Table of Contents

1. INTRODUCTION:

This Supplementary Material contains additional Tables, Figures and Methods to further support the accompanying manuscript: "***INTEGRATING COMMON AND RARE GENETIC VARIATION IN DIVERSE HUMAN POPULATIONS***" by the **International HapMap 3 Consortium**. The material is organized by section, corresponding to the organization of the main paper, and each section includes all the relevant material for that section. (NB: Supplementary Figure Legends are included in each section here: the Figures are included in a separate file'hapmap3_suppfigs').

Sample collection details: All of the samples (**Table S1**) were collected following extensive informed consent and community engagement processes. A template consent form was developed and adapted for use at each sampling site to make the document consistent with local cultural and social norms. An extensive process of community engagement was conducted at each site, to give members of the participating communities an opportunity to discuss issues of possible broader concern. No identifying, clinical, or phenotype information is available for these samples. Researchers may obtain the samples from the non-profit Coriell Institute for Medical Research (http://ccr.coriell.org/Sections/Collections/NHGRI/hapmap.aspx?PgId=266).

Methodologies for community engagement ranged from the use of extended, semi-structured individual interviews and focus groups to large public meetings and public attitudinal surveys. The processes were designed to elicit the views of a range of people within each community regarding a variety of issues relating to the HapMap Project and to genetic research more generally. Participants were given an opportunity to raise concerns about proposed recruitment methods, privacy and confidentiality risks, risks of discrimination and group stigmatization, policies regarding commercialization and intellectual property, and other topics.

As an outgrowth of these community engagement processes, a Community Advisory Group (CAG) was established in each donor community. The CAGs provided input into various aspects of the project, including how the samples from their populations should be labeled. The CAGs also serve as a liaison between the community and the Coriell Institute, where the samples are stored. The Coriell Institute provides them with quarterly reports that list the investigators who have requested their samples and the nature of the research those investigators plan to conduct with their samples.

HapMap 3, like all genetic variation research, carries the potential for group stigmatization and other ethical concerns. For example, if a variant found to be associated with a particular disease or trait has a higher frequency in groups from a particular geographical location, and if this information is over-generalized to all or most members of that group or to related groups, entire groups can be stigmatized. Stigmatization can also occur when reports of the findings of genetic association studies are not placed in context to make clear that non-genetic factors may also make important contributions to disease risk. Finally, an overemphasis on group allele frequency differences can (at least in some social and cultural groups) create the misleading impression that there are precise boundaries between groups of people, thus reinforcing racial or ethnic biases.

Investigators who reference HapMap 3 data or who order the samples included in the project for use in future studies are asked to be especially sensitive to the possible implications of their research for the sample donors and the communities and populations of which they are a part. Investigators are asked to describe the findings of their studies with care and attention to the potential broader implications of their research. Investigators are specifically asked to use the population labels (and abbreviations) listed in the main text when referring to these populations in future publications or presentations. See also http://www.hapmap.org/citinghapmap.html .

As in HapMap I[1] and II[2], the samples from some of the HapMap 3 populations were combined into analysis panels (for example, JPT+CHB+CHD, and CEU+TSI). These combined analysis panels reflect the similarities of the allele frequencies in the sets of samples. However, these analysis panels should not be confused with the populations themselves. None of the sample sets can be considered completely representative of a larger population, nor certainly of an entire continent. Thus, for example, references to the "African," "Asian," or "European" "populations" should be avoided when referring to these samples.

In addition, for this reason and to respect the preferences of the populations sampled regarding how they wished to be labeled, we recommend using a specific local identifier to describe a set of samples initially (for example, "Gujarati Indians in Houston, Texas"), and thereafter to use the designated abbreviation for that population (for example, GIH). Additional information relevant to the labeling of the HapMap 3 populations can be found in the Population Descriptions for each population, available at http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11.

Table S1. Numbers of samples successfully genotyped and sequenced for each population.

Population	Genotyping					ENCODE Sequencing
	Sample design	QC samples	Phased QC chromosomes	QC SNPs	QC SNPs polymorphic	QC samples / attempted
ASW	Trio	83	126	1,656,877	1,565,172	35 / 55
CEU	Trio	165	234	1,648,653	1,416,121	119 / 119
CHB	Unrelated	84	168	1,662,767	1,332,120	90 / 90
CHD	Unrelated	85	170	1,646,894	1,309,662	30 / 30
GIH	Unrelated	88	176	1,652,907	1,411,455	60 / 60
JPT	Unrelated	86	172	1,663,087	1,300,764	91 / 91
LWK	Unrelated	90	180	1,649,904	1,533,540	60 / 60
MXL	Trio	77	104	1,585,624	1,413,654	27 / 27
MKK	Trio, unrelated	171	286	1,635,780	1,541,375	0 / 0
TSI	Unrelated	88	176	1,655,975	1,423,618	60 / 60
YRI	Trio	167	230	1,652,198	1,505,108	120 / 120
Total		1184	2022	1,472,130	1,440,616	692 / 712

Definitions of genetic variant frequency classes used in this study: To achieve consistency and clarity, we used the variant frequency classes described in Table S2 throughout the entire study.

Table S2. Variant frequency classes

Name of Class	**Frequency Range**	**Population Issues**	**Technical Issues**	**History/Comment**
Common variants	≥ 5.0%	- Generally shared across global populations; - Often show high LD; - Highly amenable to imputation; - High tagSNP portability across populations.	Easy to discover in shallow sequence surveys.	Well studied in HapMap I and II.
Low frequency variants	0.5% – 5.0%	- Often shared between related populations but at variable frequencies; - Somewhat amenable to imputation.	Requires deep sequencing for discovery, but could be discovered by deep sequencing of other populations.	Inadequately sampled so far, the 1000 Genomes Project will find many such variants.
Rare variants	0.05% - 0.5%	- Often population specific; - Cannot be imputed easily; - Not readily tagged by other variants.	Requires deep sequencing in the specific population where it is found to be discovered.	Inadequately sampled.
Private variants	Singletons - 0.05%	- Typically private to individuals or families; - Frequent class among Mendelian disease (and *de novo* neutral variants).	Requires high precision for discovery, genotyping, etc.	Revealed in personal genomes and pedigree based family studies.

2. LARGE SCALE GENOTYPING

Array QC: Genotypes were called using BirdSeed[3] and Illumina's calling algorithms[4]**.** Array results were removed if they were of low quality (< 90% call rate) or redoes/duplications of lower call rate; in total, 233 arrays failed (160 Affymetrix, 73 Illumina). After this initial filtering, 1326 Affymetrix samples assayed at 909,622 SNPs (98.9% call rate) and 1211 Illumina samples assayed at 1,055,111 SNPs (99.6% call rate) were available for data merging.

For each SNP genotyped on both platforms, we designated the merged call as the consensus call if they were concordant and missing if they were not, as implemented in PLINK merge-mode 1[5]. The overall platform genotype concordance was 99.5% (across 250,000 overlapping SNPs) at a call rate of 99.8%. Genotypes were then aligned to the forward/(+) strand of genome build 36 and, using the array annotations, SNPs that did not map uniquely to the genome were removed. Due to ambiguity of strandedness, A/T and C/G SNPs that were present only on the Illumina array were also removed. Samples were discarded if they were discordant across platforms (< 95% concordance) or of low quality (< 95% merged call rate). SNP filtering was implemented on a population-specific basis: call rate < 95%, Hardy-Weinberg equilibrium pvalue < $1.0\text{x}10^{-6}$, > 2 Mendelian errors across all transmissions (only considered in ASW, CEU, MXL, MKK, and YRI).

Phasing Methods: Phasing was completed in two stages. During the first, family information (where available) was employed to deterministically resolve phase by transmission, where possible. During the second, sites with unresolved phase and missing data were phased statistically using IMPUTE v2[6]. For unrelated individuals, phasing was carried out using IMPUTE v2 using the phased trio parents as a reference panel. On average, 28 % (range 26.3% – 30.8 %) of the genotypes of each sample are heterozygotes and therefore require phasing. Missing data varies between 0.074 – 1.95%, and Mendel errors (for TRIOS) between 0.0127 - 0.139%. Family information allows about 80% of the heterozygotes to be deterministically resolved, and 75-87% of the missing data to be inferred. For TRIOS and DUOs, 94% and 85%, respectively, of heterozygous and missing alleles are deterministically resolved. Rates of heterozygosity among typed SNPs, missing data and Mendel errors were similar among populations.

IMPUTE v2 has been shown to perform well against other recently developed methods, when tested on unrelated samples[6]. Additional comparison of IMPUTE v2 performance against PHASE[7] on phasing chromosome 20 of CEU TRIOS showed that there is an average difference of 3.3% in the phasing outcome for alleles whose phase could not be deterministically resolved. IMPUTE v2 returns posterior probabilities for the phasing of each allele, which were used to resolve phase without overriding the family information.

Parental genotypes of TRIO samples were phased without a reference panel, with the exception of ASW. A combined CEU and YRI TRIOS reference panel was used for ASW TRIOS, due to the small sample size for that panel. Genotypes from DUO and UNR samples were phased using the phased TRIOS of the same population as reference, where available. Phased haplotypes of CEU TRIOS were used for GIH, TSI, CHD, CHB and JPT samples, and phased haplotypes from YRI TRIOS were used for LWK samples. The effective population sizes used were 17094 for YRI, ASW, MKK and LWK, and 11418 for CEU and TSI (estimates from HapMap Phase II). For other populations a value of 15000 was used, after experiments showed that the phasing results are insensitive to differences within a factor of 2. 110 iterations were used, with 120 conditioning states. Unrelated samples were phased in blocks of approximately 8,000 SNPs, due to memory requirements. Both IMPUTE v2 and our routines used additional SNPs at both flanks to account for edge effects and combined the phased SNPs into one file per chromosome.

The phased haplotypes can be found online at http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/, split by population and by family status (TRIOS, DUOS, UNR). Additional details on the phasing process and on naming conventions can be found at the same location, in the file hapmap3_r2_phasing_summary.doc.

3. RARE ALLELE CALLING BIAS

Ever since the first large-scale genome-wide association studies began, a subtle technical bias has been consistently observed against rare alleles. This effect manifests itself most obviously in family-based studies as a systematic bias against transmission of rare alleles.

While both missing data[8] and genotyping errors[9] can lead to artificial under-transmission of rare alleles, these biases have been surprisingly evident in GWAS data even after very stringent data cleaning procedures, in part explicable because Mendelian inheritance and departure from Hardy-Weinberg equilibrium are not powerful screening tools for low-frequency variants.

For example, in a recent GWAS of autism[10] with ~1,200 trios (using the Affymetrix 5.0 array), when looking at QC-passing SNPs with MAF < 5% and call rate > 98%, the authors observed 19,291 SNPs had the minor allele over-transmitted, but 27,112 SNPs had the minor allele under-transmitted; this is an astronomically significant departure from the expected 50-50 split between over- and under-transmission of any particular allele category. Likewise, the GAIN-ADHD study[11] done at Perlegen has reported concordant, highly significant biases, suggesting these artifacts are not obviously specific to any particular genotyping platform.

We show here that the bias against calling the minor allele of rare variants seems to affect Affymetrix and Illumina arrays equally in the HapMap 3 genotype data. We evaluate a TDT test of the CEU trios using PLINK[5]. While rare SNPs do not show any highly significant associations given the small sample size, we can look at their bulk properties across the genome to observe unusual distortion.

If we take SNPs whose minor allele occurs only in exactly two heterozygous parents in the CEU sample (roughly 1% MAF) assuming calling is complete and perfect, we should expect a 25% - 50% - 25% transmission proportion that corresponds to 2-0 (minor allele over-transmitted), 1-1, and 0-2 (major allele over-transmitted), respectively. Instead, we observe highly significant deviations from this expectation on both platforms (post-QC and ignoring the SNPs that appear on both platforms) (**Table S3a**). Similarly, we see far more 0-3 (major allele over-transmissions) than 3-0 (minor allele over-transmissions) on both platforms (**Table S3b)** for SNPs with a total of three heterozygous parents**.** As expected, the skew is reduced at higher minor allele frequency, and the bias begins to become more distributed and is no longer trivially "visualized" (that is, more 1-3 than 3-1, 1-2 than 2-1, etc.). Summarizing across all SNPs in the HapMap 3 data, we observe a highly significant excess of TDTs with OR < 1 for both platforms (**Table S3c).**

Table S3.

a.

Platform	Observed transmissions (minor overtransmission /heterozygote/major overtransmission)	Observed transmissions (%)
Affymetrix 6.0	1456 – 3130 – 1910	22.4 – 48.2 – 29.4
Illumina 1M	1244 – 2758 – 2026	20.6 – 45.8 – 33.6

b.

Platform	Observed transmissions (minor 3-0/ major 0-3)	P-value (binomial departure from 50-50)
Affymetrix 6.0	1117 – 1417	1.7×10^{-9}
Illumina 1M	938 – 1290	4.6×10^{-14}

c.

Platform	OR > 1	OR < 1
Affymetrix 6.0	213,929	223,657
Illumina 1M	278,561	291,976

The overall observation is that there is no significant difference between Affymetrix and Illumina in terms of this bias, but that bias against rare alleles is still evident in these data. This has important implications for family-based association studies. The evaluation of this bias as a function of allele frequency and number of samples included in the genotype clustering is likely an important follow-up for genotype calling methods, particularly as we consider advancing genotyping arrays to incorporate rarer genetic variants.

