Global distribution of genomic diversity underscores rich complex history of continental human populations

Adam Auton, Katarzyna Bryc, Adam R. Boyko, Kirk E. Lohmueller, John Novembre, Andy Reynolds, Amit Indap, Mark H. Wright, Jeremiah Degenhardt, Ryan N. Gutenkunst, Karen S. King, Matthew R. Nelson, Carlos D. Bustamante

February 1, 2009

SUPPLEMENTAL MATERIAL

Minor Allele Frequency Spectrum

The Affymetrix GeneChip provides a non-random sample of SNPs in the genome, with SNPs selected based on the catalog of known variants, frequency, and assay design considerations. The observed minor allele frequency (MAF; Supplementary Figure S2A) spectrum is therefore not representative of the underlying true population allele frequency distribution. Nonetheless, patterns of correlated allele frequencies among populations (which largely reflect the history of divergence and migration between populations) can provide novel insights into average genealogical relationships among individuals from different populations (Supplementary Figure S2B). From comparing the joint site-frequency spectra of common variation, we find SNP frequencies are more strongly correlated between Europe and South Asia than between East and South Asia. This result is consistent with a more severe founding bottleneck in the history of East Asian populations as well as less gene flow between South and East Asia than between South Asia and Europe.

Estimation of F_{ST}

F_{ST} was calculated using the 'strict' individuals from each population. We estimated F_{ST} for each SNP using the method of Weir and Cockerham [Weir and Cockerham 1984]. Specifically, we use equation 6 in that paper, and for clarity, we repeat the formula here:

$$\hat{F}_{ST} = \frac{s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}\,(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{\bar{h}}{4}\right]}{\left[1 - \frac{\bar{n}C^2}{(\bar{n}-1)r}\right]\bar{p}\,(1-\bar{p}) + \left[1 + \frac{\bar{n}(r-1)C^2}{(\bar{n}-1)r}\right]\frac{s^2}{r} + \left[\frac{C^2}{(\bar{n}-1)r}\right]\frac{\bar{h}}{4}}$$

where s^2 is the sample variance of allele frequencies over populations, $\bar{n}$ is the mean sample size, $\bar{p}$ is the mean sample allele frequency, r is the number of sub-populations, $\bar{h}$ is the mean heterozygote frequency in the sample, and C^2 is the squared coefficient of variation of the sample sizes. Further details are given in the cited paper. X chromosome estimates were obtained using only the female individuals in the study. To obtain a single estimate of F_{ST} for the complete data set, we combined estimates from all SNPs with a defined F_{ST} estimate using the weighted average scheme described in same paper (c.f. equation 10 in [Weir and Cockerham 1984]).

To estimate the expected value of F_{ST} for the X chromosome based on autosomal F_{ST}, we use a standard result from population genetics that for an idealized Wright-Fisher population with migration among many demes, the expected value of F_{ST} is simply:

$$E(F_{ST}) = \frac{1}{1 + 4Nm}$$

where $2Nm$ is the number of migrants entering each deme every generation (see, for example, [Hartl and Clark 2007]). Under this condition, one can invert the expression to estimate Nm for the autosomes as $\widehat{Nm} = \frac{1}{4}\frac{1-F_{ST}}{F_{ST}}$. Under equal migration of males and females, equal variance in offspring number, and equal population size of the two sexes, the expected value for the X chromosome based on autosomal F_{ST} is then $\frac{1}{1+3\widehat{Nm}}$.

Subcontinental Population Structure Analysis

With the exception of Europe, sub-continental population structure analyses are described below. European population structure in the POPRES has been discussed elsewhere [Novembre et al. 2008], with the large sample size allowing population structure to be observed at a fine-scale. However, the POPRES provides evidence of structure in the other continental populations, even with their smaller sample sizes. In the following section, we describe patterns of population structure at a subcontinental level using both *STRUCTURE* and Principal Component Analysis (PCA). Note that markers were selected independently in each of the following analyses.

East Asia: For East Asia, we analyzed the POPRES individuals combined with the Han Chinese (CHB) and Japanese (JPT) samples from the HapMap. Using the subset of 271 individuals from East Asia, we ran *STRUCTURE* on 6,422 randomly selected SNPs with MAF > 0.2 (within East Asia) spaced 400kb apart. The results are shown in Supplementary Figure S3B. As expected, at $K = 2$ we see two clear clusters separating the Japanese populations from the Chinese. At $K = 3$ we see that sections of the two different HapMap populations cluster together, reducing the proportion of genomes differentiated between Japanese and Chinese individuals. Further increasing K increases substructure within our POPRES samples not corresponding to known geographic structure.

In the PCA of the East Asian populations, we see clear separation between the Japanese and Taiwanese/Chinese samples (Figure 1C), with PC 1 separating the Japanese samples from Taiwan and the CHB — a pattern also seen in the *STRUCTURE* analysis. The second PC separates Taiwan from the HapMap Han Chinese, reflecting the geographic distance between these populations. To

a much lesser extent, the second PC also separates the POPRES Japanese from the HapMap Japanese. We note that the HapMap individuals were sampled in Tokyo, Japan, whereas the POPRES Japanese were sampled in Sydney, Australia [Nelson et al. 2008]. In the absence of further ancestral information, it is difficult to assess whether the small observed separation between the Japanese samples is due to subtle genotyping platform differences or true genetic differences.

South Asia: South Asian individuals were sampled as part of the LOLIPOP study in London, England, and we do not have data regarding parental or grand-parental ancestry of these individuals. However, we do have information regarding self-identified country of origin and spoken language with 10 languages represented. For categorization purposes, we note that Malayalam and Tamil are Dravidian languages, and Konkani and Sinhalese both have borrowed words from Dravidian languages [Emeneau and Burrow 1962], and we therefore group these languages into a "Dravidian Influenced" group. The six remaining languages we simply term as "Non-Dravidian Influenced" in subsequent analyses (Supplementary Table S1). The Dravidian Influenced languages are predominately spoken in southern India (Supplementary Figure S4).

We ran *STRUCTURE* using the same parameters as the global analysis but using 315 individuals from India and Sri Lanka having excluded individuals with no language information, have English as a primary language or are related. We used 6,542 SNPs having MAF > 0.2 (within South Asia) and separation of at least 400kb. Individuals were classified by self-reported language spoken. We excluded individuals without language information and those whose primary self-reported language was English. The results are shown in Supplementary Figure S3C. At $K = 2$, there is no structure consistent with language groups. However, at $K = 3$, we note that languages spoken in the south of India and Sri Lanka, including the Dravidian languages Malayalam and Tamil cluster together, as well as some Gujarati individuals. Sinhalese and Tamil are the two officially recognized languages of Sri Lanka. Malayalam is spoken along the tropical Malabar Coast of southwestern India, near Sri Lanka. Konkani is mostly spoken along the section of the south-western coastline of India known as Konkan, also near Sri Lanka. Further increasing the number of clusters to $K = 4$ increases admixture without any geographic or linguistic correlation.

Mexico: As discussed in the main text, we quantified the admixture in Mexicans using a *STRUCTURE* analysis of Mexicans, Europeans and East Asians. We extracted 778 individuals from the POPRES, comprising of 107 Mexican individuals, 400 randomly selected European individuals with known European grandparents and 271 East Asian individuals (including 90 HapMap individuals). We used 6,557 SNPs with MAF in these populations of > 0.2 (within Mexico) and spaced at least 400kb apart. The results are shown in Supplementary Figure S3D. At $K = 2$, the Mexican individuals appear admixed between a predominately European cluster and a predominately East Asian cluster, with slightly greater membership in the former cluster. However, at $K = 3$, the Mexicans form their own cluster and no longer share East Asian admixture, but retain a 'European' admixture component. The average proportion of European admixture in Mexican individuals with $K = 3$ is 32.5% with a standard deviation of 17.4%. Further increasing K only reveals further admixture among European populations or separates the Japanese and Chinese populations.

