Genomics 96 (2010) 199–210

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping

Jinchuan Xing [a], W. Scott Watkins [a], Adam Shlien [b,1], Erin Walker [b,1], Chad D. Huff [a], David J. Witherspoon [a], Yuhua Zhang [a], Tatum S. Simonson [a], Robert B. Weiss [a], Joshua D. Schiffman [c], David Malkin [b], Scott R. Woodward [d], Lynn B. Jorde [a,*]

[a] Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, 84112, USA
[b] Department of Genetics and Genome Biology, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada M5G 1X8
[c] Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, 84112, USA
[d] Sorenson Molecular Genealogy Foundation, Salt Lake City, UT 84115, USA

ARTICLE INFO

Article history:
Received 21 May 2010
Accepted 13 July 2010
Available online 16 July 2010

Keywords:
Single nucleotide polymorphism array
SNP
Population structure
Population diversity
Human population history

ABSTRACT

High-throughput genotyping data are useful for making inferences about human evolutionary history. However, the populations sampled to date are unevenly distributed, and some areas (*e.g.*, South and Central Asia) have rarely been sampled in large-scale studies. To assess human genetic variation more evenly, we sampled 296 individuals from 13 worldwide populations that are not covered by previous studies. By combining these samples with a data set from our laboratory and the HapMap II samples, we assembled a final dataset of ~250,000 SNPs in 850 individuals from 40 populations. With more uniform sampling, the estimate of global genetic differentiation (F_{ST}) substantially decreases from ~16% with the HapMap II samples to ~11%. A panel of copy number variations typed in the same populations shows patterns of diversity similar to the SNP data, with highest diversity in African populations. This unique sample collection also permits new inferences about human evolutionary history. The comparison of haplotype variation among populations supports a single out-of-Africa migration event and suggests that the founding population of Eurasia may have been relatively large but isolated from Africans for a period of time. We also found a substantial affinity between populations from central Asia (Kyrgyzstani and Mongolian Buryat) and America, suggesting a central Asian contribution to New World founder populations.

Published by Elsevier Inc.

Introduction

Every major demographic event in a population's history (*e.g.*, population bottlenecks, expansions, and migrations) leaves an imprint on the population's collective assemblage of DNA sequences. Consequently, studies of DNA variation have illuminated many aspects of human population history. Because the genetic variation responsible for disease is a subset of genetic variation in general, these studies are also providing a foundation for important biomedical studies [1,2]. Large-scale genotyping efforts using high-density SNP microarrays have generated an unprecedented amount of human population genetic data. In addition to their application in whole-genome association studies, these data have been used to address issues such as the evolutionary history of human populations [3–10], estimation of individual ancestry [3,11–15], and patterns of natural selection in populations [16–20].

In contrast to the rapid pace of technological development, progress in collecting human DNA samples has been slow and uneven. All existing human genetic diversity datasets, including the HapMap collection, the Coriell collection, and the Human Genome Diversity Project (HGDP-CEPH), are only partial representations of worldwide human diversity. For example, the HGDP database, one of the most widely used resources, lacks coverage in India. Other major regions, such as Eastern Europe and central/north Asia, are also under-represented in databases of human genetic variation.

To help achieve a more uniform sampling of worldwide human genetic diversity, we genotyped a sample of 296 individuals from 13 populations using Affymetrix 6.0 microarrays (~900,000 SNPs and 946,000 copy number variation (CNV) probes). We included populations from West Africa (Dogon and Bambaran), Central Europe (Slovenian), West Asia (Iraqi), Central Asia (Kyrgyzstani and Buryat), South/Southeast Asia (Pakistani, Napalese, and Thai), Polynesia (Tongan and Samoan), and America (Bolivian and Totonac). By adding these populations from previously under-represented regions to existing datasets, we sought to achieve two goals: first, a more comprehensive understanding of the distribution of human genetic diversity; second, a more detailed inference of human demographic

* Corresponding author. Fax: +1 801 585 9148.
E-mail address: lbj@genetics.utah.edu (L.B. Jorde).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2010.07.004

history, such as the mode and tempo of the out-of-Africa diaspora, the peopling of South Asia, and the peopling of America.

Materials and methods

DNA samples

DNA samples from 13 worldwide populations were collected by the Sorenson Molecular Genealogy Foundation (SMGF) and genotyped (Fig. 1, Table 1). Informed consent was obtained from all study subjects at the sampling location, and the Western Institutional Review Board approved all procedures. The sampling locations of these populations are: Bambaran: southwest Mali; Dogon: central Mali in the state of Mopti; Slovenian: several locations in Slovenia; Iraqi Kurds, born in Akra, northern Iraq (collected in Baghdad); Pakistani: Arain agriculturalists from the Punjab region; Nepalese: collected from Kathmandu, Nepal (samples consist of 16 Brahman, 2 Magar, 2 Chhetri, 2 Newar,1 Madhesi, and 2 Nepalese with unknown ethnicity); Kyrgyzstani: collected from Bishkek, the capital of Kyrgyzstan, having origins in several states in northeast Kyrgyzstan; Thai: 19 samples from the Moken ethnic group, and ten from Phuket, Thailand; Buryat: Buryat ethnic group from northeastern Mongolia; Samoan: ethnic Samoans sampled in Samoa; Tongan: ethnic Tongans sampled in Tonga; Totonac: agriculturalists living near Vera Cruz, Mexico; and Bolivian: high-altitude Native American Aymara speakers living near La Paz. Most of these DNA samples were collected from saliva, with the exception of 22 Tongans and Samoans from whom blood samples were obtained.

SNP genotyping, genotype calling and quality control

High-throughput microarray genotyping of approximately 906,000 SNPs was performed using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA). Previous comparisons using this array indicate that DNA derived from saliva samples yield SNP genotypes of quality comparable to DNA derived from blood samples [21]. The recommended protocol described by Affymetrix was followed to construct DNA libraries. Samples were then injected into microarray cartridges and hybridized in a GeneChip® Hybridization Oven 640, followed by washing and staining in a GeneChip® Fluidics Station 450. Mapping array images were obtained using the GeneChip® Scanner 3000 7G (Affymetrix).

Genotypes of 302 microarrays that passed the initial QC were called with the Birdseed algorithm (version 2) in the Affymetrix Power Tools package (http://www.affymetrix.com/support/developer/powertools/index.affx) with default parameters. Because our samples contain no females, CEL files from 15 unrelated CEU female samples were included in the calling process following the manufacturer's recommendation. After genotype calling, we calculated pairwise allele-sharing genetic distances between each pair of individuals. Five comparisons showed unusually small genetic distances, indicating close relatedness between these pairs of individuals. Therefore, one individual was excluded from each pair

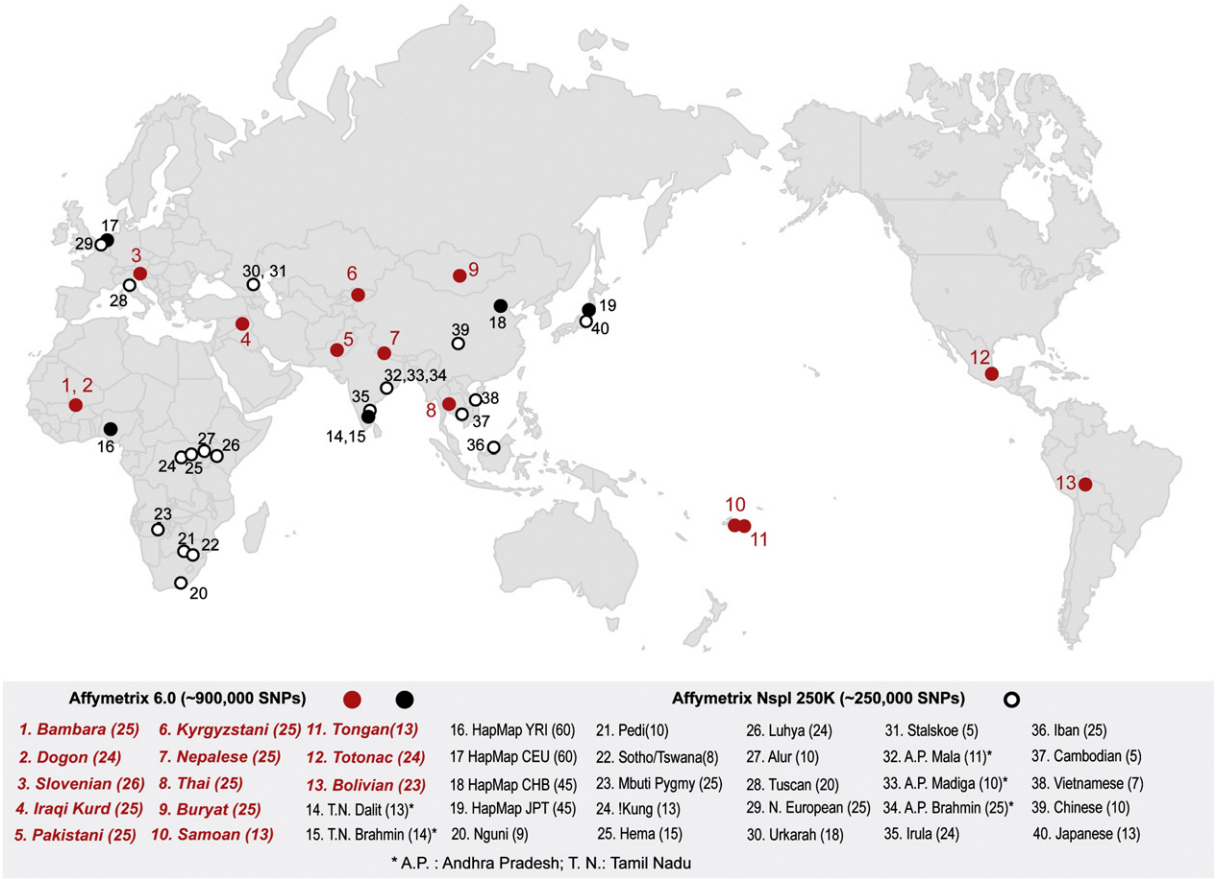


Fig. 1. Population samples analyzed in this study. The number of individuals sampled in each population is shown at the bottom of the figure. Populations genotyped in this study (1–13) are colored in red and populations obtained from the HapMap project and Xing et al. [7] are shown in black. Populations are labeled with filled (Affymetrix 6.0 array) or empty circles (Affymetrix NspI 250 K array) on the map based on the genotyping platforms.

Table 1
Populations and their average SNP and haplotype heterozygosities.

Continental groups	Population	No. of samples	SNP Het	Population group	No. of samples	Hap Het
Africa	!Kung	13	26.64%	–		–
	Alur	10	28.96%	Nilotic	25	91.68%
	Hema	15	30.18%	Nilotic		
	Bambaran	**25**	**28.91%**	**Bambaran**	**25**	**90.66%**
	Dogon	**24**	**28.74%**	**Dogon**	**24**	**89.30%**
	Luhya	24	29.61%	Luhya	24	91.42%
	Nguni	9	29.45%	Bantu	27	91.95%
	Pedi	10	29.12%	Bantu		
	Sotho/Tswana	8	29.69%	Bantu		
	Pygmy	25	25.20%	Pygmy	25	89.23%
	YRI	60	29.08%	YRI	60	90.53%
Europe	CEU	60	27.70%	CEU	60	81.54%
	N. European	25	28.13%	N. European	25	82.07%
	Slovenian	**26**	**28.05%**	**Slovenian**	**26**	**81.97%**
	Stalskoe	5	29.27%	Daghestani	23	82.48%
	Urkarah	18	27.15%	Daghestani		
	Tuscan	25	28.41%	Tuscan	25	82.76%
West Asia	**Iraqi Kurd**	**24**	**27.49%**	**Iraqi Kurd**	**24**	**82.62%**
Central/South Asia	AP Brahmin	25	27.62%	AP Brahmin	25	83.23%
	AP Madiga	10	27.65%	Mala/Madiga	21	82.78%
	AP Mala	11	27.84%	Mala/Madiga		
	Irula	24	26.49%	Irula	24	80.31%
	Kyrgyzstani	**25**	**27.82%**	**Kyrgyzstani**	**25**	**81.94%**
	Nepalese	**25**	**28.35%**	**Nepalese**	**25**	**83.71%**
	Pakistani	**25**	**27.46%**	**Pakistani**	**25**	**83.16%**
	TN Brahmin	14	27.81%	TN Brahmin		–
	TN Dalit	13	27.29%	TN Dalit		–
East Asia	**Buryat**	**25**	**26.54%**	**Buryat**	**25**	**79.72%**
	CHB	45	25.61%	CHB	45	79.89%
	Iban	25	25.49%	Iban	25	78.43%
	JPT	45	25.43%	JPT	45	79.58%
	Thai	**24**	**26.78%**	**Thai**	**24**	**80.58%**
	Cambodian	5	26.45%	–		–
	Chinese	10	25.74%	–		–
	Japanese	13	27.02%	–		–
	Vietnamese	7	25.64%	–		–
Polynesia	**Samoan**	**13**	**24.65%**	**Tongan/Samoan**	**26**	**75.30%**
	Tongan	**13**	**24.75%**	**Tongan/Samoan**		
America	**Bolivian**	**23**	**24.31%**	**Bolivian**	**23**	**73.91%**
	Totonac	**24**	**23.83%**	**Totonac**	**24**	**72.38%**

Populations sampled in this study are in bold. SNP Het: SNP heterozygosity; Hap Het: Haplotype Heterozygosity; Population Group: for populations that have been combined in the haplotype analysis, the same population group name is shown for all populations within the group (e.g., populations "Alur" and "Hema" form the population group "Nilotic").

in order to retain a set of unrelated individuals. One additional individual was removed because of ambiguous population information. The remaining 296 samples from 13 populations compose our dataset for analyses.

