, we assessed the error rate for the phasing of pairs of heterozygote SNPs. This analysis was performed using 42 individuals in the sample of 927 who have a parent or child in the full dataset (although a few individuals are part of larger families[4], this analysis only used parent/offspring pairs.) Suppose that at two SNPs an individual in the phased sample is a double heterozygote: 0/1, 0/1. If the parent is a double homozygote (for example 0/0, 0/0) then the parental genotype determines the haplotype phase of the child (haplotypes 0-0 and 1-1 for the parent above). Using this logic, we determined the error rate for phasing such genotype configurations. It is worth noting that requiring the parent to be a double homozygote is expected to shift the allele frequencies somewhat. Keeping this caveat in mind, the procedure provides a simple tool for assessing phasing accuracy.

Phasing accuracy as assessed by determining the fraction of doubly heterozygous genotypes correctly resolved is generally very high for SNP pairs within 10 kb (error rates from 0.2% to 6.6% for different geographic regions), but deteriorates substantially as spacing increases to 50 kb (regional error rates from 1.8% to 18.9%).

Third, we used the trio data from the HapMap to assess the accuracy of fastPHASE at estimating a measure of pairwise LD, r^2. The basic idea was to determine parental haplotypes using the trio data, and then to estimate the haplotypes using fastPHASE with the offspring genotypes hidden. We find that the concordance in estimated r^2 between the two methods of phase estimation (fastPHASE, and haplotypes inferred from trios) is extremely high. For example, in both the HapMap CEU and YRI samples, 88% of SNP pairs with $r^2 > 0.5$ have identical r^2 values by both methods.

3.3 Missing data imputation

Because we observed low imputation error rates in our missing data simulations (described above), we have used fastPHASE to impute all missing genotypes in the dataset. Recall that missing genotypes represent only 0.1% of the entire dataset. All haplotype-based analyses in the paper are based on this reconstructed dataset that contains no missing values.

4 Haplotype visualization

In order to visualize the haplotypes in each genomic region we used the following algorithm. Our method was conceptually motivated in part by the model developed by Scheet and Stephens[11]; however, 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 by Rosenberg et al.[9]. 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. We start in the physical center of the genomic region, and identify the largest segment that exactly matches one template. That segment is colored according to the color of the template. Next, we move immediately to the right of the colored segment, and color the largest possible segment that exactly matches one of the templates and that has a left-hand edge at the right edge of the region that has already been colored. This process is continued until the right-hand end of the genomic region is reached. An analogous process is then performed to the left of the central block. Note that sometimes a rare allele is not found on any template. We ignore these rare alleles when creating the mosaic structure; however one version of our program (not used in the main text) colors such minor alleles separately in a unique color.

Finally, for each population shown in Figure 1, 20 haplotypes were sampled without replacement from among the total number 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.

5 Estimation of recombination rates

The reversible jump Markov Chain Monte Carlo (rjMCMC) method of McVean et al.[13] (LDhat v2.0) was used to estimate maps of the population-genetic recombination rate ρ from the (unphased) SNP data for each genomic region. This approach uses a method similar to that of Hudson[14] to evaluate the likelihood of the recombination rate between every pair of SNPs, and then computes the product of these likelihoods. This "composite likelihood" is an approximation to the true likelihood. To allow for variation of the recombination rate within a genomic region, the method assumes a piecewise block-like structure to the recombination rate within a region. It then uses rjMCMC to explore a range of rates within these blocks as well as a range of possible partitions of the region into blocks.

The program requires a choice of value for a smoothing parameter that determines the penalty for introducing a new block; following McVean et al.[13], 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 2000), following a burn-in of 10^5 iterations. Multiple runs of the algorithm of equally many or larger numbers of iterations, or starting from different initial maps produced little variation in the final estimated recombination map (results not shown).

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 (this may vary across populations due to monomorphic SNPs at the edge of the region in some populations). 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 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

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

where d_{reg} is deCODE's pedigree-based estimate of recombination rate per kb estimate 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 zero. The model was fitted to minimize $\sum_{pop,reg} \epsilon^2_{pop,reg}$ by a hill-climbing algorithm, where the sum ranges over 52×31 population-region combinations (52 populations, 31 autosomal regions excluding region 1).

There is good agreement among the recombination maps for different populations. The populations have much shared history, and therefore, the extent to which this agreement indicates constancy of the present-day map over populations is difficult to assess. In view of the recent findings of polymorphic hotspots[15,16], the prevalence of some hotspots will probably differ between populations. However in the genomic regions in our data — in 52 populations with different demographies and varying amounts of shared history — the significance of any observed differences in recombination maps is hard to assess formally.

6 Haplotype summary statistics

For haplotypes in the genomic core regions and various values of the "window size" w, we computed haplotype summary statistics based on haplotypes within genomic windows with 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 $a+w$). If the position $a+w$ for a particular haplotype locus was beyond the last SNP of its core region, the haplotype locus was discarded. For the window size termed "full length," the window size was set to be one base pair longer than the full length of each core region (so that both ends would be included in each haplotype locus). For the analyses of haplotype summary statistics, we subdivided genomic core regions 30, 31 and 32 each into two core regions, as these regions each contained a large gap (130 kb, 375 kb, and 250 kb, respectively). Thus, because the four X-chromosomal regions were excluded, the number of genomic core regions used for these analyses was 35. In the analysis of full-length haplotypes, the number of haplotype loci equaled this total number of core regions. 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 in eq. 7 in the main text), 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 from those of the other statistics in that estimates involving the HapMap excluded from consideration SNPs not among the 2078 in our dataset that were contained in the HapMap.

7 Tag SNP analysis

For analysis of tag SNP portability, we used overlapping SNPs with the HapMap Phase II data (release 19, `http://www.hapmap.org`) for 29 regions (X-chromosomal regions 23-26, and regions 30-32 with gaps were excluded). Of the SNPs typed in the current study, 2078 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 only the parental genotypes in the case of the CEU and YRI samples (that is, offspring genotypes were excluded). CHB and JPT samples were combined into one 90-sample population for phasing and all subsequent analyses. The following sections describe the analyses presented in the main text; for additional related analyses see the **Supplementary Note**.

7.1 Tagging strategy

For each HapMap population separately (CEU, YRI, CHB+JPT), we selected 333 LD-based tag SNPs using a method related to that described by Carlson et al.[17] (**Supplementary Note**). The method of Carlson et al. is a greedy algorithm that identifies bins of SNPs such that pairwise LD is high for SNPs within bins, but low for SNPs in different bins. A single tag SNP is selected from each bin. Only "core" SNPs are considered as potential tag SNPs, but we assess how well each potential tag SNP captures all HapMap variation that was typed in our HGDP samples. The number of core SNPs present in the HapMap ranges from 27 to 58 per region, out of a total possible 60. A SNP is considered to be "tagged" by another SNP if the r^2 value between the two SNPs is greater than 0.85.

To focus the analysis, we decided to select tag SNPs at a density intermediate to that found on evenly-spaced 300,000-SNP and 500,000-SNP tagging panels. Using 2.8 Gb as the size of the "genotypable" euchromatic genome, a uniformly spaced 300,000-SNP chip would have an approximate density of 1 SNP per 9333 bp, while a 500,000-SNP chip would have a density of 1 SNP per 5600 bp. For the main analysis, we settled on a tagging panel size of 333 SNPs. Assuming that the total size of the core regions examined in this analysis is 2.61 Mb (29 core regions × 90 kb), a set of 333 core SNPs would have a density of 1 SNP per 7800 bp, simulating a chip with a density between those of 300,000-SNP and 500,000-SNP panels.

7.2 Assessment of tag portability

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 as 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 $n = 29$.

7.3 Sample size correction in computation of PVT

Because sample sizes vary across populations (and geographic regions), 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 an alternate approach. For each geographic region, we selected the population with the largest sample size, subsampled this population over a grid of values of the sample size (30 times for each value) and calculated the average PVT for each subsample size. We fit a simple linear model to these data (PVT on sample size) and used the estimate of the regression coefficient as a correction factor for small samples from within the same geographic region. For a population from geographic region i with $n < 36$ chromosomes, we calculated a corrected PVT as

$$PVT_{corrected} = PVT_{raw} + (36 - n)\beta_i,$$

where β_i is the regression coefficient for geographic region i and n is the number of chromosomes sampled in the population.

References

1. Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlander, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).

2. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nature Rev. Genet.* **6**, 333–340 (2005).

3. Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., and Feldman, M. W. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, 660–671 (2005).

4. Rosenberg, N. A. Standardized subsets of the HGDP-CEPH Human Genome Diversity Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* **70**, doi:10.1111/j.1469–1809.2006.00285.x (2006).

5. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).

6. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., R.Kautzer, C., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N. P., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).

7. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).

8. Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

9. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

10. Weir, B. S. *Genetic Data Analysis II.* Sinauer, Sunderland, MA, (1996).

11. Scheet, P. and 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).

12. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., and Donnelly, P. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).

13. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).

14. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).

15. Jeffreys, A. J. and Neumann, R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genet.* **31**, 267–271 (2002).

16. Jeffreys, A. J. and Neumann, R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.* **14**, 2277–2287 (2005).

17. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2005).

Supplementary Note for “A worldwide survey of haplotype variation and linkage disequilibrium in the human genome”

This supplement contains additional results about phasing, recombination rate estimation, haplotype sharing, tag SNPs, and determinants of tag SNP portability. Additional methods can be found in the **Supplementary Methods**.

Contents

1 Haplotype phasing

The primary dataset consists of 2834 genotyped in 927 individuals. These individuals represent a set of unrelated individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel (the "HGDP") representing 53 human populations (two Bantu groups that were grouped together in other analyses were kept separate during phasing runs that used population labels).

Also described below are analyses using the CEPH European-American (CEU) and Yoruba (YRI) samples from the Phase II HapMap for 2078 SNPs that overlap with our data, and analyses of 42 additional parents or children of individuals in the sample of 927 unrelateds who form our primary data set.

1.1 Phasing performance

A novelty of fastPHASE is that it models haplotypes as being shared across populations, but allows the frequencies and jump rates to vary across populations. Although this feature can be advantageous for phasing performance, there are many possible ways of grouping the populations for the phasing analysis. We considered in detail four different methods: (1) no population labels; (2) individuals assigned one of 7 regional labels: Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, and the Americas; (3) individuals assigned one of 53 distinct population labels; (4) each of the 53 populations phased in a completely separate fastPHASE analysis.

As described below, method 2 (regional labels) had the best performance by most measures, and the main analyses are based on this method. Method 1 (no labels) tends to underestimate the haplotype differences between populations, while method 4 (separate analyses) and perhaps also method 3 (population labels) tend to exaggerate the differences between populations.

Phasing performance was assessed in three different ways, as described in the **Supplementary Methods**. First, we masked 10% of the genotypes and then used fastPHASE to impute the missing data. Our results, shown in Table SN.1, are very similar for the HGDP Europeans to those reported by Scheet and Stephens[1] for the HapMap CEU group. Overall the accuracy is high. The error rate in the entire sample, as inferred from the fraction of missing genotypes correctly imputed, is only 4.4%. As might be expected, the error rates are highest in the populations with lowest linkage disequilibrium (LD). The error rate in our analysis of the HapMap samples is slightly higher (6.6%); this result is probably a consequence of the slightly lower SNP density in the data that we extracted from the HapMap.

Next, we assessed the error rate for the phasing of pairs of heterozygote SNPs by comparing haplotypes inferred using parent/offspring pairs to those inferred using fastPHASE (**Supplementary Methods**). Table SN.2 shows the switch error rates as determined by this method for SNPs within 10 kb of each other; Table SN.3 shows error rates for SNPs at spacing of 10–50 kb. Phasing accuracy is generally very high at the shorter distance, but deteriorates substantially at the longer distance (however it is still much better than random). As expected, phasing accuracy is also generally increased in the populations with higher LD.

	No lab.	**Reg.**	Pop.	SepPops	Strawman	X chr.	HapMap	Worst
WORLD	**4.9**	**4.4**	**4.6**	**7.1**	**38.4**	**2.9**	**6.6**	**11.0 (1)**
AFRICA	10.2	**8.8**	9.0	12.5	37.8	6.2	10.5	15.2 (1)
EUROPE	4.0	**3.7**	4.0	6.8	39.4	2.5	5.3	10.0 (1)
MIDDLE EAST	4.8	**4.6**	4.6	5.4	40.9	3.5		11.1 (1)
C/S ASIA	4.3	**4.2**	4.3	6.2	40.7	2.6		10.0 (1)
EAST ASIA	4.0	**3.6**	4.0	7.8	36.6	2.2	4.7	11.8 (1)
OCEANIA	4.5	**3.5**	3.5	5.7	33.4	2.6		10.9 (6)
AMERICA	3.2	**2.3**	2.3	4.1	37.8	2.3		8.7 (6)

Table SN.1: Error rates (percent) for imputing hidden genotypes using fastPHASE. Error rates are shown both for the entire sample (WORLD) and broken down by region. The first four data columns correspond to different methods of grouping the data in the fastPHASE analysis (see text). "Strawman" indicates the error rate when missing data are replaced by the most common genotype in the region; this provides a baseline for comparison[1]. Also shown are error rates in the HapMap samples using the set of SNPs that are found in both the HGDP and HapMap datasets, error rates for the four X chromosome regions (using regional labels), and the highest observed error rates at any single genomic region (regions 1 and 6).

	No lab.	**Reg.**	Pop.	SepPops	X chr. (Reg.)
AFRICA (6)	3.9	**3.7**	5.7	11.9	2.2
EUROPE (1)	6.7	**6.6**	6.7	10.1	0.0
MIDDLE EAST (2)	0.5	**0.2**	0.2	1.5	0.0
C/S ASIA (2)	7.2	**1.6**	1.4	1.6	0.0
EAST ASIA (2)	2.2	**2.5**	2.0	13.6	0.0
OCEANIA (7)	0.8	**0.6**	0.5	15.4	0.0
AMERICA (22)	1.4	**2.0**	1.4	7.4	0.5

Table SN.2: Error rates (percent) for estimating the relative phase of pairs of heterozygous SNPs within 10 kb of one another. These error rates are based on parent/offspring pairs present in the full dataset — note the small sample sizes in some regions, especially in Europe. Random phasing would produce expected error rates of 50%. The number of parent/offsping pairs for each geographic region is as indicated. See **Supplementary Methods** for further explanation.