4. DEEP PCR SEQUENCING

SNP discovery methods and QC filters applied in the regional resequencing data: PCR amplification reactions were overlapped with each reaction spanning ~600-700 bases. SNPs were discovered in the raw sequence data using 'SNP Detector 3.0' software[12] and the data filtered by removing low quality sequence reads, amplicons with too many polymorphic sites (an indication of noisy sequencing), and polymorphic sites with conflicting allelic calls. After the initial filtering, we identified 11,399 polymorphic sites, among which 10,076 were bi-allelic SNPs. We next implemented a SNP QC procedure similar to that used previously in HapMap Phase I and II. The specific QC filters included a) sample quality (outliers were identified with significantly low SNP call rate), b) completeness > 80% for each SNP in each population, and c) Hardy-Weinberg equilibrium p > 0.001. After the 'HapMap style' QC step, 20 samples were removed due to their significantly low SNP call rates, and 5,758 bi-allelic SNPs passed the filters.

We also implemented a "qualitative genotype confidence score system" to indicate various levels of stringencies used when calling different categories of genotypes. Specifically, a genotype labeled with a caret sign ("^") signifies the genotype called based solely on the chromatograms of its own DNA sample, which is theoretically more stringent in calling a minor allele for a particular sample, by not depending on other incidences of the minor alleles in the interrogated sample collection, and therefore not relaxing the thresholds in calling a minor allele. This step was especially important to improve the quality of rare allele calls (data not shown). Different number of asterisks were also used: "***" representing homozygous genotypes of major alleles that show the highest confidence, and "**" and "*" representing genotypes with minor alleles (either homozygous or heterozygous) that show intermediate and low confidence respectively (**Table S5a**). The genotypes annotated with quality scores of "^", "***" and "**" were later used in the analyses shown in this paper. They represent the bulk of the data set and provide robust genotype calls with the genotype concordance rates at 92.5%, 99.8% and 85.2% respectively, estimated by compared to the genotypes in the Broad genotyping validation experiment using Sequenom.

In total, 77% of the discovered SNPs were novel (i.e., not in dbSNP build 129), and 99% of those had a MAF < 5%. The known SNPs on average account for 86% of heterozygosity, ranging from 77% in LWK to 90% in CHD.

Table S4. ENCODE regions sequenced

Region	Chromosome	Coordinates (NCBI build 36)	Status	# SNPs in Release
ENm010	7	27,124,046 – 27,224,045	ENCODE I and III	1041
ENr321	8	119,082,221-119,182,220	ENCODE I and III	1098
ENr232	9	130,925,123-131,025,122	ENCODE I and III	840
ENr123	12	38,826,477-38,926,476	ENCODE I and III	748
ENr213	18	23,919,232-24,019,231	ENCODE I and III	899
ENr331	2	220,185,590-220,285,589	ENCODE III	0
ENr221	5	56,071,007-56,171,006	ENCODE III	567
ENr233	15	41,720,089-41,820,088	ENCODE III	28
ENr313	16	61,033,950-61,133,949	ENCODE III	0
ENr133	21	39,444,467-39,544,466	ENCODE III	460

ENCODE3 validation experiments: We assessed the genotyping accuracy for SNPs in three classes: 1) low frequency SNPs seen in multiple divergent populations from different continents; 2) low frequency SNPs ascertained only in a single population; and 3) already known, mostly common SNPs. The first category can be expected to have the lowest accuracy, because real SNPs of this type are unusual and false positives will be a larger fraction of the total. The second category should be representative of SNPs seen with low frequency in a certain population.

The datasets used for assessing accuracy for the three categories were as follows. 1) 100 SNPs that had 2 – 4 copies of the minor allele, spread across at least two divergent populations were chosen and re-genotyped by BCM-HGSC using Roche 454 pyrosequencing technology; 2) 500 SNPs genotyped by Broad using Sequenom; these were chosen to have 2 – 6 copies of the minor allele in either CEU or YRI; and 3) all SNPs that also appear in the HapMap 3 chip data were compared to measure concordance. The results are shown in **Table S5b**. The final data set showed high validation rates. For rare SNPs, the genotype concordance rate was 88% and the SNP validation rate was 89. For rare SNPs spanning multiple populations, the genotype concordance rate was 73% On a per-SNP basis, the validation rate improved to 89%. For SNPs that were already identified and included on the HapMap 3 chips (that is, mostly the common SNPs), the genotype concordance rate was 99.23% (**Table S5b**).

In addition, we assessed the validation rates of genotypes with minor alleles, as a function of their minor allele frequency, by comparing to the 1000 Genomes Project Illumina genotype chip validation data (www.1000genomes.org) that overlapped with ENCODE by 293 samples and 3,350 SNP sites. Overall, the validation rates were in concordance with the results obtained from Sequenom (for rare SNPs) and HapMap data sets (for common SNPs) and showed an exceedingly high genotype concordance rate (**Table S5**c). The lower validation rate in SNPs with 40% < MAF ≤ 50% (79%) reflects the lower stringency threshold applied in the calling the homozygous reference genotypes.

Table S5

a.

	Homozygotes major allele			Heterozygotes			Homozygotes minor allele			
SNP call category	#genotype	quality	annotation	#genotype	quality	annotation	#genotype	quality	annotation	total
1	478	high	^	43389	high	^	12353	high	^	56220
2	3742631	high	***	97057	medium	**	24457	medium	**	3864145
3	1180	high	***	22	low	*	15	low	*	1217

b.

	Rare SNPs spanning ≥ 2 continents (Baylor 454 validation)		Rare SNPs (Broad genotype data)			Common SNPs (compared to HapMap 3)	
Concordance	SNP validation (%)	Genotypes with minor alleles (%)	SNP validation (%)	All genotypes (%)	Genotypes with minor alleles (%)	All genotypes (%)	Genotypes with minor alleles (%)
ENCODE data	85	73	89	99.5	88	99.2	86.8

c.

	Concordance rate for genotypes with minor alleles (%)
Minor allele count = 1	93.6
Minor allele count = 2	93.1
Minor allele count = 3	91.2
Minor allele count = 4	84.4
Minor allele count = 5	89.0
Minor allele count <= 10	91.0
Minor allele frequency <= 10%	90.8
Minor allele frequency <= 205	89.3
Minor allele frequency <= 30%	90.4
Minor allele frequency <= 40%	86.3
Minor allele frequency <= 50%	79.0

5. COPY NUMBER VARIATION ANALYSIS

CNP discovery used two algorithms, QuantiSNP[13] (QS) and Birdseye[3] (BE), that enable joint discovery using the combined dataset while modeling data from each array platform (Affymetrix 6.0 and Illumina 1M) separately. In regions of the genome where data from only a single platform present, discovery was based on the available data[14]. We used comparisons with much higher resolution tiling-oligo Comparative Hybridization data made available by the Human Genome Structural Variation consortium on an overlapping set of 34 individuals to define confidence thresholds for each discovery algorithm to obtain an estimated FDR of ~10%. For an FDR of 10% the determined threshold for QS was log (Bayes Factor) > 18 which resulted in 57,589 autosomal calls (mean 47 per sample). Similarly, in BE a threshold of log(Odds Ratio) > 3 gave an approximate FDR of 10%, resulting in 60,512 autosomal calls (mean 51 per sample).

These sample-specific calls were then collapsed into discrete CNP segments. For our subsequent analysis, we focused on variation that was observed in at least 1% of the samples (reflecting a putative minor allele frequency > 0.5%). In order to refine the CNP breakpoint definitions using many samples simultaneously, we developed an approach utilizing the correlation structure of the probe-intensity data across samples. First, we agglomeratively clustered overlapping CNP calls to identify a series of discrete regions for more-detailed follow-up. We then analyzed each such region (together with 100 kb of flanking sequence on each side) individually. Each region involved a set of samples with putative CNPs; for the following analysis of that region, we utilized those samples together with an equal number of randomly selected samples. We built a probe-by-probe correlation matrix for the region, with each entry in the matrix containing the Pearson correlation of the intensity measurements for those two probes (across the selected set of samples). We identified CNP regions as square submatrices (symmetrical over the diagonal) of statistically significant ($p < 10^{-4}$) positive correlation.

To genotype these CNP regions (determine integer copy number per diploid genome) in the HapMap 3 samples, we used two algorithms. The “one-dimensional” approach utilized a previously published method, CNVtools[15], adapted to allow fitting mixtures of Student t distributions. A novel, two-dimensional genotyping approach treated the data as bivariate with the X,Y axes representing the Affymetrix and Illumina signals respectively. A two-dimensional Gaussian mixture model was fit to determine the most likely copy number assignments.

To critically evaluate and combine data from the one- and two-dimensional approaches to CNP genotyping, we then developed the following meta-approach. We generated draft genotype-cluster assignments using each approach (one-dimensional and two-dimensional clustering) separately. We removed (from each data set) CNPs that had call rates less than 90% or minor allele frequency less than 0.5%. For CNPs that had qualified genotype calls using both approaches (90% of CNPs), we then compared these call sets. For 96% of these CNPs, the genotype calls were concordant between one- and two-dimensional clustering (discrepancies in < 1% of samples); we combined the data sets by accepting concordant calls and changing discordant calls to no-calls. For the remaining 4% of CNPs, which showed more discrepancies between one- and two-dimensional clustering, we selected one call set over the other based on the following tiered criteria (with ties broken by dropping to the next criterion): (1) lowest rate of deviation from Mendelian inheritance in trios; (2) lack of significant ($p < 0.01$) Hardy-Weinberg test statistic in any population; (3) maximum average genotype confidence, with confidence inferred by fitting the intensity data and genotype calls to a Gaussian mixture model.

6. POPULATION ANALYSES

We characterized the relationships among the populations by using the SNP genotype of 988 unrelated individuals to carry out a principal components analysis (PCA) using the EIGENSOFT software[16] (**Figure S2**). PCA results indicate that CEU, TSI, YRI, JPT, CHB, and CHD are of relatively homogeneous ancestry (**Figure S2a,b**), while ASW, MKK, LWK, MXL and GIH are admixed populations in which individuals have varying continental ancestry proportions (**Figure S2a, c, and d**). (One ASW sample, NA19625, appeared to have a contribution of East Asian-related ancestry and was removed from subsequent PCA analyses.) A PCA run of CEU, TSI, YRI, JPT, CHB, and CHD (**Figure S2b**) confirms that these six populations have homogeneous continental ancestry and suggests that for many purposes the populations of European ancestry can be grouped together (CEU+TSI), and similarly the populations of East Asian ancestry (JPT+CHB+CHD). This is further supported by a low F_{ST} of 0.004 between CEU and TSI, and of 0.001, 0.008, and 0.007 between CHB and CHD, JPT and CHB, and JPT and CHD, respectively (**Table S6**). An analysis of each population separately (data not shown) indicates that while CEU, TSI, YRI and JPT are very homogeneous, CHB and CHD each show very subtle population structure, consistent with previous findings[**2**]. However, the deviation from homogeneity is slight.

For ASW, MKK, LWK, and YRI, which have genetic proximity to Africa, with F_{ST} only as high as 0.027 between each pair of these populations (**Table S6**), we ran PCA together with CEU (**Figure S2c**). ASW individuals occupy a range between YRI and CEU but are closer to YRI. To estimate admixture proportions, we approximated ASW allele frequencies as a mixture of YRI and CEU allele frequencies, which resulted in estimates of 78% African and 22% European ancestry, consistent with previous studies of African-American ancestry[17,18]. However, we note that a very high variability of admixture proportions between ASW individuals is suggested both by the PCA analysis (**Figure S2c**) and by F_{ST}: While F_{ST} between ASW and CEU is 0.102, individual ASW samples exhibit an F_{ST} as low as 0.053 and as high as 0.142 from the CEU population (**Table S6**).

MKK individuals occupy a wide range between YRI and an unsampled population, suggesting that these individuals are of admixed ancestry, likely with an unsampled East African ancestral component and a West African ancestral component that is captured by YRI. We hypothesize that (1) the position of the unsampled East African ancestral population on PC1 (**Figure S2c**)—lying somewhat in the direction of CEU—may be the result of an ancient Neolithic farming migration from Europe or the Middle East into

East Africa[19]; (2) the variation in the amount of YRI-related ancestry in MKK—resulting in a wide range of F_{ST} of 0.006 to 0.043 between MKK individuals and the YRI population (**Table S6**)—may be the result of the Bantu expansion from West Africa, which reached some parts of East Africa quite recently[19]; and (3) the position of some MKK samples being closer to CEU than would be expected based on their position on the YRI-related cline may be the result of recent Arab admixture in East Africa[19]. The same patterns are evident to a lesser extent in the LWK individuals, except that the LWK show no evidence of recent Arab admixture and lie much closer to YRI on the YRI-related cline. F_{ST} between LWK and YRI is 0.008 (compared to 0.027 between MKK and YRI), and ranges between 0.002 and 0.014 among LWK individuals. This is consistent with the Bantu (West African origin) linguistic affiliation of the LWK as opposed to the Nilotic (East African origin) linguistic affiliation of the MKK; however, studies of other East African populations have shown that population relationships are not always concordant with linguistic affiliations[20]. Since the level of admixture in LWK is relatively slight, it may be acceptable to group LWK with YRI in some analyses.

For MXL and GIH, which are admixed populations with genetic proximity to Europe, F_{ST} shows a wide range of admixture proportion: F_{ST} between MXL and CEU is 0.031, and ranges between 0 and 0.077 among MXL individuals; F_{ST} between GIH and CEU is 0.035, and ranges between 0.017 and 0.049 among GIH individuals (**Table S6**). We ran PCA of MXL, GIH, and CEU, which supports a very wide range of admixture proportions of MXL samples (**Figure S2d**) and is consistent with recent admixture[21]. PCA supports a wide range of admixture proportions of GIH samples as well (**Figure S2d**), which is unlikely to be due to recent admixture[22], but instead may be the result of ancestry from multiple Gujarati populations with varying levels of ancient European-related admixture. Indeed, PCA of CEU and GIH alone clearly splits GIH into two distinct clusters, consistent with ancestry from multiple Gujarati populations (**Figure S3a**). Lastly, joint analyses with CHB indicate that for both MXL and GIH, the non-European admixture component is distinct from East Asia (**Figures S3b, c**).

