www.sciencemag.org/cgi/content/full/319/5866/1100/DC1

Supporting Online Material for

Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation

Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman,* Luigi L. Cavalli-Sforza,* Richard M. Myers*

*To whom correspondence should be addressed. E-mail: marc@charles.stanford.edu (M.F.); cavalli@stanford.edu (L.L.C.S.); myers@shgc.stanford.edu (R.M.M.)

Published 22 February 2008, *Science* **319**, 1100 (2008)
DOI: 10.1126/science.1153717

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/319/5866/1100/DC1)

Supporting Online Material

Table of Contents

1. Methods

1.1 Samples and quality control

The HGDP-CEPH panel consists of 1,064 individuals, of which 1,043 were successfully genotyped at call rates > 98.5% (based on GenCall Score of 0.25; Gencall is a quality score assigned to each genotype call based on how closely the A-B allele intensities agree with the canonical clustering patterns). We also genotyped two unrelated chimpanzee genomic DNA samples. Samples that failed the call rate cutoff were run a second time; most still failed, suggesting low DNA quality. The panel contains a number of individuals who are first- or second-degree relatives of others in the dataset. In all, there are 952 samples from unrelated individuals (i.e., the H952 set in (*1*), in which no pair is closer than first cousins). Of these, 938 are among the 1,043 that we successfully genotyped. Most of the analyses were based on these 938 samples, consisting of 615 males and 323 females.

The 938 individuals represent 51 populations, whose names, sample size, and sampling locations are listed in the Supplementary Table. The sampling locations are shown on a world map in (*2*). We combined the northern and southern Han Chinese as one population.

1.2. Marker selection

The markers on the Illumina HumanHap650K Beadchips were chosen to maximize "tagging" of additional common SNPs that are in linkage disequilibrium with the genotyped SNPs. According to data from the International HapMap Project, the 650,000 markers effectively tag more than 90% of SNPs with minor allele frequency above 5% in a European sample, and 88% and 67%, respectively, for East Asian and African samples. Because of the preferential selection of common variants in Africa, Europe, and East Asia, the allele frequency spectra and SNP heterozygosities across 51 populations are biased with the highest heterozygosities seen in Europe, Middle East, Central/South Asia and the three farming groups in Africa (Bantu, Yoruba, and Mandenka), followed by the three hunter-gatherer groups in Africa (San, Mbuti Pygmy, and Biaka Pygmy) and East Asia (see Supplementary Table). Because of this moderate degree of

ascertainment bias, we reported haplotype-based heterozygosities instead of SNP-based heterozygosities (Fig. 3B), and, in analyzing ancestral allele frequencies, we have compared our results with the ENCODE region data from the HapMap Project (Fig. S7).

1.3. Genotyping and SNP quality control

The genotyping experiments were carried out according to manufacturers' specifications. Of all the samples run in duplicate, 30 had call rate >98.5% on both chips. The rate of discordance among these 30 high-quality pairs was used to define a replication error rate for each SNP. Genotyping calls were made from the two-channel bead-type data, which are the averages of all beads for the same assay. For the canonical clustering positions we used (1) Illumina's default clustering file, which specifies the expected clustering positions of the three genotypes of all SNPs based on a set of ~110 samples run by Illumina, and (2) self-clustering based only on our samples. The final genotypes were determined by a hybrid approach: for each SNP, we compared its call rate and replication error between the two clustering methods, determined the better method for each SNP, and chose the calls made by that method. SNPs were filtered out if they contained any replication error or if their call rates across 938 samples were < 95%. In all, 1,642 markers (0.25%) failed QC, leaving 659,276 markers, of which 16,400 are on the X chromosome, 642,690 are on autosomal chromosomes, the rest are on ChrY or mtDNA. The average call rate among the 938 samples and 642,690 autosomal SNPs was 99.923%. There are 12,855 autosomal SNPs having worldwide minor allele frequency < 0.01.

1.4. Hardy-Weinberg Equilibrium

We tested Hardy-Weinberg Equilibrium (HWE) by the exact test (*3*) for the 51 populations separately. Among 642,690 autosomal loci, the number of SNPs with $P < 0.001$ varied in the range of 0-570, depending on the population, and showing a positive correlation with sample size ($r = 0.93$). The 19 populations with 10 or fewer samples have no SNPs with $P < 0.001$. This is because, with the exact test, a small sample size has low power to detect deviations from HWE. Across 51 populations, less than 1% of SNPs (maximum of 4,914 of 642,690) had $P < 0.01$, and less than 0.1% (maximum of 570 of 642,690) had $P < 0.001$. A majority of the SNPs that

showed small P values had lower-than-expected heterozygosity counts, suggesting some level of relatedness (i.e., baseline inbreeding). There are 5,227 SNPs having $P < 0.001$ in at least one population, of which only 110 have $P < 0.001$ in at least two populations. We did not remove these 110 from subsequent analysis, as they had a negligible impact on the results amidst the entire set of 642,690 SNPs. The population with the largest number (570) of $P < 0.001$ SNPs is the Bedouin, which we have shown to contain two subgroups (Fig. 1A and 2B). Summary statistics for HWE analysis are included in the Supplementary Table.

1.5. Chimpanzee sequences and definition of ancestral alleles

We genotyped two chimpanzee samples and combined the results with the ancestral allele definitions published by Voight et al. (from Haplotter homepage, see Ref. (*4*)). The consensus chimpanzee allele is defined for a SNP locus if both chimpanzee genotypes are homozygotes, agree with each other, and agree with the definition in (*4*). Of the 642,690 autosomal SNPs, we were able to define the consensus chimpanzee allele for 614,088 (95.5%).

For the ENCODE region analysis shown in Fig. S7, we obtained the ancestral alleles of all ENCODE SNPs by downloading the orthologous chimpanzee alleles from the UCSC Genome Browser. The SNPs in the ENCODE regions were discovered by sequencing 48 individuals: 16 Yorubans (YRI) from Ibadan, Nigeria, 16 European Americans from Utah, eight Chinese from Beijing (CHB) and eight Japanese from Tokyo (JPT).

1.6. Ancestry analysis

A number of MCMC- (*5, 6*) or likelihood-based methods (*7-10*) have been proposed for inferring the proportional ancestry of each individual from multi-locus genotype data without using prior knowledge of the geographic location (hence the likely clustering) of the samples. We inferred individual ancestry proportions for 938 samples by using the maximum likelihood method developed in (*9*), which is implemented with an EM algorithm in the program *frappe* (http://www.fhcrc.org/labs/tang/). All analyses are based on 642,690 autosomal SNPs. No prior ancestry information was assumed. The program was allowed to run for 10,000 iterations, with

pre-specified cluster numbers, from K=2 to 7. Empirically we observed convergence both in the estimated cluster membership (i.e., ancestry proportion) and in data likelihood. Independent runs yield consistent results (i.e., highly correlated estimates). With K=8 or above, additional clusters emerge, usually representing outlier populations in a certain region, but with lower consistency across independent runs.

Different HGDP populations differ in sample size, which affects (among other things) the observed linkage disequilibrium (LD) patterns. One may question the validity of the ancestry estimates, as *frappe* does not model linkage disequilibrium among SNPs. We show in (*9*) that as long as the markers are relatively evenly distributed along the chromosomes, the point estimates of ancestry are unbiased. In our simulation in that paper we used 120 clusters of tightly linked SNPs and found that not only are the point estimates unbiased, but also that the bias in confidence intervals is not severe (90% CI has coverage probability of 70-86%). In the present situation, if we assume that every 10 consecutive SNPs are in perfect LD, but LD=0 if the two SNPs are more than 10 apart, then only 0.004% of all possible pairs are in LD, and the overwhelming majority of pairs are not in strong LD. Even this is actually an overestimate of the levels of LD, as the Illumina panel eliminates most markers in such strong LD. Thus, we do not think the clustering in ancestry estimates is an artifact of LD. Additional discussions of the ancestry inference can be found in 2.2.

