Genome-wide patterns of population structure and admixture among Hispanic/Latino populations

Katarzyna Bryc[a,1], Christopher Velez[b,1], Tatiana Karafet[c], Andres Moreno-Estrada[a,d], Andy Reynolds[a], Adam Auton[a,2], Michael Hammer[c], Carlos D. Bustamante[a,d,3,4], and Harry Ostrer[b,3,4]

[a]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14850; [b]Human Genetics Program, Department of Pediatrics, New York University School of Medicine, New York, NY 10016; [c]Arizona Research Laboratories Division of Biotechnology and Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; and [d]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

Hispanic/Latino populations possess a complex genetic structure that reflects recent admixture among and potentially ancient substructure within Native American, European, and West African source populations. Here, we quantify genome-wide patterns of SNP and haplotype variation among 100 individuals with ancestry from Ecuador, Colombia, Puerto Rico, and the Dominican Republic genotyped on the Illumina 610-Quad arrays and 112 Mexicans genotyped on Affymetrix 500K platform. Intersecting these data with previously collected high-density SNP data from 4,305 individuals, we use principal component analysis and clustering methods *FRAPPE* and *STRUCTURE* to investigate genome-wide patterns of African, European, and Native American population structure within and among Hispanic/Latino populations. Comparing autosomal, X and Y chromosome, and mtDNA variation, we find evidence of a significant sex bias in admixture proportions consistent with disproportionate contribution of European male and Native American female ancestry to present-day populations. We also find that patterns of linkage-disequilibria in admixed Hispanic/Latino populations are largely affected by the admixture dynamics of the populations, with faster decay of LD in populations of higher African ancestry. Finally, using the locus-specific ancestry inference method *LAMP*, we reconstruct fine-scale chromosomal patterns of admixture. We document moderate power to differentiate among potential subcontinental source populations within the Native American, European, and African segments of the admixed Hispanic/Latino genomes. Our results suggest future genome-wide association scans in Hispanic/Latino populations may require correction for local genomic ancestry at a subcontinental scale when associating differences in the genome with disease risk, progression, and drug efficacy, as well as for admixture mapping.

human genomics | population genetics | Hispanic/Latino

The term “Hispanic/Latinos” refers to the ethnically diverse inhabitants of Latin America and to people of Latin American descent throughout the world. Present-day Hispanic/Latino populations exhibit complex population structure, with significant genetic contributions from Native American and European populations (primarily involving local indigenous populations and migrants from the Iberian peninsula and Southern Europe) as well as West Africans brought to the Americas through the trans-Atlantic slave trade (1, 2). These complex historical events have affected patterns of genetic and genomic variation within and among present-day Hispanic/Latino populations in a heterogeneous fashion, resulting in rich and varied ancestry within and among populations as well as marked differences in the contribution of European, Native American, and African ancestry to autosomal, X chromosome, and uniparentally inherited genomes.

Many key demographic variables differed among colonial Latin American populations, including the population size of the local pre-Columbian Native American population, the extent and rate at which European settlers displaced native populations, whether or not slavery was introduced in a given region, and, if so, the size and timing of introduction of the African slave populations. There were also strong differences in ancestry among social classes in colonial (and postcolonial) populations with European ancestry often correlating with higher social standing. As a consequence, present-day Hispanic/Latino populations exhibit very large variation in ancestry proportions (as estimated from genetic data) not only across geographic regions (1, 2), but also within countries themselves (3, 4). In addition, the process of admixture was apparently sex-biased and preferentially occurred between European males and Amerindian and/or African females; this process has been shown to be remarkably consistent among countries and populations including Argentina (5), Ecuador (6), Mexico (7), Cuba (8), Brazil (9), Uruguay (10), Colombia (11), and Costa Rica (11).

The rich diversity of variation in ancestry among Hispanic/Latino populations, coupled with consistent differences among populations in the incidence of chronic heritable diseases, suggests that Hispanic/Latino populations may be very well suited for admixture mapping (12, 13). For example, differences in relative European ancestry proportions correlate with higher susceptibility in Puerto Ricans to asthma as compared with Mexicans (14). Data have also shown an increased risk of breast cancer in Latinas with greater European ancestry (15) and an interplay between African ancestry and cardiovascular disease and hypertension in Puerto Ricans from Boston (16). Hispanic/Latinos are also likely to play an increasingly important role in multi- and transethnic genetic studies of complex disease. Genome-wide scans have identified candidate markers for onset of type 2 diabetes in Mexican-Americans from Texas (17) as well as a region on chromosome 5 associated with asthma in Puerto Ricans (18).

Quantifying the relative contributions of ancestry, environment (including socio-economic status), and ancestry by environment interaction to disease outcome in diverse Hispanic/Latino populations will also be critical to applying a genomic perspective to the practice of medicine in the United States and in Latin America. For example, whereas European ancestry was associated with increased asthma susceptibility in Puerto Ricans

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution IV: The Human Condition," held December 10–12, 2009, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and audio files of most presentations are available on the NAS Web site at www.nasonline.org/SACKLER_Human_Condition.

Author contributions: C.D.B. and H.O. designed research; K.B., C.V., T.K., and M.H. performed research; K.B., A.A., M.H., C.D.B., and H.O. contributed new reagents/analytic tools; K.B., T.K., A.M.-E., A.R., A.A., and C.D.B. analyzed data; and K.B., C.V., A.M.-E., C.D.B., and H.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Genotype data from 100 Hispanic/Latinos have been deposited in the Gene Expression Omnibus (GEO) series record GSE21248.

[1]K.B. and C.V. contributed equally to this work.

[2]Present address: Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK.

[3]C.D.B. and H.O. contributed equally to this work.

[4]To whom correspondence may be addressed. E-mail: cdbustam@stanford.edu or harry.ostrer@nyumc.org.

This article contains supporting information online at www.pnas.org/cgi/content/full/0914618107/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0914618107

(14), it was also shown that the effect was moderated by socioeconomic status (19). This suggests that quantifying fine-scale patterns of genomic diversity among diverse U.S. and non-U.S. Hispanic/Latino may be critical to the efficient and effective design of medical and population genomic studies. A fine-scale population genomics perspective may also provide a powerful means for understanding the roles of ancestry, genetics, and environmental covariates on disease onset and severity (13).