SNP selection

Several criteria were applied to select SNPs for the analyses. First, we excluded all SNPs on the X, Y, and mitochondrial chromosomes, as well as SNPs whose chromosomal locations were unknown (38,456 SNPs). Then, SNPs with more than 10% missing data were removed (5742 SNPs). We next divided all individuals into four major groups (Africa, Asia, Europe, and India) and tested each SNP for deviations from Hardy–Weinberg Equilibrium (HWE) for populations within each group using the hweStrata algorithm [22]. The continent-level HWE p-values were combined using Stouffer's z-average method [23], and 213 SNPs that deviated from HWE at $p<5.5\times10^{-8}$ (Bonferroni correction: 0.05/900,000) were excluded from subsequent analyses. To combine our dataset with HapMap II samples, Affymetrix SNP Array 6.0 genotypes of the 210 unrelated HapMap samples were obtained from the HapMap project website (http://hapmap.ncbi.nlm.nih.gov), and the same SNP selection criteria were applied to HapMap samples. The filtered HapMap dataset was combined with the dataset generated in this study and a dataset from an earlier study [7] using the Affymetrix NspI 250 K microarrays, resulting in a final dataset containing 246,554 autosomal loci genotyped in 850 individuals from 40 populations.

CNV genotype calling

The microarray data for the 296 DNA samples were analyzed for CNVs using two complementary algorithms: a genomic segmentation algorithm (Partek, MO) and Birdsuite [24]. The use of two complementary CNV detection algorithms increases the robustness of CNV detection [25]. To minimize batch variability, an internal baseline was generated from all 296 samples and used in the segmentation CNV detection. A minimum of ten consecutive probes was required to detect a copy number change. CNVs were removed if the probe density was <1 probe/5000 bp, in order to remove potentially spurious CNV calls that cover centromeric regions. The Canary and Birdseye algorithms were used in Birdsuite version 1.5.3. We restricted our analysis to autosomal CNV calls that had a LOD score greater than or equal to 10, and that were greater than 1 kb in length. To obtain a conservative set of CNV regions, we removed any CNVs not found by both algorithms, leaving a stringent set of copy number regions for each individual. Genotypes of all samples in the final dataset (including both SNP and CNV genotypes) are available as a supplemental file on our website (http://jorde-lab.genetics.utah.edu/) under Published Data. The pre-filtering raw dataset is available upon request.

Data analysis

Haplotype diversity

To standardize the population sample sizes, we combined several closely related populations into population groups and excluded remaining populations that had fewer than 20 individuals (see Table 1 for details). The combined population groups are: Nilotic (Alur and Hema), Bantu (Nguni, Pedi, and Sotho/Tswana), Daghestani (Stalskoe and Urkarah), Mala/Madiga (AP Madiga and AP Mala), and Tongan/Samoan (Tongan and Samoan). Then, we randomly chose 20 individuals from each population group to equalize the sample sizes. The genome was divided into consecutive 100 kb windows, and the number of SNP loci in the dataset was determined for each window. Windows with fewer than 10 loci in the final dataset were excluded. For windows containing more than ten SNPs, we calculated the haplotype heterozygosity [26] in each population using the MATLAB Population Genetics & Evolution Toolbox [27].

Population tree

Distances between populations were calculated from allele frequency data as Nei's genetic distance implemented in the PHYLIP software package [28]. The dataset contains 232,114 SNPs with known ancestral state for 40 world populations. Dendrograms were constructed using the neighbor-joining method. All ancestral allele states were obtained from the orthologous base in chimpanzee, or orangutan plus macaque if chimpanzee was unknown, as obtained from the UCSC Genome Browser database (hg19, snp130). Each dendrogram was rooted by this chimpanzee–orangutan–macaque outgroup. One thousand bootstrap runs were performed for each dataset to generate the consensus tree and obtain the confidence value for each branch.

F_{ST} estimates and principal component analysis (PCA)

F_{ST} estimates between populations were calculated by the method described by Weir and Cockerham [29]. To obtain the confidence interval of F_{ST} values in each continental group, 60, 60, and 90 individuals were randomly sampled 1000 times (with replacement) from Africa, Europe, and Asia (to match the sample sizes of the HapMap II populations), respectively. Pairwise allele-sharing genetic distance calculation and PCA were performed using MATLAB (ver. r2008a).

ADMIXTURE analysis

A model-based algorithm implemented in *ADMIXTURE* [30] was used to determine the genetic ancestries of each individual in a given number of populations without using information about population designation. To eliminate the effect of SNPs that are in LD, we first filtered out SNPs that had $r^2>0.2$ within 100 kb using PLINK [31], as recommended by the authors of *ADMIXTURE*. The pruned data set contains 86,273 SNPs.

CNV analysis

CNV data were analyzed using internally developed software (available upon request) and SPSS 15.0 (SPSS, IL). We required a minimum of 75% reciprocal overlap between pairs of CNVs to consider that two individuals shared the same CNV region. A pairwise comparison of shared CNV regions allowed us to identify those CNVs that were private to individuals, private to specific populations, and those CNVs that were shared across multiple populations. To adjust for outlier effects, individuals above the 95th percentile for CNV number were removed from the analysis. A principal component analysis performed on all individuals indicated that the DNA samples from different DNA sources (*i.e.*, blood versus saliva) may have different CNV calling results (Supp. Fig. S1). Because DNA only from the 22 Tongan and Samoan samples was derived from blood, we excluded these subjects from the CNV analysis.

Table 2
F_{ST} and proportion of polymorphic SNPs shared among continents.

Sample set	250 K SNPs		866 K SNPs	
	F_{ST}*	Shared SNPs	F_{ST}*	Shared SNPs
HapMap	15.9% (15.8%–16.1%)	74.9%	15.9% (15.8%–16.0%)	72.7%
HapMap + Affy6.0	12.3% (11.8%–12.9%)	83.6%	12.2% (11.7%–12.8%)	81.9%
HapMap + Affy6.0 + Affy250k	11.2% (10.7%–11.7%)	88.2%	–	–

* 95% confidence intervals (CI) are shown in parentheses.

Results

Population samples

We sampled 296 individuals from 13 worldwide populations, including populations from West Africa, Central Europe, West Asia, Central Asia, South Asia, Southeast Asia, Polynesia, and America (Fig. 1, Table 1 populations in bold). All samples were genotyped using the Affymetrix 6.0 array and we will refer to this individual set as the "Affy6.0" set in the following analysis. We then combined these samples with 344 individuals from 23 populations in our previous study [7] (Fig. 1, populations in black), in which the Affymetrix 250 K NspI array was used ("Affy250K" set), and 210 individuals from four HapMap populations (YRI, CEU, CHB and JPT, "HapMap" set). The final dataset contains 246,554 autosomal loci genotyped in 850 individuals from 40 populations (see Materials and methods for details of SNP selection and merging criteria). To determine the effect of using only this subset of the SNPs, we compared pairwise F_{ST} between each pair of populations in the HapMap and the Affy6.0 sample set using the 246,554 SNP set and the whole SNP set (~866,000 SNPs). The F_{ST} values between all population pairs are virtually identical for the two SNP sets (overall correlation coefficient $r=0.99998$, $p\ll10^{-50}$), suggesting that the 250 K SNP set is sufficient for examining inter-population relationships.

Decrease in population differentiation with more uniform sampling

To assess the effect of more even sampling on the degree of population differentiation, we compared the F_{ST} values between three major continental groups (Africa, Europe and Asia) from three individual sets: HapMap, HapMap + Affy6.0, and HapMap + Affy6.0 + Affy250K. To match the sample sizes of the HapMap set, we randomly sampled 60, 60, and 90 individuals from Africa, Europe and Asia, respectively, in each individual set. Our results show that the overall F_{ST} value decreases substantially with the inclusion of geographically intermediate populations, dropping from 15.9% for HapMap, to 11.2% for HapMap + Affy6.0 + Affy250 K with non-overlapping confidence intervals (Table 2). Adding the American and Polynesian individuals into the HapMap + Affy6.0 + Affy250 K set increased F_{ST} slightly (to 11.3%) because of substantial founder effects and genetic drift in these populations. Nevertheless, the F_{ST} value in all individuals is still significantly lower than the F_{ST} value of the HapMap individual set. These statistically significant F_{ST} differences illustrate the important effects of population sampling. A decrease in population differentiation with more even sampling is also demonstrated by an increase in the proportion of SNPs whose minor alleles are shared in all three continental groups. This value increases from 74.9% for HapMap to 88.2% for HapMap + Affy6.0 + Affy250K (Table 2).

For individuals that were genotyped for more than 866,000 autosomal SNPs using the Affymetrix 6.0 array (HapMap and Affy6.0), we also determined the F_{ST} values and proportion of polymorphic SNPs using all genotyped autosomal SNPs. In both individual sets the F_{ST} values using all SNPs are comparable to F_{ST}

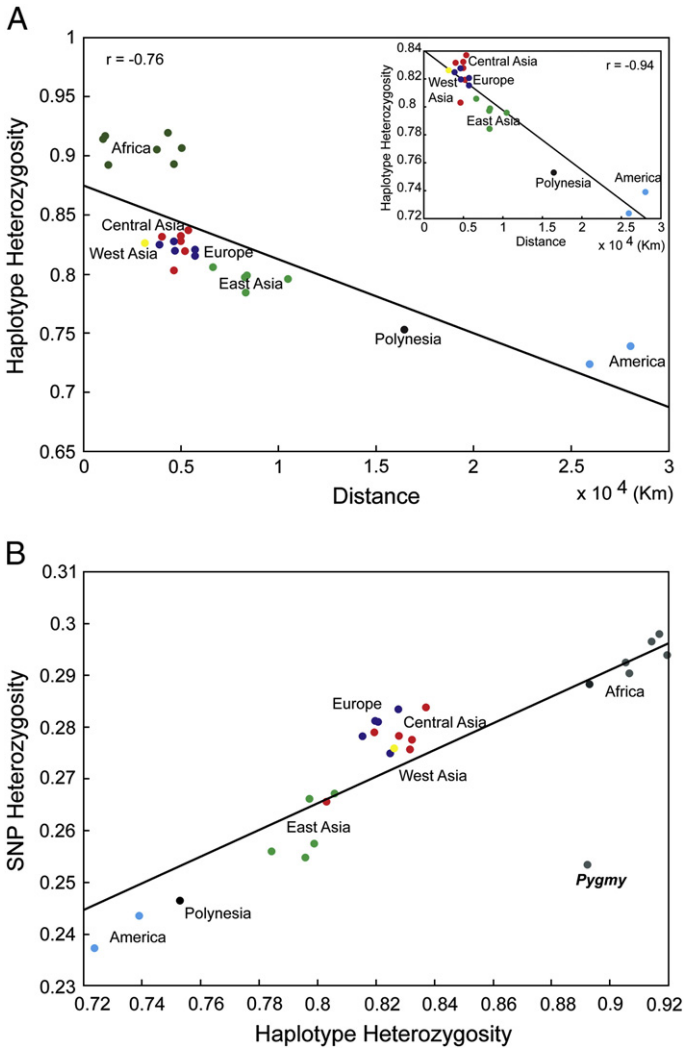


Fig. 2. SNP diversity. A) SNP haplotype diversity as a function of geographic distance from East Africa. The correlation improved substantially when African populations were excluded ($r = -0.94$, upper right panel). B) SNP haplotype diversity versus SNP heterozygosity.

values using the ~250,000 SNPs (Table 2). The difference between the two individual sets remains significant. The percentage of shared polymorphic SNPs decreases slightly in both datasets (Table 2), reflecting a relatively higher proportion of low-frequency SNPs in the Affymetrix 6.0 array.