We repeated the analysis using the 'supervised' *STRUCTURE* mode, having pre-assigned European and East Asian individuals to their respective populations. A $K = 3$, we found this method

to give similar results to the unsupervised mode, with a European admixture component of 35.0% (standard deviation 16.8%) in Mexican individuals.

The first two principal components of the same individuals demonstrates a similar patter (Figure 1B), with Mexican individuals forming a distinct cluster between the European and East Asian Clusters in the first principal component. However, the second PC further differentiates the Mexican individuals from the East Asian individuals without substantially increasing the separation from Europeans.

Comparison with HGDP

While the global *STRUCTURE* analysis reveals broad patterns of population differentiation (Supplementary Figure S3), the method is limited to using a small fraction of the available SNPs due to high computational cost. Furthermore, as the number of specified clusters is increased, the patterns of population structure become increasingly difficult to interpret. As an alternative means for analyzing population structure, we conducted a PCA of the genotype data [Patterson et al. 2006]. This method has the advantage of being able to analyze many more SNPs and can flexibly summarize patterns of both discrete [Patterson et al. 2006] and continuous spatial [Novembre and Stephens 2008] population structure. PCA analysis of the POPRES alone is considered in Nelson et al. 2008. To investigate how the POPRES complements known patterns of global diversity, we combined the 2,943 "strict" individuals in the POPRES dataset with 479 individuals from the HGDP genotype data [Jakobsson et al. 2008] for a combined total of 3,448 individuals. Although the two datasets were generated on separate genotyping platforms, more than 73,520 SNPs are shared even after pruning SNPs in high linkage disequilibrium (LD) and those with more than 5% missing data.

The first two principal components (PCs) of the combined dataset separate individuals into clusters largely determined by geographic origin (Supplementary Figure S1A), which is consistent with a previous analysis of the HGDP dataset [Li et al. 2008]. Individuals from East Asia and Europe in the POPRES tend to cluster more tightly than those from the HGDP study. This is to be expected, as the POPRES samples are taken from presumably well-mixed urban populations whereas the HGDP sample is largely composed of diverse isolated populations (e.g. Basque, Sardinian, and Orcadians within Europe). Both the Mexican and South Asian individuals cluster between the European and East Asian clusters in this projection. The next two PCs reveal further structure within the Asian / American clusters, separating the Asian individuals from the American individuals (Supplementary Figure S1B). Notably, the POPRES Mexican individuals form a new cluster between the predominately European cluster and the Native American cluster, which is indicative of the historical admixture of Europeans with Native Americans.

Phasing of the Data

For the estimation of haplotype diversity and population recombination rates, we first used the program *BEAGLE* version 2.1.3 to phase the genotype data [Browning and Browning 2007]. This

method was chosen as it is currently one of the few available methods that can phase a dataset of this size in a reasonable time. Each sub-continental population in the strict dataset was phased separately. The default parameters were used with the exception of the European samples, for which we set nsamples=1 as recommended in the documentation for large samples. We phased the X chromosome separately, using an unpublished version of *BEAGLE* (version 2.2.0) that makes use of the known phase of the male samples.

Haplotype Diversity

To test whether the mean of the distribution of the number of haplotypes is informative of recent population demography, we conducted coalescent simulations using *ms* [Hudson 2002]. We considered a family of demographic models (Supplementary Figure S6) where in the present day there are two separate subpopulations, one of size N_c = 10,000 and the other of size N_c = 5,000. These two subpopulations do not exchange any migrants. Going back in time, at τ years ago, the two populations join and form an ancestral panmictic population of size N_a = 10,000. We examined a range of four different values of τ (0, 5,000, 10,000, and 20,000 years ago) for the population split times. To match our observed data, we sampled 146 chromosomes from each subpopulation and simulated 5,000 independent 500kb regions with an average per-generation recombination rate of 1cM/Mb. The *ms* command line used for these simulations is:

./ms 292 5000 -t 300 -r 200 500001 -I 2 146 146 0 -en 0 2 0.5 -ej τ 2 1 -F 29

where τ varies between simulations. Note that the mutation rate is set to be an arbitrary value, and does not matter given our sampling strategy of selecting a subset of SNPs (see below). We converted τ from generations to years assuming 20 years per generation.

In our analysis of the observed data, we only considered SNPs with MAF $>$ 10% in all subpopulations. We implemented a similar filtering strategy in our simulations. In each simulation replicate, we selected a subset of 25 SNPs with MAF $>$ 10% in both of the subpopulations. We selected the same set of SNPs for each subpopulation. Using these SNPs, we parsed the haplotypes found in each subpopulation and then counted the number of haplotypes in each subpopulation for each of the 5,000 simulation replicates.

Supplementary Figure S6 shows the results of this analysis. Note that if the two populations (going backwards in time) joined immediately (τ = 0), we do not see a difference in the distribution of the number of haplotypes between the two populations. However, for the other values of τ, we consistently see that for the smaller subpopulation (dotted lines), the distribution of the number of haplotypes is lower than that for the larger population (solid lines). We also see that as the time since the population split increases, the smaller subpopulation has fewer and fewer haplotypes (compare τ = 5,000 years to τ = 20,000 years) as expected. These results suggest that the distribution of the number of haplotypes can be informative about recent demographic history.

In the main text, we analyzed populations with at least 73 individuals. For this reason, the Dravidian Influenced group was not included. However, using a thinned sample of 20 individuals

per population, we were able to compare the Dravidian Influenced group to the other populations S2. We see that the two South Asian populations have similar levels of haplotype diversity. For the other populations, the relative levels of diversity are nearly identical to the analysis using 73 individuals.

To understand how the haplotype diversity statistics are influenced by SNP ascertainment bias, we conducted additional coalescent simulations using the same two-population split model with τ fixed at 20,000 years. We simulated a genotype sample of 146 chromosomes from each population and a SNP discovery sample of four chromosomes in each population. The two genotype samples did not include any of the chromosomes used for SNP discovery. We considered four different ascertainment strategies relevant for the Affymetrix 500K data: 1) only considering SNPs polymorphic in two discovery chromosomes from the smaller population, 2) only considering SNPs polymorphic in four discovery chromosomes from the smaller population, 3) only considering SNPs polymorphic in four chromosomes from the larger population or the smaller population (e.g. using four SNP discovery chromosomes from each population), and 4) complete ascertainment in both populations. These ascertainment strategies are meant to mimic the actual ascertainment process where the genotyped SNPs are likely to be at high frequency due to discovery in a small number of chromosomes. Equally important, we considered differences in SNP discovery between populations, as SNP discovery was not uniform across all the populations considered in our study (e.g. little or no SNP discovery has been conducted in the South Asian population).

We simulated a single set of 5,000 independent regions and then implemented the four ascertainment strategies described above. Any differences in the distribution of the number of haplotypes among ascertainment strategies are therefore not due to the evolutionary variance among different coalescent simulation replicates, as the same simulation replicates were used for all ascertainment strategies. For each region, we selected a random subset of 25 SNPs with MAF $> 10\%$ in both populations. As in our analysis of the real data, the same set of SNPs was used in both populations. Importantly, haplotypes under each ascertainment strategy all consist of 25 SNPs. Thus any differences in the number of haplotypes among different ascertainment strategies are not due to the fact that we are missing many SNPs when a small SNP discovery sample was used.