Here, we introduce a larger, high-density SNP and haplotype dataset to investigate historical population genetics questions—such as variation in sex-biased ancestry and genome-wide admixture proportions within and among Latino populations—as well as provide a genomic resource for the study of population substructure within putative European, African, and Native American source populations. Our dataset includes three Latino populations that are underrepresented in whole-genome analyses, namely, Dominicans, Colombians, and Ecuadorians, as well as Mexicans and Puerto Ricans, the two largest Hispanic/Latino ethnic groups in the United States. This allows comparison of patterns of population structure and ancestry across multiple U.S. Hispanic/Latino populations. Our dense SNP marker panel is formed by the intersection of two of the most commonly used genotyping platforms, allowing for the inclusion of dozens of Native American, African, and European populations for ancestry inference. Our work expands on high-density population-wide genotype data from the International HapMap Project (HapMap) (20, 21), the Human Genome Diversity Panel (HGDP) (22), and the Population Reference Sample (POPRES) (23) that have representation of Mexicans but not other Hispanic/Latino groups either from the Caribbean or from South America, with a resulting gap for analyzing admixture in those populations. This project, therefore, represents an important step toward comprehensive panels for US-based studies that can more accurately reflect the diversity within various Hispanic/Latino populations.

Results

Population Structure. We applied the clustering algorithm *FRAPPE* to investigate genetic structure among Hispanic/Latino individuals using a merged data set with over 5,000 individuals with European, African, and Native American ancestry genotyped across 73,901 SNPs common to the Affymetrix 500K array and the Illumina 610-Quad panel (*Materials and Methods*). *FRAPPE* implements a maximum likelihood method to infer the genetic ancestry of each individual, where the individuals are assumed to have originated from K ancestral clusters (24). The plots for $K = 3$ and $K = 7$ are shown in Fig. 1 and for all other values of K in Fig. S1 $K = 3$. We observed clustering largely by Native American, African, and European ancestry, with the Hispanic/Latino populations showing genetic similarity with all of these populations. However, significant population differences exist, with the Dominicans and Puerto Ricans showing the highest levels of African ancestry (41.8% and 23.6% African, SDs 16% and 12%), whereas Mexicans and Ecuadorians show the lowest levels of African ancestry (5.6% and 7.3% African, SDs 2% and 5%) and the highest Native American ancestries (50.1% and 38.8% Native American, SDs 13% and 10%). We also found extensive variation in European, Native American and African ancestry among individuals within each population. A clear example could be observed in the Mexican sample, in which ancestry proportions ranged from predominantly Native American to predominantly European (with generally low levels of African ancestry). Similar results were found in Colombians and Ecuadorians, whereas Dominicans and Puerto Ricans showed the greatest variation in the African ancestry (Fig. 1). Interestingly, at K = 7, we were able to capture signals of continental substructure such as a Southwest to Northeast gradient in Europe and a Native American component that is absent in the two Amazonian indigenous populations (Karitiana and Surui) but that substantially contributes to all other studied Latino populations. We also note that several of the individuals from the Maya and Quechua Native American samples (and to a lesser extent Nahua and Pima) from the Human Genome Diversity Panel (CEPH-HGDP) show moderate levels of European admixture, consistent with previous studies of these populations (25). Interestingly, this is not the case for the Aymara and Quechua samples genotyped by Mao et al. (26).

We also undertook principal component analysis (PCA) of the autosomal genotype data from Hispanic/Latino and putative ancestral populations using the *smartpca* program from the software package *eigenstrat* (Fig. 2*A*) (27). The first two principal components of the PCA strongly support the notion that the three ancestral populations contributing to the Hispanic/Latino genomic diversity correspond exactly to Native American, European, and African ancestry. The Hispanic/Latino populations showed different profiles of ancestry, as exemplified by the fitting of ellipses to the covariance matrix of each population's first two PCs (Fig. 2*C*). Subsequent PCs showed substructure within Africa, Native Americans, and Europeans (Fig. S2). PCA on the X chromosome markers (Fig. 2*B*) showed a similar pattern, although because there are only 1,500 markers, this PCA had greater variance, which is illustrated in the fitted ellipses as well (Fig. 2*D*).

We also ran the Bayesian clustering algorithm *STRUCTURE* in "assignment mode" (28), and used a training set of Europeans, Africans, and Native Americans to estimate ancestral allele frequencies and assess admixture proportions within and among the Hispanic/Latino populations. Using *STRUCTURE* analysis of the autosomes (Fig. 3, *Upper*) and the X chromosome (Fig. 3, *Lower*), we found that, again, Puerto Ricans and Dominicans showed the greatest proportion of African ancestry whereas Colombians, Ecuadorians, and Mexicans showed extensive variation in European and Native American ancestry among individuals. We calculated LD decay curves for all populations with at least 10 individuals, choosing subsets of 10 individuals, and averaging more than 100 random subsets of the data. Patterns of decay of LD were consistent with previously published results (25), with Native American populations showing the highest levels of LD and African populations the lowest (Fig. 4*A*). Interestingly, the Hispanic/Latino populations demonstrated rates of decay of LD that correlated strongly with the amount of Native American, European, and African ancestry (Fig. 4*B*). Specifically, the populations with the most Native American ancestry, Mexican and Ecuadorian, exhibited higher levels of linkage disequilibrium among SNP markers, whereas the populations with the highest proportions of African ancestry, the Dominican and Puerto Rican samples, had the lowest levels of LD.

