Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations

Fernanda S. G. Kehdy[a,1], Mateus H. Gouveia[a,1], Moara Machado[a,1], Wagner C. S. Magalhães[a,1], Andrea R. Horimoto[b], Bernardo L. Horta[c], Rennan G. Moreira[a], Thiago P. Leal[a], Marilia O. Scliar[a], Giordano B. Soares-Souza[a], Fernanda Rodrigues-Soares[a], Gilderlanio S. Araújo[a], Roxana Zamudio[a], Hanaisa P. Sant Anna[a], Hadassa C. Santos[b], Nubia E. Duarte[b], Rosemeire L. Fiaccone[d], Camila A. Figueiredo[e], Thiago M. Silva[f], Gustavo N. O. Costa[f], Sandra Beleza[g], Douglas E. Berg[h,i], Lilia Cabrera[j], Guilherme Debortoli[k], Denise Duarte[l], Silvia Ghirotto[m], Robert H. Gilman[n,o], Vanessa F. Gonçalves[p], Andrea R. Marrero[k], Yara C. Muniz[k], Hansi Weissensteiner[q], Meredith Yeager[r], Laura C. Rodrigues[s], Mauricio L. Barreto[f], M. Fernanda Lima-Costa[t,2], Alexandre C. Pereira[b,2], Maíra R. Rodrigues[a,2], Eduardo Tarazona-Santos[a,2,3], and The Brazilian EPIGEN Project Consortium[4]

[a]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; [b]Instituto do Coração, Universidade de São Paulo, 05403-900, São Paulo, Sao Paulo, Brazil; [c]Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, 464, 96001-970 Pelotas, Rio Grande do Sul, Brazil; [d]Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, 40170-110, Salvador, Bahia, Brazil; [e]Departamento de Ciências da Biointeração, Instituto de Ciências da Saúde, Universidade Federal da Bahia, 40110-100, Salvador, Bahia, Brazil; [f]Instituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, Bahia, Brazil; [g]Department of Genetics, University of Leicester, LE1 7RH, Leicester, United Kingdom; [h]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110; [i]Department of Medicine, University of California, San Diego, CA 92093; [j]Biomedical Research Unit, Asociación Benéfica Proyectos en Informática, Salud, Medicina y Agricultura (AB PRISMA), 170070, Lima, Peru; [k]Departamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Santa Catarina, Brazil; [l]Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; [m]Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121 Ferrara, Italy; [n]Bloomberg School of Public Health, International Health, Johns Hopkins University, Baltimore, MD 21205; [o]Laboratorio de Investigación de Enfermedades Infecciosas, Universidade Peruana Cayetano Heredia, 15102, Lima, Peru; [p]Department of Psychiatry and Neuroscience Section, Center for Addiction and Mental Health, University of Toronto, Toronto, ON, Canada M5T 1R8; [q]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, 6020 Innsbruck, Austria; [r]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 20850; [s]Department of Infectious Disease Epidemiology, Faculty of Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; and [t]Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, 30190-002, Belo Horizonte, Minas Gerais, Brazil

Edited by Marcus W. Feldman, Stanford University, Stanford, CA, and approved May 27, 2015 (received for review March 8, 2015)

While South Americans are underrepresented in human genomic diversity studies, Brazil has been a classical model for population genetics studies on admixture. We present the results of the EPIGEN Brazil Initiative, the most comprehensive up-to-date genomic analysis of any Latin-American population. A population-based genome-wide analysis of 6,487 individuals was performed in the context of worldwide genomic diversity to elucidate how ancestry, kinship, and inbreeding interact in three populations with different histories from the Northeast (African ancestry: 50%), Southeast, and South (both with European ancestry >70%) of Brazil. We showed that ancestry-positive assortative mating permeated Brazilian history. We traced European ancestry in the Southeast/South to a wider European/Middle Eastern region with respect to the Northeast, where ancestry seems restricted to Iberia. By developing an approximate Bayesian computation framework, we infer more recent European immigration to the Southeast/South than to the Northeast. Also, the observed low Native-American ancestry (6–8%) was mostly introduced in different regions of Brazil soon after the European Conquest. We broadened our understanding of the African diaspora, the major destination of which was Brazil, by revealing that Brazilians display two within-Africa ancestry components: one associated with non-Bantu/western Africans (more evident in the Northeast and African Americans) and one associated with Bantu/eastern Africans (more present in the Southeast/South). Furthermore, the whole-genome analysis of 30 individuals (42-fold deep coverage) shows that continental admixture rather than local post-Columbian history is the main and complex determinant of the individual amount of deleterious genotypes.

Latin America | population genetics | Salvador SCAALA | Bambuí Cohort Study of Ageing | Pelotas Birth Cohort Study

Latin Americans, who are classical models of the effects of admixture in human populations (1, 2), remain underrepresented in studies of human genomic diversity, notwithstanding recent studies (3, 4). Indeed, no large genome-wide study on admixed South Americans has been conducted so far. Brazil is the largest and most populous Latin-American country. Its over 200 million inhabitants are the product of post-Columbian admixture between Amerindians, Europeans colonizers or immigrants, and African slaves (1). Interestingly, Brazil was the destiny of nearly 40% of the African diaspora, receiving seven times more slaves than the United States (nearly 4 million vs. 600,000).

Here, we present results of the EPIGEN Brazil Initiative (https://epigen.grude.ufmg.br), the most comprehensive up-to-date genomic analysis of a Latin-American population. We genotyped nearly 2.2 million SNPs in 6,487 admixed individuals from three population-based cohorts from different regions with distinct demographic and socioeconomic backgrounds and sequenced the whole genome of 30 individuals from these populations at an

Author contributions: E.T.-S. designed research; F.S.G.K., M.H.G., M.M., W.C.S.M., A.R.H., B.L.H., R.G.M., M.L.B., M.F.L.-C., A.C.P., M.R.R., and E.T.-S. performed research; T.P.L., R.Z., R.L.F., C.A.F., T.M.S., G.N.O.C., S.B., D.E.B., L.C., R.H.G., M.Y., L.C.R., M.R.R., and T.B.E.P.C. contributed new reagents/analytic tools; F.S.G.K., M.H.G., M.M., W.C.S.M., A.R.H., R.G.M., T.P.L., M.O.S., G.B.S.-S., F.R.-S., G.S.A., H.P.S.A., H.C.S., N.E.D., G.D., D.D., S.G., V.F.G., A.R.M., Y.C.M., and H.W. analyzed data; F.S.G.K., M.H.G., M.M., W.C.S.M., R.G.M., M.R.R., and E.T.-S. wrote the paper; F.S.G.K. coordinated the ancestry team of the project; W.C.S.M. coordinated the inputation team of the project; A.R.H. coordinated the basic analyses team of the project; B.L.H. coordinated the 1982 Pelotas Birth Cohort; M.L.B. coordinated the SCAALA (Social Changes, Asthma and Allergy in Latin America Program) cohort; M.F.L.-C. coordinated the Bambui cohort; A.C.P. and E.T.-S. supervised the genome analysis group of the project; and M.R.R. coordinated the bioinformatics team of the project.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the European Nucleotide Archive (PRJEB9080 (ERP010139) Genomic Epidemiology of Complex Diseases in Population-Based Brazilian Cohorts), accession no. EGAS00001001245, under EPIGEN Committee Controlled Access mode.

[1] F.S.G.K., M.H.G., M.M., and W.C.S.M. contributed equally to this work.

[2] M.F.L.-C., A.C.P., M.R.R., and E.T.-S. contributed equally to this work.

[3] To whom correspondence should be addressed. Email: edutars@icb.ufmg.br.

[4] A complete list of the Brazilian EPIGEN Project Consortium can be found in *SI Appendix*.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1504447112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1504447112

Significance

The EPIGEN Brazil Project is the largest Latin-American initiative to study the genomic diversity of admixed populations and its effect on phenotypes. We studied 6,487 Brazilians from three population-based cohorts with different geographic and demographic backgrounds. We identified ancestry components of these populations at a previously unmatched geographic resolution. We broadened our understanding of the African diaspora, the principal destination of which was Brazil, by revealing an African ancestry component that likely derives from the slave trade from Bantu/eastern African populations. In the context of the current debate about how the pattern of deleterious mutations varies between Africans and Europeans, we use whole-genome data to show that continental admixture is the main and complex determinant of the amount of deleterious genotypes in admixed individuals.

average deep coverage of 42× (Fig. 1*B* and *SI Appendix*, sections 1, 2, and 8). By leveraging on a population-based approach, we (*i*) identified and quantified ancestry components of three representative Brazilian populations at a previously unmatched geographic resolution; (*ii*) developed an approximate Bayesian computation (ABC) approach and inferred aspects of the admixture dynamics in Northeastern, Southeastern, and Southern Brazil; (*iii*) elucidated how aspects of the ancestry-related social history of Brazilians influenced their genetic structure; and (*iv*) studied how admixture, kinship, and inbreeding interact and shape the pattern of putative deleterious mutations in an admixed population.

Results and Discussion

Populations, Continental Ancestry, and Population Structure. We studied the following three population-based cohorts (Fig. 1*B*). (*i*) SCAALA (Social Changes, Asthma and Allergy in Latin America Program) (5) (1,309 individuals) from Salvador, a coastal city with 2.7 million inhabitants in Northeastern Brazil that harbors the most conspicuous demographic and cultural African contribution (6). We inferred (7) that this population has the largest African ancestry (50.8%; SE = 0.35) among the EPIGEN populations, with 42.9% (SE = 0.35) and 6.4% (SE = 0.09) of European and Amerindian ancestries, respectively. Notably, this African ancestry is lower than that usually observed in African Americans (8, 9). (*ii*) The Bambuí Aging Cohort Study (10), ongoing in the homonymous city of ∼15,000 inhabitants, in the inland of Southeastern Brazil (1,442 individuals who were 82% of the residents older than 60 y old at the baseline year). We estimated that Bambuí has 78.5% (SE = 0.4) of European, 14.7% (SE = 0.4) of African, and 6.7% (SE = 0.1) of Amerindian ancestries. (*iii*) The 1982 Pelotas Birth Cohort Study (11) (3,736 individuals; 99% of all births in the city at the baseline year). Pelotas is a city in Southern Brazil with 214,000 inhabitants. Ancestry in Pelotas is 76.1% (SE = 0.33) European, 15.9% (SE = 0.3) African, and 8% (SE = 0.08) Amerindian.

By comparing autosomal mtDNA and X-chromosome diversity, we found across the three populations the signature of a historical pattern of sex-biased preferential mating between males with predominant European ancestry and women with predominant African or Amerindian ancestry (12) (*SI Appendix*, sections 6.6 and 6.9, Fig. S12, and Table S18). We determined (13) that individuals from Salvador and Pelotas were, with few exceptions, unrelated and have low consanguinity (Fig. 1*A* and *SI Appendix*, Figs. S1 and S2). Conversely, the Bambuí cohort has the highest family structure and inbreeding [Fig. 1*A* and *SI Appendix*, section 4.1 (discussion about the age structure of this cohort) and Figs. S1 and S2]. Bambuí includes several families with more than five related individuals showing at least one second-degree (or closer) relative. Bambuí mean inbreeding coefficient (0.010; SE = 0.0008) (*SI Appendix*, Fig. S2) is comparable with estimates observed in populations with 15–25% of consanguineous marriages from India (14). Interestingly, inbreeding in Bambuí was correlated with European ancestry ($\rho_{\text{Spearman}} = 0.20$; $P < 10^{-15}$). These higher inbreeding and kinship structures are consistent with Bambuí being the smallest and the most isolated of the EPIGEN populations.

Continental genomic ancestry in Latin America (and specifically, in Brazil) is correlated with a set of phenotypes, such as skin color and self-reported ethnicity, and social and cultural features, such as socioeconomic status (15–17). We observed a positive correlation across the three EPIGEN populations between SNP-specific Africans/Europeans F_{ST} (a measurement of informativeness of ancestry) and SNP-specific F_{IT} (a measurement of departure from Hardy–Weinberg equilibrium)

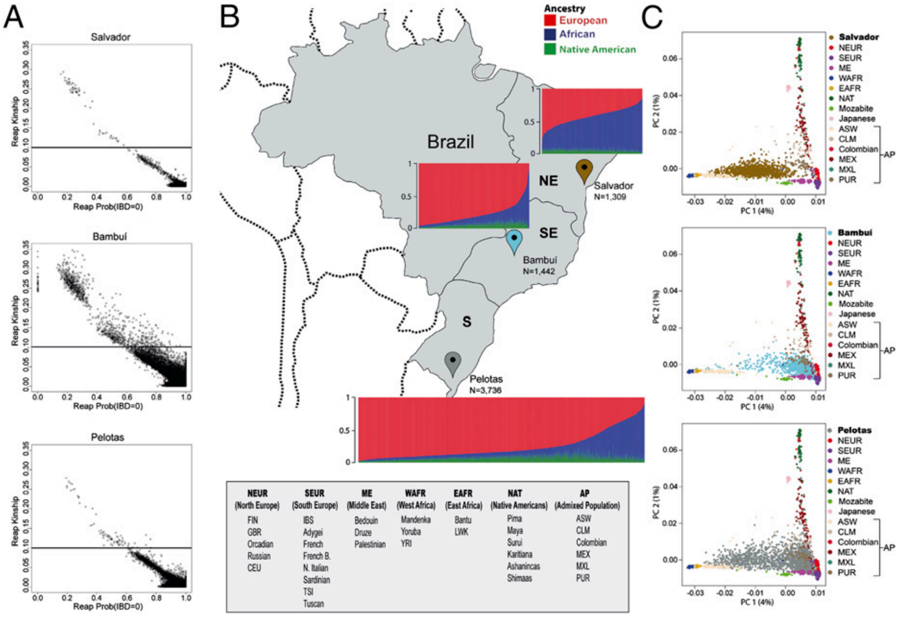


Fig. 1. Continental admixture and kinship analysis of the EPIGEN Brazil populations. (*A*) Kinship coefficient for each pair of individuals and the probability that they share zero identity by descent (IBD) alleles (IBD = 0). Horizontal lines represent a kinship coefficient threshold used to consider individuals as relatives. (*B*) Brazilian regions, the studied populations, and their continental individual ancestry bar plots. *N* represents the numbers of EPIGEN individuals in the Original Dataset (including relatives; detailed in *SI Appendix*, section 6). (*C*) PCA representation, including worldwide populations and the EPIGEN populations, using only unrelated individuals (Dataset U; explained in *SI Appendix*, section 6). The three graphics derive from the same analysis and are different only for the plotting of the EPIGEN individuals. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PC, principal component; PUR, Puerto Ricans from Puerto Rico; S, south; SE, southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa.

(*SI Appendix*, Fig. S3). This finding indicates that, after five centuries of admixture, Brazilians still preferentially mate with individuals with similar ancestry (and its correlated morphological phenotypes and socioeconomic characteristics), a trend also observed in Mexicans and Puerto Ricans (18). Interestingly, the highest correlations were found in Pelotas and Bambuí, consistent with their higher proportion of individuals with a clearly predominant ancestry (European or African) compared with Salvador (Fig. 1 *B* and *C*). Conversely, in Salvador, despite its highest mean African ancestry, individuals are more admixed (Fig. 1 *B* and *C*), probably because of a combination of a longer history of admixture (see below) and the lower and more homogeneous socioeconomic status of this cohort (5).

Three outcomes illustrate how population subdivision and inbreeding (both partly ancestry-dependent) interact to shape population structure in admixed populations with different sizes (*SI Appendix*, Figs. S1 and S3). First, Bambuí (the smallest city) has the strongest departure from Hardy–Weinberg equilibrium (F_{IT} =0.016; SE = 0.00003) because of both inbreeding (F_{IS} = 0.010; SE = 0.0008) and ancestry-based population subdivision ($\rho_{FIT\text{-}FST}$ = 0.18; $P < 10^{-16}$). Second, Pelotas (a medium-sized city; F_{IT} = 0.012; SE = 0.00002) has negligible inbreeding (F_{IS} = −0.001; SE = 0.0002) but the strongest ancestry-based population subdivision ($\rho_{FIT\text{-}FST}$ = 0.38; $P < 10^{-16}$). Third, the large city of Salvador shows the lowest inbreeding and ancestry-based population subdivision (F_{IT} = −0.003; SE = 0.00002; F_{IS} = −0.001; SE = 0.0003; $\rho_{FIT\text{-}FST}$ = 0.08; $P < 10^{-16}$).

Overall, the EPIGEN populations studied by a population-based approach exemplify how ancestry, kinship, and inbreeding may be differently structured in small (Bambuí), medium (Pelotas), and large (Salvador) admixed Latin-American populations. These populations fairly represent the three most populated Brazilian regions (Northeast, Southeast, and South) with their geographic distribution and continental ancestry (Fig. 1) and are good examples of the Latin-American genetic diversity with their ethnic diversity.

Differences in Admixture Dynamics. We estimated the continental origin of each allele for each SNP along each chromosome of the EPIGEN individuals (19) (*SI Appendix*, section 6.7) and calculated the lengths of chromosome segments of continuous specific ancestry (CSSA) (Fig. 2*A*), with distribution that informs how admixture occurred over time. By leveraging on the model by Liang and Nielsen (20) of CSSA, we developed an ABC framework to infer admixture dynamics (*SI Appendix*, section 6.8). We simulated CSSA distributions generated by a demographic history of three pulses of trihybrid admixture that occurred 18–16, 12–10, and 6–4 generations ago, conditioning on the observed current admixture proportions of each of the EPIGEN populations. This demographic model conciliates statistical complexity and the real history of admixture. We inferred the posterior distributions of nine parameters $m_{n,P}$, where m is the proportion of immigrant individuals entering in the admixed population from the n ancestral population (African, European, or Native-American ancestry) in the P admixture pulse.

Interestingly, ABC results (Fig. 2*B*) show that the observed low Native-American ancestry was mostly introduced in different regions of Brazil soon after the European Conquest of the Americas, which is consistent with the posterior depletion of the Native-American population in Brazil. Also, we inferred a predominantly earlier European colonization in the Northeast (Salvador) vs. a more recent immigration in Southeastern and Southern Brazil (Bambuí and Pelotas), consistent with historical records (brasil500anos.ibge.gov.br/). Conversely, African admixture showed a decreasing temporal trend shared by the three EPIGEN populations (21). Complementary explanations are continuous local immigration into the admixed populations from communities with high African ancestry already settled in Brazil [for example, quilombos (i.e., Afro-Brazilian slave-derived communities in Brazil) (22)].