1.7. Fst calculation, molecular phylogeny, and PCA,

We calculated Fst's for all pairs of populations by using the population allele frequencies across all 642,690 autosomal SNPs. The algorithm adjusts for unequal sample size following (*11*). We performed a principal component analysis on the 51-X-51 Fst matrix to obtain Fig. S3A, C, and D. To construct the phylogenetic tree shown in Fig. 1B, we used the maximum likelihood method, CONTML, in the PHYLIP package, with allele frequencies at 150,000 SNPs, which is the maximal number of markers allowed by the software. We tested four non-overlapping sets of 150,000 markers and obtained nearly identical trees, one of which is shown in Fig. 1B. The Neighbor-Joining method yields a highly similar dendrogram. The branch lengths of the Neighbor-Joining tree (measured from the root) are correlated with estimated effective

population sizes (Spearman's $\rho = 0.82$, $P<0.0001$), suggesting that random genetic drift is a major contributor to the pattern of genetic variation that we observe.

We computed the Identity-by-State (IBS) matrix among the 938 individuals by using PLINK (*12*), producing a 938-by-938 matrix. We then performed a Principal Component Analysis on this IBS matrix for all samples and for seven regions separately, and used the top components to illustrate the genetic relatedness among individuals. We show the PC1-PC2 plots for all samples in Fig. S3B and six of the seven regions in Figs. 2 and S4.

1.8. AMOVA

We performed the analysis of molecular variance using Arlequin v3.11 (*13*) for the 22 autosomes and ChrX separately. The 938 individuals were assigned to two levels of subdivision: 51 populations, and 7 geographical regions, with the Uygur people assigned to Central/South Asia rather than East Asia. Un-phased genotype data and default Arlequin setting were used. For ChrX, we combined random pairs of males in each population to generate pseudo-females. For populations with an odd number of males, one male was discarded.

1.9. Haplotype heterozygosity

To determine haplotype heterozygosities, we used Haploview (*14*) to obtain haplotype frequencies for Chr16 and ChrX data in 51 populations separately. Chr16 was chosen arbitrarily as an example of an autosomal chromosome. As the natural haplotype blocks vary in length in different populations, we applied a common definition of haplotype boundaries across all populations by using a block definition file, generated from population-based recombination rates (*15*) that are available from the HapMap website (http://www.HapMap.org). As the HapMap data have more SNPs than we had, we calculated the cumulative recombination rates in every SNP interval in our panel by summing over the rates in corresponding HapMap intervals. In our panel, SNP intervals with recombination rates > 0.5 cM/Mb are considered low-LD (i.e., high recombination rate) regions, between which are the high-LD regions. There are 594 high-LD regions of varying length on Chr16, of which 295 have 5-15 SNPs. These 295 regions were

included in the block definition file so that comparisons of haplotype heterozygosities would be based on the same blocks. In each population, the haplotype frequencies were obtained for each block in Haploview, the expected haplotype heterozygosities were calculated by assuming HWE, averaged, and compared across populations (Fig. 3B) by plotting against geographic distance from Addis Ababa, Ethiopia, a hypothetical but plausible point of origin for human expansion (*16*). Choosing the complementary, low-LD regions or altering the recombination rate threshold from 0.5 to 2 did not significantly change the relationship between haplotype heterozygosity and distance from Addis Ababa, suggesting that analyzing additional regions from other chromosomes is unlikely to alter the main results seen with Chr16. For example, the slope in Fig. 3B would change from -1.198×10^{-5} (in units of km^{-1}, the same below) for 295 blocks of recombination rate < 0.5 to -1.237×10^{-5} for 146 blocks of recombination rate > 0.5, or to -1.144×10^{-5} for 323 blocks of recombination rate > 2, all blocks having 5-15 SNPs. ChrX results were similar to those for Chr16. Slope is -1.338×10^{-5} for 453 blocks of recombination rate < 0.5, each block having 2-15 SNPs. In comparison, microsatellite heterozygosities studied in (*16*) would have a slope of -0.65×10^{-5} and r = -0.97.

1.10. Recombination rates and effective population size

We used LDhat (*15*) to estimate population-based recombination rates for the first 20 1-Mb segments in Chr10 (arbitrarily chosen) for 51 populations separately. In running LDhat, we set the estimated starting value for $4N_e r$ to 400. The penalty for a change in recombination rate along a chromosome was set at 20. The number of iterations of the MCMC algorithm was set to 10^6. The "thinning interval" was set to 2,000 and the "burn-in" value was set to 50.

The effective population size N_e was estimated as in (*17*), except that we used the slope of simple linear regression (of population-based on family-based recombination rates) without the intercept term. The slopes were well estimated for most populations, as the standard errors are relatively small, and the P values for testing slope = 0 are less than 0.01 for all but Surui individuals (Supplementary Table), the population with the smallest N_e. These results indicate that analyzing additional genomic regions (beyond the 20 1M regions we have analyzed) is unlikely to significantly alter the N_e results The N_e estimates from this study are positively correlated with those in (*17*) with r = 0.84, and with the estimates from the first 20 1-Mb

segments in ChrX with r = 0.88. The N_e estimates from ChrX are about half of those for Chr10 (see Supplementary Table), indicating a smaller effective population size for ChrX. This is due to at least two factors: First, in gender-balanced populations, there are three X chromosomes for every four autosomes. Second, of the three X chromosomes, only two are in females and are recombining; therefore, at a given time, only half the number of X chromosomes as autosomes are undergoing recombination.

The effective population sizes show a steady reduction with geographic distance from Addis Ababa (see Supplementary Table), with a correlation of -0.75, consistent with the serial founder effect, which postulates that during each episode of expansion, a new colony was established by a small number of founders, resulting in a stepwise decrease of population size and concomitant reduction of genetic diversity, as measured by haplotype heterozygosity (Fig. 3B).

The ChrX analysis with LDHat requires diploid genotypes. We therefore combined random pairs of males to generate pseudo-females. For populations with an odd number of males, one male was discarded. This resulted in an effective reduction of sample sizes by about a third (not to be confused with the effective population sizes that we are estimating). We found that sample size has a relatively minor impact on the estimated N_e's, as was reported by Conrad at el (*17*). For Bedouin and Palestinian individuals, we tested sample sizes of 45, 40, …, 5, in steps of 5, and found that a one-third reduction in sample size is associated with an approximately 12% decrease of estimated N_e.

2. Supplementary description and discussions

2.1. Complementary methods for studying population structure

In this study, we analyzed genotype data from several different yet complementary perspectives: A global view of individual ancestry, genetic relatedness among populations, and fine-scale population structure. Historically, populations that have migrated away from each other often settle in geographically distinct regions, and the longer the time since divergence, the less genetically similar they tend to be. Consequently, genetic differentiation between populations

reflects both history and geography. Using the same genetic distance matrix, we constructed a phylogenetic tree (Fig. 1B) to represent the likely historical order of population divergence, and used the PCA plots to depict the genetic relatedness among the populations (Fig. 2, also Figs. S3-5). The tree supports the consensus model of human migration, while the PCA plots often resemble the true geographic distribution of the sampled locations. The interpretation of genetic diversity can be carried out not only at the level of regions or continents, as has been done previously, but also among individuals from closely related populations because of the increased number of markers analyzed in this study. Unlike the population dendrogram, which classifies 51 populations based on their inferred order of divergence assuming no horizontal gene flow, our analysis of ancestry allows each individual to be represented by proportional contributions from multiple sources, producing the most sensitive inference to date of individual ancestry (Figs. 1A and 2). Individuals from different populations can often be distinguished, including highly similar ones such as Han Chinese recruited in northern China versus those recruited in the US (who are mostly southern and central Chinese), Bantus in Kenya versus those in South Africa, two subgroups of Bedouins, as well as Pathan versus Sindhi, and Brahui versus Makrani. In comparison, substructure was identified in the sub-Saharan African, American, and Oceanic populations with 387 microsatellite markers, but it was much more difficult to detect finer subdivisions of Eurasia and East Asia (*18*).

2.2 Additional notes on ancestry analysis

Frappe analysis reveals that, at K = 7 and with a 2% threshold, 21 of the 51 populations derived ancestry from at least two ancestral components. In Figure 1A, the Mozabite from the northern Sahara bear contributions from sub-Saharan Africa, the Middle East, and Europe; this group in fact originates from the Middle East. In Europe, only the Adygei, who live to the north of the Caucuses, have a significant South/Central Asian component, whereas the Russian individuals have minor contributions from South/Central Asia, East Asia, and America. In the Middle East, a small subset of the Bedouins appears to have substantially higher Middle Eastern ancestry than the Palestinians, Druze and the other Bedouins.