Supplementary Figure S7 shows the distribution of the number of haplotypes for the small (N_c = 5,000; dotted lines) and in the large (N_c = 10,000; solid lines) populations for the four different ascertainment strategies. For all four ascertainment strategies, we see that the distribution of the number of haplotypes is higher for the larger population, indicating that haplotype diversity is related to population size, even when there is no SNP discovery from the larger population. While the overall means of the distributions appear quite similar regardless of ascertainment strategy, the distributions do differ for different ascertainment strategies. For example, using only two chromosomes from the smaller population for SNP discovery (ascertainment strategy 2, red lines in Supplementary Figure S7) results in more regions with a smaller number of haplotypes for both populations. Increasing the number of SNP discovery chromosomes from 2 to 4 greatly reduces this problem (compare the blue lines to the red lines). These simulations, in agreement with previous empirical evidence [Conrad et al. 2006], suggest that qualitative patterns of haplotype diversity such as the number of haplotypes averaged over many windows of the genome are largely robust to ascertainment bias. We caution that other haplotype or LD statistics may be more sensitive to ascertainment bias and additional investigation of their properties may be warranted.

Identification of Runs of Homozygosity

To assess the robustness of the method to issues regarding SNP ascertainment, we conducted a simulation study using a similar scheme to that adopted for the haplotype diversity simulation study. Using the program *GENOME* [Liang et al. 2007], we simulated chromosomes of 5cM in two populations that separated 1,000 generations ago. As before, the ancestral population had an effective population size of 10,000, and the two sampled populations had effective population sizes of 10,000 and 5,000. We set the recombination rate to be equal to 1cM/Mb and the mutation rate to 1 x 10^{-8} per bp. We randomly combined pairs of simulated chromosomes to create simulated individuals. By chance, some of these individuals will have regions of autozygosity, and we tested the robustness of the method to detect these regions under a variety of SNP ascertainment schemes.

Using the unthinned simulated data (with approximately 7,000 to 8,000 SNPs per simulation), we estimated the cumulative LROH in each individual (cROH) without ascertainment of any kind. We selected 253 and 115 individuals from the small and large populations respectively with more than 1cM cROH. We then created 4 simulated data sets using different SNP discovery schemes: 1) SNPs discovered in a panel of 4 chromosomes from the large population, 2) SNPs discovered in a panel of 4 chromosomes from the small population, 3) SNPs discovered in 2 chromosomes from each population, 4) SNPs discovered in 4 chromosomes from each population. Once all the SNPs had been ascertained, we further thinned to 1,000 SNPs that approximately match the mean genetic distance between SNPs and frequency spectra observed in our study using the Affymetrix 500K chip. This was achieved by first constructing a site frequency spectrum of both our observed data and the simulated data. We then repeatedly removed SNPs from over-represented frequency classes until only 1,000 SNPs remained in the simulated data set.

We re-estimated cROH using the ascertained data. Robustness of the method to SNP ascertainment was measured by calculating the correlation coefficient between the cROH estimate using the unthinned data and the estimate under the ascertainment scheme for all individuals with more than 1cM of cROH . For comparison, we estimated a similar correlation coefficient using the inbreeding coefficient of these individuals, F, as estimated by *PLINK*.

We find that the HMM is largely robust to ascertainment scheme. In absolute terms, the estimated cROH under each ascertainment scheme was within 2% of the value estimated using the unthinned data. The correlation between the ascertained estimate and the unthinned estimate is very high (Supplementary Table S7), especially in comparison to F. Depending on the ascertainment scheme, the cROH method has correlation coefficients in the range of 0.966-0.978 for the small population, and 0.985-0.994 for the large population. In comparison, the F method has correlation coefficients in the range of 0.815-0936 for the small population, and 0.911-0.954 for the large population.

A potential confounding factor in the detection of LROHs using SNP genotype data is that SNPs occurring within copy number variable regions may appear to be homozygotic. For example, a hemizygous deletion of a region containing a SNP would potentially cause the SNP to be called as a homozygote. For this reason, we have attempted to remove SNPs within hemizygous regions by analyzing samples for copy number variable regions. To locate regions of hemizygous deletion

we used the *CNAT 4.0* copy number tool command line version (Affymetrix). Individual CEL files were normalized using the quantile normalization method. One hundred random females were used to generate the pooled reference sample and *CNAT 4.0* was run with the Gaussian smoothing option on and band width set to 100kb. All other options were set to default. Hemizygous deletion regions were then called for each individual as regions showing 3 or more SNPs in the hemizygous state with p-value less than 10^{-3}. Further, regions called hemizygous which contained large gaps in SNP coverage or low SNP density were removed before comparison to the regions of autozygosity.

To assess if we would expect to observe Highly Homozygous Regions (HHRs) under a standard coalescent model, we simulated 100 datasets using the program *ms* [Hudson 2002], each consisting of 20Mb regions in 250 individuals. Simulations were conducted with an effective population size of 10,000 with a 90% bottleneck between 1,600 and 2,400 generations in the past. We used a population mutation rate (θ) of 400 / Mb, and a recombination rate of 1cM/Mb (ρ = 400 / Mb). Using this simulated data, we applied the HMM method and called LROH over 1cM in length and containing at least 50 SNPs. We then looked for regions where the LROH of overlap in 5% or more individuals. We found 7 of the 100 simulations contained regions of LROH in more than 5% of individuals. However, these regions we all below 0.85Mb in length and no region was homozygous in more than 6% of individuals. Repeating the study unsing an increasing bottleneck of 95% gave 61 simulations with homozyogous regions in more than 5% of individuals, with 5 simulations achieving homozygous regions in more than 10% of individuals. We therefore suggest that many of the observed HHRs occurring at high frequency are the result of strong foundational bottlenecks.

We considered the possibility that HHRs contain large inversions. Recombination is expected to be repressed within inversions [Stefansson et al. 2005], and hence would not break down the linkage between SNPs on the inversion haplotype. Assuming a given inversion reaches intermediate frequency in a population, then a fraction of individuals are likely to be homozygous at the inversion locus. However, comparisons with published lists of inversions [Tuzun et al. 2005, Bansal et al. 2007] do not suggest that any of our top HHRs contain previously identified inversions.

References

Bansal, V., Bashir, A., and Bafna, V. 2007. Evidence for large inversion polymorphisms in the human genome from hapmap data. *Genome Res.* **17**: 219–230.

Browning, S.R. and Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084–97.

Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**: 1251–60.

Emeneau, M. and Burrow, T. 1962. *Dravidian borrowings from Indo-Aryan.* University of California Press, Berkeley; Los Angeles.

Hartl, D. and Clark, A. 2007. *Principles of population genetics, Fourth Edition.* Sinauer Associates Sunderland, Mass.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–8.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–4.

Liang, L., Zöllner, S., and Abecasis, G.R.R. 2007. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics* **23**: 1565–1567.

Nelson, M., Bryc, K., King, K., Indap, A., Boyko, A., Novembre, J., Briley, L., Maruyama, Y., Waterworth, D., Waeber, G., et al. 2008. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *The American Journal of Human Genetics* **83**: 347–358.

Novembre, J., Johnson, T., Bryc, K., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., Nelson, M., Stephens, M., et al. 2008. Genes mirror geography within Europe. *Nature* .

Novembre, J. and Stephens, M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40**: 646–9.

Patterson, N., Price, A.L., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Reddy, G. 2007.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. 2005. A common inversion under selection in europeans. *Nat Genet* **37**: 129–137.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, A.V., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732.

Weir, B.S. and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population-structure. *Evolution* **38**: 1358–1370.

