Supplementary Materials for

The genetics of Mexico recapitulates Native American substructure and affects biomedical traits

Andrés Moreno-Estrada*, Christopher R. Gignoux, Juan Carlos Fernández-López, Fouad Zakharia, Martin Sikora, Alejandra V. Contreras, Victor Acuña-Alonzo, Karla Sandoval, Celeste Eng, Sandra Romero-Hidalgo, Patricia Ortiz-Tello, Victoria Robles, Eimear E. Kenny, Ismael Nuño-Arana, Rodrigo Barquera-Lozano, Gastón MacÍn-Pérez, Julio Granados-Arriola, Scott Huntsman, Joshua M. Galanter, Marc Via, Jean G. Ford, Rocío Chapela, William Rodriguez-Cintron, Jose R. Rodríguez-Santana, Isabelle Romieu, Juan José Sienra-Monge, Blanca del Rio Navarro, Stephanie J. London, Andrés Ruiz-Linares, Rodrigo Garcia-Herrera, Karol Estrada, Alfredo Hidalgo-Miranda, Gerardo Jimenez-Sanchez, Alessandra Carnevale, Xavier Soberón, Samuel Cañizales-Quinteros, Héctor Rangel-Villalobos, Irma Silva-Zolezzi, Esteban Gonzalez Burchard*, Carlos D. Bustamante*

*Corresponding author. E-mail: cdbustam@stanford.edu (C.D.B.); morenoe@stanford.edu (A.M.-E.); esteban.burchard@ucsf.edu (E.G.B.)

This PDF file includes:

Table of Contents

Materials and Methods

Sample collection and genotyping

Institutional review board (IRB) approval for this project was obtained from Stanford University (File: NOT03H02) for obtaining and analyzing de-identified DNA specimens from participating institutions. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions: the University of Guadalajara, the National Institute of Medical Sciences and Nutrition Salvador Zubirán (INNSZ), and the National Institute of Genomic Medicine (INMEGEN). Samples were collected over several years by researchers from these institutions under protocols consistent with biomedical and/or population genetics studies aimed at characterizing the genetic diversity of Mexican populations. Sampling locations and summary data for the populations included in the study are detailed in **Table S1**. A total of 362 samples from 15 indigenous populations were genotyped at the University of California, San Francisco (UCSF) by using Affymetrix (Mountain View, CA) 6.0 arrays. These samples comprise the Native Mexican Diversity Panel (NMDP) of the study. An additional 466 samples were genotyped at the National Institute of Genomic Medicine (INMEGEN) by using a combination of Affymetrix GeneChip 500K and Illumina (San Diego, CA) HumanHap550 arrays. Samples genotyped at INMEGEN include 370 cosmopolitan samples from 10 different Mexican states and 96 samples from three indigenous populations, which were collected as part of the Mexican Genome Diversity Project (MGDP). A subset of the MGDP samples were previously genotyped on the Affymetrix 100K platform (*24*). All participants recruited in cosmopolitan locations were required to have all four grandparents born in the same state. Overall, this combined genotyping effort generated SNP array data for 828 newly genotyped samples from 28 different Mexican populations. All samples were genotyped from genomic DNA extracted from blood.

Data curation

Curation of Native Mexican samples: a total of 458 samples were initially genotyped (362 by using Affymetrix 6.0 arrays and 96 by using Affymetrix 500K arrays). The number of markers included in the Affymetrix 6.0 SNP array determined our starting SNP density before intersecting with data from additional arrays. A total of 909,622 SNPs were successfully genotyped. We removed 2,919 SNPs with duplicate marker names, 1,217 SNPs with no physical position in the NCBI Build 36.1 human reference sequence (hg18 assembly), and 8,087 SNPs failing Hardy-Weinberg equilibrium at $1x10^{-5}$. We restricted to autosomal SNPs and samples with more than 90% of genotyping rate. We removed three samples due to evidence of being duplicates of another sample. As part of the recruiting strategy, 40 trios and 6 duos were included to improve phasing accuracy of haplotype-based analyses and ancestral reference panels for admixture deconvolution (see below). One trio showed an excess of Mendelian errors and was excluded from trio phasing. Subsequently, the 46 individuals constituting the offspring of all trios and duos were removed from most of the population genetic analyses. We did not systematically filter for second-degree or lower relatives as part of our initial curation given that some of the subsequent analyses make use of IBD information to describe within- and between-population connections among pairs of individuals across Native Mexican populations (see sections below). We also excluded 8 individuals due to a high proportion (>30%) of non-Native

ancestry, as these are likely to correspond to sampling exceptions rather than being part of the population's admixture pattern. This was confirmed by PCA analysis where these samples appeared to be outliers relative to others from the same population. Since the scope of the study is to assess the population structure, including the characterization of recent admixture events among Native Mexicans, we did not initially filter genomic segments or individuals with some degree of non-Native ancestry. However, more stringent filters were applied as needed for particular analyses as detailed in the subsequent sections below. After data curation, the number of Native Mexican samples genotyped for this study was 401 (**Table S1**). Illumina 550K data were also available for the subset of three MGDP Native American populations genotyped by INMEGEN (*8*) and integrated as described below.

Curation of Cosmopolitan Mexican samples: Of the 370 cosmopolitan samples genotyped at INMEGEN, 313 were genotyped by using both Affymetrix 500K arrays and Illumina 550K arrays (covering seven Mexican states), and 57 samples were genotyped by using Illumina 550K arrays only (covering three additional Mexican states). For the subset of cosmopolitan samples genotyped with both arrays, genotype data for nearly 1 Million SNPs were available for analyses.

Data integration

To combine our dataset with additional preexisting data and assemble continental reference panels of potential ancestral populations relevant to the Mexican admixture process, our data were integrated with previously genotyped datasets from various sources. Additional Mexican data included Affymetrix 500K genotypes for 53 Native individuals from two Mexican indigenous populations (*33*), Affymetrix 6.0 genotypes from 49 Mexican-Americans (MXL) sampled in Los Angeles, California as part of the International HapMap project phase 3, and Affymetrix 500K genotypes for 50 Mexicans of admixed origin sampled in Guadalajara, Jalisco included in the Population Reference Sample (POPRES) data set. European data were obtained from a selected subset of 204 European samples from POPRES to be included as part of the reference panel of ancestral populations. Inclusion criteria were based on maximizing geographic representation of regions within Europe and equalizing sample sizes to those available for the Native Mexican populations (i.e., approximately 20, see **Table S1**). The collections and methods for the POPRES Sample are described by Nelson et al. (*17*). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v1.p1 through dbGaP accession number phs000145.v1.p1. Additional European populations from Spain (n=55) included Basque, Andalusian, and Galician (*22*), and additional HapMap samples included 25 Tuscans (TSI) and 25 Utah residents of Northern European descent (CEU). Finally, 50 Yorubas from Ibadan, Nigeria (YRI) from HapMap were included as reference panel for West African ancestry. A total of 511 additional samples were integrated from previously generated datasets. The dataset analyzed here is the result of merging autosomal SNP array data from these different sources and consists of up to 1,282 samples, including 454 Native Mexicans from 20 indigenous populations, 469 cosmopolitan Mexican samples from 12 locations, and 359 ancestral European and West African populations.

Three main working datasets with variable SNP densities were constructed after merging multiple datasets and reapplying data quality control filters (now raised to 95% of call rate for SNPs and samples and excluding SNPs with uncertain strandedness). Namely, one dataset was

constructed considering the intersection of Affymetrix and Illumina data (71,581 SNPs), one consisting of Affymetrix data only (372,692 SNPs), and one combining the union of both Affymetrix and Illumina arrays (785,663 SNPs).

Table S2 describes the details of these three datasets. Most of the analyses presented here are based on the Affymetrix dataset (including data from 500K and 6.0 arrays), as this combination offered the best balance between SNP density and number of populations included (both indigenous and cosmopolitan). Nonetheless, we also used the combined dataset of Affymetrix and Illumina arrays in those analyses that were more robust to lower marker densities and where maximizing the number of populations was essential. Likewise, we used the union of these platforms in those analyses requiring the densest dataset across a limited number of populations.

The steps described above correspond to our initial data curation and the resulting datasets (listed in **Table S2**) constituted the base of all population structure analyses. Additional filters were applied to exclude certain samples or integrate additional data for particular analyses as described below.

Population structure of Native Mexicans

We used the Affymetrix dataset (372,692 SNPs) in all the analyses focused on Native Mexican populations. We restricted these analyses to individuals having >90% of Native American ancestry (average proportion of Native American ancestry among remaining individuals was 97.26%). Therefore, in addition to the initial data curation steps described above, 78 samples were removed from the Affymetrix dataset as detailed in **Table S3.** A total of 376 samples remained and all 20 populations were represented, with an average sample size of 19 individuals.

Principal component analysis (PCA) and population differentiation: We used EIGENSOFT(*34*) to perform PCA, and R was used to generate plots. Pairwise F_{ST} values for each population comparison (**Table S4**) were calculated using the estimator of Weir & Cockerham (*35, 36*), and ggplot2 (*37*) was used to create the plots.

Runs of Homozygosity (ROH) estimation: To infer estimates of autozygosity and relative recent population size, we estimated runs of homozygosity using a sliding window approach as implemented in PLINK (*38*) with a similar set of criteria to Nalls et al. (*39*). In keeping with standard practice in the field for estimating ROHs in humans, we assigned ROH given a minimum window size of 1 Mb, allowing for no more than two missing sites and one heterozygous SNP per window. To find the fraction of the genome in ROH, we divided the total ROH per person by the length of the autosome (approximately 2.8 Gb). We ran all samples through the same analysis for plotting and interpretation per population. In Fig. S2 we present the mean and confidence interval (generated with 1000 bootstraps, sampling individuals with replacement) for each Native Mexican population. In Fig. S3 we present histograms of ROH tracts for each population on a log scale. As all populations have some amount of average ROH, varying window size slightly did not affect the overall pattern of results.

Rejection algorithm and demographic estimation: To infer population genetic parameter estimates for each Native Mexican population, including bottleneck strength and current N_e values, we implemented a demographic inference estimation method using approximate Bayesian computation via rejection algorithm as built into REJECTOR2 (*40*). We focused on a tract length statistic sensitive to bottlenecks known as cumulative Runs Of Homozygosity

(ROH). We assigned ROH based on sliding windows with a minimum of 50 SNPs in a tract, and allowing for no more than one heterozygous SNP per 500-Kb window. This set of criteria is standard practice in the field for estimating ROHs in humans (see (*39, 41, 42*)) with the additional step that we imputed genotypes from BEAGLE to fill in missing data. Given that ROH tracts are length-based and not dependent on the site frequency spectrum they are unlikely to be highly affected by ascertainment bias. Indeed previous simulations indicate accurate recovery of demographic parameters is possible using ROH statistics calculated from array genotypes (*42*).

We generated a set of simulations similar to Henn et al. (*42*): moving forward in time, we begin with a fixed large population size, then the population experiences a bottleneck and subsequent recovery to modern day deme size, with demographic parameters (Current effective size and bottleneck strength) drawn from uniform priors. We used the computationally efficient approximate coalescent simulator MaCS (*43*) for simulation, and a tolerance (alpha level) of 20% between the observed and simulated sequences for either ROH or the variance in ROH to accept or reject simulations. To make simulations tractable, we only investigated ROH on chromosome 1, and to use the maximum density of genotyped SNPs, we restricted to Native Mexican populations for which Affymetrix 6.0 array data was available (see **Table S1**). For each population we generated 100,000 simulated data sets. For estimating final parameters, we employed density-based smoothing in R (with default parameters) over each histogram of accepted runs to estimate modes and 95% confidence intervals of each parameter of interest based on the profile approximate likelihoods. We then created plots with both the real histograms and the smoothed density values, plotting the informative portion of the accepted runs, both in summary form (Fig. S4) and the individual profiles (Fig. S5).

Demographic simulations: Investigations of effective population size rely on a measure of genetic diversity. Most are informed by the full site frequency spectrum of sequence data, which typically does not reflect the site frequency spectrum of genome-wide array SNPs With genome-wide genotype data, we can still use tract-length statistics, where the tracts are long enough (hundreds of Kb to Mb) that high-density genotyping can capture the signal accurately. Summary values of ROH in the true data were calculated using the Rejstats module in Rejector2, using the same settings as the real genotype data. First, estimates of ROH and IBD both provide a signal of low effective population size. Fig. S6 (panel A) is a coalescent simulation-based distribution of the relationship between Deme Size and the proportion of the genome in ROH, with modern population size varying between 2 and 15,000, and bottleneck strength from 0.01 to 20-fold. This relationship between the two variables is quite strong (Spearman's ρ -0.94).

We first simulated large chromosomes in MaCS as six 40-Mb segments (similar in total size to chromosome 1), with no bottleneck, with an N_e between 2 and 1800. A total of 90,000 simulations were tested for acceptance and approximated as a distribution using the same criteria as with the Native American data (20% tolerance, histogram fit to a continuous distribution using the default density function in R). One hundred simulations were generated separately using the same parameters and used for testing. Here, the 95% credible intervals for all 100 tests cover the true value, and Spearman's ρ between the true values and the maximum density values is 0.96 (Fig. S6 panel B). Incorporating a varying bottleneck from 1-20-fold reduction, uniformly distributed, does weaken recovery however 94/100 tests have 95% credible intervals covering the real data, and Spearman's ρ between the true values and the maximum density values is 0.75 (Fig. S6 panel C).

As seen in simulations, given that we can vary bottleneck strength extensively and still recover ROH values we see relative insensitivity between bottleneck strength and ROH. We include it as a varying parameter to allow for uncertainty, and it is significantly associated (Spearman's ρ of 0.03, $p<10^{-16}$ across 66,502 simulations); however, the correlation is too low to have much predictive power. In keeping with this, we can see limited evidence for precision in bottleneck estimates in the real data for most samples (Fig. S5 bottom panel of each population). Modeling the bottleneck could benefit from the increased precision available from leveraging the inferred N_e first, however, as this analysis requires another condition, this could require far more simulations. As shown previously (*42*), we saw limited precision in estimating the bottleneck time so we did not investigate it further.

Identity-by-descent (IBD) analysis: Genotype data were phased using BEAGLE (*44, 45*) with available duos and trios used as training sets. We estimate the amount of DNA shared identically by descent (IBD) using the GERMLINE software (*46*), with a 5-cM threshold to minimize false-positive IBD matches. All 5 cM or greater segments shared IBD between pairs of individuals were summed, and binned into nine categories as detailed below. We then used the graph visualization software ShareViz [http://www.cs.columbia.edu/~itsik/sharevizWeb/shareviz.html] to visualize within- and between-population relationships of pairs of individuals at different bin thresholds (Fig. S8). This particular display shows all individuals in a given population regardless of the degree of IBD sharing. In contrast, to construct a network of relatedness that is informative of the degree of sharing between pairs of individuals within a population, we used Cytoscape 3.1.0 to apply a force-directed algorithm in which nodes (samples) repel each other and edges (connections) attract them proportionally to the total amount of shared IBD. We focused on individuals sharing >13 cM of the genome IBD to ease interpretation of the major between-population interactions (Fig. 1C), denoting the number of individuals with shared IBD for each population in the label. It should be noted that while the display of within-population connections is the result of the force-directed layout, the length of between-population connections is not informative of the amount of IBD as clusters have been localized to their approximate sampling location to provide geographical context.

Estimated degrees of relatedness and IBD binning

% shared	cM	Relation°	IBD range (cM)	binning
100	3000	Self	-	-
50	1500	1°	> 1300	9
25	750	2°	650 - 850	8
12.5	375	3°	325 - 425	7
6.75	188	4°	163 - 213	6
3.37	94	5°	80 - 110	5
1.69	47	6°	40 - 53	4
0.85	23.5	7°	20 - 27	3
0.42	11.75	8°	10 - 13	2
0.21	5.875	9°	5 - 7	1

Population Tree analysis: Trees have been widely used in population genetics to visualize the relationships among populations. While providing a valuable initial assessment of population

relationships, a bifurcation tree might be a simplistic representation of human population history as it assumes population splits with no further gene flow between them. To overcome this problem, new methods have been recently developed allowing for the inclusion of gene flow between edges and representing population relationships by means of a reticulated graph rather than a strict bifurcation tree. Here, we used TreeMix v1.0 (*14*) to infer patterns of population splitting and mixing from genome-wide allele frequency data. This approach estimates the maximum likelihood tree for a given set of populations given a Gaussian approximation to allele frequencies, and then attempts to infer a number of admixture events. Before adding migration, we ran TreeMix with our set of 20 Native Mexican populations and HapMap continental populations (YRI, CEU, and ASN) as outliers to help us set the root of the tree in subsequent runs (Fig. S9). Although not representing a perfect fit to the data, we used the maximum likelihood tree without migration to evaluate the general topology and the extent of population drift in terms of allele frequency shift from an ancestral population. We then used the residuals matrix to identify pairs of populations showing poor fits in the initial tree. These are then considered as candidates around which we add migration edges and try new rearrangements of the tree, now accounting for *n* number of migration events. As a test run, we first used our previous panel (Native Mexicans plus CEU and YRI), adding MXL from HapMap as a population with known recent admixture. The resulting graphs allowing migration events showed the strongest signal of gene flow arising from CEU (i.e., European) into MXL (i.e., Mexican samples in HapMap), consistent with known historical records of these populations. Given that recent admixture can bias the signals detected by TreeMix, we restricted further runs with migrations to individuals with ≥98% of Native American ancestry in order to infer historical admixture events among Native Mexican populations. This filter removed 130 samples in addition to the ones removed by the 90% filter (**Table S3**).

Population structure of cosmopolitan samples

We used the combined Affymetrix + Illumina dataset (71,581 SNPs) to perform cluster-based analysis and PCA on the full set of samples listed in **Table S1**. This allowed us to include the maximum number of cosmopolitan samples and evaluate the impact of Native American substructure in the composition of admixed Mexican genomes.

Structure analysis: We used the block relaxation algorithm implemented in ADMIXTURE (*18*) to estimate individual ancestry proportions given *K* ancestral populations. We initially ran the algorithm from *k*=2-20 using the global dataset with the maximum number of available individuals to explore general clustering patterns. We then filtered first- and second-degree relatives and selected subsets of HapMap and POPRES individuals to roughly equalize sample sizes to those available for Native Mexican populations (**Table S1**). We found extensive substructure not only among the ensemble of recently admixed cosmopolitan Mexican samples, but also among the different ancestral populations. This was true not only for Native Mexican populations, but also for Europeans showing varying proportions from different clusters within Europe (Fig. S10). Therefore, rather than using *reference* individuals as supervised training samples (which are assumed to have 100% ancestry from some ancestral population), we ran an unsupervised analysis to let ADMIXTURE estimate ancestry values across all samples. We used the default setting (folds=5) to perform ADMIXTURE's cross-validation procedure for evaluating fit of different values of K. Fig. S11 shows the cross-validation error for each run, where *k*=9 showed the lowest error estimates (0.49798), indicating that sub-continental

clustering levels are a sensible modeling choice for Mexican populations. Additionally, we found constantly increasing Log likelihood values for all runs from k=2 to k=10 (Fig. S11), where k=9 showed the maximum number of population-level clusters among Mexicans. An additional European sub-continental component was detected at k=10 and found to be restricted to the Basque population and shared to a limited extent with other Iberian populations (Fig. S10). At k=11, a group of three MXL samples clustered apart showing full membership to their own component, reflecting possible cryptic relatedness among them. Due to their shared ancestry with other Mexican cosmopolitan samples, residual proportions of this "MXL component" were also assigned to most of the remaining individuals, which is probably not the best description of their actual ancestral components given the observed patterns at smaller k values. This is also reflected in the subtle drop of the Log likelihood curve when compared to all other runs. This component remained stable across higher values of k, while other population-specific components appeared among Native Mexicans from k=12 through 20, but with less clear contribution into the admixed Mexican genomes (Fig. S10). Likewise, all clusters detected at k=9 remained constant throughout the rest of runs up to k=20. In conclusion, as a result of the observations detailed above, we found k=9 to be the most informative run for purposes of characterizing sub-continental ancestry of Mexican populations, and therefore, several subsequent analyses described below were based on ADMXTURE proportions at k=9.