In addition to analyzing the entire HGDP populations, we also applied *frappe* to subsets of individuals from two geographic regions: East Asia and South/Central Asia. Using the same

individuals but with far fewer markers, previous analyses were not able to describe finer level genetic structure (18). Fig. S2A shows that populations in East Asia with high northern ancestry include Mongola, Oroqen, Hezhen, Daur, Tu, Xibo, and Japanese. These groups reside in high latitude areas and speak languages of the Altaic family (*17*). In contrast, the right-hand side of the figure includes populations with lowest northern ancestry, Dai, Lahu and Cambodian, who live in or near southwestern China. The Han and northern Han Chinese can be distinguished, although the former is most likely a mixture of southern and more central individuals. Naxi and Yi are from the Yunnan Province in Southwest China, but unlike other southern populations to the east (She and Miao), they have high northern ancestry, possibly due to their shared ancestry with the nomadic Qiang, an ethnic group from the Tibetan plateau. The membership coefficients in Fig. S2A for Han and northern Han are correlated with PC1 of the same individuals in Fig. S5A with Pearson's r = 0.97.

Fig. S2B shows a second example of within-group ancestry analysis. At K=3, South/Central Asian individuals fall into several genetic clusters that agree with the predefined populations. First, the Kalash were previously observed to be genetically isolated (*18*), and the Hazara likely separated due to strong East Asian ancestry. Second, groups with relatively low East Asian ancestry (Makrani, Balochi, and Brahui) are distinguished from those with moderate-to-high levels of East Asian ancestry (Burusho, Pathan and Sindhi). Third, Burusho individuals can be differentiated from the Pathan and Sindhi.

It is important to emphasize that the ancestry proportions inferred from this analysis are affected by the populations used in the study, as well as by the markers analyzed. If one of the continental groups had not been represented in the DNA panel, its contribution to some mixed populations might not have been estimated accurately. Conversely, had there been large samples from genetic isolates, our analysis might have singled out these isolates as separate clusters. Despite these caveats, the HGDP-CEPH panel has extensive coverage of the world's major geographic regions. Extensive and rigorous analyses have demonstrated that the estimated genetic clusters are not artifacts of non-continuous sampling of people (*19*).

The genetic markers used also have an impact on the ancestry analysis. For example, when we restricted *frappe* analysis to a subset of markers that were chosen to have high heterozygosities within Africans, our seventh cluster, instead of separating out the Middle Eastern populations (as

occurs when all 650K SNPs were used), distinguished San and Pygmy populations from the Bantu and Yoruba populations. Taken as a whole, however, the 650K SNP panel is relatively unbiased: even though it preferentially contains SNPs having high minor allele frequencies in African, European, and Asian populations, most SNPs do not have heterozygosities that are particularly high or low in only one or a few populations. We found that individual ancestry can be robustly estimated from the genetic data presented here, is consistent with population labels that are derived primarily from historical and cultural facts, and can lead to biological insights when interpreted carefully.

2.3. Additional notes on the PCA plots

Individuals from sub-Saharan Africa are seen to separate into seven clusters, entirely in accordance with the seven pre-defined populations (Fig. S4A). This is true for closely related ones that have been difficult to separate in previous studies, such as Yoruba and Mandenka, or Bantu individuals from South Africa and Kenya. Fig. S4B shows 17 East Asian populations, with the Yakut (on the left) driving the first PC. The three southern-most populations, Cambodian, Lahu and Dai, are depicted on the upper right quadrant. The individuals in the lower right quadrant are aligned such that the northern populations (such as Mongolia, Hezhen and Oroqen) are on the lower left, and the southern ones (such as Han, Miao and She) on the upper right. Northern and Southern Han Chinese can be distinguished if Han are analyzed separately (Fig. S5A, see also S2A). In Fig. S4C, South/Central Asian populations are shown with Kalash on the left, Hazara and Uygur to the lower right. The rest of the populations fall into three clusters: from the middle to the top, Burusho, Pathan-Sindhi, and Balochi-Brahui-Makrani, all in good agreement with the ancestry analysis (Fig. S2B). Pathan and Sindhi can be further distinguished if this sub-cluster is analyzed separately (Fig. S5B). Likewise, Brahui can be separated from Makrani, but Balochi individuals overlap with either Brahui or Makrani (Fig. S5C). In Fig. S4D, the five Native American populations are resolved, with Surui and Karitiana as outlying groups, and a single Colombian individual falling into the Mayan cluster. Not included is the Oceanian group, where Melanesian and Papuan individuals are clearly distinguishable from each other and from other populations.

It is not surprising that HGDP-CEPH samples tend to aggregate in PCA plots according to geographical proximity, as physical distance between the sampling points of two populations is one of the strongest factors influencing their genetic distance (*16*), because migration between continents has been infrequent for much of the human history (*19*). Among the 51 populations, the Mantel correlation coefficient between Fst for our SNPs and great circle geographic distance (incorporating waypoints to reflect the likely paths of migrations) (*16*) is 0.79 ($P < 0.01$). Two other measures of genetic distance, PSA and Nei's genetic distance, both yield a correlation of 0.74 ($P < 0.01$) with geographic distance. By comparison, the pairwise genetic distances based on microsatellite data are correlated with geographic distances at 0.89, 0.82, and 0.83, for Fst, PSA, and Nei's distance, respectively.

2.4. Further interpretations of the phylogenetic tree

Fig. 1B shows that in Africa, the three hunter-gatherer populations (Biaka Pygmies, Mbuti Pygmies, and San) can be separated from the three Bantu groups (Yoruba, Mandenka, and Bantu) that acquired farming about 3000 years ago. The north-African group, Mozabite, is intermediate between its nearest African and Middle East-Europe-South/Central Asia groups, reflecting its unique migration history. In South/Central Asia, the Kalash constitutes an outlier, in agreement with previous results (*18*) and the isolated nature of this ethnic group. Pathan and Sindhi are on adjacent branches, in accordance with their shared history and the fact that they both speak Indo-European languages. Hazara from Pakistan and Uygur from Xinjiang in northwestern China are close genetic neighbors (they correspond to the two overlapping dots at the top of Fig. S3A, intermediate between the Eurasian and East Asian clusters), suggesting a common ancestry despite their current geographic separation. This can be seen more clearly in the direct analysis of ancestry components (Fig. S2B).

The branching orders are mostly consistent across different tree construction methods (UPGMA, Minimal-Evolution, Neighbor-Joining, maximum likelihood) and different genetic distance measures (Proportion of Shared Alleles (PSA), Fst, Nei's distance), with minor differences involving the arrangements of the Middle Eastern, European, and Central/South Asian populations. The branch lengths, on the other hand, are determined by genetic distances, which, in this case, are Fst's summarized over all 642,690 autosomal SNPs. The branch lengths thus

depict the degree of separation between populations and, under the assumption of constant mutation rates in all branches, provide a rough estimate of the divergence time between lineages. The assumption of constant rate, however, may not be met, as different populations may have experienced different demographic dynamics, and hence have different effective population sizes. A smaller population, for example, tends to have a higher rate of random genetic drift and higher likelihood of fixation. Continuing admixture between two populations, on the other hand, will lead to a shortening of branch lengths between them. We calculated the effective population size by contrasting the population-based recombination rates and the pedigree-based meiotic recombination rates (*17*), and confirmed that the Eurasian populations tend to have smaller effective sizes. As expected, these populations have longer branch lengths in Fig. 1B. Spearman's correlation between effective population sizes and branch lengths is 0.82. Population size can account for nearly three-fourths of the branch length variation, with other factors, such as selection, migration, and admixture, probably contributing to the rest. Note that the genetic distance also depends on the markers used; the panel of more than 640,000 SNPs we used represents neither a complete nor a random set, as they are biased towards common polymorphisms discovered in African, European and East Asian populations. Had we added the chimpanzee branch to the diagram, its length would have been underestimated because our marker set is enriched for variant loci within humans as compared to divergent loci between humans and chimpanzees.