	No lab.	**Reg.**	Pop.	SepPops	X chr. (Reg.)
AFRICA (6)	13.3	**14.2**	14.6	17.0	16.2
EUROPE (1)	19.6	**18.9**	19.1	31.9	0.0
MIDDLE EAST (2)	3.7	**1.8**	2.3	21.4	0.03
C/S ASIA (2)	15.4	**5.6**	4.4	4.0	—
EAST ASIA (2)	3.8	**6.4**	6.5	16.4	0.6
OCEANIA (7)	4.2	**3.4**	2.7	21.4	0.0
AMERICA (22)	4.6	**5.1**	3.8	13.3	2.2

Table SN.3: Error rates (percent) for estimating the relative phase of pairs of heterozygous SNPs between 10–50 kb of one another. These error rates are based on parent/offspring pairs present in the full dataset — note the small sample sizes in some regions, especially in Europe. Random phasing would produce expected error rates of 50%. The number of parent/offsping pairs for each geographic region is as indicated. See **Supplementary Methods** for further explanation.

Third, we used the trio data from the HapMap to assess the accuracy of fastPHASE at estimating a measure of pairwise LD, r^2. The basic idea was to determine parental haplotypes using the trio data, and then to estimate the haplotypes using fastPHASE with the offspring genotypes hidden. Recall that the HapMap data used for this analysis include only those SNPs that are also in both the HapMap *and* in our dataset, and hence the accuracy of the fastPHASE reconstructions is slightly lower for this dataset than for our HGDP data (Table SN.1).

In order to compare these two phasing methods, we used only genotypes for which phase could be determined unambiguously from the trios (that is, excluding sites that are triple heterozygotes, and excluding some missing data configurations). We then estimated haplotypes using fastPHASE, and deleted any genotypes that were missing from the trio-phased data so that the available genotypes matched exactly across the two datasets. It is plausible that these procedures of dealing with ambiguities create some bias in r^2, but this should not be of serious concern here, because the main goal is to establish the concordance of r^2 between the two methods of estimation. Finally, r^2 was estimated for every pair of SNPs that are in the same one of our genomic regions, separately for the trio-phased and fastPHASEd datasets.

We find that the concordance in estimated r^2 between the two methods of phase estimation is extremely high (Figures SN.1 and SN.2). It is apparent from the figures that there is a slight downward bias in the estimates of r^2 by fastPHASE; however this is a small effect. For example, 88% of SNP pairs with $r^2 > 0.5$ have identical r^2 values by both methods (for both CEU and YRI). In the tag SNP analysis in the paper, we use fastPHASE results to determine whether r^2 exceeds a specified threshold for any given SNP pair. Among SNP pairs with substantial LD ($r^2 > 0.5$) we find that less than 2% of SNP pairs have $r^2 > 0.9$ by one phasing method and $r^2 < 0.9$ by the other phasing method (for both CEU and YRI). Hence we conclude that the phase inference has minimal impact on the results of the tag SNP analysis, and surely much less impact that the inherent sampling variability of r^2 in finite samples.

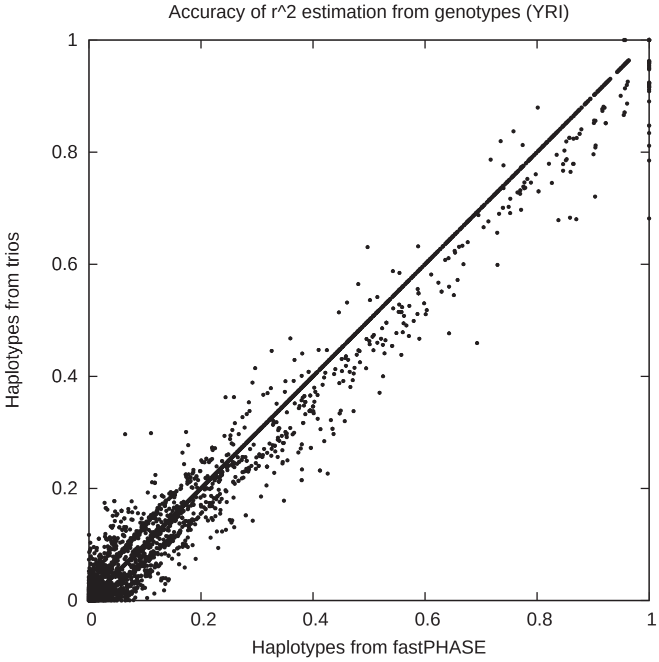


Figure SN.1: Comparison of estimates of r^2 for pairs of SNPs in the Yoruba HapMap sample, based on haplotypes from trios and on haplotypes from fastPHASE.

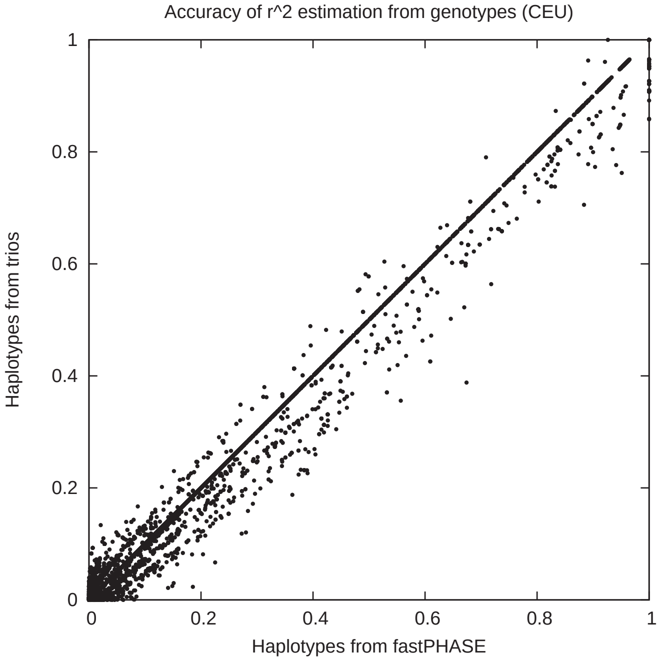


Figure SN.2: Comparison of estimates of r^2 for pairs of SNPs in the CEPH European-American HapMap sample, based on haplotypes from trios and haplotypes from fastPHASE.

1.2 Summary

We find that fastPHASE provides accurate phase reconstruction for our data over modest distances (less than 10 kb), and somewhat less accurate reconstruction over longer distances (that is, distances at which the data are presumably uninformative). fastPHASE provides extremely accurate assessments of the actual level of LD, which is perhaps of most importance for our analyses. Reconstructed haplotypes are likely to be quite accurate when there is extensive LD, whereas long-range haplotypes in populations with low LD will be less accurate. However, there is not likely to be a substantial bias in the extent of LD.

2 Estimation of recombination rates

In this section we describe additional analyses performed in estimating recombination rates and effective population sizes. We show that our results are robust to the range of sample sizes and number of monomorphic SNPs found across populations in our data, and to the method used to infer the population recombination rates. In addition we show that using the X chromosome regions to estimate effective population leads to similar results to those found using the autosomal regions. Figures showing the estimated recombination maps and the average population-genetic recombination rates plotted against the pedigree-based rate estimates for all populations will be made available at `http://pritch.bsd.uchicago.edu/dataArchive.html`.

2.1 The average recombination rate in a genomic region

We estimated ρ per 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 (this may vary across populations due to monomorphic SNPs at the edge of the region in some populations). To estimate the underlying recombination rate for each genomic region we fitted a log-linear model $\rho_{pop,reg} = 4N_{pop}r_{reg}\epsilon_{pop,reg}$, where $\rho_{pop,reg}$ is the average population-genetic recombination rate per kb in region reg and population pop, N_{pop} is the effective population size, r_{reg} is a genomic region effect that is free to vary, and $\epsilon_{pop,reg}$ is a multiplicative error term. In Figure SN.3, r_{reg} is plotted against the pedigree-based rate estimate for each genomic region. The figure indicates a high degree of concordance between the estimated population-genetic recombination rates and the pedigree-based estimates of recombination rate for each of our genomic regions. A number of factors can contribute to the “noise” in the correlation between the pedigree-based estimate of the recombination rate for a genomic region and r_{reg}: for example, the rates are imperfect estimates of the underlying population-genetic recombination rates, and the pedigree-based estimates are estimates of the average rate over larger distances than those spanned by our regions. Genomic region 1 clearly has an unusually high estimated population-genetic recombination rate given its pedigree-based estimate. This genomic region is distal to the first deCODE marker on chromosome 1p, so the pedigree-based recombination rate is unreliable. Genomic region 1 is therefore considered an outlier, and is excluded from the analyses presented. The inclusion of region 1 leads to slightly inflated estimates of population sizes for all populations, but has little effect on the relative order of the estimates across populations (results not shown).

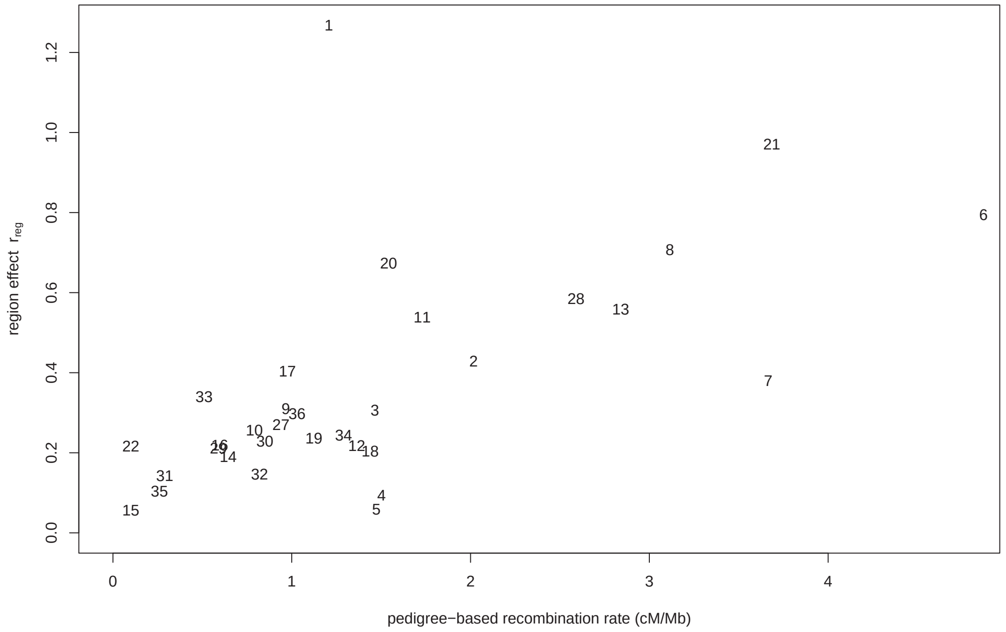


Figure SN.3: The genomic region effect r_{reg} plotted against the pedigree-based recombination rate estimate. Each non-X-chromosomal genomic region is numbered.

2.2 Impact of ascertainment scheme

To assess whether using SNPs ascertained in a worldwide panel compared to a mixture of ascertainment strategies affected our estimates of ρ, we investigated whether there was a systematic difference between estimates based Patil et al.[2] SNPs (regions on chromosome 21) and those based on SNPs in the other autosomal regions. We regressed the estimates of r_{reg} from the log-linear model described above on the pedigree-based rates. The difference in mean of the squared residuals between the chromosome 21 regions and the other regions was calculated. The significance of this difference was then assessed by permuting chromosome 21 and other autosome labels $10,000$ times among genomic regions and determining how often the difference observed in the data was exceeded in the permuted data sets. This process yielded a P-value of 0.38. To assess whether the variance of our ρ estimates was affected by the difference in ascertainment between chromosome 21 and the other autosomes, we applied a similar procedure with the difference in the variance of the residuals between chromosome 21 and the other autosomes, obtaining $P = 0.39$. Thus, there is no appreciable difference in estimated recombination rate between the mixed ascertainment SNPs and the SNPs ascertained in the global panel.

2.3 Impact of gene density

To assess whether gene density had an affect on our rate estimates, we calculated the Spearman rank correlation coefficient between the residuals of the regression of r_{reg} on the pedigree-based estimates and the gene density of each region. A Spearman rank correlation coefficient of -0.27 was found. The significance of this correlation was assessed by permuting the gene densities across regions and determining how often a more negative correlation was observed. This correlation was found to have a suggestive P-value of 0.07, consistent with the idea that increased gene density might lead to a small local reduction of effective population size via natural selection.

2.4 Estimating effective population sizes from recombination rates

Two similar strategies were used to estimate the effective population size N_{pop} for each population. In the first method (method 1) the pedigree-based rate estimate was assumed to be correct, and the population size was simply estimated by fitting a linear model (with an intercept of zero) of ρ/kb for a genomic region from a population against the pedigree-based estimated rate for the region. The N_{pop} value estimated from this first method is plotted against the microsatellite heterozygosity for each population in Figure SN.4B.

The second method (method 2) was inspired by the observation from Figure SN.3 that the pedigree-based estimate of the recombination rate for a region is in some cases a relatively poor predictor of ρ/kb over populations (the most extreme example of which is the excluded region 1). In the second model, we assume

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

where $\rho_{pop,reg}$ is the average population-genetic recombination rate per kb in region reg and population pop, d_{reg} is the pedigree rate estimate for the region, and b_{reg} is the "error" in the pedigree-based rate estimate for a region when used at a local scale. To constrain this model, we require that the sum of b_{reg} across regions is zero. This model was fitted to minimize $\sum_{pop,reg} \epsilon^2_{pop,reg}$ by a hill-climbing algorithm. Different initial conditions for the values of N_{pop} and b_{reg} led to very similar results, and there was good agreement between the two different methods of estimating N_{pop}. The second method is based on a more appropriate model, and therefore was used to estimate the effective population sizes used in Figure 5 of the main paper (shown again in Figure SN.4A). The N_{pop} values estimated by method 2, along with the Spearman rank correlations between the population-genetic and pedigree-based estimates, are given in Tables SN.4 and SN.5.