Dissecting European Ancestry. To dissect the ancestry of Brazilians at a subcontinental level, we applied (*i*) the ADMIXTURE method (7) by increasing the number of ancestral clusters (*K*) that explains the observed genetic structure (*SI Appendix*, Figs. S4 and S5) and (*ii*) the Principal Component Analysis (PCA) (23) (Figs. 1*C* and 3 *B* and *D* and *SI Appendix*, Fig. S6). To study biogeographic ancestry, we excluded sets of relatives that could affect our inferences at the within-continent level (24). We developed a method based on complex networks to reduce the relatedness of the analyzed individuals by minimizing the number of excluded individuals (*SI Appendix*, section 6.1). Using this method, we created the Dataset Unrelated (Dataset U), including 5,825 Brazilians, 1,780 worldwide individuals, and no pair of individuals closer than second-degree relatives. Hereafter, PCA and ADMIXTURE results are relative to Dataset U.

Brazil received several immigration waves from diverse European origins during the last five centuries (brasil500anos.ibge.gov.br/): Portuguese (the first colonizers), who also arrived in large numbers during the last 150 y; Italians (mostly to the South and Southeast); and Germans (mostly to the South). In our PCA representation (Fig. 3*B*), the European component of the genomes of most Brazilians is similar to individuals from the Iberian Peninsula and neighboring regions. The resemblance in within-European ancestry of individuals from Pelotas (South) and Bambuí (Southeast) to central North Europeans and Middle Easters, respectively (Fig. 3*B*), reflects a geographically wider European ancestry of these two populations with respect to Salvador. Considering the total European ancestry estimated by ADMIXTURE, we inferred a higher proportion of North European-associated ancestry in Pelotas (40.2%) than in Bambuí (35.8%) and Salvador (36.7%; $P < 10^{-15}$; Wilcoxon tests) (Fig. 3*A*, red cluster in *K* = 7). We confirmed these results by analyzing a reduced number of SNPs with a larger set of

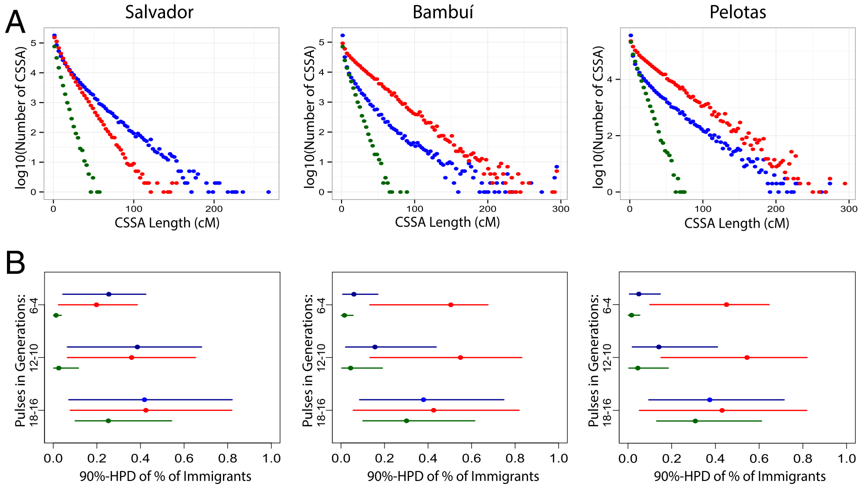


Fig. 2. Distributions of lengths of chromosomal segments of (*A*) CSSA and (*B*) admixture dynamics inferences estimated for three EPIGEN Brazilian populations. (*A*) CSSA lengths were distributed in 50 equally spaced bins per population. Red, blue, and green dots represent a European, an African, and a Native-American CSSA, respectively. (*B*) We inferred the posterior densities of the proportions of immigrants (with respect to the admixed population) from each origin, and we show their 90% highest posterior density (HPD) intervals. Inferences are based on a model of three pulses of admixture (vertical axis) simulated based on the model of CSSAs evolution by Liang and Nielsen (20). Inferences are based on approximate Bayesian computation. Ancestry color codes are red for European, blue for African, and green for Native American.

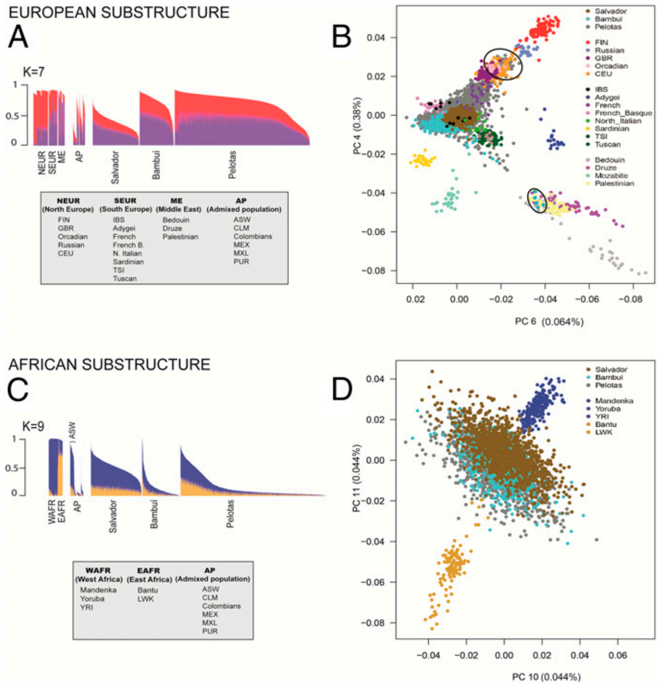


Fig. 3. European and African ancestry clusters in the Brazilian populations. We show (*A* and *C*) relevant ADMIXTURE individual ancestry bar plots and (*B* and *D*) plots of principal components (PCs) that dissect ancestry within (*A* and *B*) Europe and (*C* and *D*) Africa. We performed the analyses using Dataset U (unrelated Brazilians and worldwide individuals). We only plot individuals from relevant ancestral populations. Complete ADMIXTURE and PCA results are represented in *SI Appendix*, section 6 and Figs. S4–S6. Black ellipses in *B* show some individuals from Pelotas (Southern Brazil) clustering with northern European individuals toward the top and individuals from Bambuí (Southeastern Brazil) clustering with Middle Eastern individuals toward the bottom. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PUR, Puerto Ricans from Puerto Rico; S, south; SE, southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa.

European individuals and populations (25, 26) (*SI Appendix*, section 6.2).

Brazil, the Main Destination of the African Diaspora. African slaves arrived to Brazil during four centuries, whereas most arrivals to the United States occurred along two centuries, and the geographic and ethnic origin of Brazilian slaves differ from Caribbeans and African Americans (27). In fact, the Portuguese Crown imported slaves to Brazil from western and central west Africa (the two are the major sources of the slave trade to all of the Americas) as well as Mozambique. We detected two within-Africa ancestry clusters in the current Brazilian population (Fig. 3*C*, $K = 9$ and *SI Appendix*, section 6.3): one associated with the Yoruba/Mandenka non-Bantu western populations (Fig. 3*C*, blue) and one associated with the Luhya/HGDP (Human Genome Diversity Project) Bantu populations from eastern Africa (Fig. 3*C*, mustard). Interestingly, the proportions of these ancestry clusters, which are present across all of the analyzed African and Latin-American populations, differ across them. The blue cluster in Fig. 3*C* predominates in African Americans and in Salvador, accounting for 83% and 75% of the total African ancestry, respectively (against 17% and 25%, respectively, of the mustard cluster in Fig. 3*C*) (*SI Appendix*, Table S17). Comparatively, the mustard cluster in Fig. 3*C* is more evident in Southeastern and Southern Brazil (36% and 44% of African ancestry in Bambuí and Pelotas, respectively). These results are consistent with the fact that a large proportion of Yoruba slaves arrived in Salvador, whereas the Mozambican Bantu slaves disembarked primarily in Rio de Janeiro in Southeastern Brazil (21). These results show for the first time, to our knowledge, that the genetic structure of Latin Americans reflects a more diversified origin of the African diaspora into the continent. Interestingly, the two within-African ancestry clusters in the Brazilian populations (showing an average F_{ST} of 0.02) are characterized by 3,318 SNPs, with the 10% top F_{ST} values higher than 0.06, and include 38 SNPs that are hits of genome-wide association studies (*SI Appendix*, section 7 and Table S25).

Pattern of Deleterious Variants: Effect of Continental Admixture, Kinship, and Inbreeding. Based on whole-genome data from 30 individuals (10 from each of three EPIGEN populations), we identified putative deleterious nonsynonymous variants (28) (*SI Appendix*, section 8). There are recent interest in and apparently conflicting results on whether Europeans have proportionally more deleterious variants in homozygosis than Africans (29–32). Lohmueller et al. (29) explained these differences as an effect of the Out of Africa bottleneck on current non-African populations. Out of Africa would have enhanced the effect of genetic drift and attenuated the effect of purifying natural selection, preventing, in many instances, the extinction of (mostly weakly) deleterious variants in non-Africans.

We investigated how European ancestry shapes the amount of deleterious variants in homozygosis (a more likely genotype for common/weakly deleterious variants) and heterozygosis in admixed Latin-American individuals. We observed three patterns (Fig. 4). (*i*) Considering all (i.e., weakly and highly) deleterious variants, for a class of individuals with high European ancestry (>65%; from Bambuí and Pelotas), the individual number of deleterious variants in homozygosis is correlated with European ancestry, but importantly, this correlation is not observed among individuals with intermediate European ancestry (from Salvador) (Fig. 4*A*). (*ii*) The individual number of deleterious variants (both all and rare classes) in heterozygosis (Fig. 4 *B* and *D*) decreases linearly with European ancestry, regardless the cohort of origin. This result is also observed for rare deleterious variants in homozygosis, although the pattern is not very clear in this case (Fig. 4*C*). (*iii*) There are no differences in the amount of deleterious variants between individuals from Bambuí and Pelotas. These populations have similar continental admixture proportions and dynamics, but different post-Columbian population sizes and histories of isolation, assortative mating, kinship structure, and inbreeding. Taken together, our results are consistent with the results and evolutionary scenario proposed by Lohmueller et al. (29) and Lohmueller (31), and suggest that, in Latin-American populations, the main determinant of the amount of deleterious variants is the history of continental admixture, although in a more complex fashion than previously thought (pattern *i*). Comparatively, the role of local demographic history seems less relevant.

Conclusion

A thread of historical facts has modeled the genetic structure of Brazilians. Our population-based and fine-scale analyses revealed novel aspects of the genetic structure of Brazilians. In 1870, blacks were the major ethnic group in Brazil (21), but this scenario changed after the arrival of nearly 4 million Europeans during the second one-half of the 19th century and the first one-half of the 20th century. This immigration wave was encouraged by Brazilian officials as a way of "whiting" the population (33), and it transformed Brazil into a predominantly white country, particularly in the Southeast and South. Consistently, (*i*) we observed that larger chromosomal segments of continuous European ancestry in the southeast/south are the signature of this recent European immigration, and (*ii*) we traced the European ancestry in the Southeast/South of Brazil to a wider geographical region (including central northern Europe and the Middle East) than in Salvador (more

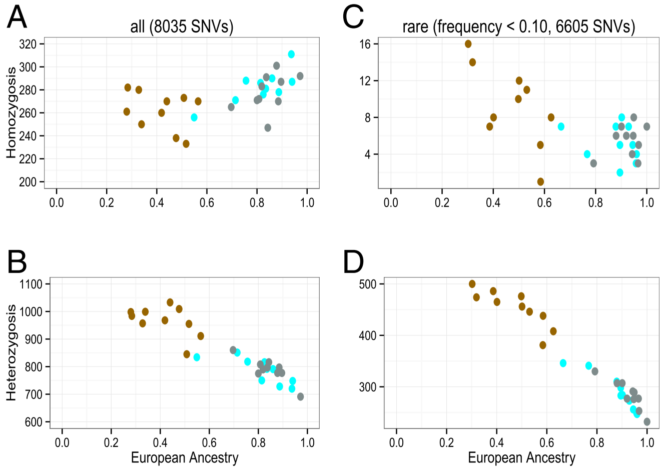


Fig. 4. Individual numbers of genotypes with nonsynonymous deleterious variants in homozygosis and heterozygosis vs. European ancestry based on the whole-genome sequence (42×) of 30 individuals (10 from each population): Salvador (Northeast; brown), Bambuí (Southeast; cyan), and Pelotas (South; gray). Deleterious variants were identified using CONDEL (28) and corrected for the bias reported by Simons et al. (30). Spearman correlation between European ancestry and the number of all deleterious variants in homozygosis for Bambuí and Pelotas individuals was 0.57 (P = 0.009). The numbers of genotypes considering all deleterious variants in homozygosis or heterozygosis are in *A* and *B*, respectively, and considering only rare deleterious variants are in *C* (in homozygosis) and *D* (in heterozygosis). SNVs, single nucleotide variants.

restricted to the Iberian Peninsula). However, neither this massive immigration nor the internal migration of black Brazilians have concealed two components of their African ancestry from the genetic structure of Brazilians: one associated with the Yoruba/Mandenka non-Bantu populations, which is more evident in the Northeast (Salvador), and one associated with central east African/Bantu populations, which is more present in the Southeast/South. This result broadens our understanding of the genetic structure of the African diaspora. Furthermore, we showed that positive assortative mating by ancestry is a social factor that permeates the demographic history of Brazilians and also, shapes their genetic structure, with implications for the design of genetic association studies in admixed populations. For instance, because mating by ancestry produces Hardy–Weinberg disequilibrium, filtering SNPs for genome-wide association studies based on the Hardy–Weinberg equilibrium conceals real aspects of the genetic structure of these populations. Finally, in Latin-American populations, the history of continental admixture rather than local demographic history is the main determinant of the burden of deleterious variants, although in a more complex fashion than previously thought. We speculate that future studies on populations from Northern Brazil (including large cities, such as Manaus, next to the Amazon forest) or the Central-West may reveal larger and different dynamics of Amerindian ancestry. Also, fine-scale studies on large urban centers from the Southeast and South of Brazil, such as Rio de Janeiro or Sao Paulo, that have been the destination of migrants from all over the country during the last decades, may show an even more diversified origin of Brazilians, including Japanese ancestry components, for instance, that we did not identify in our study. The EPIGEN Brazil initiative is currently conducting studies to clarify how the genetic variation and admixture interact with environmental and social factors to shape the susceptibility to complex phenotypes and diseases in the Brazilian populations.

Methods

Genotyping and Data Curation. Genotyping was performed by the Illumina facility using the HumanOmni2.5–8v1 array for 6,504 individuals and the HumanOmni5-4v1 array for 270 individuals (90 randomly selected from each cohort). After that, we performed quality control analysis of the data using Genome Studio (Illumina), PLINK (34), GLU (code.google.com/p/glu-genetics/), Eigenstrat (35), and in-house scripts. This study was approved by the Brazilian National Research Ethics Committee (CONEP, resolution 15895).

Whole-Genome Sequencing and Functional Annotation. We randomly selected 10 individuals from each of the three EPIGEN populations. The Illumina facility performed whole-genome sequencing of these individuals from paired-end libraries using the Hiseq 2000 Illumina platform. CASAVA v.1.9 modules were used to align reads and call SNPs and small INDELs (insertion or deletion of bases). Each genome was sequenced, on average, 42 times, with the following quality control parameters: 128 Gb (Gigabase) of passing filter aligned to the reference genome (HumanNCBI37_UCSC), 82% of bases with data quality (QScore) ≥30, 96% of non-N reference bases with a coverage ≥10×, a HumanOmni5 array agreement of 99.53%, and a HumanOmni2.5 array agreement of 99.27%. Functional annotation was performed with ANNOVAR (August 2013 release) with the refGene v.hg19_20131113 reference database in April of 2014. The nonsynonymous variants were predicted to be deleterious using CONDEL v2.0 (cutoff = 0.522) (28), which calculates a consensus score based on MutationAssessor (36) and FatHMM (37). These results were corrected for the bias reported in the work by Simons et al. (30), which evidenced that, when the human reference allele is the derived one, methods that infer deleterious variants tend to underestimate its deleterious effect (*SI Appendix*, section 8).

Relatedness and Inbreeding Analysis. We estimated the kinship coefficients for each possible pair of individuals from each of the EPIGEN populations using the method implemented in the Relatedness Estimation in Admixed Populations (REAP) software (13). It estimates kinship coefficients solely based on genetic data, taking into account the individual ancestry proportion from K parental populations and the K parental populations allele frequencies per each SNP. For these analyses, we calculated individual ancestry proportion and K parental populations allele frequencies per each SNP using the ADMIXTURE software (7) in unsupervised mode assuming three parental populations (K = 3). Inbreeding coefficients were also estimated for each individual using REAP. We represented families by networks, which were defined as groups of individuals (vertices) linked by kinship coefficient higher than 0.1 (edges).

***F* Statistics.** The F_{IS} statistic for each population is estimated as the average of the REAP inbreeding coefficients across individuals. For each SNP i and each population, we estimated the departure from Hardy–Weinberg equilibrium as $F_{IT(i)} = (He_i - Ho_i)/He_i$, where Ho_i and He_i are the observed and the expected heterozygosities under Hardy–Weinberg equilibrium for the SNP i, respectively. We estimated the population F_{IT} by averaging $F_{IT(i)}$ across SNPs. We estimated the F_{ST} for each SNP between the YRI and CEU populations using the R package hierfstat (38). The correlation between YRI vs. CEU F_{ST} and F_{IT} values for each SNP was calculated by the Spearman's rank correlation-ρ using the R cor.test function.

Population Structure Analyses. To study population structure, we applied (*i*) the ADMIXTURE method (7), increasing the number of ancestral clusters (K) that explains the observed genetic structure from K = 3, and (*ii*) PCA (35) (Figs. 1C and 3 and *SI Appendix*, section 6 and Figs. S4–S6). To study biogeographic ancestry, we have to exclude sets of relatives that could affect our inferences at within-continental level (24). We conceived and applied a method based on complex networks to reduce the relatedness of the analyzed individuals by minimizing the number of excluded individuals (*SI Appendix*, section 6.1). Applying this method, we created Dataset U, with 5,825 Brazilians, 1,780 worldwide individuals, and no pairs of individuals closer than second-degree relatives (REAP kinship coefficient >0.10) (*SI Appendix*, Table S13). We performed ADMIXTURE analyses with both the Original Dataset and Dataset U (*SI Appendix*, section 6 and Figs. S4 and S5).

PCA and ADMIXTURE analyses were performed with integrated datasets comprising the three cohort-specific EPIGEN working datasets and the public datasets populations described in *SI Appendix*, section 5. For the PCA and ADMIXTURE analyses, we used the SNPs shared by all of these populations, comprising a total of 8,267 samples and 331,790 autosomal SNPs (called the Original Dataset).

Analyses with X-chromosome data used only female samples from the Original Dataset. To perform such analyses, we integrated genotype data of shared SNPs from the X chromosome of EPIGEN female samples (from all three cohorts) and the X chromosome of female samples from the public datasets populations described in *SI Appendix*, section 5. This data integration yielded genotyping data with 5,792 SNPs for 4,192 females.