To check for possible convergence variation, we performed 10 additional runs using different random seeds per run, and the program converged after detecting the same clusters previously observed in all cases. We also estimated parameter standard errors using 200 bootstrap replicates per run. In general, standard errors were lower for individuals showing complete membership to highly divergent populations, such as Yoruba, Seri, Triqui, Tojolabal, and Lacandon (average error <0.01). In contrast, the two components accounting for most of the error at k=9 were Northern versus Southern European (standard error =0.029). The average error across all individuals and components was 0.016. The number of markers used is known to affect the performance of cluster-based algorithms. According to the ADMIXTURE guidelines (*18*), 10,000 markers suffice for continental-level distinction, while numbers closer to 100,000 are recommended for within-continent separation, assuming for instance European populations (i.e., $F_{ST} < 0.01$). Given that we are using more than 71,000 markers (using our global Affymetrix + Illumina dataset) and that all ancestral populations involved have $F_{ST} > 0.02$, we expect our ancestry estimates to be reasonably accurate. To test this assertion formally, we reran k=2 through k=20 using the global Affymetrix dataset (>370,000 markers) using the same settings described above and observed no significant differences in parameter estimates for individuals in both datasets.

Correlation of cluster membership and geographic coordinates: From the clustering patterns observed across Mexican states in the ADMIXTURE analysis, a clear correlation can be appreciated between the geographic location of samples and their membership to the six main Native Mexican clusters. To formally test for significance with Latitude and Longitude we performed a linear regression for each component using individual admixture values against their sampling location along a 45^{o} NW-SE axis across the country. We transformed latitude and longitude to create estimates across the "long axis" of Mexico, running NW-SE to better summarize the geography of Mexico in a single distance rather than latitude or longitude alone. Because the southern component decreases both northwards and towards the Yucatan peninsula, the correlation is less pronounced when Campeche and Yucatan samples are included.

Admixture maps: We used Kriging methods to interpolate ADMIXTURE proportion values for displaying the six native components identified at k=9 across both Native Mexican and cosmopolitan samples (Fig. S12). ADMIXTURE values from cosmopolitan samples (which usually show varying proportions of non-native admixture) were adjusted so that the sum of ancestry proportions coming from Native American components equals 1. Contour maps were created using MapViewer (Golden Software).

Local ancestry estimation

We used a PCA-based admixture deconvolution method (PCAdmix, (*20*)) to estimate local ancestry across the genome. This method uses phased genotype data to estimate posterior probabilities of ancestry for windows along each chromosome. First, ancestral populations are thinned for SNPs with $r^2<0.8$ in order to remove highly linked alleles from different populations, which can overfit and lead to spurious ancestry transitions. Second, chromosomes for each individual in a population are joined *in silico* to create two extended chromosomal haplotypes; this step allows us to use the full genome for PCA, and it is of special relevance when masking ancestry-specific portions of the genome (see below). Then, PCA on a number $k\leq3$ of ancestral populations is performed and the admixed population is projected into the determined $k\leq3$ PCA space. PC loadings are used as weights in a weighted average of the allele values in a window of 40 SNPs. These haploid window scores are then used as observed values in a Hidden Markov Model (HMM) to assign posterior probabilities to the ancestry in each window (where chromosome were considered separately). Two complementary algorithms, Viterbi and forward-backward, are used to compute estimates for each window. PCAdmix was implemented in C++ and is available at https://sites.google.com/site/pcadmix/. Additional performance testing and details of the implementation for this approach are available in (*20, 47, 48*).

The choice of k=3 ancestral populations for running PCAdmix was informed by ADMIXTURE results and is consistent with other investigations of ancestry in Latinos (Fig. 2B). Although continental-level ancestral populations are a good model at k=3, we observed that PCAdmix performance was improved when including reference panels representing a diverse set of haplotypes. In Mexicans, we expect most of the ancestry variation to come from the Native American (NAT) component rather than the European (EUR) or African (AFR) components. To empirically test the performance of different NAT reference panels in our Mexican dataset, we ran PCAdmix on a subset of 30 random samples using separately the different populations for which we had available trio data: Tepehuano (TEP), Nahua (NAH), and Maya (MYA). We limited our analysis to available trio data, as PCAdmix takes phased data as input. When comparing the three different possible NAT ancestral populations, we observed that comparable results were obtained when the populations were run separately. However, the proportion of windows called "unknown" was lower when using all three NAT populations combined. Therefore, we constructed our reference panel by combining five trios from each NAT population (those five showing the highest proportions of NAT global ancestry, 15 trios total), plus 15 CEU and 15 YRI trios as continental reference samples. We then separately ran PCAdmix on two groups of admixed Mexican samples: the 23 complete MXL trios from HapMap3, and the combined set of 362 unrelated cosmopolitan samples (N=312 from MGDP with available Affymetrix data plus N=50 from POPRES). The former set was trio-phased using BEAGLE, whereas the latter was population phased using phased MXL haplotypes as the

training set. Fig. S13 shows a schematic diagram of the workflow to assign local ancestry and further analyze ancestry-specific fractions of the genome.

Local ancestry scan: We plotted Viterbi posterior probabilities per window against physical distance along autosomal chromosomes to identify peaks of ancestry enrichment across the genome. We limited this analysis to EUR and NAT ancestries because AFR ancestry values were based on a much lower number of counts, making deviations from the mean incomparable. The R package ggplot2 was used to visualize normalized ancestry proportions (Fig. S20).

Ancestry-specific PCA (ASPCA)

We implemented a modified version of the subspace PCA (ssPCA) method originally described by Raiko et al. (*49*) to handle the large amount of missing data resulting from masking ancestry-specific segments across the genome of multiple individuals. Previous implementations have adapted the same algorithm to genotype data (*21*), thus limiting the analysis to loci of homozygous ancestry. In contrast, our method has been implemented for applying subspace PCA to haplotype data. To analyze ancestry-specific haplotypes derived from the admixed genomes of Mexican cosmopolitan samples we restricted to individuals with more than 25% of their genomes inferred from each continental ancestry. Continental reference panels were constructed to project Native American and European blocks separately. Three populations (Seri, Lacandon, and Tojolabal) were excluded from the Native American panel due to evidence of extreme divergence compared to the rest of the populations (and no NAT segments from admixed genomes were projected onto those clusters). The final panel consisted of 17 Native American parental populations. Our European reference panel included 1,387 POPRES individuals from throughout Europe with four grandparents from the same country (*10, 17*) plus 55 additional samples from Spain (*22*). We did not project AFR segments due to the low number of haplotypes across the population sample. To validate the consistency of our ASPCA results, we performed a supervised structure analysis using *frappe* (*50*) and observed clustering patterns in agreement with our ancestry-specific distribution in PCA space. Our implementation of the method is described in (*19*). For clarity, we quote the following passage as it appears in Text S1 from (*19*) detailing the algorithm:

"Overview of the ASPCA method (subspace learning algorithm): The method we describe here is a close adaptation of the *subspace learning algorithm* described in (*49*) to haplotype data. In contrast to the standard approach, which computes all principal components, the subspace algorithm does away with the covariance matrix altogether and computes the first d principal components, where $1 \leq d \leq n$. Specifically, given an m x n matrix of haplotypes, the algorithm seeks to obtain the decomposition $\mathbf{X} \approx \mathbf{AS}$, where $\mathbf{S}$ is a m x d matrix, and $\mathbf{A}$ is a d x n matrix containing the top *d* principal component loadings for every individual in the sample. For our purposes, we are interested in obtaining the latter to approximate PCA. In the absence of missing data, this decomposition can be obtained iteratively by gradient descent. Starting with random matrices $\mathbf{A}$ and $\mathbf{S}$, the following update rules are alternatively applied to each matrix until convergence is achieved:

$$\mathbf{A} \leftarrow \mathbf{A} + \gamma(\mathbf{X} - \mathbf{AS})\mathbf{S}^{\mathbf{T}}$$

$$\mathbf{S} \leftarrow \mathbf{S} + \gamma\mathbf{A}^{\mathbf{T}}(\mathbf{X} - \mathbf{AS})$$

where γ controls the learning rate. Note that the resulting matrices are not necessarily orthogonal. However, orthogonalization can readily be performed post-hoc. For instance, one can orthogonalize A by Singular Value Decomposition (SVD). Letting $\mathbf{A} = \mathbf{UDV}^{\mathbf{T}}$, the orthogonalization is computed as:

$$\mathbf{A}^* = \mathbf{UV}^{\mathbf{T}}$$

The progression of the algorithm towards convergence can be followed by tracking the change in the cost function C at every iteration, where C is defined as:

$$C = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \sum_{k=1}^{d} a_{ik}s_{kj})^2$$

Intuitively, this is the mean square error between the data matrix **X** and its estimate **AS**. Throughout the algorithm, C is expected to converge to a local optimum in a monotonically decreasing fashion.

Focusing on a specific ancestry component: introduction of missing data: Given this framework, the above equations can be readily adapted to the presence of missing data, corresponding to regions of the genome that have been masked out to enable the study of a specific ancestral component of admixture. Specifically, instead of iterating over all possible entries of the haplotype matrix, we now only focus on those that are non-missing (i.e., those determined by the ancestry deconvolution algorithm to be derived from the desired admixture component). Thus, the cost function becomes:

$$C = \sum_{(i,j)\in O}(x_{ij} - \sum_{k=1}^{d} a_{ik}s_{kj})^2$$

where O now denotes the set of all observed values in the haplotype matrix X. Concordantly, the update equations corresponding to the gradient descent algorithm becomes:

$$a_{ik} \leftarrow a_{ik} - \gamma\frac{\partial C}{\partial a_{ik}} = a_{ik} + \gamma\sum_{j|(i,j)\in O}\sum_{k=1}^{d}(x_{ij} - a_{ik}s_{kj})^2 s_{kj}$$

$$s_{kj} \leftarrow s_{kj} - \gamma\frac{\partial C}{\partial s_{kj}} = s_{kj} + \gamma\sum_{i|(i,j)\in O}\sum_{k=1}^{d}(x_{ij} - a_{ik}s_{kj})^2 a_{ik}$$

Implementation: Our implementation of the algorithm, which we packaged into the software PCAmask, follows the guidelines of Raiko et al. paper quite closely. Specifically, we adapted the standard gradient descent outlined above to include a speed-up term for faster convergence. We achieved this by multiplying the gradient by the inverse of the second order derivatives of the cost function, as described in Raiko et al.:

$$a_{ik} \leftarrow a_{ik} - \gamma \left(\frac{\partial^2 C}{\partial a_{ik}^2} \right)^{-1} \frac{\partial C}{\partial a_{ik}}$$

$$s_{kj} \leftarrow s_{kj} - \gamma \left(\frac{\partial^2 C}{\partial s_{kj}^2} \right)^{-1} \frac{\partial C}{\partial s_{kj}}$$

Finally, we followed the guidelines to set the convergence term γ. At the beginning of the algorithm, we set γ = 1. At every iteration, γ is then updated based on the new value C_{next} of the cost function. If $C_{next} < C$, we set $\gamma' = 1.1\gamma$; otherwise, the update of **A** and **S** is rejected and $\gamma' = \gamma/2$. This approach ensures that smaller steps are taken as the process nears the local optimum" (quoted from (*19*)).

Importantly, we implemented this method on haplotypes rather than genotypes as this allowed us to use much more of the genome (rather than just the parts estimated to have two copies of a certain ancestry).

Ancestry-specific clustering analysis

We implemented a modified version of the *frappe* clustering algorithm (*50*) in order to accommodate partial missing data resulting from masking specific sites of the genome (Fig. S16). Our analyses of ancestry-specific segments of the genomes in the Mexican individuals rely on haplotype data. This leads to the generation of heterozygous missing sites at SNPs inferred to be heterozygous for the desired ancestry. Since the original *frappe* method developed by Tang et al. (*50*) cannot process partially missing genotypes, we adapted the algorithm to process haplotype data. The algorithm relies on an EM algorithm to jointly infer overall ancestry proportions in admixed individuals and the ancestral allele frequencies at all sites used in the panel. While the standard *frappe* implementation integrates over the two observed alleles at every genotype, this integration is eliminated for haplotype data. Specifically, in the **M** step, an estimate for the ancestral allele frequencies is obtained from the best guess for ancestry proportions using the modified equation:

$$p_{mk}^{n+1} = \frac{\sum_{i \in O} h_{im} E_{imk}^{n}}{\sum_{i \in O} E_{imk}^{n}}$$

where p_{mk} is the allele frequency for ancestral population *k* at marker *m*, h_{im} is the observed allele on haplotype *i* (0/1-based), and *O* is the set of all haplotypes carrying the desired ancestry at marker *m*. (Note: for ease of notation, we drop the superscript which denotes iteration n or n+1). E_{imk} is a computational device indicating the expected ancestral contribution of ancestor *k* at haplotype *I* on marker *m*. Similarly, an estimate for the overall ancestral contribution q_{ik} of ancestral population *k* at haplotype *i* is obtained from:

$$q_{ik}^{n+1} = \frac{\sum_{m=1}^{M} E_{imk}^{n}}{\sum_{m=1}^{M} \sum_{i \in O} 1}$$

where the denominator simply corresponds to the total number of unmasked sites across all haplotypes used in the analysis. Finally, in the **E** step of the EM algorithm the quantity E_{imk} is updated based on the new estimates for overall ancestry proportion and estimated allele frequencies:

$$E_{imk}^{n+1} = \frac{p_{mk}^{n+1} q_{ik}^{n+1}}{\sum_{k'=1}^{K} p_{mk'}^{n+1} q_{ik'}^{n+1}}$$

This step is identical to the original version of the algorithm.

Biomedical associations with ASPCA values

We leveraged two studies of childhood asthma in Mexicans and Mexican Americans to determine important pulmonary associations with ancestry-specific PCA values. In particular, we focused on lung function as measured via spirometry using standard clinical measurements as ancestry has been shown previously to affect lung function (*29*). Both studies were trio-based, ensuring long-range phase determination in the probands. The genotypes included the same thresholds for quality control filtering as described in (*51*). For continuous lung function measurements, we transformed raw spirometric values into percent predicted values, which adjust for typical anthropometric measurements (i.e. age, sex, and height) (*28*). Informed consent was obtained from all individuals at the study sites prior to sample collection. Both studies have been described in detail elsewhere. We briefly describe each study below.

The Genetics of Asthma in Latino Americans (GALA I) study is a study of childhood-onset asthma in case-parent trios, including balanced recruitment of Mexican individuals from both the San Francisco Bay Area and Mexico City. (*52*). Additional Puerto Rican individuals in GALA I were excluded from our analyses. Samples were genotyped on the Affymetrix 6.0 array (*32, 53*). For this study we filtered to individuals sampled in Mexico City and the San Francisco Bay Area with four grandparents who all identified as Mexican or Mexican American. A goal of study design was to recruit equal numbers of children with "mild" vs. "moderate-severe" asthma, according to ATS standards. The minimum age at recruitment was 8, and the maximum was 40, with a median of 13. Lung function testing was performed according to ATS guidelines, both after withholding bronchodilator medications for eight hours (a pre- measurement) and after age-stratified dosage of bronchodilator (a post- measurement). Additional details of spirometry measurement are available in (*52*). We used PCAdmix for local ancestry estimation with HapMap CEU, YRI and Native Americans from (*33, 54*) as reference, combined with global admixture modeling via ADMIXTURE (*18*). After filtering for individuals with spirometry data and adequate levels of Native American ancestry for use with ASPCA, we were left with 68 individuals from Mexico City and 120 from the Bay Area.

The Mexico City Childhood Asthma Study (MCCAS) consists of trio-based sampling of individuals with asthma along with their parents, genotyped on the Illumina 550 platform (*31, 55*). All sampling was performed at a single site within Mexico City. The minimum age at recruitment was 5, and the maximum age was 17, with a median of 9 years old. Pulmonary function tests were done using the EasyOne® spirometer (nDD Medical Technologies, Zurich, Switzerland) according to American Thoracic Society guidelines. Three tests were performed on each subject and the best test was used in analysis. Testing was done on subjects who reported no respiratory symptoms on the day of testing. Genotyping was performed on the Illumina 550K Bead Chip, with details of the genotyping available in (*31, 55*). Reference data from HapMap CEU, YRI and HGDP Native Americans were used for local ancestry inference in PCAdmix. As these samples were generated on an Illumina platform, we used the Native Mexican samples from the Human Genome Diversity Panel (*56*) combined with CEU and YRI genotypes for local ancestry estimation using PCAdmix. We used global ancestry estimates from *frappe* (*50*) estimated previously (*31*). After filtering for individuals with spirometry and adequate levels of Native American ancestry we included 341 individuals in downstream analysis.

Studies were chosen to ensure long-range phasing from the trio designs. However, studies differed in age ranges, recruitment sites, and genotyping platforms. PCAmask can incorporate data from various platforms; however, with such a large number of masked samples compared to reference individuals, individual PCs can become distorted (*57*). To counteract this effect, we ran PCAmask separately on MCCAS and GALA I, resulting in two different PCA analyses with different coordinates (Fig. S17) from the limited set of intersecting SNPs available with the MCCAS Illumina data. Therefore, to combine analyses, we chose our standard unit of measurement in ancestry-specific PC space to be a standard deviation for each study. FEV_1 for both studies was transformed to percent-of-predicted-normal from equations derived for Mexican Americans (*28*), including age, sex, and height^2. Using percent-predicted values allows one to focus on simpler, more clinically relevant associations with FEV_1.

First, as GALA I includes individuals from both Mexico City and the San Francisco Bay Area, we wanted to investigate whether ASPCA values were associated with recruitment location. To do this, we used a likelihood ratio test of two different logistic regression models: a full model with ASPC1 & 2 along with global ancestry covariates; and a restricted model with simply the global ancestry terms. The statistic 2*log (likelihood ratio full:restricted) follows a 2-degree of freedom chi-squared distribution (one for each ASPC). We performed marginal tests for each ASPC using t-tests. We also estimated the raw Area Under the Curve (AUC) for a Receiver-Operator Characteristic curve including the two ASPCs using the *epicalc* package in R.

Next, for each study, we ran a separate robust linear model (rlm via *MASS* in R) to predict forced expiratory volume in the first second (FEV_1), using the ASPC values and adjusting for global ancestry covariates. We used robust linear models rather than OLS as PCA can have outliers that could potentially bias OLS estimation. Given normalized ASPC1 & 2 z-scores, the regressions took the form:

$$\%(predicted)FEV_1 \sim \beta_0 + \beta_1 z(ASPC1) + \beta_2 z(ASPC2) + \beta_3 African + \beta_4 Native + \varepsilon$$

Age, sex and height are incorporated in the percent predicted values to be able to compare effects across the entire growth curve in children. Global ancestry terms are used to adjust for any residual population stratification, and to ensure that overall levels of Native American ancestry do not confound potential associations with ASPCs 1 and 2. An additional regression was performed to estimate parameters for European ancestry by placing European ancestry in for the Native term, as the two ancestries are highly collinear.

We performed these regressions separately for GALA I and MCCAS, then combined the effect sizes for ASPC1 and 2 via random effects meta-analysis in the R package *metafor*. These values were then used for the reported p-values as they represent the largest combined sample and were independent replication with different recruiters, study designs, and genotyping arrays. Individual-study estimates for the association between ASPC1 and FEV_1 are available in Table S6. We extrapolated based on the ASPC1 association to the data from eight states to determine the change in FEV_1 due to differences in the origin of Native American ancestry. For context we then compared our inferred changes in FEV_1 with that explained by change in lung function due to age (*28*) and African ancestry levels in African Americans (*29*).

Because lung function measurement was performed with two different approaches in the two studies, we chose to use the post-bronchodilator measurement in GALA I as 1) ancestry associations with FEV_1 had been observed previously in healthy individuals, reflecting underlying physiological lung function rather than asthma specifically; and 2) GALA I by design over-sampled moderate and severe asthma. This feature is important because in Mexicans, asthma is known to be milder than in other Latino populations such as Puerto Ricans (*52*). However, as in most asthma studies, GALA I Mexican pre- and post- measurements are highly collinear (Pearson's R^2: 82%).