2.5. AMOVA results and autosome-ChrX comparison

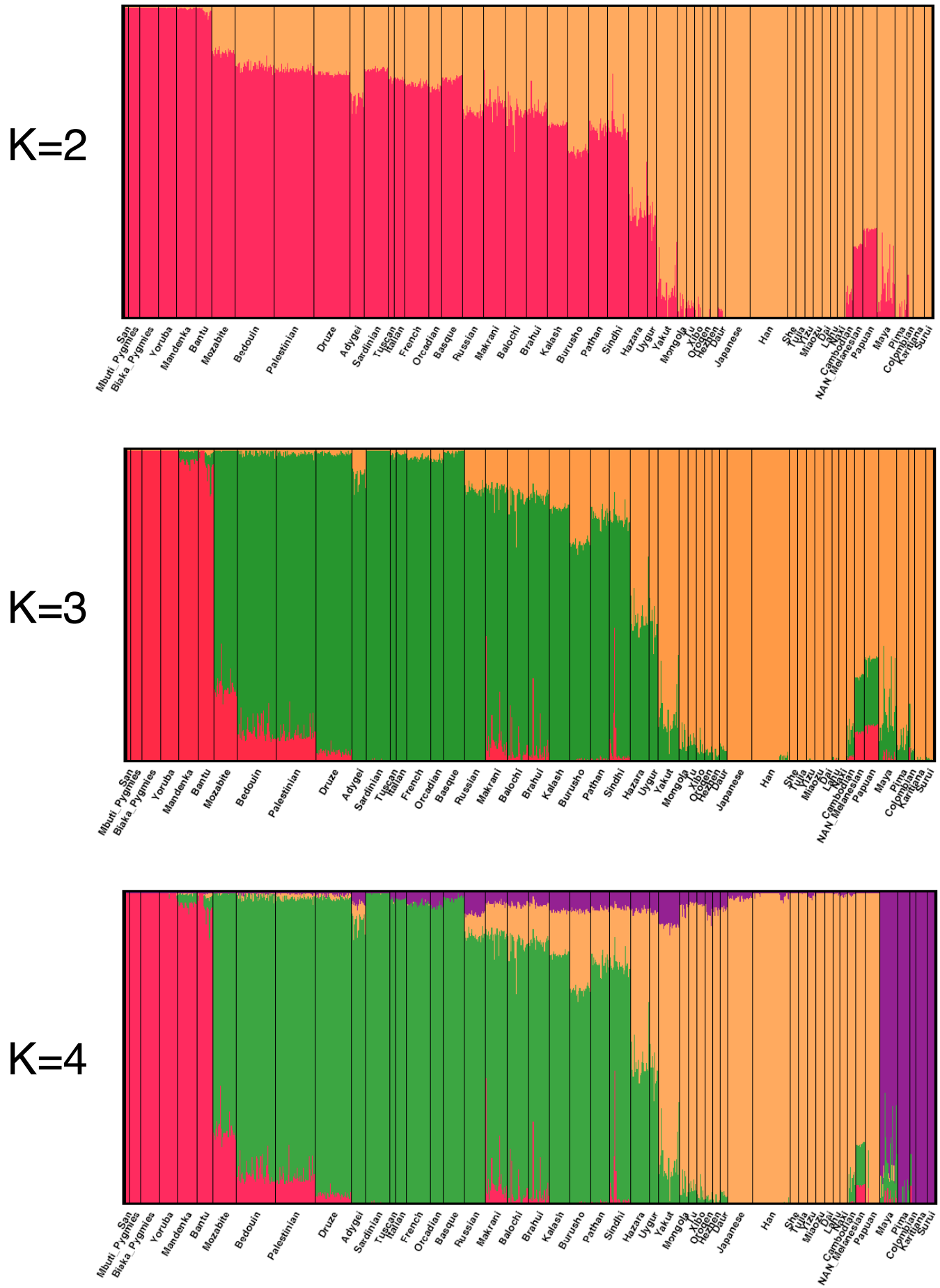
Compared to autosomes, the greater AG component for ChrX (Fig. 3A) relates to its smaller effective population size (i.e., stronger drift), lower mutation and recombination rates, and greater selective pressure in males (20), although the relative contributions of these different evolutionary forces are difficult to determine. Male-female differences in demographic parameters may also play a role, as females often have a higher rate of migration, shorter generation time, and lower reproductive variability (21-23). For ChrX, the total amount of population differentiation is likely higher, and the size of SNP discovery panels smaller, than for autosomes. Data from the 51 populations show that the effective population size of ChrX based on population recombination rates is 50% of that of autosomes (see Supporting Table). Fst's are

~52% higher for ChrX than for autosomes, indicating increased population differentiation for the former (Fig. S6) (19, 24). The ChrX-autosome differences are larger between sub-Saharan African and non-African populations (the red dots in Fig. S6) than among non-African populations (the black dots in Fig. S6). Among the latter, ChrX Fst's are only 26% higher than autosomal Fst's. The reasons for a higher African than Non-African differentiation for ChrX are complex, and may involve male-specific bottlenecks and/or selective pressure around the time of migration out of Africa.

References

1. N. A. Rosenberg, *Ann. Hum. Genet.* **70**, 841 (2006).
2. L. L. Cavalli-Sforza, *Nat. Rev. Genet.* **6**, 333 (2005).
3. J. E. Wigginton, D. J. Cutler, G. R. Abecasis, *Am. J. Hum. Genet.* **76**, 887 (2005).
4. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol.* **4**, e72 (2006).
5. D. Falush, M. Stephens, J. K. Pritchard, *Genetics* **164**, 1567 (2003).
6. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).
7. S. Purcell, P. Sham, *Hum. Hered.* **58**, 93 (2004).
8. G. A. Satten, W. D. Flanders, Q. Yang, *Am. J. Hum. Genet.* **68**, 466 (2001).
9. H. Tang, J. Peng, P. Wang, N. J. Risch, *Genet. Epidemiol.* **28**, 289 (2005).
10. X. Zhu, S. Zhang, H. Tang, R. Cooper, *Hum. Genet.* **120**, 431 (2006).
11. J. Reynolds, B. S. Weir, C. C. Cockerham, *Genetics* **105**, 767 (1983).
12. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, *Am. J. Hum. Genet.* **81**, 559 (2007).
13. L. Excoffier, G. Laval, S. Schneider, *Evolutionary Bioinformatics Online* **1**, 47 (2005).
14. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, *Bioinformatics* **21**, 263 (2005).
15. G. A. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, *Science* **304**, 581 (2004).
16. S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, *Proc. Natl. Acad. Sci. U S A* **102**, 15942 (2005).
17. D. F. Conrad, M. Jakobsson, G. Coop, X. Wen, J. D. Wall *et al.*, *Nat. Genet.* **38**, 1251 (2006).
18. N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, *Science* **298**, 2381 (2002).
19. N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard *et al.*, *PLoS Genet.* **1**, e70 (2005).

Figure S1


K=2
K=3
K=4
San
Mbuti_Pygmies
Biaka_Pygmies
Yoruba
Mandenka
Bantu
Mozabite
Bedouin
Palestinian
Druze
Adygei
Sardinian
Tuscan
Italian
French
Orcadian
Basque
Russian
Makrani
Balochi
Brahui
Kalash
Burusho
Pathan
Sindhi
Hazara
Uygur
Yakut
Mongola
Xibo
Oroqen
Hezhen
Daur
Japanese
Han
She
Tujia
Yizu
Miao
Dai
Lahu
Naxi
Cambodian
NAN_Melanesian
Papuan
Maya
Pima
Colombian
Karitiana
Surui

