Supplementary Material for "Genotype, haplotype, and copy-number variation in worldwide human populations"

Contents

1 Preparation of SNP data

1.1 Overview

The study design involved the high-resolution genotyping of a diverse sample of individuals at genome-wide single-nucleotide polymorphisms (SNPs). The set of SNPs included SNPs spread across all autosomes, as well as SNPs on the X chromosome, the pseudoautosomal region on the X and Y chromosomes, the nonrecombining proportion of the Y chromosome, and the mitochondrion. The individuals genotyped were drawn from the HGDP-CEPH Human Genome Diversity Cell Line Panel[1,2] (the "HGDP-CEPH panel" henceforth), and were augmented for some analyses with individuals taken from the International Haplotype Map Project[3] (the "HapMap"). Following a series of quality control steps (Figure S17), an initial design using 513 HGDP-CEPH individuals was reduced to a final dataset of 485 individuals and 525,910 genome-wide SNPs.

1.2 Genotyping

We selected a geographically broad collection of 513 HGDP-CEPH samples from 29 populations for genotyping. DNA was derived from Epstein-Barr virus immortalized lymphocyte cell lines (LCL) maintained as part of the HGDP-CEPH panel[1]. Genotyping was performed using Infinium HumanHap550 Genotyping BeadChips (Illumina Inc., San Diego, CA). Samples were assayed along with ongoing experiments in batches of 48. For each sample, 1μg of DNA was used as template and the experiments were performed following manufacturer instructions. Our previous work has established that genetically, LCLs remain highly faithful to the source tissue used for immortalization[4].

Of the 513 samples, 316 were typed using HumanHap550 version 1 BeadChips and 197 were typed using HumanHap550 version 3 BeadChips. Raw data from HumanHap550 version 1 and version 3 chips were loaded as separate projects into Beadstudio version 3.1.4. Reclustering of SNP genotype calls was performed, discarding all genotypes below a no-call threshold of 0.15. All samples within each BeadStudio project (version 1 or version 3) were then reanalyzed using the newly derived genotype clusters.

1.3 Initial quality control

We generated a total of ∼275 million diploid genotypes in the 513 samples. After reclustering, 18 samples with a call rate <95% were permanently excluded from further analysis. The genotype call rate threshold of 95% resulted in the removal of 13 samples typed on version 1 BeadChips and 5 samples typed on version 3 BeadChips. The number of unique SNPs in common between version 1 and version 3 BeadChips is 545,066. To test genotype concordance across HumanHap550 version 1 and version 3 BeadChips, we genotyped both BeadChip types in each of three replicate samples. Analysis of these replicates produced a mean genotype concordance rate of 0.999938 (range 0.999909 to 0.999954); the average number of called genotypes across replicates was 539,161 of 545,066 attempted (range 538,112 to 540,476), and the average number of discordant calls was 33 (range 25 to 46).

A locus-specific genotype call rate threshold of 98% (after reclustering) resulted in the removal of 18,667 of the 545,066 SNPs; thus 526,399 unique SNPs were successfully typed across 495 samples. The mean genotype call rate across these samples was 99.75% after reclustering (range 96.24% to 99.95%; median 99.86%). Post-reclustering call rates were extremely high in most individuals (Figures S18 and S19), exceeding 98% in all except seven cases and exceeding 99% in all except 17 cases.

At this point, two HGDP-CEPH samples — a Palestinian and a Papuan — were discarded. On the basis of a comparison of an initial version of our genotypes to data on 122 SNPs from Conrad *et al.*[5], these samples were suspected of having been mislabeled during the course of our project. Data preparation proceeded using the remaining 493 HGDP-CEPH samples and 526,399 SNPs, incorporating genotypes of 112 HapMap individuals previously genotyped by Illumina (using version 1 BeadChips).

1.4 Individuals

HGDP-CEPH panel. To verify the identities of the 493 remaining HGDP-CEPH individuals, we first verified that sex inferred on the basis of X-chromosomal heterozygosity and Y-chromosomal missing data matched the sex previously reported for each individual[1,6]. All individuals previously reported as male had at most 2.05% heterozygous SNPs on the X chromosome and at most 2 of 10 SNPs with missing data on the Y chromosome, whereas all individuals previously reported as female had at least 11.74% heterozygous SNPs on the X chromosome and at least 5 of 10 SNPs with missing data on the Y chromosome (at least 9 of 10 in all except two cases).

We then compared genotypes at 122 autosomal SNPs that overlapped with the study of Conrad *et al.*[5] Each of 1039 HGDP-CEPH individuals from the Conrad *et al.*[5] dataset was compared with each of 493 HGDP-CEPH individuals genotyped in the current study. Except for four individuals that were not genotyped by Conrad *et al.*[5] (Adygei 1383, Adygei 1384, Biaka Pygmy 980, Russian 890), for each individual typed in the current study, the genotypes of the 122 SNPs obtained using Illumina BeadChips almost exactly matched those associated by Conrad *et al.*[5] with the same individual label (or its duplicate, in cases where only one member of a duplicate pair was genotyped by Conrad *et al.*[5]). In some cases, up to 2 of the 122 SNPs had a discrepancy in which one dataset produced a heterozygote and the other produced a homozygote. However, other than known duplicates and pairs with the same individual label, no other pairs of individuals involving genotypes from the Conrad *et al.*[5] study and genotypes from the current study had more than 87% of SNPs in which both alleles agreed. Because the only pairs of individuals with a high level of genotype concordance between studies were those expected on the basis of identical individual labels, it was assumed that no new sample labeling errors or sample duplicates occurred in any of the 493 HGDP-CEPH samples since the time of the earlier Conrad *et al.*[5] study.

HapMap. For all 112 HapMap individuals, sex information inferred on the basis of X-chromosomal heterozygosity and Y-chromosomal missing data (using the same criteria as for the HGDP-CEPH samples) matched the previously reported sex. The sample from the HapMap did not contain any two individuals found by the HapMap Consortium[3] in their Supplementary Table 15 to have an "unreported relationship," although it did contain two Japanese individuals inferred to have relatively high inbreeding coefficients. Included as part of the HapMap sample were 28 parent/parent/offspring trios — 16 from the CEU sample, and 12 from the YRI sample. For all pairs of individuals in the HapMap sample, computations of P_0, P_1, and P_2 — the fractions of autosomal SNPs with 0, 1, and 2 alleles shared identical in state — were used to verify that relative pairs matched those expected[3]. This computation utilized 7734 SNPs on chromosome 21. For all parent/parent/offspring trios previously reported, parent/offspring relationships were in fact inferred between the offspring in each trio and each of the two parents ($P_0 < 0.0006$ for each parent/offspring pair), and no other parent/offspring relationships involving two HapMap samples were identified ($P_0 > 0.04$ for all other pairs). No two HapMap individuals were found to be sample duplicates ($P_2 < 0.73$ for all pairs).

Final set of individuals. The sample of 493 HGDP-CEPH individuals included seven pairs of duplicate samples; for each duplicate pair the individual not in the H1048 subset of the HGDP-CEPH panel[6] was excluded from the final set for data analysis (Druze 589, Bedouin 652, Melanesian 659, Melanesian 826, Biaka Pygmy 981, Biaka Pygmy 1087, and Biaka Pygmy 1092). Biaka Pygmy 980 was also excluded due to a previously reported labeling error[6,7]. The final set of individuals for data analysis included 485 HGDP-CEPH and 112 HapMap individuals.

The set of 485 HGDP-CEPH individuals included 440 individuals from the H952 subset, which contains no first- or second-degree relatives[6]. Among the relatives, four parent/parent/offspring trios were included: Melanesian 655 (father), 656 (mother), and 657 (daughter); Melanesian 788 (father), 660 (mother), and 789 (son); Melanesian 788 (father), 660 (mother), and 824 (son); Pima 1037 (father), 1038 (mother), and 1039 (son). Analyses that excluded relatives were restricted to 84 HapMap individuals — excluding offspring from trios — and 443 HGDP-CEPH individuals. The HGDP-CEPH set of "unrelated" individuals included the 440 individuals from the H952 set without close relatives[6], together with Pima individuals 1046 and 1049 and Maya 866. These three individuals were retained, as none of their close relatives were among the individuals genotyped (or, in the case of Maya 866, the close relative that if genotyped would have led to her exclusion was not genotyped). Sample sizes for the various populations are displayed in Table S5.

1.5 Populations

The populations studied are shown on the map in Figure S1 at the coordinates used in Rosenberg *et al.*[8]. These coordinates (Table S6) match those of Cann *et al.*[1], except that an averaging procedure was used for populations given by Cann *et al.*[1] with a range for the latitudes and longitudes, and an updated location

was used for Mongola. Populations were classified by geographic region in the same manner as in previous work with the same individuals[7]. For some analyses, the geographic regions of Europe, Middle East, and Central/South Asia were grouped into a "Eurasia" region. "Africa" refers to Sub-Saharan Africa. The HapMap Chinese (CHB) and Japanese (JPT) samples were included with East Asia; HapMap Yoruba (YRI) individuals were included with Africa, and HapMap European Americans (CEU) were included with Europe.

1.6 SNPs

After the set of individuals was established, apparent heterozygotes among males for X-chromosomal loci were recoded as missing data, as were heterozygotes at mitochondrial SNPs and non-missing Y-chromosomal genotypes among females. Similar procedures to those of Conrad *et al.*[5] were then applied to remove SNPs with lower quality data. Separate quality checks were applied to the recoded dataset, and upon completion of all checks, those SNPs that did not pass any one of the tests were excluded.

Monomorphic SNPs. In the final set of 597 individuals used in data analysis (485 HGDP-CEPH and 112 HapMap), 42 monomorphic SNPs were identified. The remaining SNPs included 48,723 AC, 214,444 AG, 214,751 CT, and 48,439 GT polymorphisms.

SNPs with missing data. Considering the set of 527 unrelated individuals (443 HGDP-CEPH and 84 HapMap), 161 SNPs with at least 10% missing data were identified. For autosomal, pseudoautosomal, and mitochondrial SNPs, the fraction of missing data was calculated as the total fraction of individuals whose genotypes were missing. 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 (202), the number of males considered (325), the number of females with missing genotypes, and the number of males with missing genotypes. For Y-chromosomal SNPs, the missing data rate was calculated as the fraction of males whose genotypes were missing.

SNPs were also identified for which one or more populations had a sample size of fewer than 5 alleles in the sample of 527 unrelated individuals. This criterion, which was not applied to the Y chromosome or the mitochondrial genome, led to the identification of 135 SNPs — six autosomal and 129 X-chromosomal SNPs.

SNPs not in Hardy-Weinberg equilibrium. From the set of 527 unrelated individuals, two population groupings with relatively low levels of population structure in previous work[7] were constructed: a Middle East group consisting of Bedouin, Druze, and Palestinian samples (107 individuals), and a sub-Saharan Africa group consisting of the Bantu (Southern Africa), Bantu (Kenya), Mandenka, and Yoruba populations (63 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[9]. For X-chromosomal SNPs, males were included in the tabulation of allele frequencies for the computation of expected genotype frequencies, but they were ignored in the hypothesis test. Only SNPs with at least four copies of the minor allele in both of the population groups were considered as possible candidates for exclusion. Among such SNPs, those SNPs that had either or both of the following properties were identified: (1) the chi-squared test statistic was greater than 19.51142 ($P < 10^{-5}$ for a χ^2_1 distribution) in either of the two population groups; (2) the chi-squared test statistic was greater than 6.634897 ($P < 10^{-2}$ for a χ^2_1 distribution) in both of the population groups (Figure S20). Using these criteria, 198 SNPs were identified.

SNPs discordant between duplicates. For each SNP, concordance of genotypes was evaluated between the two individuals in each duplicate pair. For all SNPs, non-identical genotypes in which neither individual had missing data were declared discordant. Two SNPs were identified in which two of the seven duplicate pairs had discordant genotypes.

SNPs with Mendelian incompatibilities. For each of the 32 trios (4 HGDP-CEPH, 16 HapMap CEU, and 12 HapMap YRI), SNPs were tested for Mendelian incompatibilities. A total of 26 SNPs with at least three Mendelian incompatibilities among the 32 trios were identified.

Summary of excluded SNPs. Not taking into account the fact that some SNPs failed more than one of the checks, the total number of SNPs identified by the various quality checks was 564 — 42 that were monomorphic, 161 with a high overall missing data date, 135 with considerable missing data in at least one population, 198 with Hardy-Weinberg disequilibrium, 2 with discordance between duplicates, and 26 with Mendelian incompatibilities. Accounting for SNPs that failed more than one of the quality checks, 489 distinct SNPs were identified and were discarded from the final dataset for analysis. Excluding these SNPs, the SNP set for data analysis included 525,910 SNPs — 512,762 autosomal, 13,052 X-chromosomal, 9 Y-chromosomal, 15 pseudoautosomal, and 72 mitochondrial SNPs (Table S7).

1.7 Missing data rate

Of the $2(597)(512,762) = 612,237,828$ autosomal genotypes possible in the full sample of 597 individuals, the number of missing genotypes was 759,570. Similarly, the proportions of missing genotypes for the Y chromosome, the pseudoautosomal region, and the mitochondrial genome were 48 of 3285, 92 of 17,910, and 1194 of 42,984 possible genotypes. Of the $(365 + 2 \times 232)(13,052) = 10,820,108$ X-chromosomal genotypes possible, the number missing was 36,982. Combining all 525,910 SNPs, the missing data rate was

$$\frac{759,570 + 48 + 92 + 1194 + 36,982}{612,237,828 + 3285 + 17,910 + 42,984 + 10,820,108} = \frac{797,886}{623,122,115} \approx 0.13\%.$$

Most individuals and SNPs had a very low missing data rate (Tables S8 and S9). None of the individuals had more than 3.5% missing data, as determined using all 525,910 SNPs.

1.8 Genotyping error rate

A rate of genotype discrepancy was determined based on duplicate samples, using 513,008 autosomal SNPs (prior to removal of 246 autosomal SNPs that failed quality checks). Considering autosomal SNPs for which both individuals in a duplicate pair were genotyped, 174 discrepancies were observed in which one individual was homozygous and the other was heterozygous, and 1 discrepancy was observed in which the two individuals were homozygous for different alleles. Therefore, the rate of discrepancies per diploid genotype was $175/3,580,862 \approx 4.92 \times 10^{-5}$.

Comparing the genotypes at the overlapping 122 SNPs to the data from Conrad *et al.*[5], there were 482 individuals who were genotyped in both studies (not counting seven cases in which one study contained the duplicate cell line of an individual but not the identical cell line). Excluding comparisons in which one study produced missing data, the rate of discrepancies between studies per diploid genotype was $54/57,764 \approx 9.35 \times 10^{-4}$. This rate is somewhat higher than the discordance rate for duplicate cell lines genotyped in the present study, but note that the Conrad *et al.*[5] study used a different genotyping technology.

2 Population-genetic analysis of unphased SNP genotype data

2.1 Allele frequencies

For comparisons of SNP allele frequencies between geographic regions, we used the 443 unrelated HGDP-CEPH individuals and the 512,762 autosomal SNPs. To produce bivariate graphs of allele frequencies in two geographic regions (Figure S21A), we plotted the number of SNPs with minor allele frequency in each of a series of bins. We employed a resampling procedure to adjust for sample size differences among geographic regions. For each SNP, we sampled 40 alleles (with replacement) from each of the geographic regions. For a given pair of regions, the minor allele was identified from the pooled sample of 80 alleles for the two regions. In cases where both alleles had exactly 40 copies in the combined sample, the minor allele was chosen randomly. The numbers of SNPs in each of the $41 \times 42/2 = 861$ possible bivariate frequency categories were then tabulated. To obtain values proportional to SNP density per unit area, for triangular bins along the diagonal (frequency sum of one for the two population groups), this number of SNPs was doubled to account for the arbitrary decision about which allele was the minor allele and to account for the fact that the triangular bins had only half the area of usual square bins. Univariate allele frequency spectra (Figure S21B) were based on the same resamples as those used in producing the bivariate spectra. Correlation coefficients of allele frequencies, obtained based on the values plotted but using both alleles at each SNP, are shown in Table S10.

2.2 Linkage disequilibrium

Linkage disequilibrium (LD) was measured for the unphased data using the HR^2 statistic[10], a measure analogous to the r^2 statistic for phased data that for a pair of SNPs considers a normalized squared difference between the proportion of double homozygotes expected under linkage equilibrium and the proportion of double homozygotes observed. For each population, we computed HR^2 for all pairs of autosomal SNPs with physical distance <70.5kb. We used a resampling procedure to adjust for possible influence of sample size on homozygosity-based LD statistics[11]. For each pair of SNPs, a random set of five individuals was sampled without replacement in each population, and the LD computation was performed using those five individuals (ignoring missing data among the five individuals). Pairs of SNPs in which the five individuals chosen were monomorphic at one or both SNPs were excluded from the computation. In the computations of HR^2, the homozygosity of a locus was obtained using all individuals who had genotypes present at the locus, while the

double homozygosity for a pair of loci was obtained considering all individuals who had genotypes present at both loci. As certain missing data configurations can produce $HR^2 > 1$ with this approach, a small proportion of values of HR^2 that exceeded 1 were set to 1.

SNP pairs were placed in bins of size 250bp (including the lower endpoint in the bin). For each population, starting at 500bp, the mean HR^2 value in 1kb windows was computed every 250 base pairs. As an example, the value plotted at 10kb in Figure 2b (and 2c) is the mean HR^2 for all SNP pairs with physical distance in [9500,10500). The standard error plotted in Figure 2c is obtained using this same collection of pairs of SNPs.

Regression of LD on geographic distance from East Africa was performed using mean HR^2 values for particular physical distance bins. Geographic distance was computed using the approach of Rosenberg *et al.*[8], starting from Addis Ababa (9°N 38°E) and traveling along the waypoint routes of Ramachandran *et al.*[12]. Paths involving the Americas all passed through 64°N 177°E and 54°N 130°W, paths involving Oceania passed through 11°N 104°E, and paths involving Africa (including the Mozabite population) passed through 30°N 31°E. Paths from Europe (excluding Adygei) to Africa, the Middle East (excluding Mozabites), or Oceania also passed through 41°N 28°E.

2.3 Geographic distribution

We used the rarefaction approach to assess the distributions of alleles across geographic regions while adjusting for differing sample sizes[13]. This method, which extends the rarefaction formula of Kalinowski[14] for private allelic richness, examines the mean number of alleles with each of a set of possible geographic distributions, considering all possible subsamples with equal size g from each of the geographic regions. The analysis used the 443 unrelated individuals and 512,509 autosomal SNPs, omitting those SNPs with >10% missing data in any of the five major geographic regions. The omission of SNPs with >10% missing data made it possible to consider larger values of the standardized sample size g. The proportions of SNPs with particular geographic distributions were obtained by averaging estimates across loci at $g = 35$.

2.4 `Structure`

The Bayesian clustering software `Structure`[15,16] was used to cluster individuals using their SNP genotypes. Replicate `Structure` runs used a burn-in period of 20,000 iterations followed by 10,000 iterations from which estimates were obtained. All runs were based on the admixture model, in which each individual is assumed to have ancestry in multiple genetic clusters, using the F model of correlation in allele frequencies across clusters. Graphs of `Structure` results were produced using `Distruct`[17].

This analysis used the 443 HGDP-CEPH unrelated individuals, as well as subsets of this collection corresponding to individual geographic regions. Four SNP subsets were obtained, each containing ∼1% of the autosomal SNPs. Chromosomes were placed in numerical order, and SNPs were ordered on each chromosome using the build 36.2 human genome sequence (dbSNP build 127). With SNPs ordered from 1 to 512,762, the four subsets consisted of SNPs in numbered positions 1 mod 100, 26 mod 100, 51 mod 100, and 76 mod 100.

For a given subset of individuals and value of K, the number of clusters considered, ten replicate analyses with `Structure` were performed for each 1% collection of SNPs. The 40 replicates for each subset of individuals and each K were then analyzed with `CLUMPP`[18] to identify common modes among the replicates, using a procedure similar to that of Wang *et al.*[19]. `CLUMPP` analysis proceeded using the *LargeKGreedy* algorithm with 10,000 random permutations. A set of runs was classified as characterizing a single mode if all pairs in the set had a symmetric similarity coefficient $G' > 0.9$. Note that because of possible nontransitivity of the criterion $G' > 0.9$, it sometimes occurred that a run was considered part of two or more distinct modes.

For each mode identified, we ran `CLUMPP` a second time (using the *LargeKGreedy* algorithm and 10,000 random permutations), using only the replicates belonging to the mode. From this analysis, for each mode, we obtained the mean across replicates of the cluster membership coefficients of each individual. For each subset of individuals and each value of K, we identified the most frequently occurring mode, breaking ties using the `CLUMPP` H' score. We also obtained mean log likelihood scores across replicates in the most frequent mode. Further details regarding the `Structure` and `CLUMPP` analyses, including additional results and a description of the basis for selecting some of these results for display in Figure 1c, are provided in Section 7.2.

2.5 Population tree

A neighbor-joining tree of populations[20] was obtained based on pairwise allele-sharing distance among populations. This analysis used 512,762 autosomal SNPs and the 443 unrelated individuals. Confidence values were obtained using 1000 bootstrap resamples across loci. The computation of bootstrap distances was performed using `microsat`[21], and the tree was obtained using the `neighbor`, `consense`, and `drawtree` programs

in the `phylip` package[22]. The production of the consensus tree used extended majority rule consensus (greedy consensus[23]), as implemented in `consense`. External branches are drawn with equal lengths, and internal branch lengths are proportional to bootstrap support.

2.6 Multidimensional scaling

A matrix of pairwise distances was constructed for the 443 unrelated HGDP-CEPH individuals, using the 512,762 autosomal SNPs. Between-individual distances were obtained using allele-sharing distance, $P_0 + P_1/2$, where P_k represents the proportion of loci at which the individuals shared exactly k alleles identical in state. The overall distance between individuals was obtained as the average across loci. Classical metric multidimensional scaling[24,25] was applied to the individual distance matrix to provide a representation of the matrix in two dimensions. The resulting coordinates were then rotated 225° to place the populations in an approximate geographic orientation. This analysis utilized the `cmdscale` function in `R`[25]. Two goodness-of-fit criteria for the proportion of the distance matrix explained by the MDS scaling are $\alpha_{1,2}$ and $\alpha_{2,2}$ (eqs. 14.4.7 and 14.4.8 of Mardia *et al.*[24]). For the plot shown for SNP data, $\alpha_{1,2} = 19.5\%$ and $\alpha_{2,2} = 88.7\%$.

2.7 Genetic and geographic distance

We analyzed the relationship of genetic and geographic distance for pairs of populations. For the 512,762 autosomal SNPs, genetic distance was computed with F_{ST}, using eq. 5.3 of Weir[9]. Geographic distance between populations used the same waypoint routes as were used in Section 2.2.

3 Preparation of haplotype data

3.1 Haplotype estimation with geographic region labels

Haplotypes and missing genotypes were estimated with `fastPHASE`[26] version 1.3, ordering SNPs on each chromosome according to positions from build 36.2 of the human genome sequence. As in Conrad *et al.*[5], for estimating haplotypes and missing genotypes, geographic region labels (Table S5) were applied during the model fitting procedure to enhance accuracy ("population labels" in Scheet & Stephens[26]). The number of haplotype clusters was set to 20, and we employed the default setting of 20 runs of the EM algorithm. This analysis was used to generate a "best guess" estimate of the true underlying patterns of haplotype structure.

We included all 597 available individuals during haplotype estimation (485 HGDP-CEPH and 112 HapMap). This combined sample included related individuals; however, during haplotype estimation and model fitting (Section 4.3), we treated all individuals as unrelated. Haplotype phase was estimated for all autosomes as well as for the pseudoautosomal region; for the X chromosome the haplotype estimation procedure treated males as having known haplotype. As described below, we removed relatives from the phased haplotype data to create a dataset of 527 unrelated individuals (443 HGDP-CEPH and 84 HapMap).

The "best guess" estimate of haplotype structure was used in the analyses of LD (Section 4.1) and of haplotype length and frequency (Section 4.2), as well as in the plots of haplotype structure (Section 4.4).

3.2 Haplotype datasets for analysis of population structure

To generate haplotype datasets for analyses of population structure (geographic distribution of haplotypes, `Structure`, population tree, and multidimensional scaling) we performed additional `fastPHASE` model fitting without using the geographic labels. For these analyses, we wanted to make inferences regarding population structure, rather than leverage a known or assumed structure for more accurate genotype estimation. As above, we included all 597 available individuals for model fitting, only afterwards removing relatives. The procedure for generating the haplotype datasets for population structure analyses is described in Section 4.3.