Additional analyses: Previous work on GALA I has argued that there is an association between Native American/European ancestry and FEV_1 (*30*). It is important to note that this previous study focused on pre-bronchodilator FEV_1 as a proxy for clinical severity (degree of lung function impairment with medication withheld), rather than overall lung function as here. Pre-FEV_1 was not measured in MCCAS; therefore, in the current study, we are looking at a different (but somewhat correlated) measure. We repeated the same analysis in the set of samples used for the ASPCA association with FEV_1. In univariate tests, European ancestry is positively correlated and Native American ancestry is negatively correlated with post-FEV_1 ($p=0.029$ and $p=0.0056$, respectively), consistent with previous work. However, when included in multiple regression models including African ancestry and either other ancestral term, the only significant ancestry term is the African one, demonstrating that in GALA I, there is a signal of association between continental ancestry proportions and post-FEV_1 primarily driven by the African component ($p=6.4\times10^{-5}$, $r^2=8.3\%$). This post-FEV_1 association does not replicate in MCCAS, however.

Multiple measurements of lung function exist, each focusing on different aspects of exhalation. Both GALA I and MCCAS include multiple measurements beyond FEV_1. Another standard measure is Forced Vital Capacity (FVC), based on the total volume exhaled during a breath. FVC is known to be a less robust measurement than FEV_1, particularly in children, as it requires compliance through the entire exhalation rather than the single second needed for FEV_1. We repeated the same analyses for two other standard measures of lung function: forced vital capacity (FVC, a measure of the overall size of the lungs) and the FEV_1/FVC ratio; however, neither of these values were significantly associated with either ASPC1 or ASPC2 in any marginal test or meta-analysis thus they were not investigated further.

Supplementary Results

Demographic simulations

We estimated effective population sizes (N_e) of different indigenous Mexican populations based on the REJECTOR algorithm models for the distribution of runs of homozygosity (ROH) (*40*) (see Methods), (Fig. S4 and S5). Acceptance rates varied between ~1-3% among runs. Across groups we found unimodal distributions for both bottleneck strength and current population size, suggesting a robust inference method. However we did identify differences between groups. For example, we estimate N_e = 1200 chromosomes for the current population size of the Seri, one of the most historically isolated groups. In contrast, larger ethnic groups, such as the Maya, have expanded to effectively more than 3,500 chromosomes (Fig. S4). The current N_e values as measured by ROH are consistent with a strong historical bottleneck as with previous estimates on the number of founders of the Americas (*13*). The relative similarity speaks to the consistent demographic scenario tested in all models (initial population, bottleneck into the Americas, and subsequent expansion).

Population differentiation and IBD

To measure population genetic differentiation among extant groups we computed overall pairwise F_{ST} across all autosomal sites (Fig. 1B). The highest value was observed between Seri and Lacandon (0.14), followed by Seri vs. Tojolabal (0.12) and Seri vs. Triqui (0.10). Both Seri and Lacandon also showed elevated F_{ST} values across all other populations, while lowest F_{ST} values were observed among groups from central Mexico and within the Yucatan peninsula (Fig. 1B).

To evaluate the impact of population isolation in genetic similarity, we measured the total length of segments inferred to be identical by descent (IBD) among all possible pairs of individuals using GERMLINE (*46*) with a minimum threshold of 5cM (see Methods). We visualized both between- and within-population connections binned into nine levels of relatedness (Fig. S8). Fig. 1C shows the approximate location of sampled populations and their connections among individuals sharing segments of total IBD above 20cM (corresponding to the genomic equivalent of 3rd cousins or closer relatives). We observed high within-population IBD levels compared to between-populations, indicating that after splitting, indigenous populations have largely remained isolated. Some exceptions include either Nahua (i.e. NAJ, NXP, NAG) or Mayan (i.e., MYA.C, MYA.Q, MYA.Y) populations, both of which are some of the most populous indigenous groups in Mexico, resulting in a lower probability of observing within-population connections in our sample. Two groups of closely related populations show higher number of between-population connections: Totonac and Nahua from Puebla (NXP and NFM), and Tzotzil, Tojolabal, and Lacandon from Chiapas (Fig. 1C).

Population phylogeny analysis

To formally evaluate the probability of gene flow between populations after splitting, we used TreeMix (*14*) to construct a maximum likelihood tree allowing for a fixed number of migration events between populations. Fig. 1D shows the splitting pattern without migration, which

recapitulates the North/South gradient of differentiation observed in our previous analyses with Seri and Lacandon showing the highest levels of drift from the ancestral population, followed by Tojolabal. Shared clades denote clear regional relationships, such as all northern populations branching out from the same initial split at the root, followed by individual population splits and two major clades: one grouping all populations from the southern states of Guerrero and Oaxaca (Triqui, Zapotec south, Zapotec north, Mazatec, and Nahua Guerrero), and the other all six Mayan speaking populations from the state of Chiapas and the Yucatan peninsula (Tzotzil, Tojolabal, Lacandon, Maya Campeche, Maya Quintana Roo, and Maya Yucatan). When running TreeMix allowing for migration edges in the tree, the matrix of residuals is used to infer pairs of populations with the poorest fit, thus becoming candidates for testing a better fit involving migration between them. Recent admixture can bias these estimations so we removed all indigenous samples with more than 2% of European ancestry as inferred by ADMIXTURE (*18*). We focused on the maximum likelihood trees for the top three events of migration (m=1 to 3) inferred from the data (Fig. S9). Interestingly, the first migration inference (m=1) involves gene flow from the Maya in Yucatan (MYA.Y) to the node of the Totonac (TOT), whose ancestors are believed to have built the large pre-Columbian city of El Tajin, located near the coast of the Gulf of Mexico, revealing a possible coastal corridor of gene flux between the Yucatan Peninsula and Central/Northern Mexico. The strongest migration rate (consistently greater than 50%) was detected between two closely related Nahua populations (NXP and NFM) both at m=2 and m=3. In the latter case an additional gene flow event was inferred from the Totonac to the neighboring Nahua in Puebla (NXP), consistent with the IBD patterns observed in Fig. 1C.

It is noteworthy that the different Nahua groups, while unified by historically speaking the same language, stem from different nodes in the tree. For example, NAJ from Jalisco is separated from the node giving rise to NXP and NFM (both from Puebla), and NAG from Guerrero is grouped together with Zapotec and other groups from southern Mexico. This translates into a lack of a single ancestry relating all the studied Nahua groups (as opposed to the Mayan groups, for instance), suggesting that current groups identified as Nahua are likely the result of linguistic and cultural assimilation over genetically distinct groups, probably as a result of the extended domination of the Nahua-speaking Aztec empire in pre-Columbian times.

Population substructure analysis

We used ADMIXTURE (*18*) to analyze the combined dataset of continental source populations (including our 20 native Mexican populations, 16 European populations, and 50 West African Yorubas) and 420 admixed individuals from 11 Mexican states as well as 49 Mexican Americans from the Los Angeles area (Fig. 2A and 2B). At K=3, each set of reference parental groups gets its own cluster, with the exception of some Native Mexican groups such as Nahua and Maya, previously documented to have considerable proportions of European admixture (*33, 56*).

Across the Mexican cosmopolitan samples we observe a clear gradient of increasing Native American and decreasing European ancestry moving southwards, consistent with previous genome-wide reports of Mexican admixture patterns (*24*). African ancestry proportions are low on average (4.9%) and remain similar across most regions, with the exception of the coastal states of Veracruz and Guerrero. Both states are known to have had increased slave trade activity (*58*), and some individuals from these states today show considerably higher proportions of African ancestry (up to 34%), also consistent with previous analyses of a subset of these samples

at K=3 (*24*). However, more in-depth analyses of ancestry were not possible in the initial study as a single Native Mexican group, the Zapotec, was used as potential source population, precluding any further detection of sub-continental ancestry.

With a larger reference panel of 20 native populations, we observe more detailed substructure at higher K values. We explored clustering patterns from K=2 through 20 (Fig. S10) and focus on K=9 for showing the lowest cross-validation error across runs (Fig. S11). At this level, the Native cluster breaks down into six separate Native American components (Fig. 2B). Three of them are restricted to isolated populations (Seri, Lacandon, and Tojolabal), showing little sharing with neighboring indigenous groups. The other three show a wider but geographically well-defined distribution. First, there is a northern component represented by Tarahumara, Tepehuano, and Huichol, which gradually decreases southwards until is virtually absent in Oaxaca and beyond. The second component is represented by southern populations from Oaxaca including Triqui, Zapotec, and Mazatec, reaching 99.9% in most Triqui individuals, and gradually decreasing northwards. In contrast, there is a sudden disruption moving towards the Yucatan peninsula, where this southern component is limited to an average of 20% of the genome and is mostly replaced by a local Mayan component, the third major component observed (Fig. 2B, bottom panel). Interestingly, this Mayan component is also present at ~10-20% in central native populations, but not in southern Oaxaca, supporting the hypothesis of a coastal or maritime route of gene flow between the Yucatan peninsula and central Mexico bypassing the mountain range of the Tehuantepec isthmus.

Additionally, we detected substructure within the European component at K=9, with a clear gradient of differentiation between northern European and southern Mediterranean populations, in agreement with previous analyses (*10, 59*). In all Mexican samples, the majority of European ancestry comes from the southern Mediterranean component, consistent with historical records about the admixture process between Spanish Europeans and native Mexicans. The map in Fig. 2A summarizes individual admixture proportions into population averages for each continental ancestry at K=3 and each native component at K=9 (see **Table S5**). For instance, Oaxaca and Campeche share similar continental patterns, showing the highest averages of native ancestry at K=3 (85% and 80%, respectively). However, when broken down at K=9, we reveal that their native proportion is composed of completely different profiles, dominated by their corresponding local native components.

Haplotype sharing analysis

For haplotype sharing analysis, we used the densest dataset (785,663 SNPs) consisting of 674 unrelated samples genotyped on both Affymetrix 500K and Illumina 550K SNP arrays. This included a combined group of 71 Native Mexicans (Tepehuano n=20, Zapotec n=21, and Maya n=30), as well as 312 cosmopolitan Mexican samples from the states of Guerrero (n=50), Guanajuato (n=48), Sonora (n=48), Tamaulipas (n=17), Veracruz (n=50), Yucatan (n=49), and Zacatecas (n=50). Sampling locations are reported in **Table S1**. To evaluate the level of haplotype sharing with diverse populations from other regions of the world, we also included a subset of HapMap continental reference samples: CEU (n=62), YRI (n=100), MXL (n=44), and CHB+JPT (n=85). Merged and curated genotype data were phased using BEAGLE software (*44, 45*). To phase the Mexican mestizo samples, we used the 22 MXL trios from HapMap3 as training set. The Tepehuano and Maya trios were used to improve phasing of the Zapotec, with

Tepehuano and Maya trios (n=10 and n=15 trios, respectively) phased separately using the trio sampling structure. HapMap populations with available trio data (CEU n=31 trios, YRI n=50 trios, and MXL n=22 trios) were also trio phased, whereas for CHB+JPT (n=85 unrelated individuals), we performed population phasing.

Genome-wide haplotype sharing (GWHS)**:** To determine the potential use of Mestizo and Native population data as reference for the genetic analysis of candidate regions and GWAS in Mexicans, we performed GWHS analysis using all available SNP genotypes within 100Kb windows of the genome. We used BEAGLE phased genotype data and then estimated all plausible haplotypes within each segment across populations using PHASE (*60, 61*). GWHS was assessed by comparing the number of common haplotypes (with frequency >5% across populations) shared between Mexican Mestizos and the different HapMap populations as well as Native Mexicans (Fig. S18).

The proportions shared between Mexicans and HapMap populations were comparable (SD from 1.4 to 3.0) across chromosomes. On average, Mexicans shared 21.6% with YRI, 54.8% with CHB+JPT, 59.3% with CEU, 78.6% with Native Mexicans and 81.2% with MXL. The proportion of shared haplotypes with CEU+CHB+JPT was 76.2%, and this was increased to 90.5% when the MXL group was added, and finally to 98.8% when Mexican Natives were included as reference (Fig. S18). These results indicate more sharing than those previously reported (*24*) due to a higher density of markers included in the analysis (capturing more variants in linkage disequilibrium [LD]) and due to the availability of data from Native Mexicans.

Tag SNP selection efficiency in candidate regions: To determine the potential use of Mestizo and Native Mexican tagSNPs for targeted studies, 10 candidate gene regions were selected for containing SNPs previously associated to diseases or traits of clinical interest, including: non-alcoholic fatty acid disease (*PNPLA3*), dyslipidemias (*ABCA1*), age-related macular degeneration (*ARMS2*), response to hepatitis C treatment (*DDRGK1*), Crohn's disease (*NOD2*), asthma (*PTGDR*, *NOTCH4* and *GC*), metabolic syndrome (ApoB) and systemic lupus erythematosus (*IKZF1*). All genes are included in the Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies); in addition, two of them, *ABCA1* (*62*) and *PNPLA3* (*63*) house genetic variants that have been identified in Mexicans or Hispanic populations.

Across all populations analyzed, we identified tag SNPs in these 10 candidate gene regions using Tagger, the tag SNP selection algorithm from Haploview software (*64*), with SNPs of frequency >5%, considering pairwise tagging only and an r^2 threshold of 0.8. We evaluated the performance of tag SNPs and their underlying coverage by estimating coverage from tag SNPs to the rest of the SNPs available in each gene using a pairwise r^2 approach. In a similar fashion to the GWHS analysis, we evaluated the mean best r^2 coverage based on the tag SNPs determined using various reference panels. Of the 10 candidate loci, two had fewer than 10 SNPs and were dropped for this analysis, resulting in eight genes evaluated using multiple reference ancestral groups. While the individual results vary from gene to gene, using the whole reference panel of Mexican Mestizos resulted in the best tagging performance overall, even better than using the MXL population from HapMap3 (Fig. S19). The results of this analysis underline the importance of using reference datasets of populations with the same LD structure for a better analysis of genetic variation in recently admixed populations such as Mexicans.

To search for a potential relationship between the enrichment in a particular ancestral component in the region with the haplotype sharing and tagging results, we analyzed the local ancestry estimations for each of the 10 regions included in this analysis (Fig. S20). We did not find any clear relationship between local ancestry and proportion of shared haplotypes. Looking at more detail in the haplotype diversity observed in these regions, we observed that in those regions with the highest European or Native American ancestral contribution, corresponding respectively to *ABCA1* and *ARMS2*, differential ancestry is not related to differences in haplotype diversity or tagging performance. In both cases, ancestral contribution differences are clearly related to differences in the frequency of specific haplotypes that, even if shared with all other populations, show distinct frequency differences in ancestral groups. This pattern is clearly illustrated in the *ARMS2* gene region where all common haplotypes (>1%) present in either Mestizo and Native Mexican groups are shared with least one HapMap group, but two haplotypes are enriched in Native Mexicans (87%) and Mestizo (72%) as compared to CEU (50%).

The results of the genome-wide and candidate region haplotype diversity showed that Mexican Native and Mestizo groups show a haplotype structure not fully represented in continental groups of the HapMap3 reference population set, which is comparable to other publicly available resources such as 1000 Genomes in terms of the Mexican diversity represented. Even including the closely related MXL population as reference does not achieve the optimal effect of using the combined Mexican groups. This is likely due to the fact that Mexican-Americans included in the MXL sample have a heterogeneous origin and thus a genetic structure of limited representation when compared to a comprehensive sample across the country. These results support the fact that a deep genetic characterization and inclusion in association studies of recently admixed populations such as Mexicans represent a great opportunity to discover new genetic variation of relevance for biological trait and disease gene mapping.

The selection of tag SNPs in candidate regions is of critical relevance for the improvement of genetic studies in Latin America, as this approach would enable the selection of small sets of SNPs for cost-effective study designs in candidate regions derived from GWAS or WGAS in other populations, with the aim of looking for new variants or haplotypes contributing to the genetic structure of biological traits or disease risk. Our results show that using the Mexican dataset generated here as a reference population translates into a better haplotype capture than using SNP sets based on the use of combinations of population groups from currently available catalogs of variation.

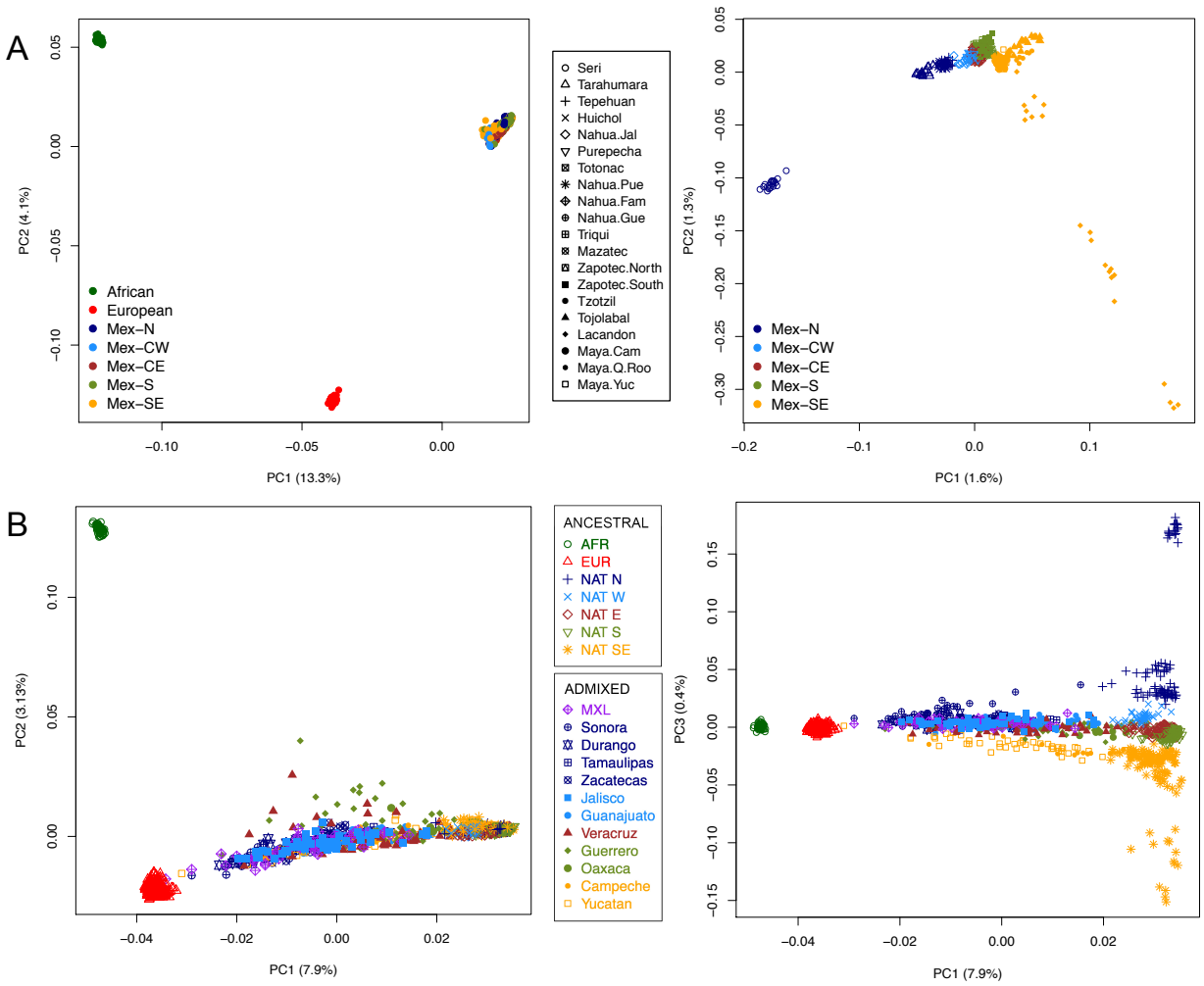

Figure S1: Principal component analyses based on the global dataset of ancestral and admixed Mexican populations. (**A**) *Left*: Global dataset of Native Mexicans combined with HapMap3 YRI African and CEU European samples. *Right*: Global dataset of Native Mexicans alone. (**B**) Combined dataset of ancestral reference samples (African, European, and Native Mexican) and admixed Mexican samples from cosmopolitan populations throughout Mexico and Mexican-Americans in the Los Angeles area (MXL). Populations are color-coded by geographic regions as follows: North (N), Central west (CW), Central east (CE), South (S), and Southeast (SE). *Left*: we observe a continuous dispersion of admixed individuals between the European and native Mexican cluster along PC1, reflecting their genome-wide average of native ancestry. PC2 separates a few individuals with higher African ancestry, predominantly from the coastal states of Veracruz and Guerrero. *Right*: along PC3, cosmopolitan samples from different states tend to be separated by the different native clusters in a north-to-south direction. For example, Yucatan and Campeche individuals form an elongated cluster that is clearly pulled in the direction of the Mayan individuals. Likewise, Sonora individuals with higher native proportions fall closer to northern native clusters. However, the separation is much more subtle among states from central Mexico, probably because standard PCA methods rely on genome-wide averaged signals from diploid genomes, making it difficult to ascertain finer scale patterns of differentiation.