We ran the HAPMIX algorithm[23] to produce local ancestry estimates (0, 1 or 2 copies of European-related ancestry at each location in the genome) for ASW, MKK and LWK, using CEU and YRI as reference populations. We verified previous work showing that African-Americans are accurately modeled as a linear combination of CEU and YRI by computing an F_{ST} of 0.001 between ASW and the optimal linear combination of 79% YRI and 21% CEU (nearly identical to the admixture calculated above). For MKK and LWK, our PCA results suggested that they were less accurately modeled by YRI and CEU. Indeed, we computed F_{ST} values of 0.014 between MKK and the optimal linear combination of 74% YRI and

26% CEU, and 0.006 between LWK and the optimal linear combination of 94% YRI and 6% CEU. However, HAPMIX has been shown to produce accurate local ancestry estimates even when the reference populations used are somewhat inaccurate, with an F_{ST} from the true ancestral populations as large as 0.0213. We found that chromosomal segments of European-related ancestry typically spanned megabases in MKK and LWK, while spanning tens of megabases in ASW, consistent with African-American admixture being more recent.

We evaluated the coverage that HapMap 3 provides of worldwide genetic diversity by comparing HapMap 3 data to data from the Human Genome Diversity Project[24-27]. We ran PCA on a set of SNPs that were genotyped for both the HGDP sample[24] and the HapMap 3 sample by restricting analysis to Illumina 650Y SNPs. Coverage of worldwide genetic diversity as captured by the top six principal components is similar for the two data sets (**Figure S4**). At this level of granularity, the main differences are that Oceanian diversity (Papuan and Melanesian) is covered by HGDP but not by HapMap 3 (principal component 4; **Figure S4b**) and that non-Bantu East African diversity (MKK) is covered by HapMap 3 but not by HGDP (principal component 6; **Figure S4c**). The ancestries of most other HGDP populations that were not sampled in HapMap 3 are still captured by admixed populations. For instance, Native American ancestry is represented by the admixture component of MXL (principal component 3; **Figure S4b**). Additional principal components would no doubt reveal much fine structure in the HGDP's wider range of populations that is invisible in HapMap 3.

Table S6. F_{ST} between each pair of populations (symmetric). Estimates are based on all autosomal SNPs in the genotype data, considering only unrelated individuals. Standard errors (in parentheses) are based on 1,000 moving block bootstraps in order to account for the dependency due to linkage disequilibrium[16]. After the pairwise F_{ST} value, the table provides the range of F_{ST} across all unrelated individuals in one population (indicated by the row), which is based on estimating F_{ST} between each individual in that population and the entire sample from each other population (indicated by the column) in a way that is not biased by sample size differences between the two samples[16]. The range of F_{ST} estimates points to variation in ancestry among individuals in one (row) population as far as this ancestry is related to the second (column) population.

	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MXL	MKK	TSI	YRI
ASW		.1018 (.0006) .053 - .142	.1419 (.0007) .090 - .172	.1429 (.0007) .091 - .173	.0947 (.0005) .050 - .129	.1433 (.0007) .091 - .173	.0100 (.0001) .002 - .020	.0938 (.0005) .043 - .129	.0145 (.0005) .003 - .022	.0988 (.0005) .051 - .138	.0092 (.0004) .000 - .023
CEU	.1018 (.0006) .094 - .119		.1105 (.0007) .102 - .131	.1123 (.0007) .104 - .132	.0349 (.0003) .026 - .056	.1124 (.0007) .104 - .132	.1457 (.0008) .138 - .162	.0310 (.0001) .022 - .052	.1034 (.0005) .096 - .121	.0040 (.0001) .000 - .025	.1573 (.0007) .150 - .174
CHB	.1419 (.0007) .136 - .148	.1105 (.0007) .103 - .118		.0010 (.0001) .000 - .008	.0761 (.0006) .069 - .082	.0070 (.0001) .000 - .012	.1751 (.0007) .169 - .182	.0692 (.0005) .061 - .075	.1428 (.0007) .137 - .149	.1108 (.0007) .104 - .117	.1853 (.0007) .179 - .192
CHD	.1429 (.0007) .135 - .154	.1123 (.0007) .104 - .126	.0010 (.0001) .000 - .019		.0768 (.0006) .068 - .090	.0080 (.0001) .000 - .027	.1759 (.0008) .169 - .187	.0709 (.0005) .063 - .083	.1436 (.0007) .136 - .155	.1122 (.0007) .104 - .125	.1862 (.0007) .179 - .197
GIH	.0947 (.0005) .087 - .107	.0349 (.0003) .017 - .049	.0761 (.0006) .070 - .093	.0768 (.0006) .071 - .095		.0773 (.0005) .072 - .095	.1321 (.0006) .126 - .144	.0350 (.0002) .024 - .049	.0946 (.0005) .087 - .107	.0340 (.0002) .017 - .048	.1434 (.0007) .137 - .156
JPT	.1433 (.0007) .136 - .163	.1124 (.0007) .105 - .132	.0070 (.0001) .000 - .032	.0080 (.0001) .000 - .034	.0773 (.0005) .070 - .099		.1764 (.0008) .169 - .196	.0700 (.0005) .063 - .091	.1442 (.0007) .137 - .164	.1125 (.0007) .105 - .132	.1866 (.0009) .179 - .206
LWK	.0100 (.0001) .004 - .016	.1457 (.0008) .135 - .153	.1751 (.0007) .165 - .182	.1759 (.0008) .165 - .182	.1321 (.0006) .122 - .139	.1764 (.0008) .166 - .183		.1329 (.0006) .122 - .140	.0170 (.0001) .008 - .024	.1415 (.0007) .130 - .148	.0080 (.0001) .002 - .014
MXL	.0938 (.0005) .077 - .128	.0310 (.0001) .000 - .077	.0692 (.0005) .056 - .135	.0709 (.0005) .057 - .136	.0350 (.0002) .013 - .065	.0700 (.0005) .057 - .136	.1329 (.0006) .115 - .166		.0958 (.0005) .079 - .131	.0320 (.0002) .000 - .080	.1434 (.0007) .125 - .176
MKK	.0145 (.0005) .001 - .023	.1034 (.0005) .082 - .122	.1428 (.0007) .126 - .158	.1436 (.0007) .127 - .159	.0946 (.0005) .073 - .111	.1442 (.0007) .128 - .160	.0170 (.0001) .000 - .031	.0958 (.0005) .076 - .112		.0980 (.0006) .076 - .117	.0270 (.0001) .006 - .043
TSI	.0988 (.0005) .092 - .106	.0040 (.0001) .000 - .012	.1108 (.0007) .103 - .119	.1122 (.0007) .104 - .120	.0340 (.0002) .027 - .042	.1125 (.0007) .105 - .121	.1415 (.0007) .136 - .149	.0320 (.0002) .025 - .039	.0980 (.0006) .092 - .105		.1532 (.0006) .147 - .161
YRI	.0092 (.0004) .005 - .017	.1573 (.0007) .150 - .164	.1853 (.0007) .179 - .193	.1862 (.0007) .180 - .194	.1434 (.0007) .137 - .150	.1866 (.0009) .181 - .195	.0080 (.0001) .004 - .015	.1434 (.0007) .136 - .150	.0270 (.0001) .022 - .033	.1532 (.0006) .145 - .160	

7. RECURRENT SNPS

All ENCODE3 SNPs (n=5,758) were filtered for those with only 2-6 copies of the minor allele that were present in at least two different HapMap populations. In total 862 SNPs were parsed out as rare recurrent variants in at least two different populations. Subsequently, we compared the haplotypes to identify the potentially different backgrounds, which might suggest that some SNPs had arisen independently in different lineages. Haplotype phasing was done for each population using fastPHASE 1.2[28]. Haplotypes of each of the individuals carrying rare SNPs were aligned and analyzed using a window size of 21 SNPs, including 10 flanking SNPs on either side of the rare SNP loci. Rare SNPs were considered to be in different haplotypes if haplotypes were less than 85% identical and differed in at least one SNP at the 4 positions that were immediately flanking the tested rare SNP.

After applying these criteria, 51 SNPs were identified as candidates. The sequencing chromatograms for all these SNPs were visually examined and 78% (40/51) of the rare SNPs were confirmed to be in different haplotypes in different populations. The average percent identity for the haplotypes including these rare SNPs was 83% (**Table S7**). These SNPs were considered as putatively independent mutations that arose in different ancestral haplotype backgrounds. The time of occurrence could be recent after the time of the population split; they are good candidates for independent occurrence of mutation at the same site.

TABLE S7: **SNPs with evidence for recurrence**.

#SNP_id	Visually verified	Maj_allele	Min_allele	Chr	Position	Strand	CpG?	DiffHap?	Av%Id	Populations
EN_9614523_328_SD3_1	Yes	C	G	chr12	38879713	+	no	yes	85	LWK=1 YRI=1
EN_9628617_83_SD3_1	No	C	T	chr7	27126286	+	yes	yes	85	ASW=2 CEU=1 GIH=1 LWK=1
EN_9629032_346_SD3_1	Yes	C	T	chr7	27154132	+	yes	yes	85	CEU=2 TSI=1
EN_9629032_371_SD3_1	Yes	C	T	chr7	27154157	+	yes	yes	80	ASW=1 LWK=1
EN_9629180_572_SD3_1	Yes	C	T	chr7	27162249	+	yes	yes	85	ASW=1 TSI=1
EN_9629192_172_SD3_1	Yes	C	G	chr7	27163861	+	no	yes	78	ASW=1 LWK=1 YRI=1
EN_9629192_312_SD3_1	Yes	C	A	chr7	27164001	+	no	yes	76	ASW=1 LWK=1
EN_9629355_110_SD3_1	Yes	A	G	chr7	27174452	+	no	yes	85	CHB=1 JPT=1
EN_9630073_167_SD3_1	Yes	T	G	chr12	38831407	+	no	yes	85	CHB=1 MXL=1
EN_9630073_357_SD3_1	Yes	C	T	chr12	38831597	+	yes	yes	80	ASW=1 YRI=1
EN_9630082_648_SD3_1	No	T	A	chr12	38833681	+	no	yes	85	ASW=1 CEU=2
EN_9630325_93_SD3_1	Yes	A	G	chr12	38854305	+	no	yes	85	CEU=1 TSI=1
EN_9630763_590_SD3_1	Yes	T	C	chr12	38881031	+	no	yes	85	LWK=1 YRI=4
EN_9630873_48_SD3_1	Yes	A	G	chr12	38891471	+	no	yes	85	CEU=2 GIH=2 TSI=1
EN_9631003_291_SD3_1	Yes	C	T	chr12	38897011	+	yes	yes	85	GIH=1 TSI=1
EN_9631193_116_SD3_1	Yes	A	G	chr12	38912452	+	no	yes	80	LWK=1 YRI=1
EN_9631193_118_SD3_1	Yes	T	C	chr12	38912454	+	no	yes	76	LWK=1 YRI=1
EN_9631193_497_SD3_1	Yes	T	G	chr12	38912833	+	no	yes	85	LWK=1 YRI=1
EN_9631297_593_SD3_1	No	C	T	chr12	38921230	+	yes	yes	85	ASW=1 LWK=1
EN_9633097_529_SD3_1	No	T	G	chr18	23929088	+	no	yes	83	CEU=2 TSI=1 YRI=2
EN_9633159_495_SD3_1	No	A	G	chr18	23931555	+	no	yes	83.33	CEU=2 TSI=1 YRI=1
EN_9633171_99_SD3_1	Yes	G	A	chr18	23934059	+	yes	yes	80	CHD=1 GIH=1
EN_9633586_514_SD3_1	Yes	C	A	chr18	23969114	+	no	yes	85	ASW=1 LWK=1 YRI=3
EN_9633592_317_SD3_1	Yes	C	T	chr18	23970746	+	yes	yes	85	TSI=1 YRI=4
EN_9634045_344_SD3_1	Yes	G	A	chr18	23996542	+	yes	yes	85	LWK=1 YRI=3
EN_9634154_83_SD3_1	Yes	C	T	chr18	23997568	+	yes	yes	85	CEU=1 GIH=1 YRI=4
EN_9635060_261_SD3_1	No	C	T	chr9	130943348	+	yes	yes	85	CHB=1 YRI=1
EN_9635060_406_SD3_1	Yes	C	T	chr9	130943493	+	yes	yes	85	CHB=1 JPT=1
EN_9635276_579_SD3_1	No	G	A	chr9	130955546	+	yes	yes	82.5	CEU=1 JPT=1 LWK=1
EN_9635282_424_SD3_1	Yes	A	G	chr9	130956443	+	no	yes	85	CHB=1 JPT=1