Locus-Specific Ancestry. To reconstruct local genomic ancestry at a fine scale, we used the ancestry deconvolution algorithm LAMP (29), allowing for a three-way admixture and focused on the four Hispanic/Latino populations genotyped on the Illumina 610-Quad platform—Dominicans, Colombians, Puerto Ricans, and Ecuadorians (*Materials and Methods*). Because this same SNP panel had also been genotyped across the HGDP samples (1,043 individuals from 53 populations), the merged data set containing more than 500,000 markers provided a unique resource for investigating the extent of subcontinental ancestry among diverse Hispanic/Latino populations. We found that individual average ancestries are in agreement with *FRAPPE* and *STRUCTURE* results in which Ecuadorians have the highest Native American proportions, followed by Colombians (showing greater European contribution), and with Puerto Ricans and Dominicans showing the highest African ancestry—specially Dominicans, who show very low contribution from Native Americans (Fig. 1). We also used the PCA-based methods of Bryc et al. (30) to infer ancestry at each locus for the samples genotyped on the Affymetrix 500K, which included more than 100 Mexican samples genotyped by

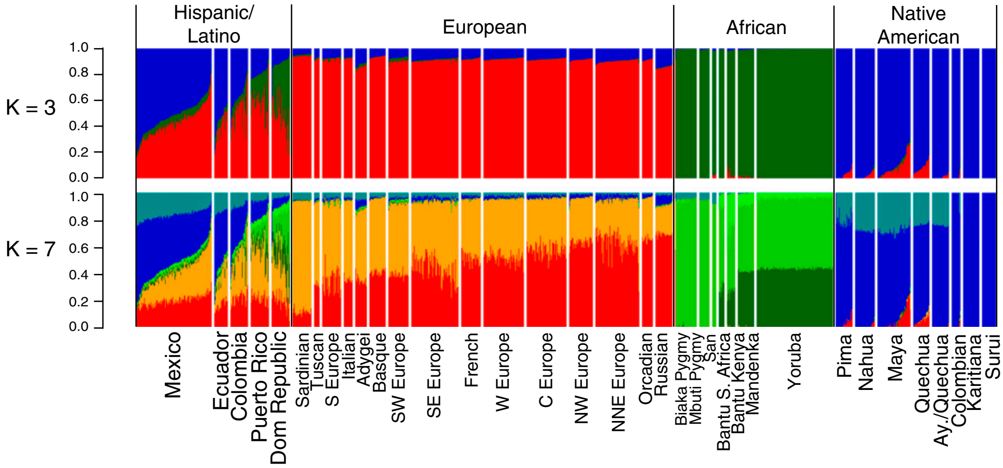

Fig. 1. *Frappe* clustering illustrating the admixed ancestry of Hispanic/Latinos shown for $K = 3$ and $K = 7$. Individuals are shown as vertical bars colored in proportion to their estimated ancestry within each cluster. Native American populations are listed in order geographically, from North to South.

the POPRES project (23) and diverse Native American populations genotyped by Mao et al. (26). The local admixture tracks for each individual are in large agreement with the genome-wide average ancestry proportions (Fig. 3, *Middle*).

To investigate the genetic relationships among admixed Hispanic/Latino populations and putative ancestral groups, we compared patterns of population divergence among the inferred segments of European, African, and Native American ancestry and corresponding putative source populations using Wright's F_{ST} measure. Specifically, we used LAMP to reconstruct for each individual in our data set, segments of European, African, and Native American ancestry across both the maximal SNP data set

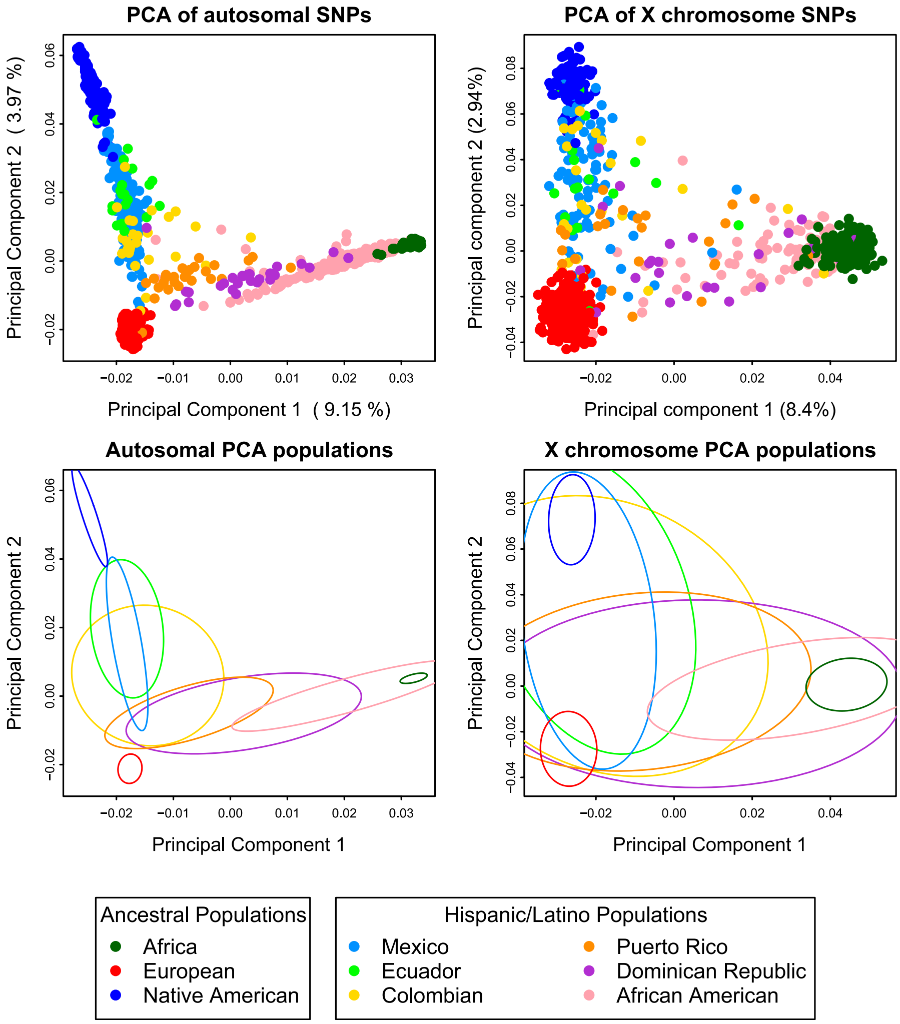

Fig. 2. Principal component analysis results of the Hispanic/Latino individuals with Europeans, Africans, and Native Americans. PC 1 vs. PC 2 scatter plots based on autosomal markers (*Upper Left*) and based on X chromosome markers (*Upper Right*). Ellipses are fitted to the PCA results on the autosomes (*Lower Left*) and to results from the X chromosome markers (*Lower Right*).