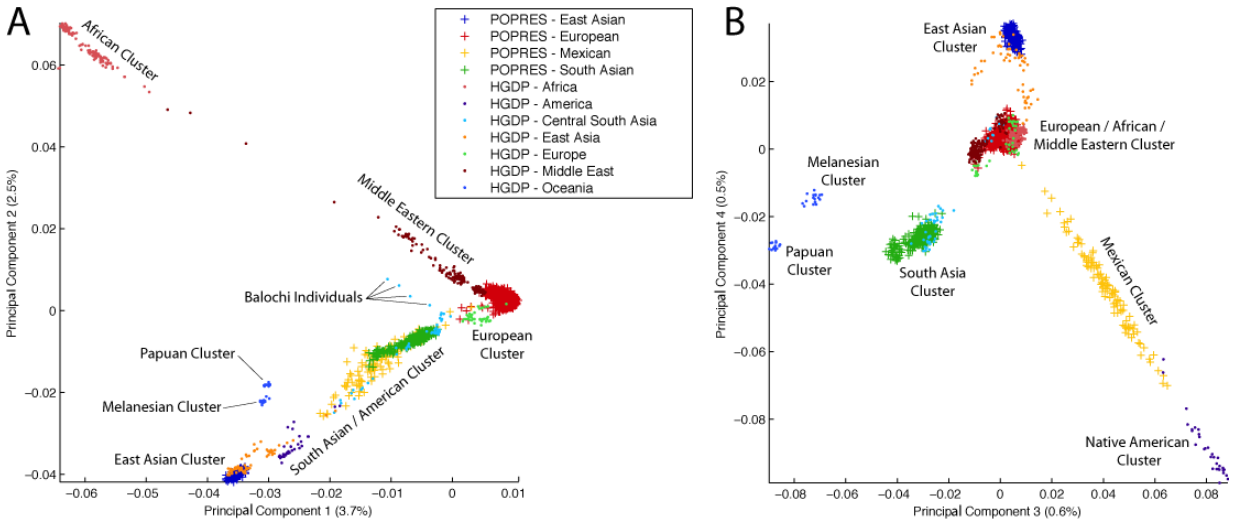

Figure S1: (A) First two principal components of global PCA analysis using 73,000 common SNPs from the POPRES and HGDP datasets. (B) Third and fourth principal components. The percentage of variance explained by each principal component is shown in brackets.

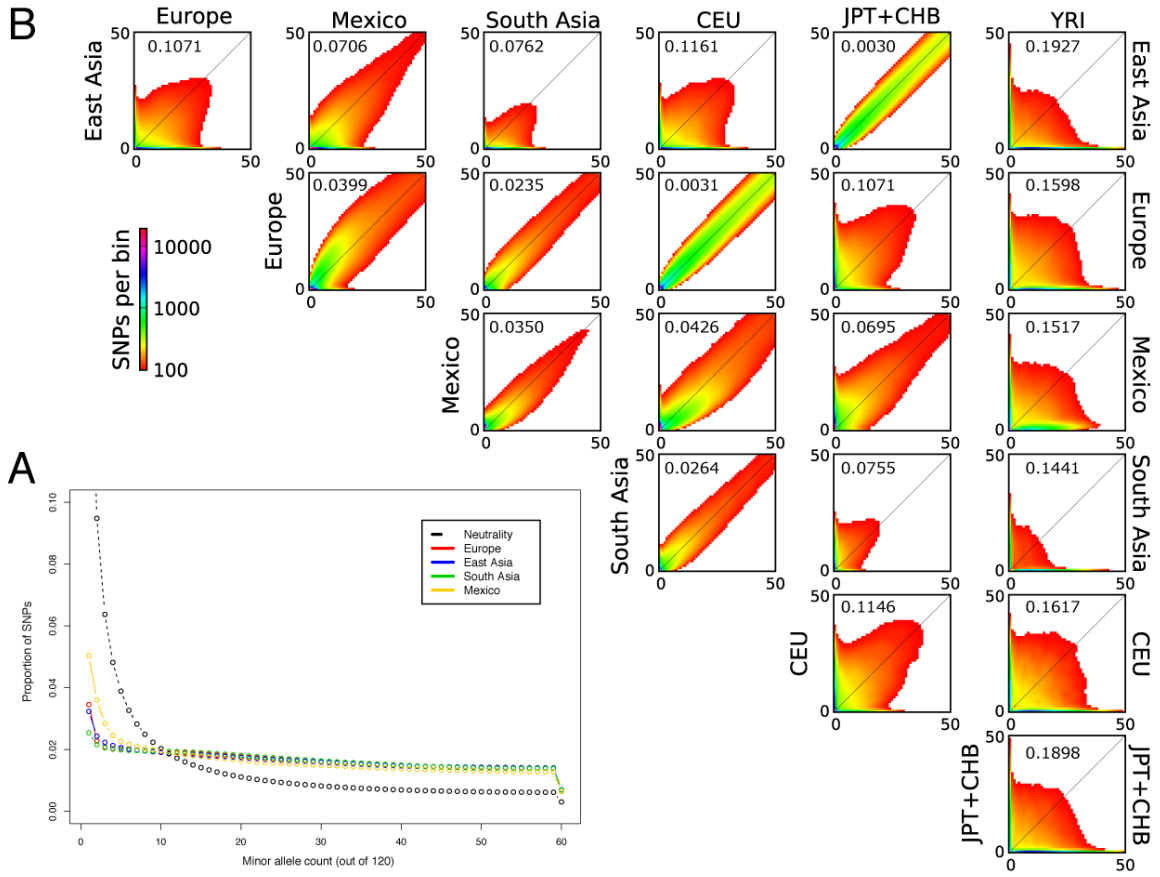

Figure S2: Frequency spectra of the POPRES populations. (A) Minor Allele Frequency Spectra for the four sub-continental populations. The spectrum expected under neutrality is also shown in black. To account for differences in sample size, each sample was projected down to 120 chromosomes using the hypergeometric distribution. (B) Two-dimensional joint frequency spectra for each pairwise sub-continental population comparison. In this case, each sample was projected down to 100 chromosomes using a hypergeometric distribution. For each plot, the minor allele is defined from the total frequency in the two populations. Colors represent the number of SNPs within each bin. Entries in the spectra containing less than 100 SNPs are shown in white. Autosomal estimates of F_{ST} for each comparison are shown in the upper left hand corner of each figure.