4 Population-genetic analysis of haplotype data

4.1 Linkage disequilibrium

LD was measured from the haplotype data using the r^2 statistic[27]. This analysis used the autosomal haplotype estimates, restricting attention to unrelated individuals. For each population, we computed r^2 for all pairs of autosomal SNPs with physical distance <70.5kb. Analogously to the computation of HR^2 for the unphased data, we used a resampling procedure to adjust for possible influence of sample size on r^2. For each pair of SNPs, a random set of ten haplotypes was chosen in each population, and the LD computation was performed

using those ten haplotypes. Pairs of SNPs in which the ten haplotypes chosen were monomorphic at one or both SNPs were excluded from the computation. SNP pairs were placed in overlapping bins in the same manner as in the HR^2 computations for unphased data, and the mean was taken for each bin. The decay of mean r^2 with physical distance is plotted in Figure S4. Regression of LD on geographic distance from East Africa was performed using r^2 values in the same manner as in the regressions involving HR^2.

4.2 Joint distribution of haplotype length and frequency

Two haplotype properties that are inherently connected are the length of a haplotype and the frequency of that haplotype. Long haplotypes tend to have lower frequencies than do short haplotypes. To assess both of these properties simultaneously without applying predefined window sizes of haplotype lengths, we devised a method that computes the length of every observed haplotype as well as its corresponding frequency.

For this analysis, we used the "best guess" phased data for 527 unrelated HGDP-CEPH and HapMap individuals, analyzing the whole autosomal genome. Let the n SNPs of a given chromosome arm belong to the set S, where s_i denotes SNP i ($i = 1, ..., n$) and p_i denotes the position of SNP s_i. The set S is ordered so that $p_i < p_{i+1}$ and so that each SNP has two possible states, "0" and "1". For the set of SNPs $\{s_i, ..., s_j\}$ (with $i < j \leq n$), denote the set of possible haplotypes by h_{ij}. In this context a haplotype is defined as the series of states (0 or 1) along a specific chromosome, for some SNPs s_i to s_j. If two chromosomes have identical states at all SNPs from s_i to s_j, then the two chromosomes have identical haplotypes in $[i, j]$; otherwise, the chromosomes have different haplotypes. The number of possible haplotypes for the SNP set $s_i, ..., s_j$ is 2^{j-i+1}; this number grows quickly as $j - i + 1$ increases, but in practice, the number of unique haplotypes that are observed is relatively small. Denote the number of observed unique haplotypes for the set h_{ij} of haplotypes by K_{ij}. For the SNP set $\{s_i, ..., s_j\}$, each observed unique haplotype is denoted h_{ijk}, where $k = 1, ..., K_{ij}$.

Starting from the first SNP s_1, we move to SNP s_2, and we compute (and store) the length in base pairs ($\ell_{1,2} = p_2 - p_1 + 1$) of the haplotypes in the set $h_{1,2}$, and the frequencies of all haplotypes $h_{1,2,k}$. We then proceed to SNP s_3, and compute the length and frequency for all haplotypes $h_{1,3,k}$. This procedure is repeated for all sets of haplotypes $h_{1,j}$, $j = 2, ..., n$ (in practice, we truncate the calculation when the set of observed unique haplotypes has the same size as the number of sampled chromosomes — which occurs well before the end of the chromosome). Thus, for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$, we compute the length and the frequency for each haplotype h_{ijk}, which gives us the joint distribution of haplotype lengths and haplotype frequencies without using window sizes to define haplotype length.

We computed the joint haplotype length and frequency distribution for each of the 29 HGDP-CEPH populations as well as for the 4 HapMap populations (Figure S5). To adjust for sample size differences across populations, in each population for each set of haplotypes h_{ij}, we performed the analysis using 12 randomly chosen chromosomes (sampled without replacement). For convenience, we ignored haplotypes of frequency 1 of 12, so that the normalization used in computing the fraction of haplotypes that lie in a given length and frequency bin is based only on haplotypes of frequency at least 2 of 12.

4.3 A model for local clustering of haplotypes

Motivation. Here we introduce a novel model-based approach for describing and displaying haplotypic variation within and among populations. Our approach, which is based on the model underlying `fastPHASE`[26], can be viewed as a summary of common haplotype frequencies in a sample.

In approaches to haplotype variation that consider windows of a given haplotype length (such as in Section 4.1), haplotypes are first estimated, then they are binned within windows of a given size, with the choice of window size having a sizeable effect on the haplotype frequency spectrum. In such analyses, it is important to investigate multiple haplotype lengths, as it may be difficult to determine the ideal window size for analysis.

An alternative approach for circumventing the issue of window size is to summarize variation using an LD model that locally captures the natural extent of haplotypes[26]. Over short physical distances, haplotypes sampled from a population of chromosomes can be clustered into groups of similar haplotypes; these "haplotype clusters" then summarize the overall variation in the population. Our approach enables the clustering process to be applied to an entire chromosome by using a hidden Markov model for the underlying "haplotype cluster" memberships of individual haplotypes in the sample. The model has been used previously for estimating haplotypes and missing genotypes, as implemented in the software package `fastPHASE`[26].

Here we utilize the machinery of `fastPHASE` to obtain the frequencies of the latent haplotype clusters, which are represented by their cluster centers. These centers, or "fuzzy haplotypes," represent locally the common haplotypes in a random sample of chromosomes from a population (or multiple populations). Use of these model-based cluster frequencies allows marker-wise summaries of haplotype variation to be produced in the form of the "frequency distribution" of haplotype clusters.

The model. We recapitulate some notation from Scheet & Stephens[26]. We assume unphased individual multilocus diploid genotypes g, observed at M SNP markers in n diploid individuals. We assume that there are K haplotype clusters, which we estimate from the data. For convenience we set K equal to 20.

Let $z_{im}^{:}$ denote the unobserved pair of haplotype cluster memberships for individual i at SNP marker m. Let $p_{mk}(i)$ denote the relative frequency of haplotype cluster k $(1, \ldots, K)$, in individual i at marker m $(1, \ldots, M)$. To calculate $p_{mk}(i)$ we integrate over the possible pairs of cluster memberships, given the observed data g and model parameters ν, as follows:

$$p_{mk}(i) = \frac{\left[\sum_{k'=1}^{K} P\big(z_{im}^{:} = \{k, k'\}|g, \nu\big)\right] + P\big(z_{im}^{:} = \{k, k\}|g, \nu\big)}{2},$$

where $P(z_{im}^{:}|g, \nu)$ is given by Scheet & Stephens[26]. The quantities $p_{mk}(i)$ for $k = 1, ..., K$ can be viewed as the relative cluster frequencies for a very small population of chromosomes (of size 2). The integration over possible pairs of cluster memberships amounts to integrating over uncertainty in haplotypic phase.

Now suppose that instead of a homogeneous sample of diploid individuals, we sample individuals from S predefined "populations." We can calculate $p_{mk}^{(s)}$, the relative frequency of cluster k in population s $(1, \ldots, S)$, by averaging $p_{mk}(i)$ over members of this population as follows:

$$p_{mk}^{(s)} = \frac{1}{n_s} \sum_{i \in \mathcal{I}_s} p_{mk}(i).$$

In this equation, $\mathcal{I}_s$ is the subset of individuals who belong to population s, and n_s is the number of elements in $\mathcal{I}_s$ (that is, the sample size for population s). Calculation of $P\big(z_i|g, \nu\big)$ can be accomplished efficiently with a dynamic programming algorithm, and the parameters ν are estimated via an EM algorithm[26].

Once we have obtained the common haplotype frequencies $\{p_{mk}^{(s)}\}$ at a particular marker m, we can use them to summarize the haplotype variation at that marker, within and among populations. Although we are assessing haplotype variation, and we are therefore inherently modeling genetic variation at multiple SNPs simultaneously, the information may be conveniently summarized pointwise at each marker, thus avoiding the problem of choosing window sizes.

Sampling latent cluster memberships. Because the EM algorithm generally obtains local modes of the likelihood $P(g|\nu)$, we run the EM algorithm T times, obtaining sets of parameter estimates $(\hat{\nu}_{(1)}, \ldots, \hat{\nu}_{(T)})$. From each of these parameter sets, we can sample an instantiation from the conditional distribution of the chromosome-wide list of cluster memberships z, given the estimated parameters and the observed genotype data. An algorithm for sampling from $P\big(z|g, \nu\big)$ is given by Scheet & Stephens[26].

For use in analyses of population structure, we generated a single sample of haplotype cluster memberships from each of $T = 10$ model fittings (that is, ten independent runs of the EM algorithm). That is, we generated ten datasets so that at each SNP position across the genome, each individual was given a pair of haplotype cluster memberships, with each cluster membership equaling an integer ranging from 1 to 20. These datasets were then analyzed in the same manner as one would analyze unphased multiallelic datasets. As noted above, the ten model fittings were obtained by treating the sample of individuals as homogeneous, rather than by using $S = 7$ populations characterized by the geographic regions in Table S5.

4.4 Haplotype cluster plots

Figures 3 and S6 each contain visualizations of the haplotype cluster frequencies $\{p_{mk}^{(s)}\}$ in the 527 unrelated HGDP-CEPH and HapMap individuals, across different populations and geographic regions, for particular regions of the genome. Each plot is based on one of the 20 parameter estimate sets used to obtain the "best guess" haplotypes used in Sections 4.1 and 4.2. Within each box (corresponding to a population), cluster frequencies in the population are arranged vertically at consecutive SNPs. Each SNP is indicated by a horizontal position, and the 20 colors indicate the frequencies for 20 haplotype clusters. No attempt is made to model the degree of similarity among the different haplotypes represented by different clusters; thus, no meaning is intended by the similarity or dissimilarity of the colors referring to different haplotype clusters.

Although it is difficult to visually ascertain exact haplotype cluster frequencies at individual SNPs, the gradual change in frequencies due to the gradual decay of LD allows the information at adjacent SNPs to blend together smoothly. One natural summary of each haplotype cluster visualization, which is largely continuous across each picture, is haplotype cluster homozygosity. We computed this homozygosity using the haplotype cluster frequencies (treated as parametric frequencies), averaging across the ten haplotype cluster datasets to obtain an overall estimate. For comparison, we also computed a standardized haplotype cluster homozygosity for each population, subtracting the genome-wide mean haplotype cluster homozygosity and then dividing by the standard deviation of haplotype cluster homozygosity across the genome (Figure S7).

4.5 Geographic distribution

We used rarefaction[13] on the ten imputations of the haplotype clusters, averaging results across imputations to obtain the final estimates. This analysis was performed with the 443 unrelated HGDP-CEPH individuals in the same manner as for the SNP data. As the haplotype datasets contain no missing data, analysis of geographic distributions was performed at each autosomal locus, and results were averaged across loci. Rarefaction corrects for sample size only after production of haplotype datasets; sample size may have a small influence during dataset production, as larger samples contribute more information during model fitting.

4.6 `Structure`

Analysis of imputed haplotype clusters using `Structure` and `CLUMPP` with 443 unrelated HGDP-CEPH individuals proceeded in a similar manner to the analysis with unphased SNPs (using the $G' > 0.9$ criterion). Each of the ten imputations of cluster memberships (one from each of the ten model fittings described above) was used in the `Structure` analysis. For each of the ten datasets, two subsets of the SNP data were obtained, each containing ∼1% of the autosomal SNPs. SNPs were ordered on each chromosome using the build 36.2 human genome sequence, and separately on each chromosome, SNPs in numbered positions 1 mod 100 were placed in one subset, and SNPs in positions 51 mod 100 were placed in a second subset. A total of 40 `Structure` runs were performed for each value of the model parameter specifying the number of clusters in the `Structure` analysis — two replicates for each combination of one of the ten imputations and one of the two 1% subsets of SNPs. `CLUMPP` analysis was performed on these 40 replicates, in the same manner as for the SNP dataset. Details of the results are described in Section 7.2.

4.7 Population tree

A neighbor-joining tree of populations was obtained based on the haplotype cluster membership data in the same manner as with the SNP data, using the 443 unrelated HGDP-CEPH individuals. This analysis was restricted to every 10th SNP marker across the autosomes (on each chromosome, the SNPs chosen were those in 1 mod 10 positions when enumerated according to the build 36.2 genome sequence, starting from 1). Confidence values were obtained by combining 1000 bootstrap-resampled distance matrices — 100 resamples for each of the same ten imputations used in the `Structure` analysis.

4.8 Multidimensional scaling

To generate a distance matrix for use in multidimensional scaling, we computed a "haplotype distance" between all pairs of individuals. We define $d^h_m(i,j)$ as the haplotype distance between the haplotype cluster probability vectors for individuals i and j at marker m, calculated in the following manner:

$$d^h_m(i,j) = \sqrt{\sum\nolimits_{k=1}^{K} \left(p_{mk}(i) - p_{mk}(j)\right)^2}.$$

Finally, we obtained a haplotype genetic distance between individuals by averaging over multiple SNP markers and multiple model fittings.

Implicitly, $p_{mk}(i)$ is associated with a single model fit, or a single set of parameters ν. For producing the plots in Figure 1d, we calculated an average haplotype distance between all pairs of individuals by averaging $d^h_m(i,j)$ from every 10th SNP marker across the autosomes (on each chromosome, the SNPs chosen were those in 1 mod 10 positions when enumerated according to the build 36.2 genome sequence, starting from 1) over ten sets of model parameters (estimated from the same ten independent model fittings used in the `Structure` and population tree analyses). The final average haplotype distance was produced from

$$d^h(i,j) = \frac{1}{10|\mathcal{M}|} \sum_{t=1}^{10} \sum_{m \in \mathcal{M}} d^h_m(i,j)_t,$$

where $\mathcal{M}$ is the set containing every 10th autosomal SNP marker, $|\mathcal{M}|$ is the number of elements in this set, and $d^h_m(i,j)_t$ represents the haplotype distance calculated from EM run t $(1, \ldots, 10)$.

Multidimensional scaling was then applied to the distance matrix in the same manner as for the SNP data. Goodness-of-fit statistics[24] for the haplotype plot in Figure 1d were $\alpha_{1,2} = 19.5\%$ and $\alpha_{2,2} = 80.2\%$.

5 Preparation of CNV data

5.1 Detecting CNVs using PennCNV

We applied the PennCNV algorithm as an experimentally validated CNV detection approach[28]. PennCNV was developed for the CNV analysis of genotyping intensity data on high-density SNP arrays (such as Illumina HumanHap). PennCNV integrates multiple information sources, including the normalized total signal intensity for each marker (the "Log R Ratio"), the allelic signal intensity ratio (the "B Allele Frequency"), the SNP allele frequency, the physical distance between neighboring markers, and pedigree information when available.

We used the previously validated default quality control criteria, excluding samples with a log R ratio standard deviation of >0.28, a median B allele frequency of >0.55 or <0.45, or a B allele frequency drift of >0.002 (for more details see Wang *et al.*[28]). As the PennCNV algorithm is more sensitive and specific to CNVs covering greater numbers of SNPs in the HumanHap550 array[28], use of a minimum number of SNPs in CNV detection increases the reliability of CNV calls (with a consequent reduction in calls per individual). We set 10 SNPs as the minimum detection threshold in the algorithm (≥10). Using high-quality HapMap samples, we have previously shown that a 10-SNP threshold (>10) results in ∼9% offspring CNV calls (excluding immunoglobulin regions) not detected in parents; this value provides a combined false positive and false negative rate measure for CNV calling accuracy[28]. As we describe below (Section 5.3), we further estimate from concordance of replicates that the false positive rate for CNV detection is no more than 0.7%.

5.2 Data cleaning

Considering the the 485 HGDP-CEPH individuals used in the SNP analysis, 42 did not meet the quality thresholds of PennCNV, leaving 443 individuals for CNV analysis. Previous work suggests that CNVs longer than 1Mb are likely to be artifacts of the lymphoblastoid cell line creation process or subsequent transition to clonality[4]. Thus, to be conservative, we removed all CNVs longer than 1Mb in size from further analysis (26 autosomal CNV observations, 1 X-chromosomal CNV observation).

Of the remaining variants we removed 400 CNV observations that occurred in regions where V(D)J-type recombination is known to occur (chr2p11 — 64 observations, chr14q11.2 — 7 observations, chr14q32.33 — 129 observations, chr22q11.22 — 200 observations). In total 427 CNV observations were removed. Analysis of the remaining variants revealed 3552 CNVs at 1428 non-overlapping copy-number-variable loci. This collection contained 3503 autosomal CNVs (1394 non-overlapping loci) and 49 X-chromosomal CNVs (34 non-overlapping loci).

Of the 443 individuals in whom CNVs were analyzed, 405 are contained in the subset of the unrelated individuals used in the SNP analysis. We therefore constructed a CNV dataset consisting of 405 unrelated individuals. Upon removing relatives, 92 autosomal and 3 X-chromosomal CNV loci are no longer polymorphic, leaving 1302 autosomal and 31 polymorphic X-chromosomal CNV loci for the subset of 405 individuals, and 3024 autosomal and 45 X-chromosomal CNVs. Excluding loci with only one observation of a CNV, the number of autosomal CNV loci is 396. Thus, the dataset used for CNV population-genetic analysis — which, like the corresponding SNP and haplotype datasets does not include the X chromosome — consists of 405 individuals and 396 CNV loci (2118 CNVs; 1470 deletions at 262 loci and 648 duplications at 134 loci). For use of this dataset in the population-genetic analysis, at each autosomal CNV locus (that is, at each genomic region in which some individuals had a copy-number variant), genotypes were coded as homozygous 00 if no CNV was observed, 01 if a heterozygous deletion or duplication was observed, and 11 if a homozygous deletion or duplication was observed. Each genomic region in which both deletions and duplications were observed was treated as two separate CNV loci.

5.3 False positives and false negatives

This section describes the basis for our estimate that the false positive rate for CNV detection is less than 0.7%. To investigate the fractions of false positive and false negative CNV calls for the 396 CNV loci in the dataset used in the population-genetic analysis, we employed the strategy of relying on concordance of replicates. This approach arises in various problems relating to categorical data analysis and medical diagnostic testing[29,30,31,32]. Similarly to our setting, each of these contexts also contains situations in which the false positive and false negative rates of tests are of interest, but in which the "truth" of individual observations is viewed as unknown. In medical diagnostics, a typical situation involves repeated diagnostic tests on the same individual when the true disease status of the individual is unknown; in psychological statistics, multiple observers may assess the same subject for a condition when the truth about whether the individual has the condition is unknown. In our case, two cell lines originating from the same individual are assessed for CNVs when the true copy-number status is unknown.

Variants of the replication-based strategy for estimating error rates have recently been devised in the context of detection of CNVs[33,34,35]. The approach we use places bounds on the false positive and false negative rates, taking into account the level of concordance of replicate observations together with the fractions of assignments made in each of the possible observational classes. Following the notation of Pepe[32], let Y be the CNV status of a CNV call at a particular copy-number-variable locus. Thus, $Y = 1$ if an allele is called as a duplication or deletion, whichever is relevant at the locus, and $Y = 0$ if the allele is called as not being a duplication or deletion (for the remainder of the section we use the term "CNV" to refer to whichever type of copy-number variant is relevant at a locus). Let D be the true CNV status — $D = 1$ if the true allele is a CNV and $D = 0$ otherwise. Denote the (unknown) false positive rate — the fraction of non-CNV alleles called as CNVs — by $\alpha = \mathbb{P}[Y = 1|D = 0]$. Denote the (unknown) false negative rate — the fraction of CNV alleles not called as CNVs — by $\beta = \mathbb{P}[Y = 0|D = 1]$. We have the following table:

	$D = 0$	$D = 1$
$Y = 0$	$1 - \alpha$	β
$Y = 1$	α	$1 - \beta$

Let $\rho = \mathbb{P}[D = 1]$ denote the (unknown) probability that a CNV is truly present for a given allele. Denote the probability that an allele is called as a CNV, $\mathbb{P}[Y = 1]$, by τ. Then

$$\begin{aligned} \mathbb{P}[Y = 1] &= \mathbb{P}[Y = 1|D = 0]\mathbb{P}[D = 0] + \mathbb{P}[Y = 1|D = 1]\mathbb{P}[D = 1] \\ \tau &= \alpha(1 - \rho) + (1 - \beta)\rho. \end{aligned} \tag{1}$$

A second equation can be obtained using the concordance of replicates. Let Y_1 and Y_2 denote two separate calls of the same allele — that is, calls in two replicate cell lines from the same individual. Denote

$$\chi = \frac{\mathbb{P}[Y_1 = 1 \cap Y_2 = 1]}{\mathbb{P}[Y_1 = 1 \cup Y_2 = 1]}.$$

As genotyping of the two replicates proceeds independently, we assume conditional independence of the two genotype calls given the true CNV status of the allele. Thus, the numerator of χ is

$$\begin{aligned} \mathbb{P}[Y_1 = 1 \cap Y_2 = 1] &= \mathbb{P}[Y_1 = 1 \cap Y_2 = 1|D = 0]\mathbb{P}[D = 0] + \mathbb{P}[Y_1 = 1 \cap Y_2 = 1|D = 1]\mathbb{P}[D = 1] \\ &= \alpha^2(1 - \rho) + (1 - \beta)^2\rho. \end{aligned}$$

Y_1 and Y_2 are identically distributed. Therefore, the denominator of χ is

$$\begin{aligned} \mathbb{P}[Y_1 = 1 \cup Y_2 = 1] &= \mathbb{P}[Y_1 = 1 \cap Y_2 = 1] + 2\mathbb{P}[Y_1 = 1 \cap Y_2 = 0] \\ &= \alpha^2(1 - \rho) + (1 - \beta)^2\rho + 2\mathbb{P}[Y_1 = 1 \cap Y_2 = 0|D = 0]\mathbb{P}[D = 0] \\ &\quad +2\mathbb{P}[Y_1 = 1 \cap Y_2 = 0|D = 1]\mathbb{P}[D = 1] \\ &= \alpha^2(1 - \rho) + (1 - \beta)^2\rho + 2[\alpha(1 - \alpha)(1 - \rho) + \beta(1 - \beta)\rho]. \end{aligned}$$

Simplifying the equation for the denominator, we obtain

$$\chi = \frac{\alpha^2(1 - \rho) + (1 - \beta)^2\rho}{(2\alpha - \alpha^2)(1 - \rho) + (1 - \beta^2)\rho}. \tag{2}$$

We can solve equations 1 and 2 for α and β in terms of ρ, τ, and χ to obtain

$$\alpha = \frac{\tau - \rho\tau + \tau\chi - \rho\tau\chi - \sqrt{\rho\tau(1 - \rho)(1 + \chi)(2\chi - \tau - \tau\chi)}}{(1 - \rho)(1 + \chi)} \tag{3}$$

$$\beta = \frac{\rho - \rho\tau + \rho\chi - \rho\tau\chi - \sqrt{\rho\tau(1 - \rho)(1 + \chi)(2\chi - \tau - \tau\chi)}}{\rho(1 + \chi)}. \tag{4}$$

Note that in obtaining these values we take the negative root of a quadratic equation for α. The positive root leads to the nonsensical result that the false positive rate increases with ρ and the false negative rate decreases with ρ.

Equations 3 and 4 provide a basis for estimating the false positive and false negative rates as functions of the unknown parameter ρ, as the quantities τ and χ can be estimated and the estimates $\hat{\tau}$ and $\hat{\chi}$ inserted into eqs. 3 and 4. The false positive and false negative rates are estimated with respect to the particular collection of 396 CNV loci in the study; thus, we are estimating the false positive and false negative rates for

CNV calls at the particular collection of 405 individuals and 396 CNV loci. This is sensible, as our interest is in the extent to which erroneous calls might affect population-genetic analysis of this dataset.

An estimate for τ is obtained as the fraction of possible genotypes at the 396 CNV loci called as CNVs:

$$\hat{\tau} = \frac{2118}{2 \times 396 \times 405} = \frac{353}{53460} \approx 0.0066.$$

To estimate χ, concordance of CNV calls was evaluated for five pairs of duplicate samples for which CNV data were obtained on both members of the pair. In each case, one member of the pair was included in the collection of 405 individuals used in other CNV analyses, while the other was excluded from all other CNV analyses. Concordance was evaluated with the same 396 CNV loci used in population-genetic analysis, as it is for this collection of loci that error rates are of interest. Averaging across pairs, the fraction of CNVs called in at least one member of a pair that were called in both members of the pair equaled $\hat{\chi} \approx 0.89$ (Table S11).

Inserting our estimates $\hat{\tau}$ and $\hat{\chi}$ into eqs. 3 and 4, we can plot estimates $\hat{\alpha}$ and $\hat{\beta}$ as functions of the unknown parameter ρ. Considering many loci, the value of ρ represents the true mean frequency of CNVs across the loci under consideration. Although ρ is unknown, previous studies suggest that CNVs tend to have quite low frequencies[36,37]. We consider values of ρ extending from 0 to $\sim 7.5\hat{\tau}$ — that is, from a value at which no CNVs exist to a value at which only a very small fraction of true CNVs are detected at the loci, and the true frequency is 7.5 times the estimated value. Within this range, we find that the false positive rate is easily bounded above by 0.7% (Figure S10A), while relatively little information is available about the false negative rate due to the uncertainty in ρ (Figure S10B). Note further that under the assumption for any useful test that the true positive rate $1 - \beta$ is larger than the false positive rate α, a rearrangement of eq. 1 has the consequence that $\alpha < \tau$. Thus, the fraction of data points with $D = 0$ that are erroneously called as CNVs is bounded above by the overall proportion of data points called as CNVs, or 0.66%. Over most of the range of ρ values considered, the best estimate of the false positive rate is actually equal to 0.