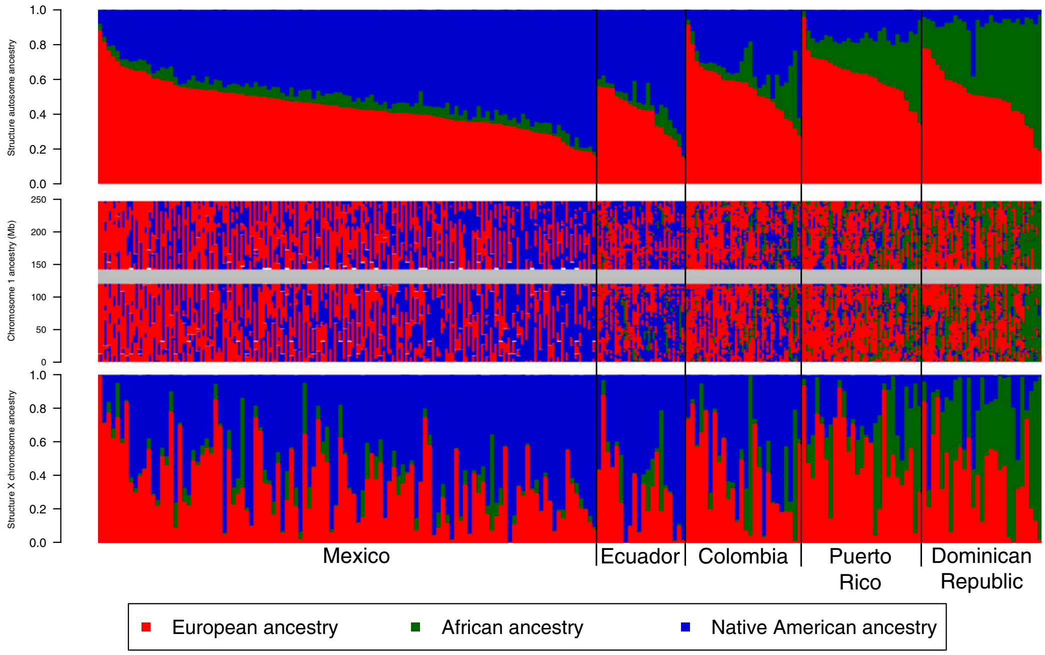

Fig. 3. Genome-wide and locus specific ancestry estimates for Mexicans, Ecuadorians, Colombians, Puerto Ricans, and Dominicans. Shown for $K = 3$, clustering of the Hispanic/Latino individuals on the autosomes (*Top*) and on the X chromosome (*Bottom*). Individuals are shown as vertical bars colored in proportion to their estimated ancestry within each cluster. Local ancestry at each locus is shown for each individual on chromosome 1 (*Middle*). The X chromosome shows greater Native American ancestry (blue) and greater variability in African ancestry (green), with reduced European ancestry (red).

for all of the admixed and putative source population individuals (i.e., either the 610K Illumina for Puerto Rican, Ecuadorian, Columbian, and Dominican or 500K for Mexicans from Guadalajara) as well as ~70 K SNPs common to both platforms. To calculate F_{ST} at a given SNP for a given pair of populations, we included only individuals with unambiguous ancestry assignment (i.e., individuals with two European-, two Native American–, or two African-origin chromosomes). One potential confounder for this analysis is that sample sizes differ substantially among subpopulations within major continental regions (e.g., in the Native American set, we have sample sizes that range from $n = 7$ for Colombian indigenous Americans in HGDP to $n = 29$ for Nahua from Mexico in the Mao et al. dataset). To minimize the potential bias of differences in sample size, we randomly selected $n = 7$ individuals from all potential subpopulations and recomputed Wright's F_{ST}. As seen in Table 1, we found that consistent with historical records, our results show that African segments of the Hispanic/Latino populations are more closely related to the Bantu-speaking populations of West Africa than other populations. Specifically, we found that the Colombians and Ecuadorians are most closely related to the Kenyan Bantu populations, whereas the Puerto Ricans and Dominicans are most close to the Yoruba from Nigeria. Likewise, European segments show the lowest F_{ST} values when compared with Southwest European populations (individuals from Spain and Portugal), as well as French and Italian individuals. Native American segments of the Hispanic/Latino individuals show the least genetic differentiation with Mesoamerican (e.g., Maya and

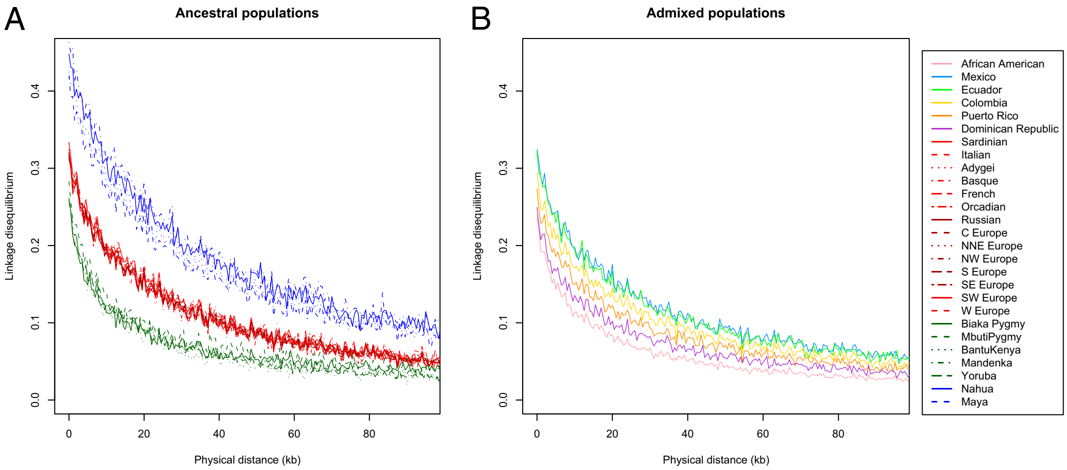

Fig. 4. Linkage disequilibrium, genotype r^2 estimated by PLINK, by population as a function of physical distance (Mb). (*Left*) Native American, European, and African populations. (*Right*) Hispanic/Latino populations. Scale is the same.