2.5 Robustness of estimates to changes in SNP spacing

The SNPs placed in each genomic region were designed around a core of 60 SNPs with an average spacing of 1.5 kb between SNPs to have two flanking regions of 120 kb with an average spacing of 10 kb between SNPs. The fitting of rate variation by LDhat might be affected by this SNP layout, so we re-estimated the recombination rates using only the core SNPs for each genomic region (note that the unusual spacing of SNPs in genomic regions 30-32 meant that these regions were excluded from the analysis). This procedure will lead to a better estimate of ρ per kb, as the core regions have a higher density of SNPs. However, the pedigree-based estimates are likely to be less appropriate for these smaller regions. The values of N_{pop} (estimated with method 2 using only the core of each genomic region) are plotted against microsatellite heterozygosity in Figure SN.4C. There is a slight tendency for the values of N_{pop} estimated from the core of each genomic

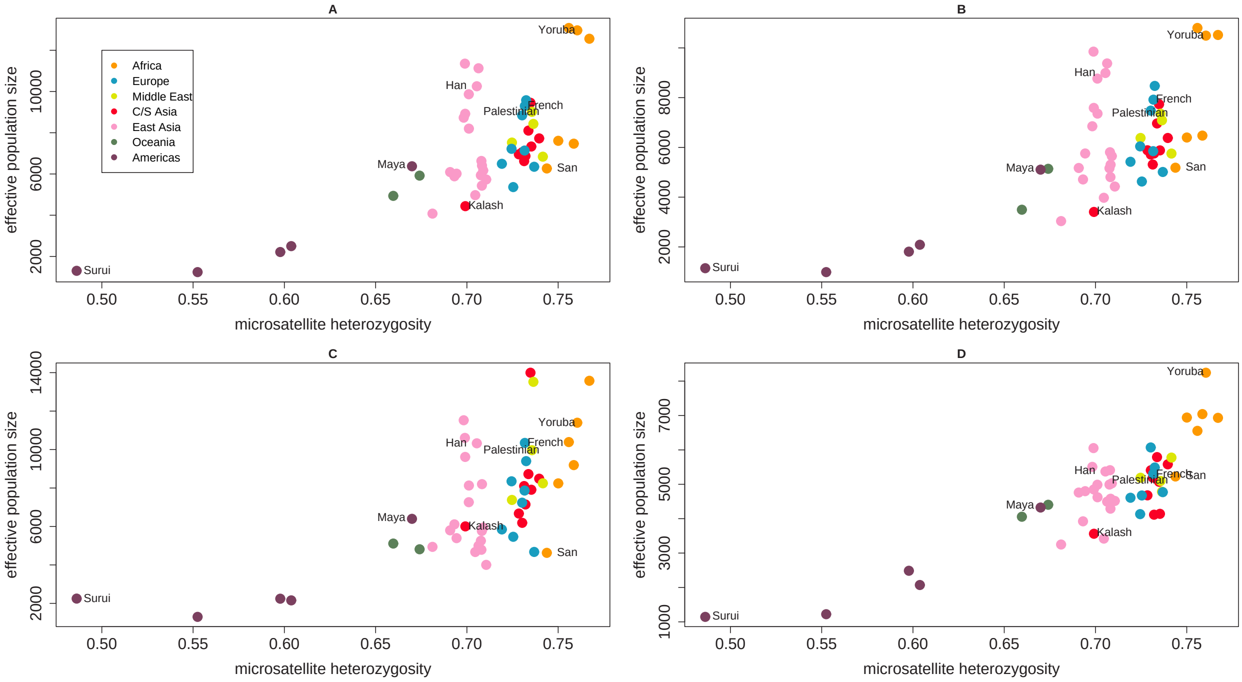


Figure SN.4: N_{pop}, estimated in four different ways, plotted against microsatellite heterozygosity. **(A)** Method 2. **(B)** Method 1. **(C)** Only the SNPs from the core regions, using method 2. **(D)** The average of reduced samples and SNP sets, using method 1 (see text).

region to be higher than those estimated from the whole region. The sparse SNP spacing in the flanking region presumably leads LDhat to place relatively few changes of rate within these flanking regions, producing a slightly lower recombination rate estimate and hence a slightly lower N_{pop} on average. Our results remain reasonably consistent whether the complete region or only the core regions are used, although the values of N_{pop} seem somewhat more variable across populations when considering only the core region.

2.6 Robustness of estimates to changes in sample size

A concern is that since LDhat penalizes the likelihood of changes in rate, populations with larger sample size and more polymorphic SNPs might be able to introduce more changes in rate than those with smaller samples and more monomorphic SNPs. Populations with smaller samples and fewer SNPs might therefore be biased toward constancy and toward the background rate, because they cannot overcome the penalty for introducing additional variation in the recombination rate. Many of the populations that have low estimates of N_{pop} also have small sample size and a large number of monomorphic SNPs.

To investigate this problem, all population samples were dropped to a sample size of 8 individuals, the Surui sample size. The Surui sample was chosen as a basis for this analysis as they have among the lowest estimated N_{pop} and the most monomorphic SNPs. Then, for each sample, each genomic region had SNPs removed at random until it had the same number, or fewer, SNPs than the Surui sample for this genomic region (note that some populations had fewer SNPs than the Surui for particular regions). This procedure was performed 10 times. Method 1 was used to estimate N_{pop} for each of these reduced data sets, as it allows us to estimate N_{pop} for each reduced data set separately, rather than requiring estimates over populations. The average N_{pop} estimates from the 10 reduced data sets are plotted against microsatellite heterozygosity in Figure SN.4D. Whereas the populations with either larger sample sizes or more polymorphic SNPs are biased down by this reduction in sample size and SNP number, the relative order of the population estimates is robust to the effects of sample size and numbers of monomorphic SNPs.

2.7 Robustness of estimates to changes in the LDhat smoothing parameter

We also tried several values of the LDhat smoothing parameter for a subset of populations with a range of estimated N_{pop} (results not shown). Larger values of the smoothing parameter in general led to somewhat lower estimates of ρ per kb, as less variation in the rates would be permitted. These lower values in turn resulted in smaller estimates of N_{pop}. However, different values of the smoothing parameter made little difference to the relative ordering of the N_{pop} values across populations, and thus, our qualitative results are robust to the smoothing parameter used.

2.8 Alternative methods for estimating population-genetic recombination rates

The recombination estimation methods *maxdip*[3] and PHASE v2.1[4] were also applied to the unphased data (without imputing missing data) to estimate population-genetic recombination rates for all the autosomal genomic regions and all populations (to reduce the computational load we excluded a number of East Asian populations — Miao, Oroqen, Daur, Mongola, Hezhen, Xibo, Dai, Lahu, She, Naxi, and Tu). The *maxdip* program assumes a constant rate across the whole region, while PHASE assumes that every interval between SNPs has a separate rate. The estimates of N_{pop} from *maxdip* and PHASE (calculated by method 2) are plotted against microsatellite heterozygosity in Figure SN.5. The estimates of N_{pop} based on *maxdip* tend to be lower than those from either PHASE or LDhat, presumably because by estimating a constant rate for a region, the rates are biased towards the background rate for the region. The Spearman rank correlation coefficients between quantities estimated by the three methods are given in Table SN.6. The correlations between the estimates obtained by the three methods are reasonably high, and the qualitative picture of the relationship between N_{pop} and microsatellite heterozygosity can be seen to hold irrespective of the recombination estimation method used.

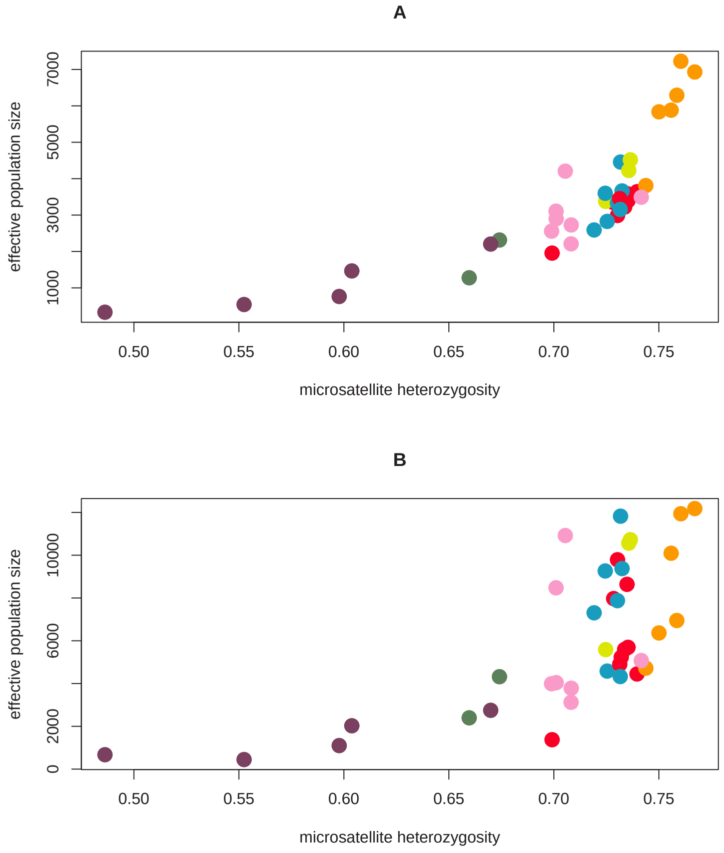


Figure SN.5: N_{pop}, estimated in two different ways, plotted against microsatellite heterozygosity. **(A)** Estimates based on *maxdip*. **(B)** Estimates based on PHASE v2.1. The coloring of populations follows the same scheme as in Figure SN.4.

2.9 Recombination rates for X-chromosomal regions

Four of our genomic regions are located on the X chromosome. LDhat v2.0 does not allow a mixture of genotype information and known haplotype information. Therefore, for this analysis the haplotypes were obtained by fastPHASE (with missing data imputed) and the LDhat method was used assuming these haplotypes to be known. Because males have only one copy of the X chromosome, this procedure leads to a substantial drop in the sample size for some populations. To estimate N_{pop} from the X chromosome we assumed that

$$\rho_{pop,reg} = 2N_{pop}(d_{reg} + b_{reg}) + \epsilon_{pop,reg}. \tag{2}$$

We fit this model as in the autosomal case using a hill-climbing algorithm, with the constraint $\sum_{reg} b_{reg} = 0$. The effective population sizes for populations with 15 or more haplotypes are plotted against microsatellite heterozygosity in Figure SN.6. The relationship between N_{pop} and microsatellite heterozygosity can be seen to hold for the X chromosome as on the autosomes.

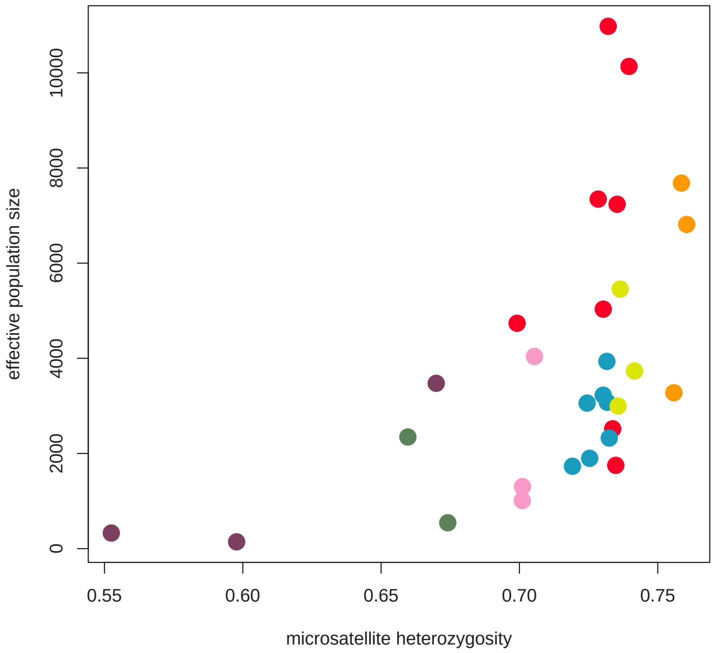


Figure SN.6: N_{pop}, estimated for the X chromosome, plotted against microsatellite heterozygosity. Only populations samples with 15 or more haplotypes are shown. The coloring of populations follows the same scheme as in Figure SN.4.

Population	N_{pop}	Spearman	P-value
Brahui	6872	0.58	0.00082
Balochi	8104	0.62	0.00026
Hazara	6954	0.58	0.00076
Makrani	7726	0.49	0.0057
Sindhi	7332	0.44	0.013
Pathan	9452	0.51	0.0037
Kalash	4439	0.49	0.0058
Burusho	7056	0.57	0.00088
Mbuti Pygmy	7609	0.59	0.00058
Biaka Pygmy	7469	0.36	0.045
Melanesian	4939	0.14	0.47
French	9298	0.6	0.00048
Papuan	5920	0.57	0.00099
Druze	7524	0.52	0.0031
Bedouin	8428	0.59	0.00058
Sardinian	7220	0.58	0.00071
Palestinian	9014	0.51	0.0039
Colombian	2500	0.43	0.018
Cambodian	5433	0.5	0.0045
Japanese	9867	0.55	0.0014
Han	10256	0.55	0.0016
Orcadian	5364	0.41	0.024
Surui	1303	0.21	0.26
Maya	6374	0.54	0.0022
Russian	9576	0.62	0.00024
Mandenka	13082	0.5	0.0044
Yoruba	12968	0.52	0.0028
Yakut	8200	0.61	0.00035
San	6268	0.48	0.007
Karitiana	1240	0.26	0.15
Pima	2215	0.34	0.062
Tujia	11352	0.15	0.42
Italian	8844	0.63	0.00023
Tuscan	6350	0.63	0.00018

Table SN.4: The estimated effective population size of each of the 52 populations, estimated using method 2. The Spearman rank correlation coefficient between the pedigree-based rate estimate and the LDhat ρ/kb estimate for each genomic region, and the P-value of the correlation coefficient, are also shown.