A more conservative estimate of χ that separately averages the numerators (CNVs called in both members of duplicate pairs) and denominators (CNVs called in at least one member of a duplicate pair) rather than averaging the ratios gives greater weight to a single pair of individuals with a large number of CNVs, and produces $\hat{\chi} \approx 0.76$. However, using this estimate in eqs. 3 and 4 leads to results nearly identical to those obtained with the less conservative estimate of $\hat{\chi} \approx 0.89$ (Figure S11).

Thus, we have shown that the intuitive result that a concordance of CNV calls among duplicate samples vastly exceeding the proportion of CNV calls in any single individual implies a low false positive rate, no more than $\sim$0.66%. Because of the greater magnitude of the false negative rate compared to the false positive rate, we can be reasonably certain that for the particular CNV loci in our study, the vast majority of errors are false negatives. This result accords well both with the low false positive rates and higher false negative rates estimated via concordance of replicates both by Wong *et al.*[35] and by subsequent articles based on their data[33,34]. It also matches closely with the validation performed by Wang *et al.*[28] using the same `PennCNV` algorithm employed for identifying CNVs in our study.

5.4 Summary of detected CNVs

We identified 2398 deletion CNVs in 426 individuals (2386 autosomal, 12 X-chromosomal), 2236 single-copy deletions and 81 homozygous deletions; 1928 of these deletions (80.4%) occurred at previously reported CNV loci. The deletions ranged from 2kb to 934kb in size (mean 82.7kb, median 58.5kb). The 2398 deletions occurred in 863 non-overlapping CNV loci. The most common deletion was at a locus on chromosome 6, occurring in 112 individuals, 22 of whom were homozygous; on average each deletion was observed 2.68 times (median 1). Of the 2398 deletions, 1491 (62.2%) were within or across genes.

We identified 1154 duplication CNVs in 402 individuals (1117 autosomal, 37 X-chromosomal) 1084 single-copy duplications and 35 double-copy duplications; 889 of these duplications (77.0%) occurred at previously reported CNV loci. The duplications ranged from 5.6kb to 998kb in size (mean 130.4kb, median 81.1kb). The 1154 duplications occurred in 565 non-overlapping CNV loci. The most common duplication was at a locus on chromosome 10, occurring in 36 individuals, 3 of whom were homozygous; on average each duplication was observed 1.98 times (median 1). Of the 1154 duplications, 791 (68.5%) were within or across genes.

Considering all 1428 CNV loci, the total number of loci that had not been previously reported was 507 (495 autosomal, 12 X-chromosomal). Of the 1428 loci, 49 had at least one individual homozygous for the CNV (47 autosomal, 2 X-chromosomal). In the final autosomal dataset used for population-genetic analysis, which did not contain relatives, 44 CNV loci had this property. Considering all 3552 CNVs, five individuals did not have any CNVs detected (Balochi 74, Balochi 78, Kalash 321, Yi 1183, Mongola 1230). A simple Poisson calculation suggests that it is not entirely unexpected to observe five individuals without CNVs. Most populations have on average 3-7 CNVs per individual, and the Poisson zero class for a distribution with

mean 5 suggests that perhaps 2-3 individuals are expected to have no CNVs in a dataset of 400 individuals. It is noteworthy, however, that one of the individuals with no CNVs was from the Kalash population, whose individuals in general had high numbers of CNVs.

Summaries are shown in Table S12 and Figures S12 and S13 of the number of CNVs identified in individual populations, the frequency spectrum of CNVs in the full collection of 405 HGDP-CEPH individuals, and the frequency spectrum of CNVs by geographic region. Figure S22 provides the distribution of CNVs by length.

5.5 Duplications on the X chromosome in males

In males, the non-pseudoautosomal part of the X chromosome is hemizygous and its SNPs are expected to be genotyped by the HumanHap technology as homozygous. While occasional heterozygous genotypes occur due to genotyping error, long stretches of male X chromosomes genotyped as having many heterozygous SNP genotypes may result from the presence of duplications. Thus, as a second approach for identifying one particular type of CNV, we used male X-chromosomal SNPs to search for duplications. This analysis used data from an intermediate stage in the preparation of the final SNP genotypes, consisting of 493 HGDP-CEPH individuals and 13,203 X-chromosomal SNPs prior to conversion of male heterozygotes to missing data.

We scanned the male X-chromosomal SNP genotypes, counting heterozygous SNPs in 10-SNP sliding windows. We then examined the spatial distribution of heterozygous SNPs, searching for windows with at least four heterozygous SNPs. This approach identified eight individuals each with a region of the genome in which multiple neighboring windows had four or more heterozygous SNPs (Figures S8 and S9).

We compared our list of duplication variants identified from X-chromosomal heterozygosity to the CNV calls based on intensity data. This comparison used the initial set of 443 individuals employed in CNV analysis. Of the 8 duplications detected by X-chromosomal heterozygosity, 7 were also detected from intensities. The eighth duplication occurred in an individual not included in the `PennCNV` analysis, Papuan 545, indicating that all duplications detectable by X-chromosomal heterozygosity were identified by `PennCNV`. The total number of X-chromosomal duplications detected by `PennCNV` from genotype intensity in males was 12 (of which 7 were also detected by X-chromosomal heterozygosity). Note that not all duplications observed from intensity data are detectable using the X-chromosomal heterozygosity method, as duplications too short to produce sufficient stretches of heterozygosity and duplications with genotypically identical copies would not be found.

6 Population-genetic analysis of CNV data

6.1 Geographic distribution

We applied the rarefaction approach[13] to the CNV dataset in the same manner as for the SNP and haplotype datasets. This analysis used the 405 unrelated individuals and the 396 non-singleton autosomal CNV loci. Similarly to the analysis of the haplotype dataset, because the CNV dataset contains no missing data, analysis of geographic distributions was performed by averaging across all 396 CNV loci.

6.2 `Structure`

`Structure` and `CLUMPP` analysis of the CNV data proceeded in a similar manner to the analysis of SNPs and haplotypes. This analysis used the 405 unrelated individuals and the 396 non-singleton CNV loci. For each value of K, 40 replicate `Structure` analyses were performed, and analysis with `CLUMPP` proceeded using the output of these 40 replicates. Because modes with the CNV data were in many cases not clearly defined, `CLUMPP` analysis with the CNV data utilized a lower cutoff of $G' > 0.8$ for identification of modes compared to the higher cutoff of 0.9 used for SNPs and haplotypes. Otherwise, `CLUMPP` analysis proceeded similarly to the SNP and haplotype `Structure` analyses. Further details are provided in Section 7.2.

6.3 Population tree

A neighbor-joining tree of populations was obtained in the same manner as with the SNP data, using the 405 unrelated individuals in the CNV dataset and the 396 autosomal non-singleton CNV loci.

6.4 Multidimensional scaling

The mean allele-sharing distance across loci[38,39], as computed using `microsat`[21], was used as the basis for multidimensional scaling with the CNV data. This analysis used the 405 unrelated individuals in the CNV dataset and the 396 autosomal non-singleton CNV loci to generate a genetic distance matrix between

populations. Multidimensional scaling proceeded in the same manner as with the SNP and haplotype datasets, except that the rotation applied was of magnitude 315° and was followed by a reflection across the vertical axis. The values of the goodness-of-fit statistics for the plot shown in Figure 1d were $\alpha_{1,2} = 33.0\%$ and $\alpha_{2,2} = 60.6\%$. This plot has three outlier populations removed. The plot that retains the outliers is shown in Figure S14; this plot was rotated and reflected in the same manner as the plot without the outliers, and it has $\alpha_{1,2} = 49.4\%$ and $\alpha_{2,2} = 86.5\%$.

7 Additional analysis of multiple data types

7.1 Linkage disequilibrium

To compare the pattern of LD across populations observed using HR^2 applied to the unphased SNP data with the pattern obtained from r^2 and the phased haplotype data, we ranked populations by LD in each of the overlapping 1kb bins in which LD was plotted. We then computed the Spearman correlation of the two lists of ranks, one for HR^2 and one for r^2 (Figure 2d). The two statistics show qualitatively the same pattern of decay of LD (Figures 2b and S4). When mean values of r^2 and HR^2 at equivalent physical distance are plotted, an approximately linear relationship is seen for mean r^2 as a function of mean HR^2 (Figure S23).

Regression coefficients for regression of LD on geographic distance from Africa were compared for regressions that used HR^2 applied to the unphased data and those that used r^2 applied to the phased data. Values of the coefficient of determination (R^2) are shown in Table S13 for various choices of physical distance. Due to the greater spread in values of r^2 (Figure S4) compared to HR^2 (Figure 2b), the regression coefficients with r^2 are greater than with HR^2. For both statistics, although the declining LD with physical distance alters the regression coefficients when physical distance changes, the similar values of R^2 indicate that geographic distance explains similar proportions of variation in LD at various choices of physical distance.

7.2 `Structure`

The `Structure` manual and past applications of `Structure`[7,8] suggest that for large and complex datasets, a sensible use of the method is to examine the behavior of `Structure` at several small values of K, and to then identify additional substructure by applying `Structure` to subsets of individuals. This approach arises from the fact that for datasets containing a large number of populations, `Structure` does not always dissect population structure at finer-scale levels when the full data are used, and such substructure may appear only in analyses of subsets of the data. The hierarchical approach to clustering with `Structure` has been investigated previously[7,40], and we adopted it for use in the current study. Because additional substructure exists that is not detected when using `Structure` with the full dataset, no single value of K provides a full description of the population structure. Therefore, for display in Figure 1c, we chose an approach in which we showed the results on worldwide SNP, haplotype, and CNV datasets for the five smallest nontrivial values of K. Within geographic regions, for which considerably less population structure exists and for which a perspective of "inferring" K is more sensible, we then chose a single value of K for display in Figure 1c. This value was chosen as the value whose most frequent mode had highest mean log likelihood, as described below.

Here we provide a more complete description of the `Structure` results, both those shown in Figure 1c as well as those for values of K not included in Figure 1c. For the worldwide data, Figure S24 displays the most frequently occurring modes for each K from 2 to 10. For each geographic region, Figure S2 displays each value of K from 2 to a value chosen as the smallest value in $\{5, 8\}$ that exceeded the number of predefined populations in the geographic region. The plots in Figure 1c have been extracted from Figures S24 and S2.

For a given dataset and value of K, a clustering mode was chosen for display in Figure S24 or S2 using the following approach. If a single mode appeared most often, then the `CLUMPP` average across replicates producing that mode was displayed. If two or more modes were tied, then the `CLUMPP` average for the mode with highest `CLUMPP` H' score was displayed. In the single instance for which no modes appeared in more than one replicate (Europe, $K = 5$), the highest-likelihood replicate (among 40 total) was displayed.

The numbers of replicates producing the modes in Figure S24 are given in Table S14. For $K = 2$ and $K = 3$, all three datasets had a single mode common to all 40 replicates. Modes appearing in at least 25% of replicates appeared for SNPs with $K \leq 6$, for CNVs with $K \leq 4$, and for all K values for haplotypes. Thus, it appears that one difference between the SNP-based and haplotype-based `Structure` analysis is a greater degree of replicability when using haplotypes.

Within geographic regions, the frequencies of the modes in Figure S2 are given in Table S15. Similarly to the results seen for the worldwide dataset, the most frequent mode appeared with high frequency for low values of K, and with decreasing frequency for higher values. For each region, the most frequent mode for the value of K selected for display in Figure 1c appeared in at least 25% of replicates.

We noted above that a single value of K may not provide a fully informative summary of the `Structure` results with a large and complex worldwide dataset. Figure S25 plots the log likelihoods for the various runs as functions of the number of clusters K. Figure S26 plots the subset of points corresponding to the most frequent mode. These figures illustrate a considerable degree of variation in log likelihood, both for individual K values as well as across values of K. Considering only the most frequent mode, the mean log likelihood is shown in Table S16. If highest mean log likelihood for the most frequent mode had been used as a criterion for selecting a single value of K, then $K = 6$ would have been selected for SNPs and $K = 5$ would have been selected for haplotypes. For CNVs, the likelihood plot does not have a clear peak, a situation described by the `Structure` manual as a case in which it is sensible to focus on smaller values of K that capture "most" of the structure in the data.

Similar likelihood plots for individual geographic regions appear in Figures S27 and S28, and the mean log likelihoods for the most frequent mode are summarized in Table S17 for each geographic region. Table S17 provides the basis for selecting the value of K for display in Figure 1c.

The collection of plots in Figure S24 illustrates that to a large extent, the SNP results match previous inferences based on microsatellites[7], except that the separations of Oceania, the Americas, and the Kalash population occurred in a different sequence, and the yellow cluster was spread more broadly across Central/South Asia rather than corresponding exclusively to the Kalash population. The main differences in the haplotype analysis were a separation of African hunter-gatherers from other Africans, which occurred at multiple values of K, rather than a separation of the Native Americans, which did not occur until quite a high value of K. For some small values of K, the CNV plots are similar to the corresponding SNP and haplotype plots with one fewer cluster; to some extent, with high values of K, the CNV plots identify the clusters corresponding to African hunter-gatherers, Native Americans, and populations from Oceania.

In some geographic regions, additional substructure is found beyond that reported at the high-likelihood value of K shown in Figure 1c. In Africa with $K = 6$, nearly all populations are somewhat separable, including the Mandenka and Yoruba populations, who had clustered together in previous analysis[7]. In Europe with $K = 3$, the three populations form distinct clusters; similarly, the four populations in Central/South Asia form distinct clusters when $K = 4$.

7.3 Population tree

The CNV population tree contained a surprising grouping of the Kalash, Melanesian, and Papuan populations (Figure 1b). These populations were also among the groups with the greatest numbers of CNVs (Figure 4b).

To understand how the large numbers of CNVs give rise to a Kalash-Melanesian-Papuan grouping, we can consider the effect of the number of CNVs on the allele-sharing distance and the neighbor-joining algorithm. The first observation we can make is that populations with large numbers of CNVs tend to have high allele-sharing genetic distances with all other populations (Table S1). Recall that genotypes are coded as 00 if no CNV was observed, 01 if a heterozygous deletion or duplication was observed, and 11 if a homozygous deletion or duplication was observed. For two populations with few CNVs, nearly all genotypic comparisons of one individual from one population and one individual from the other population involve two individuals with genotype 00. Such comparisons of 00 genotypes produce zero genetic distance, and consequently, allele-sharing genetic distances between pairs of populations with relatively few CNVs are quite small. However, if at least one of the two populations has a large number of CNVs, then the number of comparisons involving a 01 genotype from that population and a 00 genotype from the other population will be higher, producing a higher overall pairwise genetic distance. Thus, the greater numbers of CNVs in the Kalash, Melanesian, and Papuan populations can explain why genetic distances involving these populations are relatively high.

We can now consider the effect of these high genetic distances on tree construction by using a simple example that mimics the distance matrix in Table S1. Consider two types of populations, "A" populations and "B" populations. Suppose there are n_A populations of type A and n_B populations of type B, with $n_A > n_B \geq 2$. Suppose further that the genetic distance between two A populations is d_A, and the genetic distance between a B population and any other population is d_B, with $d_B > d_A$. This scenario approximates the matrix in Table S1, with Kalash, Melanesian, and Papuan being of type B, and all other populations being of type A. The fact that nearly all genetic distances to the Kalash, Melanesian, or Papuan populations are greater than nearly all distances among other populations suggests that it is reasonable to approximate that distances to the Kalash, Melanesian, and Papuan populations equal d_B, while other distances equal d_A.

What occurs during the first agglomerative step in tree construction using the neighbor-joining algorithm? Three possibilities exist: either two A populations can group together, two B populations can group together, or an A and a B populations can group together. Following the notation of Felsenstein[41] (p. 167), for an A

population, the normalized sum of the entries of its row of the symmetrized genetic distance matrix is

$$u_A = \frac{(n_A - 1)d_A + n_B d_B}{n_A + n_B - 2}.$$

Similarly, the corresponding quantity for a B population is

$$u_B = \frac{(n_A + n_B - 1)d_B}{n_A + n_B - 2}.$$

The pair of populations that are grouped together in the first agglomerative step of the neighbor-joining algorithm is the pair (i, j) that minimizes $C_{ij} = D_{ij} - u_i - u_j$, where D_{ij} represents the distance between the populations. For a pair of A populations, we have $D_{AA} = d_A$; for a pair of B populations, $D_{BB} = d_B$; for an A population and a B population, $D_{AB} = d_B$. We therefore have

$$\begin{aligned} C_{AA} &= \frac{(-n_A + n_B)d_A - 2n_B d_B}{n_A + n_B - 2} \\ C_{BB} &= \frac{(-n_A - n_B)d_B}{n_A + n_B - 2} \\ C_{AB} &= \frac{(-n_A + 1)d_A - (n_B + 1)d_B}{n_A + n_B - 2}. \end{aligned}$$

Using the inequalities $n_A > n_B \geq 2$ and $d_B > d_A$, it can be shown that $C_{BB} < C_{AA} < C_{AB}$. In particular, because C_{BB} is the smallest of the three values, the first two populations to be joined by neighbor-joining will be two of the most divergent, distinctive populations from the B class. This grouping is a form of "long-branch attraction" common to tree inference algorithms, in which taxa that are truly distant from each other and from other taxa unexpectedly group together[41]; similar computations to the example above have previously identified long-branch attraction with neighbor-joining[42].

Table S2 describes the sequences of agglomeration of populations into the neighbor-joining trees for the 1000 bootstrap replicates that underlie the CNV tree in Figure 1b. We can observe that these sequences reflect the predictions based on the example matrix. For most replicates, Kalash and Papuan — two groups from the B class in the example — are the first two taxa to agglomerate; the next agglomeration usually combines the Kalash-Papuan grouping with the Melanesian population (the last B population). Subsequently, as predicted by the fact that $C_{AA} < C_{AB}$, pairs from the A class begin agglomerating before A and B populations group together — first Pima and Maya in most replicates, often followed by Yoruba and Mandenka.

In conclusion, this analysis of genetic distance and the neighbor-joining algorithm shows that the clustering of the Kalash, Melanesian, and Papuan populations is a consequence primarily of the high numbers of CNVs detected in these populations, and should not be taken as evidence of a meaningful biological grouping. Indeed the genetic distances of Kalash to Melanesian and Papuan lie in the range of the distances to other populations (Table S1), and they do not suggest a noteworthy similarity of the Kalash population and the populations from Oceania. Thus, the unusual features of the CNV neighbor-joining tree support the strategy of relying on multiple statistical approaches in the investigation of population structure.

7.4 Multidimensional scaling

To investigate the results of multidimensional scaling within geographic regions, we analyzed submatrices of the pairwise distance matrix, restricting attention to pairs of individuals from the same region (again rotating coordinates by 225°). These analyses were performed both for SNPs and for haplotypes (Figure S3).

As was true in the full worldwide analysis (Figure 1d), analyses based on SNP and haplotype datasets produced highly concordant results. Both for SNPs and for haplotypes, for almost all populations, individuals separated in the MDS plot from individuals belonging to other populations, often to a greater extent than was observed for `Structure` in Figure S2. In the Middle East, the Bedouin, Druze, and Mozabite populations were largely separable, but they overlapped to some extent with Palestinians. In East Asia, Mongola and Daur were placed in partially overlapping regions of the plot. In general, however, because individual clusters in the within-region MDS plots correspond largely to the individuals of distinct populations, the plots indicate an ability to separate individuals even from closely related groups from the same geographic region.

7.5 Genetic and geographic distance

We analyzed the relationship of genetic and geographic distance for autosomal haplotypes and CNVs using the same approach as was used for SNPs. Results for haplotypes were averaged across the ten haplotype cluster datasets. The non-singleton autosomal CNV data were used (396 CNV loci, 405 individuals).

Considering pairs of populations in which at least one of the populations was from Africa, a relatively linear relationship of genetic and geographic distance was observed (Figure S29). A clear pattern of linear increase of genetic with geographic distance was visible for SNP and haplotype datasets, and to a lesser extent for CNVs. Although the qualitative patterns are somewhat similar, the difference in scale for the three datasets is likely due to the combination of differences in allele frequency spectra for the three datasets together with a dependence of F_{ST} values on the properties of allele frequency distributions[12,43,44,45].

8 Comparative analysis of equal-sized SNP and CNV datasets

8.1 Reduced SNP datasets

A potential explanation for different population-genetic results obtained with the SNP and CNV datasets is the differing size and information content of these datasets. To investigate whether results based on CNVs were analogous to what would be obtained from SNP datasets of similar size and information content, we constructed five subsets of the SNP dataset with the same size as the CNV dataset. For each SNP subset, we considered the same 405 unrelated individuals that were included in the population-genetic analysis of CNVs, and we selected 396 autosomal SNPs without replacement from the version of the full SNP dataset with all missing genotypes imputed (the "best guess" estimate of Section 3.1). Autosomal SNPs were chosen so that in the collection of 405 individuals, each SNP subset had the same frequency spectrum (Figure S30). This frequency spectrum was matched as closely as possible to the frequency spectrum for non-singleton CNV loci (Figure S12). Because fewer SNPs compared to CNV loci were studied in the 2/810 minor allele frequency class, it was not possible to match frequency spectra exactly. All five SNP subsets contained the same 120 SNPs in the 2/810 class, supplemented by a number of additional SNPs from the 3/810 class so that the sum of the numbers of loci in the 2/810 and 3/810 classes was the same for SNPs and CNVs.

8.2 `Structure`

We compared the population structure estimated using the SNP subsets to the corresponding population structure estimates based on the CNV dataset. Population structure analysis for the SNP subsets proceeded in a similar manner as for the CNV dataset. For each of the five SNP subsets, eight replicate runs of `Structure` were performed using each choice of K from 2 to 8. The 40 runs with a given value of K were then considered jointly when using `CLUMPP`[18] to find the most replicable mode ($G' > 0.8$ threshold). For each K, this mode was chosen for display in Figure S15. The plots in Figure S15 show a considerably greater degree of similarity to the plots of CNV population structure in Figure S24 than to the plots with the full SNP dataset in Figure S24, suggesting that the differing size of the full SNP and CNV datasets is largely responsible for the difference in SNP and CNV results in Figures S24 and 1c.

8.3 Population tree

We compared the CNV tree to corresponding trees based on the five subsets of the SNP dataset with the same size as the CNV dataset. All five trees produced groupings that largely matched geographic regions, and our comparison focused on groupings obtained within the regions (Tables S3 and S4). The CNV tree produced groupings that overlapped with those obtained using the SNP datasets, but some disagreement was apparent across the six trees. As the SNP analysis with the full collection of markers produced a tree with very high bootstrap support (1000 of 1000 bootstraps on all except one branch), the uncertainty apparent when using only 396 markers suggests that the difference of the CNV tree from the SNP tree based on the full dataset may be attributable to the smaller size of the dataset rather than to intrinsic properties of CNVs.

8.4 Multidimensional scaling

We compared the CNV multidimensional scaling plot to corresponding plots based on the five subsets of the SNP dataset with the same size as the CNV. MDS was applied to each SNP dataset in the same manner as for the CNV data. Figure S16 shows the MDS plot based on the SNP dataset of the five for which two-dimensional MDS produced the highest value of the two α statistics ($\alpha_{1,2} = 56.9\%$ and $\alpha_{2,2} = 90.1\%$). We produced a second plot of this dataset, with two outlier populations removed ($\alpha_{1,2} = 34.0\%$, $\alpha_{2,2} = 57.7\%$, Figure S16B). The plots were rotated counterclockwise by 270° for part A and 90° for part B. Although some geographic clustering is observed, likely as a result of the relatively small size of the SNP subsets, these plots illustrate a considerable difference from the MDS plot produced with the full SNP data (Figure 1d).

9 Supplementary tables and figures

Tables S1-S4 and Figures S1-S16 are cited from the main text and are numbered following the order of these citations. Additional tables and figures not cited from the main text then follow, in order of citation in the supplementary material above. The following lists give the locations where the supplementary tables and figures are first mentioned.