Table 1. Ancestry-specific F_{ST} distances between Hispanic/Latino populations and different putative source populations

African segments of the genome

	Bantu Kenya	Bantu S. Africa	Biaka Pygmy	Mandenka	Mbuti Pygmy	YRI
COL	3.191%	3.375%	6.520%	3.677%	11.217%	3.263%
DOM	1.564%	1.476%	4.657%	1.419%	8.877%	0.913%
ECU	6.098%	6.883%	10.143%	6.400%	14.702%	6.481%
PRI	2.500%	2.543%	5.761%	2.384%	10.216%	2.176%

European segments of the genome

	Adygei	Basque	EuropeanESE	EuropeC	EuropeNNe	EuropeNW	EuropeS	EuropeSE	EuropeSW	EuropeW	French	Italian	Orcadian	Russian	Sardinian	Tuscan
COL	1.836%	1.351%	1.389%	0.978%	1.253%	1.240%	1.033%	1.020%	0.863%	1.080%	0.880%	0.885%	1.410%	1.648%	1.550%	1.050%
DOM	1.560%	1.128%	1.071%	0.691%	0.919%	0.940%	0.705%	0.775%	0.537%	0.730%	0.613%	0.610%	1.093%	1.413%	1.270%	0.825%
ECU	1.669%	1.456%	1.225%	1.012%	1.212%	1.100%	1.005%	1.005%	0.838%	1.104%	0.799%	0.845%	1.417%	1.369%	1.607%	0.925%
PRI	1.811%	1.530%	1.392%	1.062%	1.345%	1.251%	1.107%	1.181%	0.916%	1.155%	0.940%	0.879%	1.508%	1.820%	1.566%	1.041%
Mexico	1.014%	0.784%	0.559%	0.335%	0.438%	0.442%	0.193%	0.307%	0.122%	0.265%	0.270%	0.271%	0.793%	0.882%	0.852%	0.336%

Native American segments of the genome

	Aymara	Colombian	Karitiana	Maya	Nahua	Pima	Quechua	Surui
COL	4.005%	5.296%	9.099%	4.724%	3.614%	8.562%	3.432%	13.803%
DOM	5.142%	5.868%	9.060%	4.262%	3.601%	9.310%	3.147%	13.736%
ECU	4.244%	5.799%	9.178%	5.446%	4.147%	9.193%	3.079%	13.765%
PRI	5.872%	6.618%	10.120%	6.624%	4.795%	10.578%	5.169%	15.093%
Mexico	2.397%	4.185%	8.197%	1.417%	0.572%	5.112%	2.086%	11.061%

Results based on ~70K overlapping SNPs between Affymetrix and Illumina platforms and equalizing population sample sizes down to seven individuals per population.

Nahua), Chibchan (e.g., Colombian), and Andean (e.g., Quechua) populations. The closest relationship is clearly observed between Mexicans from Guadalajara and Nahua indigenous individuals.

Sex Bias in Ancestry Contributions. We used the *STRUCTURE* ancestry estimates on the autosomes and X chromosome to estimate Native American, European, and African, ancestry proportions of each Hispanic/Latino individual. We then compared the estimates of ancestry for each population on the autosomes vs. on the X chromosome (Fig. 5 and Figs. S3 and S4). Whereas the Native American ancestry was significantly higher on the X chromosome than on the autosomes (including those populations with reduced Native American ancestry, i.e., Puerto Ricans and Dominicans), the autosomal vs. X-chromosome difference was more attenuated with regard to African ancestry. This reduced deviation is present even in those Hispanic/Latino populations analyzed whose non-European ancestry was principally Native American in origin (i.e., Mexicans and Ecuadorians). Furthermore, greater Native American ancestry on the X chromosome in Puerto Ricans did not necessarily imply greater Amerindian ancestry on the autosomes. This finding is similar to those observed by analyzing fine-scale genome pattern of population structure and admixture among African Americans, West Africans, and Europeans (31).

Finally, we used SNP and microsatellite genotyping to identify the canonical Y chromosome and mtDNA haplotypes for each of the Hispanic/Latino individuals that we genotyped. Details of the loci and classifications are found in Tables S1 and S2. We found an excess of European Y chromosome haplotypes and a higher proportion of Native American and African mtDNA haplotypes, consistent with previous studies (Fig. 6). In addition, we found several non-European Y chromosomal haplotypes with most likely origins from North Africa and the Middle East. We observed that African-derived haplotypes were the predominant origin of mtDNA in Dominicans (17 of 27 individuals), matching the greater African vs. Native American origins of this population on the autosomes and X-chromosomes. However, in Puerto Ricans we did not find evidence of a high African female contribution. The predominant Y chromosomal origins in the Puerto Ricans sampled were European and African; but, in contrast, 20 of 27 Puerto Rican individuals had mitochondrial haplotypes of Native American origin, suggesting a strong female Native American and male European and African sex bias contribution. Overall, in all of the Hispanic/Latino populations that we analyzed, we found evidence of greater European ancestry on the Y chromosome and higher Native American ancestry on the mtDNA and X chromosome consistent with previous findings (5–11).

Discussion

Our work has important implications for understanding the population genetic history of Latin America as well as ancestry of United States–based Hispanic/Latino populations. As has been previously documented, we found large variation in the proportions of European, African, and Native American ancestry among Mexicans, Puerto Ricans, Dominicans, Ecuadorians, and Colombians, but also within each of these groups. These trends are a consequence of variation in rates of migration from ancestral European and African source populations as well as population density Native Americans in pre-Columbian times (1). We found that Dominicans and Puerto Ricans in our study showed the highest levels of African ancestry, consistent with historical records. European settlers to island nations in the Caribbean basin largely displaced Native American populations by the early to mid 16th century and concurrently imported large African slave populations for large-scale colonial agricultural production (largely of sugar). In contrast, Colombia has wider geographic differences ranging from Caribbean coasts to Andean