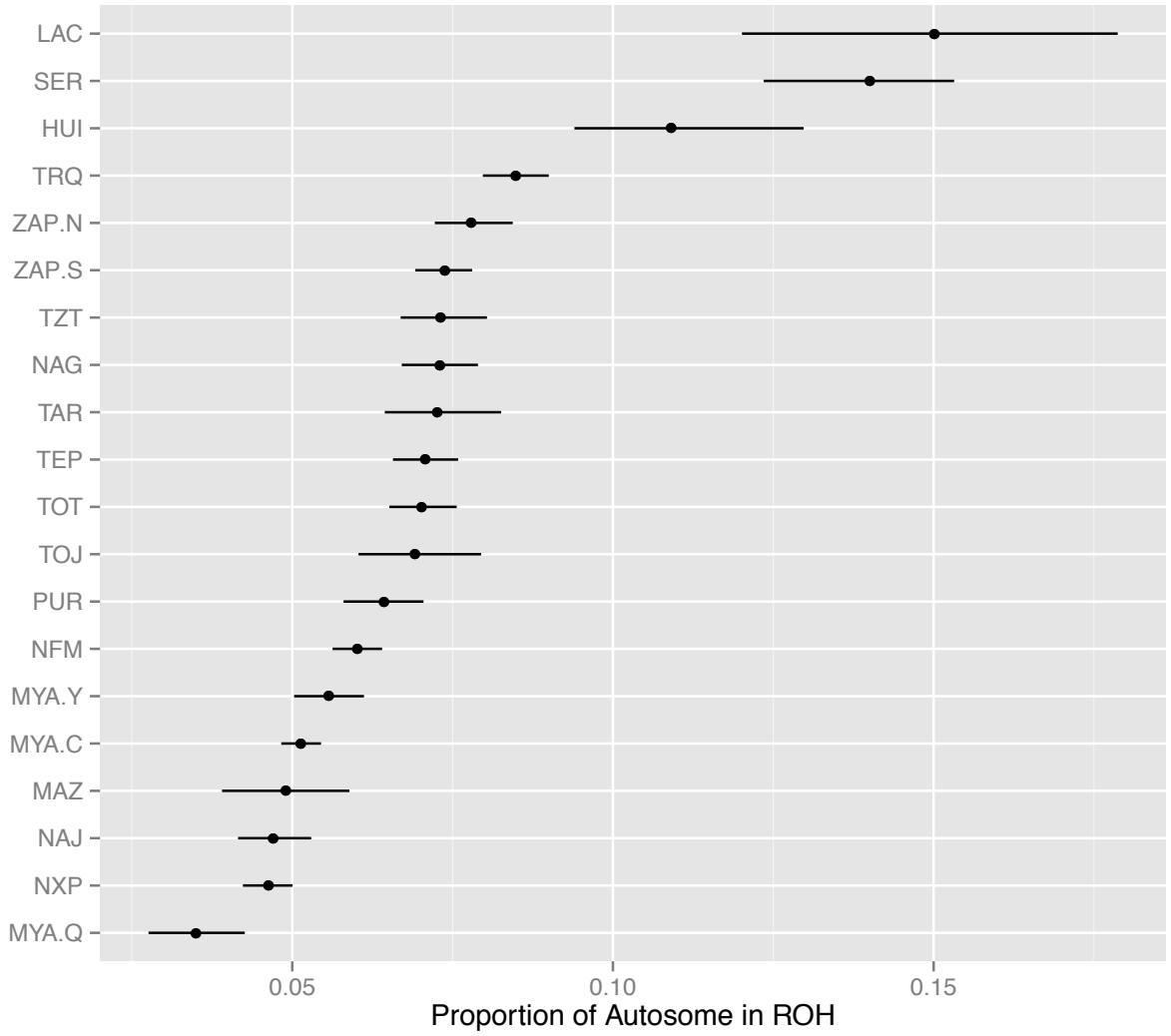

Figure S2: Proportion (%) of the genome in runs of homozygosity (ROH) per population. The total ROH per person was divided by the length of the autosomes (approximately 2.8Gb) and plotted the mean and confidence interval for each Native Mexican population. ROH sliding window values were estimated from genotype data using 372,692 SNPs in our combined Affymetrix dataset (Table S2), hence all 20 studied Native Mexican populations were used in the analysis. Samples are sorted by increasing values of ROH. Low values of ROH are mostly observed in samples from large populations such as Mayan and Nahuan groups, whereas higher ROH proportions are concentrated in more isolated populations, with the Lacandon and Seri representing extreme cases, followed by Huichol and Triqui. Population codes as in Table S1. Individual population profiles are available in Fig. S3.

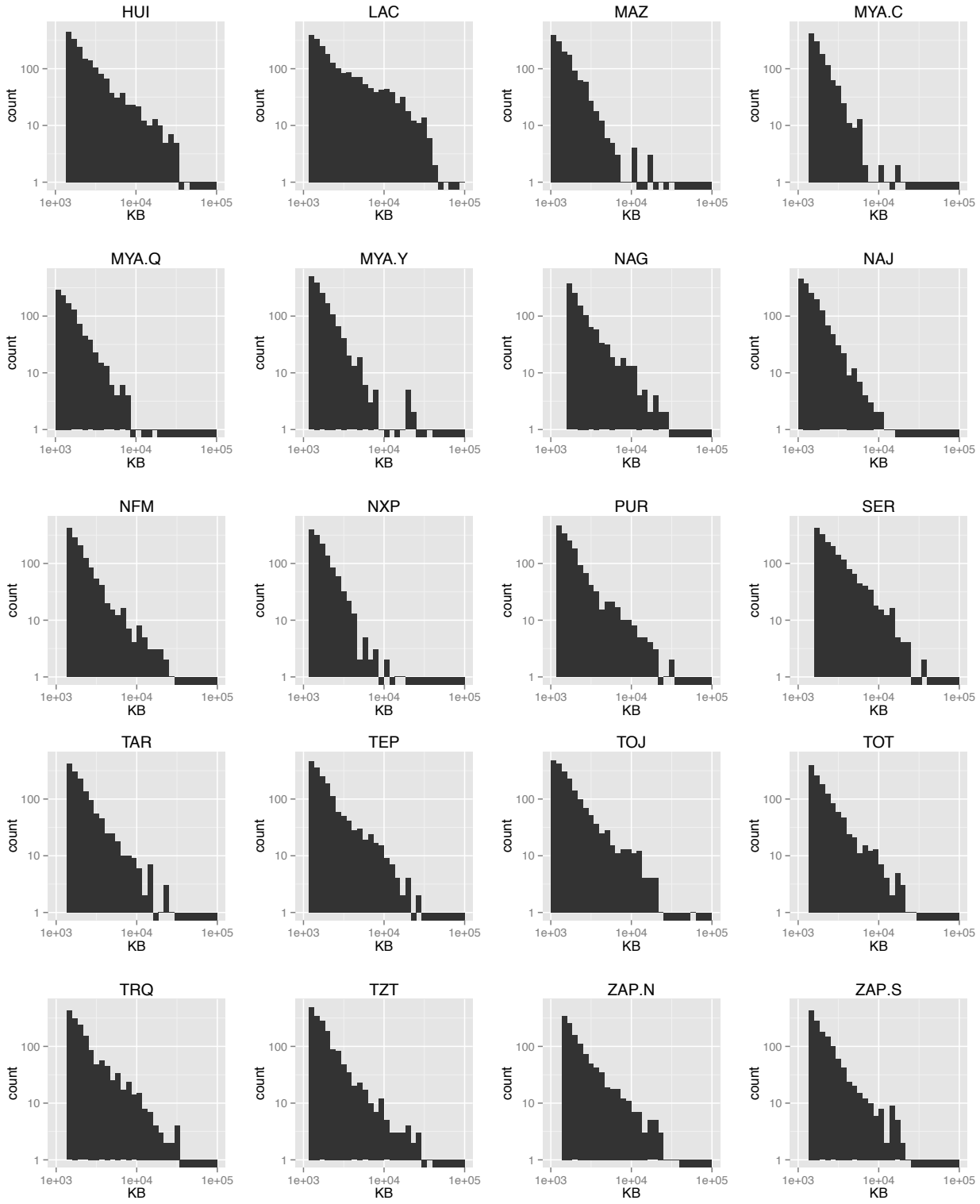

Figure S3: Individual population profiles of the distribution of ROH values in different Native Mexican samples. For each population, the histogram shows on a log scale the frequency of runs of homozygosity observed in different size ranges of physical distance (kb). Summarized means and confidence intervals per population are shown in Fig. S2.

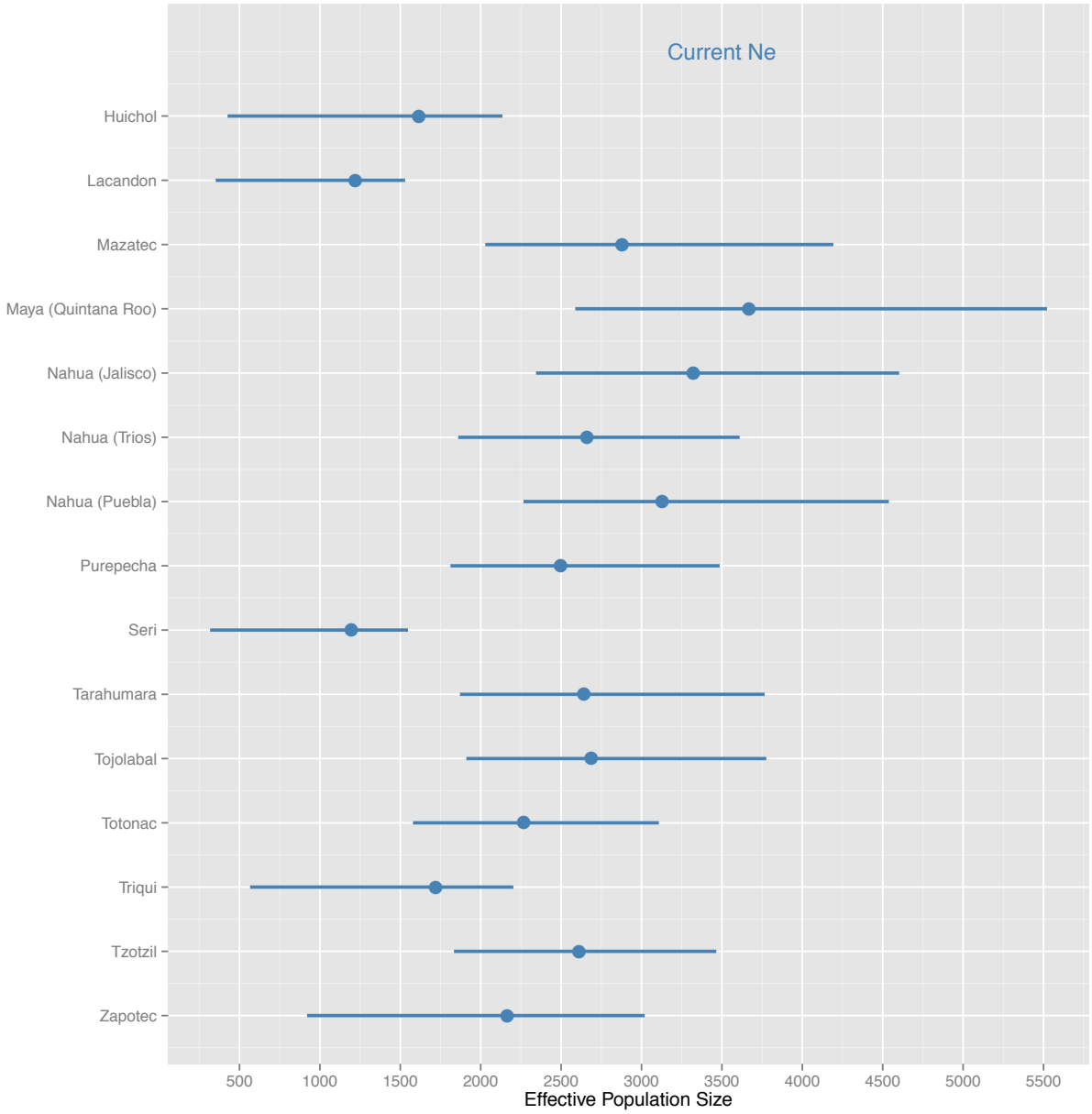

Figure S4: Summary of parameter estimates for the effective population size (N_e) in different Native Mexican population samples. Estimated current N_e is given per population showing 95% confidence intervals. Parameters were estimated from cumulative runs of homozygosity (cROH) on chromosome 1 via a rejection algorithm comparing observed and simulated data with REJECTOR (see Methods for details). In order to use the maximum density of genotyped SNPs along chromosome 1, we restricted to Native Mexican populations for which Affymetrix 6.0 array data was available (see Table S1).

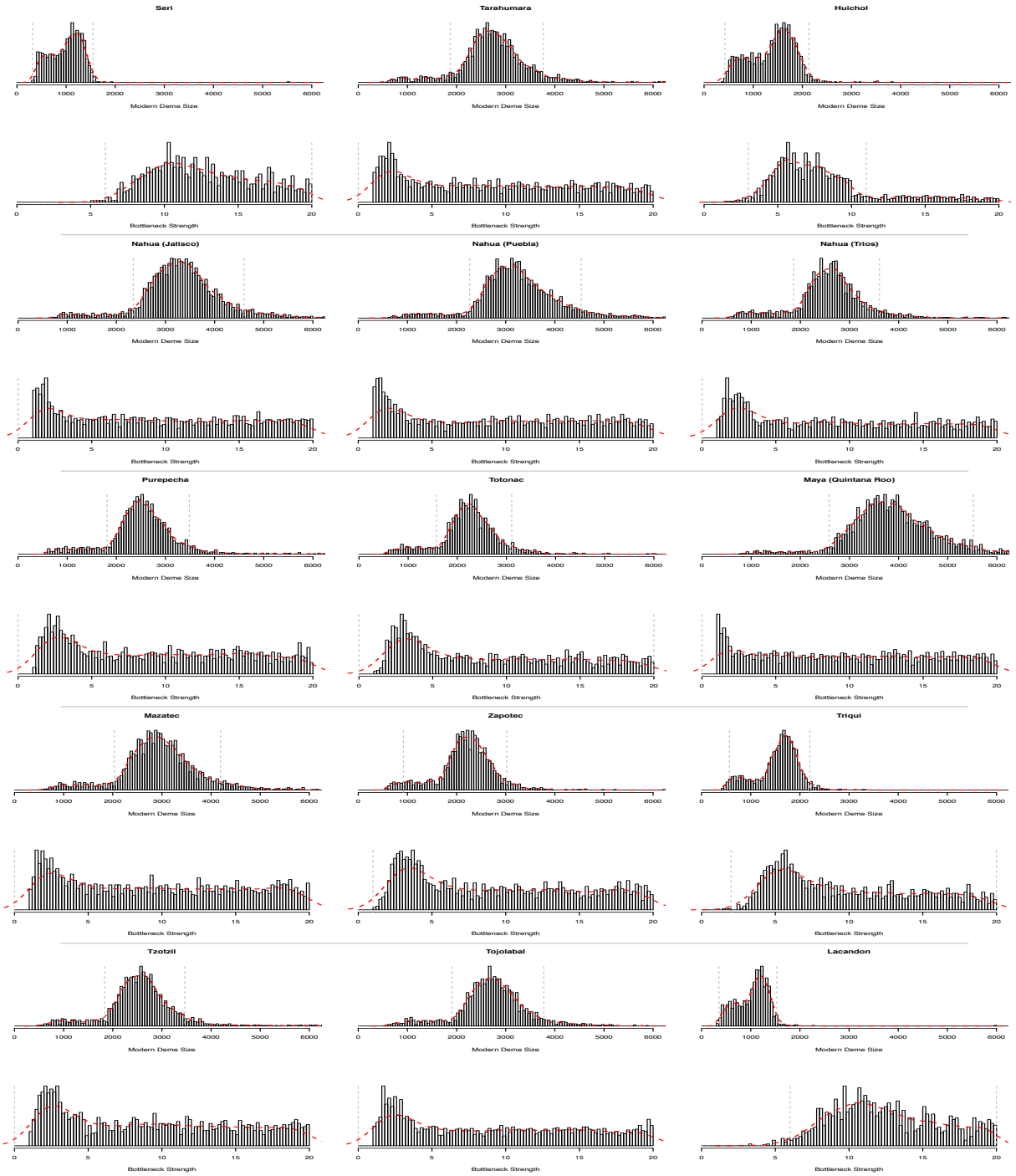

Figure S5: Individual population profiles of the simulated posterior distribution of effective population sizes in different Native Mexican samples based on cumulative runs of homozygosity (cROH). For each population, contemporary N_e (top histogram) and bottleneck strength (bottom histogram), were estimated by sampling from a uniform distribution of N_e and keeping simulated parameters within 20% of the observed cROH with REJECTOR (see Methods). Each histogram shows the frequency of accepted simulations and the smoothed density values used for estimating the final parameters shown in Fig. S4.

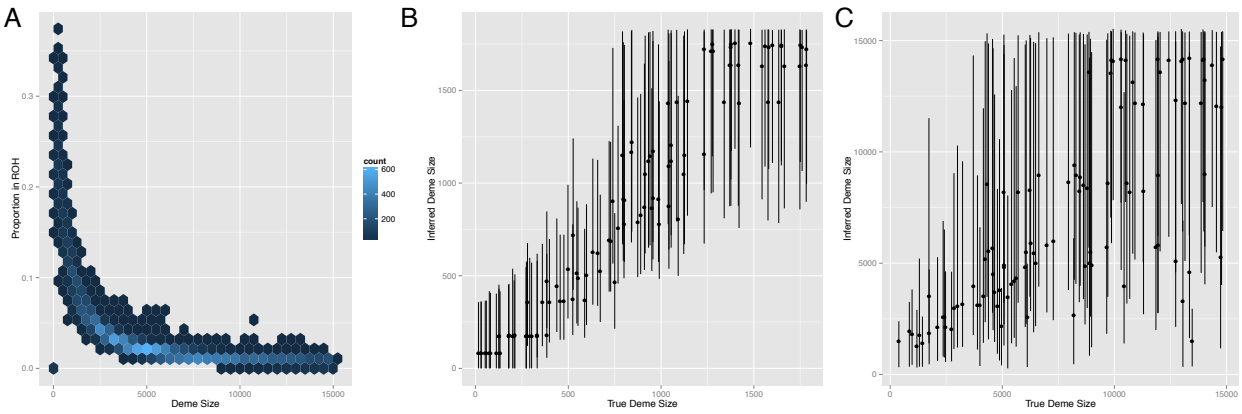

Figure S6: Testing demographic simulations. (**A**) Fraction of the genome in ROH as function of deme size based on full simulated chromosomes. (**B**) Concordance between true vs. inferred deme size based on simulated data with no bottlenecks for 2-2000 chromosomes. **(C)** Concordance between true vs inferred deme size using the same rejection algorithm parameters applied to the real Native American data and presented in Figs. S4-S5 (see Methods for details).