Local Ancestry Analyses. We inferred chromosome local ancestry using the PCAdmix software (19) and ∼2 million SNPs shared by EPIGEN (Original

Dataset) and the 1000 Genomes Project (*SI Appendix*, section 5.2). Considering our SNPs density, we defined a window length of 100 SNPs following the work by Moreno-Estrada et al. (27). PCAdmix infers the ancestry of each window. Local ancestry inferences were performed after linked markers ($r^2 >$ 0.99) were pruned to avoid ancestry misestimating caused by overfitting (4). We considered only the windows in which ancestry was inferred by the forward–backward algorithm with a posterior probability >0.90.

After local ancestry inferences, we calculated the lengths of the chromosomal segments of CSSA for each haplotype from each chromosome from each individual. The distribution of CSSA length was organized in 50 equally spaced bins defined in centimorgans and plotted for each population (Fig. 2*A*).

For the local ancestry analyses, we used phased data from the 1000 Genomes Project populations YRI and LWK (Africans) as well as CEU, FIN, GBR, TSI, and IBS (Europeans), Native-American populations Ashaninka and Shimaa [from the Tarazona–Santos group LDGH (Laboratory of Human Genetic Diversity) dataset], and the three EPIGEN populations (Original Dataset). The SHAPEIT software (39) was used to generate phased datasets.

We estimated admixture dynamics parameters using ABC. We used the model by Liang and Nielsen (20) to simulate CSSA distributions generated by a demographic history of three pulses of trihybrid admixture occurring 18–16, 12–10, and 6–4 recent generations ago conditioned on the observed admixture proportions of the EPIGEN populations. We inferred the posterior distributions of nine parameters $m_{n,P}$ (*SI Appendix*, section 6.8).

Lineage Markers Haplogroups Inferences. We performed mtDNA haplogroup assignments using HaploGrep (40), a web tool based on Phylotree (build 16) for mtDNAhaplogroup assignment. For Y-chromosome data, we inferred haplogroups using an automated approach called AMY tree (41). For Y-chromosome haplogroups, we considered the Karafet tree (42) and more recent studies to describe additional subhaplogroups. By these means, an updated tree was considered based on the information given by The International Society of Genetic Genealogy (ISOGG version 9.43; www.isogg.org).

ACKNOWLEDGMENTS. The authors thank David Alexander and Fernando Levi Soares for technical help and discussion and Rasmus Nielsen and Mason Liang for sharing their software for continuous specific ancestry simulations and feedback on its use. Centro Nacional de Processamento de Alto Desempenho em MG/Financiadora de Estudos e Projetos–Ministério da Ciência, Tecnologia e Inovação, Centro Nacional de Super Computação, and Programa de Desenvolvimento Tecnológico em Insumos para Saúde-Bioinformatics Platform at Fundação Oswaldo Cruz-Minas Gerais provided computational support. The EPIGEN Brazil Initiative is funded by the Brazilian Ministry of Health (Department of Science and Technology from the Secretaria de Ciência, Tecnologia e Insumos Estratégicos) through Financiadora de Estudos e Projetos. The EPIGEN Brazil investigators received funding from the Brazilian Ministry of Education (CAPES Agency), Brazilian National Research Council (CNPq), Pró-Reitoria de Pesquisa from the Universidade Federal de Minas Gerais, and the Minas Gerais State Agency for Support of Research (FAPEMIG).

1. Salzano FM, Freire-Maia N (1967) *Populações Brasileiras; Aspectos Demográficos, Genéticos e Antropológicos* (Companhia Editora Nacional, São Paulo, Brazil).
2. Giolo SR, et al. (2012) Brazilian urban population genetic structure reveals a high degree of admixture. *Eur J Hum Genet* 20(1):111–116.
3. Moreno-Estrada A, et al. (2014) Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344(6189): 1280–1285.
4. Eyheramendy S, Martinez FI, Manevy F, Vial C, Repetto GM (2015) Genetic structure characterization of Chileans reflects historical immigration patterns. *Nat Commun* 6:6472.
5. Barreto ML, et al. (2006) Risk factors and immunological pathways for asthma and other allergic diseases in children: Background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study). *BMC Pulm Med* 6:15.
6. Bacelar J (2001) *A Hierarquia sas Raças. Negros e Brancos em Salvador* (Pallas Editora, Rio de Janeiro).
7. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
8. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–1044.
9. 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(2):786–791.
10. Lima-Costa MF, Firmo JO, Uchoa E (2011) Cohort profile: The Bambui (Brazil) Cohort Study of Ageing. *Int J Epidemiol* 40(4):862–867.
11. Victora CG, Barros FC (2006) Cohort profile: The 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* 35(2):237–242.
12. Salzano FM, Bortolini MC (2002) *The Evolution and Genetics of Latin American Populations* (Cambridge Univ Press, New York).
13. Thornton T, et al. (2012) Estimating kinship in admixed populations. *Am J Hum Genet* 91(1):122–138.
14. Bittles AH (2002) Endogamy, consanguinity and community genetics. *J Genet* 81(3): 91–98.
15. Telles EE (2006) *Race in Another América: The Significance of Skin Color in Brazil* (Princeton Univ Press, Princeton).
16. Lima-Costa MF, et al.; Epigen-Brazil group (2015) Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci Rep* 5:9812.
17. Ruiz-Linares A, et al. (2014) Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* 10(9):e1004572.
18. Risch N, et al. (2009) Ancestry-related assortative mating in Latino populations. *Genome Biol* 10(11):R132.
19. Brisbin A, et al. (2012) PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* 84(4):343–364.
20. Liang M, Nielsen R (2014) The lengths of admixture tracts. *Genetics* 197(3):953–967.
21. Klein HS (2002) *Homo brasilis Aspectos Genéticos, Lingüísticos, Históricos e Socioantropológicos da Formação do Povo Brasileiro* (FUNPEC-RP, Ribeirão Preto, Brasil), 2nd Ed, pp 93–112.
22. Scliar MO, Vaintraub MT, Vaintraub PM, Fonseca CG (2009) Brief communication: Admixture analysis with forensic microsatellites in Minas Gerais, Brazil: The ongoing evolution of the capital and of an African-derived community. *Am J Phys Anthropol* 139(4):591–595.
23. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
24. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11(7):459–463.
25. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83(3):347–358.
26. Botigué LR, et al. (2013) Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci USA* 110(29):11791–11796.
27. Moreno-Estrada A, et al. (2013) Reconstructing the population genetic history of the Caribbean. *PLoS Genet* 9(11):e1003925.
28. González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88(4):440–449.
29. Lohmueller KE, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451(7181):994–997.
30. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. *Nat Genet* 46(3):220–224.
31. Lohmueller KE (2014) The distribution of deleterious genetic variation in human populations. *Curr Opin Genet Dev* 29:139–146.
32. Do R, et al. (2015) No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* 47(2):126–131.
33. Pena SD, et al. (2011) The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS ONE* 6(2):e17063.
34. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
35. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
36. Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8(11):R232.
37. Shihab HA, et al. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34(1):57–65.
38. Goudet J (2005) Hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol Ecol Notes* 5(1):184–186.
39. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2):179–181.
40. Kloss-Brandstätter A, et al. (2011) HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* 32(1):25–32.
41. Van Geystelen A, Decorte R, Larmuseau MHD (2013) AMY-tree: An algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. *BMC Genomics* 14(14):101–112.
42. Karafet TM, et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 18(5):830–838.

SUPPLEMENTARY MATERIAL

ORIGIN AND DYNAMICS OF ADMIXTURE IN BRAZILIANS AND ITS EFFECT ON THE PATTERN OF DELETERIOUS MUTATIONS

Fernanda S. G. Kehdy[a,1], Mateus H. Gouveia[a,1], Moara Machado[a,1], Wagner C. S. Magalhães[a,1], Andrea R. Horimoto[b] , Bernardo L. Horta[c] , Rennan G. Moreira[a] , Thiago P. Leal[a] , Marilia O. Scliar[a], Giordano B. Soares-Souza[a], Fernanda Rodrigues-Soares[a], Gilderlanio S. Araújo[a] , Roxana Zamudio[a] , Hanaisa P. Sant Anna[a] , Hadassa C. Santos[b] , Nubia E. Duarte[b], Rosemeire L. Fiaccone[d], Camila A. Figueiredo[e], Thiago M. Silva[f] , Gustavo N. O. Costa[f], Sandra Beleza[g], Douglas E. Berg[h,i], Lilia Cabrera[j] , Guilherme Debortoli[k] , Denise Duarte[l] , Silvia Ghirotto[m], Robert H. Gilman[n,o], Vanessa F. Gonçalves[p] , Andrea R. Marrero[k] , Yara C. Muniz[k] , Hansi Weissensteiner[q] , Meredith Yeager[r] , Laura C. Rodrigues[s] , Mauricio L. Barreto[f] , M. Fernanda Lima-Costa[t,2,3], Alexandre C. Pereira[b,2,3], Maíra R. Rodrigues[a,2,3], Eduardo Tarazona-Santos[a,2,3], and The Brazilian EPIGEN Project Consortium[4]

[4]The Brazilian EPIGEN Project Consortium includes: Neuza Alcantara-Neves[e], Nathalia M Araújo[a], Márcio LB Carvalho[u], Jackson Santos Conceição[f], Josélia OA Firmo[t], Denise P Gigante[d], Lindolfo Meira[v], Thais Muniz-Queiroz[a], Guilherme C Oliveira[w], Isabel O Oliveira[c], Sérgio WV Peixoto[t], Fernando A Proietti[t], Domingos C Rodrigues[u], Meddly L Santolalla[a], Agostino Strina[f], Camila Zolini[a]

[a] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, 31270-901, Belo Horizonte, Minas Gerais, Brazil;

[b] Instituto do Coração, Universidade de São Paulo, 05403-900, São Paulo, São Paulo, Brazil;

[c] Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, CP 464, 96001-970 Pelotas, RS, Brazil;

[d] Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, 40170-110, Salvador, Bahia, Brazil;

[e] Departamento de Ciências da Biointeração, Instituto de Ciências da Saúde, Universidade Federal da Bahia, 40110-100, Salvador, Bahia, Brazil;

[f] Instituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, Bahia, Brazil;

[g] Department of Genetics, University of Leicester, LE1 7RH, Leicester, United Kingdom;

[h] Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA;

[i] Department of Medicine, University of California, San Diego, CA 92093, USA;

[j] Biomedical Research Unit, Asociación Benéfica Proyectos en Informática, Salud, Medicina y Agricultura (AB PRISMA), 170070, Lima, Peru;

[k] Departamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Santa Catarina, Brazil;

[l] Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil;

[m] Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121, Ferrara, Italy;

[n] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA;

[o] Universidade Peruana Cayetano Heredia, 15102, Lima, Peru;

[p] Department of Psychiatry and Neuroscience Section, Center for Addiction and Mental Health, University of Toronto, Ontario M5T 1R8, Toronto, Canada;

[q] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, 6020, Innsbruck, Austria;

[r] Cancer Genomics Research Laboratory, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA;

[s] Department of Infectious Disease Epidemiology, Faculty of Epidemiology, London School of Hygiene and Tropical Medicine, WC1E 7HT, London, United Kingdom;

[t] Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, 30190-002, Belo Horizonte, Minas Gerais, Brazil;

[u] Laboratório de Computação Científica, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil;

[v] Centro Nacional de Supercomputação, Universidade Federal de Rio Grande do Sul, Porto Alegre, Brazil; and

[w] Grupo de Genômica e Biologia Computacional, CEBio, Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, Brazil

[1] These authors equally contributed to this article

[2] These authors equally contributed to this article

Corresponding author:

Eduardo Tarazona Santos

Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Email: edutars@icb.ufmg.br

Phone: 55 31 34092597

INDEX

1. POPULATION-BASED COHORTS

The Salvador-SCAALA (Social Changes, Asthma and Allergy in Latin America Program) project is a longitudinal study involving a sample of 1,445 children aged 4-11 years in 2005, living in Salvador, a city of 2.7 million inhabitants in Northeast Brazil. The population is part of an earlier observational study that evaluated the impact of sanitation on diarrhea in 24 small sentinel-areas selected to represent the population without sanitation in Salvador. Further details are available in Barreto et al.[5]. From these study participants, 1,309 were successfully genotyped as part of the EPIGEN Project (Genomic Epidemiology of Complex Diseases in Population-based Brazilian Cohorts).

The Bambuí cohort study of Ageing is ongoing in Bambuí, a city of approximately 15,000 inhabitants, in Minas Gerais State in Southeast Brazil. The population eligible for the cohort study consisted of all residents aged 60 years and over on January 1997, who were identified from a complete census in the city. From 1,742 Bambuí individuals older than 60 years (i.e. the eligible residents), 1,606 constituted the original cohort, and 1,442 (82.7% of the older residents) were successfully genotyped as part of the EPIGEN Project. Further details of the Bambuí study can be seen in Lima-Costa et al.[10].

The 1982 Pelotas Birth Cohort Study was conducted in Pelotas, a city in Brazil extreme South, near the Uruguay border, with 214,000 urban inhabitants in 1982. Throughout 1982, the three maternity hospitals in the city were visited daily and births were recorded, corresponding to 99.2% of all births in the city. The 5,914 live-born infants whose families lived in the urban area constituted the original cohort. From these, we have genome-wide data for 3,736 individuals. Further details are available in Victora and Barros[11].

2. DATA DESCRIPTION

The original datasets received from Illumina, as a result of 2.5M and 5M genotyping, were as follows: 2,379,855 SNPs for 6,504 individuals and 4,301,332 SNPs for 270 individuals. The 2.5M dataset was genotyped with the Illumina HumanOmni2.5-8v1 array and the 5M dataset was genotyped with the HumanOmni5-4v1 array. Both datasets contained individuals from the 3 cohorts, where 90 individuals from each cohort were randomly selected and genotyped for the 5M dataset. These 270 individuals are not present in the 2.5M dataset. All data were generated in the Illumina facility in San Diego (CA, US).

After extensive Quality Control (QC) procedures and filtering, the EPIGEN project has high quality genotyping data for a total of 6,487 Brazilian individuals.

To perform the genotyping analyses presented in this paper we used consensus datasets containing the shared SNPs between the 2.5M and 5M datasets. We also separated these consensus datasets into autosomal SNPs datasets, mitochondrial SNPs datasets, as well as X and Y chromosomal SNPs datasets. Each cohort has unique autosomal, mitochondrial, X and Y chromosomal datasets. Additionally, to allow ancestry and population structure analyses, we created a merged autosomal dataset from the autosomal datasets of the 3 cohorts to represent all EPIGEN data. This EPIGEN-autosomal dataset and the 12 cohort-specific datasets are described in the **EPIGEN Working Datasets Summary** section below.

2.1. EPIGEN Working Datasets Summary

Genotyping Data

Our genotyping data regards only SNPs and 1bp-INDELs. All analyses presented in this paper are based on 4 working datasets for autosomal SNPs, and 9 working datasets for Mitochondrial, X and Y chromosomal SNPs. All these datasets contain only consensus (shared) SNPs from the 2.5M and 5M datasets.

Summary of consensus-autosomal working datasets:

1. EPIGEN_2.5M_5M_autosomal (2,235,109 SNPs for 6,487 samples)
2. Salvador_2.5M_5M_autosomal (2,234,475 SNPs for 1,309 samples)
3. Bambui_2.5M_5M_autosomal (2,233,665 SNPs for 1,442 samples)
4. Pelotas_2.5M_5M_autosomal (2,234,985 SNPs for 3,736 samples)

Summary of consensus-X-chromosomal working datasets:

5. Salvador_2.5M_5M_X (46,906 SNPs for 1,309 samples)
6. Bambui_2.5M_5M_X (46,900 SNPs for 1,441 samples)
7. Pelotas_2.5M_5M_X (46,902 SNPs for 3,736 samples)

Summary of consensus-Y-chromosomal working datasets:

8. Salvador_2.5M_5M_Y (2,136 SNPs for 707 male samples)
9. Bambui_2.5M_5M_Y (2,115 SNPs for 562 male samples)
10. Pelotas_2.5M_5M_Y (2,144 SNPs for 1,873male samples)

Summary of consensus-mitochondrial working datasets:

11. Salvador_2.5M_5M_mitochondrial (216 SNPs for 1,308 samples)
12. Bambui_2.5M_5M_mitochondrial (213 SNPs for 1,442 samples)
13. Pelotas_2.5M_5M_mitochondrial (218 SNPs for 3,735 samples)

3. QUALITY CONTROL AND DATA CLEANING FOR GENOTYPING DATA

Quality control and data cleaning procedures start with the **Illumina SNP-Array Quality Control** and the **Data Export** steps. After that, a number of standard procedures are applied to the EPIGEN datasets, as described next in section **Data Cleaning and Quality Control**.

3.1. Illumina SNP-Array Quality Control

According to the Illumina's genotyping report, the 2.5M dataset has the following quality control parameters: Locus Success Rate (99.21%), Genotypes - Call Rate – (99.71%), and Reproducibility (99.99%). The 5M dataset has the following parameters: Locus Success Rate (98.87%), Genotypes - Call Rate – (99.81%), and Reproducibility (100.00%).

3.2. Data Export

Genotyping data for both 2.5M and 5M EPIGEN datasets were exported from Genome Studio as PED and MAP format files using the same Illumina plugin with the following parameters: (i) "UseForwardStrand" set to "True", and (ii) remove SNPs that have no signal. As a result, 18,762 SNPs were removed from the 2.5M dataset and 48,815 SNPs from the 5M dataset.

3.3. Data Cleaning and Quality Control

The QC and data cleaning processes for genotyping data were performed in four steps: (Step 1) initial data cleaning of the 2.5M and 5M datasets separately, where basic data filters and strand check procedures were applied; (Step 2) separation of autosomal, mitochondrial as well as X and Y chromosome SNPs into distinct datasets and posterior integration in four 2.5M-5M consensus datasets; (Step 3) QC and data cleaning of the consensus 2.5M-5M autosomal SNPs dataset; and (Step 4) QC of the mitochondrial, as well as X and Y chromosome SNPs datasets. Each of these steps is detailed next.

Step1: 2.5M and 5M Datasets

For the data cleaning of the 2.5M and 5M datasets, the following filters were applied: removal of SNPs with zeroed (missing) chromosome (Filter 1), and removal of repeated SNPs (Filter 2). A summary is presented in Table S1. For the removal of repeated SNPs (Filter 2), first the Illumina's "kgp" SNP identifiers were replaced by the updated correspondent "rs" identifiers, provided by Illumina. After that, SNPs with the same physical position but different identifiers in the same dataset were considered as duplicated, and for each set of duplicated SNP, those with lower call rate were removed from their respective datasets. In this step of the data cleaning, we also corrected possible strand flips in both datasets using the software PLINK[34].