EN_9635291_251_SD3_1	Yes	G	A	chr9	130957796	+	yes	yes	80	ASW=1 JPT=1
EN_9635534_560_SD3_1	No	T	G	chr9	130974079	+	no	yes	85	ASW=1 GIH=1 JPT=2
EN_9635688_35_SD3_1	Yes	A	C	chr9	130982924	+	no	yes	85	CHD=2 JPT=2
EN_9636091_55_SD3_1	Yes	G	T	chr9	131010837	+	no	yes	82.5	ASW=1 GIH=1 LWK=1
EN_9636254_421_SD3_1	Yes	C	T	chr9	131022080	+	yes	yes	80	CEU=3 GIH=1 TSI=2
EN_9639518_342_SD3_1	Yes	C	T	chr8	119093503	+	yes	yes	85	ASW=1 CEU=1
EN_9639719_431_SD3_1	No	G	A	chr8	119125502	+	yes	yes	76	CHB=1 LWK=1
EN_9640004_335_SD3_1	Yes	G	T	chr8	119173720	+	no	yes	85	CEU=1 GIH=1
EN_9640022_374_SD3_1	Yes	C	T	chr8	119176907	+	yes	yes	76	ASW=1 LWK=1
EN_9687394_205_SD3_1	No	T	G	chr18	23962666	+	no	yes	80.25	CEU=1 CHB=1 JPT=1 LWK=1 YRI=1
EN_9687517_520_SD3_1	Yes	T	A	chr12	38845409	+	no	yes	85	CEU=1 TSI=1
EN_9857396_422_SD3_1	Yes	C	T	chr21	39467775	+	yes	yes	85	ASW=1 YRI=1
EN_9857426_306_SD3_1	Yes	T	C	chr21	39484634	+	no	yes	85	ASW=1 YRI=1
EN_9857522_264_SD3_1	Yes	G	A	chr21	39535660	+	yes	yes	80	LWK=1 YRI=2
EN_9857584_570_SD3_1	Yes	T	C	chr5	56088514	+	no	yes	81.57	LWK=3 YRI=2
EN_9857586_122_SD3_1	Yes	C	A	chr5	56088953	+	no	yes	85	CHB=1 CHD=1
EN_9857586_484_SD3_1	Yes	A	G	chr5	56089315	+	no	yes	85	CHB=1 CHD=1
EN_9857633_498_SD3_1	Yes	G	A	chr5	56113913	+	yes	yes	85	LWK=1 YRI=1
EN_9857653_234_SD3_1	No	G	C	chr5	56124133	+	no	yes	80	CEU=1 GIH=1
EN_9857671_624_SD3_1	Yes	T	A	chr5	56133483	+	no	yes	80	LWK=2 MXL=1 YRI=3
EN_9858711_399_SD3_1	Yes	A	G	chr21	39540483	+	no	yes	80	ASW=1 YRI=1

8. HAPLOTYPE SHARING

Haplotype sharing: To study the extent of haplotype sharing around intermediate frequency ENCODE SNPs, we selected a subset of SNPs identified in the sequencing whose minor allele was seen between two and six times in either the YRI or CEU samples. We obtained genotypes for these SNPs by Sequenom genotyping in the full set of trios for the relevant population (CEU or YRI or both). We filtered on call rate and restricted ourselves to samples that passed the overall genotyping QC (see above) and to cases where unambiguous phase could be assigned by the trio information, which yielded 106 SNPs with 2 – 6 copies of the minor allele in the CEU sample and 272 in the YRI sample. These SNPs were inserted into the phased HapMap 3 genotype data (176 phased chromosomes for CEU, 200 for YRI, see Large Scale Genotyping for information about phasing). Starting at the ENCODE SNP, we successively added array SNPs in one direction and calculated the probability that a pair of chromosomes sharing the minor SNP allele were identical at all SNPs. We repeated the calculation independently in the other direction.

For comparison, we also selected at random ~500 array SNPs in each of four frequency bins (1%, 5%, 20% and 50%) for the same two populations, and carried out the identical analysis.

9. IMPUTATION OF UNTYPED VARIANTS

Imputation: Imputation was performed using the MACH program [18], with the "mle" and "greedy" options selected, and the number of rounds set to ten. The statistic of merit was the squared correlation between the true genotype and the (continuous-valued) imputed genotype dosage, averaged across all SNPs; this is indicative of the fraction of power retain when using imputation instead of direct genotyping in a disease association study. In all analyses, the set of samples whose genotypes were imputed did not overlap the set of samples used to construct reference panels.

For the 1958 British Birth Cohort analysis, we assessed improved imputation using the HapMap 3 (release 2) panel totalling 410 phased European-ancestry chromosomes (CEU+TSI) compared to using a smaller HapMap Phase II panel of 120 CEU chromosomes (HM2-CEU). The 58BBC samples had been previously genotyped on the Affymetrix 500K and Illumina 550K chips, so we used the 58BBC Illumina 550K genotypes in tandem with either reference panel (HM2-CEU or CEU+TSI) to impute the known (but masked) SNPs assayed on the Affymetrix 500K chip. Using the Illumina array genotypes, we imputed HapMap 3 SNPs on chromosome 20 and calculated the mean r^2 between true (called) genotype and imputed genotype dosage for each Affymetrix SNP not on the Illumina chip (**Table S8**) We present representative results based on imputing all available SNPs on chromosome 20 (**Table S9**; **Figs S9a,b**). Although chromosome 20 is only ~2% of the genome, we found that imputation on chromosome 1 (~8% of genome) gave similar results, also showing that imputation improved mainly due to SNPs with unobserved minor alleles in the HM2-CEU reference panel that became informative in the larger CEU+TSI panel (see main text).

The remaining imputation analyses were restricted to 988 unrelated individuals from 11 populations for which genotype data from 1,440,616 SNPs were available as part of HapMap 3 release 2. Genotypes were first phased using the Phase program to produce phased reference panels, as described above.

For cross-population comparisons, we used Affymetrix 6.0 genotypes to impute non-overlapping Illumina 1M genotypes. For imputation in admixed populations, we constructed reference panels of 200 chromosomes from either: one Phase II HapMap population, two Phase II HapMap populations, three Phase II HapMap populations (60 CEU + 60 CHB+JPT + 80 YRI) (COSMO1), or six HapMap 3 populations (30 CEU + 30 CHB+JPT + 30 MXL + 30 GIH + 40 YRI + 40 MKK) (COSMO2). We also constructed a reference panel of 100 chromosomes from the same population for each admixed population with data from more than 50 samples available: GIH, MKK and LWK. (Additional statistics, plus results for using Illumina 1M genotypes to impute Affymetrix 6.0 genotypes, are reported in **Tables S9**, **S10** and displayed in **Figure S7**). Coverage was consistently higher when using Illumina 1M genotypes to impute Affymetrix 6.0 genotypes than *vice versa*: 96.5% vs. 91.2% for CEU, 95.4% vs. 90.4% for CHB+JPT and 91.6% vs. 87.5% for YRI, for r^2 between imputed genotype dosage and true genotype averaged across common SNPs (MAF ≥ 5%).

For closely related populations, we compared imputation of CEU or TSI using the CEU reference panel, CHD or CHB+JPT using the CHB+JPT reference panel, and YRI or LWK using the YRI reference panel. Imputation into closely related populations worked well for common but not for low-frequency alleles (**Table S11**).

In the final set of analyses, imputation was carried out for a single SNP (or CNP) at a time, using the consensus (Affymetrix 6.0 + Illumina 1M) genotypes, and all available samples from the reference population. Target samples were imputed one at a time; when the reference and target populations were the same, the target individual was removed from the reference panel for that imputation only. To reduce the computational load, MACH was run in a two-stage process: 1) the entire reference panel was used by MACH to generate cross-over and error maps (with only the target SNP removed from the data); 2) those maps were used for imputing each target sample in turn.

The probability that pairs of SNPs were perfect proxies for each other was estimated by counting how often the minor allele occurred in the same individuals in the sample set. All frequency-matched pairs of SNPs, separated by < 20 kb, in the two sets of data (consensus array data and ENCODE sequence data). (This provides a ~0.5% overestimate of the rate of true proxies, since the minor allele could be on either chromosome in the individual.) Results are show in **Figure S9**.

Table S8: 1958 British Birth Cohort imputation results.

Typed SNPs	Imputed SNPs	MAF (copies) in CEU+TSI panel	SNP N	r^2 HM2-CEU ±SEM	r^2 CEU+TSI ±SEM	% “improved” SNPs (r^2 increase > 0.1)	Mean r^2 increase “improved” SNPs
Illumina 550K	Affymetrix 500K	~0.25% (1)	180	0.091 ±0.018	0.310 ±0.028	36%	0.61
Illumina 550K	Affymetrix 500K	~0.5% (2)	84	0.221 ±0.040	0.545 ±0.042	51%	0.64
Illumina 550K	Affymetrix 500K	~0.75% (3)	68	0.328 ±0.053	0.693 ±0.044	50%	0.73
Illumina 550K	Affymetrix 500K	~1.0% (4)	72	0.512 ±0.051	0.831 ±0.027	47%	0.68
Illumina 550K	Affymetrix 500K	~1.25 – 2.5% (5-10)	303	0.714 ±0.020	0.858 ±0.011	26%	0.52
Illumina 550K	Affymetrix 500K	~2.5 – 5.0% (11-20)	491	0.841 ±0.011	0.898 ±0.008	17%	0.29
Illumina 550K	Affymetrix 500K	All rare (<0.5%)	264	0.132 ±0.018	0.385 ±0.024	41%	0.62
Illumina 550K	Affymetrix 500K	All low-frequency (0.5%-5%)	934	0.737 ±0.011	0.865 ±0.007	25%	0.49
Illumina 550K	Affymetrix 500K	All common (>5.0%)	6185	0.946 ±0.0014	0.961 ±0.0011	3%	0.17

Table S9. Additional imputation statistics. We report results for (a) imputing CEU using CEU, (b) imputing CHB+JPT using CHB+JPT, and (c) imputing YRI using YRI. In each case, imputation was performed using a reference panel of 100 chromosomes (only nonoverlapping samples were imputed). The first nine rows of each table are based on using Affymetrix 6.0 to impute Illumina 1M, and the last nine rows are based on using Illumina 1M to impute Affymetrix 6.0. Concordance denotes average concordance between imputed and true genotypes, r^2 denotes average squared correlation between imputed and true genotypes, r^2 (dosage) denotes average squared correlation between imputed genotype dosages and true genotypes, freqdiff denotes average absolute frequency difference between imputed and true genotypes, and freqdiff (normalized) denotes average absolute frequency difference normalized by true MAF. Values of freqdiff (normalized) for bins that include 0-1% MAF were set to n/a, as the small denominator leads to large values of the statistic for this bin.

(a) CEU using CEU

Typed SNPs	Imputed SNPs	MAF bin	Concordance	r^2	r^2 (dosage)	freqdiff	freqdiff (normalized)
Affymetrix	Illumina	0-1%	99.9%	31.3%	33.0%	0.001	n/a
Affymetrix	Illumina	1-2%	98.4%	58.2%	60.3%	0.007	0.417
Affymetrix	Illumina	2-5%	98.3%	78.7%	80.7%	0.006	0.189
Affymetrix	Illumina	5-10%	97.7%	85.6%	87.5%	0.008	0.101
Affymetrix	Illumina	10-20%	97.0%	88.9%	90.4%	0.009	0.063
Affymetrix	Illumina	20-50%	95.9%	91.2%	92.3%	0.010	0.031
Affymetrix	Illumina	all rare	99.6%	63.6%	65.5%	0.002	n/a
Affymetrix	Illumina	all common	96.4%	89.9%	91.2%	0.010	0.048
Affymetrix	Illumina	ALL	97.4%	87.0%	88.4%	0.007	n/a
Illumina	Affymetrix	0-1%	99.9%	35.5%	37.3%	0.001	n/a
Illumina	Affymetrix	1-2%	98.7%	65.6%	67.4%	0.006	0.355
Illumina	Affymetrix	2-5%	98.7%	82.9%	84.5%	0.005	0.162
Illumina	Affymetrix	5-10%	99.0%	93.1%	94.0%	0.004	0.056
Illumina	Affymetrix	10-20%	98.9%	95.8%	96.3%	0.004	0.029
Illumina	Affymetrix	20-50%	98.5%	96.7%	97.1%	0.005	0.015
Illumina	Affymetrix	all rare	99.7%	67.6%	69.2%	0.001	n/a
Illumina	Affymetrix	all common	98.6%	96.0%	96.5%	0.005	0.024
Illumina	Affymetrix	ALL	99.1%	91.7%	92.4%	0.003	n/a

(b) CHB+JPT using CHB+JPT

Typed SNPs	Imputed SNPs	MAF bin	Concordance	r^2	r^2 (dosage)	freqdiff	freqdiff (normalized)
Affymetrix	Illumina	0-1%	99.9%	39.9%	41.2%	0.000	n/a
Affymetrix	Illumina	1-2%	98.6%	71.2%	72.8%	0.006	0.425
Affymetrix	Illumina	2-5%	98.5%	87.1%	88.6%	0.006	0.190
Affymetrix	Illumina	5-10%	97.8%	91.7%	93.2%	0.007	0.096
Affymetrix	Illumina	10-20%	96.7%	93.3%	94.6%	0.009	0.059
Affymetrix	Illumina	20-50%	95.3%	94.6%	95.7%	0.010	0.030
Affymetrix	Illumina	all rare	99.7%	82.2%	83.7%	0.001	n/a
Affymetrix	Illumina	all common	96.0%	94.9%	95.9%	0.009	0.045
Affymetrix	Illumina	ALL	97.5%	95.8%	96.6%	0.006	n/a
Illumina	Affymetrix	0-1%	99.9%	39.3%	40.5%	0.000	n/a
Illumina	Affymetrix	1-2%	98.8%	76.5%	77.5%	0.005	0.381
Illumina	Affymetrix	2-5%	98.9%	90.7%	91.6%	0.005	0.153
Illumina	Affymetrix	5-10%	98.7%	95.1%	95.9%	0.004	0.061
Illumina	Affymetrix	10-20%	98.4%	96.8%	97.4%	0.005	0.033
Illumina	Affymetrix	20-50%	98.1%	97.9%	98.2%	0.005	0.016
Illumina	Affymetrix	all rare	99.8%	85.3%	86.3%	0.001	n/a
Illumina	Affymetrix	all common	98.3%	97.8%	98.2%	0.005	0.026
Illumina	Affymetrix	ALL	99.0%	98.2%	98.5%	0.003	n/a