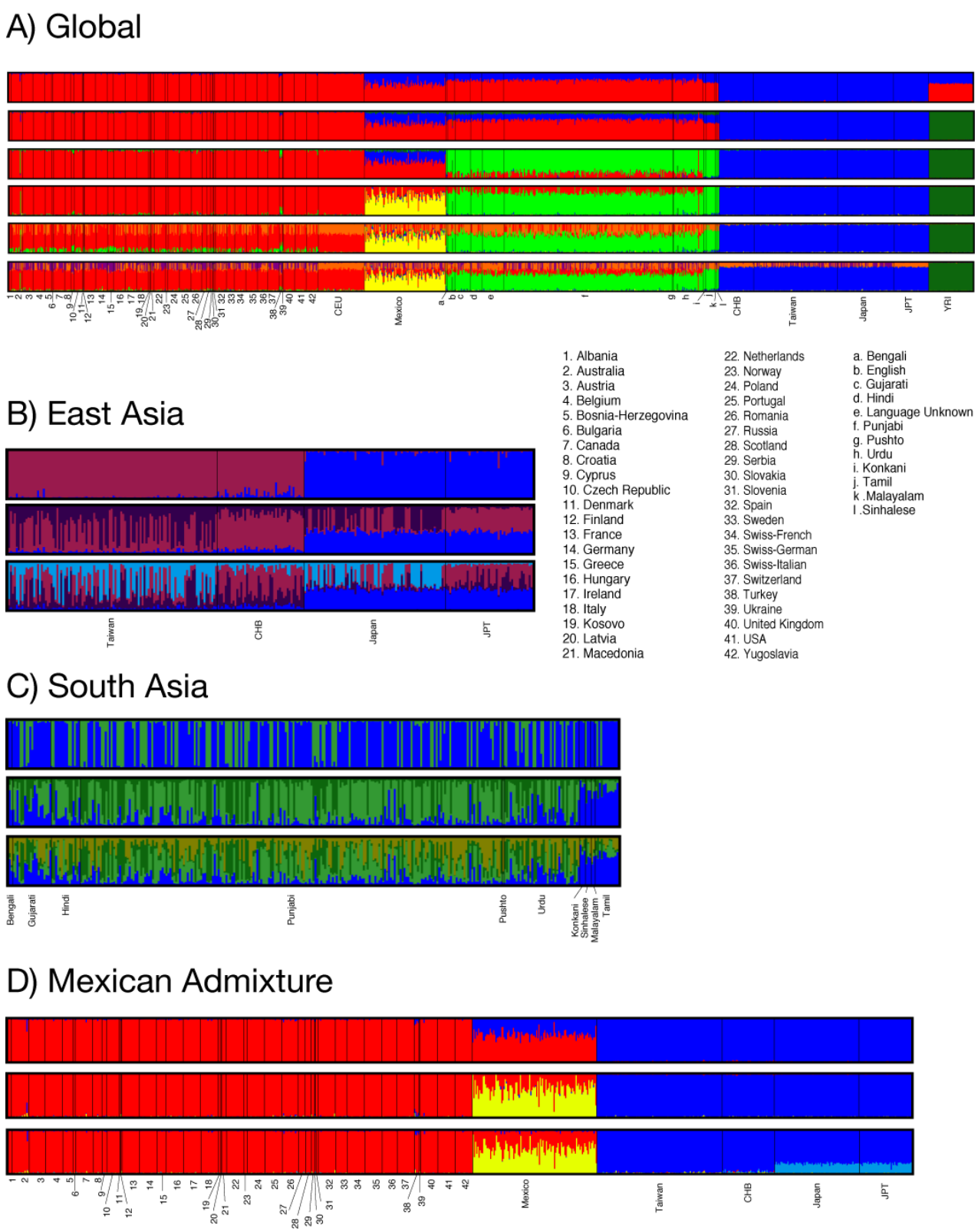

Figure S3: (A) Global *STRUCTURE* results for K=2 to K=6. Subsequent plots show regional analyses for K=2 to K=4. (B) East Asia. (C) South Asia. (D) Mexican Admixture.

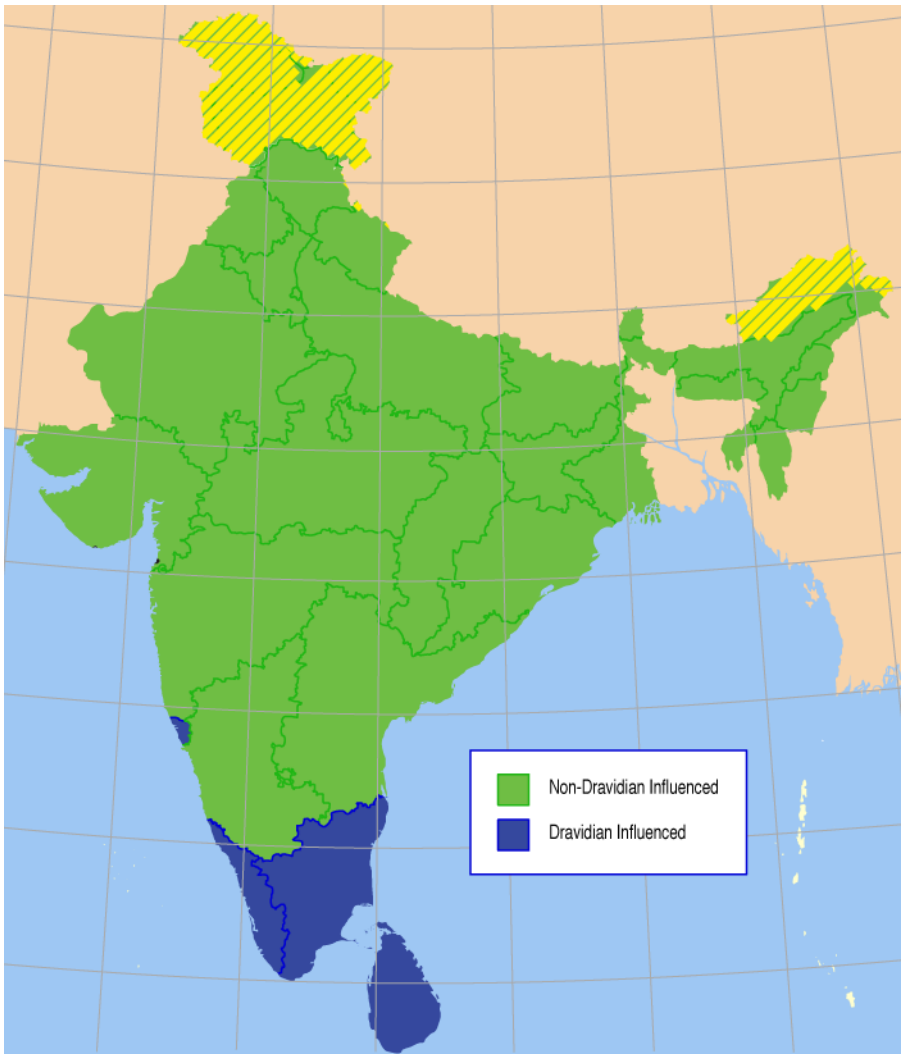

Figure S4: Map of India and Sri Lanka showing the regions in which the Dravidian Influenced languages (blue) and the Non-Dravidian Influenced languages (green) are spoken, based on the official languages of each region. Map adapted from Reddy 2007.