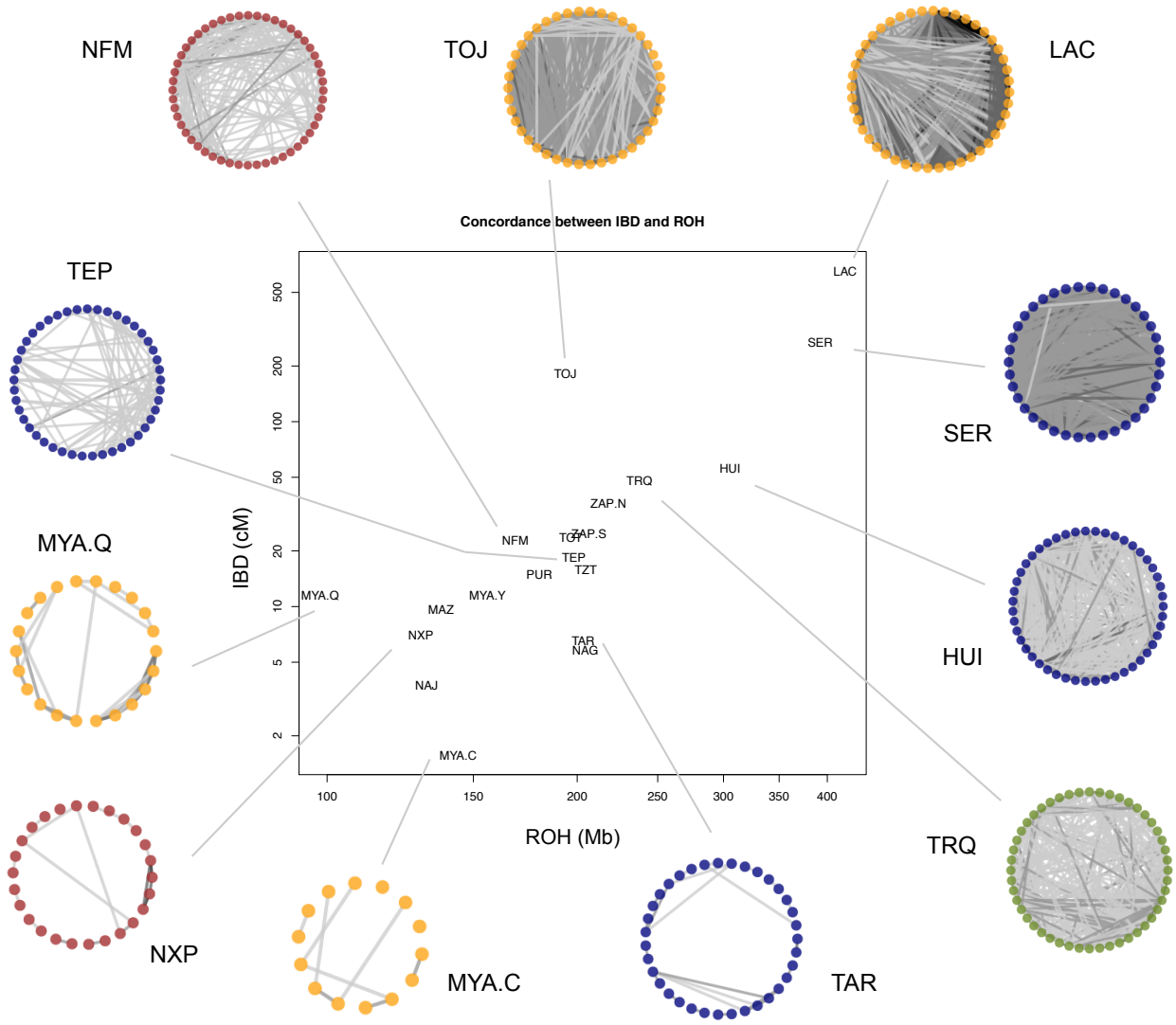

Figure S7: Concordance between Identity-by-Descent (IBD) and Runs of Homozygosity (ROH) for each population and patterns of within-population IBD sharing. The correlation on a log-log scale is quite high (Pearson's R^2 61%, p=5.4x10^{-5}), consistent with the similar information provided by each type of tract length. However half-IBD can uncover familial relationships when a population is an outlier from the overall trend. Most populations fall near the line, with the exception of the Tojolabal, suggesting that additional cryptic relatedness is potentially present in the dataset beyond that which is found in the overall population. Average IBD sharing within each population is plotted for all 20 indigenous groups represented in the combined Affymetrix dataset (Table S2). Around the plot, detail of within-population pairwise IBD matches greater than 13 cM is also shown for selected populations. Nodes represent haploid genomes of samples color-coded by geographic region as in Fig. 1C and edges are proportional to the total amount of shared IBD between pairs (darker lines denote higher sharing according to bins in Fig. S8). A force-directed network for all populations is shown in Fig. 1C.

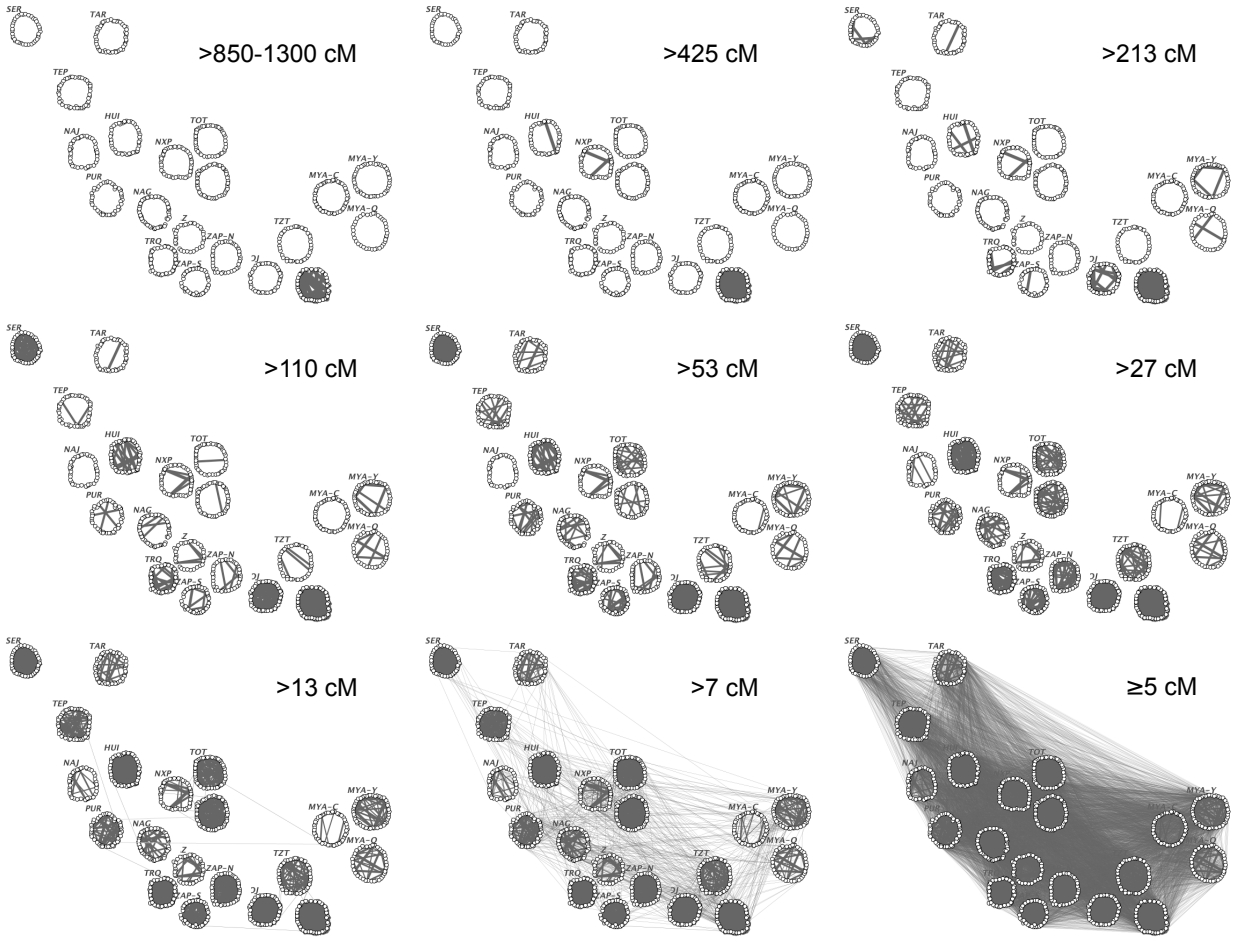

Figure S8: Patterns of relatedness within and between Native Mexican populations as measured by the total amount of segments identical-by-descent (IBD) shared between pairs of individuals. Each dot represents one haploid individual and each line denotes a pairwise match between two individuals sharing more than a given amount of total IBD. To minimize false positive IBD matches a minimum length of 5 cM was required to be considered in the analysis. Values of total IBD (in cM) were binned into consecutive categories to approximate the following proportions of the genome: 50% and above, 25%, 12.5%, 6.75%, 3.37%, 1.69%, 0.85%, 0.42%, and 0.21%, which intend to reflect the first 9 degrees of relatedness. Each plot shows the network of cumulative connections resulting from each of these IBD thresholds, meaning that lower IBD thresholds include connections at the indicated range in cM and above. In order to provide geographic context, individuals from the same population are displayed in positions that approximate the location of the sampled populations. The pattern across different populations shows high within-population sharing compared to between-populations for bins above 13 cM. At this level, between-population connections are limited to either relatively neighboring groups (e.g., Tzotzil-Tojolabal-Lacandon in the southeastern state of Chiapas or Totonac-Nahua in northern Puebla), or to ethnic groups connected to central Mexican populations like the Purepecha that show relatives with both northern Tepehuano and southern Triqui. Of note is that despite the longer distance, Mayans from Campeche in the Yucatan peninsula show IBD sharing with the Totonac near the Gulf of Mexico, supporting gene flow between these two regions.

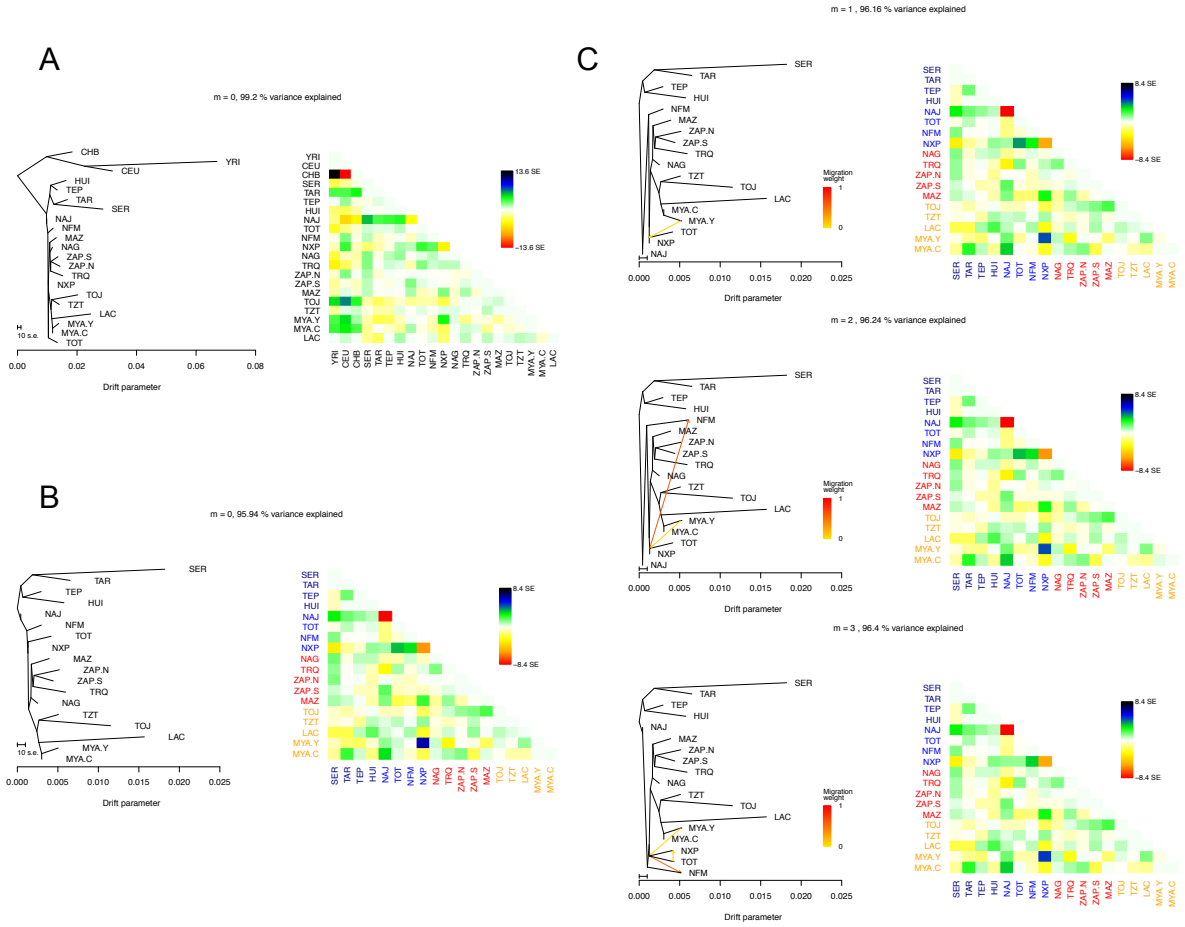

Figure S9: Maximum likelihood trees as inferred by TreeMix representing splitting patterns of Native Mexican populations and inferred migration events. (A) TreeMix graph depicting the relationships among Native Mexican populations along with three continental outgroups (HapMap YRI, CEU, and CHB). The length of the branches is proportional to the drift of each population. The resulting topology informed the position of the root in subsequent analyses (i.e., between all four Northern native populations and the rest). (B) TreeMix graph of Native Mexican populations alone without allowing for migration. The matrix next to each graph summarizes the residuals from the fit of the model to the data, where extreme values indicate populations that could be better modeled when adding migration to the model. (C) Models allowing for 1 to 3 events of migration (m = 1 through 3). Trees were constructed using the known topology from B and including samples with more than 98% of Native American ancestry. Arrows indicate migration edges and directionality of gene flow. Color intensity is proportional to the inferred amount of gene flow according to the migration weight bar. Residuals for each model are presented in pairwise matrices next to each graph.

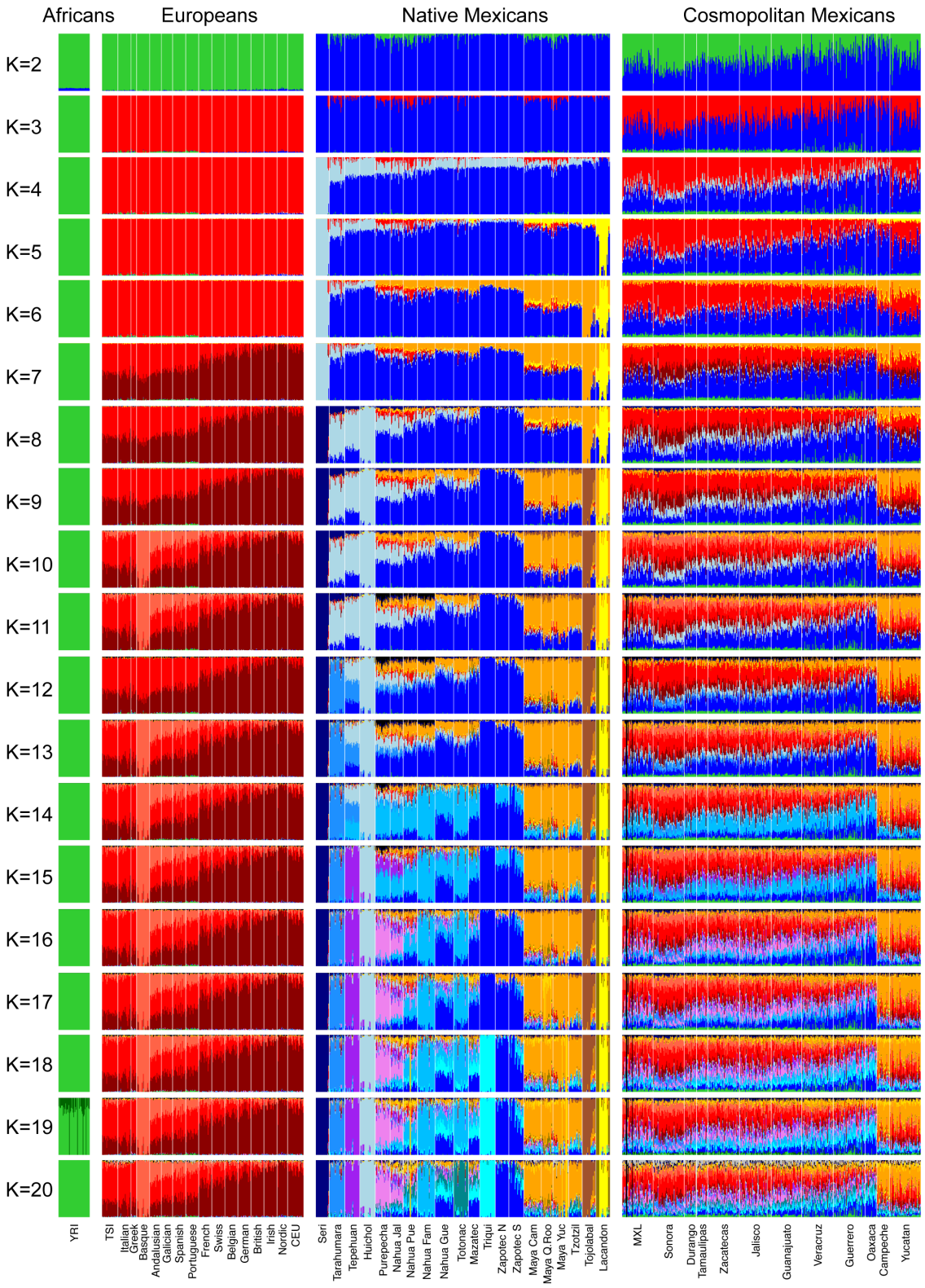

Figure S10: Unsupervised ADMIXTURE results from *k*=2—20 based on the intersection of Affymetrix and Illumina data (71,581 SNPs) from 1,282 samples (see Tables S1, S2). The

analysis includes 454 Native Mexicans, 469 Mexican mestizos, 309 Europeans, and 50 Yorubas. Each vertical bar represents an individual and the y-axis the proportion of the genome assigned to each of the ancestral clusters. Substantial substructure dominates the Native American component of both indigenous and cosmopolitan Mexican samples. European substructure is mainly driven by two sub-continental components following a North-South gradient, with the Basque clustering apart from the rest at k=10 and higher. We limited the representation of West Africans to a subset of HapMap YRI samples due to the study's focus on Native American diversity (see Tables S1, S5 for details).

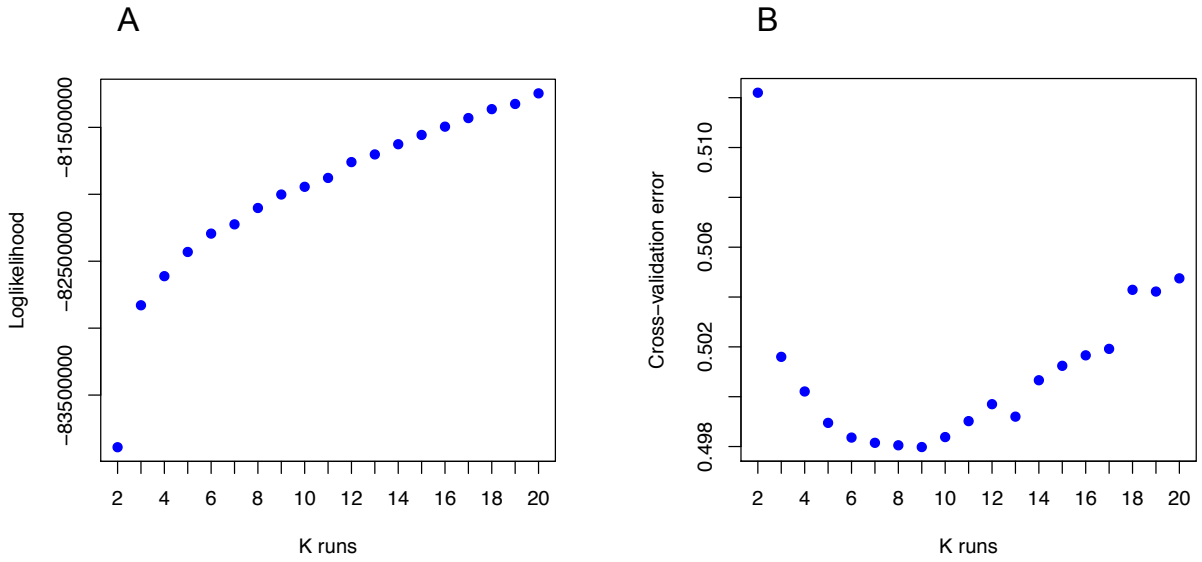

Figure S11: ADMIXTURE metrics at increasing *k* values based on Log-likelihoods (A) and cross-validation errors (B) for results shown in Fig. S10. While increasing clustering levels were associated with a continuous increase of likelihood values (*left*), *k*=9 showed the lowest error after cross validation (*right*).