(c) YRI using YRI

Typed SNPs	Imputed SNPs	MAF bin	Concordance	r^2	r^2 (dosage)	freqdiff	freqdiff (normalized)
Affymetrix	Illumina	0-1%	99.8%	46.7%	48.8%	0.001	n/a
Affymetrix	Illumina	1-2%	98.3%	69.3%	71.5%	0.007	0.463
Affymetrix	Illumina	2-5%	97.8%	82.8%	85.0%	0.009	0.254
Affymetrix	Illumina	5-10%	97.1%	89.0%	90.8%	0.010	0.134
Affymetrix	Illumina	10-20%	95.8%	91.3%	92.9%	0.012	0.084
Affymetrix	Illumina	20-50%	93.8%	92.9%	94.3%	0.015	0.045
Affymetrix	Illumina	all rare	99.2%	80.0%	82.3%	0.003	n/a
Affymetrix	Illumina	all common	94.8%	93.3%	94.6%	0.013	0.068
Affymetrix	Illumina	ALL	95.9%	93.9%	95.1%	0.011	n/a
Illumina	Affymetrix	0-1%	99.9%	55.6%	57.8%	0.001	n/a
Illumina	Affymetrix	1-2%	98.6%	75.1%	77.4%	0.006	0.393
Illumina	Affymetrix	2-5%	98.2%	86.3%	88.2%	0.007	0.213
Illumina	Affymetrix	5-10%	97.7%	91.3%	92.8%	0.008	0.112
Illumina	Affymetrix	10-20%	97.1%	93.9%	95.1%	0.009	0.065
Illumina	Affymetrix	20-50%	96.4%	95.9%	96.6%	0.010	0.032
Illumina	Affymetrix	all rare	99.4%	84.6%	86.4%	0.002	n/a
Illumina	Affymetrix	all common	96.8%	95.8%	96.6%	0.010	0.054
Illumina	Affymetrix	ALL	97.5%	96.2%	96.9%	0.008	n/a

Table S10. Strategies for imputation in admixed populations: (a) ASW, (b) MXL (c) GIH, (d) MKK and (e) LWK. Reference panels contained 200 chromosomes for most runs, but only 100 chromosomes (as indicated) for imputing GIH, MKK and LWK using the same population. Runs imputing ASW and MXL using the same population were not performed due to the lower number of samples available for those populations. In runs labeled with a *, a subset of samples included in the COSMO2 panel were excluded from the imputed samples.

(a)

Imputed population	Reference panel	r^2 for rare SNPs	r^2 for common SNPs
ASW	YRI	45.5%	83.0%
ASW	YRI+CEU	71.7%	86.5%
ASW	COSMO1	70.1%	85.4%
ASW	COSMO2	67.2%	83.9%

(b)

Imputed population	Reference panel	r^2 for rare SNPs	r^2 for common SNPs
MXL	CEU	42.8%	85.3%
MXL	CEU+(CHB+JPT)	45.9%	87.4%
MXL	COSMO1	74.8%	88.9%
MXL*	COSMO2	78.1%	90.6%

(c)

Imputed population	Reference panel	r^2 for rare SNPs	r^2 for common SNPs
GIH	CEU	62.5%	85.7%
GIH	CEU+(CHB+JPT)	69.4%	87.4%
GIH	COSMO1	72.6%	86.9%
GIH*	COSMO2	77.9%	89.4%
GIH	GIH (100)	75.1%	91.7%

(d)

Imputed population	Reference panel	r^2 for rare SNPs	r^2 for common SNPs
MKK	YRI	44.4%	76.4%
MKK	YRI+LWK	51.5%	80.6%
MKK	COSMO1	57.7%	78.6%
MKK*	COSMO2	64.8%	77.5%
MKK	MKK (100)	64.8%	87.3%

(e)

Imputed population	Reference panel	r^2 for rare SNPs	r^2 for common SNPs
LWK	YRI	47.3%	82.9%
LWK	YRI+MKK	61.8%	85.5%
LWK	COSMO1	53.6%	80.3%
LWK	COSMO2	53.8%	80.7%
LWK	LWK (100)	61.6%	86.6%

Table S11. **Effect of reference panel choice on imputation accuracy in closely related populations.** We report values of r^2 between imputed dosage and true genotype, based on a reference panel size of 100 chromosomes (only non-overlapping samples were imputed).

Imputed population	Reference panel	r^2 for low frequency SNPs	r^2 for common SNPs
CEU	CEU	65.5%	91.2%
TSI	CEU	56.0%	89.5%
CHB+JPT	CHB+JPT	56.1%	90.4%
CHD	CHB+JPT	56.6%	89.4%
YRI	YRI	65.2%	87.5%
LWK	YRI	40.4%	80.0%

10. NATURAL SELECTION

Natural selection: To examine evidence for recent positive selection, we implemented a previously published method that combines multiple tests for selection, the Composite of Multiple Signals (CMS)[29]. CMS combines multiple signals of section - long-range associations, population differentiation, and high-frequency derived alleles - to localize signals in the genome, increasing resolution by up to 100-fold over individual signals, and can do so even with incomplete genotype data. Because it integrates multiple independent tests, CMS has a very low false discovery rate.

As prior distributions for the input statistics to CMS, we used previously published empirical distributions generated from simulated regions under positive selection. The simulation parameters were taken from a previously validated demographic model[22], which included samples from three populations (African, Asian, and European). We did not explicitly simulate HapMap 3 populations that were not in HapMap II, because we did not have a detailed validated demographic model that included these populations. Instead, we assumed that TSI was sufficiently similar to CEU to allow distributions of scores determined from modeling the CEU to apply to TSI, and similarly used YRI to model LWK and MKK.

To determine confidence regions, we used the windowed approach from Grossman *et al*, and adjusted the number of high-scoring SNPs per significant window to reflect the genotyping density. We divided the region into 0.02 cM regions, each overlapping the next one by 0.01 cM, and included all windows that contain at least 1 SNP with a normalized CMS score above 0.5.

In the CEU, CHB+JPT, and YRI, we analyzed previously published regions that were identified as targets of recent positive selection in HapMap II. To evaluate the replication rate of the CMS localization in HapMap 3, we recomputed CMS scores using HapMap 3 data across the published regions identified as targets of selection in HapMap2. We defined a region to be replicated if the 95% confidence intervals for the position of the selected variants overlapped in the two datasets. In the TSI, MKK, and LWK populations, we identified regions potentially under positive selection using three previously published tests for selection, the Long-Range Haplotype (LRH)[19], the integrated Haplotype Score (iHS), and the cross-population EHH (XP-EHH) tests[20, 21], and then ran CMS to localize the signal within these regions.

To determine the significance thresholds for the selection tests, we used the *cosi* coalescent simulator to simulate 1,000 1MB autosomal regions, evolving neutrally under a previously validated demographic model[22]. We set thresholds that yielded no false positives in simulations (<0.001 FPR). The model included samples for three populations (African, Asian and European), with sample sizes matching HapMap 3 data (167, 171 and 165 samples respectively). Recombination was modeled as varying along the region, with a hierarchical model that included both regional variation in recombination rates (estimated from deCODE data) and local hotspots of recombination[2].

Thinned simulations modeling SNP ascertainment were created from full-sequence simulations by randomly removing SNPs, with the probability of removal based on minor-allele frequency. The per-frequency removal probabilities were chosen to match half of HapMap II densities of SNPs with each minor-allele frequency. Again, we did not explicitly simulate HapMap 3 populations that were not in HapMap II, and instead assumed that TSI was sufficiently similar to CEU to allow significance thresholds determined on CEU to apply to TSI, and similarly YRI to LWK and MKK.

The significance thresholds, determined from the 1000 neutral simulations, were calibrated to have a <0.1% false positive rate as follows. A 100K window was declared significant by the LRH test if over 0.1 of its SNPs had LRH significance scores of over 4.8; by the iHS test, if 0.3 of its SNPs had iHS significance scores of over 3.4; by the XP-EHH test, if at least one of its SNPs had an XP-EHH score of over 5.1 in two population comparisons. Significant windows separated by less than 50K were merged into a single significant region.

In the CEU, CHB+JPT and YRI, regions that did replicate include a number of well-known pigmentation genes, *SLC24A5*, *KITLG*, OCA2, *TYRP1* and *MATP* [30,31] (**Figure S10a-d**), and regions with the genes *LCT*, *EDAR*, *HERC1*, and *PKFP*[32].

Table S12: Signals of natural selection localized by CMS identified in the TSI, MKK, and LWK.

Chr	Start	End	Peak SNP	Size	Pop	Genes in Region
1	160308937	160388443	160380511	79506	TSI	NOS1AP
1	228203609	228274056	228274056	70447	TSI	GALNT2
1	246214811	246268967	246253723	54156	TSI	OR2L13,OR2L1P,OR2L2
1	247082090	247124862	247124862	42772	LWK	SH3BP5L,ZNF672,ZNF692
2	88691341	88881695	88700868	190354	MKK	EIF2AK3,RPIA,FLJ40330
2	104122546	104183834	104183834	61288	TSI	
2	134217677	134245826	134230091	28149	MKK	
2	135478814	136008895	135886269	530081	MKK	YSK4,RAB3GAP1,ZRANB3,R3HDM1
2	176383762	176388492	176388492	4730	LWK	
2	197618211	197713604	197687944	95393	LWK	ANKRD44
2	238001541	238066523	238001541	64982	TSI	MLPH
3	25706331	25798544	25794632	92213	TSI	NGLY1
3	47256641	48142694	47256641	886053	LWK	KIF9,KLHL18,PTPN23,SCAP,C3orf75,CSPG5,SMARCC1,DHX30,MIR1226,MAP4
3	193422610	193473269	193422610	50659	MKK	FGF12
4	41807491	41815266	41807491	7775	TSI	BEND4
5	14800247	14803481	14803251	3234	MKK	ANKH
5	109904024	110080398	109926082	176374	TSI	TMEM232
5	115913181	115913757	115913571	576	MKK	SEMA6A
5	142278349	142304596	142278537	26247	TSI	ARHGAP26
5	158512344	158600287	158512814	87943	LWK	RNF145
6	63416530	63643099	63428610	226569	LWK	
7	33672016	33735989	33672016	63973	TSI	
8	139575414	139613806	139613806	38392	TSI	FAM135B
9	31424342	31513416	31513416	89074	MKK	
9	38729591	38736937	38729591	7346	MKK	
9	90371934	90379528	90376776	7594	TSI	NXNL2
9	128866305	128972214	128950091	105909	LWK	RALGPS1,ANGPTL2
9	139097172	139100238	139100238	3066	MKK	UAP1L1,LOC100289341,MAN1B1,DPP7
9	139107171	139126312	139107171	19141	TSI	UAP1L1,LOC100289341,MAN1B1,DPP7
10	3017807	3039561	3023127	21754	TSI	
10	135219522	135227438	135227438	7916	MKK	SYCE1
12	87485844	87586557	87586557	100713	TSI	KITLG
12	110394602	110556807	110492139	162205	TSI	ATXN2
14	59865374	59880957	59865374	15583	LWK	
14	61077818	61107749	61107749	29931	TSI	PRKCH
14	62892274	62940684	62940684	48410	TSI	PPP2R5E
15	42928665	43054398	42939663	125733	TSI	C15orf43
15	57434283	57448696	57434283	14413	TSI	MYO1E
16	1590947	1757432	1656011	166485	TSI	IFT140,CRAMP1L,HN1L,MAPK8IP3
16	31602764	31680296	31629786	77532	LWK	C16orf67,ZNF720
16	31672282	31710886	31677125	38604	MKK	C16orf67,ZNF720
16	34219719	34464860	34373576	245141	MKK	UBE2MP1,LOC283914
16	64658900	64672826	64672826	13926	TSI	
17	3555271	3565280	3555271	10009	LWK	ITGAE
17	10961600	10967784	10967784	6184	TSI	
17	33785091	33818975	33789582	33884	TSI	SOCS7
17	41118038	41173230	41157478	55192	TSI	
17	41118038	41173230	41157478	55192	TSI	
18	7574294	7599137	7588656	24843	TSI	PTPRM
18	19749481	19828457	19756062	78976	TSI	LAMA3,TTC39C
18	64846196	64877649	64855865	31453	MKK	CCDC102B
18	65748313	65779636	65749159	31323	LWK	CD226,RTTN
18	65764906	65880313	65775153	115407	MKK	CD226,RTTN
20	62302138	62306628	62302138	4490	LWK	MYT1

References:

1. The Internatinal HapMap Consortium. A haplotype map of the human genome. *Nature* 437, 1299-320 (2005).
2. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-61 (2007).
3. Korn, J. M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 40, 1253-60 (2008).
4. Teo, Y. Y. et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* 23, 2741-6 (2007).
5. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559-75 (2007).
6. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5, e1000529 (2009).
7. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68, 978-89 (2001).
8. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6, 95-108 (2005).
9. Cutler, D. J. et al. High-throughput variation detection and genotyping using microarrays. *Genome Res* 11, 1913-25 (2001).
10. Weiss, L. A., Arking, D. E., Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 461, 802-8 (2009).
11. Neale, B. M. et al. Genome-wide association scan of attention deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet* 147B, 1337-44 (2008).
12. Zhang, J. et al. SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput Biol* 1, e53 (2005).
13. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35, 2013-25 (2007).
14. Peiffer, D. A. et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16, 1136-48 (2006).
15. Barnes, C. et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 40, 1245-52 (2008).
16. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* 2, e190 (2006).
17. Smith, M. W. et al. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet* 74, 1001-13 (2004).
18. Price, A. L. et al. Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet* 4, e1000294 (2008).
19. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The history and geography of human genes* (Princeton University Press, Princeton, N.J.,, 1994).