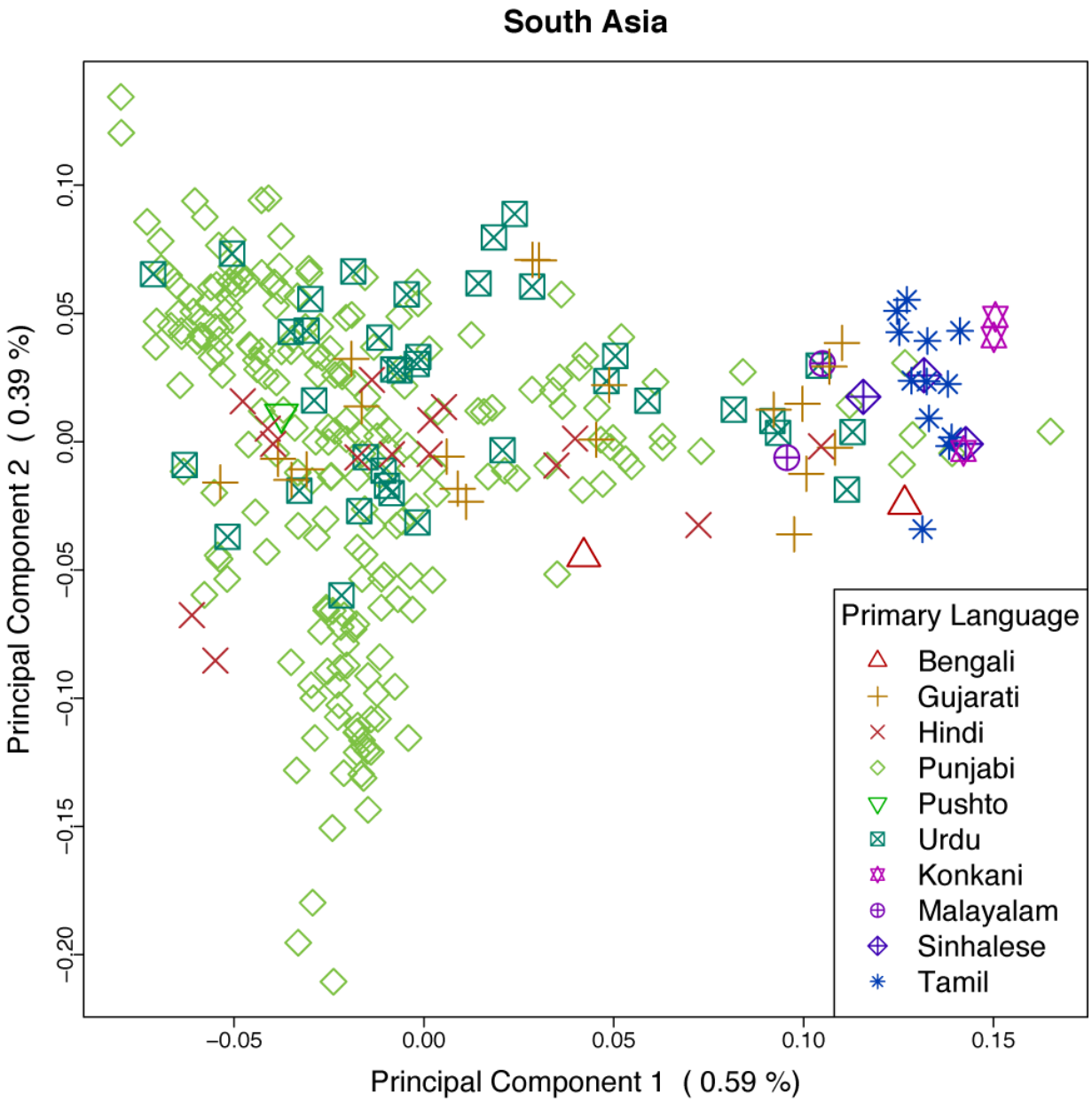

Figure S5: Principal Component Analysis of South Asia. The figure shown here is the same as Figure 1D in the main text, except that individuals are colored by spoken language.

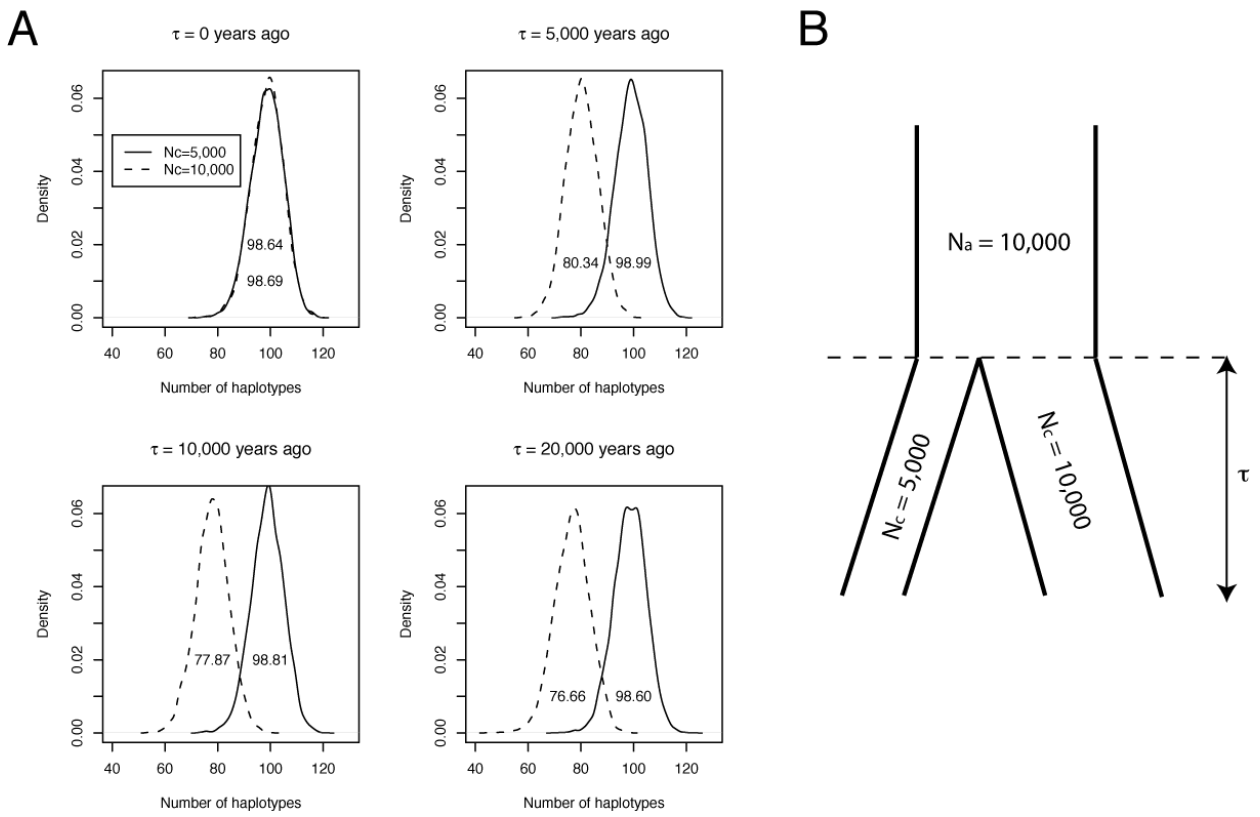

Figure S6: (A) Distribution of the number of haplotypes for different population split times. The number inside each of the density plots is the mean of the distribution of the number of haplotypes. (B) Illustration of the population demography used in the simulations.