Haplotype diversity

To compare haplotype diversity across populations, we normalized the sample size across population groups by randomly choosing 20 individuals and excluding populations with fewer than 20 samples (see methods for details). The average haplotype heterozygosity is significantly higher in African populations than non-African populations (Table 1, Wilcoxon rank test $p = 1.2 \times 10^{-4}$), and haplotype diversity decreases as geographic distance to East Africa increases (Fig. 2A, $r = -0.76$, $p = 4.3 \times 10^{-6}$). Despite the overall significant correlation, there appears to be little correlation within Africa between haplotype diversity and distance to East Africa ($r = -0.13$, $p = 0.78$). Indeed, when African populations were excluded from the analysis, a stronger correlation is obtained ($r = -0.94$, $p = 9.6 \times 10^{-10}$, Fig. 2A upper panel).

We also compared the SNP and haplotype heterozygosity values in each population (Fig. 2B). These two quantities are generally highly correlated, although there are several exceptions: first, SNP heterozygosity is higher than haplotype heterozygosity in European and Central Asian populations. This may reflect a SNP ascertainment bias, since many of these polymorphisms were historically selected to maximize heterozygosity in European populations. Second, the Pygmy sample shows a low SNP heterozygosity despite relatively high haplotype heterozygosity. This unusual pattern could be caused by stronger effects of SNP ascertainment bias in this population than in others. Indeed, a recent study of Khoisan individuals (another hunter–gatherer group from Africa) showed a similar pattern: despite high SNP heterozygosity (~60%) in whole-genome sequence data, a Khoisan individual showed low heterozygosity on the SNP microarray genotypes (~22%) [32]. Alternatively, this difference could also reflect unique attributes of population history.

Genetic structure among populations

To examine inter-population relationships, we first constructed a neighbor-joining tree based on genetic distances (Fig. 3A). Populations from major geographic regions are clustered, and most branches have very high (>95%) bootstrap support (Supp. Fig. S2). New World populations (Totonac and Bolivian) are placed between Nepalese and Kyrgyzstanis, indicating higher affinity of these American samples to central Asians than to eastern Asians. A second neighbor-joining tree was constructed by adding 40 HGDP populations (46,260 SNPs in common), producing similar patterns of population clustering (Supp. Fig. S3).

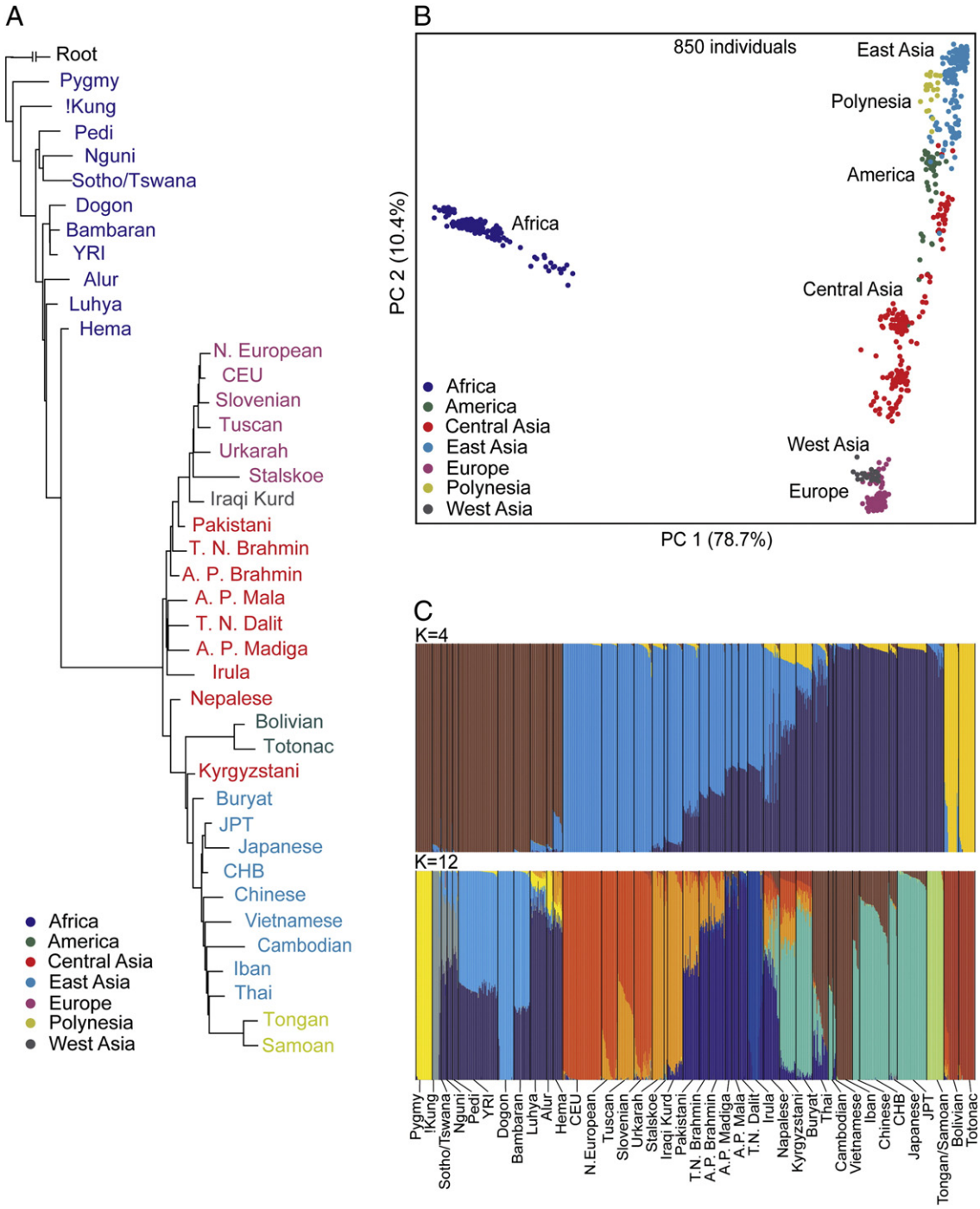


Fig. 3. Population relationships between the 40 populations. A) Neighbor-joining tree. Populations are color-coded based on their continental origins. The hypothetical ancestral population is shown. Bootstrap support values for most branches are larger than 95% (the bootstrap consensus tree is shown in Supp. Fig. S2). B) Principal component analysis. First two principal components (PCs) are shown. Each individual is represented by one dot and the color label corresponding to their regional origin. The percentage of variance explained by each PC is shown on the axis. C) Individual grouping inferred by *ADMIXTURE*. Results from K = 4 and K = 12 are shown. Each individual's genome is represented by a vertical bar composed of colored sections, where each section represents the proportion of an individual's ancestry derived from one of the K ancestral populations. Individuals are arrayed horizontally and grouped by population as indicated.

We then performed a Principal Component Analysis (PCA) based on the pairwise allele-sharing distances among all pairs of individuals (Fig. 3B). The majority of the genetic variation is found between African and non-African populations, as the first principal component (PC1) accounts for 78.7% of total variance. PC2 reflects genetic variation in Eurasia, and populations from Central and West Asia occupy the space between East Asia and Europe to form a relatively continuous distribution. The two Polynesian populations (Tongan and Samoan) show a close relationship to Southeast Asian populations (Fig. 3B). PC3 distinguishes New World populations (Bolivian and Totonac) from other populations (Supp. Fig. S4A).

At the sub-continental level, we focus first on Eurasia, where most of our samples have been selected (Fig. 4A). Overall, PC1 and PC2 mainly reflect the geographic distribution of the populations, with the majority of genetic variation accounted for by their locations. PC1 (accounting for 62.7% of the variance) reflects an east–west gradient, while PC2 (3.3% of the variance) reflects a north–south gradient. Slovenians and Iraqi Kurds show close relationships to European

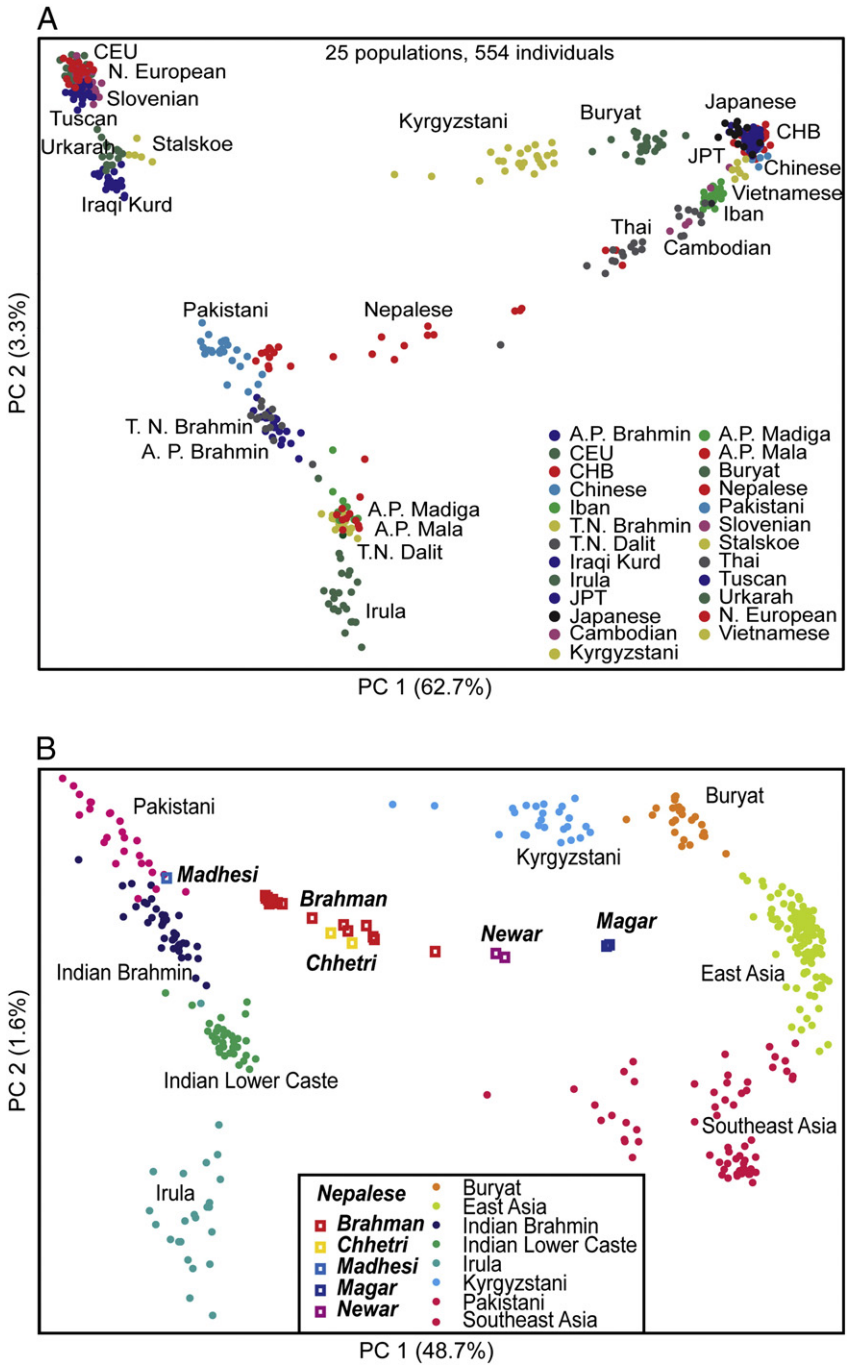


Fig. 4. Principal component analysis of population structure. First two PCs are shown. The percentage of variance explained by each PC is shown on the axis. A) Eurasia. Each individual is represented by one dot and the color label corresponding to their population. B) Nepalese and surrounding Asian populations. Nepalese individuals are represented by squares and the color label corresponding to their ethnic groups. Two Nepalese who have no ethnic group information were excluded from this plot. Other Asian individuals are represented by dots and color labels corresponding to their regional origins to improve the resolution. The regional groups include: India Brahmin (A.P. Brahmin and T.N. Brahmin); India Lower Caste (A.P.Madiga, A.P. Mala, T.N. Dalit); East Asia (CHB, Chinese, Japanese, JPT); and Southeast Asia (Cambodian, Iban, Thai, Vietnamese).

populations. A closer examination (Supp. Fig. S4B) shows that Kurds and eastern European Daghestani populations (Urkarah and Stalskoe) are clearly separated from western European populations. On the other hand, Slovenians show very little differentiation from western European populations (Supp. Fig. S4B).

Some of our populations form less defined clusters than do the HapMap populations. The Nepalese samples, in particular, are highly diverse, with some individuals showing a closer relationship to East Asian populations, while others are closer to South Asian populations. An examination of the ethnicity of the Nepalese individuals reveals that individuals from the ethnic groups derived from the caste system, including Madhesi, Brahman, and Chhetri, show a closer relationship to South Asian populations (especially Indian Brahmins). Individuals from the two indigenous Nepal ethnic groups (Newar and Magar) are closer to Central/East Asian populations (Fig. 4B). Kyrgyzstanis were also widely dispersed along the first PC, although to a lesser extent than the Nepalese samples. This dispersion is expected because Kyrgyzstan is on the trade route between Europe and Asia, where there has long been a high level of migration.