20. Ayodo, G. et al. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet* 81, 234-42 (2007).

21. Price, A. L. et al. A genomewide admixture map for Latino populations. *Am J Hum Genet* 80, 1024-36 (2007).

22. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* 461, 489-94 (2009).

23. Price, A. L. et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5, e1000519 (2009).

24. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-4 (2008).

25. Cann, H. M. et al. A human genome diversity cell line panel. *Science* 296, 261-2 (2002).

26. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* 6, 333-40 (2005).

27. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998-1003 (2008).

28. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78, 629-44 (2006).

29. Grossman, S. R. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327, 883-6.

30. Sabeti, P. C. et al. Positive natural selection in the human lineage. *Science* 312, 1614-20 (2006).

31. Lamason, R. L. et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310, 1782-6 (2005).

32. Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* 19, 711-22 (2009).

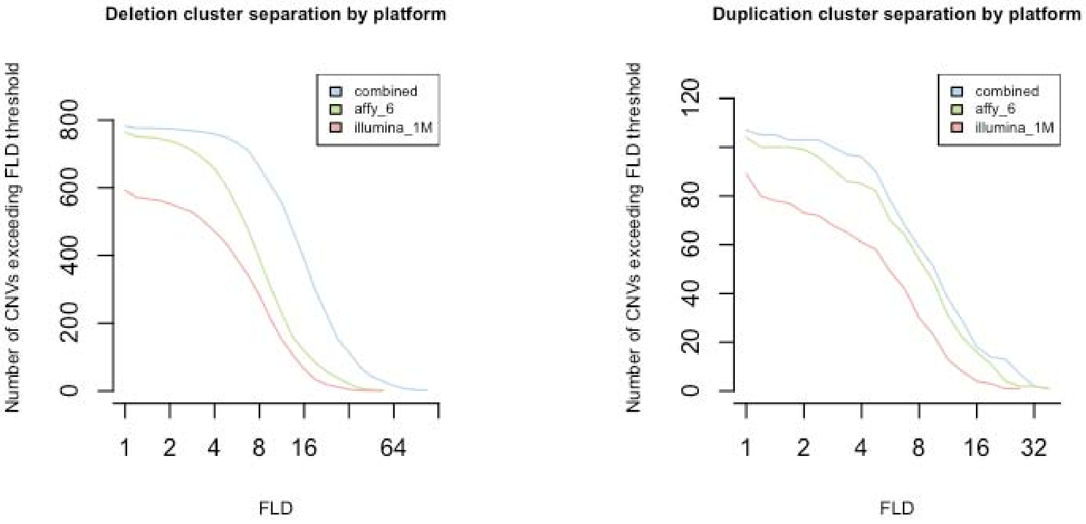
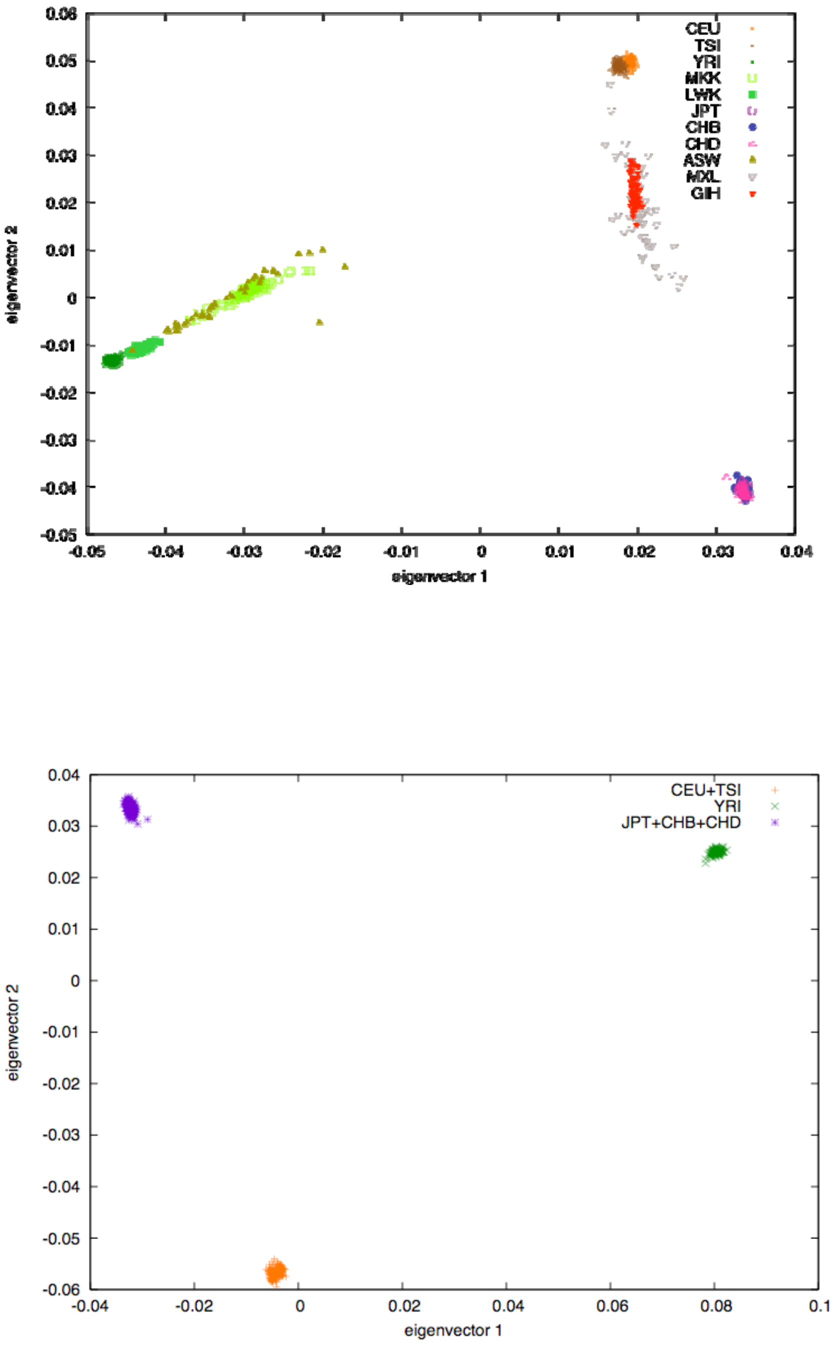


Figure S1. Resolution of copy-number genotype classes, measured using Fisher's linear discriminant (FLD). Joint utilization of the data from the two array platforms (Illumina and Affymetrix) together yielded genotype clusters that were more clearly resolved than when data from either platform was used on its own. Deletion polymorphisms showed more-effective separation of genotype classes than duplication polymorphisms did.

Figure S2. Principal components analysis.



a.

b.

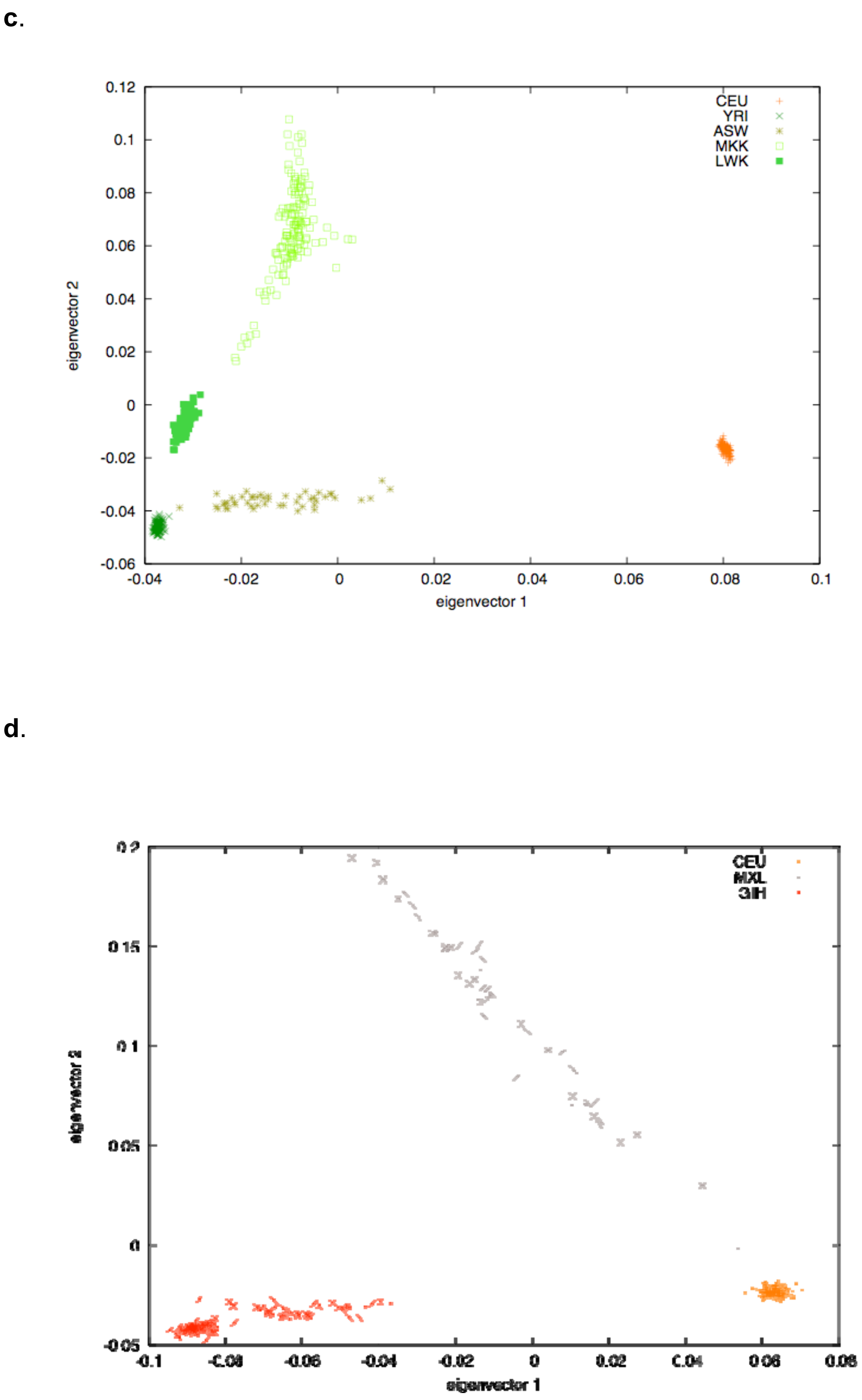
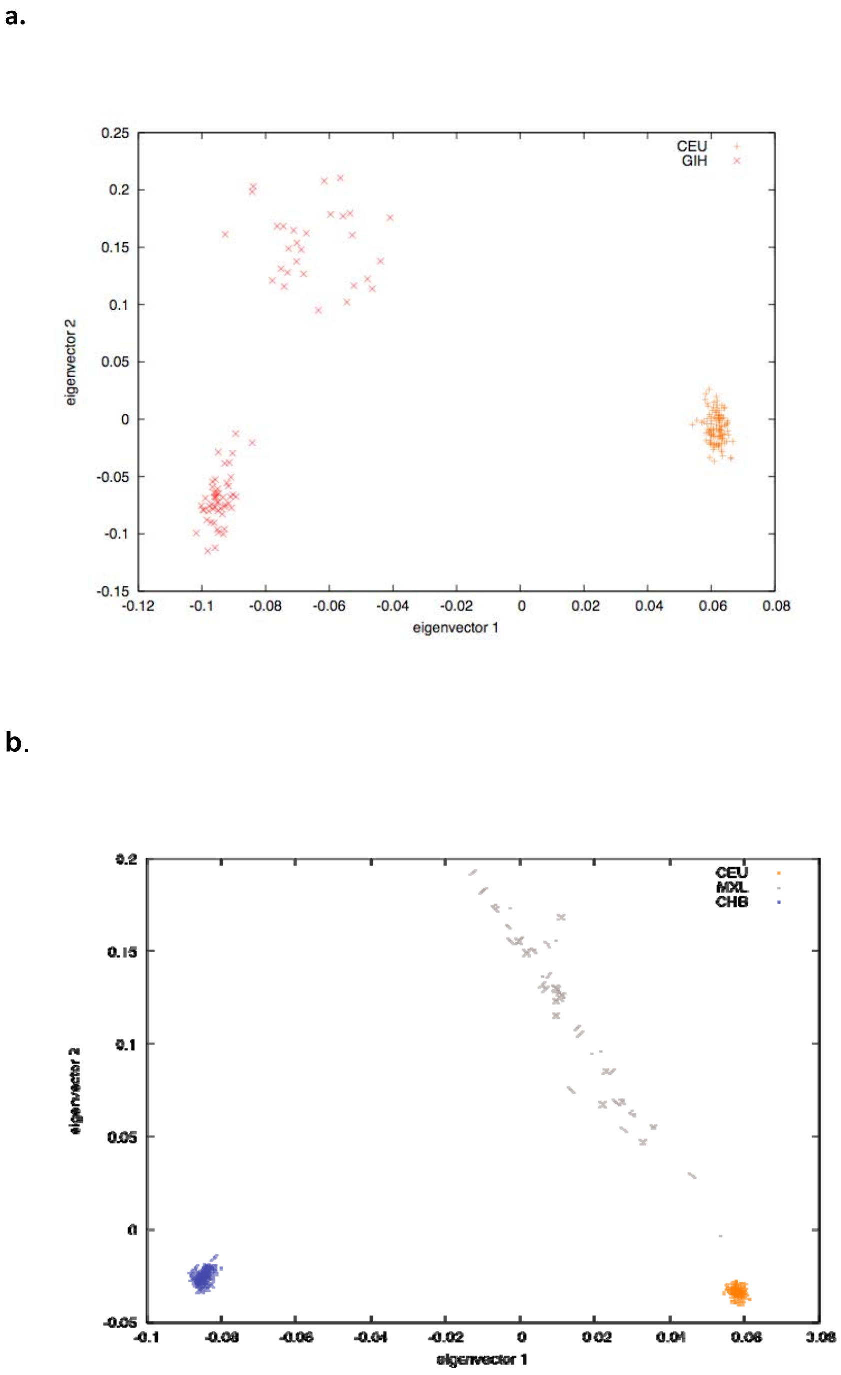


Figure S2. Principal components analysis. We plot the top two PCs for **a**. all 11 populations, **b**. 6 unadmixed populations, **c**. 3 admixed populations with genetic proximity to Africa, together with CEU and YRI, and **d**. 2 admixed populations with genetic proximity to Europe, together with CEU.