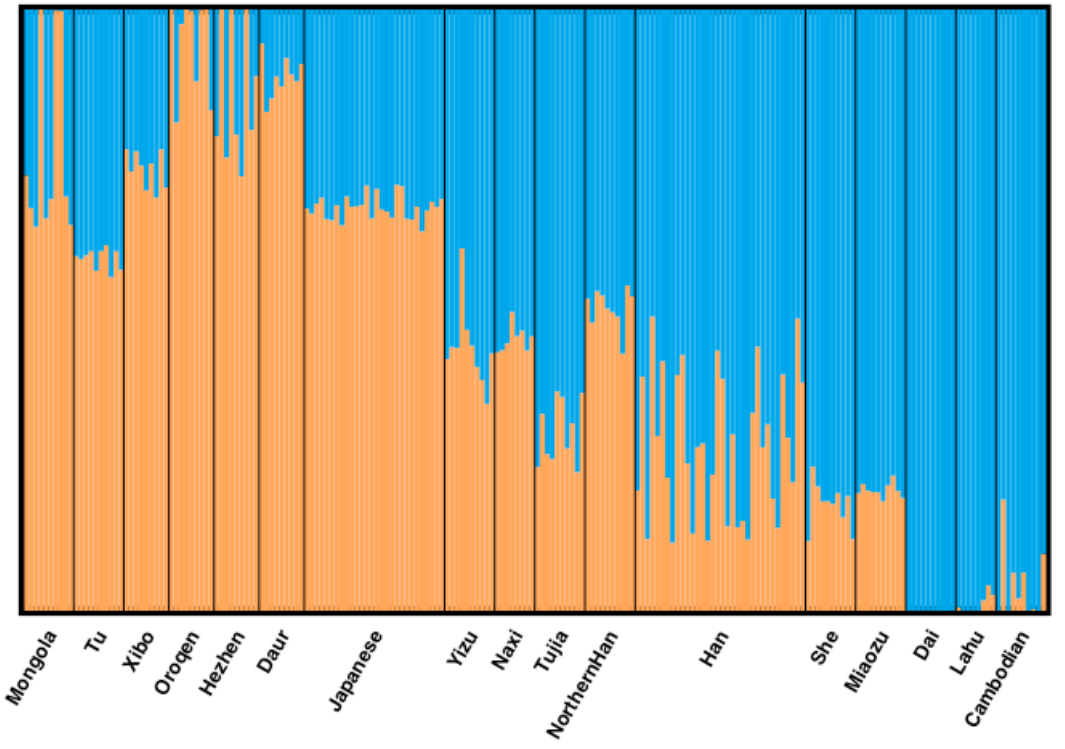
K=5
San
Mbuti_Pygmies
Biaka_Pygmies
Yoruba
Mandenka
Bantu
Mozabite
Bedouin
Palestinian
Druze
Adygei
Sardinian
Tuscan
Italian
French
Orcadian
Basque
Russian
Makrani
Balochi
Brahui
Kalash
Burusho
Pathan
Sindhi
Hazara
Uygur
Yakut
Mongola
Xibo
Oroqen
Hezhen
Daur
Japanese
Han
She
Tujia
Miaozu
Dai
Lahu
Cambodian
NAN_Melanesian
Papuan
Maya
Pima
Colombian
Karitiana
Surui

K=6
San
Mbuti_Pygmies
Biaka_Pygmies
Yoruba
Mandenka
Bantu
Mozabite
Bedouin
Palestinian
Druze
Adygei
Sardinian
Tuscan
Italian
French
Orcadian
Basque
Russian
Makrani
Balochi
Brahui
Kalash
Burusho
Pathan
Sindhi
Hazara
Uygur
Yakut
Mongola
Xibo
Oroqen
Hezhen
Daur
Japanese
Han
She
Tujia
Miaozu
Dai
Lahu
Cambodian
NAN_Melanesian
Papuan
Maya
Pima
Colombian
Karitiana
Surui