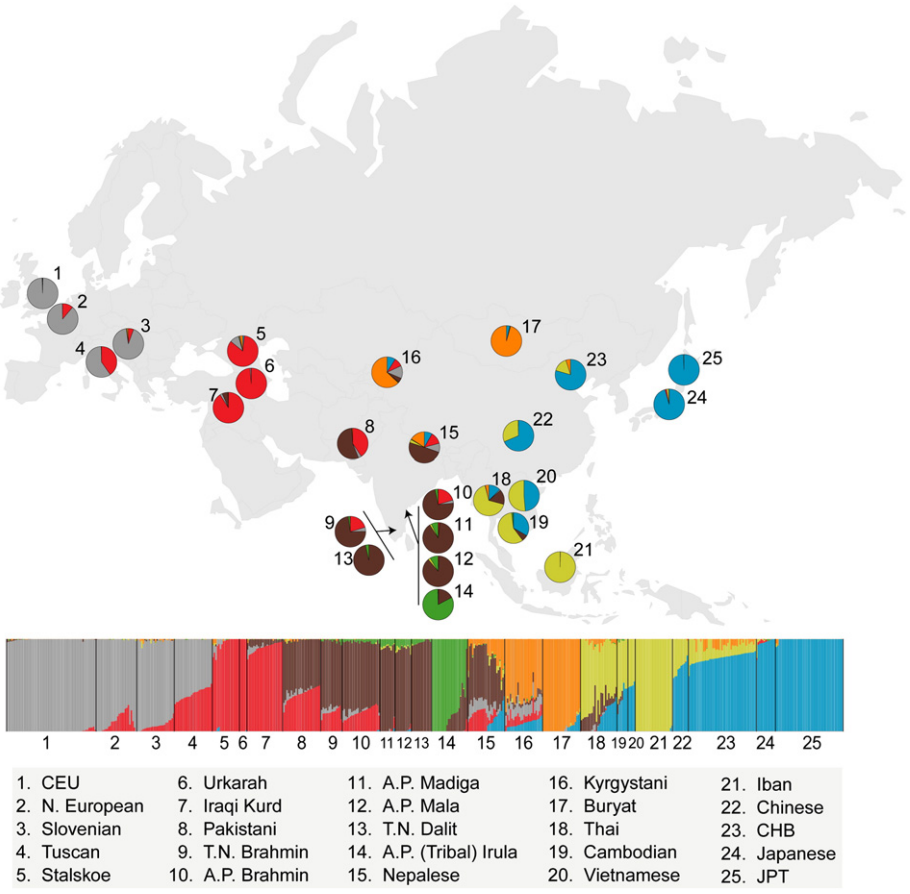


Fig. 5. *ADMIXTURE* analysis of Eurasian individuals with K = 7. Each individual's genome is represented by a vertical bar composed of colored sections (bottom of the figure), where each section represents the proportion of an individual's ancestry derived from one of the seven ancestral populations. Individuals are arrayed horizontally and grouped by population as indicated. In the map, the average ancestral components of each population are illustrated as pie charts.

Distinctive patterns can also be observed at the sub-continental level in non-Eurasian populations. Within Africa, the first two PCs separate Mbuti Pygmy and !Kung from other African populations (Supp. Fig. S4C). The remaining African populations appear to follow a north–south gradient, and the Dogon and Bambara from Mali show high similarity to the HapMap YRI from Nigeria (Supp. Fig. S4C). Within America, the two populations showed contrasting patterns: Totonacs from Mexico form a tight cluster, while about half of the Bolivian samples are separated from the Bolivian cluster, which appears to reflect European admixture (Supp. Fig. S4D).

Individual group membership

We used the program *ADMIXTURE* [30] to assess the ancestry of each individual from 3–12 inferred populations (K) (Supplemental Table S1). The results from K = 4 and K = 12 are illustrated in Fig. 3C. When K = 4, four groups corresponding to Africa, America, Europe, and Asia are identified. Unlike individuals from Africa and America, who form two relatively distinct groups, individuals from Eurasia show a mixture of Asian and European ancestry components.

When K = 12, a number of sub-continental patterns appear. In Africa, Mbuti Pygmy, !Kung, and Dogon are separated into distinct groups. Despite being sampled from neighboring regions in Mali, Bambaran and Dogon individuals show quite different ancestry. Most Dogon individuals appear to be composed of a single western African component, while Bambaran individuals contain more than 30% of a component prevalent in eastern Africa. Polynesian and American populations were separated into two distinct components. In agreement with the PCA result (Supp. Fig. S4D), some Bolivian individuals contain more than 20% European ancestry, suggesting admixture in these samples.

Within Eurasia, the patterns are more complex. To examine the relationships among Eurasian populations in detail, we performed *ADMIXTURE* analysis on the Eurasian individuals only and calculated the average ancestry components in each population. Major regional groups and geographic clines are best visualized with seven ancestral components (K = 7, Fig. 5). In Europe, a northern/western European component is predominant in HapMap CEU, the Utah Northern European, and the Slovenian samples. One Caucasus/Middle East component is predominant in Daghestani and Iraqi samples and appears to decrease clinally to the east through Pakistan and Nepal and to the west through southern and northern Europe. In southern India, this component is a major genetic signal in two independently sampled Brahmin groups (>20%) but is nearly absent in lower castes and Irula (a tribal group, <1.5%). Notably, the central Asian populations of Nepal and Kyrgyzstan have the most genetic admixture. This result is consistent with our PCA results showing a high level of genetic variation within these two populations. Another interesting observation is that Buryats and Kyrgyzstanis share about 5% ancestry with native American populations (averages of 4.4% in Kyrgyzstanis and 5.8% in Buryats; Fig. 3C), while East Asian individuals have very little of the Native American ancestry component (average <1%).

Copy number variation (CNV) profile

As a complement to our SNP analysis, we also used the same array platform to determine each individual's CNV profile. To investigate

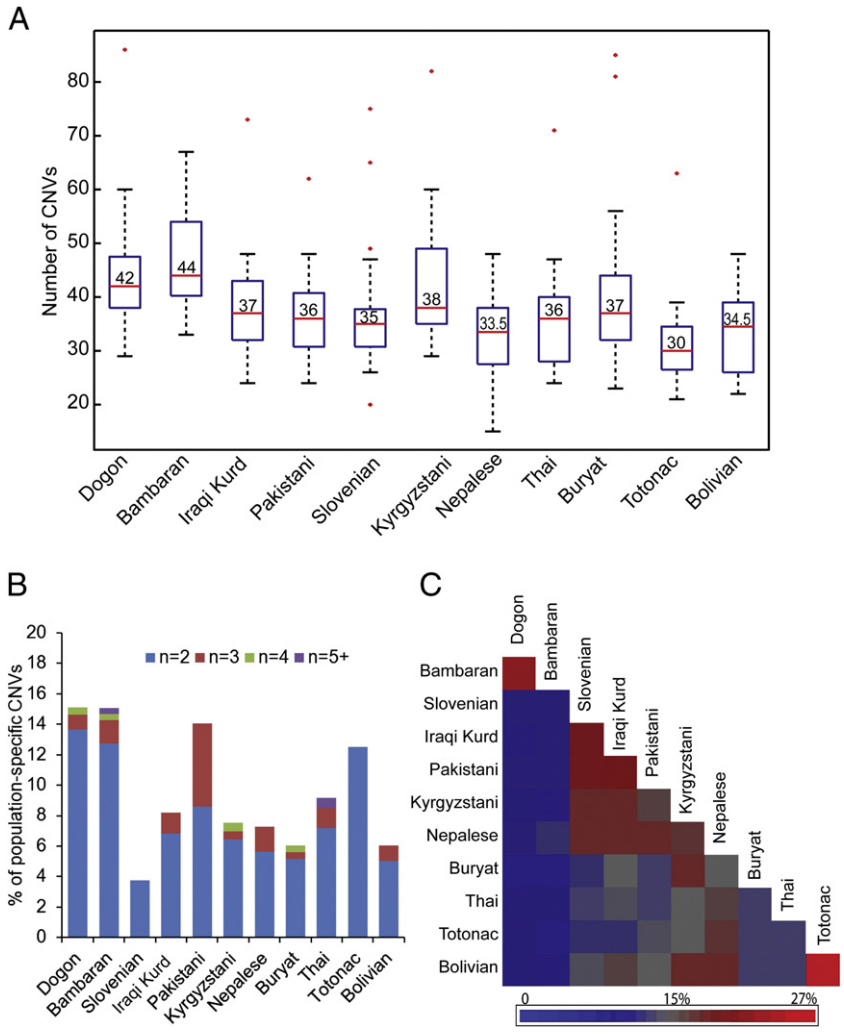


Fig. 6. CNV profile among populations. A) The median number of CNVs in each population. Red bars represent the median number of CNVs. The central boxes span the quartiles and the whiskers extend to the most extreme data points not considered outliers. Outliers are indicated by red dots. B) Population-specific CNVs found in multiple individuals. The percentage of population-specific CNVs present in multiple individuals within each population is shown. The number of CNVs present in 2 (blue), 3 (red), 4 (green), or ≥5 (purple) individuals is represented as a percentage of the total number of CNVs within that population. C) CNV sharing between population pairs. A heatmap showing the CNV overlap between each population pair is shown. The number of CNVs present in both populations was calculated as a percentage of the total number of CNVs in each individual population. The scale below the figure represents the range of percentage values, ranging from 0% (light blue) to 27% (bright red).

the overall inter-population differences due to CNVs, we determined the number of CNVs per person and the average CNV frequencies in each population (Fig. 6A). The African populations (Bambaran and Dogon) have the highest number of CNVs among all populations (median of 44 and 42 CNVs per genome, respectively). Outside of Africa, median number of CNVs varies between 38 in Kyrgyzstani to 30 in Totonac (Fig. 6A). These data are comparable with previous work, which found a higher number of CNVs in African populations [33–36], suggesting a loss of low-frequency CNV alleles due to population bottlenecks during the out-of-Africa migration and the peopling of the Americas.

Next, we identified CNVs that are specific to each population and then counted the number of individuals within each population sharing the same population-specific CNV (Fig. 6B). Within the Dogon, Bambaran, Pakistani, and Totonac populations, we found a high proportion of population-specific CNVs that were common to multiple members, with more than 12% observed in two or more individuals. The remaining populations had few population-specific CNVs in common among their members. More than 90% of detected population-specific CNVs in these populations are only present in one individual. Because most population-specific CNVs are relatively rare within each population, and there are only a small number of total CNV loci, samples from different populations do not form distinct clusters in a PCA (Supp. Fig. S5).

We also investigated CNVs that are common between pairs of populations (Fig. 6C). A comparison of the African populations (Dogon and Bambaran) revealed that 23% of CNVs were present in both populations, while both groups had little in common with any other population. There is also a relatively high proportion of CNVs in common between the Slovenian and Iraqi populations. Likewise, the Pakistanis, Kyrgyzstanis, Nepalese and Buryats all have a high percentage of CNVs in common (14–19%, Fig. 6C). This pattern is consistent with the population affinities shown by PCA and *ADMIXTURE* analysis of the SNP data (Figs. 4A and 5). Finally, the Totonac and Bolivian populations have the highest proportion of CNVs in common, with 27.1% of CNVs identified in both populations. This high proportion of CNV sharing and the relatively low number of CNVs identified in these populations may be due to the low genetic diversity in their common founding population.

Discussion

Patterns of human genetic variation are influenced by mating patterns, and the latter are in turn influenced by geographic and cultural factors (*e.g.*, mountain ranges, language, religious practices). Consequently, it is not surprising that human genetic variation, while correlated with geographic location, is not perfectly clinal [37–39]. However, between-population differences can be seriously exaggerated if human populations are sparsely sampled.

Consistent with previous studies [37,39,40], our analyses demonstrate that differentiation among human populations decreases substantially and genetic diversity is distributed in a more clinal pattern when more geographically intermediate populations are sampled. The reduction of F_{ST} values with further geographic sampling illustrates the limitations of a global F_{ST} estimate to capture the pattern of human genetic diversity. With more comprehensive population samples, our data have also led to several new observations about human demographic history and genetic relationships among human populations.