Figure S3. Principal components analysis of populations with a European admixture component.



a.

b.

c.

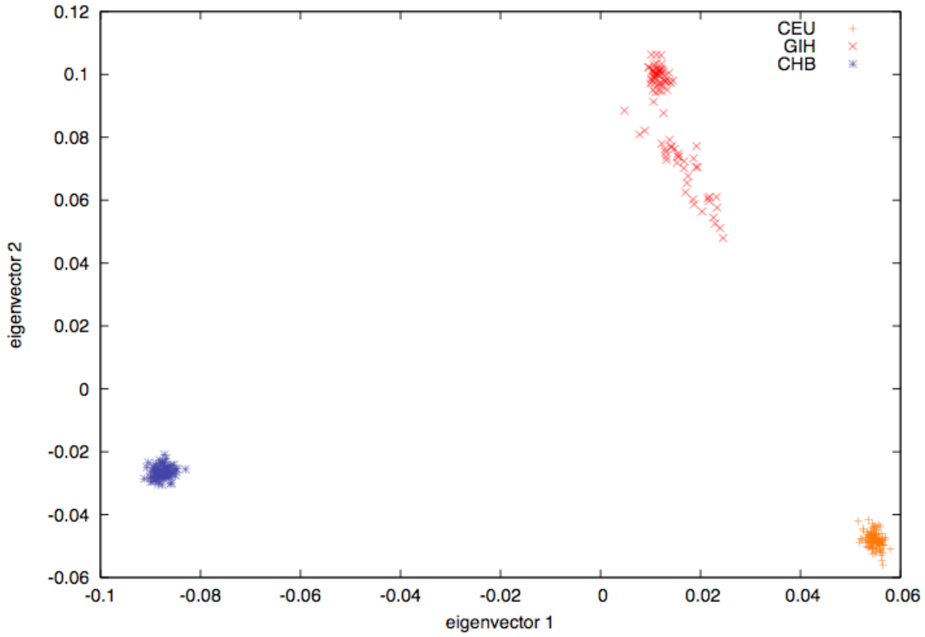
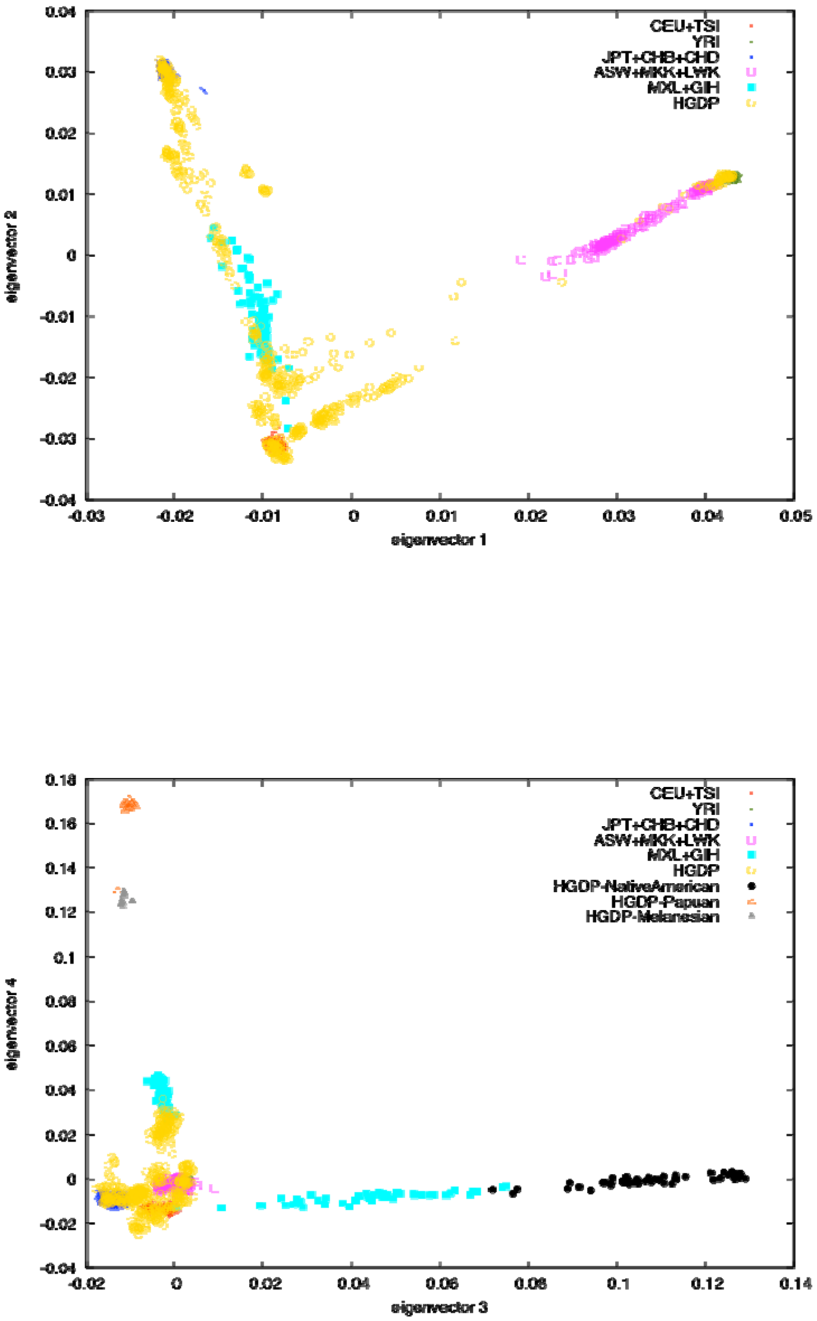


Figure S3. Principal components analysis of populations with a European admixture component. We plot the top two PCs for **a**. CEU and GIH, **b**. CEU, MXL and CHB, and **c**. CEU, GIH, and CHB.

Figure S4. Principal components analysis of HapMap3 and HGDP samples.



a.

b.

c.

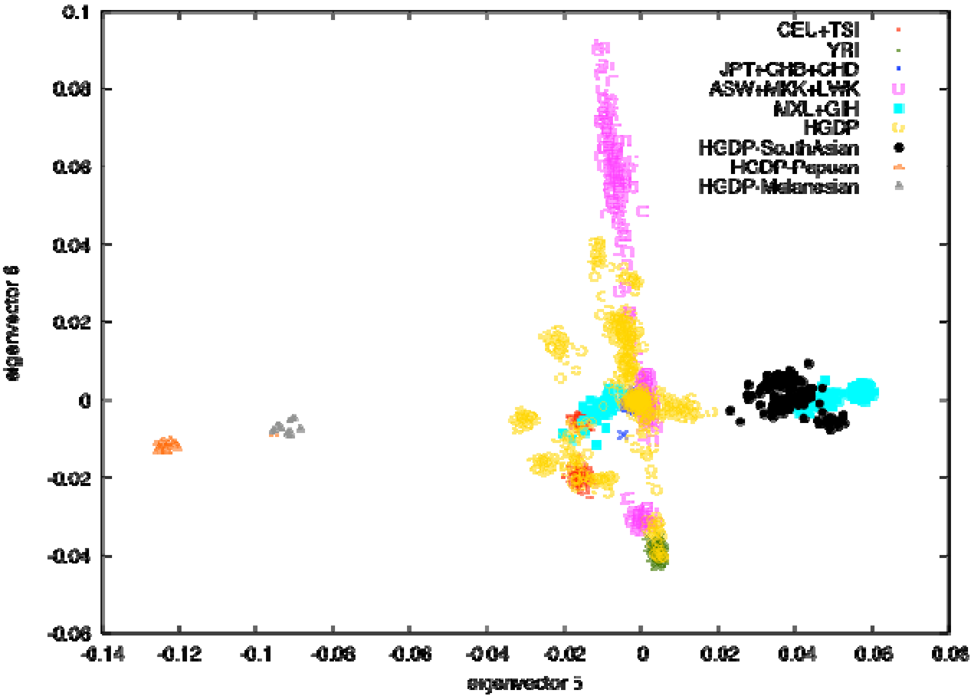


Figure S4. Principal components analysis of HapMap 3 and HGDP samples. We plot **a**. PC1 and PC2, **b**. PC3 and PC4, and **c**. PC5 and PC6.

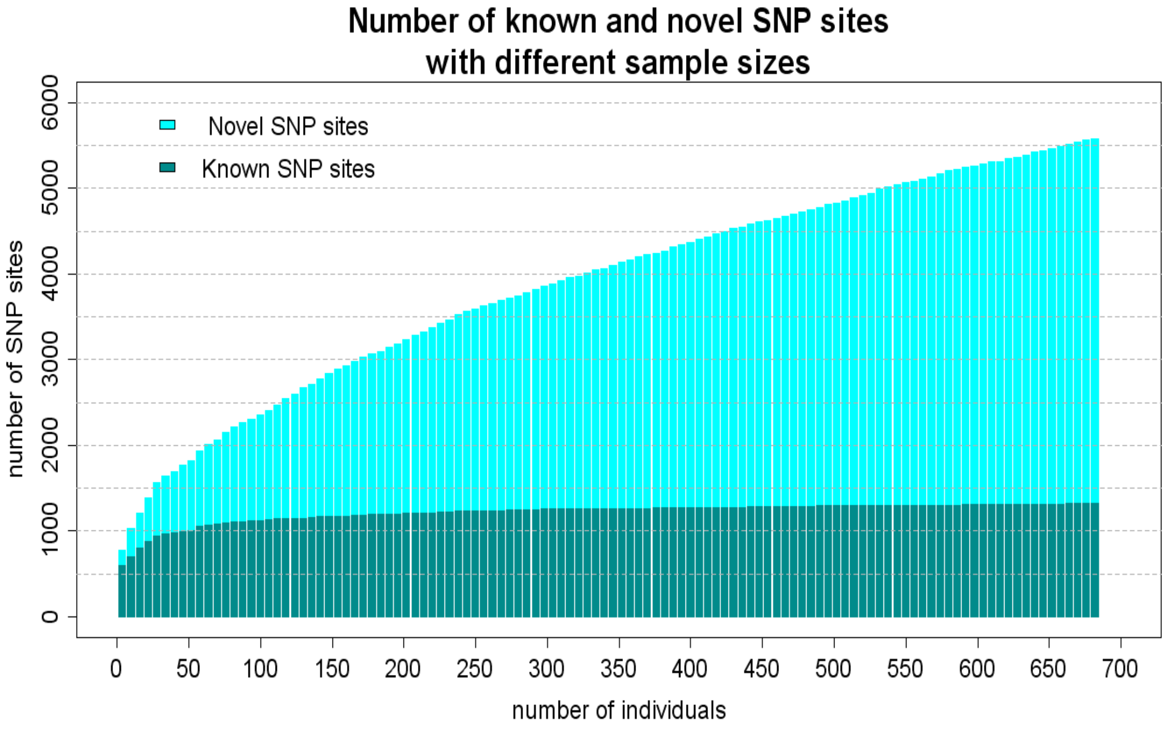


Figure S5. Number of discovered known and novel SNPs in the ENCODE resequencing data set as a function of the number of samples. We randomly sampled individuals from the ENCODE resequencing data set. We plotted the numbers of known (i.e. present in dbSNP129) and novel SNPs discovered by resequencing as a function of the number of interrogated samples.

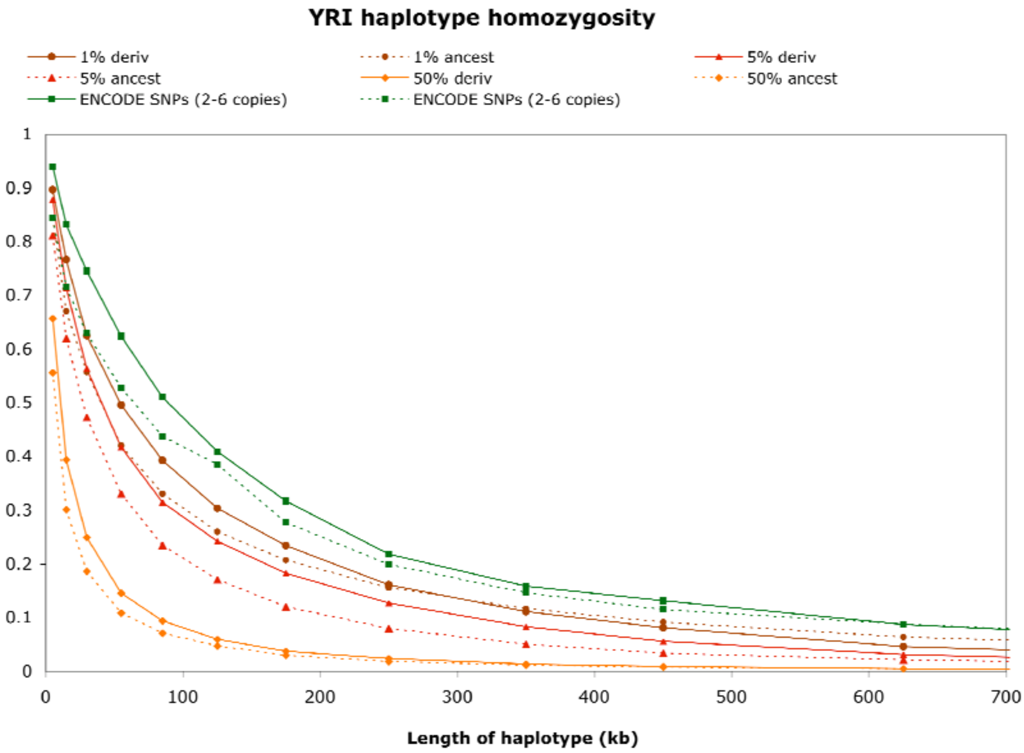


Figure S6. Effect of ancestral status on haplotype sharing. Shown is the haplotype homozygosity in YRI for different minor allele frequencies, broken down by whether the minor allele is ancestral or derived (as estimated from the chimpanzee allele). ENCODE SNPs have 2 - 6 copies of the minor allele (or MAF of 1-3%).