Table or figure	Brief description	Section where first mentioned	Page number for table or figure
	Tables cited in main text		
Table S1	Allele-sharing distance matrix for CNV data	7.3	20
Table S2	Stages at which populations are agglomerated into the CNV tree	7.3	21
Table S3	Bootstrap support for CNV tree and reduced-data SNP trees (part I)	8.3	22
Table S4	Bootstrap support for CNV tree and reduced-data SNP trees (part II)	8.3	23
	Tables cited only in supplement		
Table S5	Individuals included in the study	1.4	24
Table S6	Coordinates used in geographic analyses	1.5	25
Table S7	Distribution of SNPs by chromosome	1.6	26
Table S8	Distribution of the missing data rate across individuals	1.7	27
Table S9	Distribution of the missing data rate across SNPs	1.7	27
Table S10	Correlation coefficients of SNP allele frequencies	2.1	28
Table S11	Concordance of CNV calls in duplicate samples	5.3	29
Table S12	Summary of CNVs detected	5.4	30
Table S13	Regression of linkage disequilibrium on distance from Africa	7.1	31
Table S14	Number of replicates in the most frequent clustering mode (worldwide)	7.2	32
Table S15	Number of replicates in the most frequent clustering mode (regions)	7.2	32
Table S16	Mean log likelihood for replicates in the most frequent mode (worldwide)	7.2	33
Table S17	Mean log likelihood for replicates in the most frequent mode (regions)	7.2	33
	Figures cited in main text		
Figure S1	Map of population locations	1.5	34
Figure S2	Inferred population structure within individual geographic regions	7.2	35
Figure S3	Multidimensional scaling within geographic regions	7.4	36
Figure S4	Linkage disequilibrium measured by r^2 from phased haplotypes	4.1	37
Figure S5	Joint distribution of haplotype length and frequency	4.2	38
Figure S6	Haplotype cluster frequencies for a "typical" genomic region	4.4	39
Figure S7	Standardized homozygosity for two genomic regions	4.4	40
Figure S8	X-chromosomal heterozygosity and duplications in males	5.5	41
Figure S9	X-chromosomal heterozygosity and duplications in males (magnified)	5.5	42
Figure S10	CNV detection error rates (less conservative duplication concordance)	5.3	43
Figure S11	CNV detection error rates (more conservative duplication concordance)	5.3	44
Figure S12	Allele frequency spectrum for CNVs (worldwide)	5.4	45
Figure S13	Allele frequency spectra for CNVs (regions)	5.4	46
Figure S14	Multidimensional scaling for CNVs including outliers	6.4	47
Figure S15	Inferred structure for SNP datasets matched to the CNV data	8.2	48
Figure S16	Multidimensional scaling for SNP datasets matched to the CNV data	8.4	49
	Figures cited only in supplement		
Figure S17	Flow chart of genotyping and quality control	1.1	50
Figure S18	Quality control for HumanHap550 version 1 BeadChips	1.3	51
Figure S19	Quality control for HumanHap550 version 3 BeadChips	1.3	52
Figure S20	Hardy-Weinberg test statistics for SNPs	1.6	53
Figure S21	Allele frequency spectra for SNPs	2.1	54
Figure S22	Length distribution of CNVs	5.4	55
Figure S23	Comparison of LD based on phased and unphased data	7.1	56
Figure S24	Inferred population structure (worldwide)	7.2	57
Figure S25	Likelihood for `Structure` runs (worldwide)	7.2	58
Figure S26	Likelihood for `Structure` runs in the most frequent mode (worldwide)	7.2	59
Figure S27	Likelihood for `Structure` runs (regions)	7.2	60
Figure S28	Likelihood for `Structure` runs in the most frequent mode (regions)	7.2	61
Figure S29	Pairwise genetic distance as a function of geographic distance	7.5	62
Figure S30	Allele frequency spectra for SNP datasets matched to the CNV data	8.1	63

	Population	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	San		.008	.009	.008	.007	.008	.007	.008	.008	.009	.010	.010	.012	.009	.008	.025	.009	.007	.010	.008	.008	.006	.007	.009	.017	.021	.011	.013	.008
2	Mbuti Pygmy	.008		.009	.009	.007	.009	.009	.010	.010	.011	.011	.010	.013	.010	.008	.027	.011	.009	.011	.009	.009	.007	.008	.010	.018	.023	.012	.015	.009
3	Biaka Pygmy	.009	.009		.011	.009	.009	.009	.010	.010	.011	.012	.012	.014	.011	.009	.029	.012	.010	.012	.010	.010	.008	.009	.011	.019	.024	.014	.016	.011
4	Bantu (Kenya)	.008	.009	.011		.007	.008	.008	.010	.010	.011	.010	.011	.013	.010	.009	.026	.011	.009	.011	.010	.010	.008	.009	.010	.017	.022	.013	.014	.010
5	Bantu (S. Africa)	.007	.007	.009	.007		.006	.006	.008	.009	.010	.010	.010	.013	.009	.008	.027	.009	.007	.010	.009	.008	.007	.007	.009	.016	.022	.011	.014	.009
6	Yoruba	.008	.009	.009	.008	.006		.005	.008	.008	.009	.010	.010	.013	.009	.008	.027	.010	.008	.010	.009	.009	.007	.008	.010	.017	.022	.012	.014	.010
7	Mandenka	.007	.009	.009	.008	.006	.005		.008	.009	.009	.010	.010	.012	.009	.007	.026	.010	.008	.010	.008	.008	.007	.007	.009	.017	.022	.012	.013	.009
8	Mozabite	.008	.010	.010	.010	.008	.008	.008		.006	.007	.008	.008	.011	.007	.007	.026	.010	.008	.010	.008	.009	.006	.008	.009	.016	.022	.012	.014	.010
9	Bedouin	.008	.010	.010	.010	.009	.008	.009	.006		.007	.006	.008	.010	.007	.006	.026	.009	.008	.010	.008	.008	.006	.008	.009	.016	.022	.011	.013	.010
10	Palestinian	.009	.011	.011	.011	.010	.009	.009	.007	.007		.007	.009	.011	.007	.006	.026	.010	.009	.011	.008	.010	.007	.008	.010	.016	.022	.013	.014	.010
11	Druze	.010	.011	.012	.010	.010	.010	.010	.008	.006	.007		.009	.011	.007	.007	.026	.010	.009	.011	.008	.009	.007	.008	.009	.016	.022	.012	.014	.010
12	Basque	.010	.010	.012	.011	.010	.010	.010	.008	.008	.009	.009		.011	.007	.007	.027	.008	.008	.010	.007	.008	.007	.007	.009	.016	.023	.012	.015	.010
13	Russian	.012	.013	.014	.013	.013	.013	.012	.011	.010	.011	.011	.011		.009	.010	.026	.013	.011	.011	.012	.012	.010	.011	.013	.019	.023	.015	.013	.013
14	Adygei	.009	.010	.011	.010	.009	.009	.009	.007	.007	.007	.007	.007	.009		.007	.023	.010	.008	.008	.008	.008	.008	.006	.007	.009	.015	.021	.011	.012	.010
15	Balochi	.008	.008	.009	.009	.008	.008	.007	.007	.006	.006	.007	.007	.010	.007		.025	.008	.007	.009	.007	.008	.005	.006	.008	.016	.021	.011	.013	.009
16	Kalash	.025	.027	.029	.026	.027	.027	.026	.026	.026	.026	.026	.027	.026	.023	.025		.027	.026	.023	.026	.027	.024	.025	.027	.027	.023	.029	.030	.025
17	Burusho	.009	.011	.012	.011	.009	.010	.010	.010	.009	.010	.010	.008	.013	.010	.008	.027		.008	.011	.009	.009	.008	.008	.010	.017	.023	.012	.014	.010
18	Uygur	.007	.009	.010	.009	.007	.008	.008	.008	.008	.009	.009	.008	.011	.008	.007	.026	.008		.009	.007	.007	.005	.005	.007	.016	.021	.011	.013	.007
19	Yakut	.010	.011	.012	.011	.010	.010	.010	.010	.010	.011	.011	.010	.011	.008	.009	.023	.011	.009		.009	.008	.007	.008	.009	.016	.019	.012	.013	.010
20	Mongola	.008	.009	.010	.010	.009	.009	.008	.008	.008	.008	.008	.007	.012	.008	.007	.026	.009	.007	.009		.006	.005	.006	.007	.015	.021	.010	.012	.008
21	Daur	.008	.009	.010	.010	.008	.009	.008	.009	.008	.010	.009	.008	.012	.008	.008	.027	.009	.007	.008	.006		.005	.006	.007	.016	.021	.010	.013	.009
22	Yi	.006	.007	.008	.008	.007	.007	.007	.006	.006	.007	.007	.007	.010	.006	.005	.024	.008	.005	.007	.005	.005		.003	.004	.014	.020	.009	.011	.007
23	Cambodian	.007	.008	.009	.009	.007	.008	.007	.008	.008	.008	.008	.007	.011	.007	.006	.025	.008	.005	.008	.006	.006	.003		.005	.015	.020	.010	.012	.007
24	Lahu	.009	.010	.011	.010	.009	.010	.009	.009	.009	.010	.009	.009	.013	.009	.008	.027	.010	.007	.009	.007	.007	.004	.005		.016	.021	.011	.013	.010
25	Melanesian	.017	.018	.019	.017	.016	.017	.017	.016	.016	.016	.016	.016	.019	.015	.016	.027	.017	.016	.016	.015	.016	.014	.015	.016		.017	.018	.020	.017
26	Papuan	.021	.023	.024	.022	.022	.022	.022	.022	.022	.022	.022	.023	.023	.021	.021	.023	.023	.021	.019	.021	.021	.020	.020	.021	.017		.024	.025	.022
27	Pima	.011	.012	.014	.013	.011	.012	.012	.012	.011	.013	.012	.012	.015	.011	.011	.029	.012	.011	.012	.010	.010	.009	.010	.011	.018	.024		.011	.010
28	Maya	.013	.015	.016	.014	.014	.014	.013	.014	.013	.014	.014	.015	.013	.012	.013	.030	.014	.013	.013	.012	.013	.011	.012	.013	.020	.025	.011		.013
29	Colombian	.008	.009	.011	.010	.009	.010	.009	.010	.010	.010	.010	.010	.013	.010	.009	.025	.010	.007	.010	.008	.009	.007	.007	.010	.017	.022	.010	.013	

Table S1: Allele-sharing distance matrix for CNV data, based on 396 copy-number-variable loci in 405 individuals. For convenience, entries in the table that exceed 0.015 are highlighted in red. As is described in Section 7.3, in a matrix of this structure, populations that have large distances to other populations tend to be agglomerated first by the neighbor-joining algorithm. This matrix underlies the CNV multidimensional scaling plots in Figures 1d and S14.

	Stage of agglomeration into neighbor-joining tree																										
Population	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
San	0	0	0	4	10	19	14	41	61	68	47	61	62	72	80	59	78	67	57	55	42	29	26	19	7	14	8
Mbuti Pygmy	0	0	15	54	77	102	84	75	71	70	61	30	55	46	50	35	31	33	25	23	14	15	11	9	6	4	4
Biaka Pygmy	0	0	13	39	49	82	81	78	78	65	59	48	68	59	64	36	46	34	25	26	8	10	10	10	5	1	6
Bantu (Kenya)	0	0	37	105	130	103	94	89	75	68	53	49	39	34	20	12	18	15	14	10	9	6	5	2	4	5	4
Bantu (S. Africa)	0	1	50	163	207	149	97	77	40	46	41	26	31	20	10	11	8	1	3	2	4	3	2	0	4	0	4
Yoruba	0	1	114	288	188	109	71	46	33	32	38	26	12	14	6	6	4	3	2	0	0	2	1	1	0	2	1
Mandenka	0	2	104	275	208	109	62	42	43	35	35	23	16	14	11	4	5	3	3	1	1	1	0	0	2	0	1
Mozabite	0	0	7	32	43	43	52	51	37	51	61	46	42	47	55	32	48	40	44	45	35	37	33	35	36	20	28
Bedouin	0	0	11	64	59	62	77	77	64	66	66	60	54	44	46	34	32	33	25	28	22	15	16	19	11	5	10
Palestinian	0	1	3	26	40	46	82	65	77	84	83	79	63	48	47	41	35	29	34	17	14	10	13	9	14	16	24
Druze	0	1	6	55	55	66	79	64	92	68	61	53	61	42	48	42	29	28	20	18	15	18	15	19	12	14	19
Basque	0	0	10	31	34	25	48	45	59	54	59	70	70	72	73	78	45	61	31	39	17	25	23	13	9	5	4
Russian	0	1	88	50	59	90	73	84	69	76	66	57	44	41	32	25	24	22	22	17	10	15	9	7	5	8	6
Adygei	0	0	6	12	22	44	47	73	76	74	67	71	70	55	42	37	52	29	34	35	37	44	22	21	13	7	10
Balochi	0	0	0	2	5	3	22	21	24	39	31	49	55	61	87	79	97	76	63	70	45	35	30	28	23	29	26
Kalash	841	156	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Burusho	0	0	10	30	31	25	44	45	50	44	47	53	51	55	59	62	44	51	51	53	46	26	33	30	25	20	15
Uygur	0	0	0	0	0	2	2	5	4	12	23	31	45	31	48	47	50	49	65	69	71	77	85	81	63	49	91
Yakut	0	4	24	145	118	122	95	98	78	59	51	45	37	25	19	23	13	12	9	8	4	2	2	4	1	0	2
Mongola	0	0	1	7	11	13	28	24	37	40	43	43	43	60	62	58	66	58	54	64	64	73	48	31	27	14	31
Daur	0	0	3	9	20	16	34	39	42	48	53	44	47	54	44	49	51	57	48	51	68	51	48	35	25	33	31
Yi	0	0	7	9	13	39	34	35	41	29	53	48	52	59	53	57	79	51	45	59	61	54	25	26	15	30	26
Cambodian	0	0	13	51	61	63	60	52	67	51	55	51	61	48	36	54	42	42	37	31	36	33	16	9	13	7	11
Lahu	0	0	21	58	72	87	79	69	77	56	66	58	48	48	34	32	41	35	25	29	20	15	8	6	4	4	8
Melanesian	159	773	59	6	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Papuan	1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pima	0	63	617	85	50	29	28	16	9	12	6	15	8	3	2	13	8	5	9	7	1	5	2	3	1	2	1
Maya	0	64	693	96	58	33	17	8	6	4	3	10	1	0	0	1	2	1	2	0	0	1	0	0	0	0	0
Colombian	0	0	0	64	40	45	51	48	37	41	50	46	40	48	47	65	41	38	41	49	44	45	38	33	18	15	16

Table S2: Distribution of the stage at which populations are agglomerated into the CNV neighbor-joining tree. Each of the 1000 bootstrap neighbor-joining trees is produced by sequentially incorporating additional populations as the tree is being constructed. During the ith stage of this agglomeration procedure, where i ranges from 1 to three less than the number of populations, two branches, each corresponding to a population or to a grouping of populations, are joined; three branches are then joined in a final stage. In the table, as described in Section 7.3, it can be observed for example that the populations with the largest numbers of CNVs (Kalash, Melanesian, Papuan) usually agglomerated in the first stages. The Papuan population entered the tree at the first stage in each of the 1000 bootstrap replicates, and the Kalash and Melanesian populations usually entered the tree at either the first or second stage. For convenience, for each population, the stage at which the population most frequently entered the tree is highlighted in red (a tie occurred for the Bedouin population).

21

		CNV tree	Reduced SNP tree 1	Reduced SNP tree 2	Reduced SNP tree 3	Reduced SNP tree 4	Reduced SNP tree 5
Africa	.....**	792	627		487	951	749
	....**.			471			
	....***						467
	...*..*			333			
	...*.**		821		229		
	...**..	455				511	
	...****	595	700	792	181	840	206
	..*****		684	356		520	
	.**....	518			446		324
	.******	307	578		862	886	637
	*.*****			691			
	*******	732	496	476	790	711	616
Eurasia	....|...|.**.				492		303
	....|.**|....	304					
	....|*..|..*.	629					
	....|**.|....		883	978	826	1000	999
	....|***|....		1000	995	1000	986	971
	....|***|...*				391		
	....|***|*...		437	528		344	
	....|***|*..*				144		
	..**|...|....	306				262	404
	.*.*|...|....		390	441	348		
	.***|...|....			327	280		215
	.***|***|*..*				115		
	.***|***|****				169		
	*...|***|....						519
	..|...|....		611				
	**..|...|....	422				306	
	****|...|....	264	539	274		239	
	****|.**|....	271					
	****|.**|*...	256					
	****|***|....						368
	****|***|*...		554	264		700	903
	****|***|*..*						602
	****|***|*.*.	247					
	****|***|**..		415	591		245	
	****|***|**.*			603			
	****|***|***.		672			334	
	****|***|****				474	359	

Table S3: Bootstrap support for within-region groupings in Africa and Eurasia, for the CNV tree in Figure 1b and for five SNP trees constructed with datasets of the same size as the CNV dataset. The total number of bootstrap replicates was 1000 for each tree. A sequence of dots and asterisks indicates a particular grouping. The meaning of positions in the sequence follows the order of populations within geographic regions in Figure 1c. Africa — San, Mbuti Pygmy, Biaka Pygmy, Bantu (Kenya), Bantu (Southern Africa), Yoruba, Mandenka; Eurasia — Mozabite, Bedouin, Palestinian, Druze, Basque, Russian, Adygei, Balochi, Kalash, Burusho, Uygur. Thus, for example, in Africa, ...** corresponds to a grouping of Yoruba and Mandenka. The Middle East, Europe, and Central/South Asia are separated into lists of symbols for each subregion. The table illustrates that the population tree based on the CNV dataset has a similar degree of uncertainty to trees based on SNP datasets of the same size.

		CNV tree	Reduced SNP tree 1	Reduced SNP tree 2	Reduced SNP tree 3	Reduced SNP tree 4	Reduced SNP tree 5
East Asia	....**	502					
	...*.*				226		
	...**.		367	291		427	
	...***	534			183		
	..*.*.						447
	.*...*		339	368			
	.*.*..						498
	.**...	390			425		
	.**..*		223	204			
	.*****	309	180	82			
	....					223	206
	**...*					123	
	**.*.*						147
	***...				209		
	******				184		384
Oceania	**		572		647	625	404
America	.**			277			368
	.		685		396	745	
	**.	885					
	***	490	620	134	684	685	

Table S4: Bootstrap support for within-region groupings in East Asia, Oceania, and the Americas, for the CNV tree in Figure 1b and for five SNP trees constructed with datasets of the same size as the CNV dataset. The total number of bootstrap replicates was 1000 for each tree. A sequence of dots and asterisks indicates a particular grouping. The meaning of positions in the sequence follows the order of populations within geographic regions in Figure 1c. East Asia — Yakut, Mongola, Daur, Yi, Cambodian, Lahu; Oceania — Melanesian, Papuan; America — Pima, Maya, Colombian. Thus, for example, in East Asia, `....**` corresponds to a grouping of Cambodian and Lahu. The table illustrates that the population tree based on the CNV dataset has a similar degree of uncertainty to trees based on SNP datasets of the same size.

Geographic region	Population	Number of distinct individuals typed	Number of males included among distinct individuals typed	Number of distinct unrelated individuals	Number of males included among distinct unrelated individuals
AFRICA	San	7	7	6	6
	Mbuti Pygmy	15	13	13	11
	Biaka Pygmy	32	30	23	23
	Bantu (Southern Africa)	8	8	8	8
	Bantu (Kenya)	12	11	11	10
	Yoruba	25	13	22	12
	YRI HapMap	36	22	24	12
	Mandenka	24	16	22	15
MIDDLE EAST	Mozabite	30	20	29	20
	Bedouin	47	28	45	27
	Palestinian	26	6	24	6
	Druze	43	13	38	11
EUROPE	Basque	13	9	13	9
	CEU HapMap	48	23	32	16
	Russian	13	10	13	10
	Adygei	14	6	14	6
C/S ASIA	Balochi	15	15	15	15
	Kalash	18	14	16	12
	Burusho	7	6	7	6
	Uygur	10	8	10	8
EAST ASIA	Yakut	15	12	15	12
	Daur	10	7	10	7
	Mongola	9	7	9	7
	JPT HapMap	16	7	16	7
	CHB HapMap	12	4	12	4
	Yi	10	9	10	9
	Lahu	8	7	8	7
	Cambodian	10	6	10	6
OCEANIA	Melanesian	17	6	11	4
	Papuan	16	12	16	12
AMERICA	Pima	11	6	8	4
	Maya	13	2	10	1
	Colombian	7	2	7	2

Table S5: Individuals included in the study. Individuals with a relationship more distant than second-degree (avuncular, half sib, or grandparent/grandchild) were treated as unrelated. YRI, CEU, JPT, and CHB respectively refer to Yoruba, European American, Japanese, and Chinese samples from the HapMap.

Geographic region	Population	Latitude	Longitude
AFRICA	San	-21	20
	Mbuti Pygmy	1	29
	Biaka Pygmy	4	17
	Bantu (Southern Africa)	-25.56926433	24.25
	Bantu (Kenya)	-3	37
	Yoruba	7.995094727	5
	Mandenka	12	-12
MIDDLE EAST	Mozabite	32	3
	Bedouin	31	35
	Palestinian	32	35
	Druze	32	35
EUROPE	Basque	43	0
	Russian	61	40
	Adygei	44	39
C/S ASIA	Balochi	30.49871492	66.5
	Kalash	35.99366014	71.5
	Burusho	36.49838568	74
	Uygur	44	81
EAST ASIA	Yakut	62.98287845	129.5
	Daur	48.49753416	124
	Mongola	45	111
	Yi	28	103
	Lahu	22	100
	Cambodian	12	105
OCEANIA	Melanesian	-6	155
	Papuan	-4	143
AMERICA	Pima	29	-108
	Maya	19	-91
	Colombian	3	-68

Table S6: Coordinates used in geographic analyses. Latitudes in the northern hemisphere are listed with positive values, as are longitudes in the eastern hemisphere.

Chromosome	Number of SNPs genotyped	Number of SNPs discarded in quality checks	Number of SNPs included in final data analysis
1	39,676	21	39,655
2	42,605	17	42,588
3	35,409	10	35,399
4	31,319	19	31,300
5	32,493	14	32,479
6	34,250	18	34,232
7	28,064	14	28,050
8	29,901	19	29,882
9	25,150	7	25,143
10	27,392	7	27,385
11	25,625	19	25,606
12	25,332	12	25,320
13	19,517	15	19,502
14	17,338	11	17,327
15	15,680	5	15,675
16	15,977	10	15,967
17	13,570	9	13,561
18	15,882	4	15,878
19	8,858	1	8,857
20	13,466	3	13,463
21	7,734	4	7,730
22	7,770	7	7,763
Autosomal total	513,008	246	512,762
X	13,203	151	13,052
Y	10	1	9
XY	15	0	15
M	163	91	72
Total	526,399	489	525,910

Table S7: Distribution of SNPs by chromosome. XY refers to the pseudoautosomal region on the X and Y chromosomes, and M refers to the mitochondrial genome.

Missing data rate (%)	Number of individuals
< 0.5	574
$[0.5, 1)$	9
$[1, 1.5)$	4
$[1.5, 2)$	2
$[2, 2.5)$	1
$[2.5, 3)$	6
$[3, 3.5)$	1

Table S8: Distribution of the missing data rate across 597 HGDP-CEPH and HapMap individuals, computed from the final set of 525,910 SNPs, and binned in intervals of 0.5%.

Missing data rate (%)	Number of SNPs
< 1	514,665
$[1, 2)$	10,160
$[2, 3)$	779
$[3, 4)$	157
$[4, 5)$	65
$[5, 6)$	30
$[6, 7)$	31
$[7, 8)$	10
$[8, 9)$	11
$[9, 10)$	2

Table S9: Distribution of the missing data rate across 525,910 SNPs, computed from the final set of 597 HGDP-CEPH and HapMap individuals, and binned in intervals of 1%.

	Africa	Middle East	Europe	C/S Asia	East Asia	Oceania
Middle East	0.770					
Europe	0.706	0.962				
C/S Asia	0.731	0.947	0.949			
East Asia	0.686	0.833	0.841	0.891		
Oceania	0.647	0.755	0.751	0.792	0.809	
America	0.623	0.772	0.788	0.822	0.858	0.719

Table S10: Pearson correlation coefficients of allele frequencies for 512,762 autosomal SNPs. This analysis uses both alleles at each locus and is based on 443 unrelated HGDP-CEPH individuals. The table reflects the correlations of allele frequencies visible in Figure S21.

Population	Pair of individuals	Number of CNVs called in at least one member of pair	Number of CNVs called in both members of pair	Fraction of CNVs called in both members of pair
Bedouin	650, 652	3	3	1.000
Biaka Pygmy	452, 1087	6	6	1.000
Biaka Pygmy	457, 1092	4	3	0.750
Biaka Pygmy	472, 981	5	5	1.000
Melanesian	657, 826	49	34	0.694

Table S11: Concordance of CNV calls in five pairs of duplicate samples for which CNV data were obtained in both members of the pair. The computation is based on the 396 autosomal CNV loci used in the population-genetic analysis. The average concordance across pairs is $871/980 \approx 0.89$. Pooling the five pairs, the estimated concordance is $51/67 \approx 0.76$.