The out-of-Africa (OoA) bottleneck and the peopling of Eurasia

As observed in previous studies [4,5,7], we find that SNP and haplotype variation is highest in African populations, and that heterozygosity in non-African populations declines with geographic distance from Africa. This decline in heterozygosity has been interpreted as evidence for a worldwide serial founder effect originating in East Africa [4,41]. While serial founder effects may explain much of the pattern of worldwide variation, we note two interesting deviations from the prediction of a linear decline in heterozygosity. First, as demonstrated in Fig. 2A, there appears to be little relationship between heterozygosity within Africa and distance from the hypothesized point of East African origin ($r = -0.13$, $p = 0.78$). Second, there is a drastic decrease in diversity for all Eurasian populations immediately outside of Africa. These observations are best explained by a single bottleneck out of Africa rather than by a series of founding emigrations from Africa (Fig. 2A).

The OoA hypothesis, proposing a single OoA bottleneck followed by an expansion into Eurasia approximately 50,000 years ago, has gained extensive support from the archeological record [42,43] and genetic studies [4,5,7]. Nevertheless, many of the historical details of this diaspora remain unclear. A common interpretation is that the OoA bottleneck was the result of a migration of a small founding population into Eurasia. Given the difference in haplotype heterozygosity between African and non-African populations and the relationship between heterozygosity and effective population size, we can estimate the effective population size of such a founding population [44]. Within Africa, the average 100-kb haplotype heterozygosity in our data is 0.91. Immediately outside of Africa in Europe, the Middle East, and Central Asia, the average haplotype heterozygosity is 0.82 (Fig. 2). A reduction of heterozygosity from 0.91 to 0.82 in a one-generation bottleneck would require an effective population size of only 5.5 individuals. While a one-generation bottleneck is an oversimplification, these estimates indicate that an OoA bottleneck resulting from the migration of a small founding population would require an extremely small population size. However, given that the archeological record indicates a rapid expansion of modern humans into Europe and Asia in just a few thousand years [42,43], it seems unlikely that Eurasia could be populated so quickly by a such a small founding population.

A more likely explanation for the OoA bottleneck is that Eurasia was populated by a larger population that had been relatively isolated from other modern human populations for tens of thousands of years prior to the expansion. The first fossil evidence for modern humans outside of Africa is in the Middle East at Skhul and Qafzeh between 80,000–100,000 years ago, which is at least 20,000 years prior to the Eurasian diaspora [45]. If a population of modern humans remained in a location such as the Middle East until the expansion into Eurasia, there would have been sufficient time for genetic drift to reduce heterozygosity dramatically before the Eurasia expansion. This "Delayed expansion" hypothesis provides a robust explanation for the relative homogeneity of European and Asian populations relative to African populations (see Figs. 3A and B) and is supported by a recent maximum likelihood estimate of 140,000 years ago for the time of Eurasian–West African population separation [46]. Interestingly, a recent study of the Neandertal genome suggests that the non-African individuals, but not the Africans, contain a similar amount of admixture (1–4%) with the Neandertals [47]. The authors suggest that the admixture must have happened between the Neandertals and an ancestral non-African population before the Eurasian expansion. Given the fossil, archeological, and genetic evidence, the Delayed expansion hypothesis warrants rigorous evaluation as whole-genome sequence data become available.

Dispersion of a Caucasus/Middle East genetic component

In the *ADMIXTURE* analysis of Eurasia, we observed a clinal distribution of a Caucasus/Middle East genetic component (red component, Fig. 5) in several South Asian populations. Evidence from mitochondrial DNA, Y-chromosome, and autosomal loci suggests that the genetic composition of India has been influenced by west Eurasians [7,8,48,49]. We find that this ancestry component is most prevalent in West Asians (Iraqi Kurd) and Caucasus populations (Daghestani). The component extends eastward into Central Asia (Pakistan, Nepal, and Kyrgyzstan) and into South India, where it is more prevalent in higher castes than in lower castes. This ancestry component also extends into Europe and is more prevalent in southern Europeans than in northern Europeans. Our results suggest that the northern Indian genetic component proposed by Reich et al [8] could represent the dispersion of a genetic ancestry component originating near the Caucasus/Middle East region.

Nepalese diversity

Containing more than 100 ethnic groups, Nepal is a geographically small but diverse country [50]. Earlier genetic studies of Nepalese populations have suggested a northern Asian origin with subsequent gene flow from South Asia (*e.g.*, Hindu caste-derived groups) [51–53]. Our results are in general agreement with this view and suggest that the most prevalent ancestry component in the Nepalese is the primary ancestry component found in Indians and Pakistanis. The Nepalese, however, are highly heterogeneous and also have substantial ancestry components from Central Asia, East Asia, and Southeast Asia. Moreover, individual Nepalese from different ethnic groups have substantially different genetic composition. Hindu upper-caste Nepalese Brahman and Chhetri individuals cluster in PCA and show affinity to Indian Brahmin samples (Fig. 4B). In contrast, samples from the linguistically distinct Magar and Newar groups show affinity to populations from Central and East Asia. These results suggest that substantial population structure may exist between the major population groups of Nepal. Although our limited sample size prevents a detailed analysis of the genetic diversity among Nepalese ethnic groups, our observations suggest high levels of genetic diversity in South and Central Asian populations and underscore the need for additional genetic studies of this region.

Native American founding populations

The Americas, first peopled during the late Pleistocene, were the last continents to be colonized by modern humans. Despite general agreement that modern humans crossed a land bridge in the current Bering Strait region to populate the Americas (reviewed in [54–56]),

the exact timing, routes of colonization, and origin of the ancestral population(s) remain unclear [57–61].

Earlier studies suggest that an ancestral American population may have lived in western Siberia, rather than eastern Siberia/Northern Asia [62,63]. Congruent with this view, the two Native American populations (Totonac and Bolivian) in our samples show closer relationships to Central Asian populations (Kyrgyzstanis and Buryats from Mongolia) than East Asian populations (*e.g.*, Chinese and Japanese). This result is most apparent in the *ADMIXTURE* plot (Fig. 4B; K = 12), where Kyrgyzstani and Buryat individuals share about 5% of the American ancestry component. In contrast, East Asian individuals share very little (<1%) genetic ancestry with the American populations.

CNV population profiles

In previous studies, we have shown highly consistent patterns of population genetic structure when using different types of polymorphisms, such as restriction site polymorphisms, short tandem repeat polymorphisms, and *Alu* and L1 insertion polymorphisms [37,64–66]. Similarly, despite a very different mutational mechanism, CNVs also reveal overall patterns of genetic structure that are highly similar to those of other types of polymorphisms: first, we find that populations from Africa harbor the greatest number of CNVs, and that the average number of CNVs decreases with increasing distance from Africa. Second, we find that the degree of CNV sharing between groups reflects their population relationships. Notably, the Totonac and Bolivian populations share a high number of CNVs. The Pakistani, Kyrgyzstani, Nepalese, and Buryat populations also exhibit a high number of shared CNVs. Previous studies have also shown general agreement in genetic structure patterns revealed by SNP and CNV data [5].

Conclusion

In this study, by sampling populations from previously undersampled regions, we sought to assess the effect of more even sampling on human genetic diversity and to investigate the evolutionary history of these populations. We found support for a relationship between the initial founding populations of America and Central/North Asian populations. We demonstrated high genetic diversity in Central Asian and South Asian populations, especially in Nepal. We also found that Iraqi Kurds have a closer relationship to European populations than Asian populations. These results increase our understanding of human population relationships and evolutionary history. In addition, our data provide a resource for understanding patterns of linkage disequilibrium, natural selection and the differential distributions of SNP and CNV alleles among populations, all of which have important implications in genome-wide association studies and the identification of loci with functional, biomedical significance.

Acknowledgments

We thank Dr. Dashtseveg Tumen, National University of Mongolia; Dr. Sukkid Yasothornsrikul, Naresuan University; and Dr. Alejandro Escobar, State of Veracruz Department of Health for their help in collecting the samples. We thank Dr. Dennis O'Rourke for insightful discussion on the peopling of America. We also thank Diane Dunn and Edward Meenen for their technical support during the microarray hybridization and scanning process. This work was supported by grants from the National Institutes of Health (GM-59290 to LBJ) and the Sorenson Molecular Genealogy Foundation. Additional supports for this study were provided by grants from the Canadian Institutes for Health Research (DM). C.H. is supported by the University of Luxembourg – Institute for Systems Biology Program and the Primary Children's Medical Center Foundation National Institute of Diabetes and Digestive and Kidney Diseases (DK069513). A.S. is supported by a Canadian Institutes of Health Research Frederick Banting & Charles H. Best Doctoral Studentship Award.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.ygeno.2010.07.004.

References

[1] D. Altshuler, L.D. Brooks, A. Chakravarti, F.S. Collins, M.J. Daly, P. Donnelly, A haplotype map of the human genome, Nature 437 (2005) 1299–1320.
[2] Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls, Nature 447 (2007) 661–678.
[3] C. Tian, R.M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, L. Klareskog, A.E. Pulver, L. Qi, P.K. Gregersen, M.F. Seldin, Analysis and application of European genetic substructure using 300 K SNP information, PLoS Genet. 4 (2008) e4.
[4] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.
[5] M. Jakobsson, S.W. Scholz, P. Scheet, J.R. Gibbs, J.M. VanLiere, H.C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, R. Guerreiro, J.M. Bras, J.C. Schymick, D.G. Hernandez, B.J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H.M. Cann, J.A. Hardy, N.A. Rosenberg, A.B. Singleton, Genotype, haplotype and copy-number variation in worldwide human populations, Nature 451 (2008) 998–1003.
[6] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, C.D. Bustamante, Genes mirror geography within Europe, Nature 456 (2008) 98–101.
[7] J. Xing, W.S. Watkins, D.J. Witherspoon, Y. Zhang, S.L. Guthery, R. Thara, B.J. Mowry, K. Bulayeva, R.B. Weiss, L.B. Jorde, Fine-scaled human genetic structure revealed by SNP microarrays, Genome Res. 19 (2009) 815–825.
[8] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, L. Singh, Reconstructing Indian population history, Nature 461 (2009) 489–494.
[9] M.A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S.K. Brahmachari, G.C. Calacal, A. Chaurasia, C.H. Chen, J. Chen, Y.T. Chen, J. Chu, E.M. Cutiongco-de la Paz, M.C. De Ungria, F.C. Delfin, J. Edo, S. Fuchareon, H. Ghang, T. Gojobori, J. Han, S.F. Ho, B.P. Hoh, W. Huang, H. Inoko, P. Jha, T.A. Jinam, L. Jin, J. Jung, D. Kangwanpong, J. Kampuansai, G.C. Kennedy, P. Khurana, H.L. Kim, K. Kim, S. Kim, W.Y. Kim, K. Kimm, R. Kimura, T. Koike, S. Kulawonganunchai, V. Kumar, P.S. Lai, J.Y. Lee, S. Lee, E.T. Liu, P.P. Majumder, K.K. Mandapati, S. Marzuki, W. Mitchell, M. Mukerji, K. Naritomi, C. Ngamphiw, N. Niikawa, N. Nishida, B. Oh, S. Oh, J. Ohashi, A. Oka, R. Ong, C.D. Padilla, P. Palittapongarnpim, H.B. Perdigon, M.E. Phipps, E. Png, Y. Sakaki, J.M. Salvador, Y. Sandraling, V. Scaria, M. Seielstad, M.R. Sidek, A. Sinha, M. Srikummool, H. Sudoyo, S. Sugano, H. Suryadi, Y. Suzuki, K.A. Tabbada, A. Tan, K. Tokunaga, S. Tongsima, L.P. Villamor, E. Wang, Y. Wang, H. Wang, J.Y. Wu, H. Xiao, S. Xu, J.O. Yang, Y.Y. Shugart, H.S. Yoo, W. Yuan, G. Zhao, B.A. Zilfalil, Mapping human genetic diversity in Asia, Science 326 (2009) 1541–1545.
[10] S.A. Tishkoff, F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J.B. Hirbo, A.A. Awomoyi, J.M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M.J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, S.M. Williams, The genetic structure and history of Africans and African Americans, Science 324 (2009) 1035–1044.
[11] M.F. Seldin, R. Shigeta, P. Villoslada, C. Selmi, J. Tuomilehto, G. Silva, J.W. Belmont, L. Klareskog, P.K. Gregersen, European population substructure: clustering of northern and southern populations, PLoS Genet. 2 (2006) e143.
[12] M. Bauchet, B. McEvoy, L.N. Pearson, E.E. Quillen, T. Sarkisian, K. Hovhannesyan, R. Deka, D.G. Bradley, M.D. Shriver, Measuring European population stratification with microarray genotype data, Am. J. Hum. Genet. 80 (2007) 948–956.
[13] A.L. Price, J. Butler, N. Patterson, C. Capelli, V.L. Pascali, F. Scarnicci, A. Ruiz-Linares, L. Groop, A.A. Saetta, P. Korkolopoulou, U. Seligsohn, A. Waliszewska, C. Schirmer, K. Ardlie, A. Ramos, J. Nemesh, L. Arbeitman, D.B. Goldstein, D. Reich, J.N. Hirschhorn, Discerning the ancestry of European Americans in genetic association studies, PLoS Genet. 4 (2008) e236.
[14] S.L. Guthery, B.A. Salisbury, M.S. Pungliya, J.C. Stephens, M. Bamshad, The structure of common genetic variation in United States populations, Am. J. Hum. Genet. 81 (2007) 1221–1231.
[15] I. Silva-Zolezzi, A. Hidalgo-Miranda, J. Estrada-Gil, J.C. Fernandez-Lopez, L. Uribe-Figueroa, A. Contreras, E. Balam-Ortiz, L. del Bosque-Plata, D. Velazquez-Fernandez, C. Lara, R. Goya, E. Hernandez-Lemus, C. Davila, E. Barrientos, S. March, G. Jimenez-Sanchez, Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico, Proc. Natl Acad. Sci. USA 106 (2009) 8611–8616.
[16] J.L. Kelley, J. Madeoy, J.C. Calhoun, W. Swanson, J.M. Akey, Genomic signatures of positive selection in humans and the limits of outlier approaches, Genome Res. 16 (2006) 980–989.