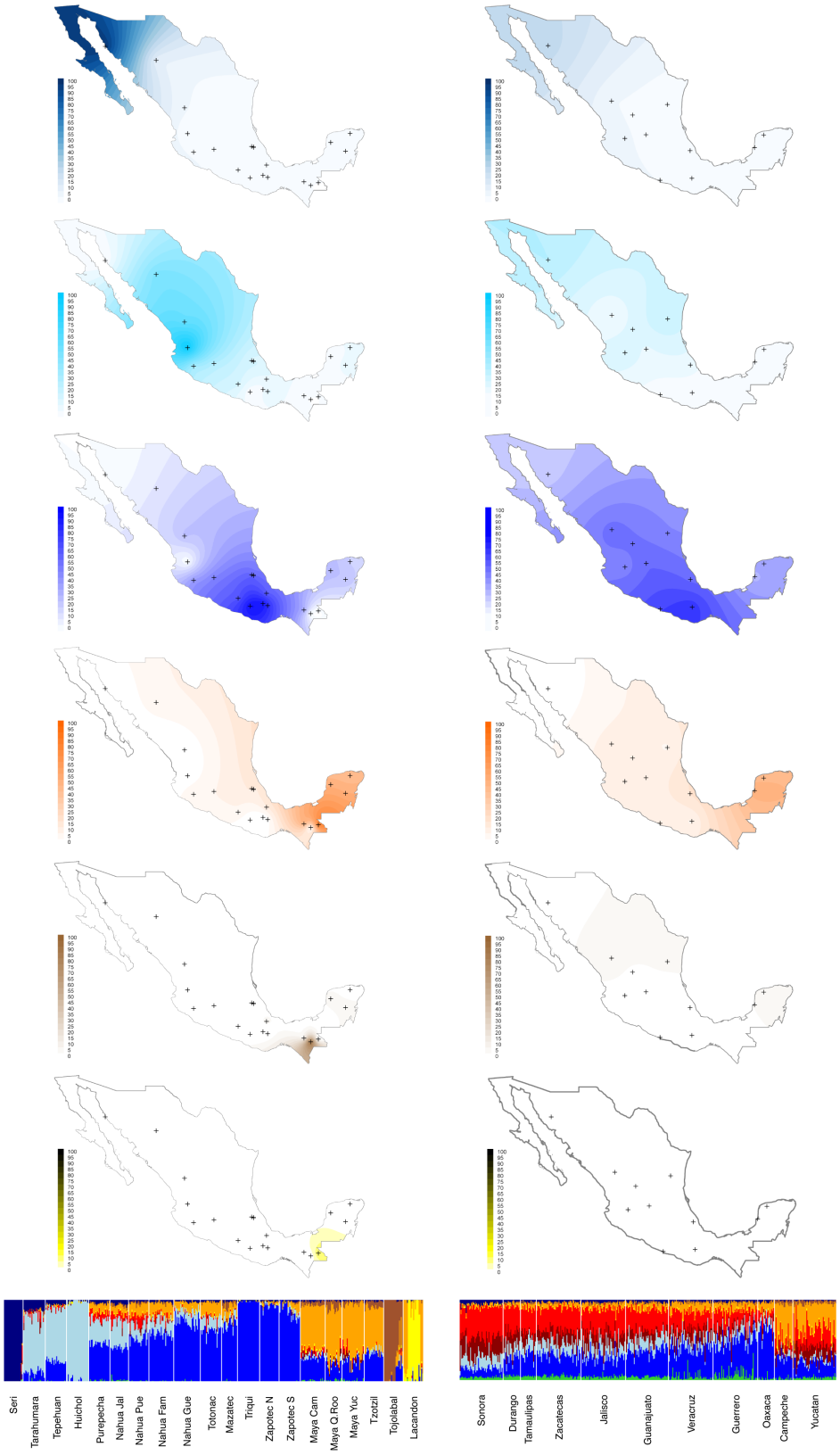

Figure S12: Spatial distribution of the major Native American components across Mexican populations. Interpolation maps are shown for ADMIXTURE values at *k*=9 observed among

indigenous (left column) and cosmopolitan (right column) samples. Black crosses on the maps of each column indicate sampling locations of indigenous and cosmopolitan populations, respectively. From top to bottom the six pairs of maps correspond to the six Native American components identified at k=9 (shown at the bottom and in Fig. S10). Contour maps were generated using Kriging interpolation methods, where intensities are proportional to ADMIXTURE values. For the group of cosmopolitan samples (thus with higher non-native admixture proportions), values were adjusted relative to the total Native American ancestry of each individual (see Methods for details).

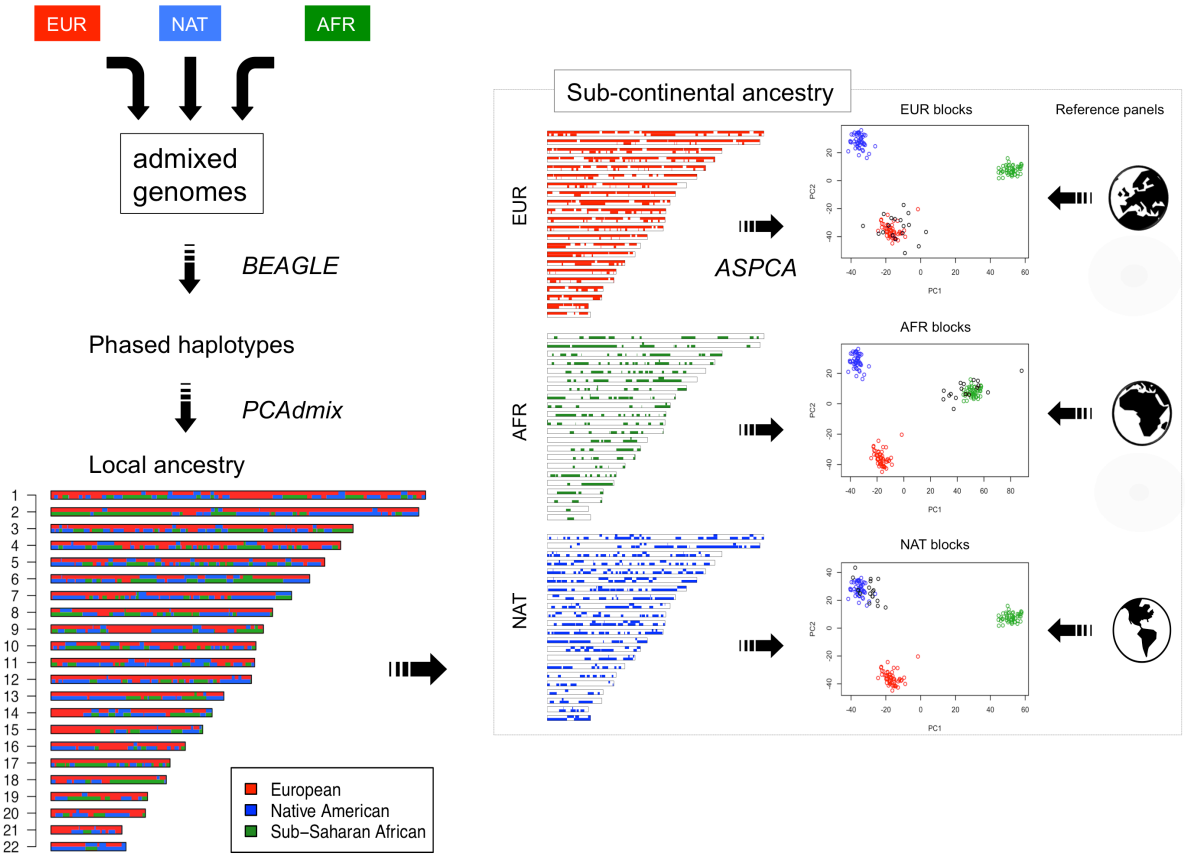

Figure S13: Diagram of the analytical strategy used for inferring sub-continental ancestry in admixed genomes (modified from (*19*)). The starting point consists of genome-wide SNP data from admixed Mexican individuals. Unrelated individuals and family trios are population phased and trio phased, respectively, using BEAGLE. Next, phased haplotypes are used to estimate local ancestry along the genome using PCAdmix and continental reference samples. Then, taking Viterbi calls at each locus, ancestry-specific regions of the genome are masked to separately analyze European, African, and Native American haplotypes in a PCA framework together with large sub-continental reference panels of putative ancestral populations (see Methods for details). We refer to this methodology as ancestry-specific PCA (ASPCA) and the code is packaged into the software PCAmask.

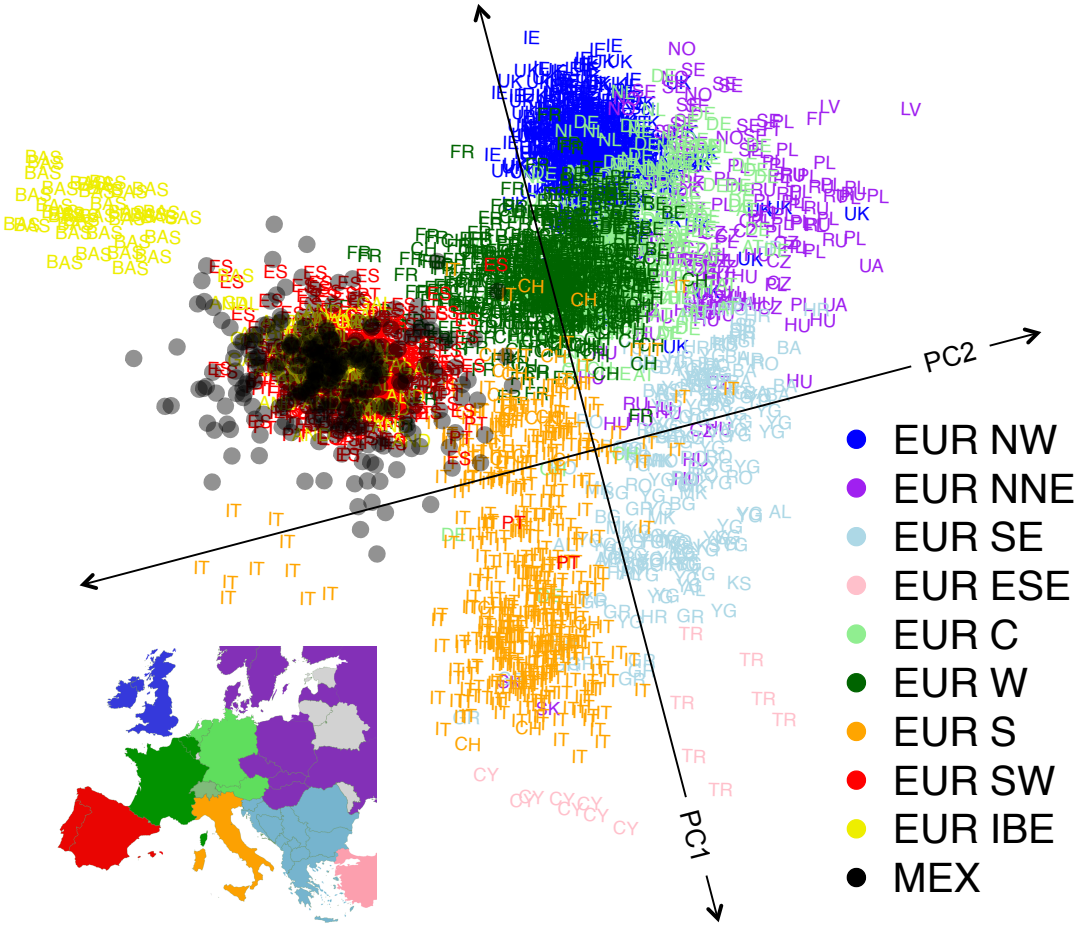

Figure S14: Sub-continental origin of European haplotypes derived from admixed Mexican genomes based on Ancestry-specific PCA (ASPCA). Plotted are European segments from cosmopolitan Mexican samples (black circles) together with our reference panel of 1,387 European individuals from POPRES (labeled by country code) plus 55 additional samples from Spain (yellow labels). Each black circle represents the combined set of Mexican haplotypes called European along the haploid genome of each sample with >25% of European ancestry. Axes were rotated 16 degrees counterclockwise to approximate the geographic orientation of population samples over Europe. Inset map shows POPRES countries of origin color-coded by region (areas not sampled in gray and Switzerland in intermediate shade of green to denote shared membership with EUR W, EUR C, and EUR S). Population codes and regions within Europe are detailed in Table S1.

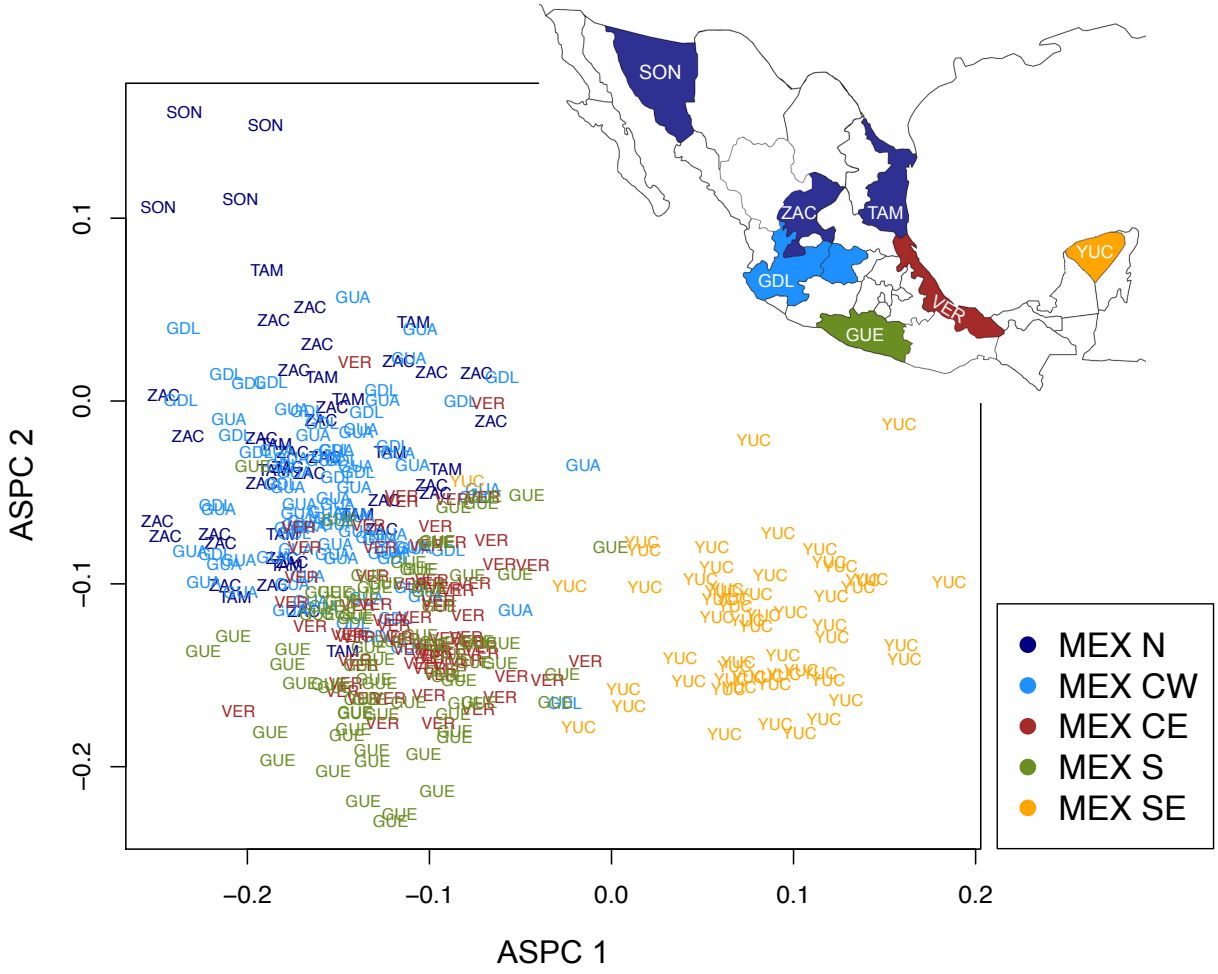

Figure S15: Sub-continental origin of Native American haplotypes derived from admixed Mexican genomes based on Ancestry-specific PCA (ASPCA). To provide detail on the Native American segments from cosmopolitan Mexican samples, the reference panel of Native American populations in Fig. 3A has been removed. All the samples have been analyzed together with the reference panel to define PCA space. Then only Mexican mestizo samples were plotted. Each data point represents the combined set of haplotypes of inferred Native American ancestry along the haploid genome of each mestizo sample with >25% of global Native American ancestry. Samples are represented by population labels and color-coded by region to ease contextualization with geography. Inset map shows Mexican states in which cosmopolitan samples were collected. Note that among the samples grouped as northern states (in dark blue: Sonora, Zacatecas, and Tamaulipas), the most extreme positive values along ASPC2 (the north-south axis) are indeed given by the Sonora individuals. See also Fig. 3B. GDL: POPRES Mexican individuals from Guadalajara, Jalisco. Population codes and regions within Mexico are detailed in Table S1.

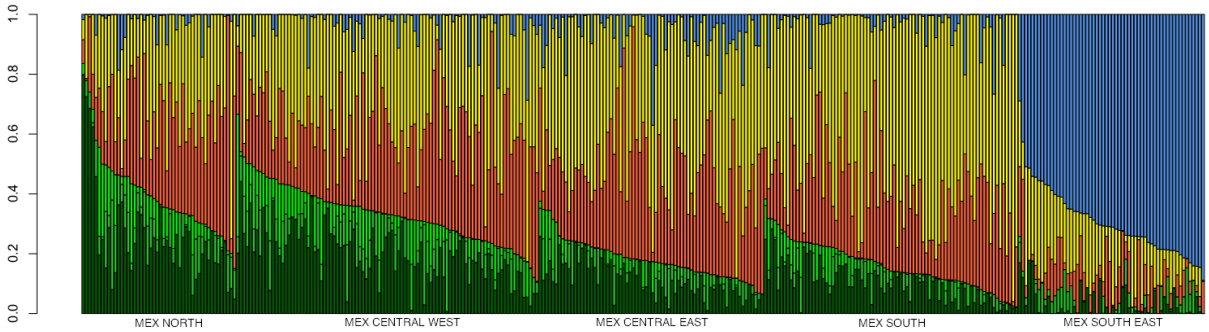

Figure S16: Supervised ancestry-specific clustering analysis of Native American haplotypes derived from admixed cosmopolitan Mexican genomes. On the x-axis bars represent haploid genomes for all admixed individuals with >25% of global Native American ancestry, that is, one individual is usually represented by two bars. The y-axis indicate native ancestry proportions at *k*=6 using our reference panel of Native Mexican populations (see Table S1). Given the low overall contribution of isolated native components into the mestizo population (as identified in Fig. 2), we excluded Seri, Lacandon, and Tojolabal from the reference panel. Since our ancestry-specific approach relies on haplotype data, we used a modified version of the FRAPPE algorithm to estimate admixture proportions in the presence of missing sites at SNPs inferred to be heterozygous for the desired ancestry (see Methods). Individuals are grouped into regions as described in Table S1. Because we required more than 25% of Native American ancestry to be included in the analysis, some regions are represented by less individuals that the actual sample size, such as mestizo individuals from Northern states of Mexico, where overall proportions of Native American ancestry are considerably lower than in the rest of the territory. The six clusters identified to run the algorithm on supervised mode were: Northern Native Mexicans, Huichol (which clustered on their own in previous analyses), Native Mexicans from Central West, Central East, South, and Southeast Mexico (excluding Seri, Lacandon, and Tojolabal). Overall, the results replicate the observations from our ASPCA analysis: on average, Mexicans sampled from different regions of Mexico derive differential ancestral contributions from each of the Native American components.