Geographic region	Population	Number of individuals in CNV dataset	Number of observed CNVs	Number of observed deletions	Number of observed duplications	Number of CNV loci with CNVs	Number of CNV loci in DBGV	Number of new CNV loci
AFRICA	San	7	38	22	16	33	27	6
	Mbuti Pygmy	14	101	49	52	70	49	21
	Biaka Pygmy	31	214	107	107	112	88	24
	Bantu (Kenya)	12	99	79	20	71	55	16
	Bantu (S. Africa)	7	42	25	17	36	30	6
	Yoruba	25	147	85	62	79	68	11
	Mandenka	24	148	98	50	96	68	28
MIDDLE EAST	Mozabite	29	159	88	71	80	68	12
	Bedouin	43	247	147	100	131	103	28
	Palestinian	25	182	126	56	121	90	31
	Druze	40	262	143	119	126	91	35
EUROPE	Basque	11	80	28	52	57	41	16
	Russian	13	153	117	36	110	73	37
	Adygei	13	80	61	19	63	42	21
C/S ASIA	Balochi	14	70	39	31	55	47	8
	Kalash	13	278	258	20	147	108	39
	Burusho	6	42	18	24	39	32	7
	Uygur	9	39	15	24	33	23	10
EAST ASIA	Yakut	12	86	58	28	71	47	24
	Mongola	9	53	27	26	45	36	9
	Daur	10	60	38	22	56	38	18
	Yi	9	36	21	15	34	22	12
	Cambodian	10	44	18	26	41	28	13
	Lahu	8	38	19	19	28	20	8
OCEANIA	Melanesian	11	332	289	43	178	132	46
	Papuan	12	246	200	46	163	121	42
AMERICA	Pima	8	70	55	15	52	26	26
	Maya	11	169	148	21	138	79	59
	Colombian	7	37	20	17	28	21	7

Table S12: Summary of CNVs detected in 443 HGDP-CEPH individuals from 29 populations. The table is based on a total of 3552 CNVs at 1428 copy-number-variable loci, and it forms the basis for Figure 4b. DBGV refers to the Database of Genomic Variants[46,47]; "new" CNV loci are those not previously reported in DBGV version hg18.v3.

	HR^2			r^2		
Physical distance (kb)	Intercept ($\times 10^{-1}$)	Slope ($\times 10^{-6}$ km)	R^2	Intercept ($\times 10^{-1}$)	Slope ($\times 10^{-6}$ km)	R^2
5	3.065	9.080	0.8381	2.800	11.97	0.8261
10	2.882	7.950	0.8327	2.434	11.28	0.8195
15	2.772	7.184	0.8264	2.192	10.75	0.8247
20	2.700	6.460	0.8347	2.019	10.02	0.8207
25	2.645	5.820	0.8315	1.887	9.432	0.8189
30	2.612	5.255	0.8147	1.798	8.837	0.8148
35	2.589	4.899	0.8382	1.724	8.364	0.8195
40	2.559	4.415	0.8277	1.657	7.837	0.8165
45	2.543	4.152	0.8198	1.602	7.307	0.8172
50	2.524	3.866	0.8092	1.556	7.051	0.8196

Table S13: Linear regression of linkage disequilibrium on geographic distance from East Africa. LD is measured using HR^2 applied to unphased data or using r^2 applied to phased data, and distance to East Africa is measured from Addis Ababa using waypoint routes. The table indicates that geographic distance explains variation in LD to a similar extent for HR^2 with unphased data and for r^2 with phased data, and that in the range shown the physical distance at which LD is measured has only a relatively slight impact on the fraction of variation explained.

	Number of replicates in most frequent mode		
K	SNPs	Haplotypes	CNVs
2	40	40	40
3	40	40	40
4	38	23	37
5	16	36	6
6	16	29	3
7	9	20	2
8	6	17	2
9	3	12	2
10	4	22	2

Table S14: Number of replicates appearing in the most frequent `Structure` clustering mode, for the worldwide SNP, haplotype, and CNV datasets (relative to a maximum of 40). Ties for the most frequent mode were broken using the `CLUMPP` H' score[18] for replicates in the mode.

	Number of replicates in most frequent mode						
K	Africa	Middle East	Europe	C/S Asia	East Asia	Oceania	America
2	39	40	23	40	39	40	33
3	19	27	24	19	15	34	32
4	11	10	2	4	20	37	16
5	8	6	1	2	14	31	15
6	3				7		
7	2				12		
8	9				10		

Table S15: Number of replicates appearing in the most frequent `Structure` clustering mode, for individual geographic regions (relative to a maximum of 40). Ties for the most frequent mode were broken using the `CLUMPP` H' score[18] for replicates in the mode.

	Mean log likelihood for replicates in most frequent mode		
K	SNPs	Haplotypes	CNVs
2	-2218914	-10834453	-11074.20
3	-2166724	-10534972	-11037.02
4	-2142764	-10499423	-11163.83
5	-2136825	-10404686	-12997.67
6	-2123684	-10436284	-10534.03
7	-2154336	-10428550	-10230.50
8	-2605349	-10985902	-10197.75
9	-2137650	-10673840	-10271.60
10	-2129510	-10542288	-10017.55

Table S16: Mean log likelihood for replicates appearing in the most frequent `Structure` clustering mode, for the worldwide SNP, haplotype, and CNV datasets.

	Mean log likelihood for replicates in most frequent mode						
K	Africa	Middle East	Europe	C/S Asia	East Asia	Oceania	America
2	-484660.9	-681107.9	-201262.9	-240451.7	-285320.0	-104755.5	-106636.2
3	-479091.7	-684091.8	-201813.5	-240042.6	-291057.2	-105559.5	-281717.5
4	-494108.2	-683420.9	-201785.5	-240308.0	-296049.0	-107332.2	-117576.8
5	-493140.8	-685805.3	no mode	-269941.8	-286742.2	-109007.6	-155811.1
6	-510799.9				-296119.6		
7	-792167.5				-286533.2		
8	-2177801				-285659.2		

Table S17: Mean log likelihood for replicates appearing in the most frequent `Structure` clustering mode, for individual geographic regions. For Europe with $K = 5$ no mode contained at least two replicates.

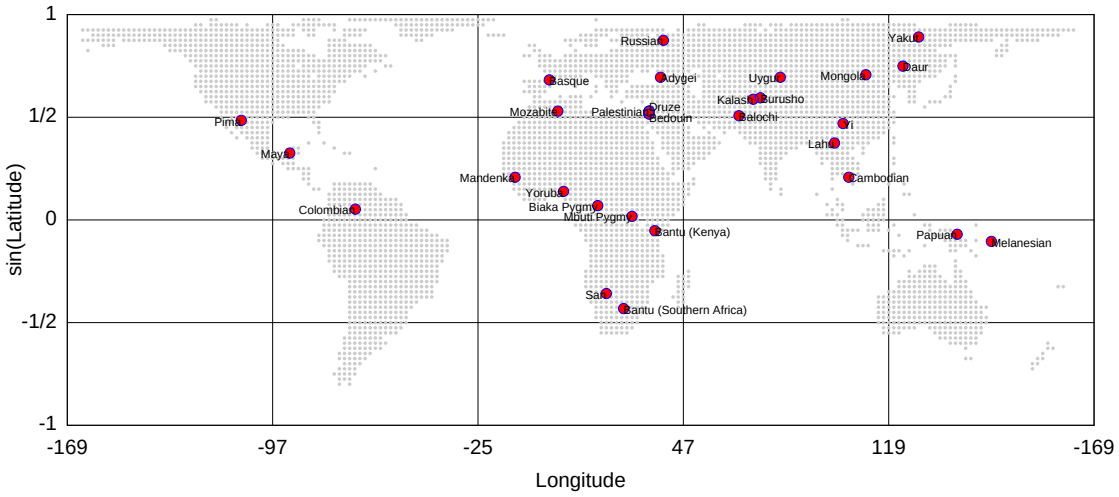

Figure S1: Map of the geographic locations of HGDP-CEPH populations included in the study. Geographic coordinates used for the populations are provided in Table S6.

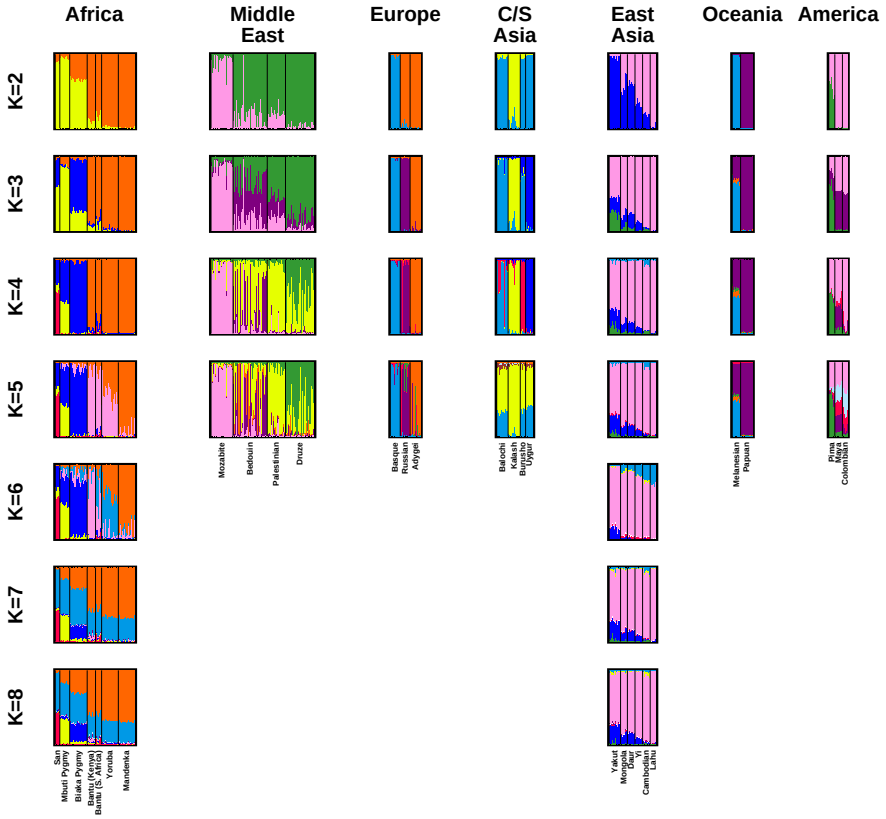

Figure S2: Population structure within geographic regions, inferred from the SNP dataset for various choices of the number of clusters, K. The plots in Figure 1c for individual geographic regions were extracted from this figure on the basis of the maximal mean log likelihood for replicates appearing in the most frequently observed mode (Table S17). If K^* is the number of predefined populations in a geographic region, K proceeds from 2 to at least $K^* + 1$.

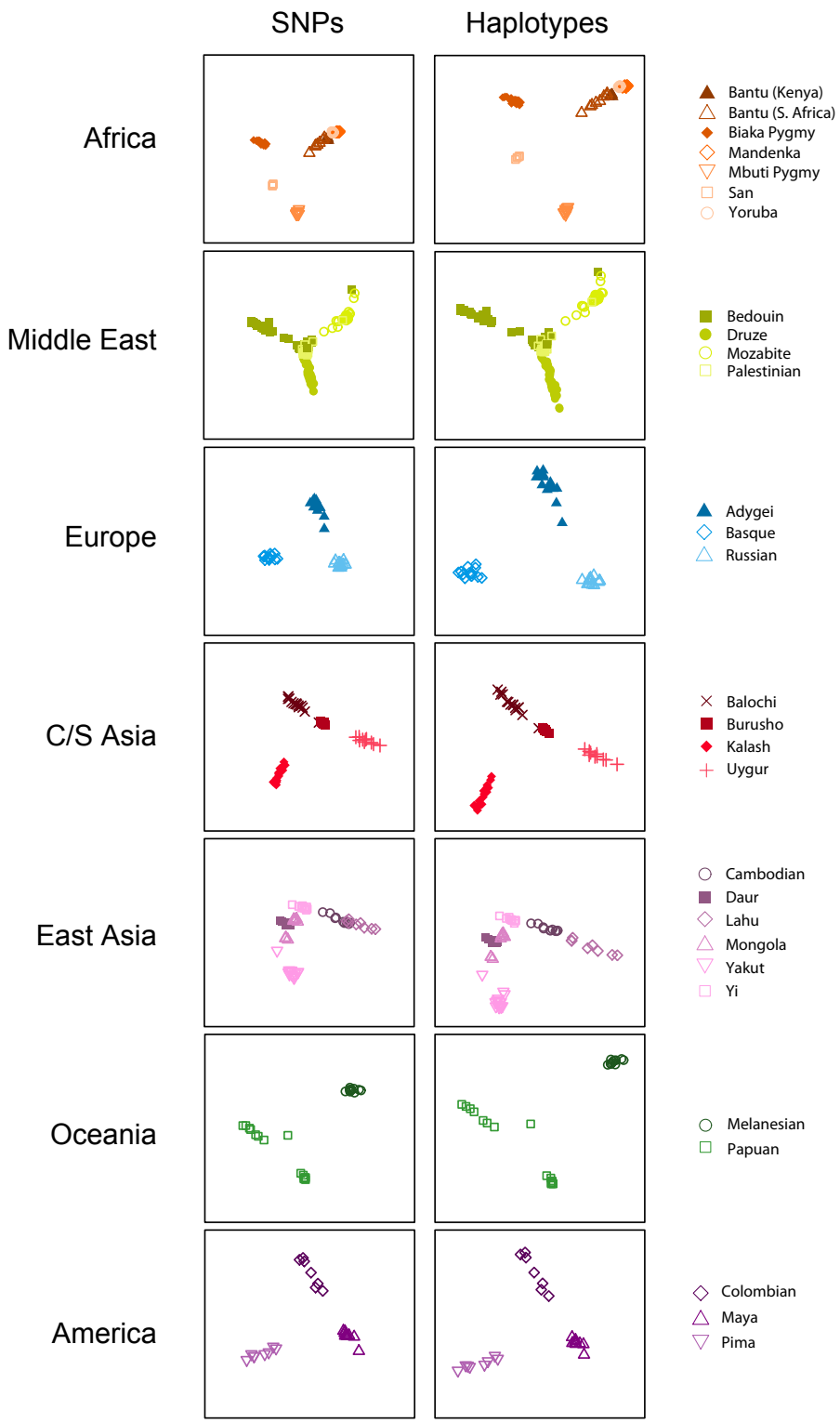

Figure S3: Multidimensional scaling representations of genetic distance matrices for individual geographic regions. Analysis was performed separately for the unphased SNP data and for the haplotype cluster data. This plot illustrates that within a given geographic region, the individuals of different populations can be further subdivided into clusters corresponding largely to distinct populations.

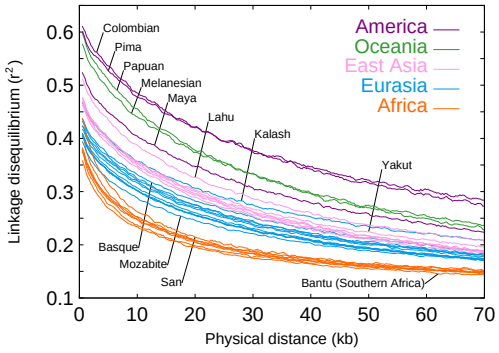

Figure S4: Linkage disequilibrium measured by r^2 as a function of physical distance. Similarly to the corresponding plot of HR^2 in Figure 2b, for each population, the mean LD across pairs of SNPs within 1kb bins is plotted at intervals of 250bp. Adjustment for sample size was performed by computing r^2 from 10 random haplotypes per population for each SNP pair considered. The figure shows a pattern of LD decay for r^2 with phased data similar to the pattern in Figure 2b for HR^2 with unphased data.

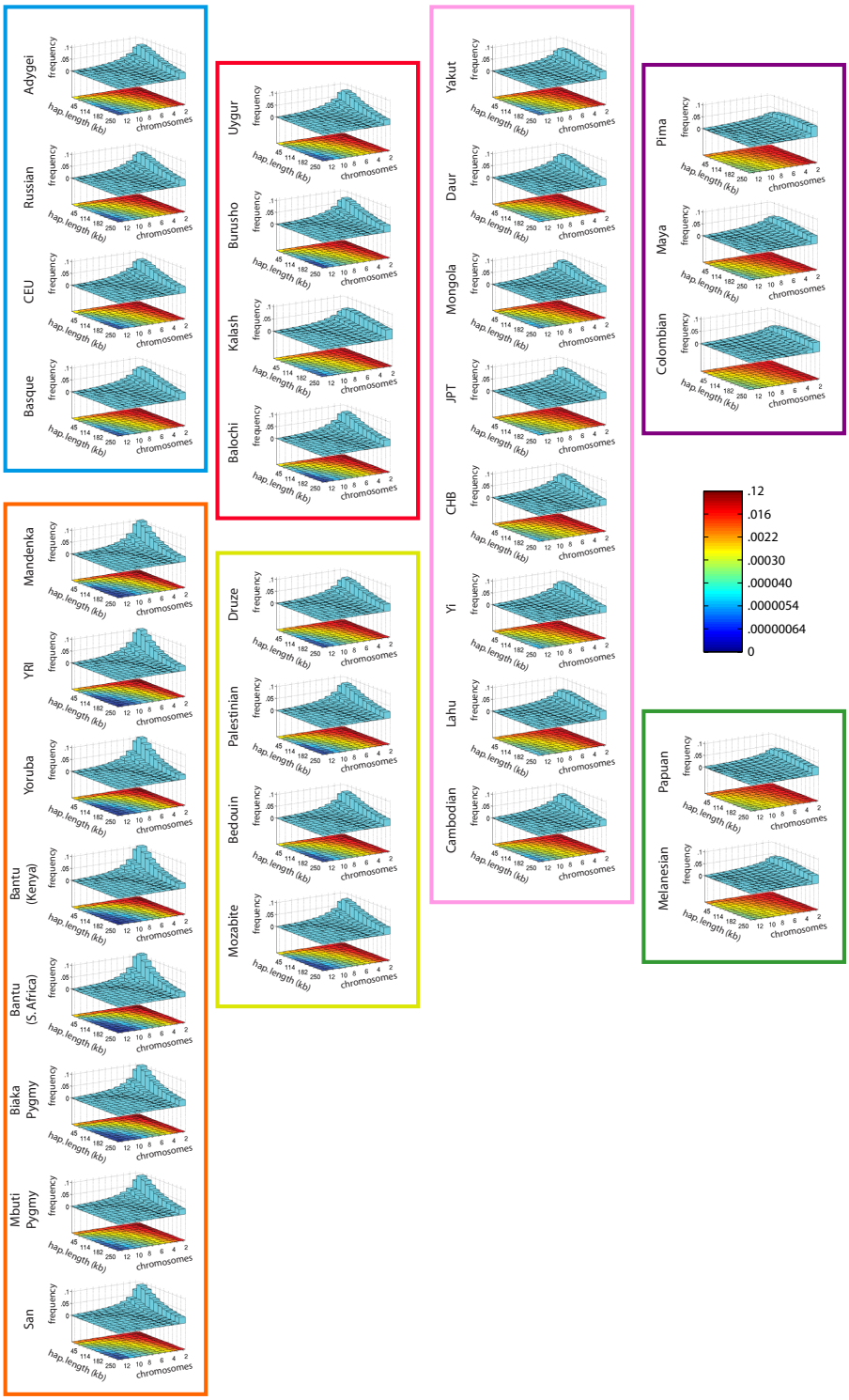

Figure S5: Joint distribution of haplotype length and frequency. The x-axis represents haplotype frequency, indicated by number of chromosomes observed for a given haplotype in a sample of 12 chromosomes (truncated at one chromosome). The y-axis represents haplotype length, and the z-axis represents the density of haplotypes of a specific length and frequency. A heat map of the density is shown on a logarithmic scale below each histogram. The figure illustrates that in Africa, short low-frequency haplotypes are common, while long high-frequency haplotypes are almost absent. Populations from Europe, the Middle East, and Central/South Asia have fewer short low-frequency haplotypes and more long high-frequency haplotypes. This trend of increasing occurrence of long high-frequency haplotypes continues into East Asia, and finally, into Oceania and America.

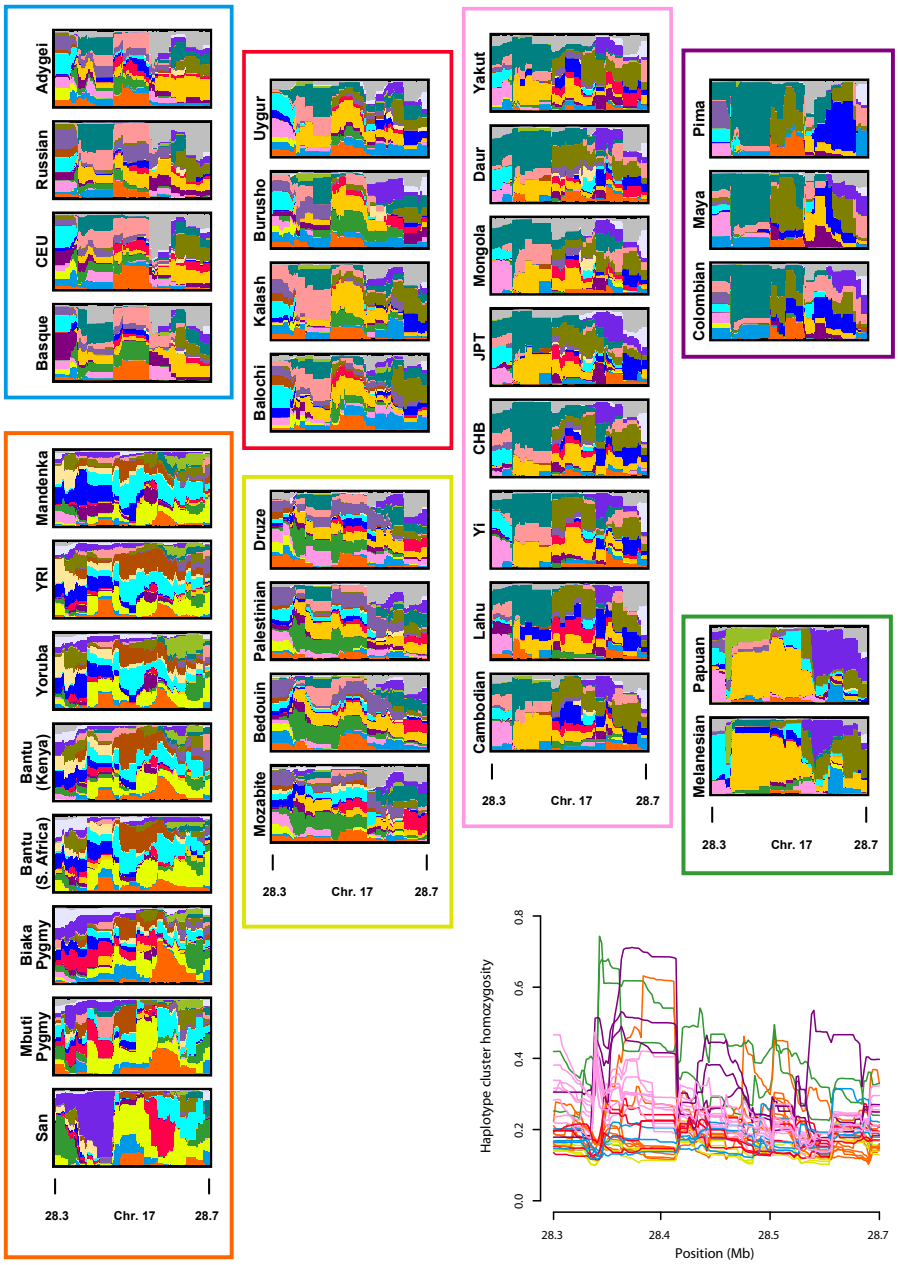

Figure S6: Haplotype cluster frequencies for a "typical" genomic region of 150 SNPs (from chromosome 17), and haplotype cluster homozygosity across the region for each population. Reduced haplotype diversity in Oceania and the Americas is consistent with founder effects via migration from East Asia — in the left part of the region, it is possible that via separate founder effects the orange cluster common in East Asia rose to a high frequency in Oceania, while the blue-green cluster rose to a high frequency in the Americas. This figure illustrates several frequently-observed features of such plots. Africa exhibits great diversity, with common haplotype clusters rare or absent elsewhere. A sampling of haplotype clusters moving outward from Africa towards Oceania and the Americas is also a typical pattern. Finally, haplotypic similarity of populations from the same geographic region is also evident.