Population	N_{pop}	Spearman	P-value
Yi	6407	0.30	0.10
Miao	8731	0.6	0.0004
Oroqen	5887	0.5	0.0049
Daur	6631	0.48	0.0075
Mongola	5732	0.40	0.025
Hezhen	4973	0.44	0.014
Xibo	5936	0.45	0.012
Mozabite	6839	0.61	0.00038
Han (N. China)	6171	0.55	0.0015
Uygur	6624	0.63	0.00022
Dai	8920	0.53	0.0027
Lahu	4077	0.23	0.20
She	6093	0.25	0.17
Naxi	6014	0.58	0.00074
Tu	11128	0.54	0.0022
Basque	6494	0.53	0.0023
Adygei	7137	0.54	0.002
Bantu	12559	0.54	0.0018

Table SN.5: The estimated effective population size of each of the 52 populations, estimated using method 2 (continued). The Spearman rank correlation coefficient between the pedigree-based rate estimate and the LDhat ρ/kb estimate for each genomic region, and the P-value of the correlation coefficient, are also shown.

	LDhat	*maxdip*	PHASE
LDhat		0.69 (2×10^{-16})	0.70 (2×10^{-16})
maxdip	0.73 (4×10^{-7})		0.74 (2×10^{-16})
PHASE	0.79 (8×10^{-8})	0.83 (5×10^{-9})	

Table SN.6: The Spearman rank correlation coefficient between recombination rate estimates for different estimation methods. The correlation is computed based on estimates for the autosomal regions, excluding region 1. The P-value of the Spearman rank correlation coefficient is given in parentheses. The upper triangular matrix contains the Spearman rank correlation coefficient between two methods in the average estimated rate for each genomic region and each population. The lower triangular matrix contains the Spearman rank correlation coefficient between two methods for the N_{pop} values calculated from their rates using method 2 (see text).

3 Haplotype sharing

Averaging across haplotype loci within genomic regions — and then averaging across the regions — we computed a sample-size corrected statistic $\phi^{(j)}_{g,c,j'}$ that measures the fraction of the common haplotypes in a population j that have the property of being common in population j' (see the "Methods" section of the main text). The value of c denotes the frequency threshold above which haplotypes are considered to be common, and the value of g is a parameter that allows the effect of sample size to be examined.

Using the ϕ statistic, we calculated the fraction of common haplotypes in each HGDP population that were also common in the HapMap populations (CEU, YRI, and CHB+JPT). Thus, j ranged over the HGDP populations and j' ranged over the HapMap populations. This analysis was performed for several choices of the window size for haplotype loci. To explore the effect of sample size on the results, we considered four ways of choosing the parameter g. These methods were as follows: (1) set g in all populations to the smallest number of sampled haplotypes studied in any population, or 12; (2) for each population, set g to the smallest number of haplotypes in the geographic region in which the population originated; (3) set g to 20, and exclude populations with sample size < 20; (4) set g individually for each population to $g = \min(N_j, N_{j'})$, where N_j and $N_{j'}$ respectively denote the numbers of sampled haplotypes in populations j and j'. Because the HapMap populations had larger sample sizes than the HGDP populations, this approach always led to $g = N_j$.

Estimates of ϕ, using common allele thresholds of $c = 0.05$ and $c = 0.1$, were very similar for these four approaches to choosing g (results not shown). We present results only for method 4, which uses the full samples from individual populations to identify common haplotypes. For each population, and for six different haplotype window sizes, Tables SN.7-SN.12 show the fractions of common haplotypes found to be common in the most similar of the HapMap populations. Table SN.13 then presents the average across HGDP populations of these proportions, and Table SN.14 shows the largest value across populations of the fraction of common haplotypes *not* found in the most similar HapMap population.

Several conclusions can be drawn from these tables. First, fairly similar results are obtained for cutoff values of 0.05 and 0.1. Second, for both thresholds and most combinations of populations and window sizes, the HapMap population with the greatest fraction of common alleles matches well with previous estimates of population structure. For populations from Europe, the Middle East, and Central/South Asia, the CEU sample generally has the largest fraction of common alleles among HapMap samples; for populations from Africa, the YRI sample has the largest fraction; and for populations from East Asia, Oceania, and the Americas, the CHB+JPT sample has the largest fraction.

Third, the coverage of common haplotypes by the most similar haplotype is extremely high in nearly all populations for short haplotypes (length 5kb or less). As the haplotype length increases, the fraction of haplotypes present in the HapMap decreases considerably, so that at length 50kb, the most distant populations have less than half of their common haplotypes present in the HapMap. In general, the populations whose common haplotypes are least contained in the HapMap are African populations such as San and Mbuti, populations from Oceania and the Americas, which are not represented in the Hapmap, and relatively distinctive Eurasian populations such as Kalash and Uygur. In the case of the Uygur it is possible that ancient admixture between populations more similar to CEU and populations more similar to CHB+JPT subdivided many of the haplotypes common in one but not the other of the main ancestral groups, so that many of the common haplotypes in Uygur could represent mosaics of common haplotypes from these ancestors.

Table SN.7: The fraction of common SNPs (equal to a window size of 1 bp) in HGDP populations that are also common in the most similar HapMap population. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.9836	2	0.9726	2
Yoruba	0.9917	2	0.9781	2
Mandenka	0.9876	2	0.9802	2
San	0.9860	2	0.9698	2
Mbuti Pygmy	0.9877	2	0.9608	2
Biaka Pygmy	0.9890	2	0.9616	2
Orcadian	0.9835	1	0.9669	1
Adygei	0.9751	1	0.9578	1
Russian	0.9920	1	0.9758	1
Basque	0.9837	1	0.9835	1
French	0.9837	1	0.9747	1
Italian	0.9863	1	0.9769	1
Sardinian	0.9884	1	0.9773	1
Tuscan	0.9872	1	0.9843	1
Mozabite	0.9540	1	0.9519	1
Bedouin	0.9716	1	0.9580	1
Druze	0.9797	1	0.9774	1
Palestinian	0.9725	1	0.9564	1
Balochi	0.9778	1	0.9590	1
Brahui	0.9823	1	0.9735	1
Makrani	0.9745	1	0.9591	1
Sindhi	0.9742	1	0.9547	1
Pathan	0.9809	1	0.9606	1
Burusho	0.9783	1	0.9678	1
Hazara	0.9871	1	0.9673	1
Uygur	0.9645	1	0.9487	1
Kalash	0.9820	1	0.9675	1
Han	0.9961	3	0.9888	3
Han (N. China)	0.9798	3	0.9784	3
Dai	0.9767	3	0.9752	3
Daur	0.9743	1	0.9729	3
Hezhen	0.9768	3	0.9812	3
Lahu	0.9897	3	0.9781	3
Miao	0.9825	3	0.9797	3
Oroqen	0.9736	1	0.9900	3
She	0.9901	3	0.9699	3
Tu	0.9693	1	0.9633	3
Tujia	0.9811	3	0.9808	3
Xibo	0.9707	3	0.9766	3
Yi	0.9746	3	0.9663	3
Mongola	0.9755	3	0.9644	3
Naxi	0.9853	3	0.9820	3
Cambodian	0.9770	3	0.9709	3
Japanese	0.9942	3	0.9815	3
Yakut	0.9746	3	0.9699	3
Melanesian	0.9818	3	0.9715	3
Papuan	0.9837	3	0.9788	3
Karitiana	0.9820	1	0.9574	3
Surui	0.9849	1	0.9693	1
Colombian	0.9786	1	0.9621	1
Maya	0.9747	1	0.9551	1
Pima	0.9834	1	0.9634	1

Table SN.8: The fraction of common haplotypes in HGDP populations that are also common in the most similar HapMap population. The window size was 1 kb and there are on average 1.85 SNPs in such a window. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.9716	2	0.9619	2
Yoruba	0.9850	2	0.9694	2
Mandenka	0.9814	2	0.9734	2
San	0.9699	2	0.9476	2
Mbuti Pygmy	0.9745	2	0.9354	2
Biaka Pygmy	0.9733	2	0.9425	2
Orcadian	0.9743	1	0.9535	1
Adygei	0.9665	1	0.9432	1
Russian	0.9866	1	0.9682	1
Basque	0.9795	1	0.9775	1
French	0.9780	1	0.9608	1
Italian	0.9786	1	0.9636	1
Sardinian	0.9833	1	0.9698	1
Tuscan	0.9788	1	0.9727	1
Mozabite	0.9316	2	0.9307	1
Bedouin	0.9580	1	0.9415	1
Druze	0.9715	1	0.9678	1
Palestinian	0.9601	1	0.9402	1
Balochi	0.9653	1	0.9463	1
Brahui	0.9755	1	0.9626	1
Makrani	0.9639	1	0.9459	1
Sindhi	0.9645	1	0.9395	1
Pathan	0.9738	1	0.9473	1
Burusho	0.9707	1	0.9537	1
Hazara	0.9794	1	0.9549	1
Uygur	0.9498	1	0.9298	1
Kalash	0.9722	1	0.9520	1
Han	0.9937	3	0.9851	3
Han (N. China)	0.9685	3	0.9693	3
Dai	0.9673	3	0.9658	3
Daur	0.9592	1	0.9606	3
Hezhen	0.9705	3	0.9735	3
Lahu	0.9822	3	0.9674	3
Miao	0.9729	3	0.9720	3
Oroqen	0.9586	1	0.9774	3
She	0.9829	3	0.9573	3
Tu	0.9567	3	0.9523	3
Tujia	0.9718	3	0.9736	3
Xibo	0.9568	3	0.9662	3
Yi	0.9605	3	0.9537	3
Mongola	0.9612	3	0.9492	3
Naxi	0.9740	3	0.9710	3
Cambodian	0.9619	3	0.9618	3
Japanese	0.9896	3	0.9754	3
Yakut	0.9643	3	0.9568	3
Melanesian	0.9750	3	0.9557	3
Papuan	0.9732	3	0.9662	3
Karitiana	0.9738	1	0.9437	3
Surui	0.9764	1	0.9531	1
Colombian	0.9694	1	0.9458	3
Maya	0.9661	1	0.9371	1
Pima	0.9766	1	0.9519	3

Table SN.9: The fraction of common haplotypes in HGDP populations that are also common in the most similar HapMap population. The window size was 2 kb and there are on average 2.66 SNPs in such a window. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.9599	2	0.9532	2
Yoruba	0.9773	2	0.9627	2
Mandenka	0.9746	2	0.9677	2
San	0.9502	2	0.9251	2
Mbuti Pygmy	0.9601	2	0.9141	2
Biaka Pygmy	0.9577	2	0.9269	2
Orcadian	0.9662	1	0.9428	1
Adygei	0.9571	1	0.9281	1
Russian	0.9825	1	0.9600	1
Basque	0.9720	1	0.9660	1
French	0.9712	1	0.9520	1
Italian	0.9718	1	0.9519	1
Sardinian	0.9752	1	0.9614	1
Tuscan	0.9701	1	0.9619	1
Mozabite	0.9131	2	0.9086	1
Bedouin	0.9440	1	0.9264	1
Druze	0.9618	1	0.9604	1
Palestinian	0.9480	1	0.9282	1
Balochi	0.9558	1	0.9333	1
Brahui	0.9669	1	0.9521	1
Makrani	0.9542	1	0.9326	1
Sindhi	0.9544	1	0.9258	1
Pathan	0.9626	1	0.9349	1
Burusho	0.9617	1	0.9366	1
Hazara	0.9698	1	0.9434	1
Uygur	0.9352	1	0.9109	1
Kalash	0.9617	1	0.9379	1
Han	0.9920	3	0.9828	3
Han (N. China)	0.9584	3	0.9582	3
Dai	0.9604	3	0.9573	3
Daur	0.9448	3	0.9493	3
Hezhen	0.9654	3	0.9659	3
Lahu	0.9735	3	0.9595	3
Miao	0.9682	3	0.9641	3
Oroqen	0.9425	1	0.9696	3
She	0.9742	3	0.9474	3
Tu	0.9468	3	0.9412	3
Tujia	0.9630	3	0.9659	3
Xibo	0.9468	3	0.9563	3
Yi	0.9472	3	0.9447	3
Mongola	0.9499	3	0.9352	3
Naxi	0.9671	3	0.9651	3
Cambodian	0.9519	3	0.9546	3
Japanese	0.9871	3	0.9697	3
Yakut	0.9578	3	0.9479	3
Melanesian	0.9689	3	0.9450	3
Papuan	0.9627	3	0.9569	3
Karitiana	0.9659	1	0.9302	3
Surui	0.9681	1	0.9353	1
Colombian	0.9567	1	0.9316	3
Maya	0.9566	1	0.9238	1
Pima	0.9659	1	0.9424	3

Table SN.10: The fraction of common haplotypes in HGDP populations that are also common in the most similar HapMap population. The window size was 5 kb and there are on average 4.72 SNPs in such a window. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.9301	2	0.9304	2
Yoruba	0.9640	2	0.9462	2
Mandenka	0.9586	2	0.9503	2
San	0.8924	2	0.8708	2
Mbuti Pygmy	0.9150	2	0.8553	2
Biaka Pygmy	0.9215	2	0.8810	2
Orcadian	0.9496	1	0.9198	1
Adygei	0.9293	1	0.8940	1
Russian	0.9694	1	0.9405	1
Basque	0.9528	1	0.9444	1
French	0.9479	1	0.9305	1
Italian	0.9530	1	0.9225	1
Sardinian	0.9571	1	0.9394	1
Tuscan	0.9440	1	0.9414	1
Mozabite	0.8801	2	0.8669	1
Bedouin	0.9150	1	0.8947	1
Druze	0.9418	1	0.9341	1
Palestinian	0.9211	1	0.8973	1
Balochi	0.9293	1	0.9021	1
Brahui	0.9475	1	0.9206	1
Makrani	0.9337	1	0.8968	1
Sindhi	0.9273	1	0.8864	1
Pathan	0.9397	1	0.9074	1
Burusho	0.9392	1	0.9065	1
Hazara	0.9525	1	0.9072	1
Uygur	0.8958	1	0.8686	1
Kalash	0.9367	1	0.8980	1
Han	0.9829	3	0.9745	3
Han (N. China)	0.9343	3	0.9463	3
Dai	0.9369	3	0.9385	3
Daur	0.9166	3	0.9307	3
Hezhen	0.9432	3	0.9489	3
Lahu	0.9539	3	0.9395	3
Miao	0.9438	3	0.9461	3
Oroqen	0.9037	1	0.9526	3
She	0.9571	3	0.9262	3
Tu	0.9176	3	0.9232	3
Tujia	0.9390	3	0.9453	3
Xibo	0.9233	3	0.9383	3
Yi	0.9237	3	0.9267	3
Mongola	0.9175	3	0.9126	3
Naxi	0.9450	3	0.9517	3
Cambodian	0.9216	3	0.9423	3
Japanese	0.9790	3	0.9600	3
Yakut	0.9383	3	0.9317	3
Melanesian	0.9485	3	0.9148	3
Papuan	0.9345	3	0.9286	3
Karitiana	0.9404	1	0.9004	3
Surui	0.9468	1	0.9061	3
Colombian	0.9269	1	0.9078	3
Maya	0.9339	1	0.9017	3
Pima	0.9448	1	0.9196	3