Step 2: Autosomal Datasets Separation and 2.5M-5M Consensus

After the initial filtering of the 2.5M and 5M datasets, we separated the autosomal from the mitochondrial and sex-chromosome SNPs in each dataset. A summary is shown in Table S2. Next, we combined the 2.5M and 5M Autosomal and Mit/X/Y datasets into one 2.5M-5M autosomal dataset and one 2.5M-5M Mit/X/Y dataset with consensus SNPs. This resulted in a consensus autosomal dataset with 2,256,647 SNPs, and a consensus Mit/X/Y dataset with 49,709 SNPs (Table S2).These datasets contain the shared SNPs between the 2.5M and 5M datasets. Since there was no sample filtering in this step of the data cleaning, the total number of samples in the consensus datasets at this point is 6,774.

Step 3: Consensus Autosomal SNPs and Samples

Since we are working with a consensus autosomal dataset, we first perform data cleaning procedures to verify and guarantee consistency between the SNPs in the 2.5M and 5M datasets. These include allele frequency checks and possible strand flip checks. From these analyses, we concluded that there were inconsistencies between the two arrays manifests due to strand flip for a number of SNPs. Particularly, we found a list of 21,624 SNPs that have both allele frequency and genotype (possible strand flip) inconsistencies. Therefore, we excluded the 21,624 SNPs from the consensus datasets (as shown in Table S3).

After that, standard QC procedures were performed for autosomal SNPs, separately for each cohort. The initial consensus autosomal-SNPs dataset had 2,256,647 SNPs and 6,774 samples (Table S2). We start by describing the sample filtering process and then the SNP filtering, as follows.

To evaluate samples, 3 filters were used: the filter –mind 0.1 from the PLINK software, to evaluate the rate of genotypic loss per individual, which eliminated a total of 214 individuals with more than 10% of missing data (Filter 1); check sample duplicates, which preserved samples with the highest call rate among duplicates, and removed a total of 68 samples (Filter 2); and the sex check filter which removed 5 individuals (Filter 3). This is detailed in Table S4.

Autosomal SNPs were evaluated with the filter –geno 0.10 from PLINK, applied to evaluate the rate of genotypic loss per marker (Filter 4). The MAF and Hardy-Weinberg equilibrium filters were not applied. Because we are working with admixed population-based cohorts, some level of internal subdivision may exist, and filtering on a customary cutoff of 10^{-4}, may conceal aspects of the genetic structure of these populations. After that, the datasets from the three cohorts were merged with PLINK, recreating the autosomal dataset with 2,256,636 SNPs and 6,487 individuals (Tables S3 and S4). Finally, the list of 21,624 SNPs identified earlier in data cleaning procedure as inconsistent were removed from all 4 datasets (Filter 5). Note that the number of SNPs excluded with the latter filter varies according to the intersection of the SNP list with each dataset. A summary is shown in Table S3.

Step 4: Consensus Mitochondrial, X and Y SNPs

Quality control for mitochondrial, X and Y chromosomal SNPs was performed separately for each cohort. From the initial 49,709 SNPs (Table S2), 46,945 are X-chromosomal SNPs, 2,153 are Y-chromosomal SNPs, 220 are mitochondrial SNPs, and 391 are pseudo-autosomal SNPs that were removed from our datasets. As before, SNPs were evaluated with the filter –geno 0.10 from PLINK (see the Excluded columns in Table S5). The MAF and HWE filters were not applied.

Regarding samples, we maintained the same list of individuals from Table S4 as the starting point, in order to achieve comparable datasets sample size, and further applied the filter –mind 0.1 from PLINK. The results are shown in Table S5.

The complete data cleaning and QC processes resulted in 4 working datasets for autosomal SNPs, and 9 working datasets for Mitochondrial, X and Y chromosomal SNPs (Table S6). These are the working datasets used in all analyses presented in this paper. Importantly, all datasets contain only consensus (shared) SNPs from the 2.5M and 5M datasets. These are exactly the same datasets that are on Section 2.1 Working Datasets Summary.

4. RELATEDNESS AND INBREEDING IN THE EPIGEN COHORTS

4.1. Relatedness

To assess the family structure, we estimated the kinship coefficients (Φ_{ij}) for each possible pair of individuals from each of the EPIGEN populations. The kinship coefficient Φ_{ij} is the probability that two alleles at a locus, randomly picked from individuals i and j, are identical by descent (IBD). We estimated kinship coefficients using the method implemented in the REAP software (Relatedness Estimation in Admixed Populations[13]). It estimates kinship coefficients solely based on genetic data, taking into account the individual ancestry proportion (IAP) from K parental populations and the K-parental populations allele frequencies per each SNP (KAF). For these analyses, we calculated IAP and KAF using the ADMIXTURE software assuming three unsupervised parental populations (K = 3, see Section 6 below for details). REAP estimation of kinship coefficients improve when larger numbers of unlinked SNPs are used[13] Assuming the EPIGEN populations as tri-hybrid, we considered the following K=3 parental samples for ADMIXTURE analysis: 174 CEU (European) and 176 YRI (African) from the HapMap Project and 89 Peruvian Native Americans (Shimaa, N=45 and Ashaninkas, N=44) from our laboratory database (Tarazona-Santos´ group LDGH), reaching 994,151 SNPs shared with all three EPIGEN populations. REAP also estimates the probability that two individuals i and j, share 0, 1 or 2 IBD alleles (δ_{ij}^{0},$\delta_{ij}^{1}$and δ_{ij}^{2}, respectively), and for each admixed individual i in the sample, it estimates the inbreeding coefficient h_i^A, which is the probability that the two alleles at a locus within an individual are IBD.

To provide a visual comparison of relatedness in the EPIGEN populations, we plotted the combination of theoretical values of Φ_{ij} and δ_{ij}^{0} for different pairs of relatives (Fig. S1A). Next, keeping in mind these theoretical values, we can envisage the level of relatedness in each EPIGEN cohort by plotting, for all pairs of individuals i and j, the kinship coefficient Φ_{ij} on the vertical axis and δ_{ij}^{0} (on the horizontal axis (Figures 1A and Figs. S1C, S1E and S1G). We established a "family"-kinship coefficient threshold $\Phi_{ij} \geq 0.1$ to consider individuals as related or not. This threshold allows us to consider as related: first-degree relatives (pair offspring and full siblings) and second-degree relatives (uncle/aunt, nephew/niece, grandparent/grandchild or half-sibling).

Bambuí is the unique among the studied cohorts that includes individuals with a wide range of age (over 60 years). We verified if its high level of family structure was an effect of its age structure. Even after excluding all pairs of related individuals ($\Phi_{ij} \geq 0.1$) with more than 5 years of difference in age, Bambuí continued showing the highest family structure level among the EPIGEN populations (429 pairs of individuals with $\Phi_{ij} \geq 0.1$ vs. 65 in Salvador and 95 in Pelotas).

4.2. Relatedness representation using a networks

To visualize the family structure of the EPIGEN populations we clustered individuals into family groups using a network approach. To do that, we model the families within each cohort like a network, where each node is an individual who connects to others by edges, that represent kinship coefficients higher than the threshold of 0.1 (Figs. S1B, S1D and S1F). We observed that Bambuí has the most conspicuous family structure with 266 families of up to 25 individuals, followed by Pelotas with 80 families of up to 5 individuals and Salvador with 61 families with up to 3 individuals. Based on these results, we represent in Figs. 1C1, 1E1 and 1G1 the inferred family size distributions in Salvador, Bambuí and Pelotas, respectively.

4.3. Consanguinity

For each individual of the studied cohorts we estimated the inbreeding coefficient h_i^A using the REAP software[13], that perform this estimation conditioning on individual admixture. Fig. S2 shows that for Salvador and Pelotas, inbreeding coefficients are centered on 0, which suggest a negligible level of inbreeding in these populations. Conversely, the highest inbreeding coefficients are observed in Bambuí.

4.4. Association between excess of observed homozygosity and ancestry

For these analyses, we constructed a dataset with the SNPs shared by the following five populations: one African population (YRI, N=88), one European population (CEU, N=85), both from the 1000 Genomes Project, and the three populations of this study.

To investigate the association between homozygosity excess and ancestry, we estimated the F_{ST} for each SNP between the YRI and CEU populations as a measure of how these SNPs are differentiated between the two main ancestry sources of the Brazilian population. We used the R package hierfstat[38] to estimate the F_{ST}. Then, we estimated the F_{IT} for each SNP for each cohort as a measure of homozygosity excess. We used GLU to calculate the observed and expected-under-Hardy-Weinberg heterozygosities (Ho and He, respectively) and then estimated F_{IT}= (He - Ho) / He. We estimated the Spearman's rank correlation rho using the cor.test function in R, to test if there was an association between the F_{ST} and F_{IT} values (Fig. S3).

To verify possible genotyping errors in our data, we plotted the F_{IT} distribution for each cohort (Fig. S3) and identified the allele frequencies of SNPs with extreme F_{IT} values > 0.6. We observed that most of these SNPs are rare, having minor allele frequency (MAF) less than 0.01. When an allele has a MAF<0.01, a small difference between the expected and observed numbers of heterozygous is enough to have a F_{IT}>0.60. For instance, for a MAF=0.001 in a sample of 1000 individuals, 1 observed and 2.5 expected heterozygous would produce F_{IT}=0.60. However, some SNPs with high F_{IT}>0.6 show MAF higher than 0.01, posing the possibility of genotyping errors. Therefore, for the correlation tests and plots presented here the latter list of SNPs were removed from the working datasets.

Considering that the Bambuí cohort has many related individuals, we removed the 516 related individuals (see Section 6) and repeated the F_{IT} vs. F_{ST} analysis. The results were very similar to the first analyses, showing a mean F_{IT} of 0.015 (as opposed to F_{IT} = 0.016 with related individuals) and rho = 0.16 (as opposed to rho = 0.18 with related individuals).

5. DATA INTEGRATION (EPIGEN AND PUBLIC DATASETS)

5.1. From Public HapMap, HGDP and 1000 Genomes Project Data to Frozen Datasets

Public data from the HapMap project[43], 1000 Genomes Project[44] and Human Genome Diversity Project (HGDP)[45] were used together with the EPIGEN datasets (in PED/MAP formats) in the form of a frozen dataset (also in PED/MAP format).

HapMap Project Datasets

We downloaded all .hapmap (phases II + III) files for all chromosomes and for the mtDNA from all available populations (at ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/latest_phaseII+III_ncbi_b36/forward/non-redundant/). These datasets were then converted to PED/MAP files. At the end of this step, we obtained 275 pairs of files (PED/MAP) representing 11 HapMap populations and 22 autosomes, sexual chromosomes and mtDNA. This generated 11 files (one per population). Table S7 shows the number of individuals and SNPs in the final HapMap frozen datasets.

HGDP Datasets

HGDP data is available in a single dataset comprising all populations and chromosomes from http://hagsc.org/hgdp/files.html. We identified and excluded SNPs with missing data for all individuals, obtaining 52 PED/MAP files, one for each population. The number of individuals and SNPs in each of these files (datasets) is shown in Table S8.

1000 Genomes Project Datasets

The 1000 Genomes project phase I data, version v3.20101123.snps_indels_svs.genotypes, are available in separate files for each chromosome, in VCF format (Variant Call Format)[46]. We only downloaded for each autosomal chromosome, SNPs that are shared with the EPIGEN dataset (see Section 2.1). As a result, we obtained new VCF files separated by chromosomes.

After filtering, the new VCF files for each autosomal chromosome were converted to PED/MAP files. These files were then merged, resulting in a dataset containing the shared autosomal SNPs with the EPIGEN autosomal dataset for all 1000 Genomes populations. The total number of SNPs and the 1000 Genomes populations are described in Table S9.

Phased 1000 Genomes Datasets

We also used phased data from 1000 Genomes Project phase I v3.20101123 snps_indels_svs.genotypes.nomono.haplotypes /.legend, comprising all populations and all autosomal chromosomes. These datasets, separated by chromosomes, were downloaded in shapeit format from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/shapeit2_phased_haplotypes/ and stored in our server. The number of individuals in each population and of SNPs in each chromosome for these phased data are presented in Tables S9 and S10.

5.2. Integrating Public Datasets with EPIGEN Datasets

Data Integration for PCA and ADMIXTURE Analyses

PCA and ADMIXTURE analyses (see Section 6 below) were performed with integrated datasets, comprising the 3 cohort-specific EPIGEN working datasets (Section 2) and the following public datasets populations: ASW, CEU, MEX/MXL, JPT, LWK, TSI and YRI (from HapMap and 1000 Genomes project datasets); CLM, FIN, GBR, IBS and PUR (from 1000 Genomes project datasets); Tuscan, French, French Basques, Sardinian, North Italian, Orcadian, Russian, Adygei, Yoruba, Bantu, Mandenka, Colombians, Pima, Maya, Surui, Karitiana, Japanese, Bedouin, Druze, Mozabite, and Palestinian (from HGDP datasets); and Peruvian Ashaninka and Shimaa (Native Americans) populations from Tarazona-Santos´ group, genotyped for the same 2.5 Omni array. We used for the PCA and ADMIXTURE analysis the SNPs shared by all these populations.

At the end, we obtained a single dataset, in PED/MAP format, containing 8,267 samples and 331,790 autosomal SNPs. Tables S11 and S12 summarize the number of individuals per population and of SNPs per chromosome (Original Dataset in the Main Text).

To avoid the bias caused by family structure in our population structure analyses, we excluded from Original Dataset (Tables S11 and S12) related samples that were identified by our methodology creating a new dataset, Dataset U (where U stand for Unrelated, see Section 4.1 for relatedness identification and Section 6 for the exclusion method). The number of individuals that were excluded from and kept in each cohort is described in Table S13.

Analyses with X-chromosome data used only female samples. To perform such analyses we integrated genotype data of shared SNPs from the X-chromosome of EPIGEN female samples (from all three cohorts) and the X-chromosome of female samples from the following public datasets populations: ASW, CEU, MEX/MXL, JPT, LWK, TSI and YRI from HapMap and 1000 Genomes; CLM, FIN, GBR, IBS and PUR from 1000Genomes and Tuscan, French, French_Basque, Sardinian, North_Italian, Orcadian, Russian, Adygei, Yoruba, Bantu, Mandenka, Colombians, Pima, Maya, Surui, Karitiana, Japanese, Bedouin, Druze, Mozabite from HGDP. The X chromosome data of HapMap and HGDP populations were extracted from our frozen datasets, while data from female samples of the 1000 Genomes were downloaded separately for these analyses. The above data integration yielded genotyping data with 5,792 SNPs for 4,192 female samples, as detailed in Table S14.

Data Integration for Tri-Hybrid Local Ancestry Analyses

For the local ancestry analyses we used phased data from 1000 Genomes Project populations YRI and LWK (Africans) and CEU, FIN, GBR, TSI and IBS (Europeans), from Native Americans populations Ashaninka and Shimaa (from Tarazona-Santos group LDGH dataset), and from the 3 EPIGEN populations (Original Dataset).

The SHAPEIT software[39] was used to generate phased datasets. The polymorphic shared SNPs between 1000 Genomes African and European populations, Native Americans and the EPIGEN cohorts were used for the local analyses. At the end, we obtained for each chromosome, 6 datasets: Africans (YRI + LWK) with N=185, Europeans (CEU + FIN + GBR + TSI + IBS)with N=379, Native Americans (Shimaa + Ashaninka) with N=89 and the three EPIGEN populations: Bambuí (N=1,442), Pelotas (N=3,736) and Salvador (N=1,309).The number of SNPs per chromosome used in the local ancestry analyses are described in Table S15.

6. POPULATION STRUCTURE

The Principal Component Analysis (PCA) was applied using EIGENSOFT 4.21[35]. We ran ADMIXTURE[7] to explore global patterns of population structure between two subsets of data: the Original Dataset with all samples (including related EPIGEN samples), and Dataset U (unrelated, see Section 5.2). We always ran ADMIXTURE in unsupervised mode, which estimates individual ancestry values solely using information from the included genotypes, without any information about which individuals belong to which population. All ADMIXTURE analyses were repeated 4 times using binary input files and different random seed numbers, and in all cases results were highly correlated.

To arrive to a dataset with only unrelated samples (Dataset U) we need to reduce the level of family structure of the Bambuí cohort. To do that without eliminating all families, we implemented a network-based approach that aims at eliminating the smallest possible number of individuals (see description in Section 6.1). We applied our method to the EPIGEN populations datasets to generate Dataset U (with only unrelated EPIGEN individuals). As a result, 63 (of 125 relatives), 516 (of 886 relatives) and 83 (of 169 relatives) individuals were removed from the Salvador, Bambuí and Pelotas cohorts, respectively.

In summary, the Original Dataset is composed by 6,487 individuals from the EPIGEN populations, including relatives plus 1,780 individuals from our integrated public dataset and 331,790 autosomal SNPs (see Section 5.2). Dataset U is composed of 5,825 individuals from the EPIGEN populations without related individuals (after the exclusion previously presented, based on family structure) plus 1,780 individuals from our integrated public dataset and the same autosomal SNPs as the Original Dataset (Section 5.2). Dataset U was the main dataset used to study the population structure of the EPIGEN cohorts.

ADMIXTURE results were shown by barplots (Figs. S4A and S5) where each bar represents an individual and the colours represent the proportion of each inferred ancestry. We ran ADMIXTURE from K = 2 to K = 15 for the Dataset U (Fig. S4A), and from K = 3 to K = 10 for the Original Dataset (Fig. S5). Using ADMIXTURE's cross-validation procedure we found that K = 6 has the lowest predicted error (Fig. S4B).

Based on the results of ADMIXTURE with K=3 and from the Principal Components 1 and 2 (PC1 and PC2), we were able to differentiate the main continental parental groups that contributed to the formation of the Brazilian population: Europeans, Africans and Native Americans (Figs. 1B and 1C, Figs. S4A, and S6 (A, D and G). The Salvador cohort presented a mean proportion of 0.43 continental European ancestry while for the Bambuí and Pelotas cohorts the values were 0.77 and 0.76, respectively. Regarding the continental African ancestry, the Salvador, Bambuí and Pelotas cohorts presented mean proportions of 0.50, 0.16 and 0.16, respectively. The mean proportion of continental Native American ancestry were similar and low for all EPIGEN cohorts: 0.06, 0.07 and 0.08 in Salvador, Bambuí and Pelotas, respectively. Also, ADMIXTURE analysis with K=4 identifies the Japanese individuals from HapMap and 1000 Genomes, but none of the Brazilian individuals showed a relevant contribution from this ancestry cluster.

6.1. Network-based method for reducing family structure

We designed and implemented a node selection algorithm based on node centrality degree statistics. This statistic was calculated using the last version of the software NetworkX (https://networkx.github.io/). The degree centrality for a node *v* is the fraction of nodes that it is connected to.