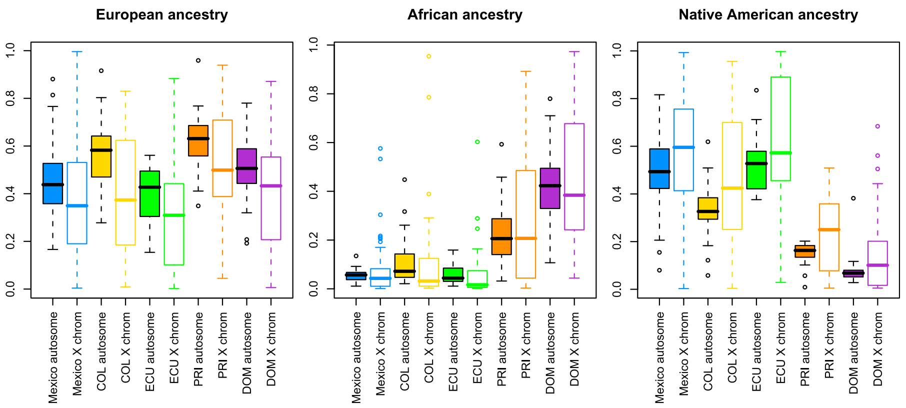

Fig. 5. Boxplots comparing autosomal vs. X chromosome ancestry proportions by population, shown for European ancestry (*Left*), Native American ancestry (*Center*), and African ancestry (*Right*). Filled boxes correspond to autosomal ancestry estimates; open boxes show X chromosome ancestry estimates. Median (solid line), first and third quartiles (box) and the minimum/maximum values, or to the smallest value within 1.5 times the IQR from the first quartile (whiskers). For each paired comparison of X chromosomes and autosomes, median Native American ancestries are consistently higher on the X chromosome in all Hispanic/Latino populations sampled, and European ancestries are lower across all populations.

valleys and mountains, which could explain the enrichment of African ancestry in some individuals and not in others, likely representing the differences in origin within Colombia. Finally, Mexico and Ecuador are two continental countries that had high densities of Native Americans during pre-Columbian times; as expected, the individuals from these two countries show the highest degree of Native American ancestry. Our findings clearly show that the involuntary migration of Africans through the slave trade appears to have left a clear trace in Hispanic/Latino populations proximal to these routes.

From the F_{ST} analysis, we found that the high-density genotype data that we have collected is quite informative regarding the personal genetic ancestry of admixed Hispanic/Latino individuals. Specifically, we found that individuals differ dramatically within and among populations and that we can reliably identify subpopulations within major geographic regions (i.e., Europe, Africa, and the Americas) that exhibit lower pairwise F_{ST} (and, therefore, higher genetic similarity) to the inferred European, African, and Native American segments for the 212 individuals studied. We found, for example, that Nahua showed the lowest F_{ST} in Mexicans, consistent with the observation that the Nahua are one of the largest Native American populations in this region and are likely to have contributed to the genomes of admixed individuals in Mexico (as opposed, for instance, to the Mexican Pima who fall outside the Mesoamerican cultural region and show considerably higher levels of differentiation). We also found that the lowest F_{ST} for the African regions of the Dominican and Puerto Rican genomes are with the Yoruba, a Bantu-speaking West African population that has been shown to be genetically similar to the African segments of African Americans sampled in the United States (30). Although we have limited Native American populations and Hispanic/Latino sample sizes and, thus, the differences in F_{ST} with different subcontinental populations suggest that there exists a reasonably strong signal of which present day populations are most closely related to the ancestral populations that contributed ancestry to each of the Hispanic/Latino populations.

When comparing inferred continental ancestry of the X and Y chromosomes and mitochondrial vs. the autosomal genome, we observed an enrichment of European Y-chromosome vs. autosomal genetic material, and a greater percentage of both Native American and African ancestry on the X-chromosomes and mtDNA compared with the autosomes for the Hispanic/Latino individuals in this study. This suggests a predominance of European males and Native American/African females in the ancestral genetic pool of Latinos, consistent with previous studies. A particularly interesting observation from our work on sex-biased admixture is that the pattern exists not only within populations but among Hispanic/Latino populations as well. In all populations studied, there is an enrichment of Native American ancestry both on the X chromosome and mtDNA compared with the autosomes. This would suggest that a greater

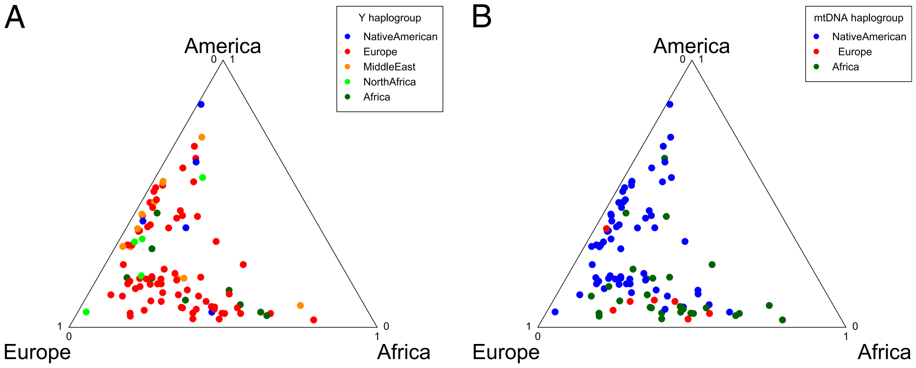

Fig. 6. Comparison of mtDNA and Y chromosome haplotypes. Each individual is represented by a point within the triangle that represents the autosomal ancestry proportions. The most probable continental location for each individual's haplotype is designated by the color of the point. The Y chromosome contains a disproportionate number of European haplotypes, whereas the mtDNA has a high proportion of Native American, slightly more African haplotypes and fewer European haplotypes, consistent with a sex bias toward a great European male and Native American/African female ancestry in the Hispanic/Latinos.

female Native American contribution to the genome of Latinos. A different result was obtained in relation to African ancestry. We found a smaller difference between mean African ancestry on the X chromosome and the autosomes, compared with the difference in Native American ancestry. Furthermore, unlike in Native American ancestry, we found an overwhelming representation of Native American mtDNA haplogroups in Puerto Ricans, even though non-European ancestry on the autosomes was largely African.