Table SN.11: The fraction of common haplotypes in HGDP populations that are also common in the most similar HapMap population. The window size was 20 kb and there are on average 12.82 SNPs in such a window. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.7868	2	0.8145	2
Yoruba	0.8841	2	0.8577	2
Mandenka	0.8720	2	0.8455	2
San	0.6585	2	0.6125	2
Mbuti Pygmy	0.6999	2	0.5967	2
Biaka Pygmy	0.7663	2	0.6861	2
Orcadian	0.8968	1	0.8564	1
Adygei	0.8491	1	0.8118	1
Russian	0.9212	1	0.8838	1
Basque	0.8831	1	0.8887	1
French	0.8882	1	0.8796	1
Italian	0.9000	1	0.8372	1
Sardinian	0.8853	1	0.8791	1
Tuscan	0.8435	1	0.8808	1
Mozabite	0.7621	1	0.7937	1
Bedouin	0.8397	1	0.8360	1
Druze	0.8798	1	0.8784	1
Palestinian	0.8550	1	0.8468	1
Balochi	0.8630	1	0.8320	1
Brahui	0.8701	1	0.8387	1
Makrani	0.8584	1	0.8088	1
Sindhi	0.8601	1	0.8098	1
Pathan	0.8630	1	0.8358	1
Burusho	0.8603	1	0.8252	1
Hazara	0.8841	1	0.8283	3
Uygur	0.7566	1	0.7460	3
Kalash	0.8548	1	0.7899	1
Han	0.9376	3	0.9293	3
Han (N. China)	0.8066	3	0.8672	3
Dai	0.8260	3	0.8626	3
Daur	0.7745	3	0.8456	3
Hezhen	0.8252	3	0.8660	3
Lahu	0.8572	3	0.8465	3
Miao	0.8336	3	0.8623	3
Oroqen	0.7917	3	0.8568	3
She	0.8513	3	0.8319	3
Tu	0.8123	3	0.8320	3
Tujia	0.8284	3	0.8598	3
Xibo	0.8090	3	0.8589	3
Yi	0.8084	3	0.8388	3
Mongola	0.7919	3	0.8303	3
Naxi	0.8374	3	0.8705	3
Cambodian	0.7919	3	0.8731	3
Japanese	0.9182	3	0.9052	3
Yakut	0.8666	3	0.8769	3
Melanesian	0.8503	3	0.7920	3
Papuan	0.7964	3	0.7830	3
Karitiana	0.8193	1	0.7571	3
Surui	0.8449	1	0.7800	3
Colombian	0.7986	1	0.7735	3
Maya	0.8243	1	0.7951	3
Pima	0.8305	1	0.8116	3

Table SN.12: The fraction of common haplotypes in HGDP populations that are also common in the most similar HapMap population. The window size was 50 kb and there are on average 27.45 SNPs in such a window. The most similar HapMap population is denoted by 1 = CEU, 2 = YRI and 3 = CHB+JPT. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

HGDP population	$c > 0.05$	HapMap	$c > 0.1$	HapMap
Bantu	0.5756	2	0.6484	2
Yoruba	0.7264	2	0.7620	2
Mandenka	0.7381	2	0.6951	2
San	0.3667	2	0.3299	2
Mbuti Pygmy	0.4230	2	0.3338	2
Biaka Pygmy	0.5517	2	0.5451	2
Orcadian	0.8154	1	0.7532	1
Adygei	0.7611	1	0.7083	1
Russian	0.8531	1	0.7813	1
Basque	0.7971	1	0.8189	1
French	0.8189	1	0.8036	1
Italian	0.8228	1	0.7555	1
Sardinian	0.8155	1	0.7741	1
Tuscan	0.6900	1	0.7683	1
Mozabite	0.6513	1	0.7146	1
Bedouin	0.7578	1	0.7034	1
Druze	0.8061	1	0.8026	1
Palestinian	0.7975	1	0.7588	1
Balochi	0.7666	1	0.7344	1
Brahui	0.7571	1	0.7330	1
Makrani	0.7406	1	0.7062	1
Sindhi	0.7818	1	0.7028	1
Pathan	0.7758	1	0.7485	1
Burusho	0.7589	1	0.7179	1
Hazara	0.7804	1	0.7724	3
Uygur	0.5887	1	0.6289	3
Kalash	0.7351	1	0.6347	1
Han	0.8657	3	0.8658	3
Han (N. China)	0.6478	3	0.7400	3
Dai	0.6645	3	0.7546	3
Daur	0.6105	3	0.7417	3
Hezhen	0.6763	3	0.7449	3
Lahu	0.7085	3	0.7307	3
Miao	0.6782	3	0.7449	3
Oroqen	0.6487	3	0.7372	3
She	0.7084	3	0.7146	3
Tu	0.6443	3	0.7043	3
Tujia	0.6617	3	0.7367	3
Xibo	0.6579	3	0.7535	3
Yi	0.6603	3	0.7185	3
Mongola	0.6267	3	0.7178	3
Naxi	0.6963	3	0.7919	3
Cambodian	0.6349	3	0.7781	3
Japanese	0.8030	3	0.8456	3
Yakut	0.7779	3	0.7877	3
Melanesian	0.7154	3	0.6316	3
Papuan	0.6670	3	0.6268	3
Karitiana	0.6280	1	0.5968	3
Surui	0.6658	3	0.6333	3
Colombian	0.6054	3	0.6424	3
Maya	0.6864	1	0.6412	3
Pima	0.6822	3	0.6575	3

Table SN.13: The fraction of haplotypes common in a randomly chosen HGDP population that are also common in the most similar HapMap population. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

window size	$c > 0.05$	$c > 0.1$
1	0.9807	0.9702
1000	0.9708	0.9576
2000	0.9611	0.9463
5000	0.9365	0.9205
20000	0.8360	0.8270
50000	0.7014	0.7091

Table SN.14: The maximum across HGDP populations of the fraction of common haplotypes absent from the most similar HapMap population. Common haplotypes were defined by having a frequency $c > 5\%$ or $c > 10\%$.

window size	$c > 0.05$	$c > 0.1$
1	0.0460	0.0513
1000	0.0684	0.0702
2000	0.0869	0.0914
5000	0.1199	0.1447
20000	0.3415	0.4033
50000	0.6333	0.6701

4 Tag SNP analysis

The ongoing HapMap Project has generated a tremendous empirical description of linkage disequilibrium across the human genome. One of the principal motivations for the HapMap Project is to enable researchers to efficiently select genetic markers for genome-wide association studies that (a) maximize coverage of the genomic regions of interest and (b) minimize the amount of information shared between markers. The current scope of the HapMap Project involves genotypes from four populations. Therefore, it is important for studies of non-HapMap populations to find a reasonable way to connect the genetics of their samples to the data gathered in HapMap populations or other populations studied at a similar density as in the HapMap[5]. Because it represents a broad sample of human populations, the HGDP-CEPH Diversity Panel presents an opportunity to refine the understanding of how patterns of LD are shared among groups.

We set out to identify which of the HapMap populations is the most appropriate for selecting tag SNPs for LD-based population-genetic analysis in each of the HGDP populations. In an ideal situation, researchers would have the ability to tailor their genome-wide marker panel to the specific patterns of LD in a population of interest. As many preselected SNP panels are already commercially available as part of large-scale genotyping platforms, it seems likely that many researchers will be analyzing SNPs selected without regard to LD considerations. Therefore, we also examined the amount of information about non-genotyped variation that is captured using random panels of SNPs, and compared this level of information to that captured by LD-directed tag SNPs.

4.1 LD-based tag selection

For each population, we selected LD-based tag SNPs using the method described in Carlson et al.[6]. We only considered "core" SNPs as potential tag SNPs, but we assessed how well each potential tag SNP captured all HapMap variation that was typed in our HGDP samples. The number of core SNPs present in the HapMap ranges from 27 to 58 per region, out of a total possible 60. For each region, we used the following algorithm to select tag SNPs:

1. Calculate pairwise r^2 between each core SNP and each of the rest of the SNPs in the region.

2. Select the core SNP with the most pairwise r^2 values above a tagging threshold, record the identification number (rs#) of this SNP and the number of SNPs that it tags (its "object SNPs"), and remove this SNP and its object SNPs from consideration.

3. If there are no core SNPs remaining, stop.

4. Return to step 1.

We ran this algorithm on each genomic region separately. We then compiled results across all regions, ranking tag SNPs based on their number of "hits" (pairwise r^2 values above the specified threshold), and we then selected tag SNPs sequentially from the ordered list until the desired SNP density was reached. We did not require that each region have a tag SNP.

In order to explore appropriate tagging thresholds, we estimated the total number of SNPs required to tag all core HapMap SNPs at a given level of r^2 (Table SN.15). Qualitatively similar results were obtained with different thresholds, and in the end we settled on an r^2 threshold of 0.85.

HapMap population	$r^2 = 0.95$	$r^2 = 0.90$	$r^2 = 0.80$	Total core
CEU	510	451	378	952
YRI	689	641	558	986
CHB+JPT	452	392	328	906

Table SN.15: Number of SNPs required to capture common variation as function of tagging threshold. For three values of r^2, we report the number of tag SNPs required to tag all common core SNP variation (minor allele frequency > 0.05) in the HapMap populations. "Total core": total number of core SNPs with $MAF > 0.05$ in each HapMap population.

4.2 Impact of sample size

There is considerable variation in the sample sizes of the populations in the 927 individuals studied from the HGDP, ranging from 6 individuals (San, Tuscans) to 45 (Bedouin). To make cross-population comparisons of tag SNP portability and the extent of LD, it is important to understand the effect of sample size on these analyses, and if necessary, to account for it. Therefore, we have investigated the effects of sampling on r^2 and on the tag portability of LD-based tags selected in each of the three HapMap populations, as measured by PVT, the proportion of variation tagged (described in the "Methods" in the main text and in the **Supplementary Methods**).

First, we conducted a preliminary tag SNP analysis of the total HGDP using a set of tag SNPs chosen in the HapMap CEU sample, and examined the results for any obvious sample size effects. The correlation coefficient between sample size and the percent of untyped variation tagged was -0.36, suggesting a substantial sample size effect. Fitting a simple linear model to these variables (PVT vs. sample size) yielded a regression coefficient of -0.003, or a 3% decrease in variation tagged for every 10 haplotypes added to a population sample. If this relationship held true for all populations, Bedouin ($n = 90$ haplotypes and PVT=0.44) would have a PVT of 0.674 at $n = 12$, while San ($n = 12$ and PVT=0.489) would have a PVT of 0.255 if we could increase their sample size to $n = 90$.

This analysis does not address the fact that we have an *a priori* expectation for several populations with a small n in the HGDP to be well-tagged regardless of sample size. To explore the effect of sample size more carefully, our strategy was to select the HGDP population with the largest sample size from each of the 7 geographic regions, and to assess sampling effects within each of these populations by subsampling real data in a Monte Carlo fashion. The actual subsample sizes selected in each case depend on the total sample size of the population, but were selected to be roughly comparable.

For each subsample size "*ss*", we analyzed tag SNP portability just as in the original (full-data) analysis, using only *ss* haplotypes from the original 90 individuals and the set of tag SNPs from the appropriate HapMap donor population. This process was repeated 30 times and the results for each *ss* averaged together. These results clearly display a linear trend between sample size and PVT (Table SN.16). This trend suggests that a sensible approach to populations with smaller than average sample sizes would be to apply a correction factor based on extrapolation from simulations with genetically similar populations but larger sample sizes.

To estimate an appropriate correction factor, we fit a simple linear model to each set of resampling data and recorded the regression coefficient. As can be seen in Table SN.16, the slopes are similar across many populations.

In all subsequent analyses (unless otherwise mentioned), we adjusted all PVT scores to the mean HGDP sample size, 36 chromosomes. For populations with fewer than 36 chromosomes, we applied a correction factor (estimated from the relevant geographic region) to the PVT score.

$$PVT_{corrected} = PVT_{raw} + (36 - n)\beta_i,$$

where β_i is the regression coefficient for geographic region i and n is the number of chromosomes sampled in the population. For samples larger 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.

Population	β	R^2	P-value
Makrani	-0.0016	0.90	0.009
Maya	-0.001	0.97	0.01
Biaka	-0.0024	0.96	0.012
Bedouin	-0.0012	0.59	0.146
French	-0.004	0.93	0.004
Papuan	-0.004	0.79	0.074
Han	-0.001	0.78	0.004

Table SN.16: Linear modeling of sample size on proportion of variation tagged. β: slope from least squares regression of PVT on sample size; R^2: the fraction of variance explained by the model; P-value: P-value testing the null model that $\beta = 0$.

4.3 Performance of alternate tag panels

For each of the HGDP populations, we scored what proportion of the "untyped" variation in the core regions would be captured by tag SNPs chosen in each of the HapMap populations (Tables SN.17 and SN.18). Overwhelmingly, haplotype variation from populations with the same geographic "region" label is best-represented by the same tag SNP set (CEU, YRI, or CHB+JPT).

Is there a "universal tag donor"? Our results clearly indicate that the best tag SNP set varies by geographic region. This observation suggests the next question, "In what geographic region does selecting the correct tag donor matter most?" On average, East Asian populations are most affected by tag donor selection; the average difference in PVT between the best- and worst-performing SNP sets for these populations is 26%. These are followed by the African populations, with approximately 21% difference between best and worst sets. C/S Asian and Middle-Eastern populations tend to be the least affected in this analysis; the average differences between the best and worst sets are 13.9% and 14.25%, respectively.