The network is generated with all individuals from a given cohort represented as nodes. Links between nodes are established if the kinship value between these nodes (individuals) is higher than the 0.1 kinship threshold ($\Phi_{ij} \geq 0.1$, for two individuals *i* and *j*). Therefore, clusters of connected nodes in the network indicate families. The goal of the algorithm is to eliminate these clusters by removing the smallest possible number of nodes (i.e. individuals), thus creating a totally disconnected network (or an edgeless network). To do that, our algorithm works in two steps. First, we iteratively (i) calculate the nodes centrality degree and (ii) eliminate those with highest centrality degree (or the most central nodes), until only pairs of nodes (like families with only two individuals) and/or unconnected nodes (or nodes with zero centrality degree) remain in the network (N1).

The second step consists of disconnecting pairs of nodes that remained in N1 from the first elimination round (the first step). This is necessary to guarantee that the final network is totally disconnected. To decide the best individual to be eliminated from each pair, we look at the individuals with a smaller degree of kinship relations. This is done by creating a new network (N2), but this time with node connections with kinship values smaller than the original threshold (0.1). These new node connections must also have a kinship value higher than 0.03, which is the minimum value for related individuals (thus, $0.03 < \Phi_{ij} \leq 0.1$). Having this new network, we calculate the degree centrality of each node in the pair. The node with highest degree centrality is eliminated from N1. At the end of this step, we have a final network, N1, with only unconnected nodes.

6.2. European ancestry in the Brazilian population

ADMIXTURE analysis with K=5 identifies European-Middle East substructure, and in fact, new clusters appear associated with Europe (Fig. S4A in red) and Middle East/Southern Europe (Fig. S4A in purple). With K=7, the purple Middle East cluster is further separated, generating a cluster more associated only with Middle East (Fig. S4A in magenta), and a Southern European-associated cluster (purple). For the sake of readiness, hereafter we call these geographically-associated ancestry clusters obtained with K=7, simply as North European (red), Middle East (magenta) and South European (purple) clusters, even if we make clear that these associations are of course not absolute, in the sense that most European and Brazilian individuals share variable percentages of each cluster. This substructure is also visualized by the PCA, where the distribution of North European, South European and Middle Eastern populations is captured by the Principal Component 4 (Fig. S6B, S6E and S6H and Figure 3B).

The Salvador population presented a mean proportion of 0.43 of the total European ancestry, while for the Bambuí and Pelotas cohorts the values were 0.77 and 0.76, respectively (Fig. S4A, K=3, red color). When analyzing the mean proportions for the sub-continental clusters of European ancestry in the Salvador population corresponding to K=7, we find values of 0.16, 0.23 and 0.04 for North European, South European and Middle Eastern clusters, respectively. For the Bambuí population, these values were 0.275, 0.425 and 0.068, and for the Pelotas population 0.307, 0.402 and 0.054, respectively (Fig. S4A, K=7).

Our results indicate a higher mean proportion of North European ancestry in the south of Brazil (40.2% of Pelotas European ancestry), in comparison to the Southeast (Bambuí, 35.8% of the European ancestry) and the Northeast (Salvador, 36.7% of the European ancestry). Consistently, the European ancestry of some Pelotas individuals matches very well that of some North European individuals (Figure 3A and 3B, K=7 and PCA plot).

In addition, the Principal Components Analysis allowed the separation of Europe in East and West (Figure 3B, PC6), while this substructure was not detected by ADMIXTURE analysis using a range of K=3 to K=15. The resemblance of most Brazilians with Southwest European individuals is consistent with its predominant Iberic colonization.

With K=8 we verify another European ancestry cluster (cyan) with its highest mean proportion in Bambuí (0.225), in comparison to Salvador (0.064) and Pelotas (0.064), (Figs. S4A). We observed that this cluster has a South European origin, since its highest mean proportions are in Sardinian (0.16), French Basque (0.16) and Iberian Spanish (0.14) populations. The South European origin of this cluster is confirmed when analyzing the distributions of the Northern and Southern clusters mean proportions in Bambuí throughout the analyses with different Ks. When comparing the mean proportions of the Northern and Southern European clusters in Bambuí for analysis with K=7 and K=8 (where the cyan component appears), we verify a more marked decrease in the Southern European cluster (0.134) when compared to the Northern European cluster (0.075). This suggests the Southern European origin of this cyan cluster.

A possible explanation for the high mean proportions of this cluster in Bambuí is a founder effect due to the small size of the Bambuí population, and therefore, more subject to genetic drift. Genetic drift was quantified through a genetic distance analysis (F_{ST}) between the different clusters (Ks) generated in analysis with K=8 (Table S16). We observed a F_{ST}=0.029 between the cyan cluster and the other both European clusters (North [red] and South [purple]). This differentiation is similar to the difference found between North and South of Europe (F_{ST}=0.030), and higher than the observed between the East and West Africa clusters with K=9 (F_{ST}=0.019, see Section 6.3 below).

With K=10 we observe a cluster with higher mean proportions in Bambuí (0.25) and Pelotas (0.30) than in Salvador (0.15) (Fig. S4A, K=10, grey color). This cluster also appears in all European populations, with its highest values in French (0.34), British in England and Scotland (GBR) (0.32) and Sardinian (0.32). This cluster appears in high proportions (>80%) in some Brazilian individuals, mainly from Pelotas, while no European individuals show this proportion of the grey cluster.

To evaluate the robustness of our results regarding European ancestry, we reproduced PCA and ADMIXTURE analyses with a different dataset including more individuals but a reduced number of 44,901 shared SNPs. We selected from Dataset U EPIGEN individuals with more than 50% of whole European ancestry (measured by K=3, Original Dataset, in ADMIXTURE analysis), merged them with POPRES (Population Reference Sample) European individuals (from Albania, Austria, Belgium, Bosnia, Bulgaria, Croatia, Cyprus, Czech Republic, England, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Macedonia, Netherlands, Norway, Poland, Portugal, Romania, Russia, Scotland, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, Yugoslavia[25]), with individuals (from Canary Islands, Spain_NW and Spain_S)[26] and with HapMap, 1000 Genomes and HGDP individuals from our Dataset U. We confirmed the patterns of ancestry observed with our Dataset U. In particular, the Bambuí associated cluster (the cyan cluster in K=8 on the Dataset U analysis), appears at K=7 in the EPIGEN-POPRES-Botiguè analysis, and is also more prevalent in Southern European populations such as Portugal-POPRES, and Spanish/Canary Island from POPRES and[26].

6.3. African Ancestry in the Brazilian population

Our worldwide dataset for comparison includes four African populations belonging to the Niger-Kordofanian linguistic macro-family, the most spread in South-Saharan Africa. Two of them are Bantu-speaking, namely, Luhya from Kenya and the scattered HGDP-Bantu from Southeastern Africa. The former descend from the large spread of farmers from near the Nigerian/Cameroon highlands across eastern and southern Africa within the past 5000 to 3000 years[8]. The other two populations included in the analysis are non Bantu-speaking populations from Western Africa, Yoruba and Mandenka, which are known for their high contribution to the African diaspora to Brazil and USA.

We detected two within-Africa ancestry clusters in the current Brazilian population (Figure 3C, K=9): The blue cluster, associated with the Yoruba/Mandenka non-Bantu Western populations; and the mustard cluster, associated with the Luhya/HGDP Bantu populations from Eastern Africa.

To verify which Principal Component better differentiates these East/Bantu and West/non-Bantu groups, we performed a correlation analysis between the values of each PC (PC10 and PC11) for each African individual and the logarithm of the ratio of mustard/blue contributions percentage (calculated from ADMIXTURE analysis with K=9). We found that PC10 and PC11 capture the African sub-continental differentiation evidenced by ADMIXTURE with K=9 (Figs. 3D and S7).

The Salvador, Bambuí and Pelotas cohorts presented, respectively, 0.50, 0.16 and 0.16 mean proportion of global African ancestry (Figs. 1B and Fig. S4A, K=3). The mean sub-continental proportions for the mustard cluster (East Africa/Bantu associated, EAFR) and blue cluster (West Africa /non-Bantu associated, WAFR) of the three Brazilian populations and the Afro-American population ASW from HapMap are in Table S17. To verify the different African contributions in different Brazilian regions, we calculate the ratio between the means of blue (WAFR) and the mustard (EAFR) clusters (Blue/Mustard) for the three Brazilian cohorts and for Afro-Americans (ASW). Blue/Mustard ratios are 4.85 for ASW, 3.00 for Salvador, 1.79 for Bambuí and 1.30 for Pelotas. Thus, there is a higher proportion of the mustard-EAFR cluster in Southeast and Southern Brazil, respect to Northeast.

To verify whether the percentage of the individual total African ancestry influenced ADMIXTURE estimates of sub-continental clusters of African ancestry, we performed ADMIXTURE analyses only with individuals showing more than 50% total African ancestry, previously inferred with ADMIXTURE by K=3, using the same parental populations. By performing this analysis, the same two clusters of African sub-structure were detected by K=7, and we estimated the ratio of the corresponding individual blue (West) to mustard (East) ancestry proportions. We compared the logarithms of the individuals Blue/Mustard ratios with values of the same variable, estimated for the same individuals in the global analysis (i.e., that incorporated the whole set of individuals), using correlation analysis. We found a strong correlation between the two estimates (r^2=0.97, p<2.2e-16, Fig. S8), revealing that the individual total African ancestry does not influence ADMIXTURE power to infer African sub-structure.

To verify whether the distribution of African sub-continental ancestry depends on the total African ancestry, we estimated the correlation between the logarithm of the individuals Blue/Mustard ratios and the total African ancestry of individuals for each Brazilian population. We observed a significant correlation between the log (Blue/Mustard) and the global African ancestry for Salvador and Pelotas (r^2=0.22, p<1.2e-14 for Salvador and r^2=0.14, p<2.2e-16 for Pelotas, Figs.S9A and S9C). For Bambuí the correlation was not significant (r^2=0.0014, p<0.96, Fig. S9B). Therefore, in Salvador and Pelotas, the West African cluster of ancestry is more present in individuals with more total African ancestry.

6.4. Native American ancestry in the Brazilian populations

Considering the low contribution of Native Americans to Brazilians, we do not further analyze the genetic structure of Native American ancestry clusters.

6.5. Clusters of relatives identified with ADMIXTURE with the Original Dataset.

In the ADMIXTURE analysis performed with the Original Dataset (before the exclusion of related individuals), we identified clusters (by K=7 to K=10) that are associated to groups of individuals that match those identified by the REAP kinship analysis as relatives (Fig. S5). For instance, by K=7 we observed two clusters that are highly associated with two Bambuí families inferred by REAP (Fig. S10, brown and black clusters). In particular, all individuals with more than 80% of the black cluster belong to a unique family of 25 individuals identified by REAP (Fig. S10). The second Principal Component obtained only with the entire Bambuí cohort also separates these related individuals (Fig. S10). Moreover, individuals from this family (i.e. belonging to the black cluster) have higher inbreeding coefficients than the entire Bambuí cohort (mean 0.042 vs. 0.012, p= 0.048 by Wilcoxon Signed-Rank test), which suggest that recurrent consanguineous marriages may be associated to specific families.

We did not detect, through ADMIXTURE and PCA analyses, any family structure in the Pelotas and Salvador cohorts, in agreement with the REAP results (Section 4).

These results evidence that family structure can be a confounding factor in studies of population structure. With enough data, ADMIXTURE and PCA analyses interpret familiar structure as ancestry clusters, if families include enough individuals.

6.6. Population structure inferred by X-chromosome data

X-chromosome diversity data are more associated to the demographic history of women, because X-chromosomes spend 2/3 of their evolutionary history in females, and only 1/3 of it in males. We applied: (i) Principal Component Analysis (PCA, EIGENSOFT 4.21, see section PCA), and (ii) unsupervised ADMIXTURE analysis by K=3, to the diploid X-chromosome data for 5,792 SNPs of 4,192 EPIGEN females (see section 5.2, Table S14). For comparison, we reanalyzed autosomal data extracted from the Original Dataset, including the same females individuals present in the X-chromosome dataset (4,192 female samples for 331,790 autosomal SNPs).

We observed that: (i) The distributions of individuals on the PC1 vs. PC2 space (the only informative clustering pattern for X-chromosome), suggest differences in the evolutionary history of males and females. For the three EPIGEN populations, we observed that compared with autosomal data (Fig. S11), a larger number of females X-chromosome cluster near the Native American and African parental populations (Fig. S11). This is consistent with the lower effective recombination rate of the X-chromosome[47], that result in a large number of X-chromosomes with a unique continental ancestry. This differential pattern between X-chromosome and autosomal markers is not evident for European ancestry, because it is the predominant continental ancestry in our sample, and therefore there is a high number of individuals with both high autosomal and X-chromosome European ancestry. (ii) Both PCA and ADMIXTURE analyses show that compared with autosomal data, the X-chromosome show a larger Native American and African contribution to extant Brazilian genomic diversity than at genome-wide level (Figs. S11 and S12A). This is due to a historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry. This pattern of mating is well documented in demographic and genetic studies across all Latin America[12]. (iii) On average, the sex-bias in admixture was larger in Salvador, and lower in Bambuí and Pelotas, and it was higher for Native American ancestry than for African ancestry (Table S18 and Figs. S12B and S12C).

6.7. European, African and Native American Local Chromosome Ancestry

We inferred chromosome local ancestry using the PCAdmix software[19] using ~2 Million SNPs shared by EPIGEN and 1000 Genomes Project (Section 5.2). Considering our SNPs density, we defined a window length of 100 SNPs, following[27]. PCAdmix infers the ancestry of each window. Local ancestry inferences were performed after linked markers (r^2>0.99) were pruned to avoid ancestry misestimating due to overfitting[4]. We considered only the windows which ancestry was inferred by the forward-backward algorithm with a posterior probability >0.90.

After local ancestry inferences, we calculated for each haplotype from each chromosome from each individual, the lengths of the chromosomal segments of continuous specific ancestry (CSSA), which distribution is informative about the admixture dynamics. The distribution of CSSAs length was organized in 50 equally spaced bins defined in cM and plotted for each population (Fig. S13 and Fig.2A). The distribution of CSSA length suggest that the admixture dynamics is similar in Bambuí (SE) and Pelotas (S), but not in Salvador (NE), where the European CSSA lengths are shorter, suggesting recent European admixture or a more pronounced ancestry-based positive assortative mating in the former than in Salvador. African admixture dynamics seems to be similar across the three cohorts.

We also looked for each population for entire chromosomes of a distinct ancestry that would suggest recent admixture and/or ancestry assortative matting. In Southeastern Brazil, and particularly in the Southern Brazil, we found a large number of individuals with European full chromosomes (Figure S14A), consistently with recent European immigrations to these regions. Interestingly, the Brazil´s Southeast and South present individuals with a larger number of African full chromosomes than in northeastern Brazil (Figure S14B), suggesting a more pronounced assortative matting based on African ancestry in South and Southeast compared to Northeast. This finds are consistent with the 2010 Brazilian census (http://censo2010.ibge.gov.br/) that showed that about 70% of Brazilian people were married to the same group of people of color/race.

6.8. Approximate Bayesian Computation (ABC) to Infer Admixture Parameters

We implemented a new approach based on Approximate Bayesian Computation (ABC)[48] and local ancestry to infer historical admixture parameters for each of the EPIGEN populations, conditioning on a model of admixture dynamics of three pulses of immigration. The main steps of this approach are:

(1) generation of an informative prior distribution of admixture parameters for each pulse, conditioning on the estimated total continental ancestry;

(2) simulation of chromosome segments of continuous specific ancestry (CSSAs), based on the prior distribution;

(3) computation of the distance between the simulated and observed CSSA distributions;

(4) estimation of the posterior distribution of the admixture parameters for each pulse by retaining the simulated CSSA distributions that are more similar to the observed distribution.

We simulated CSSA using the stochastic process described in[20] and implemented by them in the algorithm multipulses. The Liang-Nielsen model allows for at most one admixture event from a unique ancestral population per generation (i.e. European or African or Native American admixture). Considering this assumption and that European/African admixture in the Americas started 500 years ago, we constructed a model of admixture dynamics of three admixture pulses (early, intermediate and recent) distributed over 20 generations of 25 years each (Fig. S15). Each pulse has three possible proportions of immigrants (**m**) from the ancestral populations (European (EUR), African (AFR) and Native American (NAT)) arriving in consecutive generations. We called Admixture Scenarios (ASs) the combination of $\mathbf{m}_{n,P}$ (total of nine **m** parameters**),** where the positive real number **m** is the proportion of immigrants respect to the admixed population from the ancestral population ***n*** in the pulse ***P***.

To explore the space of population mean proportion of ancestry (**M)** space, we randomly generated the **m** number in each admixture pulse to produce ASs following these rules:

(a) The admixture events from the three ancestral populations are randomly sorted along the three generations of each admixture pulse;

(b) For Pulse 1: the first **m** is equal to 1 (i.e. founder population) and the sum of the next two **m** is ≤ 1;

(c) For Pulses 2 and 3: the sum of the three **m** is ≤ 1.

(d) After each immigration event defined by $\mathbf{m}_{n,P}$ is generated, the three parameters **M** corresponding to the three ancestral populations are updated.

These rules aim to avoid an unrealistic scenario in which a population is totally substituted by another population, and they allowed exploring all the **M** space from a uniform **m** over the three pulses (Fig. S16).

Initially, we randomly generated the $\mathbf{m}_{n,P}$ for 20 million of ASs and calculated the associated $\mathbf{M}_{n,P}$ over the three pulses, using the pseudocode described in Fig. S17. We retained those combinations of nine $\mathbf{m}_{n,P}$ values that generate $\mathbf{M}_{n,3}$ (current admixture proportions after the third admixture pulse) within the 5% range centered on the inferred mean proportion of European, African and Native American ancestry in Salvador (43%, 50% and 7%), Bambuí (77%, 16% and 7%) and Pelotas (76%, 16% and 8%). In this way, we generated informative prior distributions of admixture parameters **m,** ensuring that they always produce final **M** close to the observed data. It reduces the number of simulations needed in the following step, that is more computationally demanding.

Then, we used Liang-Nielsen multipulses software to simulate CSSAs distributions for the chromosomes 14, 19, 21 and 22 using the filtered ASs (~180.000 sets of $\mathbf{m}_{n,P}$) and the same number of diploid individuals (Salvador (1309), Bambuí (1442) and Pelotas (3736) for each EPIGEN population. We estimated the distance between the distributions of simulated and observed CSSAs (Section 6.7) using the Kolmogorov–Smirnov statistics Ks[49] Finally, we retained the 1% of the ASs that generated the simulated CSSA distributions closest to the observed CSSA distribution, estimating the posterior distribution of the 9 $\mathbf{m}_{n,P}$ for each EPIGEN population (Figs. S18-20). Considering the posterior probability distributions, we calculated the quantile-based probability intervals of 90% using Bayesian unimodal Highest Posterior Density (HPD) intervals (Fig. 2B).