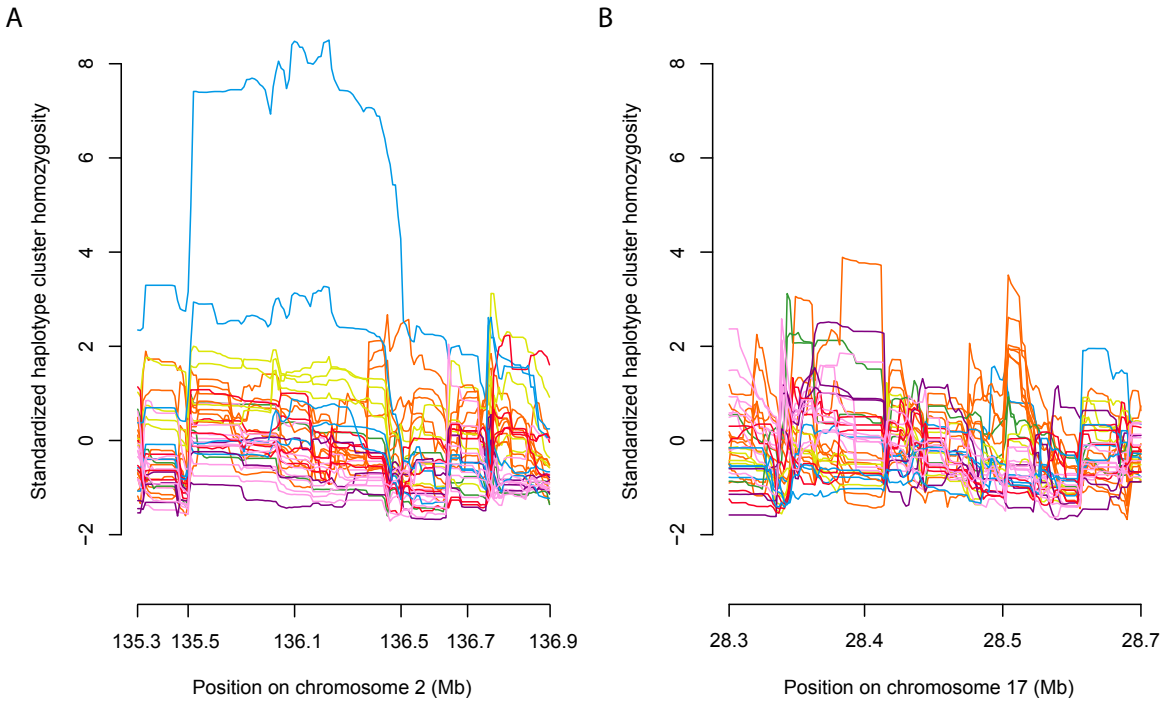

Figure S7: Standardized haplotype cluster homozygosity for (A) the lactase region, and (B) the "typical" genomic region of 150 SNPs shown in Figure S6. For each population, standardized homozygosity is obtained pointwise by subtracting the mean haplotype cluster homozygosity in the population across the genome and dividing by the standard deviation. The mean and standard deviation are obtained by using all ten haplotype cluster datasets employed throughout our population-genetic analysis. Populations are color-coded geographically as in Figures 3 and S6. Part A shows that the CEU population has unusually high homozygosity in the lactase region, but that homozygosities of other populations are less extreme. Part B provides a comparison with a typical region. The x-axis is scaled by the number of SNPs typed; corresponding physical positions are indicated on the graphs.

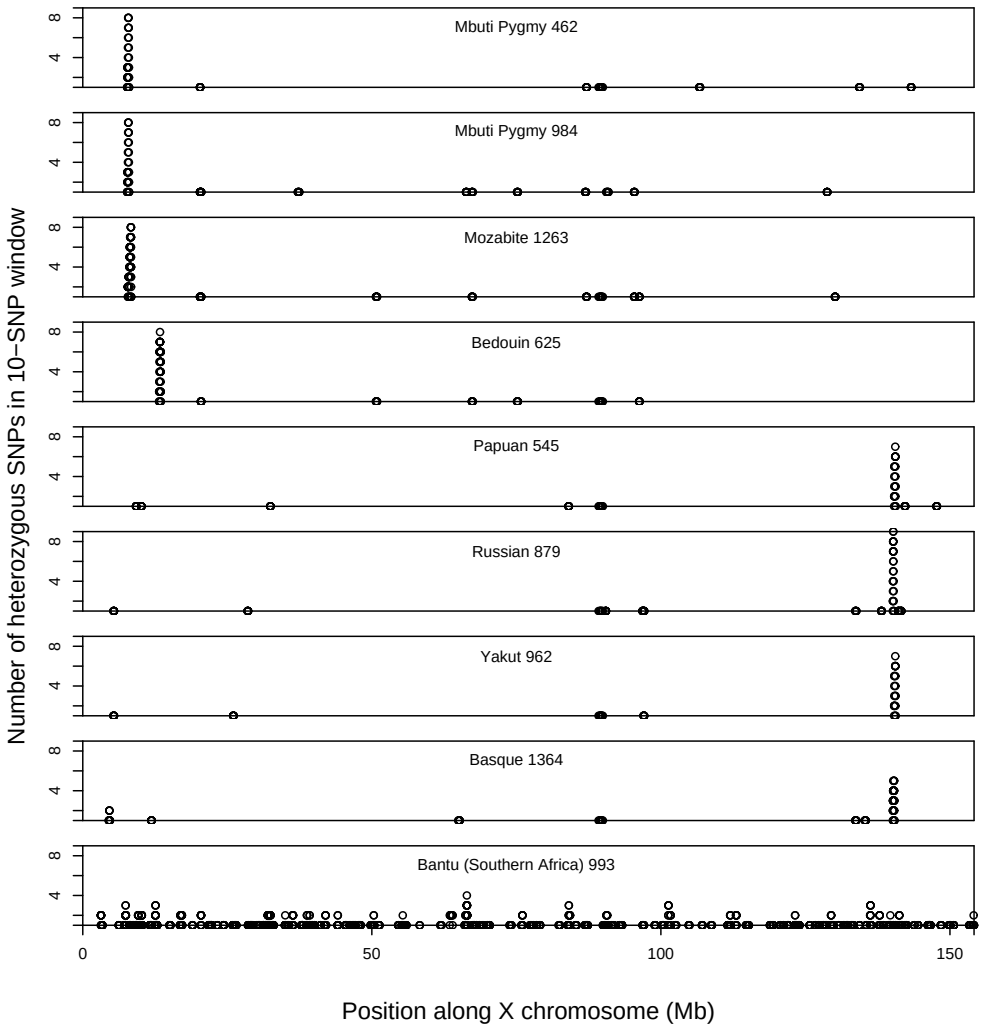

Figure S8: Identification of duplications using scans of X-chromosomal heterozygosity in males. The plot shows the number of heterozygous SNPs in sliding windows that contain 10 SNPs. Only windows of the (non-pseudoautosomal) X chromosome in which at least one SNP is heterozygous are shown, for only the nine males who had at least one window with at least three heterozygous SNPs. As duplications can lead to extended stretches of consecutive heterozygous genotypes along male X chromosomes, the high peaks around a single position in each of the top eight plots indicate the likely presence of duplications. By contrast, the plot for Bantu (Southern Africa) 993 displays no clear peaks. The higher overall level of heterozygosity in this individual is more likely to be due to genotyping error: individual 993 has genotyping call rate 0.9818 (after reclustering), compared with values ≥ 0.9977 for each of the other eight males shown. Due to poorer data quality, individual 993 is among the individuals not considered in the `PennCNV` analysis.

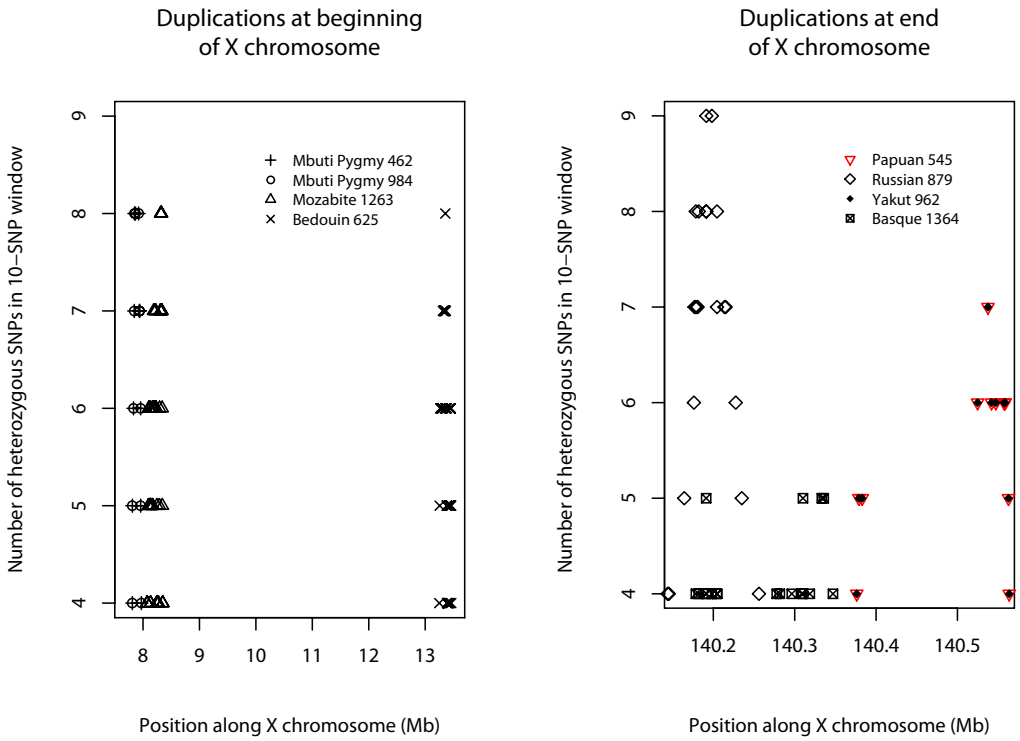

Figure S9: Identification of duplications in males on two parts of the X chromosome. The plot shows the number of heterozygous SNPs in sliding windows that contain 10 SNPs, magnified near two X-chromosomal regions in which duplications were detected. Papuan individual 545 was not included in the `PennCNV` analysis, and consequently, the duplication in this individual was not detectable by `PennCNV`. All other duplications detected by extended stretches of X-chromosomal heterozygosity in males were also detected by `PennCNV`.

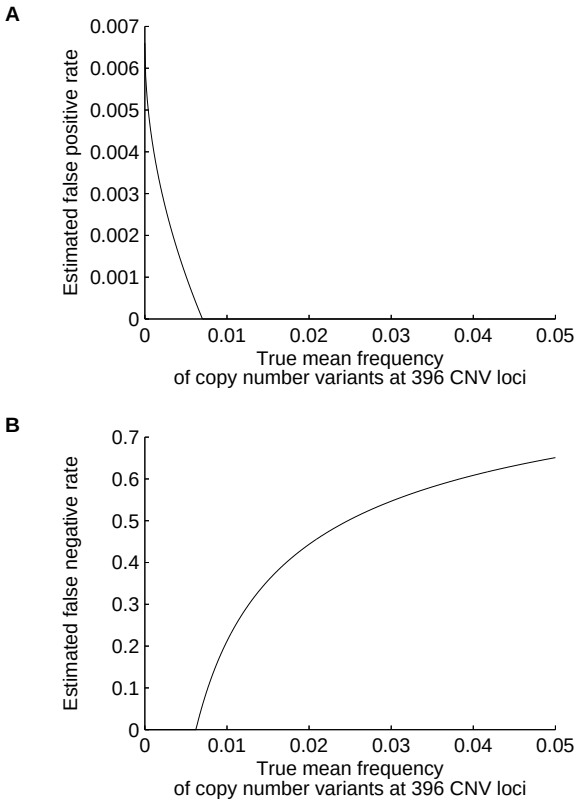

Figure S10: Estimated false positive and false negative rates as functions of the unknown true mean frequency of copy-number variants across 396 CNV loci. The plots are based on equations 3 and 4, with an estimated concordance of duplicates equal to $871/980 \approx 0.89$, the average concordance across duplicate pairs. (A) False positive rate. (B) False negative rate. The figure shows a low false positive rate for CNV calls and a comparatively higher false negative rate.

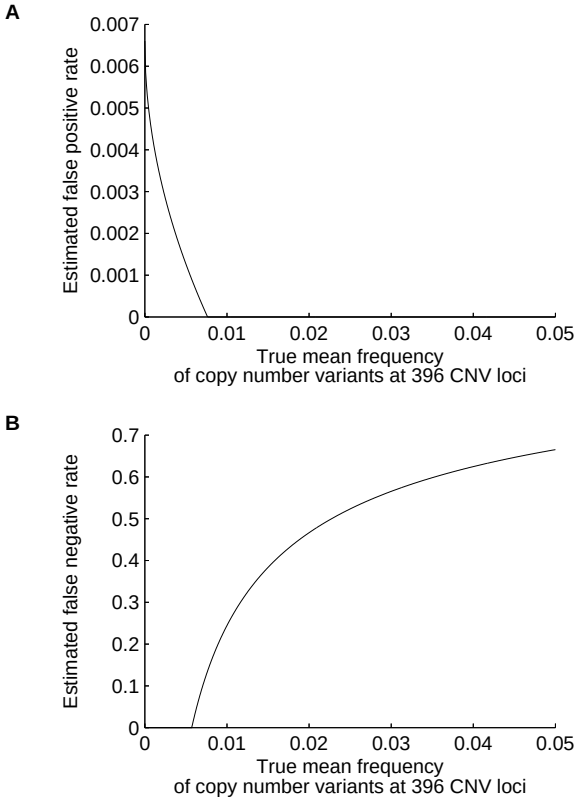

Figure S11: Estimated false positive and false negative rates as functions of the unknown true mean frequency of copy-number variants across 396 CNV loci. The plots are based on equations 3 and 4, with an estimated concordance of duplicates equal to $51/67 \approx 0.76$, a value estimated by pooling duplicate pairs. (A) False positive rate. (B) False negative rate. The figure shows a low false positive rate for CNV calls and a comparatively higher false negative rate.

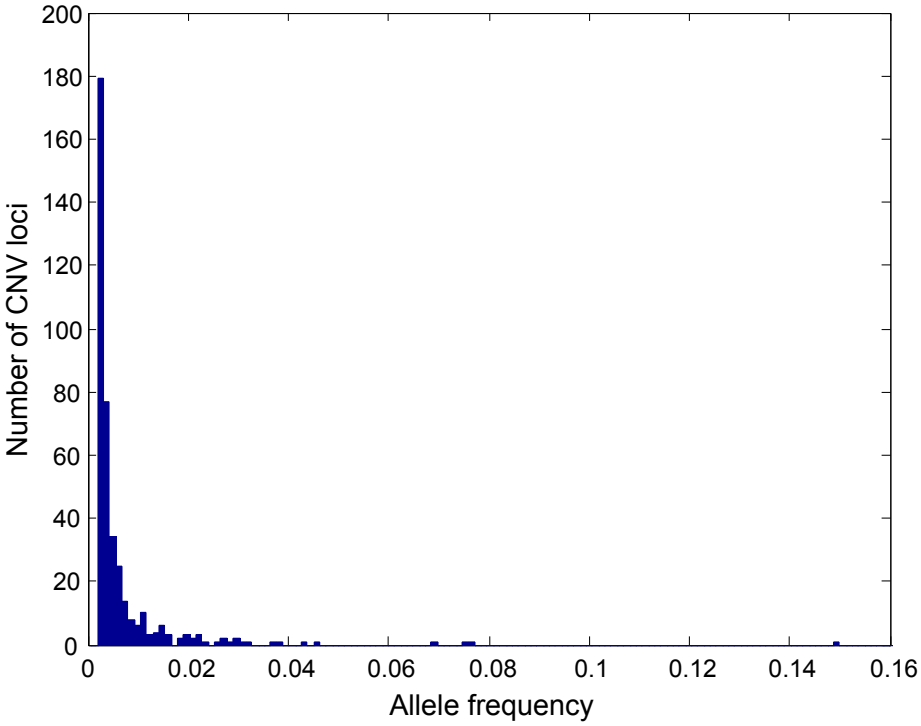

Figure S12: Allele frequency spectrum for the copy-number variants at 1302 autosomal CNV loci in 405 unrelated individuals from 29 populations. The 1/810 frequency class is not plotted and contains 906 loci. The figure illustrates that most CNVs were observed to be rare.

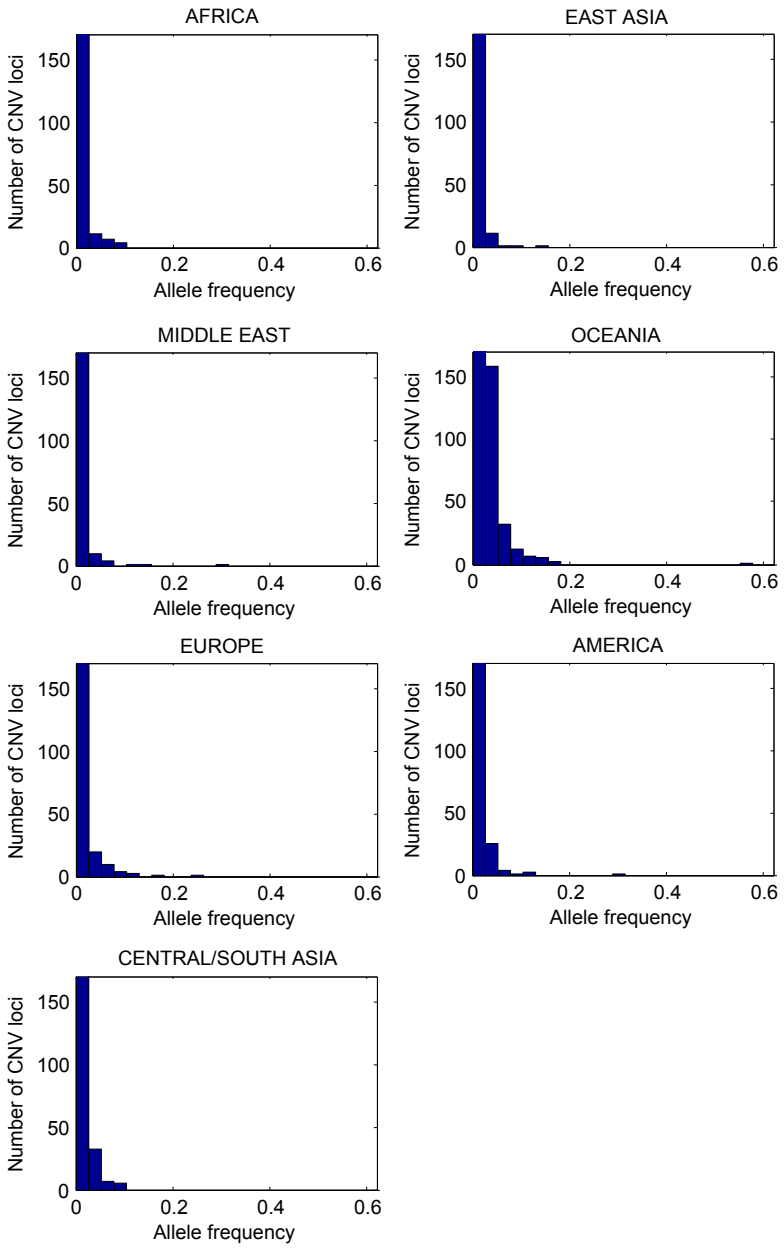

Figure S13: Allele frequency spectra for the copy-number variants at 1302 autosomal CNV loci, considering unrelated individuals in 7 geographic regions. To adjust for sample size differences among regions, we used a resampling procedure. The alleles of each individual were partitioned into two "pseudo-genomes," each containing one allele at each CNV locus. Next, for each CNV locus, 38 haploid pseudo-genomes were randomly drawn (without replacement) from each geographic region and the frequency of each CNV was calculated. An average frequency for each CNV was then computed across 1000 sets of pseudo-genomes (with the sets independently chosen for different CNV loci). Using this average frequency, loci were classified into one of 39 allele frequency bins: $[0, 1/38]$, $(1/38, 2/38]$, ..., $(36/38, 37/38]$, $(37/38, 1]$. After averaging, the numbers of CNV loci in the $[0, 1/38]$ bin were as follows: Africa — 1281, Middle East — 1285, Europe — 1266, Central/South Asia — 1259, East Asia — 1288, Oceania — 1086, and America — 1270.

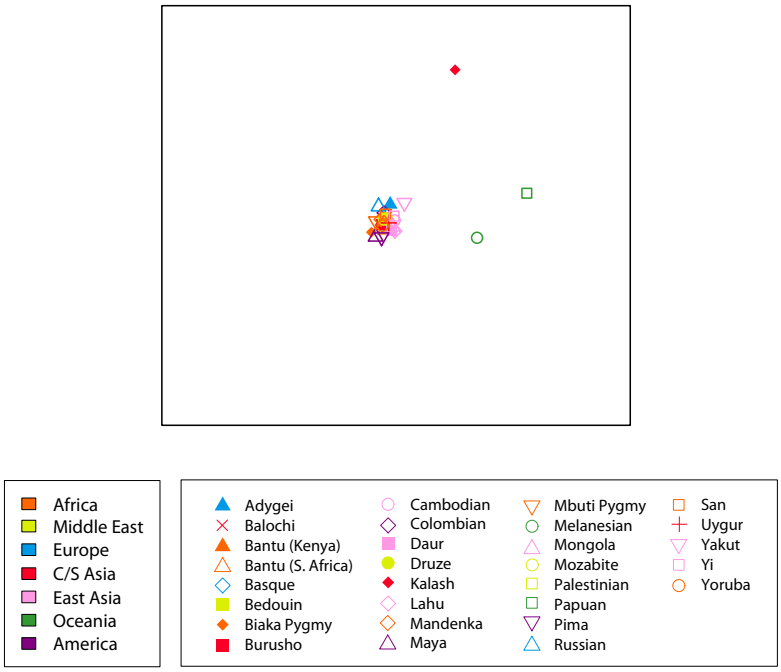

Figure S14: Multidimensional scaling representation of the genetic distance matrix for the CNV dataset. Three outliers removed from the matrix for the CNV plot in Figure 1d (Kalash, Melanesian, and Papuan) are retained in the analysis that underlies this plot. The geographic clustering visible in Figure 1d is less visible in this plot, with the outliers included.

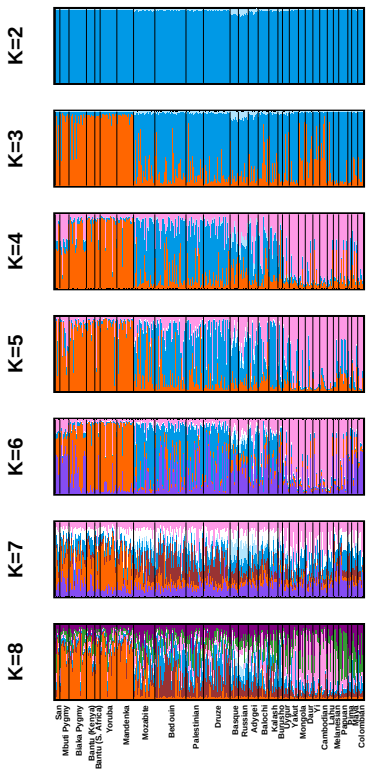

Figure S15: Population structure inferred from SNP sets of the same size and frequency spectrum as the CNV dataset, for various choices of the number of clusters, K. The figure shows that the level of "noise" in population structure plots based on the reduced SNP datasets is comparable to the corresponding level in plots based on the CNV dataset (Figures 1c and S24).

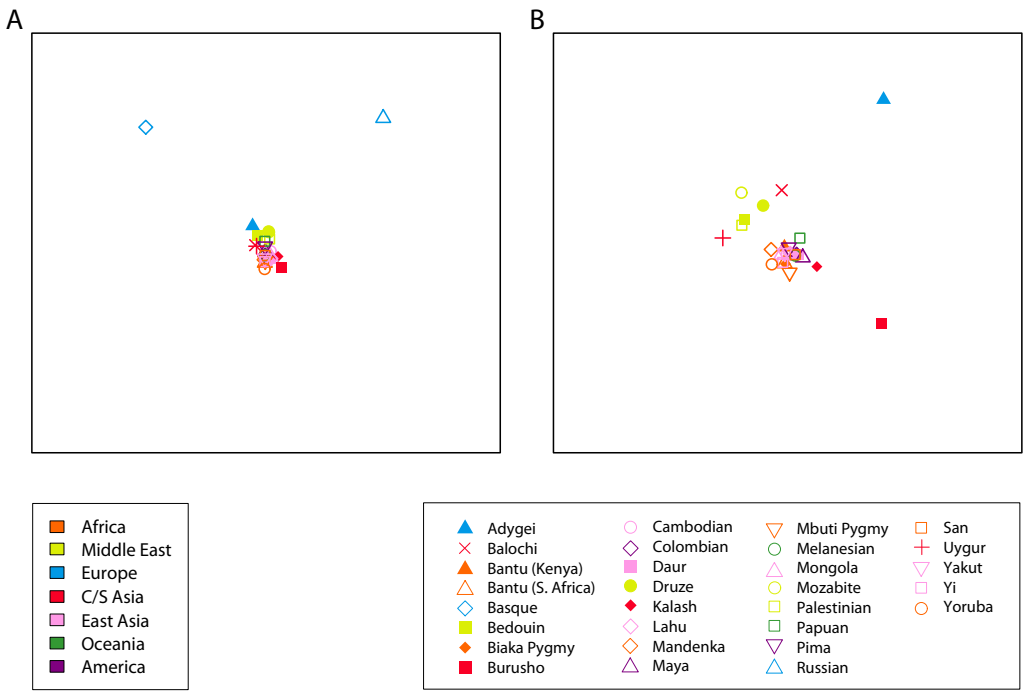

Figure S16: Multidimensional scaling representations of genetic distance matrices for individual geographic regions, based on a reduced SNP dataset of the same size and frequency spectrum as the CNV dataset. (A) All populations. (B) Two outliers removed (Basque and Russian). The figure shows that the level of "noise" in multidimensional scaling plots based on the SNP dataset is comparable to the corresponding level in plots based on the CNV dataset (Figures 1d and S14).