It may be the case that selecting the "nearest" (in terms of genetic distance) HapMap population as the source of tag SNPs for a given HGDP population may not be the most effective way of using the HapMap data (see also "Determinants of tag SNP portability" section). We tried a simple strategy of designing tag panels using each possible pair of HapMap populations, as well as a cosmopolitan set of 360 HapMap haplotypes (120 from each population), choosing the best 333 SNPs as before. The PVT improves for many of the genetically intermediate populations in C/S Asia and the Middle East (**Supplementary Figure 1**).

4.4 Impact of tagging threshold

Despite the simplicity of the concept of tag SNP analysis, the process of selecting tag SNPs and measuring their portability required many choices about the details of the analysis. In the following sections we analyze the effect of these choices on the conclusions that we draw in the main paper.

Tagging threshold is one highly visible parameter in our analysis. To evaluate the role of this parameter, we conducted a tag SNP analysis with tagging threshold 0.95 and a 452 tag SNP panel, which corresponds to the density of a 500,000-SNP chip. The qualitative results are very similar to the analysis with r^2 threshold of 0.85. The best tag donor is the same for each HGDP population, with the exception that with the 0.95 threshold the Mozabite population is best tagged by CEU. We have also performed numerous analyses with the popular r^2 threshold of 0.8. In one example using the same tagging density as the main analysis, we observed a slight departure from the trend of the "nearest" HapMap population as the best tag donor (Tables SN.19 and SN.20).

One possible limitation to the generality of these results is the manner in which the tag SNPs were selected in each HapMap population. By prioritizing tag SNPs that capture the most variation across all regions, we are emphasizing regions that have high LD and probably also those that have low recombination rates. A more balanced analysis might be to select the best N tag SNPs from each of the 29 regions. Our analysis of this alternate study design, with $N = 6$, indicates that although the proportion of variation tagged in each case drops across all comparisons, the ordering of which HapMap population is the best tag donor does not change (results not shown). The only exceptions are that in this new analysis, using an r^2 threshold of 0.95 and comparing the results to those obtained with our primary tag selection strategy, the Bedouin, Mozabite, and Sindhi populations are now best tagged by the YRI tag panel rather than by the CEU panel.

			Tagging set		
Region	Population	N	CEU	YRI	CHB+JPT
AFRICA	Bantu	38	0.252	**0.421**	0.196
	Biaka Pygmy	46	0.218	**0.393**	0.220
	Mandenka	42	0.352	**0.493**	0.284
	Mbuti Pygmy	24	0.268	**0.357**	0.175
	San	12	0.195	**0.386**	0.179
	Yoruba	44	0.299	**0.527**	0.256
AMERICAS	Colombian	14	0.769	0.684	**0.838**
	Karitiana	28	0.823	0.660	**0.827**
	Maya	42	**0.740**	0.596	0.726
	Pima	24	0.831	0.672	**0.831**
	Surui	16	0.881	0.722	**0.908**
C/S ASIA	Balochi	48	**0.671**	0.555	0.505
	Brahui	48	**0.620**	0.516	0.472
	Burusho	46	**0.660**	0.541	0.500
	Hazara	44	**0.653**	0.497	0.553
	Kalash	46	**0.728**	0.583	0.587
	Makrani	50	**0.678**	0.530	0.531
	Pathan	48	**0.684**	0.558	0.531
	Sindhi	48	**0.616**	0.532	0.497
	Uygur	20	**0.671**	0.533	0.552
E ASIA	Cambodian	16	0.636	0.559	**0.746**
	Dai	20	0.734	0.539	**0.801**
	Daur	20	0.736	0.544	**0.798**
	Han	68	0.708	0.562	**0.883**
	Han (N. China)	20	0.694	0.529	**0.834**
	Hezhen	18	0.713	0.545	**0.840**
	Japanese	58	0.671	0.541	**0.819**
	Lahu	16	0.706	0.604	**0.848**
	Miao	20	0.717	0.510	**0.858**
	Mongola	20	0.669	0.520	**0.799**
	Naxi	16	0.725	0.507	**0.767**
	Oroqen	18	0.603	0.502	**0.758**
	She	20	0.688	0.571	**0.860**
	Tu	20	0.704	0.546	**0.778**
	Tujia	20	0.689	0.511	**0.785**
	Xibo	18	0.726	0.533	**0.757**
	Yakut	50	0.696	0.578	**0.769**
	Yi	20	0.696	0.566	**0.833**

Table SN.17: Summary of Phase II HapMap tag portability. Tagging threshold is $r^2 = 0.85$. Each tagging set contains 333 SNPs selected on the basis of Phase II HapMap data. The best portability score for each population (PVT) is set in boldface. These boldface values of PVT are the ones used to construct Figure 7A in the main text. N: number of haplotypes in the HGDP sample.

			Tagging set		
Region	Population	N	CEU	YRI	CHB+JPT
EUROPE	Adygei	30	**0.673**	0.518	0.495
	Basque	48	**0.802**	0.576	0.595
	French	56	**0.749**	0.568	0.542
	Italian	24	**0.636**	0.480	0.450
	Orcadian	26	**0.651**	0.531	0.433
	Russian	48	**0.791**	0.546	0.566
	Sardinian	54	**0.759**	0.540	0.549
	Tuscan	12	**0.728**	0.509	0.520
MIDDLE EAST	Bedouin	90	**0.586**	0.516	0.439
	Druze	82	**0.701**	0.538	0.557
	Mozabite	56	0.513	**0.518**	0.400
	Palestinian	88	**0.618**	0.526	0.451
OCEANIA	Melanesian	22	0.657	0.543	**0.768**
	Papuan	32	0.704	0.565	**0.752**

Table SN.18: Summary of Phase II HapMap tag portability (continued). Tagging threshold is $r^2 = 0.85$. Each tagging set contains 333 SNPs selected on the basis of Phase II HapMap data. The best portability score for each population (PVT) is set in boldface. These boldface values of PVT are the ones used to construct Figure 7A in the main text. N: number of haplotypes in the HGDP sample.

			Tagging set		
Region	Population	N	CEU	YRI	CHB+JPT
AFRICA	Bantu (Kenya)	22	0.298	**0.464**	0.241
	Bantu (southern Africa)	16	0.276	**0.415**	0.198
	Biaka Pygmy	46	0.273	**0.413**	0.214
	Mandenka	42	0.380	**0.535**	0.28
	Mbuti Pygmy	24	0.269	**0.416**	0.186
	San	12	0.209	**0.391**	0.134
	Yoruba	44	0.337	**0.569**	0.274
AMERICA	Colombian	14	**0.814**	0.688	0.794
	Karitiana	28	0.814	0.713	**0.873**
	Maya	42	**0.747**	0.639	0.694
	Pima	24	**0.855**	0.746	0.843
	Surui	16	0.880	0.757	**0.913**
C/S ASIA	Balochi	48	**0.717**	0.598	0.505
	Brahui	38	0.276	**0.459**	0.214
	Brahui	48	**0.665**	0.567	0.466
	Burusho	46	**0.676**	0.588	0.499
	Hazara	44	**0.733**	0.550	0.541
	Kalash	46	**0.762**	0.633	0.584
	Makrani	50	**0.700**	0.584	0.515
	Pathan	48	**0.691**	0.613	0.518
	Sindhi	48	**0.678**	0.590	0.480
	Uygur	20	**0.742**	0.575	0.526
E ASIA	Cambodian	16	0.621	0.593	**0.688**
	Dai	20	0.779	0.598	**0.804**
	Daur	20	**0.787**	0.605	0.785
	Han	68	0.742	0.612	**0.857**
	Han (N. China)	20	0.761	0.601	**0.882**
	Hezhen	18	0.765	0.581	**0.862**
	Japanese	58	0.716	0.600	**0.827**
	Lahu	16	0.720	0.636	**0.839**
	Miao	20	0.754	0.592	**0.876**
	Mongola	20	0.691	0.577	**0.788**
	Naxi	16	**0.734**	0.536	0.734
	Oroqen	18	0.649	0.553	**0.745**
	She	20	0.718	0.619	**0.851**
	Tu	20	0.721	0.603	**0.811**
	Tujia	20	0.782	0.611	**0.845**
	Xibo	18	0.754	0.579	**0.762**
	Yakut	50	0.709	0.625	**0.728**
	Yi	20	0.727	0.613	**0.800**

Table SN.19: Summary of Phase II HapMap tag portability, alternate threshold. Tagging threshold is $r^2 = 0.80$. Each tagging set contains 327 SNPs selected on the basis of Phase II HapMap data. The best portability score for each population (PVT) is set in boldface. N: number of haplotypes in the HGDP sample.

			Tagging set		
Region	Population	N	CEU	YRI	CHB+JPT
EUROPE	Adygei	30	**0.702**	0.551	0.525
	Basque	48	**0.839**	0.619	0.62
	French	56	**0.793**	0.612	0.534
	Italian	24	**0.735**	0.585	0.493
	Orcadian	26	**0.689**	0.588	0.448
	Russian	48	**0.837**	0.601	0.559
	Sardinian	54	**0.787**	0.589	0.581
	Tuscan	12	**0.708**	0.545	0.490
MIDDLE EAST	Bedouin	90	**0.628**	0.564	0.465
	Druze	82	**0.733**	0.588	0.552
	Mozabite	56	0.541	**0.574**	0.442
	Palestinian	88	**0.652**	0.571	0.475
OCEANIA	Melanesian	22	0.688	0.622	**0.769**
	Papuan	32	0.727	0.610	**0.744**

Table SN.20: Summary of Phase II HapMap tag portability, alternate threshold (continued). Tagging threshold is $r^2 = 0.80$. Each tagging set contains 327 SNPs selected on the basis of Phase II HapMap data. The best portability score for each population (PVT) is set in boldface. N: number of haplotypes in the HGDP sample.

4.5 Impact of ascertainment scheme

To address the impact of different ascertainment strategies on studies of SNP and haplotype variation, we explicitly designed our genotyping panel to consist of SNPs from two general classes of ascertainment. SNPs from 16 of our regions were ascertained by resequencing of a multiethnic panel in Patil *et al.*[2], while SNPs typed in the other regions follow an ascertainment strategy that might be more representative of all SNPs in dbSNP.

A point of primary interest was the effect of ascertainment on our conclusions about tag portability. For each of our three primary tag panels (based on CEU,YRI, and CHB+JPT haplotypes), we measured the difference in mean PVT for the "Patil" regions and "non-Patil" regions separately (using an r^2 threshold of 0.85), and estimated the variance in this difference with 10,000 bootstrap replicates per HGDP population. The results are displayed as box-whisker plots in **Supplementary Figure 2**, Figure SN.7, and **Supplementary Figure 3**. Although there are definite trends for each tag panel (described in the "Worldwide portability of the HapMap" section of the main text), the bootstrap 95% confidence intervals typically contain the value of "no difference" for each HGDP population.

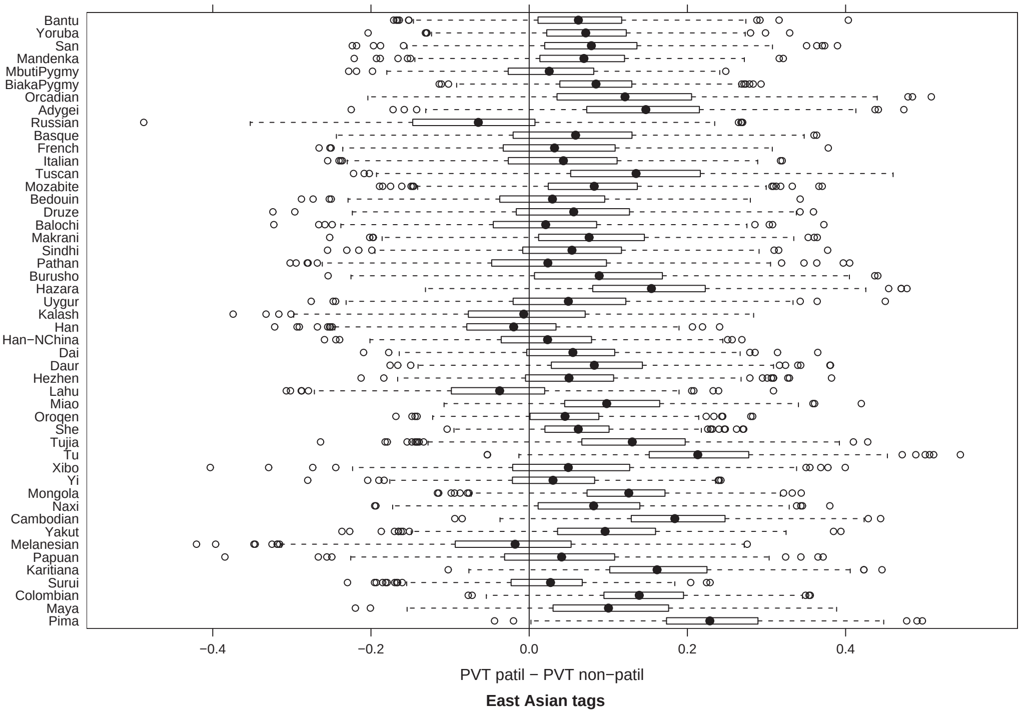


Figure SN.7: Difference in PVT between Patil and Non-Patil regions, East Asian tags. Bootstrap resampling was used to assess the difference in mean PVT of Patil regions and non-Patil Regions. Ten thousand bootstrap replicates were generated for each HGDP population, and are depicted here as box-whisker plots.

4.6 Impact of allele frequency

A mildly surprising result is that American and Oceanic populations, which show extensive LD in our model-based analyses, are not tagged more effectively by all tag sets. This result may be partly explained by the fact that numerous tag SNPs selected in HapMap populations are simply monomorphic in American and Oceanic populations.

We quantified the relative effects of LD and allele frequency on tag portability with an analysis of variance. We estimated what fraction of variance in PVT among populations was due to differences in N_e (as estimated from ρ) and to differences in the number of monomorphic tag SNPs. This analysis used 452 tag SNPs based on an r^2 threshold of 0.95. Whereas N_e was found to have a very significant effect ($P < 0.003$), number of monomorphic tag SNPs did not have a significant effect.