[17] L.B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans, Nat. Genet. 40 (2008) 340–345.
[18] P.C. Sabeti, S.F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T.S. Mikkelsen, D. Altshuler, E.S. Lander, Positive natural selection in the human lineage, Science 312 (2006) 1614–1620.
[19] R.L. Lamason, M.A. Mohideen, J.R. Mest, A.C. Wong, H.L. Norton, M.C. Aros, M.J. Jurynec, X. Mao, V.R. Humphreville, J.E. Humbert, S. Sinha, J.L. Moore, P. Jagadeeswaran, W. Zhao, G. Ning, I. Makalowska, P.M. McKeigue, D. O'Donnell, R. Kittles, E.J. Parra, N.J. Mangini, D.J. Grunwald, M.D. Shriver, V.A. Canfield, K.C. Cheng, SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans, Science 310 (2005) 1782–1786.
[20] T.S. Simonson, Y. Yang, C.D. Huff, H. Yun, G. Qin, D.J. Witherspoon, Z. Bai, F.R. Lorenzo, J. Xing, L.B. Jorde, J.T. Prchal, R. Ge, Genetic evidence for high-altitude adaptation in Tibet, Science 329 (2010) 72–75.
[21] Affymetrix, Use of Saliva gDNA for SNP GenotypingTechnical Note, http://media.affymetrix.com/support/technical/technotes/saliva_gDNA_genotyping.pdf 2008.
[22] D.J. Schaid, A.J. Batzler, G.D. Jenkins, M.A. Hildebrandt, Exact tests of Hardy–Weinberg equilibrium and homogeneity of disequilibrium across strata, Am. J. Hum. Genet. 79 (2006) 1071–1080.
[23] S.A. Stouffer, E.A. Suchman, L.C. DeVinney, S.A. Star, R.M.J. Williams, The American Soldier: Adjustment During Army Life, Princeton University Press, Princeton, 1949.
[24] J.M. Korn, F.G. Kuruvilla, S.A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P.J. Collins, K. Darvishi, C. Lee, M.M. Nizzari, S.B. Gabriel, S. Purcell, M.J. Daly, D. Altshuler, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, Nat. Genet. 40 (2008) 1253–1260.
[25] D. Pinto, C. Marshall, L. Feuk, S.W. Scherer, Copy-number variation in control population cohorts, Hum. Mol. Genet. 16 (2007) Spec No. 2 (2007) R168-73.
[26] M. Nei, F. Tajima, DNA polymorphism detectable by restriction endonucleases, Genetics 97 (1981) 145–163.
[27] J.J. Cai, PGEToolbox: a Matlab toolbox for population genetics and evolution, J. Hered. 99 (2008) 438–440.
[28] J. Felsenstein, PHYLIP (Phylogeny Inference Package) Version 3.6, Department of Genome Sciences, University of Washington, Seattle, 2004 Distributed by the Author.
[29] B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution 38 (1984) 1358–1370.
[30] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19 (2009) 1655–1664.
[31] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.
[32] S.C. Schuster, W. Miller, A. Ratan, L.P. Tomsho, B. Giardine, L.R. Kasson, R.S. Harris, D.C. Petersen, F. Zhao, J. Qi, C. Alkan, J.M. Kidd, Y. Sun, D.I. Drautz, P. Bouffard, D.M. Muzny, J.G. Reid, L.V. Nazareth, Q. Wang, R. Burhans, C. Riemer, N.E. Wittekindt, P. Moorjani, E.A. Tindall, C.G. Danko, W.S. Teo, A.M. Buboltz, Z. Zhang, Q. Ma, A. Oosthuysen, A.W. Steenkamp, H. Oostuisen, P. Venter, J. Gajewski, Y. Zhang, B.F. Pugh, K.D. Makova, A. Nekrutenko, E.R. Mardis, N. Patterson, T.H. Pringle, F. Chiaromonte, J.C. Mullikin, E.E. Eichler, R.C. Hardison, R.A. Gibbs, T.T. Harkins, V.M. Hayes, Complete Khoisan and Bantu genomes from southern Africa, Nature 463 (2010) 943–947.
[33] S.A. McCarroll, F.G. Kuruvilla, J.M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M.H. Shapero, P.I. de Bakker, J.B. Maller, A. Kirby, A.L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P.J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K.W. Jones, R. Rava, M.J. Daly, S.B. Gabriel, D. Altshuler, Integrated detection and population-genetic analysis of SNPs and copy number variation, Nat. Genet. 40 (2008) 1166–1174.
[34] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F. Grant, H. Hakonarson, M. Bucan, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, Genome Res. 17 (2007) 1665–1674.
[35] D.F. Conrad, T.D. Andrews, N.P. Carter, M.E. Hurles, J.K. Pritchard, A high-resolution survey of deletion polymorphism in the human genome, Nat. Genet. 38 (2006) 75–81.
[36] D.A. Hinds, A.P. Kloek, M. Jen, X. Chen, K.A. Frazer, Common deletions and SNPs are in linkage disequilibrium in the human genome, Nat. Genet. 38 (2006) 82–85.
[37] D.J. Witherspoon, E.E. Marchani, W.S. Watkins, C.T. Ostler, S.P. Wooding, B.A. Anders, J.D. Fowlkes, S. Boissinot, A.V. Furano, D.A. Ray, A.R. Rogers, M.A. Batzer, L.B. Jorde, Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions, Hum. Hered. 62 (2006) 30–46.
[38] M.D. Shriver, R. Mei, E.J. Parra, V. Sonpar, I. Halder, S.A. Tishkoff, T.G. Schurr, S.I. Zhadanov, L.P. Osipova, T.D. Brutsaert, J. Friedlaender, L.B. Jorde, W.S. Watkins, M.J. Bamshad, G. Gutierrez, H. Loi, H. Matsuzaki, R.A. Kittles, G. Argyropoulos, J.R. Fernandez, J.M. Akey, K.W. Jones, Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation, Hum. Genomics 2 (2005) 81–89.
[39] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, M.W. Feldman, Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1 (2005) e70.
[40] L.J. Handley, A. Manica, J. Goudet, F. Balloux, Going the distance: human population genetics in a clinal world, Trends Genet. 23 (2007) 432–439.
[41] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, L.L. Cavalli-Sforza, 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 (2005) 15942–15947.
[42] R.G. Klein, Darwin and the recent African origin of modern humans, Proc. Natl Acad. Sci. USA 106 (2009) 16007–16009.
[43] J.F. Hoffecker, Out of Africa: modern human origins special feature: the spread of modern humans in Europe, Proc. Natl Acad. Sci. USA 106 (2009) 16040–16045.
[44] S. Wright, Evolution in Mendelian populations, Genetics 16 (1931) 97–159.
[45] C.B. Stringer, R. Grun, H.P. Schwarcz, P. Goldberg, ESR dates for the hominid burial site of Es Skhul in Israel, Nature 338 (1989) 756–758.
[46] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, C.D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data, PLoS Genet. 5 (2009) e1000695.
[47] R.E. Green, J. Krause, A.W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M.H. Fritz, N.F. Hansen, E.Y. Durand, A.S. Malaspinas, J.D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H.A. Burbano, J.M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E.S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V.B. Doronichev, L.V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R.W. Schmitz, P.L. Johnson, E.E. Eichler, D. Falush, E. Birney, J.C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Paabo, A draft sequence of the Neandertal genome, Science 328 (2010) 710–722.
[48] M. Bamshad, T. Kivisild, W.S. Watkins, M.E. Dixon, C.E. Ricker, B.B. Rao, J.M. Naidu, B.V. Prasad, P.G. Reddy, A. Rasanayagam, S.S. Papiha, R. Villems, A.J. Redd, M.F. Hammer, S.V. Nguyen, M.L. Carroll, M.A. Batzer, L.B. Jorde, Genetic evidence on the origins of Indian caste populations, Genome Res. 11 (2001) 994–1004.
[49] W.S. Watkins, R. Thara, B.J. Mowry, Y. Zhang, D.J. Witherspoon, W. Tolpinrud, M.J. Bamshad, S. Tiripati, R. Padmavati, H. Smith, D. Nancarrow, C. Filippich, L.B. Jorde, Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms, BMC Genet. 9 (2008) 86.
[50] Government of Nepal, in: C.B.o. Statistics (Ed.), Statistical Year Book of Nepal, Kathmandu, 2007.
[51] S. Fornarino, M. Pala, V. Battaglia, R. Maranta, A. Achilli, G. Modiano, A. Torroni, O. Semino, S.A. Santachiara-Benerecetti, Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation, BMC Evol. Biol. 9 (2009) 154.
[52] T. Gayden, S. Mirabal, A.M. Cadenas, H. Lacau, T.M. Simms, D. Morlote, S. Chennakrishnaiah, R.J. Herrera, Genetic insights into the origins of Tibeto-Burman populations in the Himalayas, J. Hum. Genet. 54 (2009) 216–223.
[53] T. Gayden, A.M. Cadenas, M. Regueiro, N.B. Singh, L.A. Zhivotovsky, P.A. Underhill, L.L. Cavalli-Sforza, R.J. Herrera, The Himalayas as a directional barrier to gene flow, Am. J. Hum. Genet. 80 (2007) 884–894.
[54] T. Goebel, M.R. Waters, D.H. O'Rourke, The late Pleistocene dispersal of modern humans in the Americas, Science 319 (2008) 1497–1502.
[55] D.H. O'Rourke, J.A. Raff, The human genetic history of the Americas: the final frontier, Curr. Biol. 20 (2010) R202–R207.
[56] C.J. Mulligan, K. Hunley, S. Cole, J.C. Long, Population genetics, history, and health patterns in native Americans, Annu. Rev. Genomics Hum. Genet. 5 (2004) 295–315.
[57] S. Wang, C.M. Lewis, M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M.V. Parra, J.A. Molina, C. Gallo, G. Mazzotti, G. Poletti, K. Hill, A.M. Hurtado, D. Labuda, W. Klitz, R. Barrantes, M.C. Bortolini, F.M. Salzano, M.L. Petzl-Erler, L.T. Tsuneto, E. Llop, F. Rothhammer, L. Excoffier, M.W. Feldman, N.A. Rosenberg, A. Ruiz-Linares, Genetic variation and population structure in native Americans, PLoS Genet. 3 (2007) e185.
[58] C.J. Mulligan, A. Kitchen, M.M. Miyamoto, Updated three-stage model for the peopling of the Americas, PLoS ONE 3 (2008) e3199.
[59] N.J. Fagundes, R. Kanitz, R. Eckert, A.C. Valls, M.R. Bogo, F.M. Salzano, D.G. Smith, W.A. Silva Jr., M.A. Zago, A.K. Ribeiro-dos-Santos, S.E. Santos, M.L. Petzl-Erler, S.L. Bonatto, Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas, Am. J. Hum. Genet. 82 (2008) 583–592.
[60] U.A. Perego, A. Achilli, N. Angerhofer, M. Accetturo, M. Pala, A. Olivieri, B.H. Kashani, K.H. Ritchie, R. Scozzari, Q.P. Kong, N.M. Myres, A. Salas, O. Semino, H.J. Bandelt, S.R. Woodward, A. Torroni, Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups, Curr. Biol. 19 (2009) 1–8.
[61] N. Ray, D. Wegmann, N.J. Fagundes, S. Wang, A. Ruiz-Linares, L. Excoffier, A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with Asia, Mol. Biol. Evol. 27 (2010) 337–345.
[62] C.J. Kolman, N. Sambuughin, E. Bermingham, Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders, Genetics 142 (1996) 1321–1334.
[63] D.A. Merriwether, F. Rothhammer, R.E. Ferrell, Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World, Am. J. Phys. Anthropol. 98 (1995) 411–430.
[64] W.S. Watkins, A.R. Rogers, C.T. Ostler, S. Wooding, M.J. Bamshad, A.M. Brassington, M.L. Carroll, S.V. Nguyen, J.A. Walker, B.V. Prasad, P.G. Reddy, P.K. Das, M.A. Batzer, L.B. Jorde, Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms, Genome Res. 13 (2003) 1607–1618.
[65] L.B. Jorde, A.R. Rogers, M. Bamshad, W.S. Watkins, P. Krakowiak, S. Sung, J. Kere, H.C. Harpending, Microsatellite diversity and the demographic history of modern humans, Proc. Natl Acad. Sci. USA 94 (1997) 3100–3103.
[66] L.B. Jorde, W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, M.A. Batzer, The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y chromosome data, Am. J. Hum. Genet. 66 (2000) 979–988.