Our ABC approach allowed us to elucidate the admixture dynamics in Brazilians. Overall, we observed different admixture dynamics between the Northeastern Brazil (Salvador) and Southeastern/South (Bambuí and Pelotas).

The European contribution to Salvador mainly occurred during the early and intermediate admixture pulses (AP) and to a lesser extent during the recent AP. Conversely, Bambuí and Pelotas showed an even European contribution over the three AP (Fig.2B and Figs. S18-20). The African contribution to the three populations showed a decreasing trend across time, but this trend was more pronounced in Bambuí and Pelotas (Fig.2B and Figs. S18-20). The dynamics of Native American contribution was small and similar in the three studied Brazilian populations, concentrated during the early pulse (Fig.2B and Figs. S18-20). Interestingly, this is consistent with the Native American decimation after the arrival of the Portuguese settlers.

6.9. Population structure inferred from lineage markers: mitochondrial DNA and Y-chromosome

Methods for Mitochondrial DNA Analysis

Merging data sets. After variant calling and QC filters for mitochondrial DNA (mtDNA), we had the following number of SNPs and subjects for each sample: Bambuí (213 SNPs; 1,442 individuals), Pelotas (218; 3,735), and Salvador (216; 1,308). These three sets of samples were merged, for a total of 219 SNPs and 6,485 individuals.

Haplogroup assignment. We performed haplogroup assignments using HaploGrep[40] (http://haplogrep.uibk.ac.at/), a web tool based on Phylotree (build 16) for mtDNA haplogroup assignment.

Haplogroup assignment checking. We adopted two strategies to check the HaploGrep results: (a) we used Network.exe (http://www.fluxus-engineering.com/sharenet.htm) to check for outliers. The HaploGrep-output file was split in smaller files containing subjects classified as belonging to the same haplogroup. We analyzed each haplogroup-specific independently with the Network software (using median joining calculation). Outliers were manually investigated for haplogroups assignment according to Phylotree build16 (http://www.phylotree.org/). (b)

We conducted PCA of the 6,485 individuals to check if each set of samples classified in a specific-haplogroup would cluster together in the PCA plot. Also, PCA was used to verify if we would be able to reproduce the pattern of Phylotree with the 219 SNPs used for the haplogroup assignment. We calculated the four first Principal Components (adegenet package) in R, and PCA plots of the first two PCs were generated for all sample. We repeated the analysis, independently, for the set of individuals with Europeans haplogroups as well as the set of individuals with African haplogroups

Based on the haplogroup/subhaplogroups frequencies (inferred by HaploGep), population genetics analyses were performed using the Arlequin software 3.1[50].

The haplogroup assignment checking performed with the network and PCA suggest that HaploGrep was efficient in determining the haplogroup status using the set of 219 SNPs available for the analysis. Sequences classified as belonged to a specific haplogroup or sub-haplogroup were clustered together in the PCA plot, and we did not observe any outliers (i.e. potential haplogroup misclassified) in our sample. Furthermore, we were able to reproduce the mtDNA phylogeography tree through PCA, being able to distinguish among individuals from Africa, Asia/America and Europe.

HaploGrep also provides a confidence value for its haplogroup/subhaplogroup inferences, based on two components rank calculation (for details see[40]). This is however only valid for whole mtDNA genomes. We therefore classified all profiles defined in Phylotree by applying a range according to the available SNPs positions to check the reliability of the resulting haplogroups with HaploGrep. This way we found 96.1% of all 4,806 possible haplogroups to be classified in the correct Macro-Haplogroup. B4a*haplogroups in Phylotree could not be found with the available SNPs and were classified in 70 out of the 76 present false as HV, 28 of 52 Phylotree V groups ended up in the HV0 haplogroup. Also Haplogroups in the R* clade result in the HV branch. In total 35 HV haplogroups were found, with a frequency of 0.5%.

We had a total of 6,485 individuals for 124 inferred haplogroups or sub-haplogroups. Table S19 shows the frequencies of all haplogroups and subhaplogroups inferred by HaploGrep. Table S20 summarizes the population genetics results of the haplotype analyses.

To estimate admixture contributions from mtDNA, we relied on the continental tri-hybrid admixture nature of the Brazilian population and on extensive available literature on the phylogeography of mtDNA, and we performed the continental biogeographic assignments of haplogroups (Table S21). Namely, haplogroups A, B, C and D were considered as Native Americans, haplogroups H, HV, I, J, K, T, U, V, M, N, P, Y, W were considered as markers of European/Middle Eastern and Asiatic admixture, and all the L haplogroups were considered as markers of African admixture during the last five centuries. This biogeographic classification has some limitations. For instance, haplogroups H and V have been recently reported in some Sub-Saharan African populations at medium frequencies (10-15%)[51,52]. Therefore, by considering all H and V haplogroups as European, we recognize that we overestimate the European contribution and under-estimate African contribution.

Based on biogeographic assignments of Table S21, we estimated the African, European and Native American female-mediated (i.e. based on mtDNA) contributions to the three EPIGEN cohorts simply as the observed frequencies of the continental-attributed haplogroups. We considered all the Eurasian haplogroups as European contribution (including Middle East), because based on historical records, East Asian contribution should be very low. Overall, both

African and Native American ancestry estimates for mtDNA are higher than autosomal estimates across the three cohorts (Table S20), which is the result of a historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry. This pattern of mating is well documented in demographic and genetic studies across all Latin America[12]. Despite this bias, across the three cohorts the largest continental contributions are the same both for autosomal and mtDNA estimates: African for Salvador, and European for Bambuí and Pelotas, although in Bambuí, the three continental contributions are more evenly distributed for the mtDNA. This predominant African ancestry in Salvador and the predominant European ancestry in Bambuí and Pelotas are reflected in the highest differentiation of the Bambuí cohort in the F_{ST} matrix based on mtDNA (Table S22).

Subcontinental biogeographic interpretation. When we estimate the population differentiation (F_{ST}) between the EPIGEN cohorts independently for the sets of haplogroups/subhaplogroups assigned to each continental ancestry (i.e. when we exclude the effect of the higher whole-African contribution to Salvador and the higher whole-European contribution to Bambuí and Pelotas), Bambuí is consistently the most differentiated population. Because this is a general pattern of most Bambuí haplotypes, independently of their continental origin, this pattern probably reflects the recent Post-Columbian demographic history of Bambuí that, as inferred from autosomal data, has an important familiar structure and high levels of inbreeding that are likely related with a higher level of isolation respect to Pelotas and Salvador. Bambuí, independently of its higher frequencies of the African L haplogroups, is characterized by: (i) the absence of the Native haplogroup A, which is common in almost all Latin American population with a non-negligible Native American female-mediated genetic contribution). (ii) a relative high frequency of the Eurasian haplogroup N (13% vs. <1% in Salvador and Pelotas) and (iii) by presenting the L1c haplotype (more common in West-Central Africa than elsewhere in the continent[53]) as modal among the African-specific haplogroups (22%). In Salvador and Pelotas, L1c is the second most common African haplotype (12% and 15% respectively), the pan-African L2a being modal.

Respect to intra-continental sub-haplogroups distribution, Pelotas and Bambuí, despite their similar genome-wide estimates of total European ancestry, differs in the frequency of the Euroasiatic N subhaplogroups: 94.5% of the N haplogroups in Pelotas are N vs 1.3% in Bambuí and 0.05% of the N haplogroup in Bambuí are N2 vs. 68.4% in Pelotas. Also, in general the M haplogroup is rare in our samples, but the M1 subhaplogroup is common in Pelotas respect to the total of the M haplogroup (66 out of 70 copies). For African subhaplogroups, Pelotas respect to Salvador has slightly higher frequencies (relative to the pool of L haplogroups) of subhaplogroups L3e, L3 and L1c and slightly lower frequencies of subhaplogroups L1b and L2a.

The dataset for the analyses was composed by 3,142 males from Bambuí (N=562), Pelotas (N=1,873) and Salvador (N=707). From the 2,775 Y-SNPs genotyped, 1,886 were used in these analyzes.

We inferred haplogroups using an automated approach, written in Perl, called AMY-tree[41]. The assignment considers a phylogenetic tree with the root on the left and the leaves on the right side, traversing the nodes to determinate the (sub)haplogroups of each sample, due the hierarchical order of the non-recombining region of Y chromosome (NRY) variants. For the haplogroups inferences, we considered the "Karafet tree"[42] and more recent studies to describe additional sub-haplogroups, therefore, an updated tree was considered based on the information given in The International Society of Genetic Genealogy (ISOGG version 9.43, www.isogg.org accessed in 03.20.2014).

Since many SNPs may have several names, these redundancies were identified and considered only once. Capital letters were used to identify major clades and the alphanumeric nomenclature was applied to name sub-haplogroups, following[42].

From the AMY-tree output, we organized results considering each population. Tables with absolute numbers and frequencies were manually constructed, considering both major clades and sub-haplogroups. All samples were associated to at least a major clade (like T*) and, when possible, sub-haplogroup were identified (like R1b1a2a1a2b2a1*).

Using the Y-SNP dataset we determine the Y-haplogroup of all males (N=3,142) and identified 70 sub-haplogroups included in 14 major clades. Considering each population, we found 43 sub-haplogroups in Bambuí (N=562), 60 in Pelotas (N=1,873) and 51 in Salvador (N=707). Table S23 shows the frequencies of all sub-haplogroups.

Because in the tree defined by[42] there is a strong association between most haplogroups and continental distribution, we performed the following assignment (Table S24). We considered as Eurasian (i.e. European for the purpose of the recent migration into Brazil), the haplogroups D, O, G, I, J, L, N, R and T, and the sub-haplogroups E1b1b1b1* and E1b1b1b1b (common in Middle East and Jews, and in the Iberian Peninsula[54]. The most frequent European subhaplogroup is R1b1a2a (formerly R1b1b2) defined by L11 (rs9786076), described by[55] as a Western European subhaplogroup. The J clade ranks second among European haplogroups, particularly J2*. Haplogroups A, B and E (except E1b1b1b1* and E1b1b1b1b) are considered by us as Africans. Haplogroup Q is considered Native American. As in the case of mitochondrial DNA, this biogeographic classification has some limitations, because association between haplogroups and continents is not absolute. However, this biogeographic classification allows a reasonable quantification of the amount of continental admixture mediated by males during the last five centuries. A further issue in Y-chromosome continental assignment is the high frequency of the haplogroup “Root” in the Bambuí cohort. These individuals are classified as “Root” because does not hold any of the mutations that define the well-defined Y-chromosome haplogroups A-T. “Root” haplogroups are found both in Africa and Europe at low frequencies[56]. Thus, to determine whether ancestral origin of "Root" haplogroups found in EPIGEN cohorts were African or European we inferred the haplogroups of public domain y chromosomes, using the same methodology described above. Thereafter we performed a PCA using common SNPs between 1000 Genomes populations and EPIGEN. Our results showed that all “Root” haplogroups from EPIGEN clustered with the European samples from 1000 Genomes classified as R haplogroup. Therefore, all “Root” haplogroups from EPIGEN were considered European.

We estimated Y-chromosome specific continental admixture in the same way than for mitochondrial DNA. The particularly high frequency of “Root” haplogroup in Bambuí determines the highest pairwise F_{ST} observed between Bambuí and Salvador or Pelotas (~13%, Table S22).

For Salvador, Bambuí and Pelotas, consistently with the results obtained for mitochondrial DNA, we observed a higher Y-chromosome (i.e. male mediated) continental European admixture than autosomal estimate. Again, this is due to the historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry (Table S20). Also, and consistently with autosomal estimates, Salvador has relatively higher percentage of African-associated haplogroups such as E1b1a (Table S23, >20% vs. <4% in Pelotas and Bambuí).

7. SNP ANNOTATION

We used the results of ADMIXTURE analysis with K=9 to obtain SNP frequencies for the East-mustard and West-blue Africa clusters (EAFRxWAFR). We then estimated the F_{ST} values for each SNP. After that, we determined a 99% cut-off for the F_{ST} values which is 0.059 for the EAFRxWAFR SNPs. This resulted in 3,318 most differentiated SNPs between EAFRxWAFR, which were then annotated.

We used an annotation software developed by us, called MASSA, to perform annotation regarding *Diseases and Traits* from the GWAS Catalog (version March 2014), a database of genome-wide association studies hits for SNPs and Genes. The result shows 38 SNPs that are GWAs hits, as described in Table S25.

8. WHOLE GENOME DATA

8.1. Samples for Whole-Genome Sequencing and Quality Control

Sampling

We sequenced the complete genome of 30 Brazilians individuals using Illumina's methods (Illumina - Pub. No. 770-2007-002). We randomly selected 10 individuals from each of the EPIGEN cohorts, conditioning on availability of DNA quality and quantity. In total, we sequenced the genomes of eighteen men and twelve women overall. All DNA samples were obtained from peripheral leukocytes by four different DNA extraction methods (EZ-DNA isolation kit, Gentra Puregene Blood – QIAGEN, *salting-out* method, and phenol-chloroform method). A minimum of 1.75 µg of DNA (stored in a solution of 35 µl) of each sample was sent to the Illumina facility in San Diego (CA, US), where it was sequenced with the Hiseq 2000 platform (Illumina - Pub. No. 770-2009-036) and genotyped for 2.5 million SNPs using the HumanOmni2.5-8 chip, for the purpose of an internal control by the Illumina LIMS (Laboratory Information Management System).

These are the codes of the individuals whole-genome sequenced: B0078, B0516, B0741, B0987, B0990, B1097, B1102, B1149, B1261, B1282, P0026, P0075, P0078, P0086, P0176, P0227, P0377, P2110, P2829, P2953, S0421, S0509, S0527, S0534, S0541, S0636, S0637, S0638, S0647, S0649. B, P and S codes corresponds to Bambuí, Pelotas and Salvador.

Library construction

Illumina generated paired-end libraries from 500ng-1µg of genomic DNA using the TruSeq DNA Sample Preparation Kit (Illumina's Catalog #: FC-121-2001; Pub. No. 770-2012-019). This step includes the purification of genomic DNA using magnetic beads (Agencourt®AMPure® XP reagents, Beckman Coulter), fragmentation of genomic DNA, and end-pairing of fragments of approximately 300 bp (Illumina's Catalog # PE-930-1001; Part # 1005063 Rev. E). Finally, an electrophoresis is used to confirm fragments size and DNA quality.

Clustering and Sequencing

The Clustering procedure provides enough number of DNA molecules to be sequenced by the Illumina's HiSeq2000. For clustering, libraries are denatured, diluted, and clustered onto v3 flow cells using the Illumina cBot™ system (Illumina - Pub. No. 770-2009-032). This system promotes cDNA fragments amplification onto the surface of the flow cells. Fragments anneal with DNA template covalently bound onto the flow cells, where isothermal enzymes promote the extension of the attached DNA to create hundreds of millions of clusters, each containing around 1,000 identical copies of a single template molecule. cBot runs are performed based on the cBot User Guide (Illumina's Part#15006165 Rev. K), using the reagents provided in Illumina TruSeq Cluster Kit v3 (Illumina's Catalog #: PE-401-3001).

The flow cells are then loaded onto the HiSeq2000 for sequencing. Each run performs sequencing on 100 bp paired-end, non-indexed, following HiSeq 2000 User Guide, which requires using Illumina TruSeq SBS v3 Reagents. Briefly, two primers are used to sequence both ends of the fragment. While sequencing runs, each lane of the flow cell is controlled for quality to guarantee >80% of the bases with a Qscore>30. These controls are performed using manufacture's tools, such as Illumina HiSeq Control Software and Real-Time Analysis (RTA). These tools generate final sequencing files in .bcl format (Illumina - Pub. No. 770-2009-020), which comprises base callings and quality values by cycle.

Alignment and Variants Identification

Sequencing files in **.bcl** format produced by the Illumina HiSeq Control Software and Real-Time Analysis (RTA) are the initial files used by Illumina on its standard data analysis pipeline. Illumina used CASAVA v1.9 (Consensus Assessment of Sequence and Variation) to convert the .bcl files to Fastq format and to map the reads against the reference genome NCBI37/hg19 (stored at the *Assembly* folder, inside the *Genome* and *bam* subdirectories), in order to identify SNPs (Single Nucleotide Polymorphisms) and INDELS (insertions and deletions). CASAVA performs sequencing alignment using the *configureAlignment* module, which comprises a set of scripts and protocols (CASAVA v1.8.2 User Guide - Part # 15011196 Rev D). The *configureAlignment* module includes the Illumina's ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) alignment algorithm version 2 (Illumina – Pub. No. 770-2011-005). Alignment parameters used at CASAVA can be found at *Assembly/conf/project.conf* (more detailed information about parameters meaning can be found at CASAVA v1.8.2 User Guide - Part # 15011196 Rev D or at:
http://umbc.rnet.missouri.edu/resources/How2RunCASAVA.html).

After the alignment, reads of each genome are ordered by their positions and converted to BAM format (http://samtools.sourceforge.net). After this conversion to the BAM format, the CASAVA *assembleIndels* module is used to identify possible INDELS, and the *callSmallVariants* module to identify variants genotypes. For INDELS identification, CASAVA requires parameters to be provided, available at the *project.conf* file. The callSmallVariants module calls SNPs and small indels from both the sorted alignment files (sorted.bam) and optionally also from the candidate indel contigs produced by assembleIndels.

Illumina Array concordance - HumanOmni2.5-8v1

Sequenced samples were also genotyped using the HumanOmni2.5-8 chip, as an Illumina internal control, and showed an average agreement of 99.27%.

EPIGEN- QC analyses

The VCF files generated for each genome were treated following quality parameters to build final datasets suitable for posterior analyses. VCF files have quality values based on Illumina´s Qscores (Illumina – Pub. No. 770-2011-030) for each variant. This Illumina Qscore is generated according to a set of parameters, including base calling quality, its concordance with the reference genome, whether it is a beforehand known polymorphism, etc. We used a final Qscore ≥ 20 as cutoff to label variants as "PASS" and kept them in the file. In the EPIGEN project context, the VCF file was filtered using the software VCFtools[46] to create a final VCF file containing only those variants with Qscore ≥ 20.

EPIGEN VCF files filtering – SNPs variants

Illumina generates specific VCF files for different types of variants, but only high quality SNPs were considered in the following analyses. To create a final dataset, we filter only the SNPs with a Qscore higher or equal to 20.