Nonetheless, we observed a slight negative correlation (-0.12) between the number of monomorphic SNPs and PVT. Given this result, we were interested to determine if alterations to our tag SNP selection algorithm that weighed minor allele frequency (MAF) in the selection process could improve tag SNP portability, at least in the Oceanic and American populations. While the correlations in MAF between East Asian populations and populations from Oceania and America are not extremely high, there is enough shared information to believe such a modification might be successful (Figure SN.8). As a simple exploration of this principle, we again performed the tag analysis with the method of Carlson et al.[6], with the modification that the core SNP with highest MAF within a bin is selected as the tag SNP for that bin.

The results of this analysis, using the CHB+JPT HapMap population as the donor, are presented for a sample of populations in Table SN.21. The conclusion is that the algorithm which uses frequency information performs (just slightly) worse in most populations, including those from America and Oceania. The two tagging approaches selected 417/452 SNPs in common. The frequency-based algorithm basically produces the same number of monomorphic tag SNPs in the Americas as the original approach, while actually increasing the number of monomorphic tag sites in the Maya by 1. This reflects the fact that SNPs with very high r^2 tend to have correlated histories; if a SNP is fixed during a bottleneck, other SNPs in high LD with that SNP are likely to have been fixed as well.

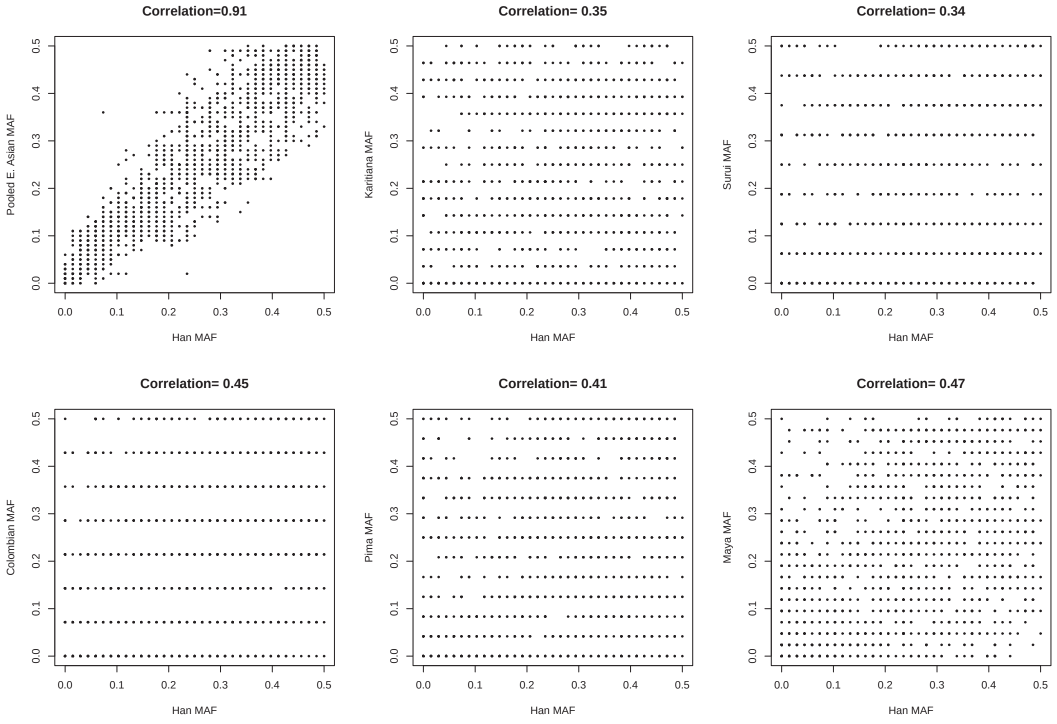


Figure SN.8: Minor allele frequency correlation between populations. Each panel displays a scatter plot of Han MAF against MAF from a second population. In the first panel at the upper left, the Han MAFs are plotted against MAFs from a pooled set of 124 chromosomes from various East Asian populations. All other panels show Han MAF against a single American population MAF. The value of the Pearson correlation coefficient between each set of MAFs is displayed above each panel.

	Frequency information		No frequency information	
Population	PVT	% polymorphic	PVT	% polymorphic
Russian	0.779	0.967	0.760	0.967
Basque	0.750	0.947	0.717	0.947
Sardinian	0.742	0.942	0.732	0.942
Orcadian	0.721	0.918	0.694	0.918
Kalash	0.719	0.938	0.692	0.938
Brahui	0.576	0.962	0.546	0.962
Druze	0.639	0.960	0.621	0.960
Sindhi	0.538	0.989	0.504	0.989
Biaka Pygmy	0.372	0.894	0.350	0.894
Mbuti Pygmy	0.442	0.774	0.442	0.774
Yoruba	0.502	0.916	0.464	0.916
Mandenka	0.452	0.907	0.441	0.907
Papuan	0.801	0.874	0.786	0.874
Maya	0.734	0.912	0.668	0.909
Colombian	0.908	0.803	0.908	0.803
Karitiana	0.878	0.739	0.872	0.739
Pima	0.908	0.777	0.905	0.777
Surui	0.936	0.646	0.936	0.646
Han	0.840	0.998	0.831	0.998
Japanese	0.839	1.000	0.787	1.000
Yakut	0.715	0.982	0.627	0.982
Melanesian	0.871	0.852	0.871	0.852
Mongola	0.890	0.978	0.881	0.978

Table SN.21: Frequency-based and LD-based tag SNP performance. Proportion of variation tagged (PVT) by SNPs selected in CHB+JPT, using the method of Carlson, et al.[6] (no frequency information) or a modified version that uses frequency information. Tagging threshold is $r^2 = 0.95$. The fraction of all tags polymorphic within each population is also listed. PVT values in this table are not sample-size corrected.

4.7 Summary

The results of these analyses, although by no means comprehensive, provide an overview both of how one might incorporate information from the HapMap when designing studies of SNP variation in non-HapMap populations, and of the types of issues that are likely to arise in the process. In general, the HapMap population “closest” to the study population should be used to design tags for a given target population. There will clearly be opportunities to leverage the HapMap in different ways in the design of tagging sets, possibly by combining information across HapMap populations. One possible example of this use of population-genetic data is for admixture mapping.

Sample size is a concern when measuring how well a tag SNP set represents variation in any given population; as the number of haplotypes sampled from a population increases, the number of sites tagged in that sample decreases. Coalescent theory predicts that this effect will reach an asymptote as the sample size of the target population is increased. The rate at which this asymptote is approached will depend on the effective population size (N_e) of the target population.

One idea that has emerged during our tag SNP analysis is that there are three central forms of information captured from the donor population when designing tag SNPs: which SNPs are likely to be polymorphic in neighboring populations, which regions of the genome have low LD (and thus require fewer tag SNPs), and the specific LD relationships between individual SNPs. Optimal tagging strategies will at least indirectly capture these three forms of information.

The first of these three factors has not been carefully considered in most tagging strategies. We observe a loss of tagging power in some bottlenecked populations, such as those in the Americas and Oceania, as a result of many tag SNPs being monomorphic in those populations. Within our current tagging scheme it is difficult to adjust for this by exploiting information about HapMap allele frequencies. It seems likely that with a lower tagging threshold there will be a larger range of MAFs from which to select tags, and thus, allele frequency may have a larger effect on tagging performance (possibly at the expense of linkage information).

Finally, although we have not explored multi-marker methods for tag SNP selection, we expect that results from such an approach to be qualitatively similar to the ones presented here.

5 Determinants of tag SNP portability

In the same manner as described under “Determinants of tag SNP portability” in the main text, we performed additional computations of the relationships between proportion of variation tagged (PVT) and the r^2-decay distance d^*, and between PVT and F_{ST} genetic distance to the HapMap population with highest PVT. To investigate whether SNP ascertainment had an effect on these relationships, we plotted PVT vs. d^* and PVT vs. F_{ST}, restricting attention to only the chromosome 21 SNPs (**Supplementary Figure 4A, 4B**), and to only the autosomal SNPs not on chromosome 21 (**Supplementary Figure 4C, 4D**). In both of these supplementary figures, the correlations between PVT and d^* and between PVT and F_{ST} were similar to the case in which all SNPs were used (Figure 8).

To assess whether the relationship held when allowing tag panels to be determined by combinations of HapMap populations rather than individual populations (see **Supplementary Figure 1**), we repeated the analysis using F_{ST} distance to the HapMap population or combination of HapMap populations that produced the highest portability. **Supplementary Figure 5** provides plots of PVT vs. d^* and PVT vs. F_{ST} for all SNPs, chromosome 21 SNPs only (“Patil SNPs”) and SNPs not on chromosome 21 only (“non-Patil SNPs”). The correlations were similar for all SNP sets, and were also similar to the setting in which only three rather than seven tag SNP panels were considered (see Table SN.22).

Finally, we investigated the robustness of the correlation between PVT and distance d^* by varying the minor allele frequency cutoff m, the r^2 cutoff c, and the percentage p of SNPs with $r^2 > c$. None of these variables had a sizeable impact on d^*, or on the qualitative nature of the relationship between PVT and d^* (results not shown).

Figure	Tag set	Genomic regions	PVT correlation with r^2	PVT correlation with F_{ST}
Supplementary Figure 4A, 4B	3-HapMap	Patil	0.66	-0.23
Supplementary Figure 4C, 4D	3-HapMap	Non-Patil	0.65	-0.16
Supplementary Figure 5A, 5B	7-HapMap	All	0.71	-0.11
Supplementary Figure 5C, 5D	7-HapMap	Patil	0.69	-0.23
Supplementary Figure 5E, 5F	7-HapMap	Non-Patil	0.62	-0.13

Table SN.22: Spearman rank correlation coefficients between tag portability (PVT) and (i) the distance at which r^2 decays below 0.5, and (ii) the F_{ST} genetic distance to the HapMap population that produces the highest tag portability. Each line corresponds to different sets of supplementary figures (“Figure”). Tag SNPs are chosen using either the 3 HapMap populations separately (“3-HapMap”) or using the best of 7 possible combinations (“7-HapMap”). The data are taken either from all regions, or from the Patil (chromosome 21) or non-Patil (autosomal non-21) regions only.

References

1. Scheet, P. and 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).

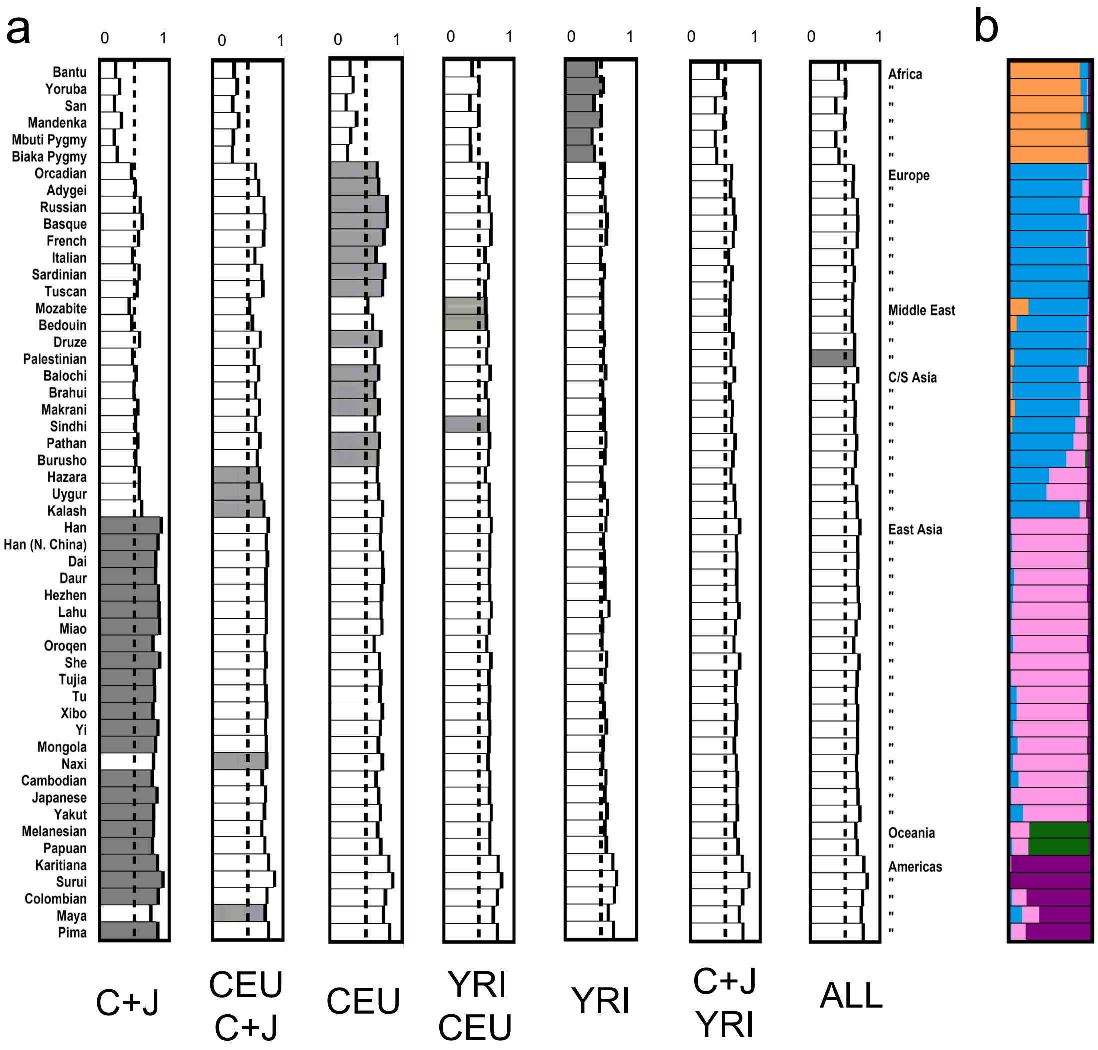
2. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., R.Kautzer, C., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N. P., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).

3. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).