Figure S2

A

K=2



B

K=3

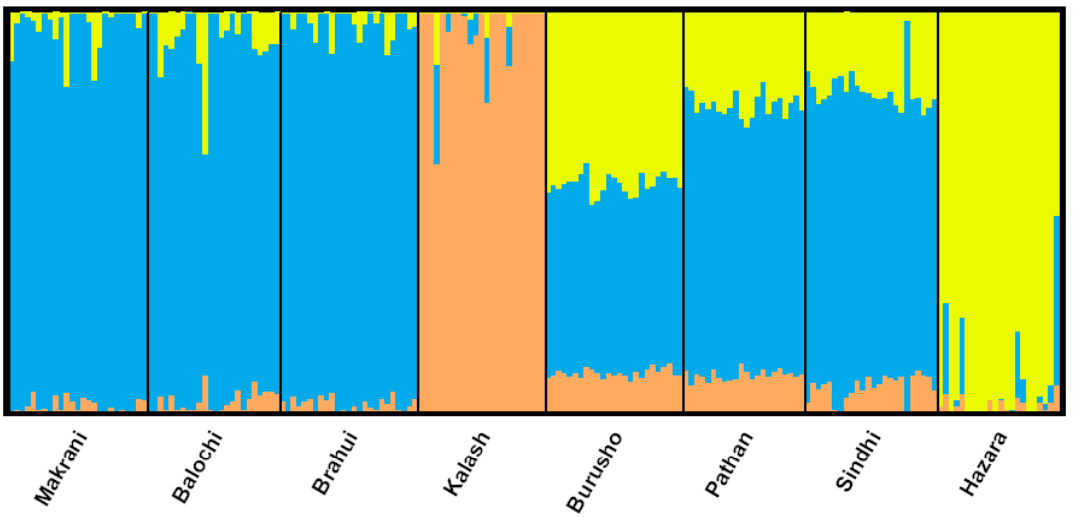


Figure S3

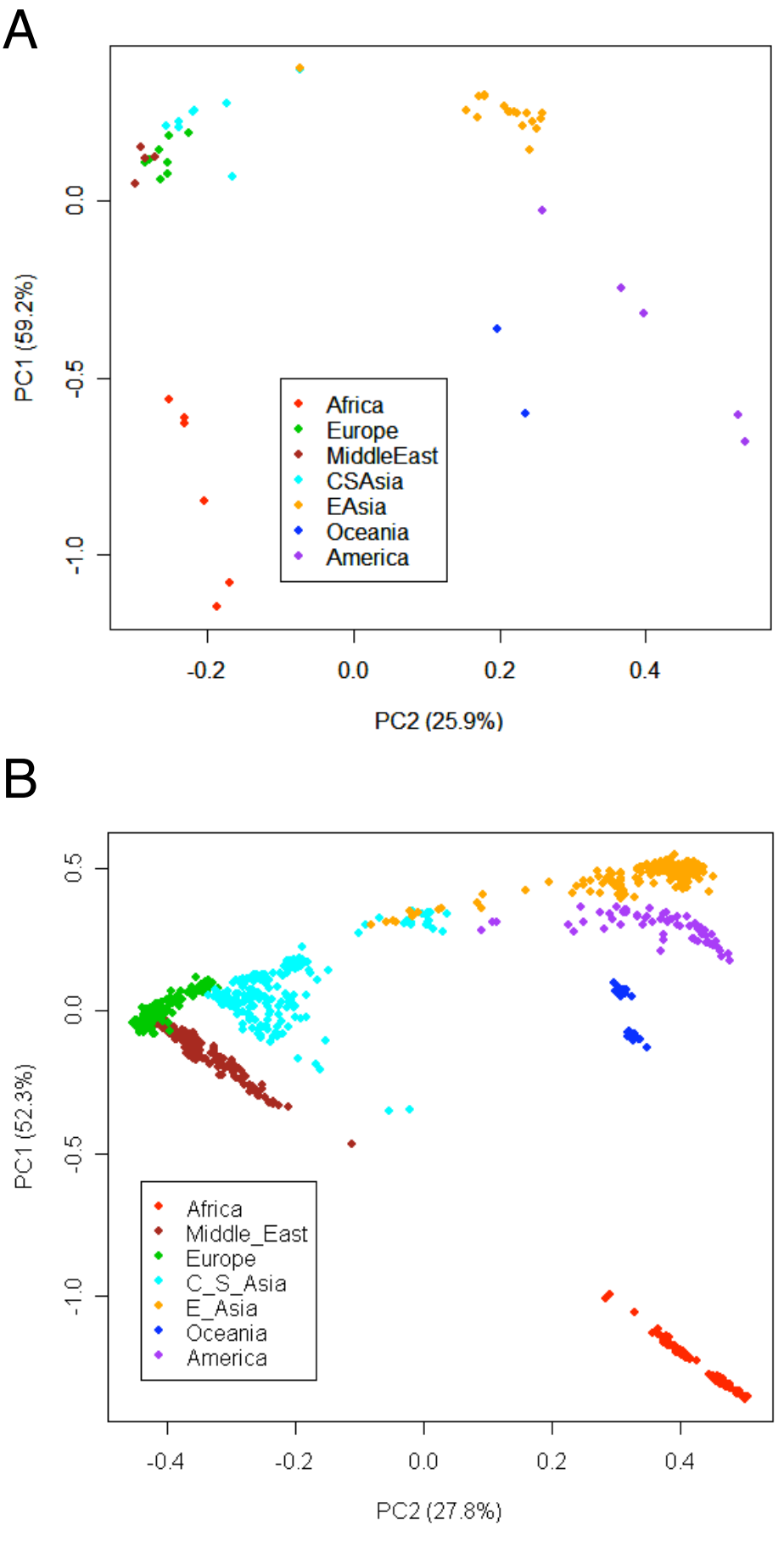
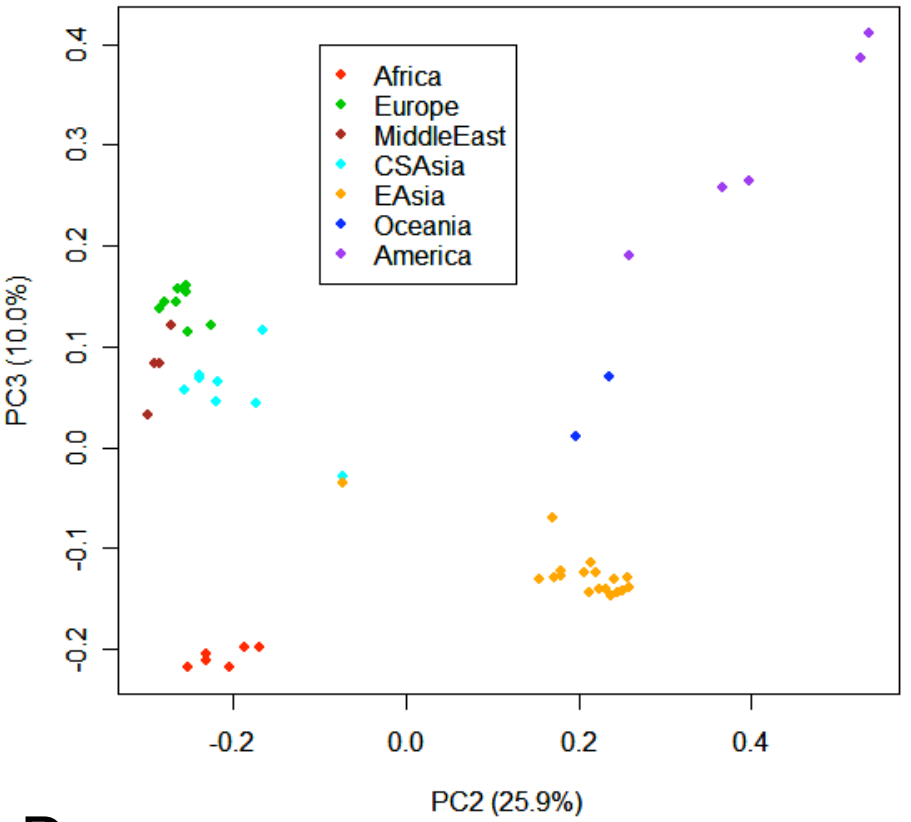

A
PC1 (59.2%)
0.0
-0.5
-1.0
-0.2
0.0
0.2
0.4
PC2 (25.9%)
Africa
Europe
MiddleEast
CSAsia
EAsia
Oceania
America
B
PC1 (52.3%)
0.5
0.0
-0.5
-1.0
-0.4
-0.2
0.0
0.2
0.4
PC2 (27.8%)
Africa
Middle_East
Europe
C_S_Asia
E_Asia
Oceania
America

Figure S3

C


Africa
Europe
MiddleEast
CSAsia
EAsia
Oceania
America
PC3 (10.0%)
PC2 (25.9%)
-0.2
0.0
0.1
0.2
0.3
0.4

D

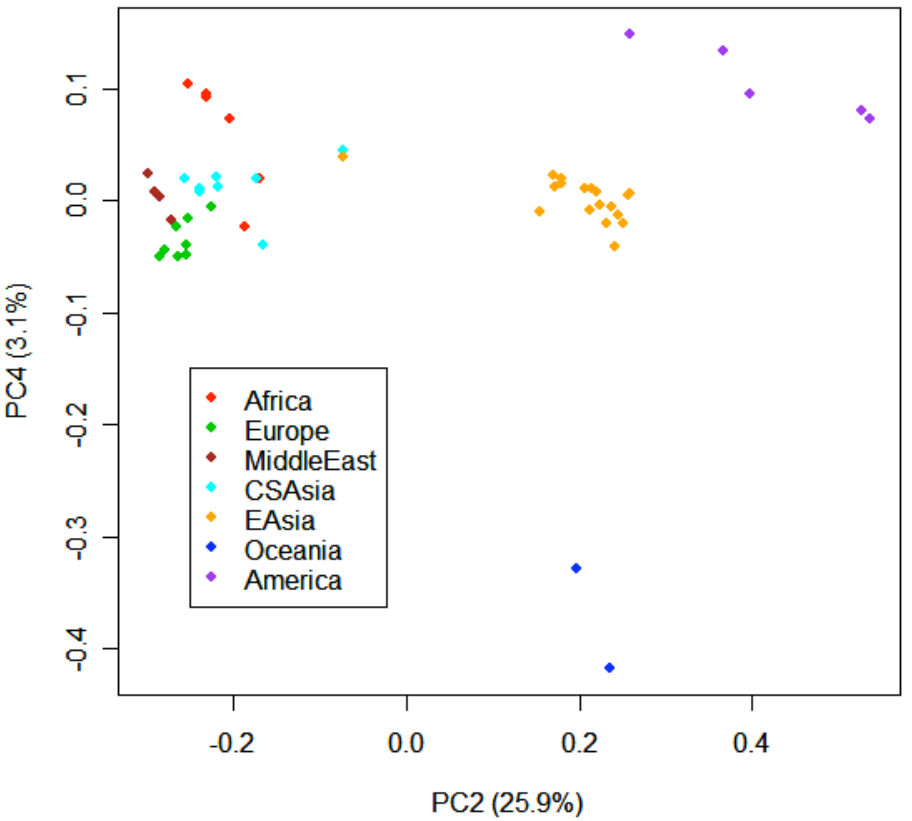

Africa
Europe
MiddleEast
CSAsia
EAsia
Oceania
America
PC4 (3.1%)
PC2 (25.9%)
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.4