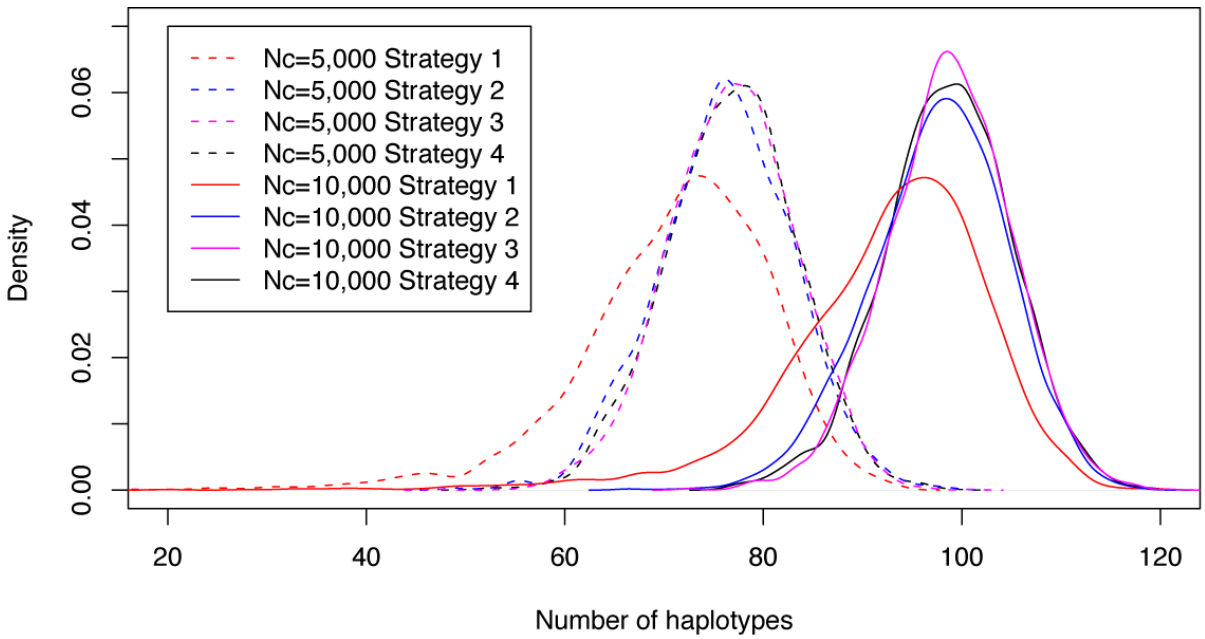

Figure S7: The effect of SNP ascertainment on the distribution of the number of haplotypes. We considered four different ascertainment strategies: 1) SNPs polymorphic in two discovery chromosomes from the smaller population (red lines), 2) SNPs polymorphic in four discovery chromosomes from the smaller population (blue lines), 3) SNPs polymorphic in four chromosomes from the larger population or the smaller population (e.g. using four SNP discovery chromosomes from each population; pink lines), and 4) complete ascertainment in both populations (black lines). Dotted lines represent the distribution of the number of haplotypes for the smaller population ($N_c = 5{,}000$) and solid lines the distribution of the number of haplotypes for the larger population ($N_c = 10{,}000$).

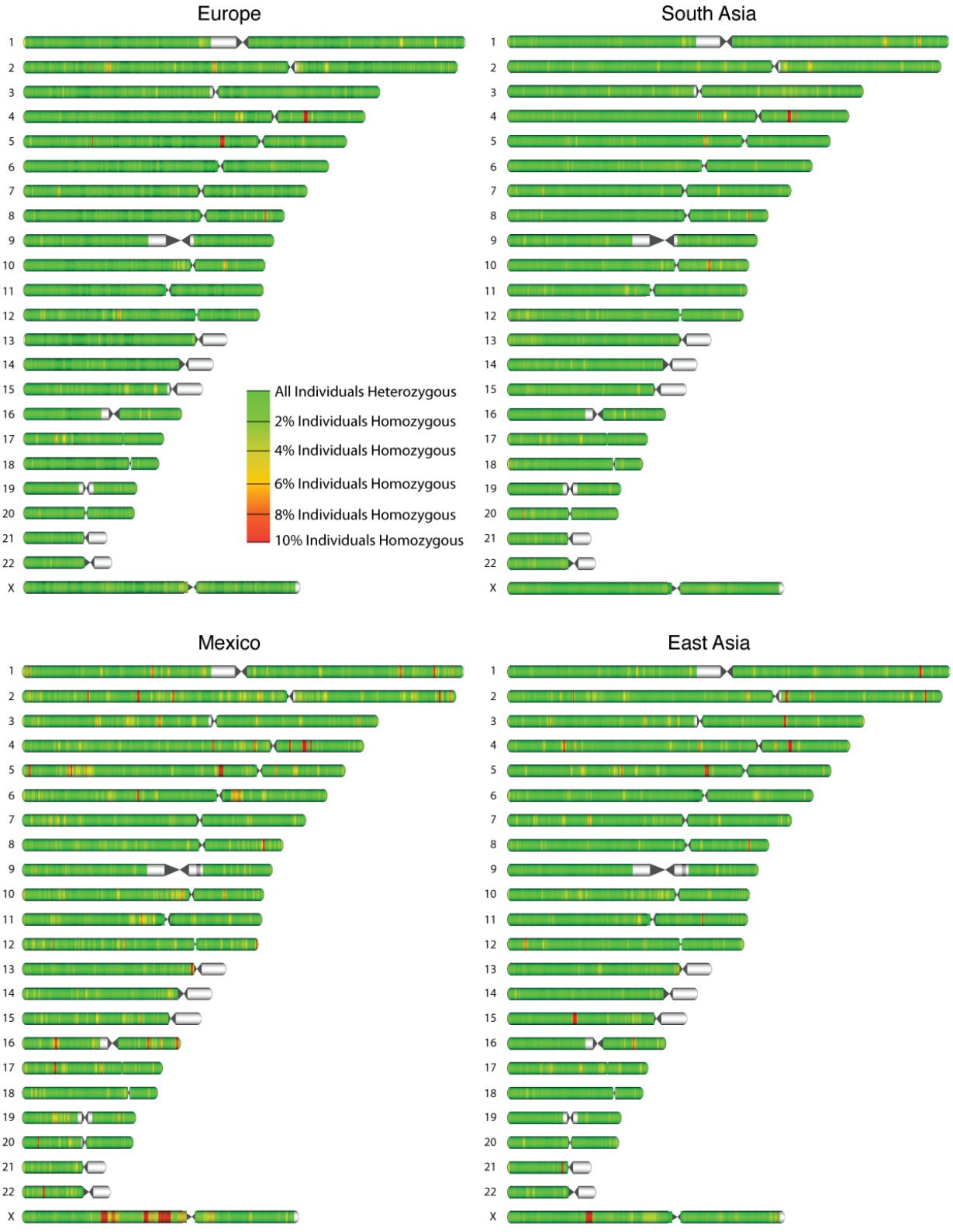

Figure S8: Genome ideograms showing the percentage of individuals with LROH in the four continental populations. Colors indicate the percentage of individuals with LROHs in each region of the genome. The most extreme regions shown in red are indicative of HHRs (LROH in more than 10% of individuals).

Population Group	Continental Group	Individuals	'Strict' Individuals	Included Countries / Language Groups
Dravidian Influenced	South Asia	20	20	Konkani, Malayalam, Sinhalese, Tamil
Non-Dravidian Influenced	South Asia	312	284	Bengali, Gujarati, Hindi, Punjabi, Pushto, Urdu
Europe (C)	Europe	190	186	Austria, Germany, Netherlands, Switzerland (German)
Europe (ESE)	Europe	10	8	Cyprus, Turkey
Europe (NNE)	Europe	78	76	Czech Republic, Denmark, Finland, Hungary, Latvia, Norway, Poland, Russia, Slovakia, Sweden, Ukraine
Europe (NW)	Europe	459	447	Ireland, UK
Europe (S)	Europe	238	232	Italy, Switzerland (Italian)
Europe (SE)	Europe	99	96	Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Kosovo, Macedonia, Romania, Serbia, Slovenia
Europe (SW)	Europe	272	264	Portugal, Spain
Europe (W)	Europe	1,063	1,042	Belgium, France, Switzerland (French)
Mexico	Central America	112	107	Mexico
Japan	East Asia	73	73	Japan
Taiwan	East Asia	108	108	Taiwan
Europe Other A	Europe	237	0	Apparent European ancestry, but self-identified from the USA, Canada or Australia
Europe Other B	Europe	18	0	Apparent European ancestry, but self-identified from elsewhere
Europe (Mixed)	-	524	0	European individuals of mixed ancestry
South Asian Other	South Asia	28	0	South Asian individuals without language information
Unknown	-	4	0	No geographic or linguistic information

Table S1: Details of population groupings. Full details of each individual are available as a supplementary table.