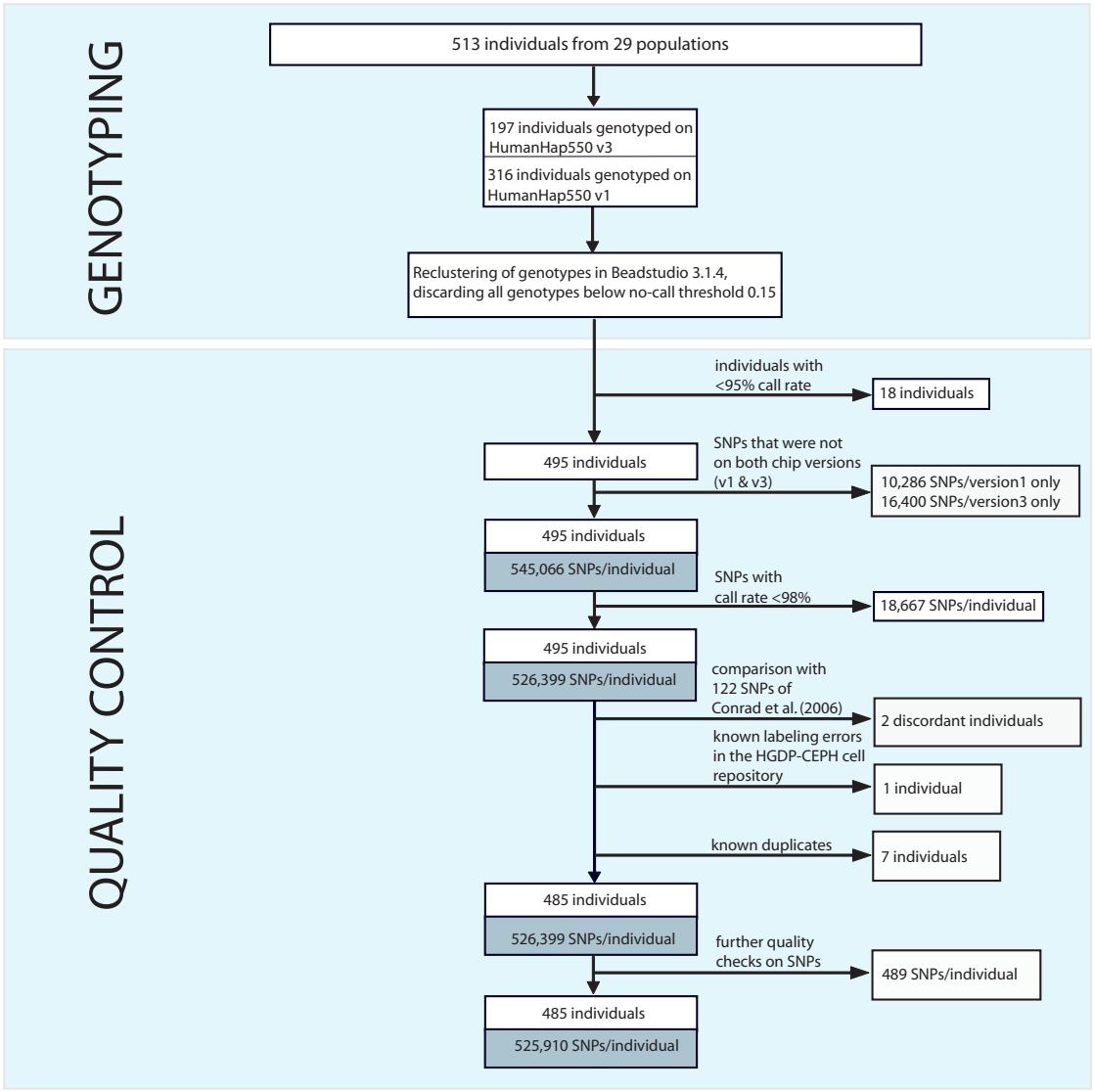

Figure S17: Flow chart of the genotyping and quality control process for SNP genotypes.

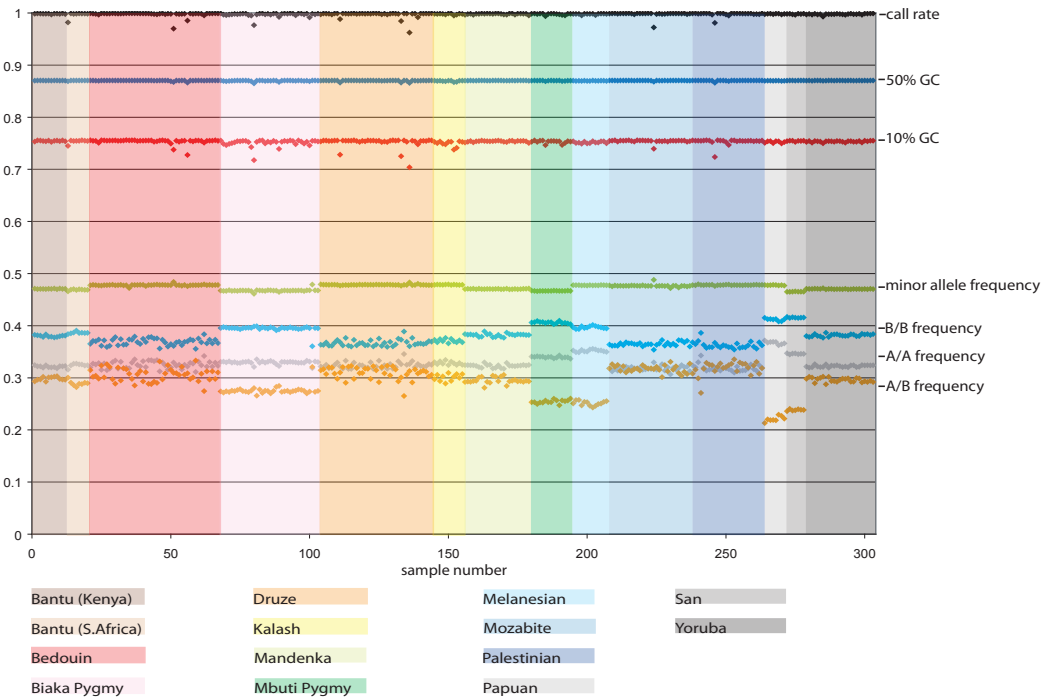

Figure S18: Quality control plot for 303 individuals genotyped on HumanHap550 version 1 BeadChips. The 303 individuals plotted are those with a genotype call rate >95% after genotype reclustering. Each background color represents the individuals from a single population, and within each population, individuals are sorted by HGDP-CEPH identification number. Black diamonds indicate call rates, blue diamonds indicate the 50% GC score (median GenCall score across SNPs), red diamonds indicate the 10% GC score (tenth percentile of the ranked GenCall scores), green diamonds indicate the minor allele frequency, turquoise diamonds indicate the frequency of B/B calls, gray diamonds indicate the frequency of A/A calls, and orange diamonds indicate the frequency of A/B calls. For a given individual and SNP, the GenCall score is a measure of data quality that takes into account the fit of the individual genotype to defined genotype clusters. Scores above 0.7 indicate high-quality genotypes and scores below 0.2 indicate low-quality genotypes. We used a GenCall threshold of 0.15 in measuring the call rate.

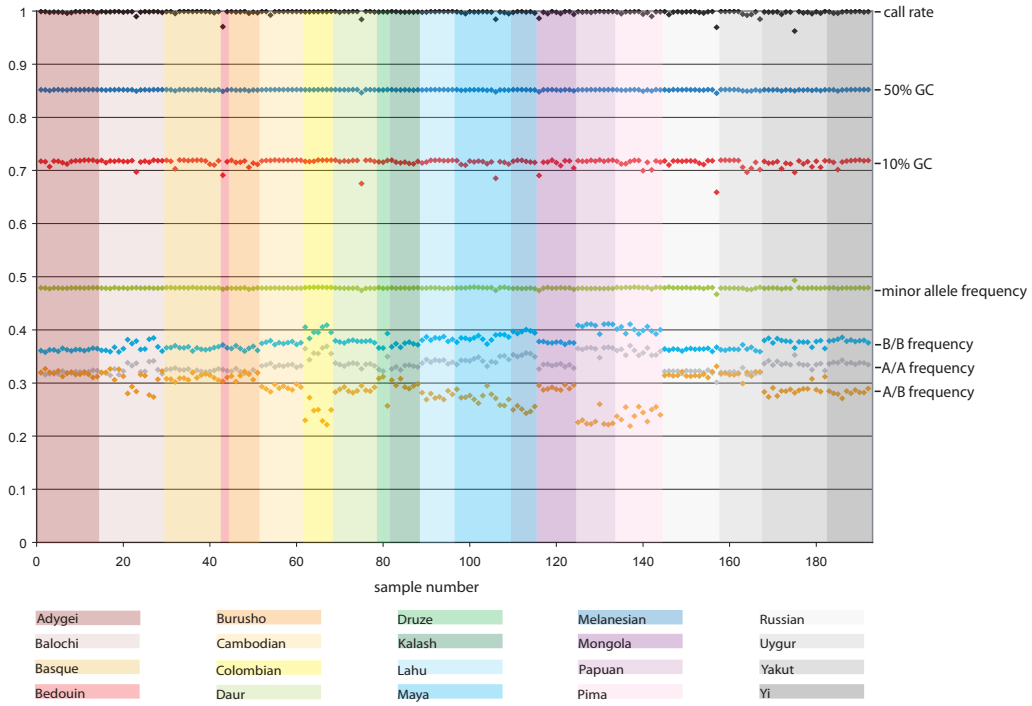

Figure S19: Quality control plot for 192 individuals genotyped on HumanHap550 version 3 BeadChips. The 192 individuals plotted are those with a genotype call rate >95% after genotype reclustering. Each background color represents the individuals from a single population, and within each population, individuals are sorted by HGDP-CEPH identification number. Black diamonds indicate call rates, blue diamonds indicate the 50% GC score (median GenCall score across SNPs), red diamonds indicate the 10% GC score (tenth percentile of the ranked GenCall scores), green diamonds indicate the minor allele frequency, turquoise diamonds indicate the frequency of B/B calls, gray diamonds indicate the frequency of A/A calls, and orange diamonds indicate the frequency of A/B calls. For a given individual and SNP, the GenCall score is a measure of data quality that takes into account the fit of the individual genotype to defined genotype clusters. Scores above 0.7 indicate high-quality genotypes and scores below 0.2 indicate low-quality genotypes. We used a GenCall threshold of 0.15 in measuring the call rate. In the quality control plots, the minor allele at a locus is defined with respect to the BeadStudio project, and may differ between version 1 and version 3 BeadStudio projects.

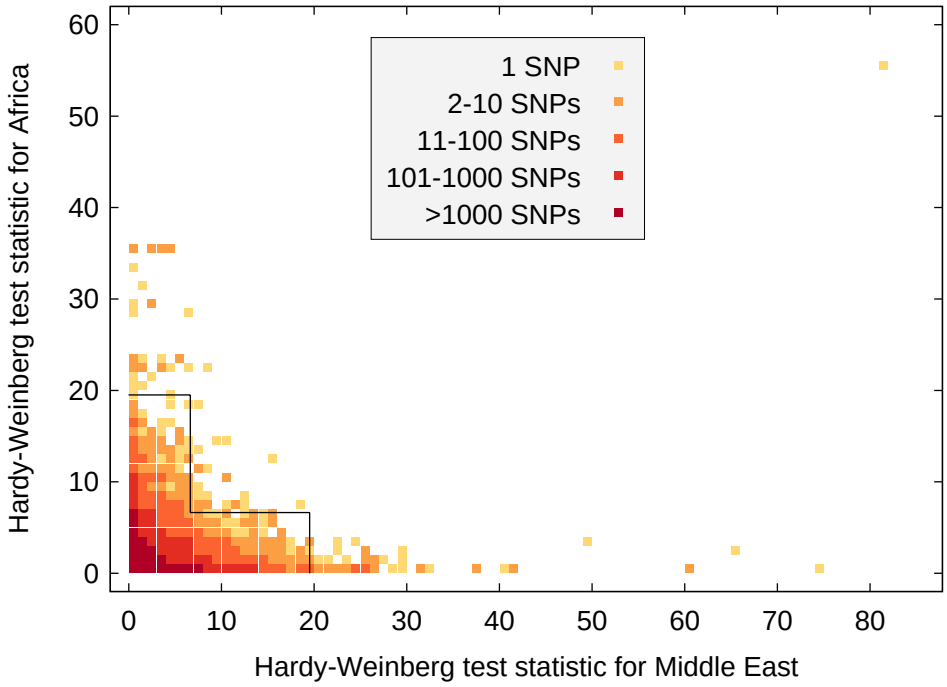

Figure S20: Hardy-Weinberg test statistics in two groups of individuals, 63 from Africa and 107 from the Middle East, for 447,215 autosomal, X-chromosomal, and pseudoautosomal SNPs with at least four copies of the minor allele in both groups of individuals. SNPs above or to the right of the cutoff lines in the graph were discarded from further analysis. The single extreme SNP in the upper right part of the plot is rs10027797.

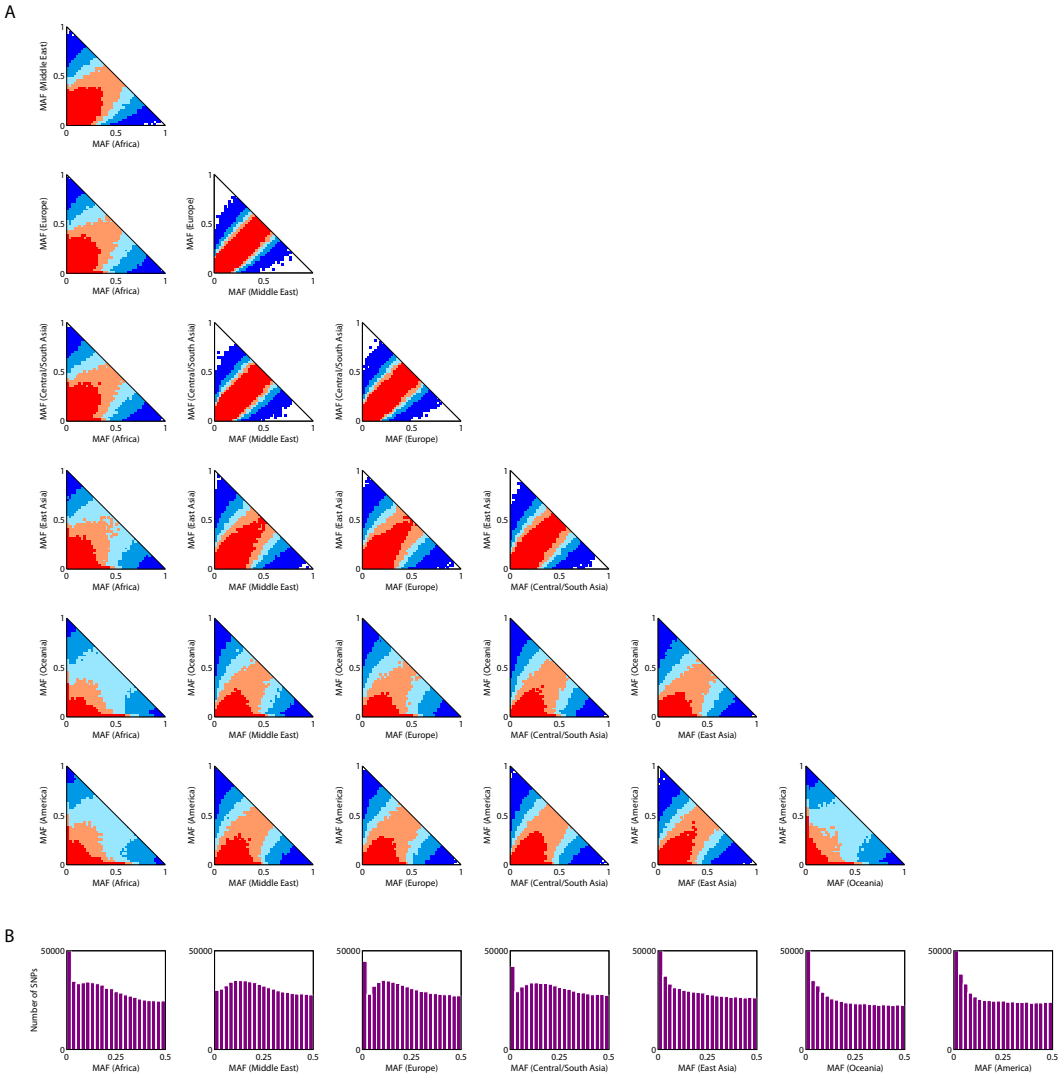

Figure S21: Allele frequency spectra for SNP loci. (A) Allele frequency in pairs of geographic regions. The minor allele for a pair of regions is defined as the allele whose average frequency in the two regions is at most 1/2, choosing arbitrarily in case of ties. From dark blue to red, colors indicate the number of SNPs in a bin: [1,125], [126,350], [351,550], [551,800], >800. A high degree of correlation is visible, particularly among the Middle East, Europe, and Central/South Asia. Relatively few SNPs are discordant in minor allele frequency. (B) Allele frequency spectra in specific regions. The zero class has 53,972 SNPs in Africa, 64,766 in East Asia, 125,733 in Oceania, and 111,671 in America. Eurasian-centered SNP ascertainment bias is visible, in that the Middle East, Europe, and Central/South Asia have fewer low-frequency variants and do not have monotonically decreasing distributions.

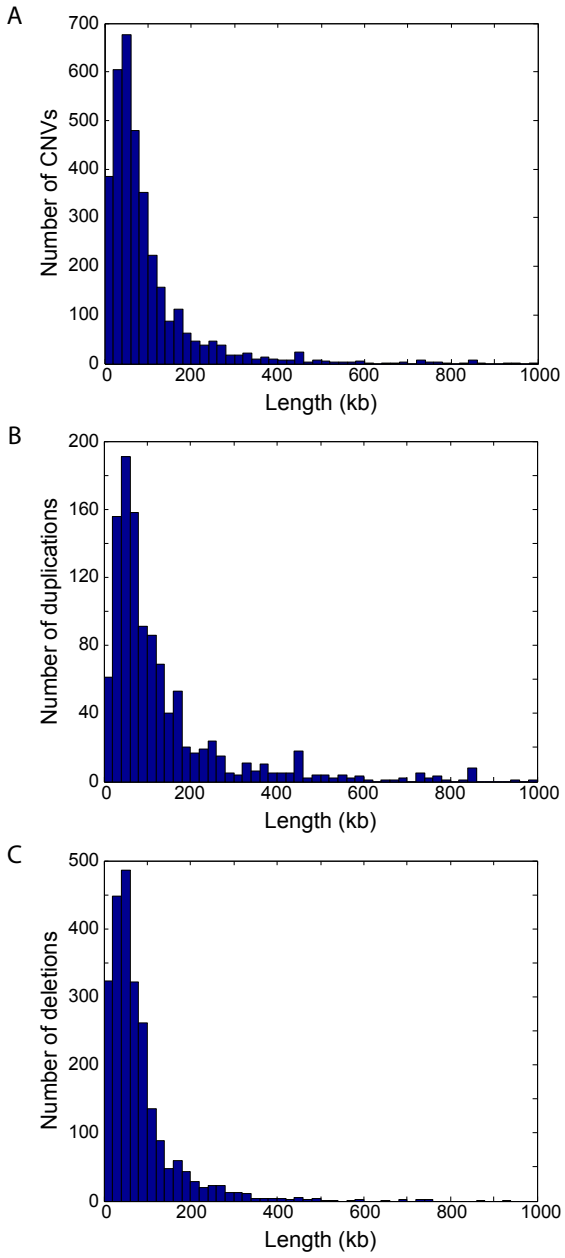

Figure S22: Distribution of the lengths of 3503 autosomal CNVs. (A) All CNVs. (B) Duplications (1117). (C) Deletions (2386). In kilobases, the bins in each histogram are (0, 20], ... , (980, 1000].

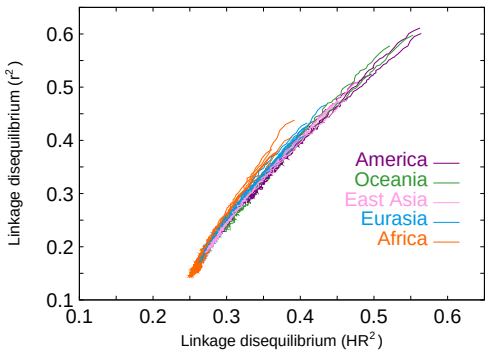

Figure S23: Comparison of LD measured by r^2 on phased data to LD measured by HR^2 on unphased data. For each population, ordered pairs (HR^2, r^2) are obtained at 250bp intervals and are connected in increasing order of physical distance. The figure illustrates a close relationship between HR^2 computed on unphased data and r^2 computed on phased data.

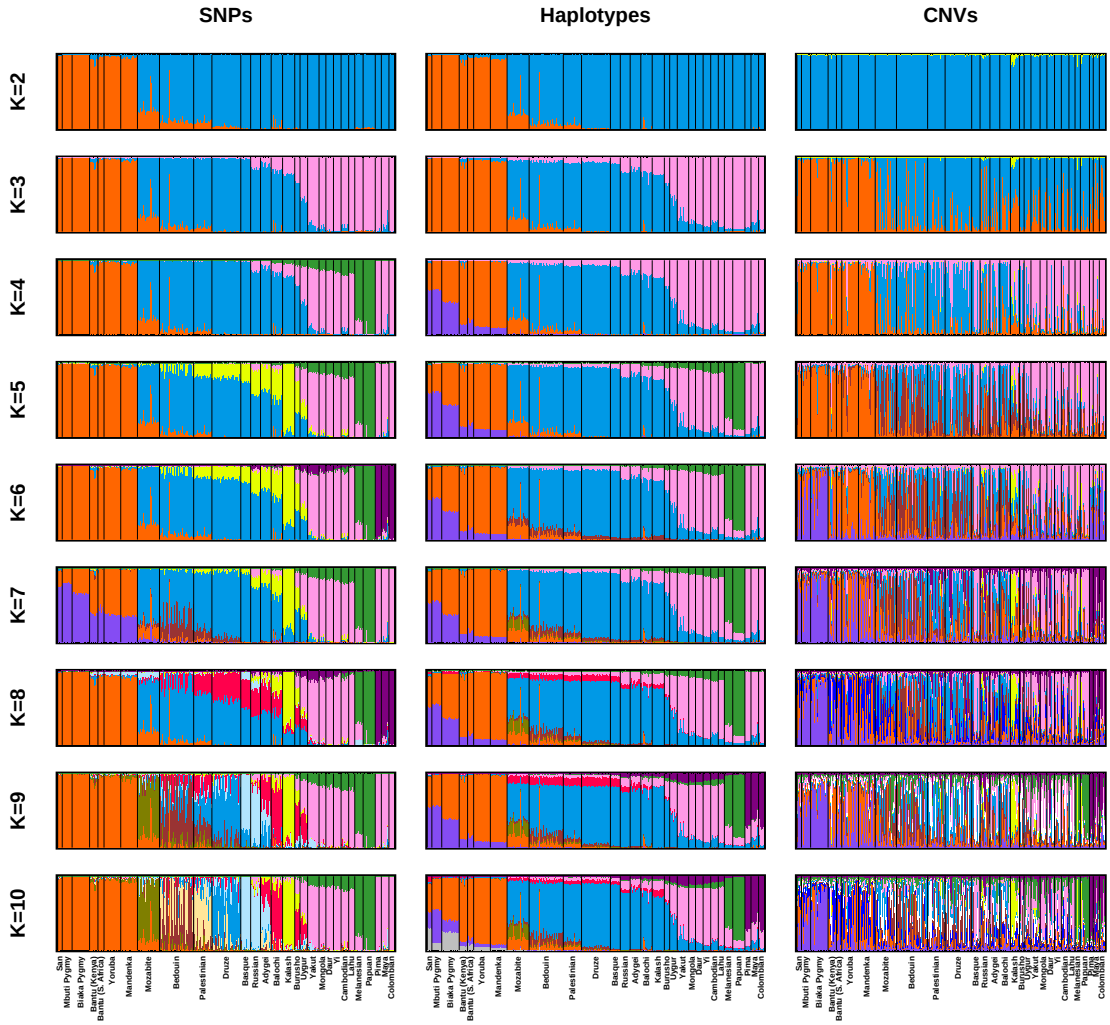

Figure S24: Population structure inferred from SNPs, haplotypes, and CNVs for various choices of the number of clusters, K. The plots with $K \leq 6$ are copied in Figure 1c. The Bedouin and two Mozabites with high estimated membership in the cluster corresponding to Africans in SNP and haplotype analyses are the same individuals seen to lie on the path connecting Africans to the remaining individuals from the Middle East in the multidimensional scaling analysis (Figure 1d). Two populations with a considerable degree of mixed membership between the clusters corresponding to Eurasia and East Asia — Burusho and Uygur — are also the two populations most intermediate in the MDS analysis between Europe and East Asia.