Figure S4

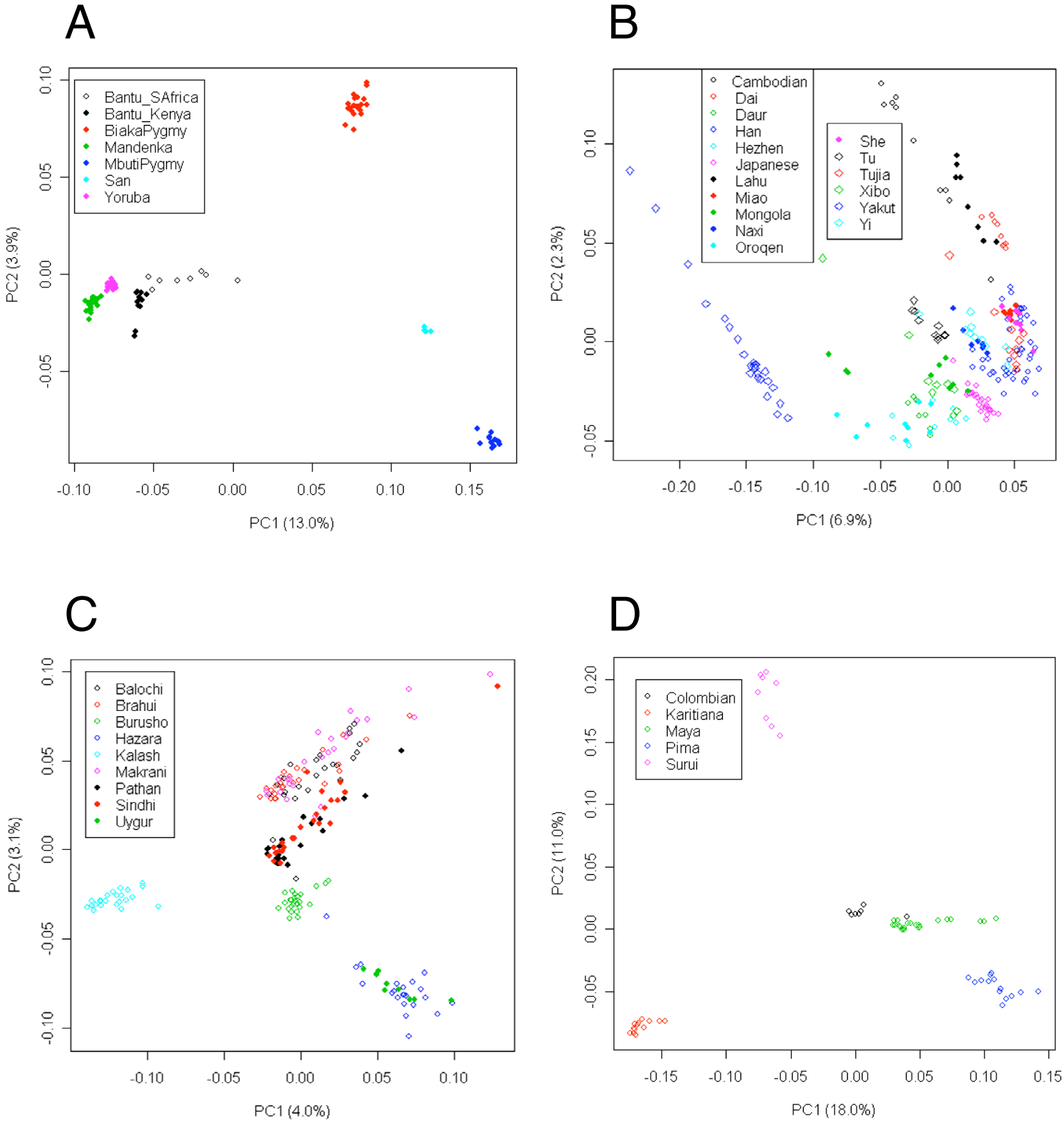

A
Bantu_SAfrica
Bantu_Kenya
BiakaPygmy
Mandenka
MbutiPygmy
San
Yoruba
PC1 (13.0%)
PC2 (3.9%)
B
Cambodian
Dai
Daur
Han
Hezhen
Japanese
Lahu
Miao
Mongola
Naxi
Oroqen
She
Tu
Tujia
Xibo
Yakut
Yi
PC1 (6.9%)
PC2 (2.3%)
C
Balochi
Brahui
Burusho
Hazara
Kalash
Makrani
Pathan
Sindhi
Uygur
PC1 (4.0%)
PC2 (3.1%)
D
Colombian
Karitiana
Maya
Pima
Surui
PC1 (18.0%)
PC2 (11.0%)

Figure S5

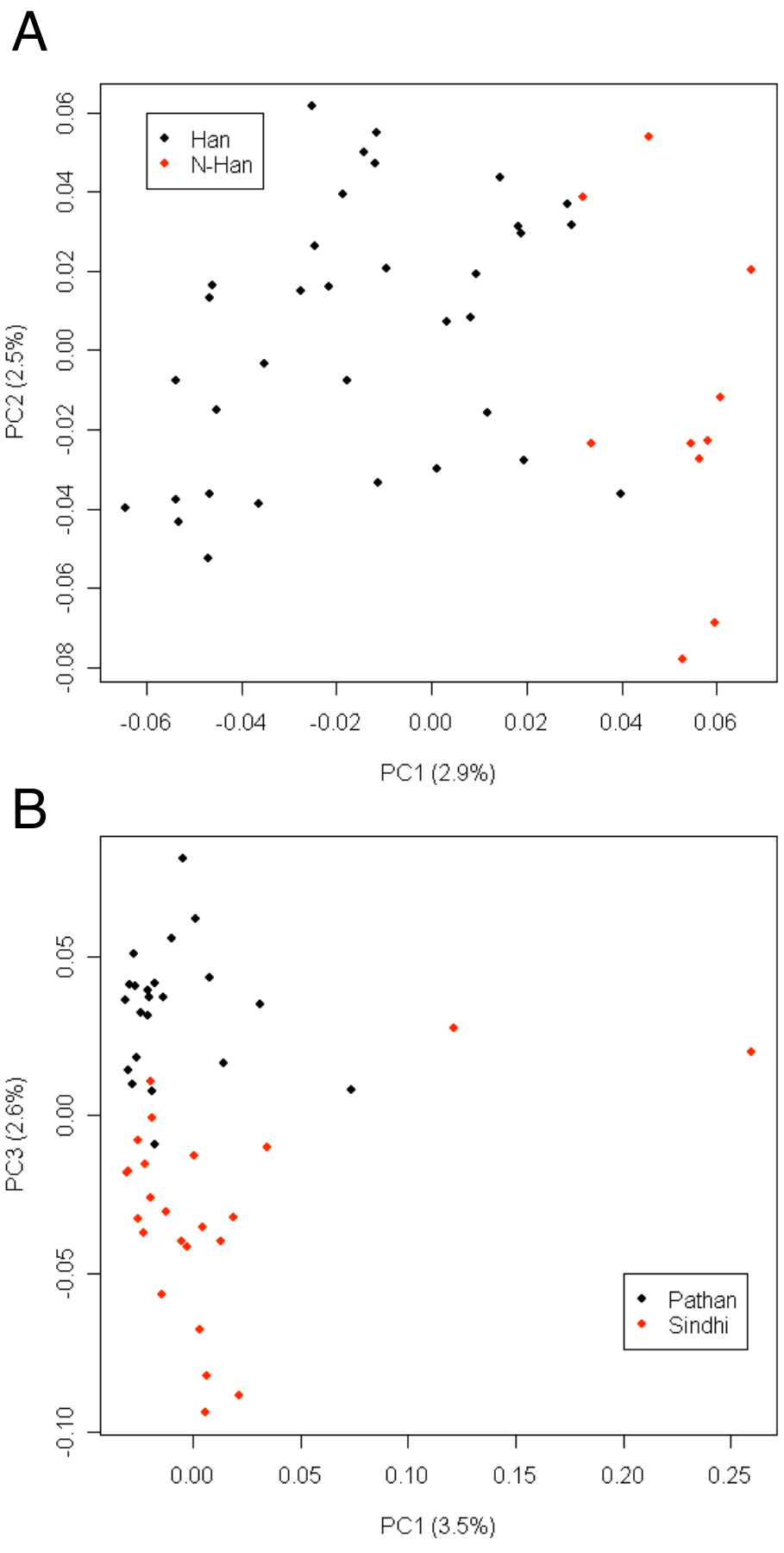

A
Han
N-Han
PC2 (2.5%)
PC1 (2.9%)
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
B
Pathan
Sindhi
PC3 (2.6%)
PC1 (3.5%)
0.05
0.00
-0.05
-0.10
0.00
0.05
0.10
0.15
0.20
0.25