Population	H_{10}	**95% Confidence Interval**	H_{25}	**95% Confidence Interval**
Non-Dravidian Influenced	**33.5341**	33.365, 33.703	**22.4679**	22.246, 22.69
Dravid Influenced	33.4043	33.233, 33.576	22.3368	22.117, 22.556
Europe (NW)	31.213	31.026, 31.4	20.0524	19.836, 20.268
Europe (C)	31.5891	31.406, 31.772	21.0207	20.806, 21.235
Europe (NNE)	31.5986	31.419, 31.778	21.0613	20.848, 21.274
Europe (W)	31.6785	31.494, 31.863	21.263	21.056, 21.47
Europe (SE)	32.0286	31.849, 32.208	21.3254	21.11, 21.54
Europe (SW)	32.2328	32.056, 32.41	21.5581	21.34, 21.776
Europe (S)	**32.5165**	32.337, 32.696	**21.5941**	21.375, 21.813
Mexico	**31.3765**	31.202, 31.551	**20.9565**	20.743, 21.17
Japan	30.6862	30.489, 30.884	19.6953	19.479, 19.912
Taiwan	**31.3138**	31.118, 31.51	**20.7644**	20.553, 20.976

Table S2: Estimates of Haplotype Diversity using a thinned sample of 40 chromosomes per population. High values within each continent are shown in bold. Confidence intervals for the haplotype counts are calculated assuming a normal distribution.

	East Asia	Europe	Mexico	South Asia	CEU	JPT+CHB	YRI
East Asia	-	0.1071	0.0706	0.0762	0.1161	0.0030	0.1927
Europe	0.1595	-	0.0399	0.0235	0.0031	0.1071	0.1598
Mexico	0.0965	0.0826	-	0.0350	0.0426	0.0695	0.1517
South Asia	0.1027	0.0426	0.0592	-	0.0264	0.0755	0.1441
CEU	0.1717	0.0056	0.0849	0.0456	-	0.1146	0.1617
JPT+CHB	0.0047	0.1560	0.0912	0.0987	0.1655	-	0.1898
YRI	0.3063	0.2640	0.2406	0.2300	0.2529	0.2928	-

Table S3: F_{ST} estimates between pairs of populations. Autosomal estimates are shown in the upper matrix triangle, whereas X chromosome estimates are shown in the lower triangle. For comparison, F_{ST} estimates using the HapMap populations and the same SNP set are also shown.

Population	Non-Dravidian Influenced	Dravidian Influenced
CEU	0.0256	0.0496
JPT-CHB	0.0764	0.0806
YRI	0.1444	0.1474
East Asia	0.0772	0.0808
Europe	0.0227	0.0457
Mexico	0.0348	0.0492
Non-Dravidian Influenced	-	0.0121
Dravidian Influenced	0.0121	-

Table S4: F_{ST} estimates between the Dravidian Influenced and Non-Dravidian Influenced populations and the other continental populations.

Population	**Percentage of YRI haplotypes shared**	**Lower 95% C.I.**	**Upper 95% C.I.**
Europe (SW)	5.52%	5.25%	5.79%
Europe (S)	5.22%	4.96%	5.48%
Europe (W)	5.19%	4.93%	5.46%
Europe (SE)	5.17%	4.91%	5.43%
Europe (C)	5.15%	4.88%	5.42%
Europe (NW)	5.10%	4.84%	5.37%
Europe (NNE)	5.10%	4.84%	5.37%

Table S5: Percentage of HapMap YRI haplotypes found in the European sample. This table is based on 25 SNP haplotypes in 2,925 windows of 0.5cM. The data was thinned to 114 chromosomes in each populations (to equal the YRI sample size).

Population	**Percentage of Mexican haplotypes shared**	**Lower 95% C.I.**	**Upper 95% C.I.**
Europe (SW)	26.43%	26.05%	26.81%
Europe (S)	26.31%	25.92%	26.69%
Europe (W)	26.29%	25.92%	26.66%
Europe (NW)	26.14%	25.78%	26.51%
Europe (C)	26.12%	25.75%	26.49%
Europe (NNE)	25.93%	25.55%	26.30%
Europe (SE)	25.85%	25.48%	26.23%

Table S6: Percentage of Mexican haplotypes shared with European populations. This table is based on 25 SNP haplotypes in 2,925 windows of 0.5cM. The data was thinned to 152 chromosomes in each populations (to equal that of the smallest European sample in the table, Europe NNE).

	Small Population (N=253)		Large Population (N=115)	
Ascertainment Scheme	cROH	F	cROH	F
Complete Ascertainment	1.0	1.0	1.0	1.0
4 chr from large population	**0.966**	0.815	**0.994**	0.911
4 chr from small population	**0.974**	0.845	**0.987**	0.921
4 chr from each population	**0.974**	0.921	**0.993**	0.952
2 chr from each population	**0.978**	0.872	**0.994**	0.928

Table S7: Robustness of HMM method to SNP ascertainment. The table shows the correlation between cROH estimated with full SNP discovery compared to the cROH estimated under 4 other ascertainment schemes. For comparison, a similar study was performed using F. Only simulated individuals with cROH $>$ 1cM (estimated under full ascertainment) were used in the calculations. Both F and cROH were estimated using within-population SNP frequencies.

Chr	Start	End	Region	# SNPs (MAF > 5%)	Mean % of Individuals with LROH	Population
1	15379784	17401898	1p36.13	136	13.5	East Asia
2	3045705	3710421	2p25.3	68	10.3	Mexico
2	8688612	10000083	2p25.1	116	22.1	East Asia
2	8899545	9854486	2p25.1	109	12.8	Mexico
2	43179074	44202272	2p21	94	11.3	East Asia
2	157991956	159379990	2q24.1	138	10.9	Mexico
2	176871680	177857610	2q31.1	94	13.9	East Asia
2	205483512	206252910	2q33.3	130	11.5	Mexico
3	43179088	45124712	3p21.33	161	10.1	East Asia
3	121522065	122818151	3q13.33	124	10.4	East Asia
3	189688953	190302800	3q28	60	11.6	East Asia
3	198067604	198958007	3q29	67	11.4	East Asia
4	29372445	29998717	4p15.1	54	11.5	Mexico
4	32250599	34658227	4p15.1	181	26.0	Europe
4	32250599	34826055	4p15.1	201	12.1	Mexico
4	32528188	34431234	4p15.1	166	12.7	South Asia
4	32555448	34431234	4p15.1	138	19.0	East Asia
4	40844073	41888547	4p13	120	10.2	Mexico
4	41017949	42342777	4p13	110	22.2	East Asia
4	158335214	160167630	4q32.1	153	11.4	East Asia
5	116125903	118646439	5q23.1	187	11.0	East Asia
6	105599748	106475582	6q21	99	10.6	Mexico
8	10509878	12039387	8p23.1	169	17.5	East Asia
8	10509878	12039387	8p23.1	260	11.5	Mexico
10	21519891	23314154	10p12.31	64	10.4	East Asia
13	18441915	19690082	13q12.11	116	10.1	Mexico
15	61189627	64122891	15q22.31	173	16.3	East Asia
16	17231173	17878102	16p12.3	68	12.0	East Asia
16	68106289	71557266	16q22.3	288	10.2	Mexico
17	53118270	54758734	17q22	91	15.8	East Asia
21	15813718	16718760	21q21.1	90	16.4	East Asia
22	34790020	35312621	22q12.3	58	10.3	East Asia
22	37049908	37855737	22q13.1	60	11.9	Mexico
22	44385321	45441994	22q13.31	51	11.4	East Asia
X	47266602	57222190	Xp11.22	222	13.7	Mexico
X	100386133	111121991	Xq22.3	342	12.6	Mexico
X	106862626	111770922	Xq22.3	123	32.1	East Asia
X	146603508	148146339	Xq28	65	17.0	East Asia
X	146603508	148146339	Xq28	94	14.4	Mexico

Table S8: Regions appearing to be LROH in over 10% of individuals within a population.