We fixed some inconsistencies regarding SNPs rs# identification numbers, such as same positions labeled with two or more different rs# numbers, which will produce error in analyses with GLU (http://code.google.com/p/glu-genetics/) and PLINK (http://pngu.mgh.harvard.edu/purcell/plink/). Also, the same rs# number was often registered for more than one physical position. We also evaluated the concordance between values of columns *max_gt* and *poly_max_gt* of VCF files generated for each genome of the EPIGEN project. Only those variants that showed a concordant value were kept in the new VCF file, increasing the dataset reliability. Therefore, the final data set in VCF format was used in the following analyses.

EPIGEN Data quality summary

Each genome was sequenced on average 42x (mean deep coverage), with an average of 128 GB of passing filter and aligned to the reference genome (HumanNCBI37_UCSC), 82% of bases with data quality Qscore>=30, 96% of Non-N reference bases with a coverage >= 10x, an HumanOmni5 array agreement of 99.53% and a HumanOmni2.5 array agreement of 99.27% (Table S26).

Figure S21 shows the Venn diagram of the distribution of the 15,033,927 biallelic SNPs identified in the 30 Brazilian genomes and its intersection with the databases dbSNP-138 and 1000 Genomes SNPs.

Figure S22 shows the distribution of the 15,033,927 identified SNPs in the three Brazilian cohorts.

8.2. Functional Annotation with ANNOVAR based on refGene

Functional annotation of the whole-genome variants was performed using ANNOVAR (August 2013 release) with refGene v.hg19_20131113 reference database and with ensGene v.hg19_20131113 reference database. ANNOVAR classifies the variants into different categories considering their functions (Table S27). ANNOVAR and the other functional annotations described below were performed on the set of 15,033,927 SNPs (14 988 895 of them are biallelic).

SNPs annotation showed that most of the SNPs were classified as intergenic (58.03%) or intronic (34.88%), whereas the remaining variants were classified in other functional categories (Fig. S23) including 101,201 SNPs (0.68%) in coding exonic regions of which 6,329 (6.25%) were not present at dbSNP138 neither at 1000 Genomes Phase1 database (hereafter called novel). We identified similar proportions of non-synonymous and synonymous exonic SNPs: 50,518 (49.91%) and 48,464 (47.88%) respectively (Tables S28 and S29), a result that is similar to other studies (Tables S28). Furthermore, of the 6,329 novel exonic SNPs, 99 (1.56%) were classified as stopgain SNPs, 1 as stoploss, 2,223 (35.12%) synonymous, 3,865 (61.07%) non-synonymous, and 141 (2.23%) as unknown.

To evaluate ANNOVAR's accuracy classifying the SNPs of 30 Brazilian genomes, we checked manually the annotation of 210 exonic SNPs in the dbSNP138 website. The results showed a high concordance between the ANNOVAR annotation and the dbSNP database since only for 1 SNPs (rs34179073) ANNOVAR and manual dbSNP checking produced inconsistent results. dbSNP gives a missense classification while ANNOVAR reports it as a synonymous mutation, once they annotate the variant based on different non-reference allele.

8.3. Functional annotations with other tools and databases

We also performed functional annotation using Variant Effect Predictor (VEP, v77, http://www.ensembl.org/info/docs/tools/vep/index.html), based on the Ensembl database (October 2014, release 77_GRCh37) and RefSeq database (October 2014, release refseq_vep_77_GRCh37). The Ensembl classification is based on the Table of functional categories available in http://useast.ensembl.org/info/genome/variation/predicted_data.html#consequence_type_table.

Importantly, when there are multiple possibilities, VEP returns the annotation for the more severe category. However, it is possible to obtain the information with all the analyzed transcripts.

8.3.1.Functional annotation using VEP (RefSeq*)*

Table S30 shows exonic SNPs classified by VEP (Ensembl) in the 30 Brazilian genomes.

8.3.2. Functional annotation using ANNOVAR (Ensembl*)*

Table S31 shows exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes.

8.4. Analysis of deleterious variants by CONDEL

First, we determined the ancestral-derived phylogenetic status for 45875 of the 49494 autosomal non-synonymous SNPs annotated with ANNOVAR and RefSeq database, by retrieving the ancestral allele information for each SNP from ancestral sequences files available in 1000 Genomes Project FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/) using BEDTools suite v2.15 (http://bedtools.readthedocs.org/en/latest/content/bedtools-suite.html). Then, these 45875 variants were predicted for deleteriousness using CONDEL v2.0[28], which calculates the scores for each SNP as a weighted average of the scores of MutationAssessor[36] and FatHMM[37]. Once CONDEL analysis fails for 869 SNPs, we, initially, treated the result file with 45006 hits removing 289 SNPs without CONDEL score. Because CONDEL shows the scores for all transcripts analyzed from the Ensembl database, we also excluded 700 SNPs with different predictions for more than one transcript. Thus, after applying these filters, our analysis included 44017 autosomal non-synonymous SNPs. We considered as deleterious mutations the derived variants of those SNPs with a CONDEL score > 0.52, as recommended by the CONDEL authors.

Simons et al.[30] reported a bias in methods that detect deleterious variants based on phylogenetic comparisons. They evidenced that when the human reference allele is the derived one, methods that identify deleterious variants tend to underestimate its deleterious effect. We confirmed the presence of this bias in our CONDEL analysis. Table S32 reports the comparison of the CONDEL scores for the derived/reference and derived/non-reference variants across different allele frequency classes estimated from our 30 genomes. Consistently with[30], across all the allele frequency classes, CONDEL scores are lower for the derived-reference than for the derived/non-reference alleles (Fig. S25). Therefore, we corrected the bias by the following procedure: for all the derived-reference variants, we added to the uncorrected CONDEL score, the value of the bias corresponding to its allele frequency class, where

bias = CONDEL score$_{\text{derived/non-reference}}$ – CONDEL score$_{\text{derived/reference}}$.

After this correction, we identified 8035 deleterious variants (versus 7451 before the correction), of which 6604 are rare deleterious variants (frequency < 0.10) and 79 are very deleterious variants (CONDEL score > 0.80) (Fig. S26).

9. REFERENCES

*The texts corresponding to these references are only in this SI Appendix

43. International HapMap 3 Consortium; et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58. Q:25

44. 1000 Genomes Project Consortium; et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65.

45. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104.

46. Danecek P, et al.; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

47. Vicoso B, Charlesworth B (2006) Evolution on the X chromosome: Unusual patterns and processes. Nat Rev Genet 7(8):645–653.

48. Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate Bayesian computation without summary statistics: The case of admixture. Genetics 181(4):1507–1519.

49. Sokal RR, James RF (2012) Biometry (Freeman, New York), 4th Ed.

50. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10(3):564–567.

51. Badro DA, et al.; Genographic Consortium (2013) Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern Middle Eastern populations with European and African populations. PLoS ONE 8(1):e54616.

52. Hernández CL, et al. (2014) Human maternal heritage in Andalusia (Spain): Its composition reveals high internal complexity and distinctive influences of mtDNA haplogroups U6 and L in the western and eastern side of region. BMC Genet 15:11.

53. Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J (2009) On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evol Biol 9:80.

54. Scozzari R, et al. (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 24(3):535–544.

55. Rocca RA, et al. (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: An online community approach. PLoS ONE 7(7):e41634.

56. Mendez FL, et al. (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92(3):454–459.

57. Lachance J, et al. (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150(3):457–469.

58. Shen H, et al. (2013) Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS ONE 8(4):e59494.

10. FIGURES

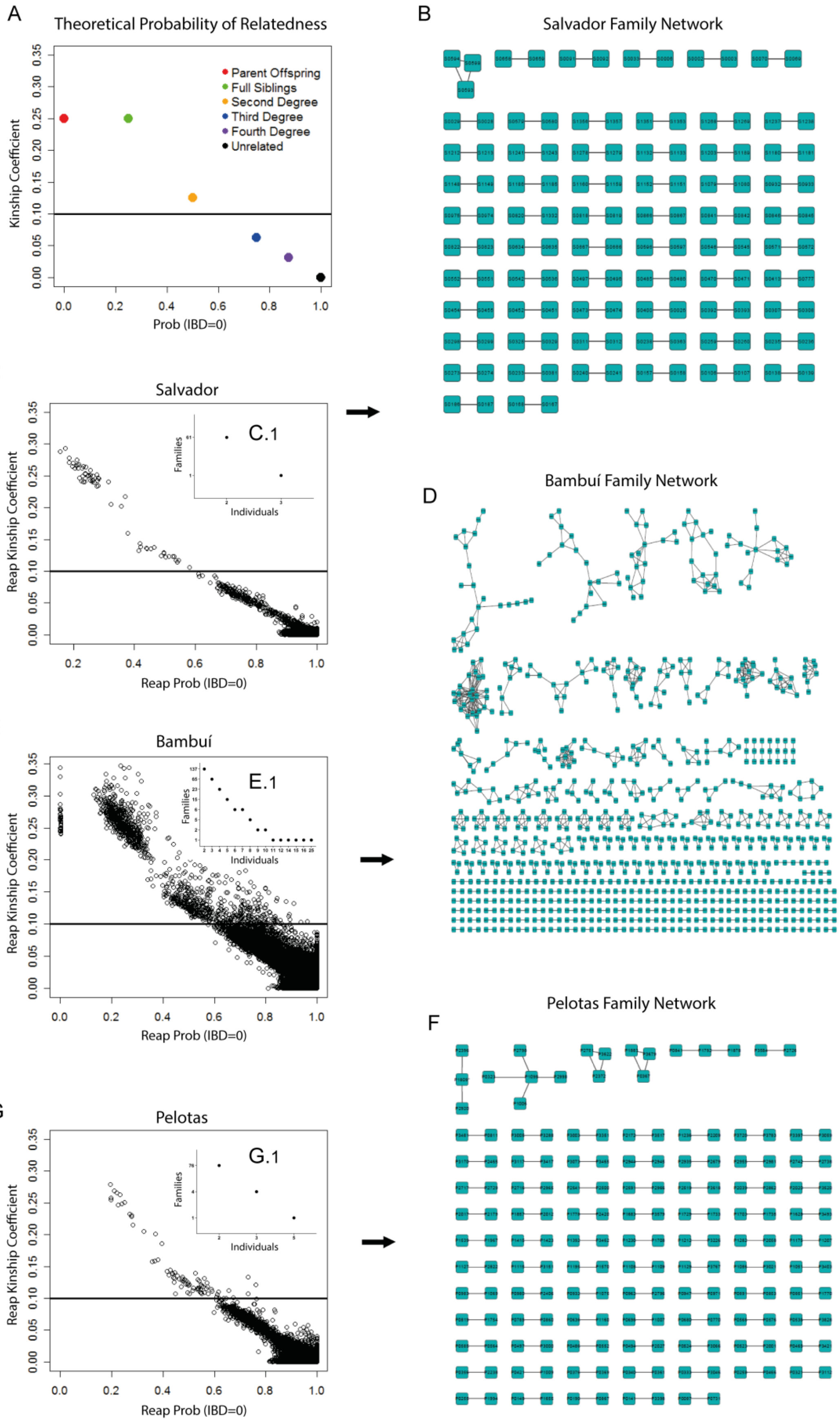
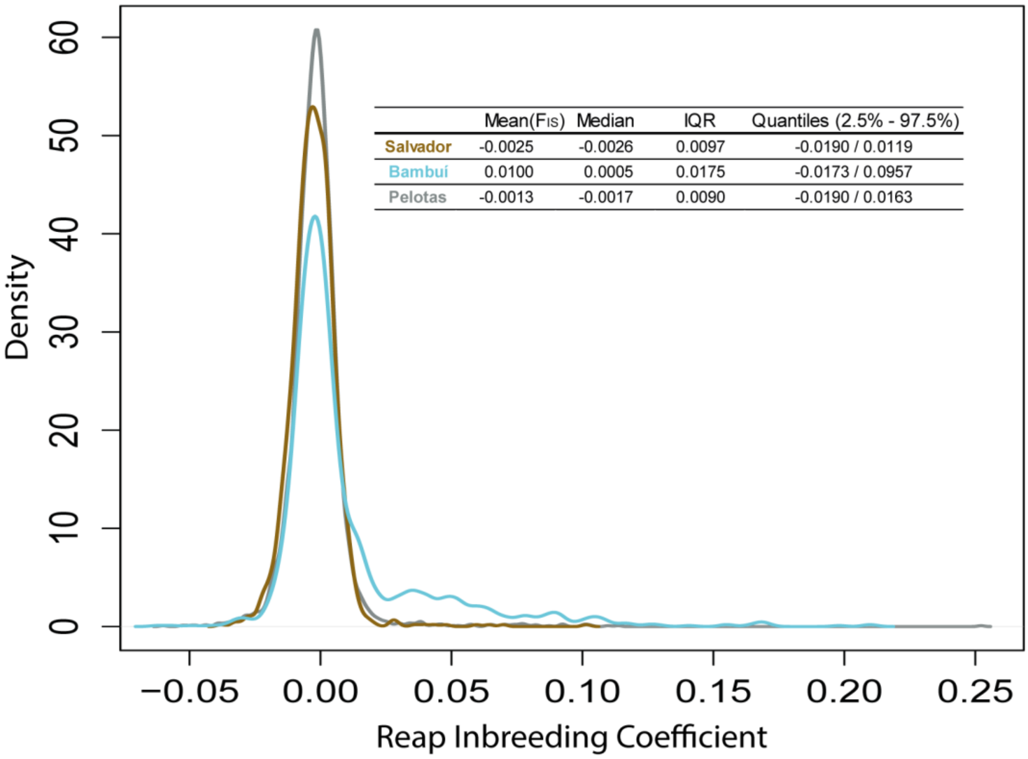


Figure S1. Relatedness in the EPIGEN cohorts. (A) The combination of theoretical values of kinship coefficients and the probability that individuals i and j share zero identical-by-descent

alleles (IBD=0) for different degrees of relatedness. These combinations describe the proportion of IBD genomic regions shared by two blood relatives. A pair of first-degree relatives (parent/offspring or full siblings) are IBD for about half of their genome. A second-degree relative of a person (uncle/aunt, nephew/niece, grandparent/grandchild or half-siblings) is IBD for about one quarter of their genomes. A third degree relative of a person (a first cousin and great-grandparent/great-grandchild) is IBD for about one eighth of their genomes. C, E and G plot kinship coefficient on the vertical axis and IBD=0 on the horizontal for Salvador, Bambuí, and Pelotas, respectively. The thick lines in the plots represent a “family”-kinship coefficient threshold Φij ≥ 0.1 established to consider individuals as related or not. B, D and F are the Salvador, Bambuí and Pelotas family networks, in this order. We model the families within each cohort like a network, where each node is an individual who connects to others by edges, which represent kinship coefficients ≥ 0.1.



	Mean(F_{IS})	Median	IQR	Quantiles (2.5% - 97.5%)
Salvador	-0.0025	-0.0026	0.0097	-0.0190 / 0.0119
Bambuí	0.0100	0.0005	0.0175	-0.0173 / 0.0957
Pelotas	-0.0013	-0.0017	0.0090	-0.0190 / 0.0163

Figure S2. REAP Inbreeding Coefficient. Distribution of individual inbreeding coefficients in the EPIGEN populations estimated using REAP software. F_{IS} is the mean of the inbreeding coefficients across individuals. IQR = interquartile range.

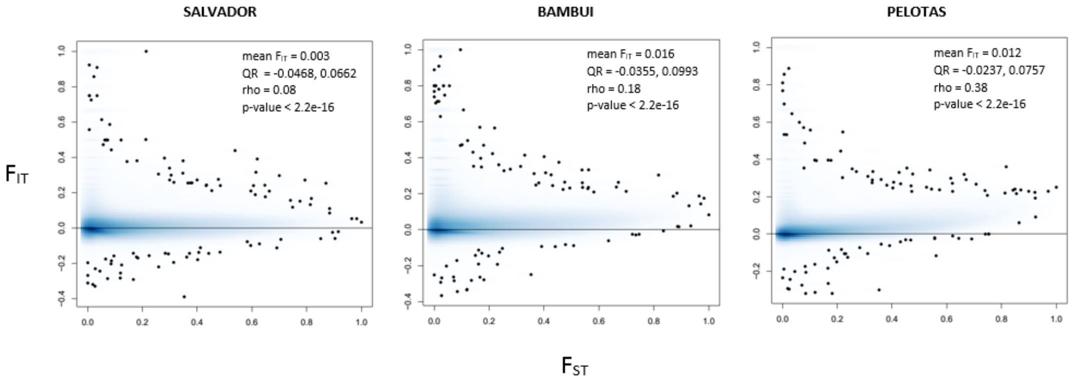


Figure S3. Homozygosity vs Informativeness for ancestry. The smoothed scatter plots represent the association between homozygosity excess and informativeness for ancestry. Homozygosity excess was measured by the F_{IT} per SNP estimated for each population. Informativeness for ancestry was measured by the F_{ST} per SNP estimated between the African and European populations. In the upper-right of the plot, we report the mean F_{IT} for the population cohort and the Spearman correlation parameter rho (cor.test function in R) between F_{IT} and F_{ST}. QR = quartile range.