Figure S5

C

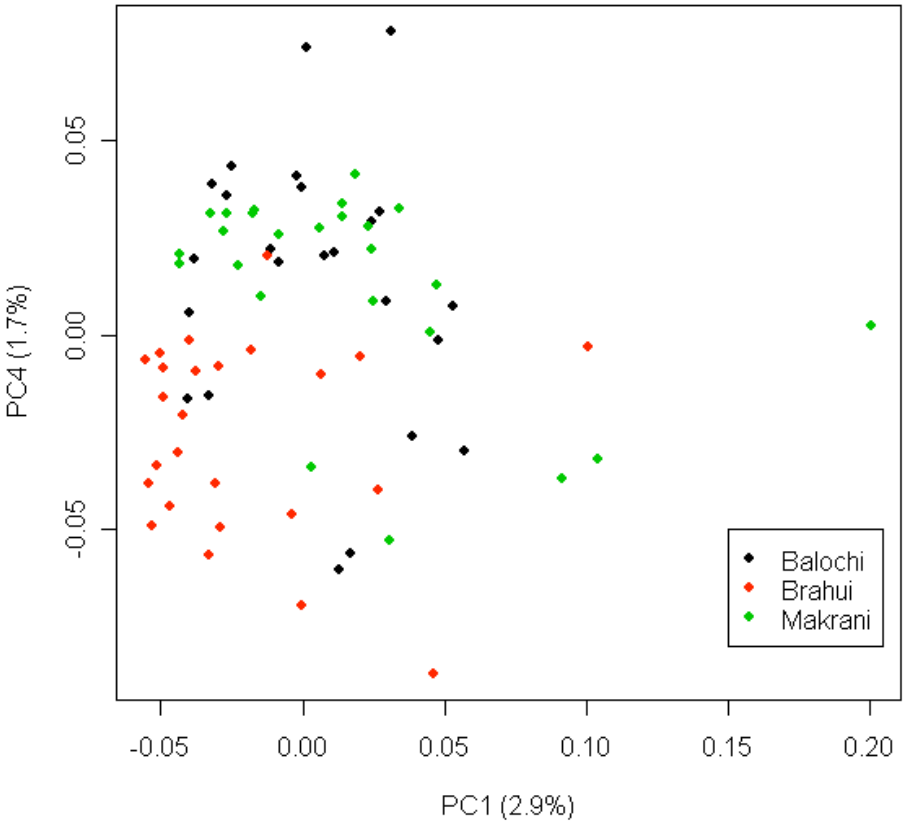

PC4 (1.7%)
0.05
0.00
-0.05
-0.05
0.00
0.05
0.10
0.15
0.20
PC1 (2.9%)
Balochi
Brahui
Makrani

Figure S6

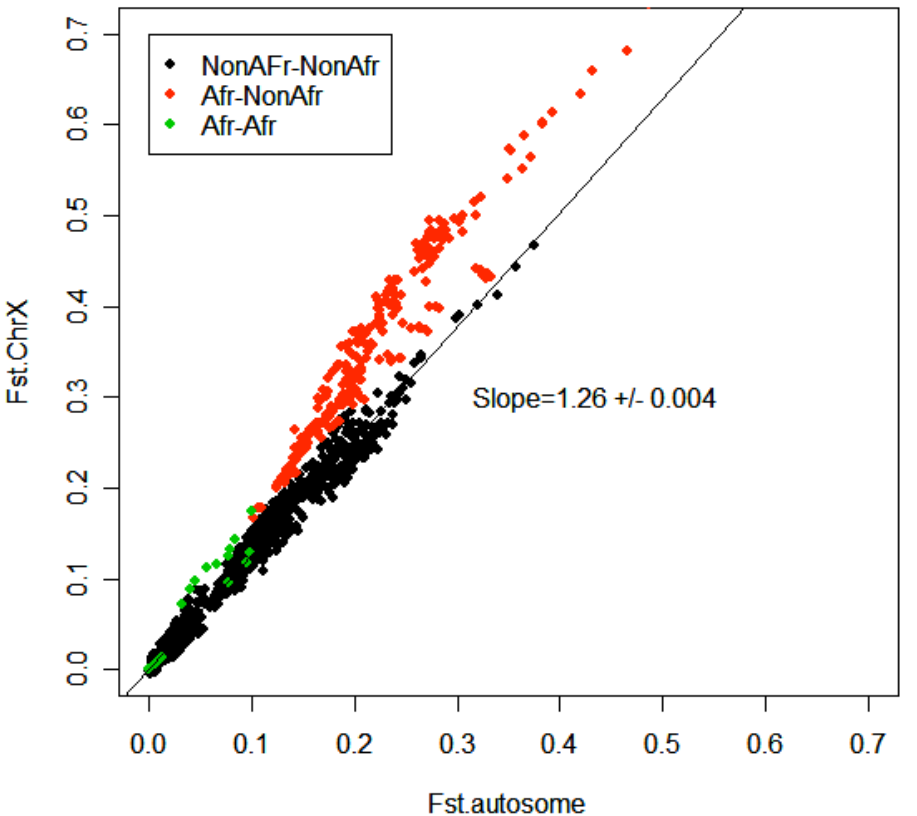
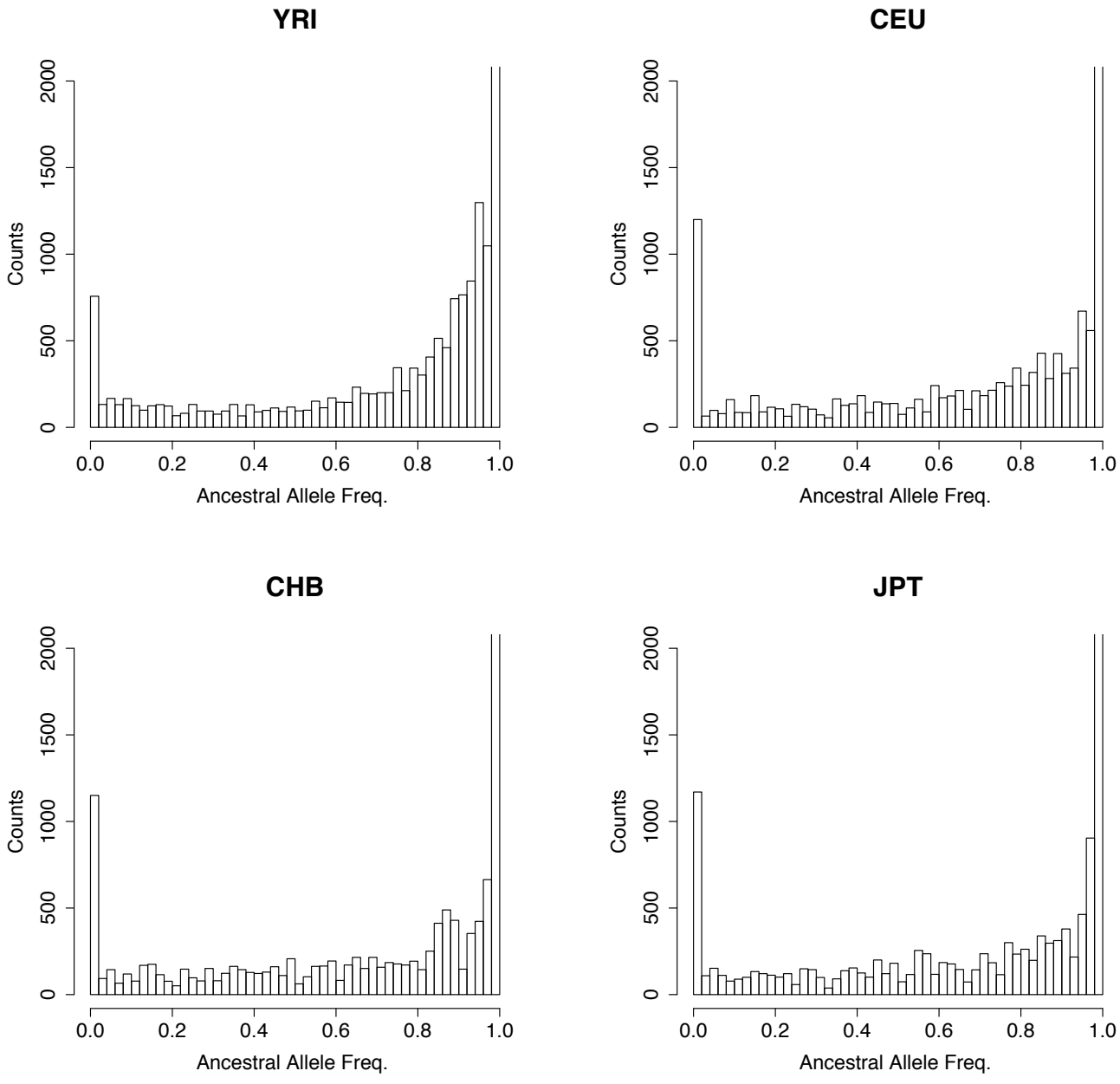

NonAFr-NonAfr
Afr-NonAfr
Afr-Afr
Slope=1.26 +/- 0.004
Fst.ChrX
Fst.autosome
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S7


YRI
CEU
CHB
JPT
Counts
Ancestral Allele Freq.
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0