Supplemental Figures for:

Toward a more Uniform Sampling of Human Genetic Diversity: A Survey of Worldwide Populations by High-density Genotyping

Jinchuan Xing, W. Scott Watkins, Adam Shlien, Erin Walker, Chad D. Huff, David J. Witherspoon, Yuhua Zhang, Tatum S. Simonson, Robert B. Weiss, Joshua D. Schiffman, David Malkin, Scott R. Woodward, and Lynn B. Jorde

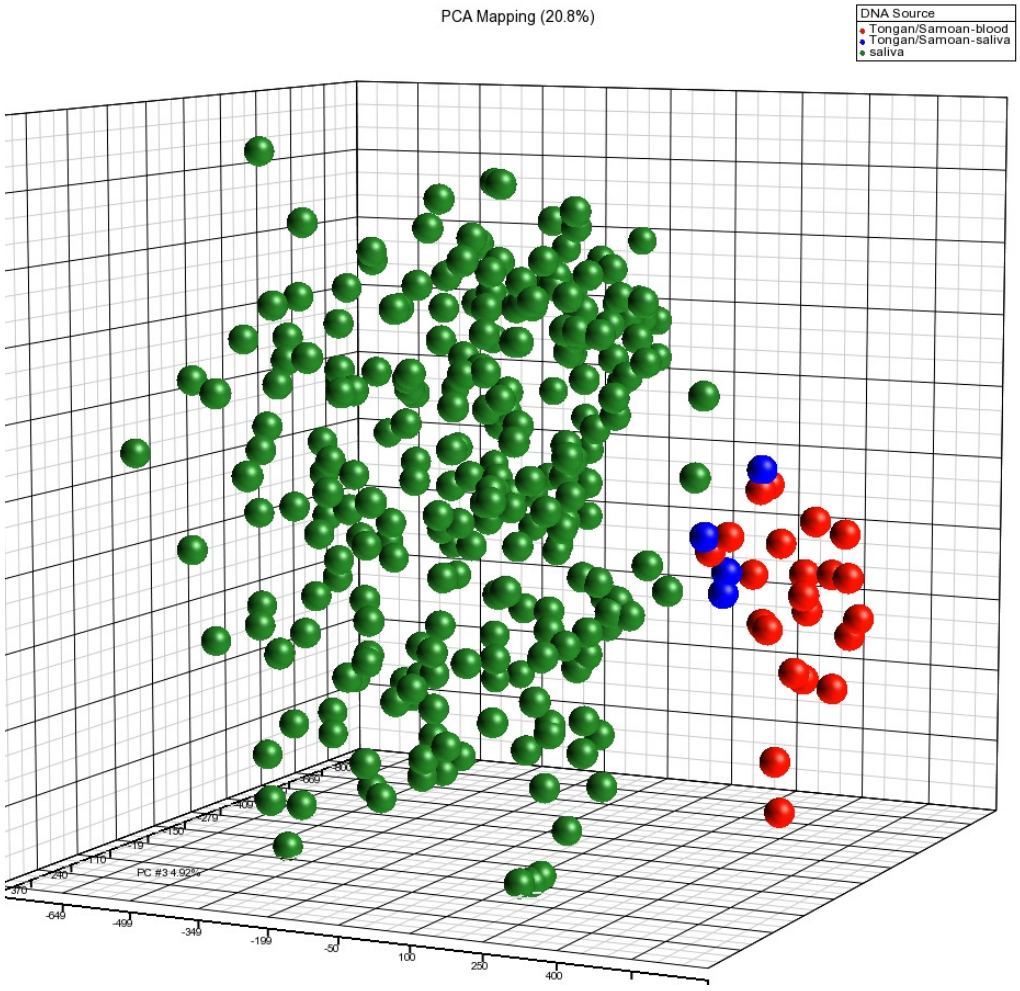
Supplemental Figure 1. PCA of blood vs saliva-derived DNA samples. PCA was performed in Partek Genomics Suite, using sample covariance as the dispersion matrix. For Tongan and Samoan samples, saliva-derived samples are in blue and blood-derived samples are in red. All other samples are saliva-derived and are in green.

Supplemental Figure 2. Bootstrap consensus tree of 40 populations. Populations are color-coded based on their continental origins. The hypothetical ancestral population is shown. Bootstrap value for each branch is shown.

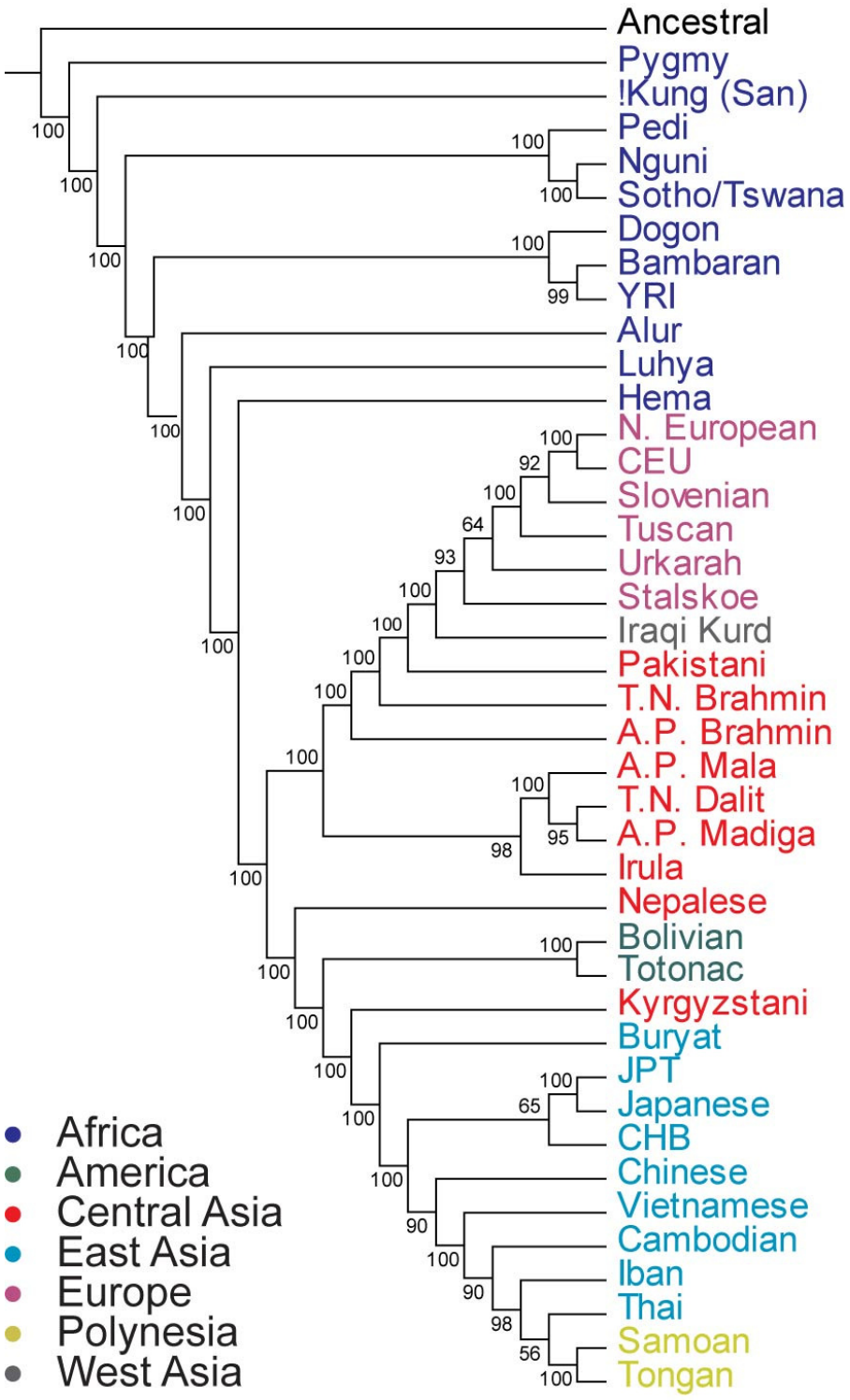
Supplemental Figure 3. Neighbour-joining tree of 80 populations. Populations are color-coded based on their continental origins. The hypothetical ancestral population is shown. Several populations in the HGDP panel were grouped together to improve the resolution. The modified group include: HGDP Bantu (Bantu Kenya, Herero, Nguni, Ovambo, Pedi, Sotho, and Tswana); HGDP N. Chinese (Oroqen, Daur, Hezhen, Mongola, Xibo, and Tu), HGDP S. Chinese (Dai, Lahu, Miao, Naxi, She, Tujia, Yi) and HGDP Han (Han and Han-NChina).

Supplemental Figure 4: PCA of population structure. A) All individuals; B) Europe and West Asia; C) Africa; D) Eurasia, Polynesia, and America. PC3 and PC4 (A) or PC1 and PC2 (B, C, D) are shown. Each individual is represented by one dot and the color label corresponding to their regional origin (A) or population (B, C, D). The percentage of variance explained by each PC is shown on the axis.

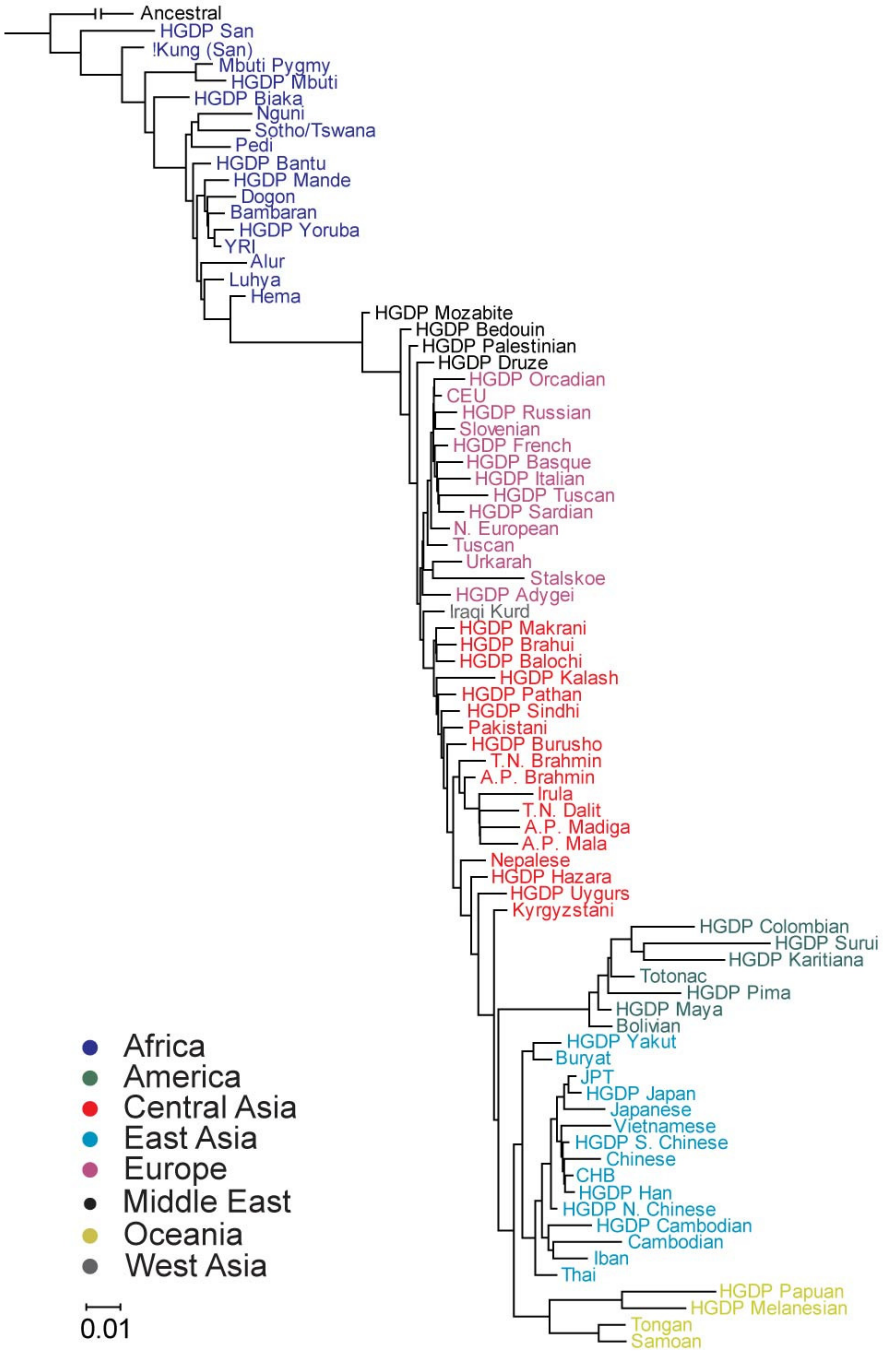
Supplemental Figure 5: PCA of genome-wide CNVs for all populations.