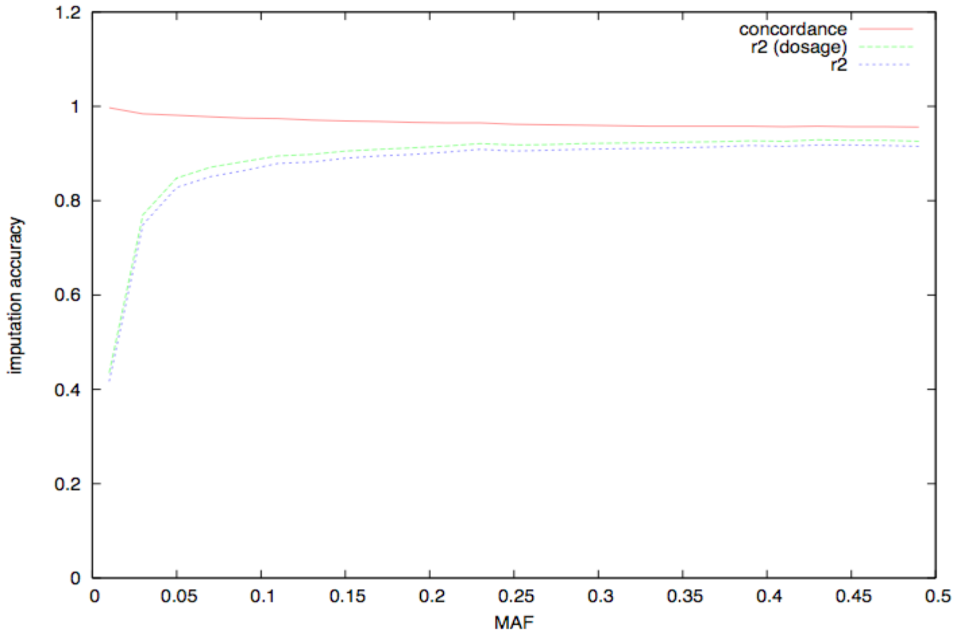


Figure S7. Imputation accuracy as a function of minor allele frequency (MAF). We report concordance, r^2 (dosage), and r^2, for each of **a**. CEU, **b**. CHB+JPT, and **c**. YRI. Results are binned in MAF bins of size 0.02, using the midpoint of each bin on the x-axis.

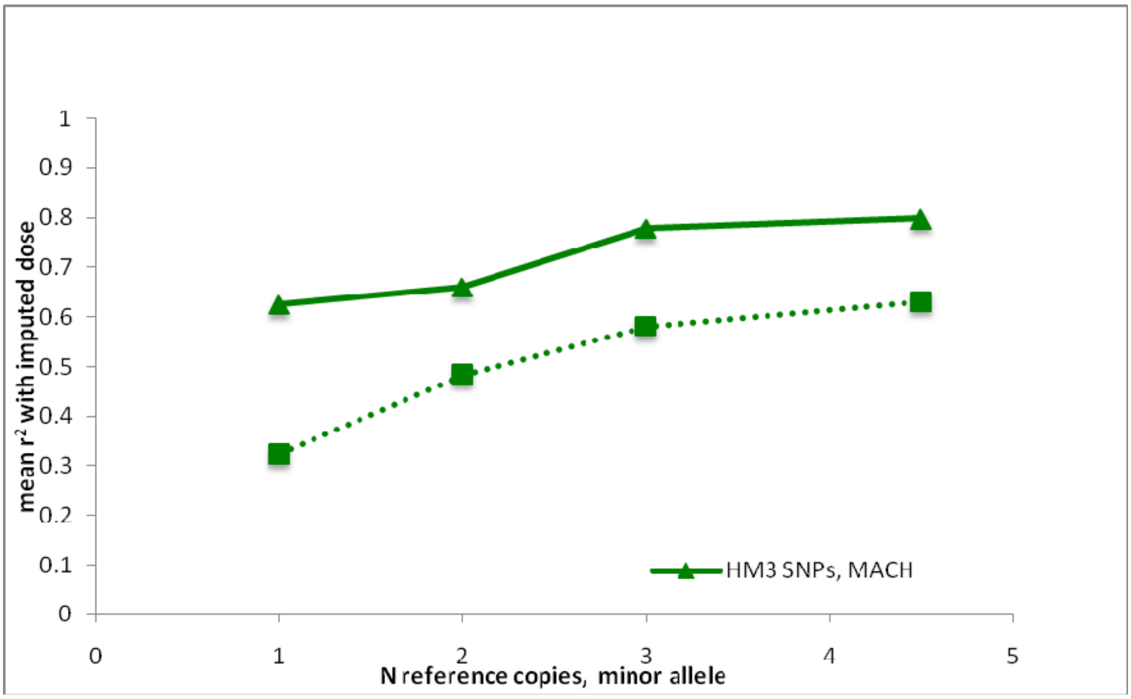


Figure S8. Effect of SNP density on low-frequency allele imputation. Mean r^2 between true and imputed genotype dosage for YRI SNPs found in sequencing, using the full set of HapMap 3 genotyped SNPs as tag SNPs (solid line), and using only SNPs present on an earlier generation array (~1/3rd the density) (dashed line).

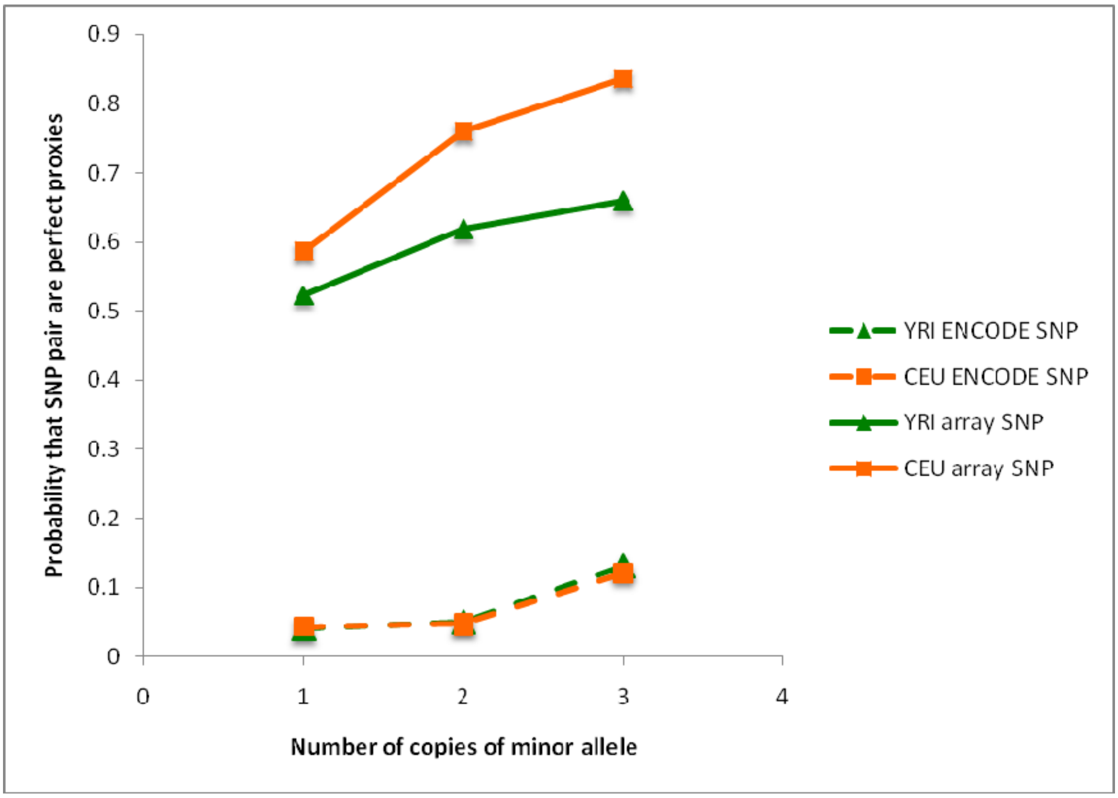


Figure S9. Proxy probability. The probability that two frequency-matched SNPS, less than 20 kb apart, are perfect proxies for each other, shown separately for pairs of SNPs on the arrays (solid) and for pairs of ENCODE SNPs (dashed).

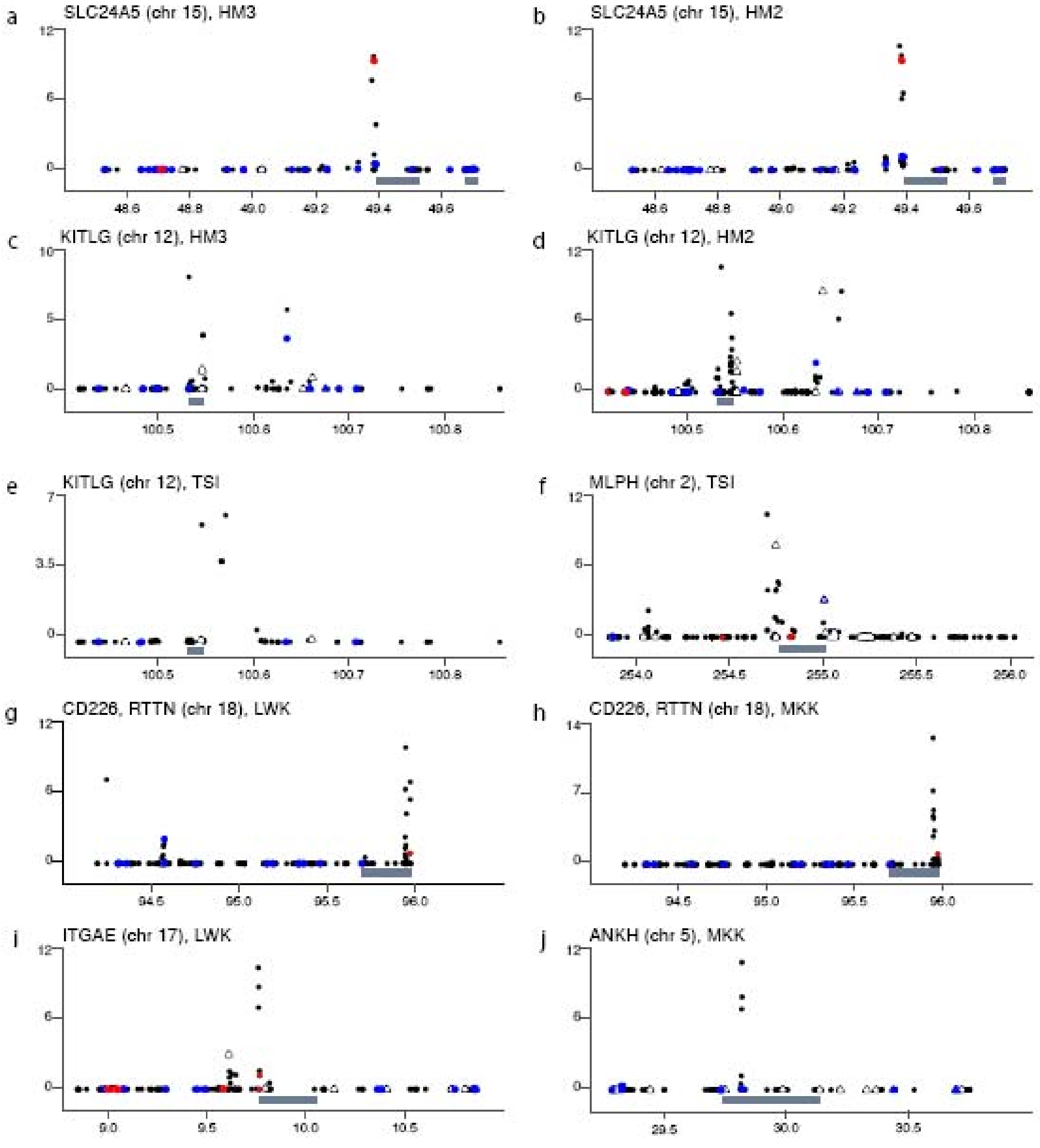


Figure S10. Signals of selection in previously identified and novel regions. CMS analysis of: *SLC24A5* in CEU in **a**. HapMap 3, and **b**. HapMap II; *KITLG* in CEU in **c**. HapMap 3, and **d**. HapMap II; **e**. *KITLG* in TSI, **f**. *MLPH* in TSI, *CD226* in **g**. LWK and **h.** MKK, **i** *ITGAE* in LWK, **j**. *ANKH* in MKK. Bars on x-axis indicate genes, black dots show CMS values, red dots indicate non-synonymous SNPs, blue dots indicate SNPs in conserved regions, white triangles indicate SNPs in putative transcription factor binding sites.