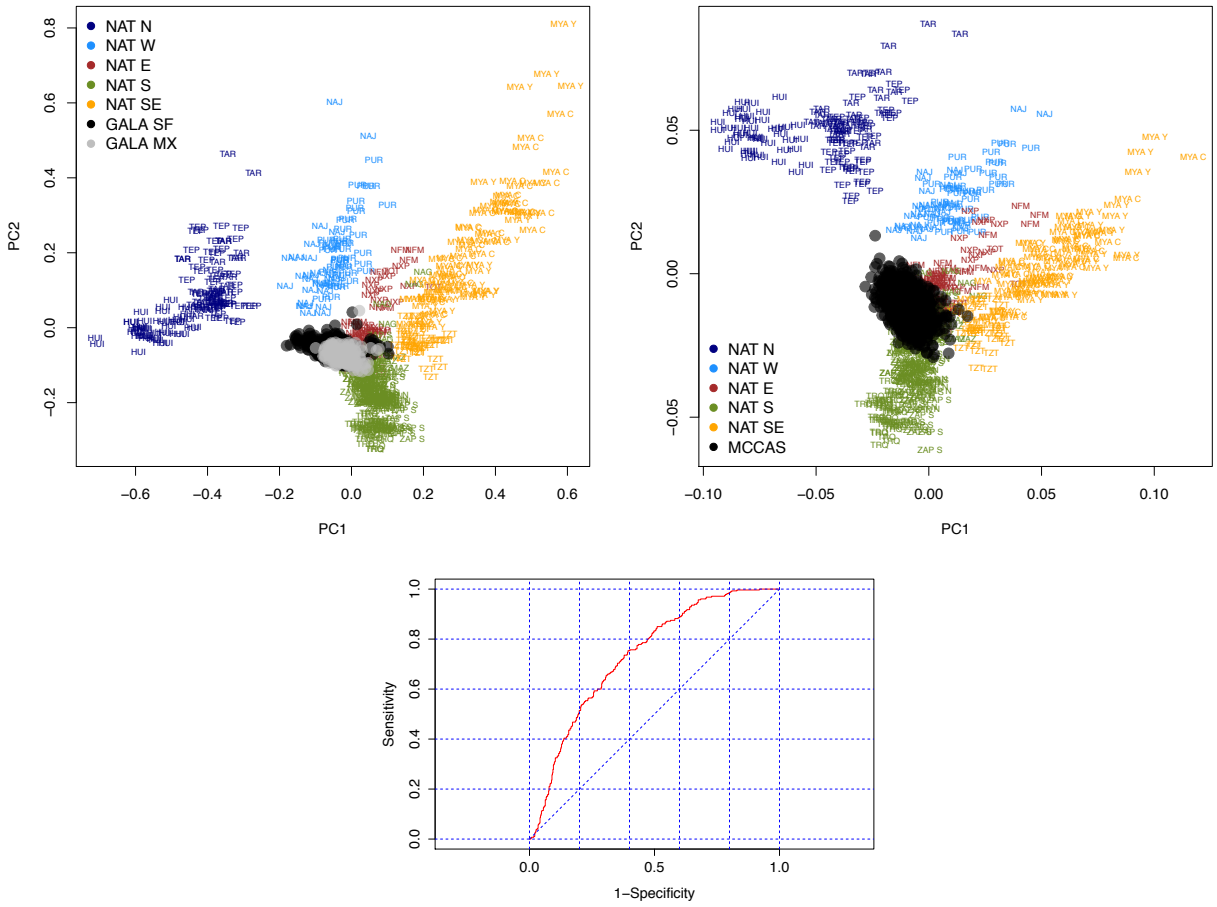

Figure S17: ASPCA analysis of Native American segments from Mexican participants of the GALA I study (*left*) sampled in Mexico City (GALA MX, gray circles) and the San Francisco bay area (GALA SF, black circles), and participants of the MCCAS study (*right*) sampled in Mexico City (black circles), analyzed together with our dataset of 20 indigenous Mexican populations (labeled by population identifier and color-coded by region of origin). Samples with >10% of non-native admixture were excluded from the reference panel as well as population outliers such as Seri, Lacandon, and Tojolabal. Here, a total of 803 phased haploid genomes (280 MX and 523 SF) represent the GALA Mexican sample and 1900 the MCCAS cohort. Bottom: ROC curve for the logistic regression of ASPCA values separating Mexico City (MX) versus San Francisco (SF) cases from the GALA I study (see main text for details).

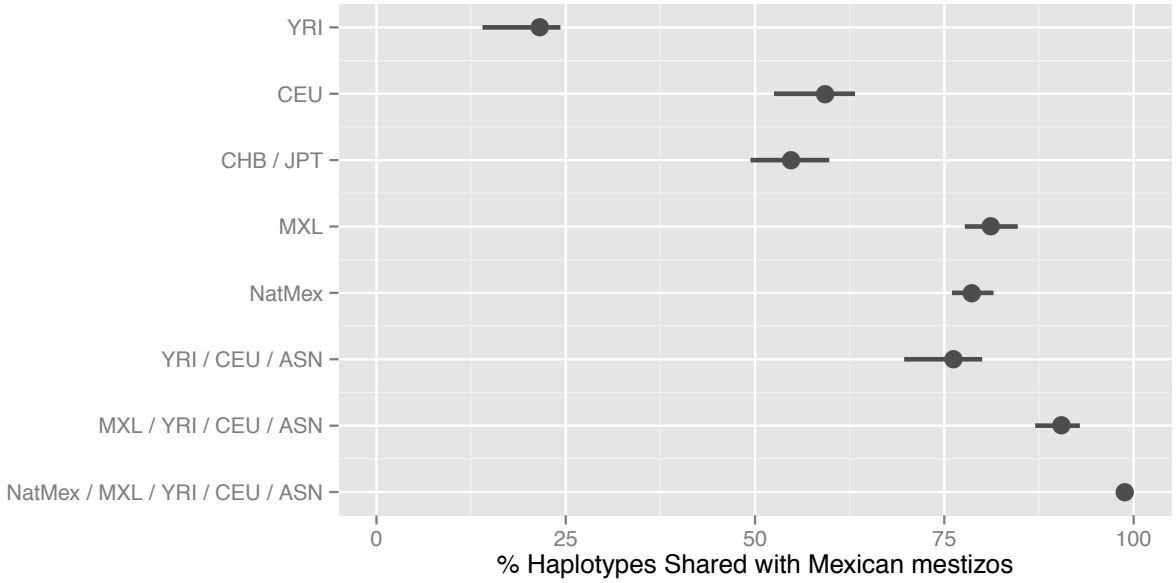

Figure S18: Global haplotype sharing analysis in autosomal chromosomes. Genome-wide proportions of haplotypes shared between a combined set of mestizo samples and different combinations of HapMap continental populations before and after including a combined set of Native American samples. Average haplotype sharing is given as a point with the range spanning the most-captured and least-captured chromosomes. Haplotype sharing analysis was performed using the subset of Mexican samples typed on both Affymetrix and Illumina platforms, hence with the largest intersection of genotyped SNPs (785,663) with HapMap3 populations, which included 312 Mexican mestizos from the Mexican Genome Diversity Project (MGDP) representing seven cosmopolitan populations (see Table S1). The "NatMex" panel here consists of 71 individuals from 3 indigenous groups typed on the same platforms, one from each of the major genetic components identified in Fig. 2: Tepehuano from Northern Mexico, Zapotec from Southern Mexico, and Maya in Campeche from the Yucatan peninsula (see Methods for details). Any of the continental source populations alone (YRI, CEU, NAT) shares a limited proportion of haplotypes with mestizo samples (21.6%, 59.3%, and 78.6%, respectively). HapMap Mexican-Americans (MXL) alone share 81.2% with MGDP mestizos and 90.5% when combined with all continental HapMap populations. After considering Native American samples in addition to this previous combination, nearly 100% of haplotypes in MGDP mestizos are captured.

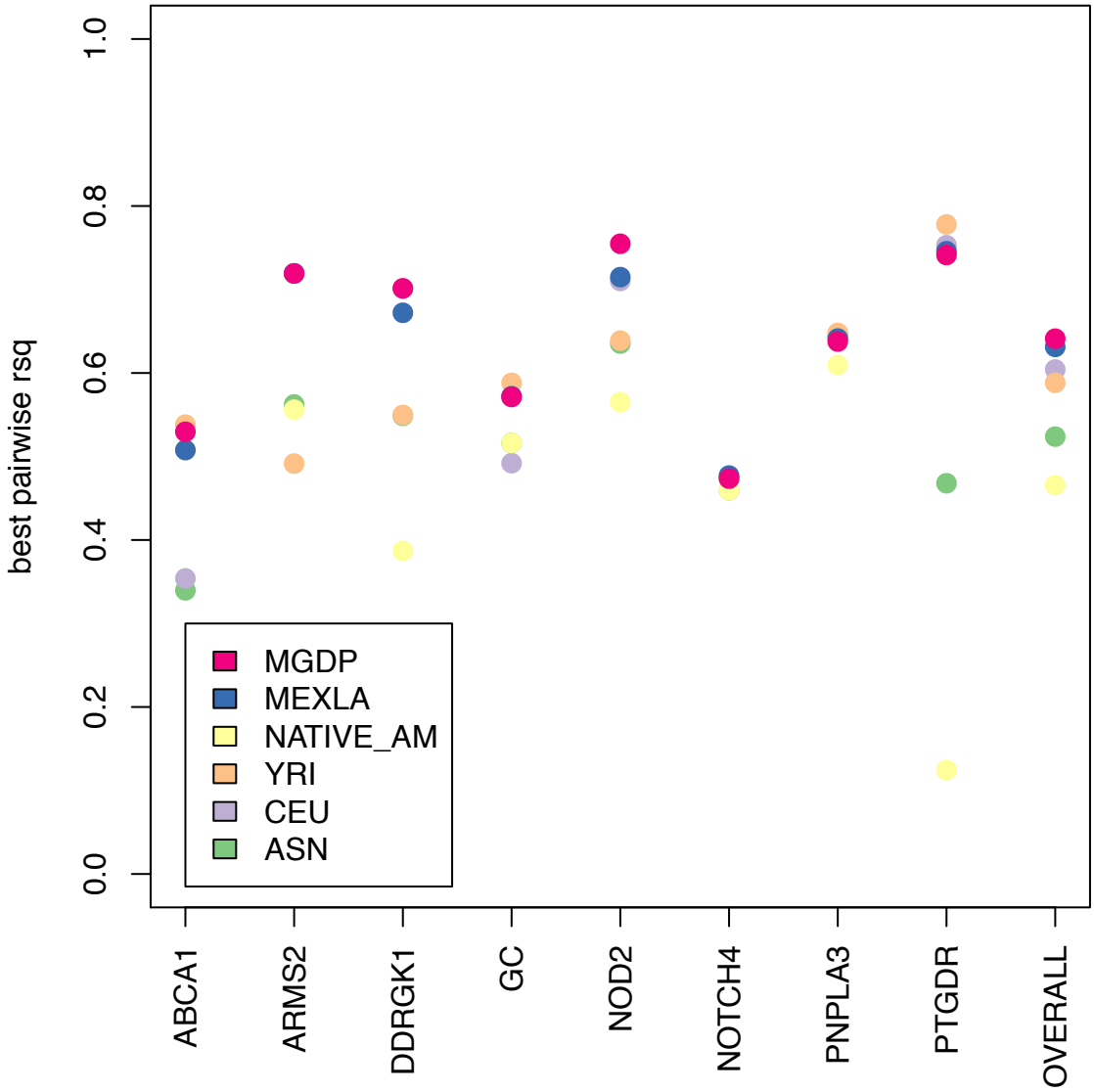

Figure S19: Tagging efficiency using Mexican Mestizos or HapMap Populations as reference. The mean best r^2 coverage based on the tag SNPs determined using various reference panels was evaluated in a subset of candidate gene regions of biomedical interest. While the individual results vary from gene to gene, using the full reference panel of Mexican Mestizos from the Mexican Genome Diversity Project (MGDP) resulted in the best tagging performance overall, notably, better than using the MXL population from HapMap3.

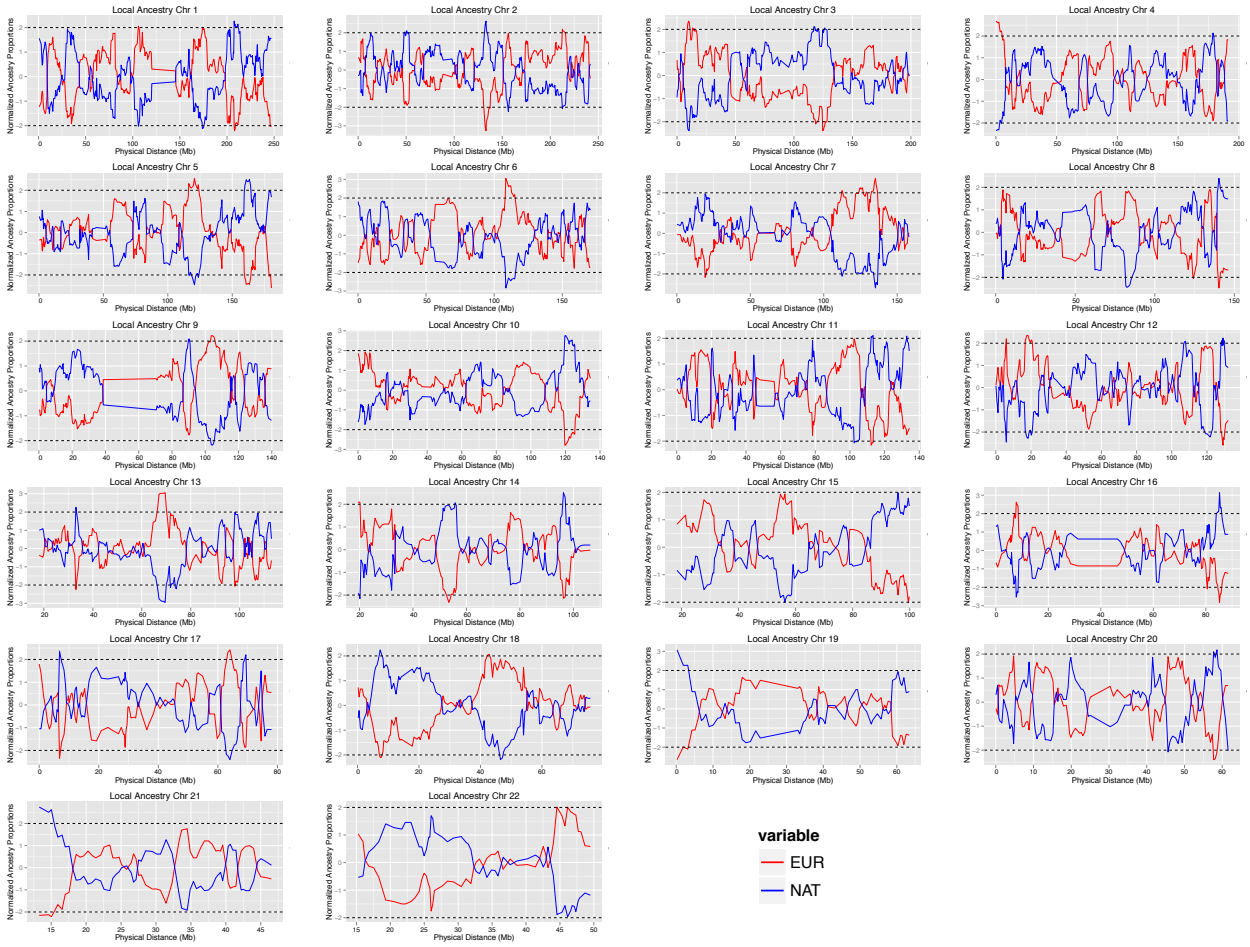

Figure S20: Local ancestry scan in the combined set of cosmopolitan Mexican samples showing normalized Z scores of Native American versus European ancestry proportions along autosomal chromosomes. African ancestry was not considered due to the small proportion of African haplotypes across individuals. Local ancestry calls were estimated using PCAdmix and counts were scaled to the total sample size. Dashed lines indicate two standard deviations away from the mean. Results are based on 372,692 SNPs and 362 samples with available Affymetrix data (see Table S1).

Table S1. Summary data for 32 Mexican populations and continental reference panels included in the study.

Population	Pop ID	N (initial)	Filtered	N (final)	Region*	Latitude	Longitude	Linguistic Family	Reference	Data
NATIVE MEXICANS										
Seri	SER	25	4	21	North Mexico	29.00	-112.15	Serian	*Present study (NMDP)*	Affymetrix 6.0
Tarahumara	TAR	25	1	24	North Mexico	27.75	-107.17	UtoAztecan	*Present study (NMDP)*	Affymetrix 6.0
Tepehuano	TEP	30	7	23	North Mexico	23.48	-104.39	UtoAztecan	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Huichol	HUI	24	0	24	North Mexico	21.17	-104.08	UtoAztecan	*Present study (NMDP)*	Affymetrix 6.0
Nahua (Jalisco)	NAJ	23	3	20	Central-West Mexico	19.50	-103.50	UtoAztecan	*Present study (NMDP)*	Affymetrix 6.0
Purepecha	PUR	23	0	23	Central-West Mexico	19.75	-101.50	Tarascan	*Present study (NMDP)*	Affymetrix 6.0
Totonac	TOT	25	2	23	Central-East Mexico	20.00	-97.80	Totonacan	*Present study (NMDP)*	Affymetrix 6.0
Nahua (Puebla)	NXP	25	3	22	Central-East Mexico	19.97	-97.62	UtoAztecan	*Present study (NMDP)*	Affymetrix 6.0
Nahua (Trios)	NFM	41	14	27	Central-East Mexico	19.93	-97.62	UtoAztecan	*Present study (NMDP)*	Affymetrix 6.0
Nahua (Guerrero)	NAG	29	0	29	South Mexico	17.89	-99.13	UtoAztecan	Mao *et al*. 2007	Affymetrix 500K
Triqui	TRQ	25	1	24	South Mexico	17.18	-97.95	Otomanguean	*Present study (NMDP)*	Affymetrix 6.0
Zapotec (North)	ZAP.N	21	0	21	South Mexico	17.41	-96.69	Otomanguean	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Zapotec (South)	ZAP.S	24	1	23	South Mexico	17.23	-96.23	Otomanguean	*Present study (NMDP)*	Affymetrix 6.0
Mazatec	MAZ	17	0	17	South Mexico	18.33	-96.33	Otomanguean	*Present study (NMDP)*	Affymetrix 6.0
Tzotzil	TZT	22	1	21	Southeast Mexico	16.83	-92.67	Mayan	*Present study (NMDP)*	Affymetrix 6.0
Tojolabal	TOJ	22	1	21	Southeast Mexico	16.50	-92.00	Mayan	*Present study (NMDP)*	Affymetrix 6.0
Lacandon	LAC	22	0	22	Southeast Mexico	16.75	-91.25	Mayan	*Present study (NMDP)*	Affymetrix 6.0
Maya (Quintana Roo)	MYA.Q	19	1	18	Southeast Mexico	19.58	-88.58	Mayan	*Present study (NMDP)*	Affymetrix 6.0
Maya (Campeche)	MYA.C	45	18	27	Southeast Mexico	20.37	-90.05	Mayan	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Maya (Yucatan)	MYA.Y	24	0	24	Southeast Mexico	21.17	-88.14	Mayan	Mao *et al*. 2007	Affymetrix 500K
TOTAL	20 groups	511	57	454						
COSMOPOLITAN MEXICANS										
Mexican-Americans	MXL	80	31	49	LA, California	34.08	-118.17	-	HapMap3 (*11*)	Affymetrix 6.0
Mexican from Sonora	SON	49	0	49	North Mexico	29.07	-110.94	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Durango	DUR	19	0	19	North Mexico	24.06	-104.66	-	*Present study (MGDP)*	Illumina 550K
Mexican from Tamaulipas	TAM	17	0	17	North Mexico	23.74	-99.14	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Zacatecas	ZAC	50	0	50	North Mexico	22.79	-102.59	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Jalisco	JAL	50	0	50	Central-West Mexico	20.67	-103.35	-	POPRES (*17*)	Affymetrix 500K
Mexican from Guanajuato	GUA	48	0	48	Central-West Mexico	21.01	-101.26	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Veracruz	VER	50	0	50	Central-East Mexico	19.57	-96.90	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Guerrero	GUE	50	0	50	South Mexico	16.88	-99.87	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
Mexican from Oaxaca	OAX	18	0	18	South Mexico	17.06	-96.72	-	*Present study (MGDP)*	Illumina 550K
Mexican from Campeche	CAM	20	0	20	Southeast Mexico	19.84	-90.53	-	*Present study (MGDP)*	Illumina 550K
Mexican from Yucatan	YUC	49	0	49	Southeast Mexico	20.98	-89.63	-	*Present study (MGDP)*	Affymetrix 500K/Illumina 550K
TOTAL	12 groups	500	31	469						
EUROPEANS										
European-Americans	CEU	25	0	25	Northern Europe	-	-	-	HapMap3 (*11*)	Affymetrix 6.0
Tuscan	TSI	25	0	25	Italy (Tuscany)	-	-	-	HapMap3 (*11*)	Affymetrix 6.0
Andalusian	AND	20	2	18	Spain (Andalusia)	-	-	-	Botigue *et al*. 2013	Affymetrix 6.0
Galician	GAL	17	0	17	Spain (Galicia)	-	-	-	Botigue *et al*. 2013	Affymetrix 6.0
Basque	BAS	20	0	20	Spain (Basque Country)	-	-	-	Henn *et al*. 2012	Affymetrix 6.0
Portuguese	PT	20	0	20	Europe SW	-	-	-	POPRES (*17*)	Affymetrix 500K
Spanish	ES	20	0	20	Europe SW	-	-	-	POPRES (*17*)	Affymetrix 500K
Italian	IT	20	0	20	Europe S	-	-	-	POPRES (*17*)	Affymetrix 500K
Greek	GR	8	0	8	Europe SE	-	-	-	POPRES (*17*)	Affymetrix 500K
French	FR	20	0	20	Europe W	-	-	-	POPRES (*17*)	Affymetrix 500K
Swiss	CH	20	0	20	Europe W	-	-	-	POPRES (*17*)	Affymetrix 500K
Belgian	BE	20	0	20	Europe W	-	-	-	POPRES (*17*)	Affymetrix 500K
German	DE	20	0	20	Europe C	-	-	-	POPRES (*17*)	Affymetrix 500K
British	GB	20	0	20	Europe NW	-	-	-	POPRES (*17*)	Affymetrix 500K
Irish	IE	20	0	20	Europe NW	-	-	-	POPRES (*17*)	Affymetrix 500K
Scandinavian**	SC	16	0	16	Europe NNE	-	-	-	POPRES (*17*)	Affymetrix 500K
TOTAL	16 groups	311	2	309						
AFRICANS										
Yoruba	YRI	50	0	50	West Africa	-	-	-	HapMap3 (*11*)	Affymetrix 6.0
TOTAL SUM	49 groups	1372	90	1282						