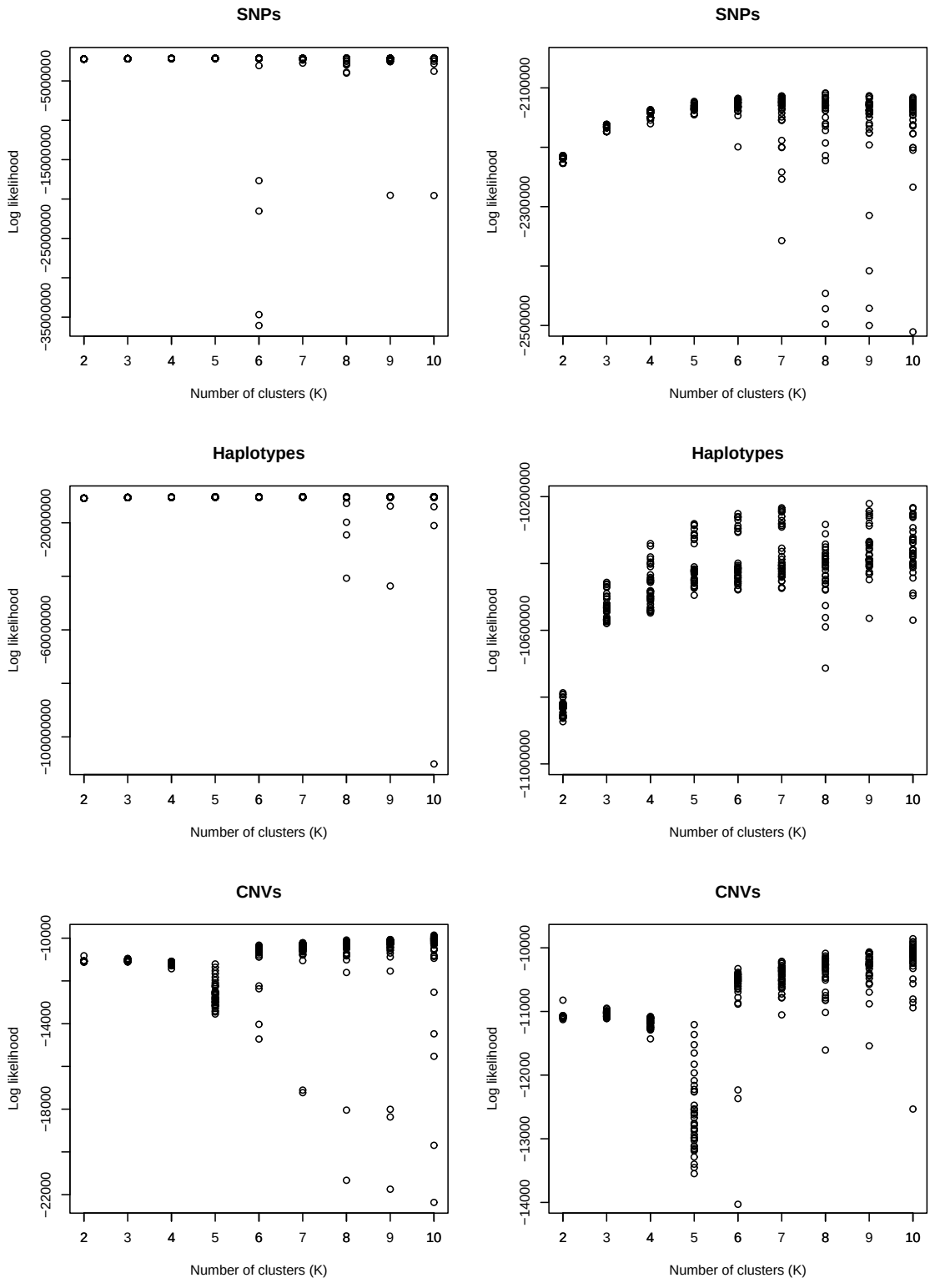

Figure S25: Log likelihood as a function of the number of clusters, based on `Structure` runs applied to the full worldwide dataset. Likelihoods for all 40 replicates at a given K are shown. The plots on the right side are magnified versions of the plots on the left.

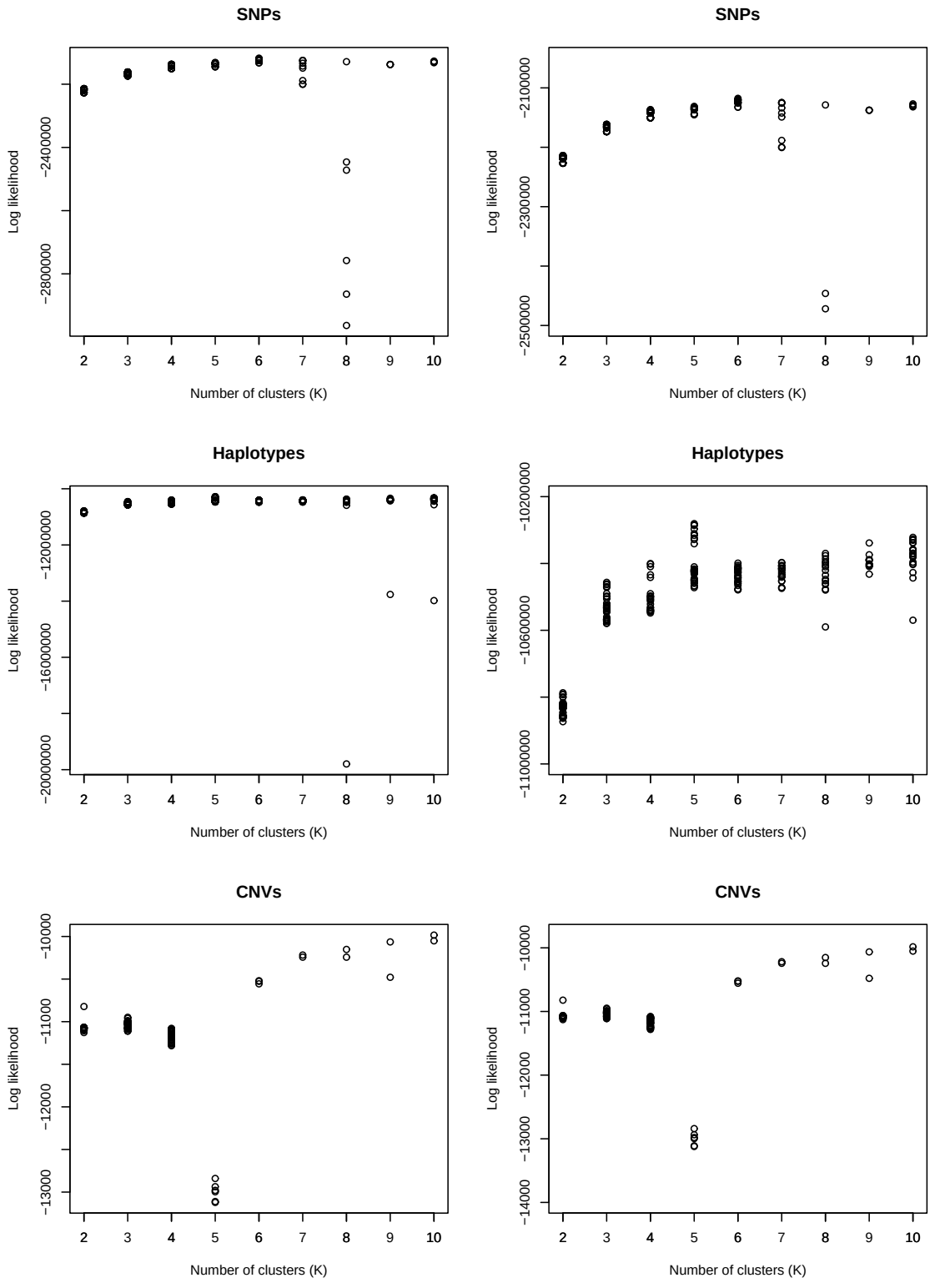

Figure S26: Log likelihood as a function of the number of clusters, based on `Structure` runs applied to the full worldwide dataset. Likelihoods are shown only for `Structure` runs in the most frequent mode. The plots on the right side are magnified versions of the plots on the left.

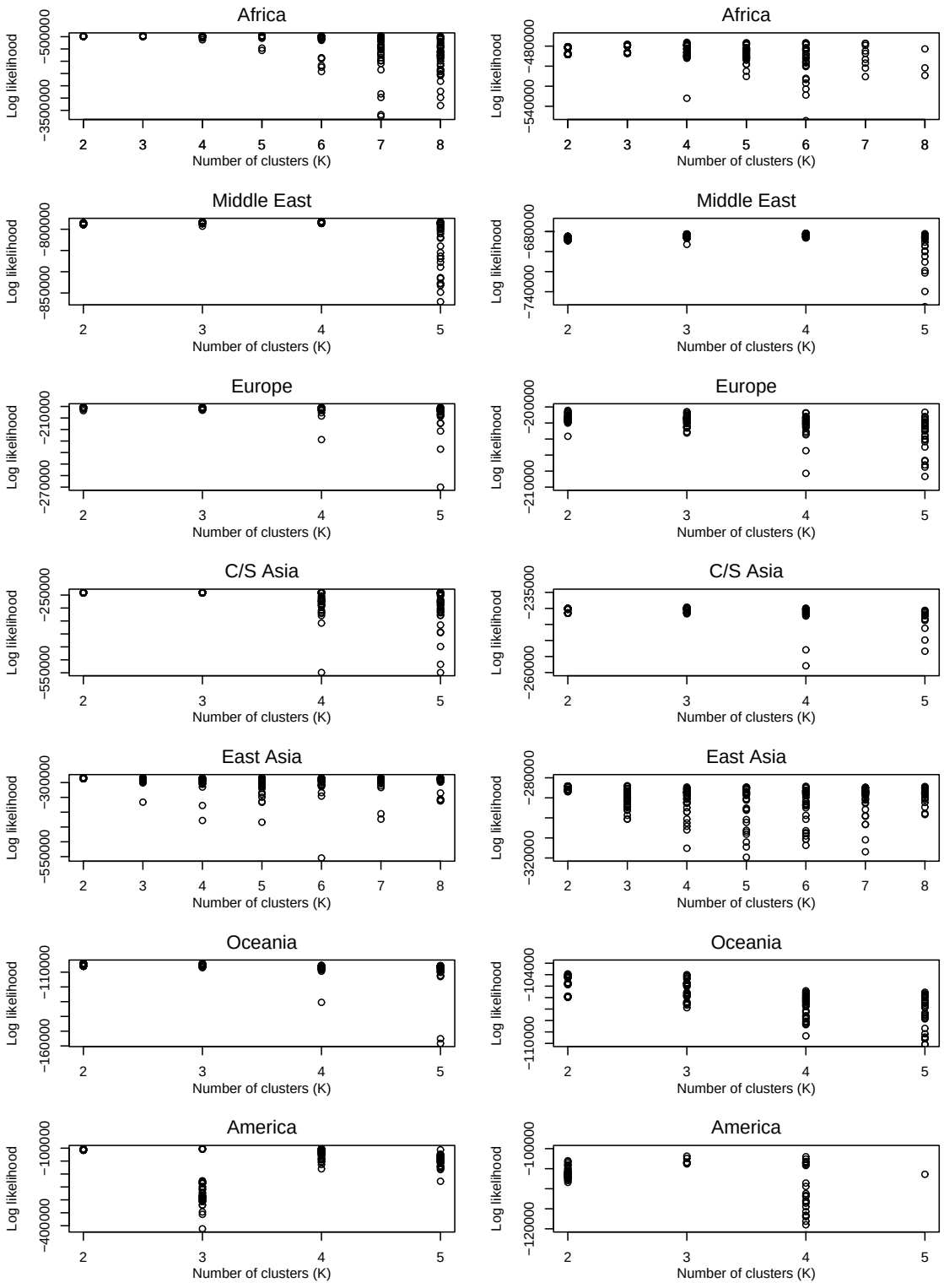

Figure S27: Log likelihood as a function of the number of clusters, based on `Structure` runs applied to individual geographic regions. Likelihoods for all 40 replicates at a given K are shown. The plots on the right side are magnified versions of the plots on the left.

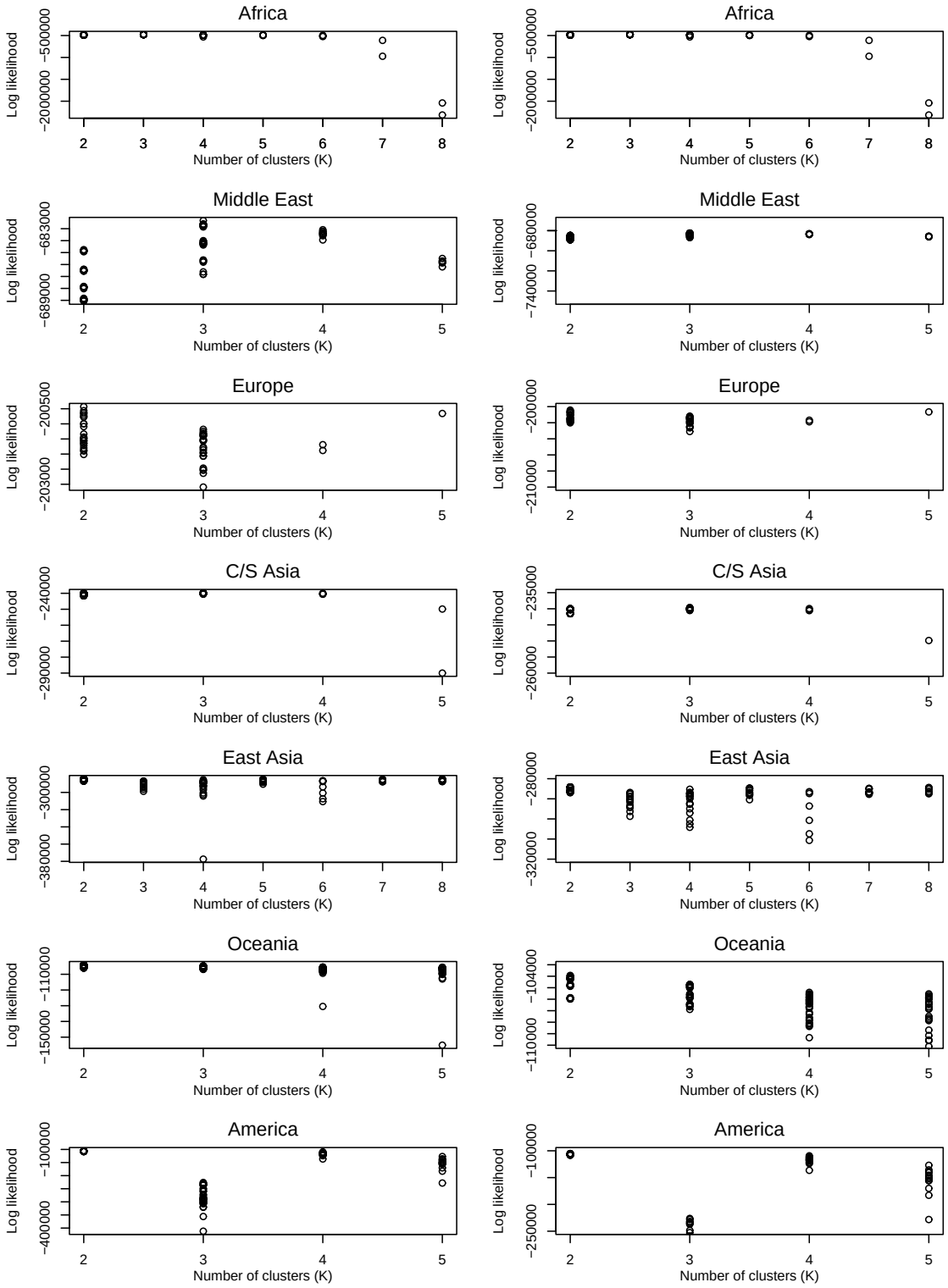

Figure S28: Log likelihood as a function of the number of clusters, based on `Structure` runs applied to individual geographic regions. Likelihoods are shown only for `Structure` runs in the most frequent mode. The plots on the right side are magnified versions of the plots on the left. For Europe with $K = 5$, there was no mode, and the point plotted corresponds to the single highest-likelihood run.

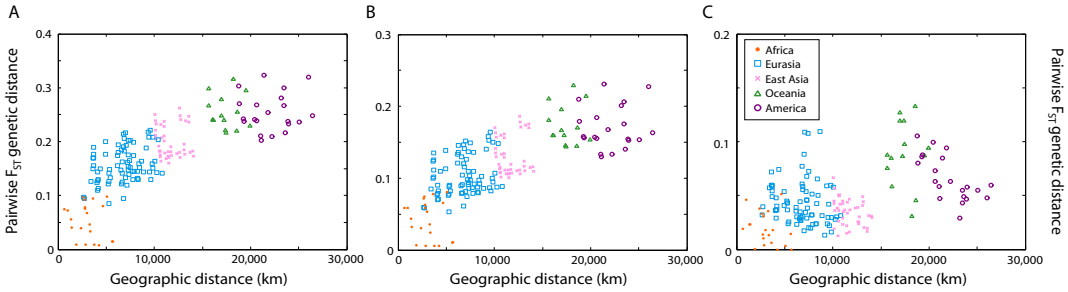

Figure S29: F_{ST} between pairs of populations, plotted as a function of geographic distance between the populations. Only population pairs that contain at least one African population are shown. (A) SNPs. The Pearson correlation between F_{ST} and geographic distance is 0.74 and the regression line is $F_{ST} = (9.057 \times 10^{-6})D + 0.045$, where D represents distance in kilometers. $R^2 = 0.54$. (B) Haplotypes. The Pearson correlation is 0.74, the regression line is $F_{ST} = (6.220 \times 10^{-6})D + 0.029$, and $R^2 = 0.54$. (C) CNVs. The Pearson correlation is 0.49, the regression line is $F_{ST} = (2.494 \times 10^{-6})D + 0.023$, and $R^2 = 0.24$. The plots indicate a reasonably close relationship between genetic distance and geographic distance for SNPs and haplotypes, and a weaker relationships for CNVs. The SNP plot is copied in Figure 2a.

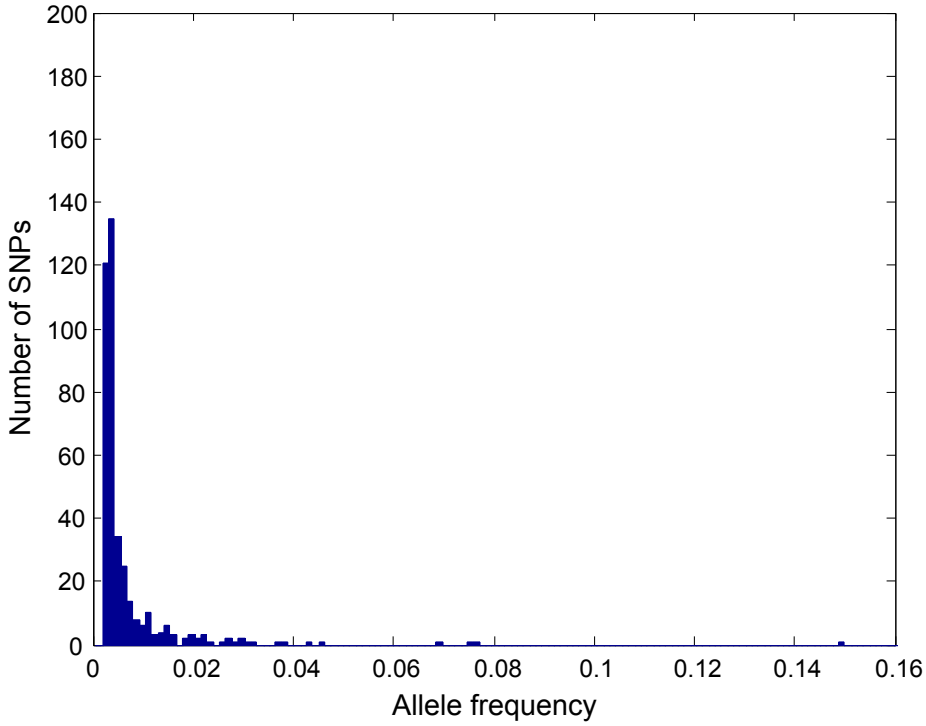

Figure S30: Allele frequency spectrum for each of five sets of 396 non-singleton SNPs in 405 unrelated individuals from 29 populations. The frequency spectrum is matched as closely as possible to the corresponding frequency spectrum of CNV loci in Figure S12.

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., Friedlaender, 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. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

4. Simon-Sanchez, J., Scholz, S., Fung, H.-C., Matarin, M., Hernandez, D., Gibbs, J. R., Britton, A., Wavrant de Vrieze, F., Peckham, E., Gwinn-Hardy, K., Crawley, A., Keen, J. C., Nash, J., Borgaonkar, D., Hardy, J., and Singleton, A. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16**, 1–14 (2007).

5. Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* **38**, 1251–1260 (2006).

6. Rosenberg, N. A. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* **70**, 841–847 (2006).

7. 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).

8. 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).

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

10. Sabatti, C. and Risch, N. Homozygosity and linkage disequilibrium. *Genetics* **160**, 1707–1719 (2002).

11. Rosenberg, N. A. and Blum, M. G. B. Sampling properties of homozygosity-based statistics for linkage disequilibrium. *Math. Biosci.* **208**, 33–47 (2007).

12. Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* **102**, 15942–15947 (2005).

13. Szpiech, Z. A., Jakobsson, M., and Rosenberg, N. A. ADZE: Allelic Diversity Analyzer version 1.0. http://rosenberglab.bioinformatics.med.umich.edu/adze.html, (2007).

14. Kalinowski, S. T. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conserv. Genet.* **5**, 539–543 (2004).

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

16. Falush, D., Stephens, M., and Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).

17. Rosenberg, N. A. DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).

18. Jakobsson, M. and Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).

19. Wang, S., Lewis Jr., C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M. W., Rosenberg, N. A., and Ruiz-Linares, A. Genetic variation and population structure in Native Americans. *PLoS Genet.* **3**, 2049–2067 (2007).

20. Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

21. Minch, E., Ruiz Linares, A., Goldstein, D. B., Feldman, M. W., and Cavalli-Sforza, L. L. MICROSAT (version 2.alpha): a program for calculating statistics on microsatellite data. Department of Genetics, Stanford University, Stanford, CA (1998).

22. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.65. Department of Genome Sciences, University of Washington, Seattle (2005).

23. Bryant, D. A classification of consensus methods for phylogenetics. In BioConsensus, Janowitz, M. F., Lapointe, F.-J., McMorris, F. R., Mirkin, B., and Roberts, F. S., editors, 163–183. American Mathematical Society, Providence (2003).

24. Mardia, K. V., Kent, J. T., and Bibby, J. M. *Multivariate Analysis.* Academic Press, London (1979).

25. Everitt, B. S. and Hothorn, T. *A Handbook of Statistical Analyses Using R.* Chapman & Hall/CRC, Boca Raton (2006).

26. 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).

27. Hudson, R. R. Linkage disequilibrium and recombination. In Handbook of Statistical Genetics, Balding, D. J., Bishop, M., and Cannings, C., editors, chapter 11, 309–324. Wiley, Chichester, UK (2001).

28. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., and Bucan, M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

29. Walter, S. D. and Irwig, L. M. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J. Clin. Epidemiol.* **41**, 923–937 (1988).

30. Hui, S. L. and Zhou, X. H. Evaluation of diagnostic tests without gold standards. *Stat. Methods Med. Res.* **7**, 354–370 (1998).

31. Enøe, C., Georgiadis, M. P., and Johnson, W. O. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.* **45**, 61–81 (2000).

32. Pepe, M. S. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford (2003).

33. Jakobsdottir, J. and Weeks, D. E. Estimating prevalence, false-positive rate, and false-negative rate with use of repeated testing when true responses are unknown. *Am. J. Hum. Genet.* **81**, 1111–1113 (2007).

34. Lynch, A. G., Marioni, J. C., and Tavaré, S. Numbers of copy-number variations and false-negative rates will be underestimated if we do not account for the dependence between repeated experiments. *Am. J. Hum. Genet.* **81**, 418–420 (2007).

35. Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., MacAulay, C., Ng, R. T., Brown, C. J., Eichler, E. E., and Lam, W. L. A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91–104 (2007).

36. Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., and Eichler, E. E. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).

37. Locke, D. P., Sharp, A. J., McCarroll, S. A., McGrath, S. D., Newman, T. L., Cheng, Z., Schwartz, S., Albertson, D. G., Pinkel, D., Altshuler, D. M., and Eichler, E. E. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).

38. Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).

39. Mountain, J. L. and Cavalli-Sforza, L. L. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* **61**, 705–718 (1997).

40. Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A. M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., and Weigend, S. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**, 699–713 (2001).

41. Felsenstein, J. *Inferring Phylogenies*. Sinauer, Sunderland, MA (2004).

42. Susko, E., Inagaki, Y., and Roger, A. J. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Mol. Biol. Evol.* **21**, 1629–1642 (2004).

43. Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15**, 538–543 (1998).

44. Nagylaki, T. Fixation indices in subdivided populations. *Genetics* **148**, 1325–1332 (1998).

45. Long, J. C. and Kittles, R. A. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* **75**, 449–471 (2003).

46. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).

47. Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., and Scherer, S. W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).