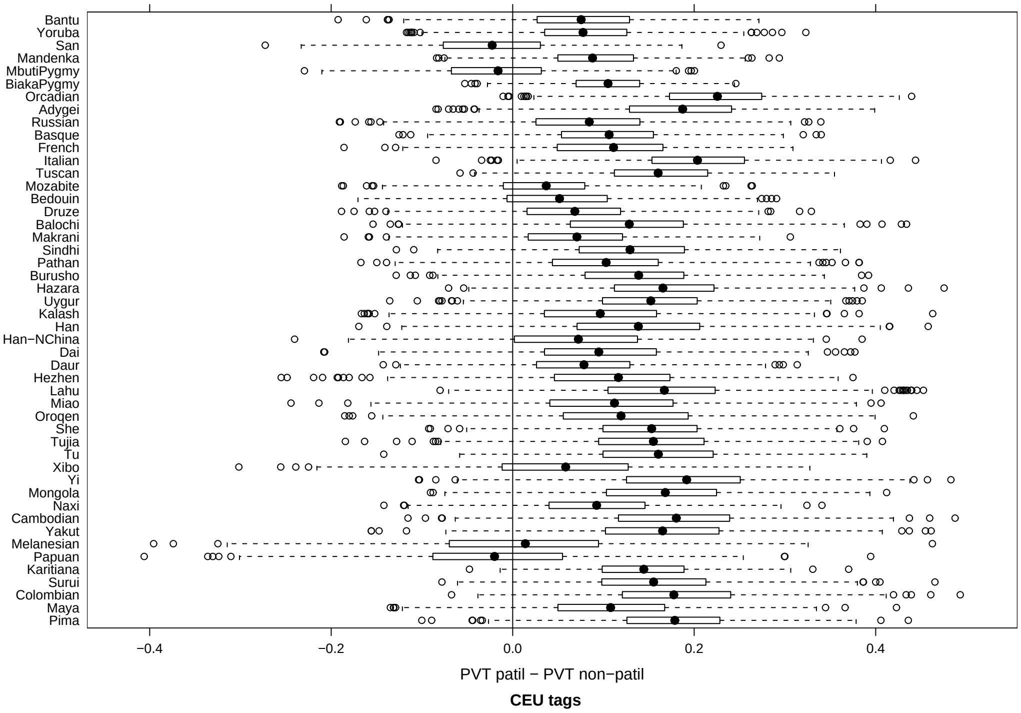
4. Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36**, 700–706 (2004).

5. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).

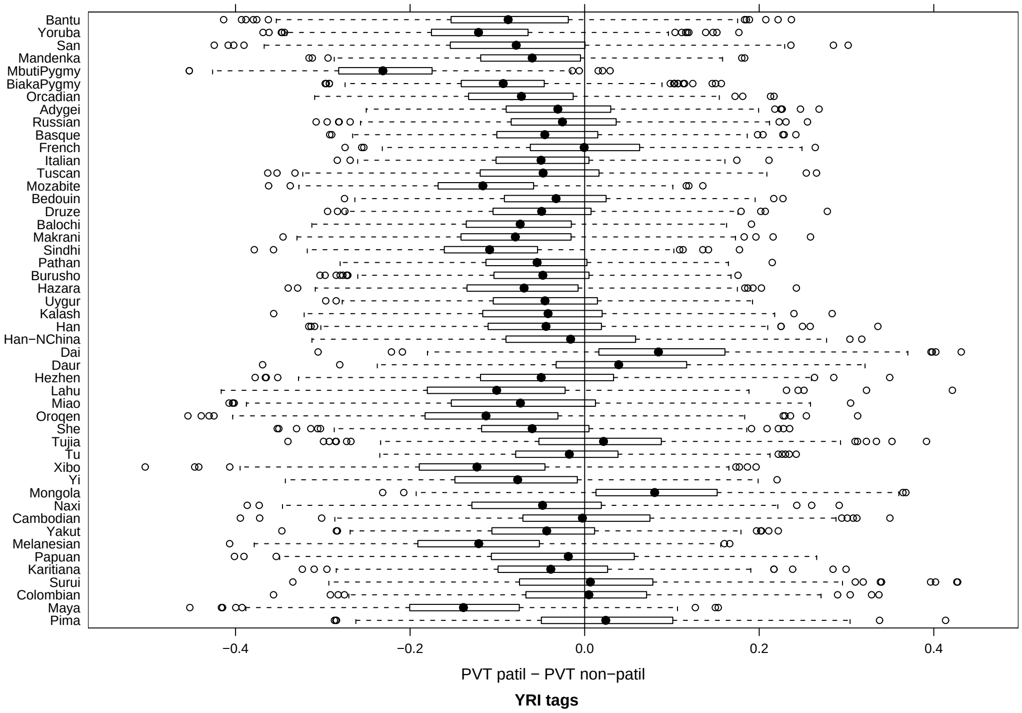
6. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2005).



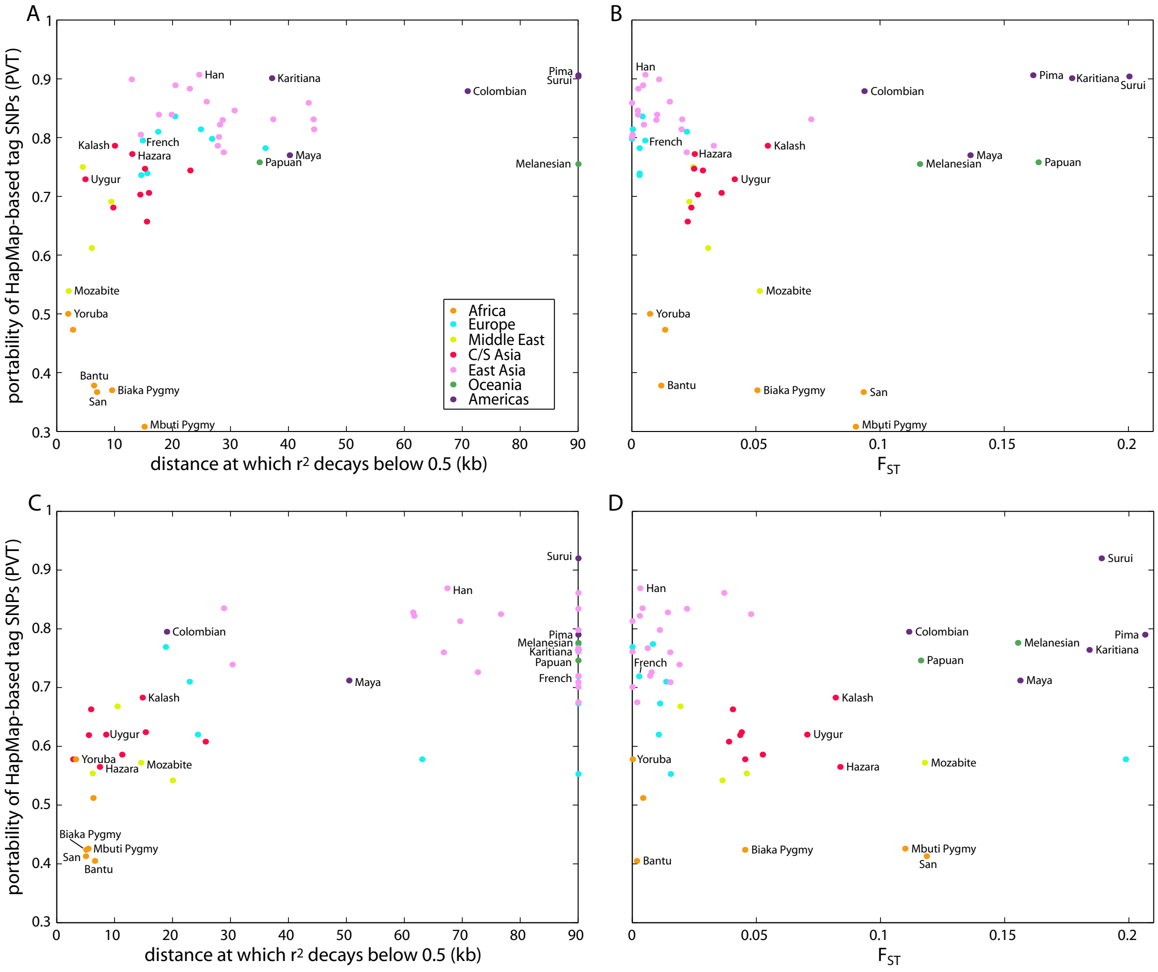
Supplementary Figure 1. Portability of tag SNPs from 7 different panels. **(A)** The first 7 panels show for each of 52 populations and each tag panel the proportion of polymorphic non-tag SNPs that have $r^2 > 0.85$ with at least one tag SNP. Panels were designed using haplotypes from a single HapMap population (CEU, YRI, CHB+JPT), pairs of HapMap populations, or 360 haplotypes from all three populations (ALL). For each population, the grey bar indicates which tag SNP set is best. **(B)** Estimated worldwide population structure based on microsatellite data from 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 5 clusters identified by the program *structure* (Rosenberg et al. 2002).



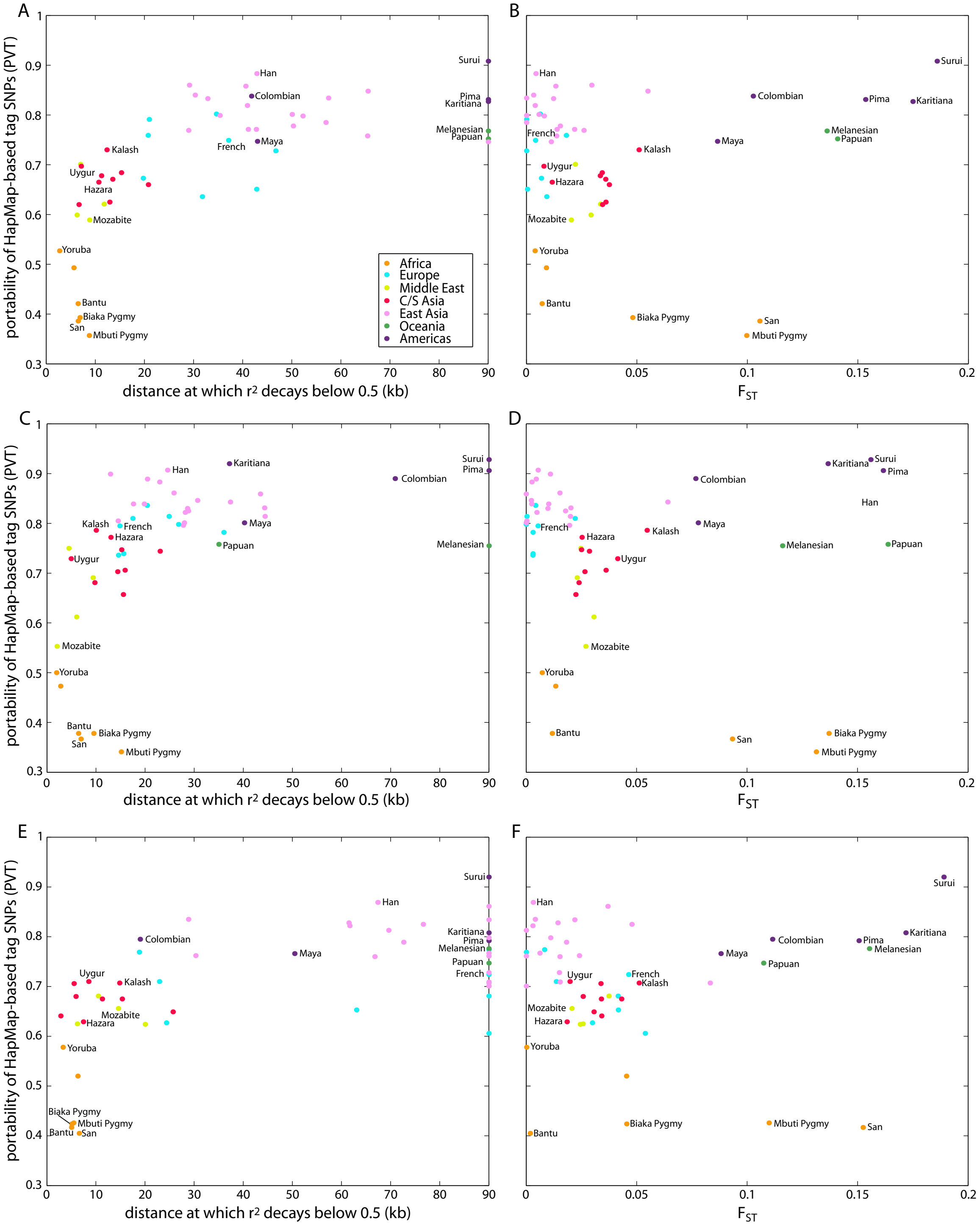
Supplementary Figure 2. Difference in PVT between Patil and Non-Patil regions, using CEU tags. Bootstrap resampling was used to assess the difference in mean PVT of Patil regions and non-Patil regions. Ten thousand bootstrap replicates were generated for each HGDP population, and are depicted here as box-whisker plots. Further details are provided in the Supplementary Note.



Supplementary Figure 3. Difference in PVT between Patil and non-Patil regions, using YRI tags. Bootstrap resampling was used to assess the difference in mean PVT of Patil regions and non-Patil regions. Ten thousand bootstrap replicates were generated for each HGDP population, and are depicted here as box-whisker plots. Further details are provided in the Supplementary Note.



Supplementary Figure 4. The relationships between [**First column**] tag portability and the distance at which the r^2 measure of linkage disequilibrium decays below 0.5, and between [**Second column**] 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. **(A)**,**(B)**: Only SNPs from the Patil regions (chromosome 21) were used. **(C)**, **(D)**: Only SNPs from the non-Patil regions (autosomal non-chromosome 21) were used. Further details are provided in the Supplementary Note.



Supplementary Figure 5. The relationships between [**First column**] tag portability and the distance at which the r^2 measure of linkage disequilibrium decays below 0.5, and [**Second column**] tag portability and F_{ST} genetic distance to the HapMap population that produces the highest tag portability. For each population, tag portability was computed as the maximum of the seven PVT values in Supplementary Figure 1. **(A)**, **(B)**: All autosomal regions were used. **(C)**, **(D)**: Only SNPs from the Patil regions (chromosome 21) were used. **(E)**, **(F)**: Only SNPs from the non-Patil regions (autosomal non-chromosome 21) were used. Further details are provided in the Supplementary Note.