It is important to note that this observation does not necessarily undermine the model of sex-biased admixture among European male and African females in the founding of Hispanic/Latino populations, especially when one considers the predominance of European Y chromosomes in all groups studied. However, it suggests that admixture between European males and Amerindian/African females has been a complex process in the formation of the various Hispanic/Latino populations. Specifically, a reduced X vs. autosome mean African ancestry compared with Native American ancestry suggests a more balanced gender contribution in the Hispanic/Latino genome by individuals of African ancestry. In the case of Puerto Ricans, the only way that one can reconcile greater African ancestry on the X chromosome vs. what would be expected on mitochondrial data would be through transmission of X chromosomes independent of mitochondrial transmission, which is plausible biologically only via males. Caution, however, should be exercised before considering such conclusions as concrete; unlike X chromosomes, which can recombine and thus represent haplotypes derived from thousands of individuals, mitochondrial DNA represents just a sole distant ancestor among these thousands. Thus, a larger mtDNA sample would be necessary compared with X chromosomes to have similar confidence that a cohort would accurately reflect the presumed diversity of ancestry in the population as a whole.

The Y chromosomal results also demonstrate the insufficiency of the paradigm of European males and Native American/African females to capture the complexity within the Latin American populations. For example, we find Y chromosomal haplotypes in Hispanic/Latinos with presumed origins in the Middle East and Northern Africa. Given that historical documentation suggests that most of the non-African and non–Native American contribution to admixed Hispanic/Latino populations is from Southwest Europe, this suggests that the contemporary populations inherited these Y chromosomes from Europeans who, in turn, were descended from Middle Eastern or North African men. Several historical events could have led to the acquisition by Europeans of non-European haplotypes, perhaps during the period of the Roman Empire when the Mediterranean Sea behaved as a conduit (not a physical barrier) between Europe, the Middle East, and North Africa or by Sephardic Jews or Moorish Muslims during the European Middle Ages/Islamic Golden Age. Alternatively, the presence of non-European Y chromosomal haplotypes originating from the Middle East and North Africa could represent the result of Iberian Jews and Muslims (themselves admixed) fleeing the peninsula for New World territories in response to discriminatory policies that strongly pressured both communities at the termination of the Reconquista. Essentially, the diversity of haplotypes in the Y chromosomes in Latinos reflects not only population dynamics from the 15th century onward, but also the historical trends of population movement occurring across the Atlantic during centuries prior.

The marked genetic heterogeneity of Latino populations shown in this study, as previously suggested by other surveys of genetic ancestry (2, 26, 32) has important implications for the identification of disease-associated variants that differ markedly in frequency among parental populations. In their study of 13 Mestizo populations from Latin America, for example, Wang et al. (2008) suggested that admixture mapping in Hispanic/Latino populations may be feasible within a two-population admixture framework, since the mean African ancestry in Mestizo populations is typically low (<10%) (2). Although this is true for Hispanic/Latino populations with origins in the continental landmass of the Americas (such as the populations studied by Wang et al.), our results show that this may not apply to Latino populations with origins in the Caribbean, as their African ancestry proportion is considerably higher and is highly variable among individuals, suggesting an extensive three-way admixture and representing additional challenges for admixture mapping. Likewise, we find subtle but reproducible differences in subcontinental ancestry among Hispani/Latino individuals, suggesting that even a three-way admixture model may not be sufficient to accurately model the dynamic population genetic history of these populations.

Another observation with important implications for designing association studies is the large variation in individual admixture estimates within certain Latino populations (e.g., Mexicans, Colombians, and Ecuadorians). One could expect such outcome when collecting samples from United States–based Latino communities, which in turn may come from different locations within their countries of origin (e.g., Colombians and Ecuadorians). However, within the Mexican sample, which has been collected in a single sampling location (i.e., Guadalajara, Mexico), we also observed large variation in European vs. Native American admixture proportions. Our findings are in agreement with previous studies on genetic ancestry from Mexico City (2, 33), supporting the idea that such urban agglomerations, in which a large number of epidemiological studies are likely to take place, continue to host a wide range of genetic variability among individuals that may self-identify as individuals from the same population. Therefore, particular attention should be paid to carefully matching representative cases and controls, as well as to carefully control for ancestry when performing association studies using Hispanic/Latino populations. We hope that our dense genome-wide admixture analysis has allowed greater insight into the population dynamics of multiple Hispanic/Latino populations and that it will provides a resource for designing next-generation epidemiological studies in these communities, opening the possibility of better understanding the genetic makeup of this growing segment of the U.S. population.

Materials and Methods

Datasets. We genotyped 100 individuals with ancestry from Puerto Rico, the Dominican Republic, Ecuador, and Colombia on Illumina 610K arrays. We extracted 400 European, 365 African American, and 112 Mexican samples from the GlaxoSmithKline POPRES project, which is a resource of nearly 6,000 control individuals from North America, Europe, and Asia genotyped on the Affymetrix GeneChip 500K Array Set (23). We randomly sampled 15 individuals from each European country where possible, or the maximum number of individuals available otherwise, to select the POPRES European individuals to be included in our study. Further description of sampling locations, genotyping, and data quality control are available elsewhere (23). We include 165 and 167 individuals from the HapMap project from the CEU and YRI populations, thinned to the same SNP set (21). We also include all European, Native American, and African individuals from the HGDP genotyped on Illumina 610K arrays (25). Finally, we include all Native American populations from the Mao et al. (2007) study genotyped on Affymetrix 500K arrays (26). For each dataset, we used annotation information to determine the strand on which the data were given and to map all Affymetrix and Illumina marker ids to corresponding dbSNP reference ids [rsids]. SNPs without valid rsids were excluded from analysis. Each dataset was then converted to the forward strand to facilitate merging of the data. Data from the various platforms were merged using the PLINK toolset, version 1.06 (34). Likewise, nonmissing genotype calls that showed disagreement between datasets were omitted. Demographic data for all individuals included in this study are available on GenBank. All samples were approved by institutional review board protocols from their respective studies.

Data Quality Control. The HapMap II release 23, HGDP, Mao et al., and POPRES samples were genotyped and called according to their respective quality control procedures (21, 23, 25, 26). Our final merged dataset contains 73,901

SNPs with genotype missingness of <0.1 and <0.05 individual missingness across 5,104 individuals.