NMDP: Native Mexican Diversity Panel. *MGDP*: Mexican Genome Diversity Project. See supplementary text for details.

* Mexicans are grouped according to major geographic areas whereas Europeans from POPRES are according to the classification in Auton *et al*. 2009 (*41*).

** Includes individuals from Sweden, Norway, Denmark, and Finland (n=11, 3, 1, 1, respectively).

Table S2. Three working datasets generated for this study.

Name	Samples	SNPs	Average call rate	Notes
global.illu.affy.unrel	1,282	71,581	98.93%	All samples as reported in Table S1
global.affy.unrel	1,224	372,692	98.85%	All samples with available Affymetrix data
mex.hapmap.unrel	674	785,663	99.30%	Samples with both Affymetrix and Illumina data

The *unrel* suffix denotes that all individuals being part of the offspring of trios or duos have been removed.

Table S3. Population sample sizes filtered for individuals with <90% and <98% Native American ancestry.

Population	N (pre-filter)	Filter 1 (<90% NAT)	N (post-filter 1)	Filter 2 (<98% NAT)	N (post-filter 2)
SER	21	2	19	0	19
TAR	24	6	18	7	11
TEP	23	3	20	10	10
HUI	24	0	24	2	22
NAJ	20	8	12	10	2
PUR	23	8	15	15	0
TOT	23	3	20	5	15
NXP	22	15	7	5	2
NFM	27	5	22	14	8
NAG	29	1	28	4	24
TRQ	24	0	24	0	24
ZAP.N	21	0	21	0	21
ZAP.S	23	0	23	2	21
MAZ	17	0	17	6	11
TZT	21	0	21	2	19
TOJ	21	1	20	6	14
LAC	22	1	21	4	17
MYA.Q	18	4	14	14	0
MYA.C	27	13	14	12	2
MYA.Y	24	8	16	12	4
TOTAL	454	78	376	130	246

Table S4. Pairwise F_{ST} values among Native Mexican populations* (symmetric). Estimates are based on all autosomal SNPs in the Affymetrix genotype data, considering only unrelated individuals with >90% of Native American ancestry.

	SER	TAR	TEP	HUI	NAJ	PUR	TOT	NFM	NXP	NAG	TRQ	ZAP.N	ZAP.S	MAZ	MYA.C	MYA.Q	MYA.Y	TZT	TOJ	LAC
SER	-	0.087	0.086	0.097	0.086	0.087	0.095	0.090	0.090	0.085	0.100	0.097	0.093	0.092	0.092	0.099	0.096	0.096	0.121	0.136
TAR	0.087	-	0.032	0.044	0.033	0.034	0.041	0.038	0.030	0.034	0.049	0.045	0.041	0.038	0.036	0.045	0.042	0.043	0.068	0.083
TEP	0.086	0.032	-	0.031	0.022	0.023	0.031	0.028	0.020	0.022	0.038	0.035	0.031	0.028	0.026	0.034	0.032	0.033	0.057	0.071
HUI	0.097	0.044	0.031	-	0.033	0.034	0.041	0.038	0.031	0.033	0.049	0.046	0.042	0.040	0.037	0.045	0.042	0.042	0.068	0.081
NAJ	0.086	0.033	0.022	0.033	-	0.016	0.025	0.020	0.011	0.016	0.032	0.028	0.023	0.020	0.018	0.026	0.024	0.025	0.049	0.065
PUR	0.087	0.034	0.023	0.034	0.016	-	0.023	0.020	0.011	0.015	0.031	0.027	0.024	0.021	0.018	0.026	0.023	0.025	0.050	0.064
TOT	0.095	0.041	0.031	0.041	0.025	0.023	-	0.024	0.013	0.018	0.032	0.030	0.025	0.024	0.021	0.029	0.026	0.026	0.053	0.066
NFM	0.090	0.038	0.028	0.038	0.020	0.020	0.024	-	0.010	0.015	0.030	0.027	0.023	0.021	0.019	0.027	0.024	0.024	0.050	0.063
NXP	0.090	0.030	0.020	0.031	0.011	0.011	0.013	0.010	-	0.007	0.023	0.018	0.015	0.011	0.009	0.018	0.014	0.016	0.042	0.059
NAG	0.085	0.034	0.022	0.033	0.016	0.015	0.018	0.015	0.007	-	0.020	0.018	0.014	0.013	0.014	0.022	0.017	0.018	0.044	0.058
TRQ	0.100	0.049	0.038	0.049	0.032	0.031	0.032	0.030	0.023	0.020	-	0.031	0.026	0.026	0.029	0.037	0.034	0.034	0.058	0.072
ZAP.N	0.097	0.045	0.035	0.046	0.028	0.027	0.030	0.027	0.018	0.018	0.031	-	0.022	0.024	0.024	0.033	0.029	0.030	0.055	0.070
ZAP.S	0.093	0.041	0.031	0.042	0.023	0.024	0.025	0.023	0.015	0.014	0.026	0.022	-	0.019	0.021	0.029	0.025	0.025	0.050	0.064
MAZ	0.092	0.038	0.028	0.040	0.020	0.021	0.024	0.021	0.011	0.013	0.026	0.024	0.019	-	0.019	0.024	0.024	0.023	0.047	0.063
MYA.C	0.092	0.036	0.026	0.037	0.018	0.018	0.021	0.019	0.009	0.014	0.029	0.024	0.021	0.019	-	0.013	0.008	0.014	0.039	0.052
MYA.Q	0.099	0.045	0.034	0.045	0.026	0.026	0.029	0.027	0.018	0.022	0.037	0.033	0.029	0.024	0.013	-	0.017	0.022	0.045	0.059
MYA.Y	0.096	0.042	0.032	0.042	0.024	0.023	0.026	0.024	0.014	0.017	0.034	0.029	0.025	0.024	0.008	0.017	-	0.018	0.044	0.057
TZT	0.096	0.043	0.033	0.042	0.025	0.025	0.026	0.024	0.016	0.018	0.034	0.030	0.025	0.023	0.014	0.022	0.018	-	0.043	0.059
TOJ	0.121	0.068	0.057	0.068	0.049	0.050	0.053	0.050	0.042	0.044	0.058	0.055	0.050	0.047	0.039	0.045	0.044	0.043	-	0.083
LAC	0.136	0.083	0.071	0.081	0.065	0.064	0.066	0.063	0.059	0.058	0.072	0.070	0.064	0.063	0.052	0.059	0.057	0.059	0.083	-

*Population codes as in Table S1 and sorted geographically as in Fig. 1B.

Table S5. Average ADMIXTURE proportions for each cosmopolitan population sample at K=3 and K=9 ancestral clusters*

Population Sample**	K=3			K=9								
	% AFR	% EUR	% NAT	West African	Northern European	Southern European	Northern Native	Southern Native	Maya	Seri	Tojolabal	Lacandon
MXL	0.0418	0.4815	0.4767	0.0418	0.1555	0.3260	0.1346	0.2492	0.0488	0.0240	0.0121	0.0080
SON	0.0350	0.6126	0.3525	0.0350	0.2299	0.3827	0.1521	0.1097	0.0216	0.0498	0.0088	0.0105
DUR	0.0538	0.5054	0.4408	0.0538	0.1833	0.3220	0.1299	0.2256	0.0383	0.0251	0.0143	0.0076
TAM	0.0469	0.4277	0.5254	0.0469	0.1203	0.3074	0.1473	0.2726	0.0591	0.0278	0.0114	0.0073
ZAC	0.0491	0.4383	0.5126	0.0491	0.1426	0.2957	0.1480	0.2676	0.0538	0.0257	0.0110	0.0065
JAL	0.0423	0.4518	0.5060	0.0423	0.1518	0.2999	0.1475	0.2464	0.0574	0.0290	0.0144	0.0112
GUA	0.0405	0.3678	0.5917	0.0405	0.0964	0.2714	0.1508	0.3250	0.0654	0.0256	0.0159	0.0090
VER	0.0419	0.3348	0.6234	0.0419	0.1183	0.2164	0.1138	0.3603	0.1073	0.0200	0.0142	0.0078
GUE	0.0691	0.2520	0.6789	0.0691	0.0725	0.1796	0.1214	0.4202	0.0906	0.0217	0.0159	0.0090
OAX	0.0204	0.1337	0.8459	0.0204	0.0386	0.0951	0.1046	0.5977	0.1043	0.0217	0.0102	0.0074
CAM	0.0192	0.1828	0.7981	0.0192	0.0636	0.1191	0.0452	0.2623	0.4137	0.0089	0.0429	0.0251
YUC	0.0324	0.3713	0.5962	0.0324	0.1521	0.2192	0.0387	0.1907	0.3079	0.0087	0.0370	0.0132

*Ancestral clusters were identified based on population affiliation of parental samples showing the highest proportions of each K across the global dataset.
**Population sample sizes and locations of cosmopolitan samples are summarized in Table S1.

Table S6: Summaries of multiple regression associations between FEV_1 and various components of ancestry in the GALA I and MCCAS studies. Significant findings in the meta-analysis are in bold, with significant heterogeneity indicated with an asterisk. Parameters for European ancestry were estimated from an additional regression with European ancestry instead of Native American ancestry, given the high collinearity between the two.

GALA I	**β**	**SE**	**t**
(Intercept)	90.6093	9.9767	9.0821
z(ASPC1)	-2.3741	1.2869	-1.8449
z(ASPC2)	0.6537	1.3028	0.5018
African	164.1133	53.6408	3.0595
Native American	-0.9299	11.0486	-0.0842
European	0.9299	11.0486	0.0842
MCCAS	**β**	**SE**	**t**
(Intercept)	75.2673	8.6412	8.7103
z(ASPC1)	-2.1278	0.9816	-2.1678
z(ASPC2)	-0.4693	0.9797	-0.479
African	-5.9006	42.9404	-0.1374
Native American	11.9151	8.4243	1.4144
European	-11.9152	8.4243	-1.4144
Meta-analysis	**β**	**SE**	**p-value**
z(ASPC1)	**-2.2184**	**0.7805**	**0.0045†**
z(ASPC2)	0.0127	0.7478	0.9864
African	79.0749	85.0069	0.3523*
Native American	7.1928	6.6991	0.283
European	-7.1929	6.6991	0.283

†β and p-values were quite similar in a joint OLS model, adjusting in addition for study of origin.

References and Notes

1. S. Gravel *et al.*, Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11983 (Jul 19, 2011).
2. S. Wang *et al.*, Genetic variation and population structure in native Americans. *PLoS Genet* **3**, e185 (Nov, 2007).
3. V. Acuna-Alonzo *et al.*, A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. *Hum Mol Genet* **19**, 2877 (Jul 15, 2010).
4. A. L. Williams *et al.*, Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, (Dec 25, 2013).
5. R. Lisker, E. Ramirez, V. Babinsky, Genetic structure of autochthonous populations of Meso-America: Mexico. *Hum Biol* **68**, 395 (Jun, 1996).
6. K. Sandoval *et al.*, Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas. *American journal of physical anthropology* **148**, 395 (Jul, 2012).
7. A. Gorostiza *et al.*, Reconstructing the history of Mesoamerican populations through the study of the mitochondrial DNA control region. *PLoS ONE* **7**, e44666 (2012).
8. D. Reich *et al.*, Reconstructing Native American population history. *Nature* **488**, 370 (Jul 11, 2012).
9. See supplementary materials on *Science* Online.
10. J. Novembre *et al.*, Genes mirror geography within Europe. *Nature* **456**, 98 (Nov 06, 2008).
11. D. M. Altshuler *et al.*, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (Sep 2, 2010).
12. B. Henn, L. Hon, J. Macpherson, N. Eriksson, Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS ONE*, (2012).
13. J. Hey, On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**, e193 (Jul 01, 2005).
14. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).
15. A. Pascual Soto, *El Tajín. En busca de los orígenes de una civilización.*, (UNAM-INAH, Mexico, 2006).
16. L. Campbell, T. Kaufman, Mayan Linguistics: Where Are We Now? *Annual Review of Anthropology* **14**, 187 (1985).
17. M. R. Nelson *et al.*, The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* **83**, 347 (Sep 01, 2008).
18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655 (Sep 01, 2009).
19. A. Moreno-Estrada *et al.*, Reconstructing the population genetic history of the Caribbean. *PLoS genetics* **9**, e1003925 (Nov, 2013).
20. A. Brisbin *et al.*, PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol* **84**, 343 (Aug, 2012).
21. N. A. Johnson *et al.*, Ancestral components of admixed genomes in a mexican cohort. *PLoS Genet* **7**, e1002410 (Dec, 2011).
22. L. R. Botigue *et al.*, Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A* **110**, 11791 (Jul 16, 2013).

23. S. Wang *et al.*, Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* **4**, e1000037 (Apr 01, 2008).
24. I. Silva-Zolezzi *et al.*, Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, (Jun 11, 2009).
25. M. A. Nalls *et al.*, Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am J Hum Genet* **82**, 81 (Jan, 2008).
26. C. A. Peralta *et al.*, The Association of African Ancestry and elevated creatinine in the Coronary Artery Risk Development in Young Adults (CARDIA) Study. *Am J Nephrol* **31**, 202 (2010).
27. L. Fejerman *et al.*, Genetic ancestry and risk of breast cancer among U.S. Latinas. *Cancer Research* **68**, 9723 (Dec 01, 2008).
28. J. L. Hankinson, J. R. Odencrantz, K. B. Fedan, Spirometric reference values from a sample of the general U.S. population. *American Journal of Respiratory and Critical Care Medicine* **159**, 179 (Jan, 1999).
29. R. Kumar *et al.*, Genetic ancestry in lung-function predictions. *New England Journal of Medicine* **363**, 321 (Jul 22, 2010).
30. K. Salari *et al.*, Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genet Epidemiol* **29**, 76 (Jul, 2005).
31. D. B. Hancock *et al.*, Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children. *PLoS genetics* **5**, e1000623 (Aug 01, 2009).
32. D. G. Torgerson *et al.*, Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J Allergy Clin Immunol*, (May 12, 2012).
33. X. Mao *et al.*, A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* **80**, 1171 (Jun, 2007).
34. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904 (Aug 01, 2006).
35. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358 (1984).
36. B. S. Weir, W. G. Hill, Estimating F-statistics. *Annual Review of Genetics* **36**, 721 (2002).
37. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Use R (Springer New York, New York, NY, 2009), pp. VIII, 213.
38. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559 (Sep 01, 2007).
39. M. A. Nalls *et al.*, Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS genetics* **5**, e1000415 (Apr 01, 2009).
40. M. Jobin, J. Mountain, REJECTOR: Software for Population History Inference from Genetic Data via a Rejection Algorithm. *Bioinformatics (Oxford, England)*, (Oct 20, 2008).
41. A. Auton *et al.*, Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res*, 1 (Mar 13, 2009).
42. B. M. Henn *et al.*, Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences* **108**, 5154 (Apr 29, 2011).
43. G. K. Chen, P. Marjoram, J. D. Wall, Fast and flexible simulation of DNA sequence data. *Genome research* **19**, 136 (Jan, 2009).
44. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**, 1084 (Nov 01, 2007).
45. S. R. Browning, Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics* **124**, 439 (Dec 01, 2008).
46. A. Gusev *et al.*, Whole population, genome-wide mapping of hidden relatedness. *Genome research* **19**, 318 (Mar 01, 2009).

47. B. M. Henn *et al.*, Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397 (Jan, 2012).
48. J. M. Kidd *et al.*, Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am J Hum Genet* **91**, 660 (Oct 5, 2012).
49. T. Raiko, A. Ilin, J. Karhunen, Principal component analysis for large scale problems with lots of missing values. *Machine Learning: ECML 2007*, 691 (2007).
50. H. Tang, J. Peng, P. Wang, N. J. Risch, Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology* **28**, 289 (Jun, 2005).
51. D. G. Torgerson *et al.*, Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nature Genetics* **43**, 887 (Sep, 2011).
52. E. G. Burchard *et al.*, Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* **169**, 386 (Feb 1, 2004).
53. J. M. Galanter *et al.*, Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 Latino populations. *The Journal of allergy and clinical immunology* **128**, 37 (Jul, 2011).
54. A. Bigham *et al.*, Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLoS genetics* **6**, e1001116 (Sep 09, 2010).
55. H. Wu *et al.*, Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol* **125**, 321 (Feb, 2010).
56. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 (Mar 22, 2008).
57. G. McVean, A genealogical interpretation of principal components analysis. *PLoS genetics* **5**, e1000686 (Oct, 2009).
58. G. Aguirre Beltran, The Slave Trade in Mexico. *The Hispanic American Historical Review* **24**, 412 (1944).
59. O. Lao *et al.*, Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**, 1241 (Aug 26, 2008).
60. M. Stephens, P. Scheet, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics* **76**, 449 (Apr 01, 2005).
61. M. Stephens, N. J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data. *American journal of human genetics* **68**, 978 (May 01, 2001).
62. M. T. Villarreal-Molina *et al.*, Association of the ATP-binding cassette transporter A1 R230C variant with early-onset type 2 diabetes in a Mexican population. *Diabetes* **57**, 509 (Feb, 2008).
63. S. Romeo *et al.*, Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature Genetics* **40**, 1461 (Dec, 2008).
64. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263 (Jan 15, 2005).