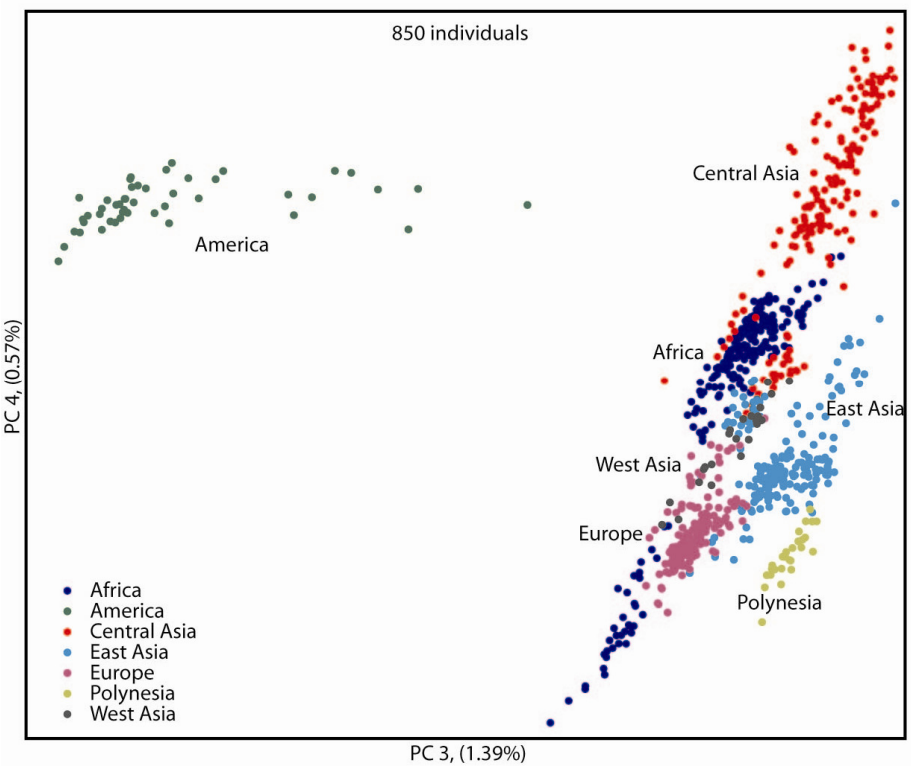
Supplemental Figure 1



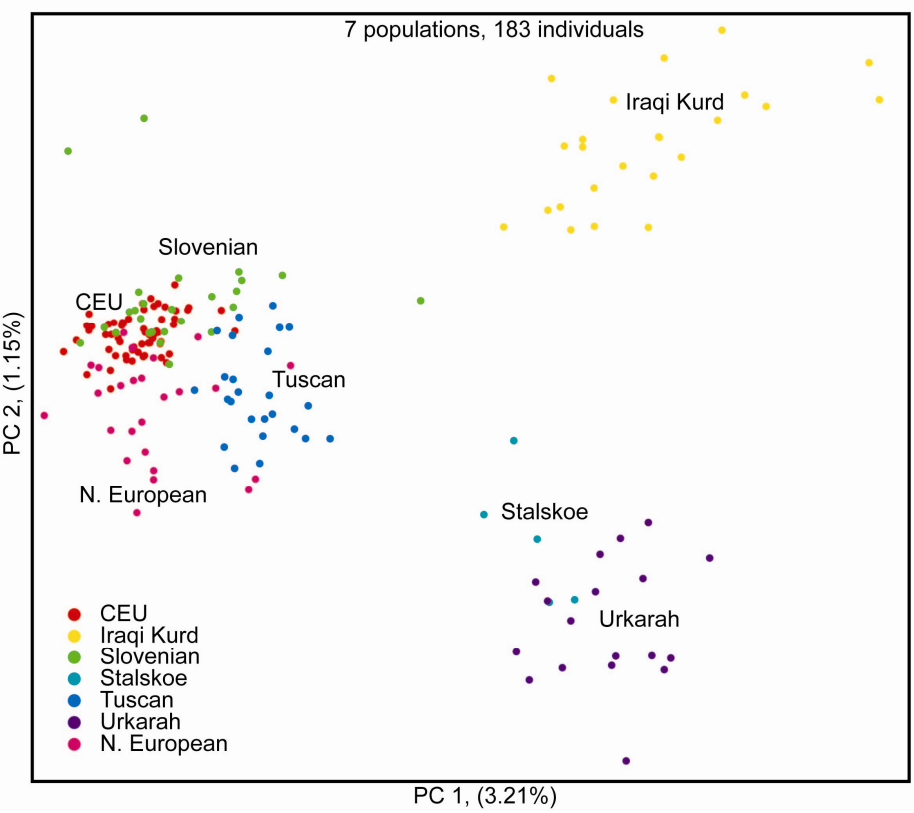
Supplemental Figure 2



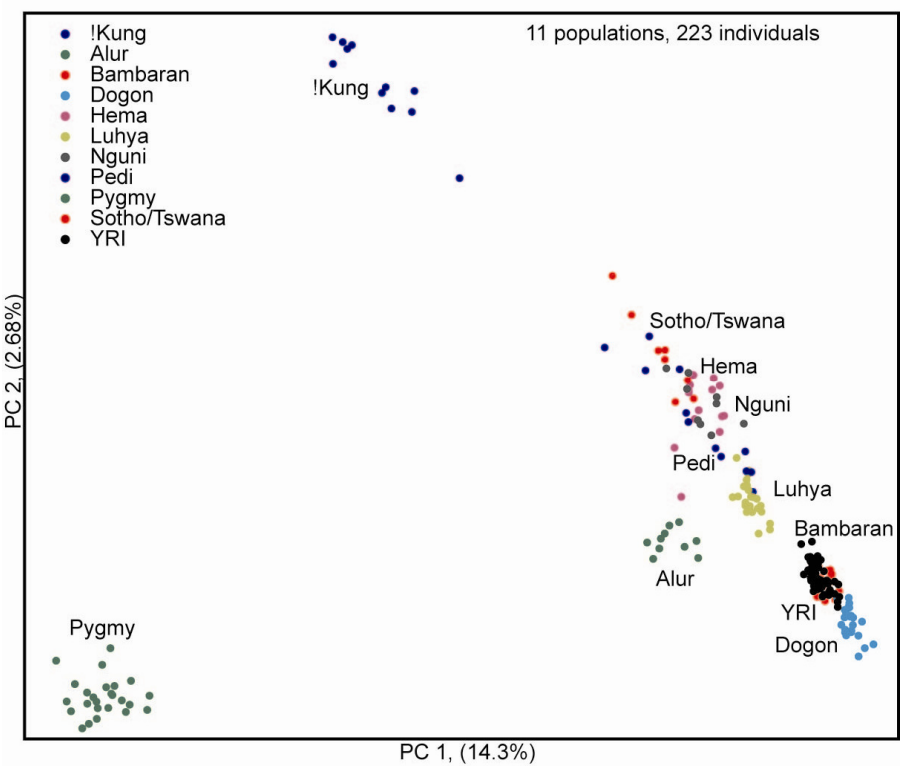
Supplemental Figure 3



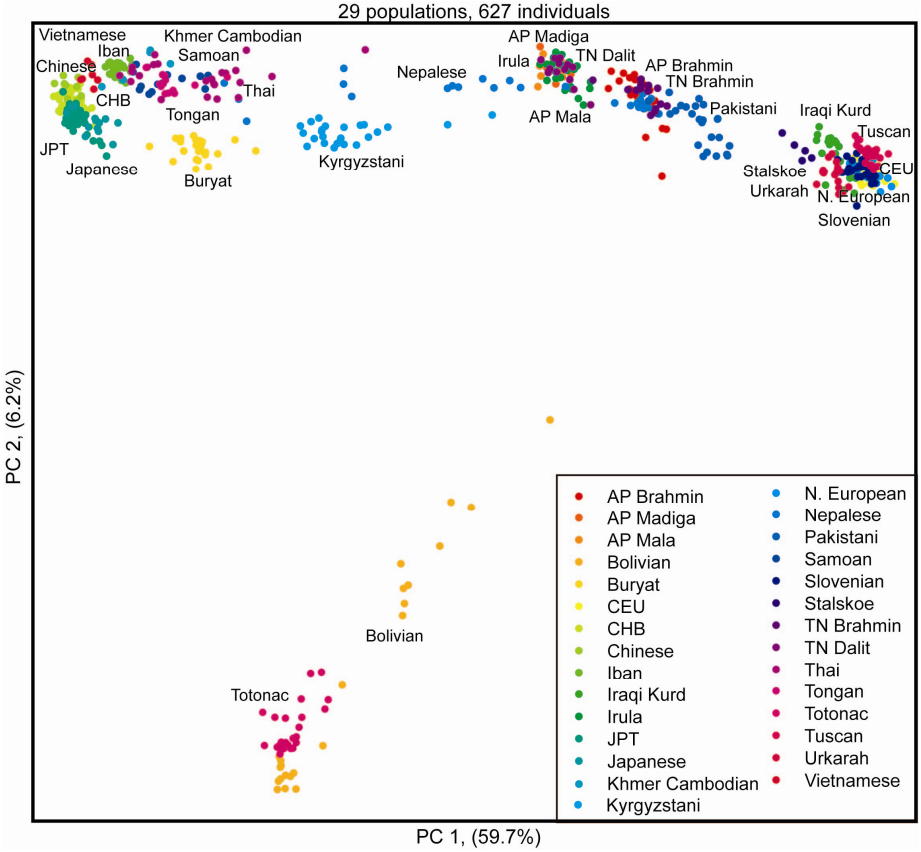
Supplemental Figure 4A



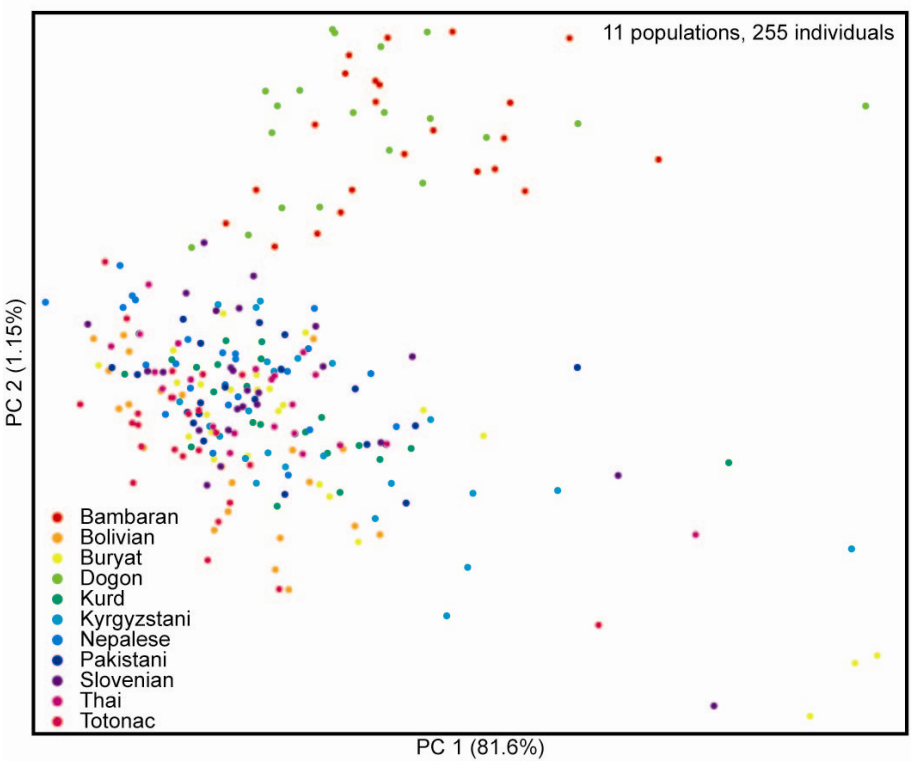
Supplemental Figure 4B



Supplemental Figure 4C



Supplemental Figure 4D



Supplemental Figure 5