Population Structure. We used the software *FRAPPE*, which implements an expectation-maximization algorithm for estimating individual membership in clusters (24). This algorithm is more computationally efficient than other MCMC methods, allowing it to analyze many more markers than, for example, STRUCTURE (24, 28). After thinning markers to have $r^2 < 0.5$ in 50 SNP windows, shifted and recalculated every 5 SNPs, we ran *FRAPPE* on all 64,935 remaining markers for 5,000 iterations. We also assessed admixture proportions for the Hispanic/Latino individuals using *STRUCTURE* on a reduced dataset of 5,440 markers after thinning for MAF > 0.2 and with a minimum separation of 400 Kb between markers. We use the F model with USEPOPINFO = 1 to update allele frequencies using only the ancestral individuals, with 5,000 burn-in and 5,000 iterations (28). We also used all 1,518 SNPs on the X chromosome for the same analysis of the X chromosome ancestry. Principal component analysis was conducted using a dataset thinned to have $r^2 < 0.8$ in 50 SNP windows, leaving 69,212 SNPs for analysis using the package *smartpca* from the software *eigenstrat*. Ellipses were fitted following the means and 1 SD of the variance–covariance matrix of the PC1 and PC2 scores of each population.

For local ancestry estimation, we used the software LAMP in LAMPANC mode providing allele frequencies for the HGDP West Africans, Europeans, and Native Americans as ancestral populations (29). A total of 552,025 SNPs were included in the analysis, and configuration parameters were set as follows: mixture proportions (alpha) = 0.2, 0.4, 0.4; number of generations since admixture (g) = 20; recombination rate (r) = 1e-8; fraction of overlap between adjacent windows (offset) = 0.2; and r2 threshold (ldcutoff) = 0.1. Local ancestry estimation for the Mexican individuals was performed using the two-way PCA-based method described in Bryc et al. (30) for both the full Illumina 610K and the Affymetrix 500K datasets, in 10 SNP windows. Only Native Americans with <0.01 European ancestry (as estimated from *FRAPPE* results) were used as the ancestral Native American individuals within their respective datasets. F_{ST} was calculated between Native American, European, and African regions of the Hispanic//Latino individuals and the respective continental populations using a C++ implementation of Weir and Cockerham's F_{ST} weighed equations as previously published (35). To eliminate bias in estimation of F_{ST} due to European ancestry shown in some of the Native Americans, we also removed regions showing European ancestry within any of the Native Americans showing >0.01 European ancestry, using the same local ancestry estimation procedure as described for the Mexican individuals. Furthermore, to avoid any potentially confounding effect of sample size, we used a random sample of 7 (the minimum sample size of the Native American populations) individuals per non-Hispanic/Latino population to calculate pairwise F_{ST}. MAF was set at a threshold >0.1 in the populations compared by F_{ST} calculations.

ACKNOWLEDGMENTS. We thank Mariano Rey for support of the project; Peter Gregersen, Carole Oddoux, and Annette Lee for technical assistance; and Marc Pybus for valuable programming support during part of the analyses. This work was supported by the National Institutes of Health (Grant 1R01GM83606) as part of the National Institute of General Medical Sciences research funding programs.

1. Sans M (2000) Admixture studies in Latin America: From the 20th to the 21st century. *Hum Biol* 72:155–177.
2. Wang S, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4:e1000037.
3. Seldin MF, et al. (2007) Argentine population genetic structure: Large variance in Amerindian contribution. *Am J Phys Anthropol* 132:455–462.
4. Silva-Zolezzi I, et al. (2009) Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc Natl Acad Sci USA* 106:8611–8616.
5. Dipierri JE, et al. (1998) Paternal directional mating in two Amerindian subpopulations located at different altitudes in northwestern Argentina. *Hum Biol* 70:1001–1010.
6. González-Andrade F, Sánchez D, González-Solórzano J, Gascón S, Martínez-Jarreta B (2007) Sex-specific genetic admixture of Mestizos, Amerindian Kichwas, and Afro-Ecuadorans from Ecuador. *Hum Biol* 79:51–77.
7. Green LD, Derr JN, Knight A (2000) mtDNA affinities of the peoples of North-Central Mexico. *Am J Hum Genet* 66:989–998.
8. Mendizabal I, et al. (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. *BMC Evol Biol* 8:213.
9. Marrero AR, et al. (2007) Pre- and post-Columbian gene and cultural continuity: The case of the Gaucho from southern Brazil. *Hum Hered* 64:160–171.
10. Sans M, et al. (2002) Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. *Am J Phys Anthropol* 118:33–44.
11. Carvajal-Carmona LG, et al. (2003) Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Hum Genet* 112:534–541.
12. Smith MW, et al. (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am J Hum Genet* 69:1080–1094.
13. González Burchard E, et al. (2005) Latino populations: A unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health* 95:2161–2168.
14. Salari K, et al. (2005) Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genet Epidemiol* 29:76–86.
15. Fejerman L, et al. (2008) Genetic ancestry and risk of breast cancer among U.S. Latinas. *Cancer Res* 68:9723–9728.
16. Lai CQ, et al. (2009) Population admixture associated with disease prevalence in the Boston Puerto Rican health study. *Hum Genet* 125:199–209.
17. Hayes MG, et al. (2007) Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* 56:3033–3044.
18. Choudhry S, et al. (2008) Genome-wide screen for asthma in Puerto Ricans: Evidence for association with 5q23 region. *Hum Genet* 123:455–468.
19. Choudhry S, et al. (2006) Ancestry-environment interactions and asthma risk among Puerto Ricans. *Am J Respir Crit Care Med* 174:1088–1093.
20. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.
21. Frazer KA, et al.; International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
22. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298:2381–2385.
23. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83:347–358.
24. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 28:289–301.
25. Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.
26. Mao X, et al. (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* 80:1171–1178.
27. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
28. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.
29. Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. *Am J Hum Genet* 82:290–303.
30. Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107:786–791.
31. Lind JM, et al. (2007) Elevated male European and female African contributions to the genomes of African American individuals. *Hum Genet* 120:713–722.
32. Price AL, et al. (2007) A genomewide admixture map for Latino populations. *Am J Hum Genet* 80:1024–1036.
33. Martinez-Marignac VL, et al. (2007) Admixture in Mexico City: Implications for admixture mapping of type 2 diabetes genetic risk factors. *Hum Genet* 120:807–819.
34. Purcell SNB, et al. (2007) PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 81:559–575.
35. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.