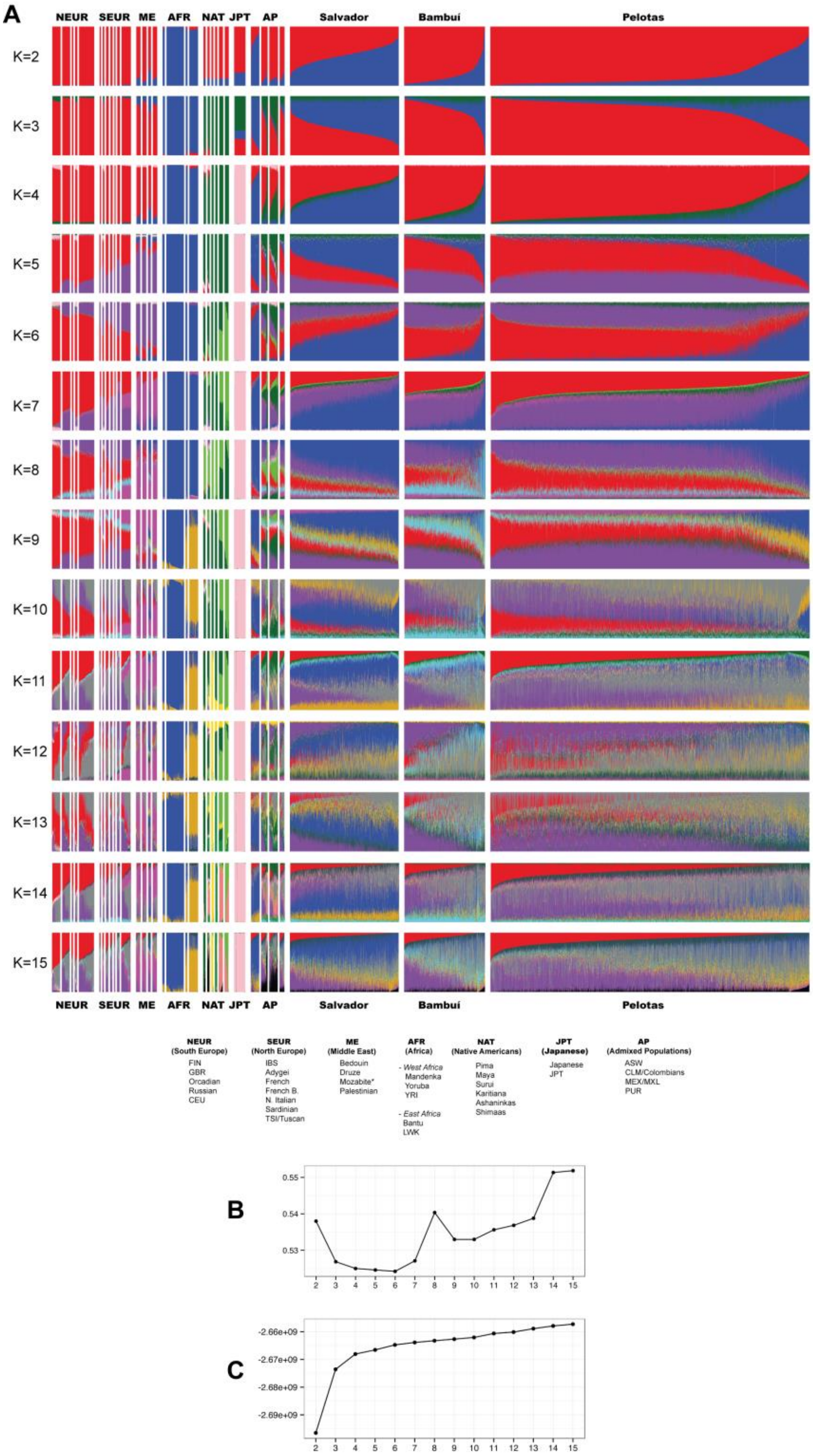


Figure S4. Barplot representation of the individual ancestry proportion for unrelated individuals inferred using ADMIXTURE. The proportions of Individual ancestry values were calculated using the number of parental K = 2 to K = 15 for the Dataset U (the main dataset used to study the population structure of the EPIGEN populations). Ancestral populations are sorted so that each one is assigned to an ethnic/geographic group, like North Europe, Middle East and Native American. The populations of each ethnic/geographic group are described at the bottom of the figure in the same order as plotted. Each bar represents an individual and each color a specific ancestry cluster. Barplots are sorted for each K by decreasing amount of the red ancestry cluster in the EPIGEN populations and individuals are not vertically aligned across the Figure. *Mozabite is a northwestern African population. ADMIXTURE cross-validation errors (B) and Log-likelihoods (C) as a function of K. Results corresponds to A.

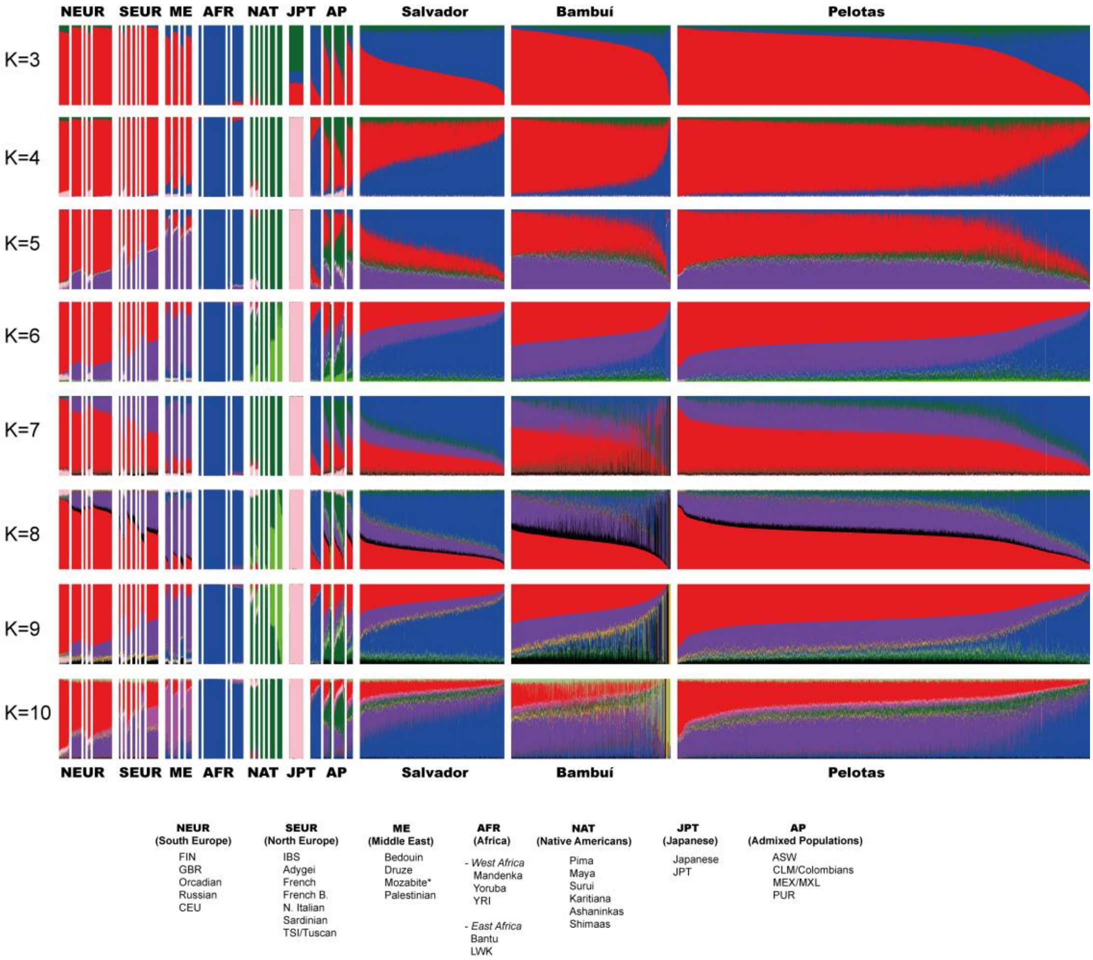


Figure S5. Barplot representation of the individual ancestry proportions for all EPIGEN individuals. The proportions of Individual ancestry values were calculated using the number of parental clusters K = 3 to K = 10 for the Original Dataset. Ancestral populations are sorted so that each one is assigned an ethnic/geographic group, like North Europe, Middle East and Native American. The populations of each ethnic/geographical group are described at the bottom of figure in the same order as plotted. Each bar represents an individual and each color a specific ancestry cluster. Barplots are sorted for each K by decreasing amount of the red ancestry cluster in the EPIGEN populations and individuals are not vertically aligned across the Figure. *Mozabite is a northwestern African population.

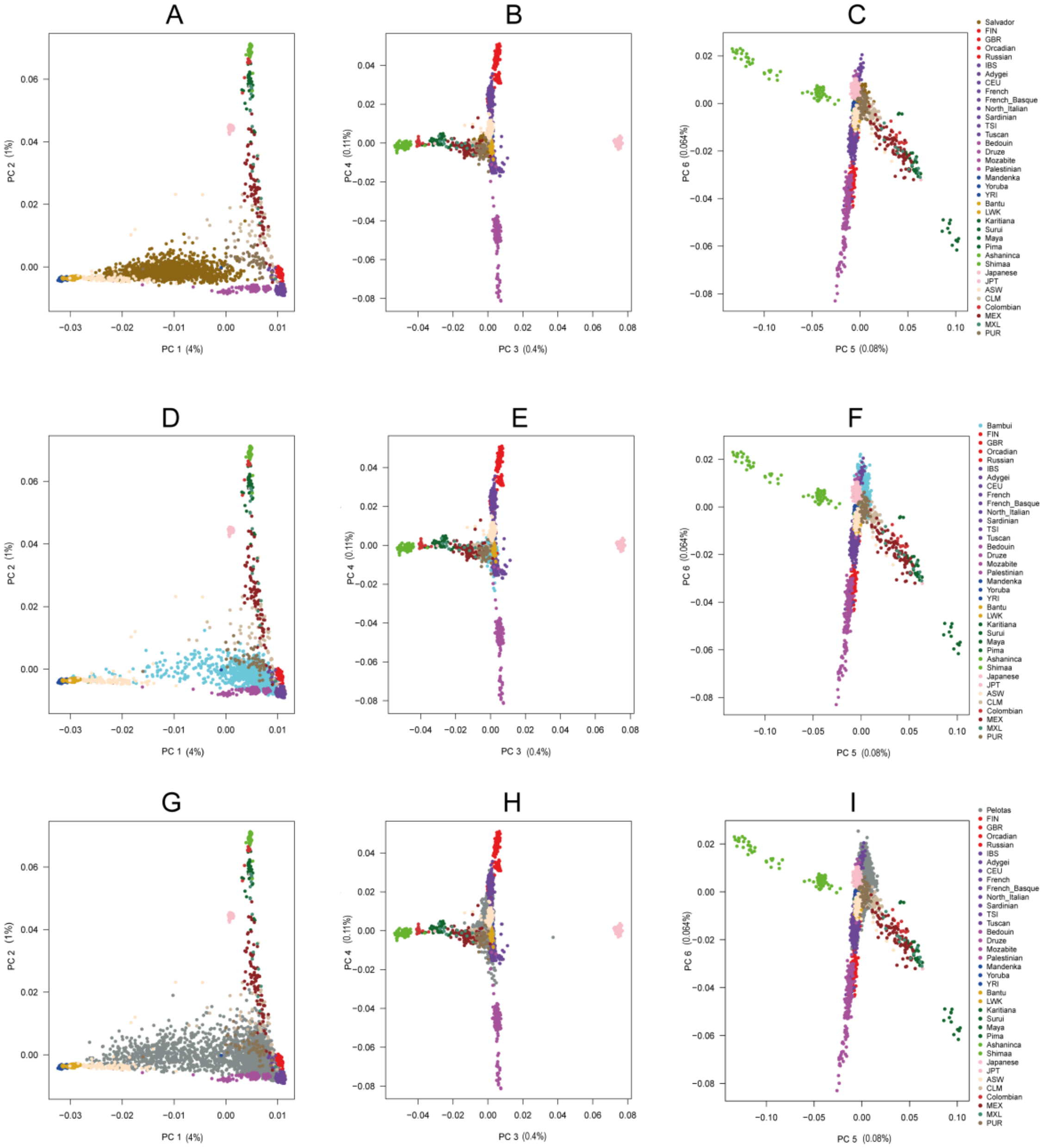


Figure S6. Principal Component Analysis (PCA) for EPIGEN and worldwide populations. PCA and the percentage of variability identified by each PC for Dataset U (the main dataset used to study the population structure of the EPIGEN populations, that does not include relatives), representing the worldwide populations and Brazil Northeast (Salvador), Southeast (Bambuí) and South (Pelotas) populations.

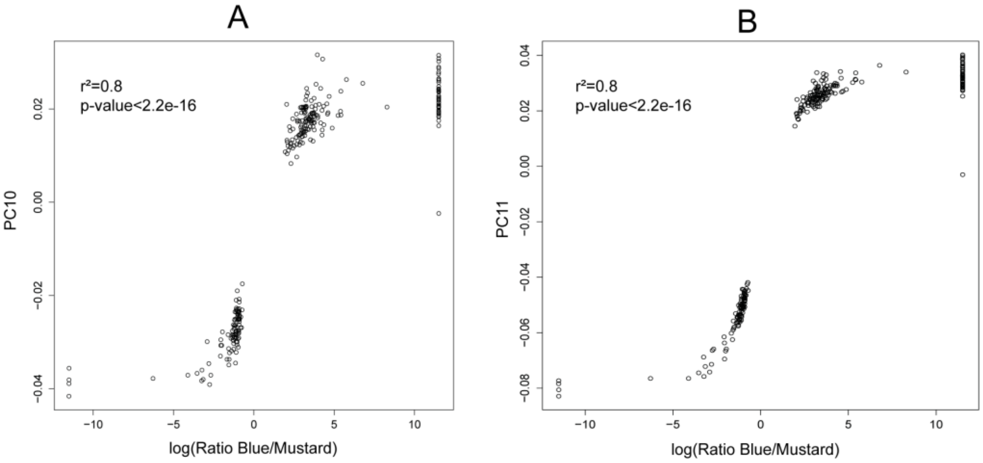


Figure S7. Correspondence between sub-continental African ancestry clusters identified by ADMIXTURE and by Principal Component Analysis. Scatterplot of the logarithm of the ratio between the blue and mustard sub-continental Africa ancestry clusters obtained from ADMIXTURE analyses (K=9) in each Brazilian individual from the EPIGEN cohorts (horizontal axes), versus the individual coordinate in the 10th (PC10, A) and 11th (PC11, B) Principal Components (vertical axes), estimated using the Dataset U (i.e. that contain no relatives). We estimated the association using Pearson's product-moment correlation coefficient. The high correlation suggest that Principal Components 10 and 11 capture the information of the within-African ancestry clusters, being correlated with the proportion of the blue ancestry component.

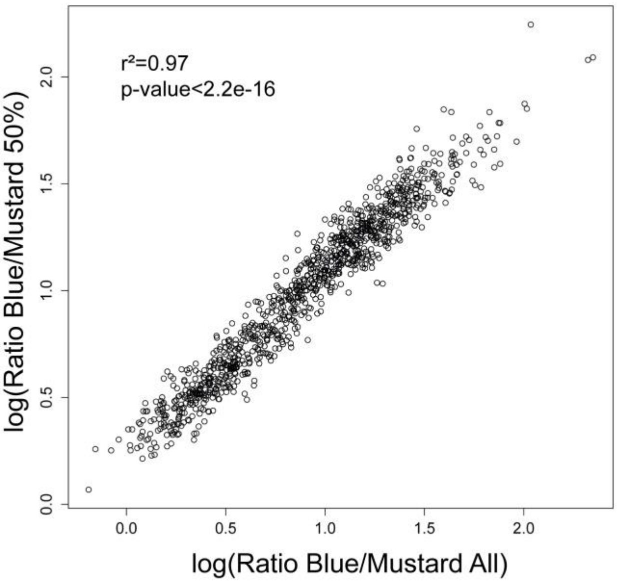


Figure S8. Logarithm of the ratio between the sub-continental African ancestry clusters. Testing the consistency of estimates of within-Africa ancestry clusters in function of total African ancestry. Scatterplot of the logarithm of the ratio between the blue and mustard sub-continental Africa ancestry clusters obtained from ADMIXTURE analyses (K=9) for the individuals from the EPIGEN populations with more than 50% of total African ancestry, estimated using the Dataset U (that contain no relatives). In the horizontal axis is represented the estimates obtained from ADMIXTURE when the run was performed including all the individuals (independently of their amount of African admixture). In the vertical axis is represented the estimates obtained from and ADMIXTURE analysis using only individuals with >50% of total African ancestry. The high Spearman correlation suggests that the estimates of

the blue and mustard within-Africa cluster of ancestry do not depend on the level of individual total African ancestry.

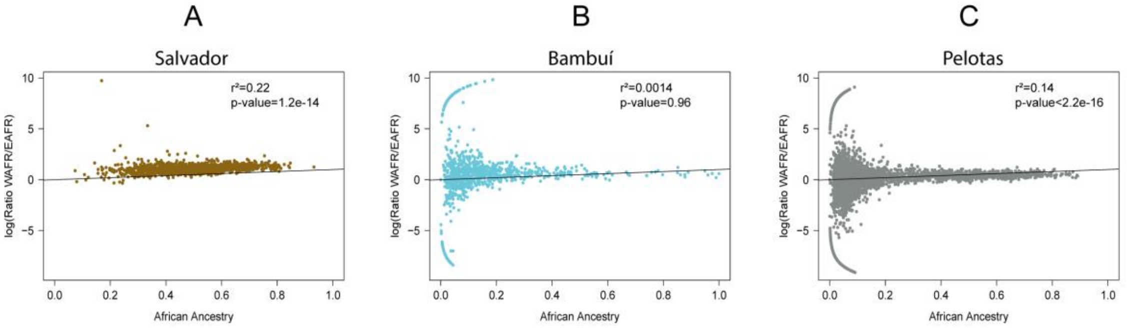


Figure S9. Testing correlation between sub-continental African ancestry clusters and total African ancestry in the EPIGEN cohorts. Scatterplot of the logarithm of the Blue/Mustard ancestry components ratio and the total African ancestry of individuals from Salvador (A), Bambuí (B) and Pelotas (C). We used Pearson's product-moment correlation coefficients to measure these correlations. In Salvador and Pelotas, individuals with more total African ancestry, tend to have proportionally more of the Blue ancestry cluster, which is associated to West Africa and non-Bantu populations.

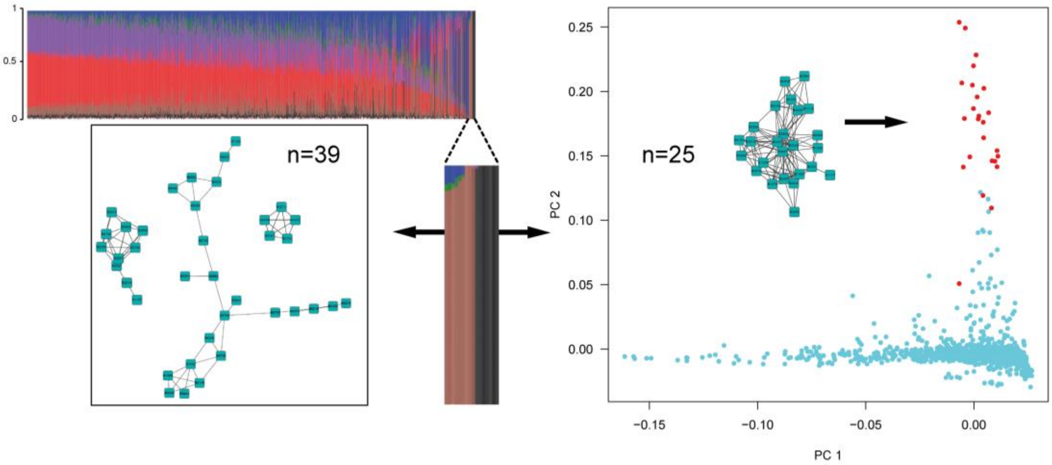


Figure S10. Familiar structure in Bambuí consistently identified by REAP, ADMIXTURE and Principal Component Analysis (PCA). When we used the entire set of EPIGEN individuals (Original Dataset), ADMIXTURE (K=7) identifies ancestry clusters (brown and black) that match a set of relatives identified by REAP kinship analysis and by our network approach (Section 6). Individuals from the black cluster were also identified by the second component of the PCA (red points) performed only for the Bambuí cohort.

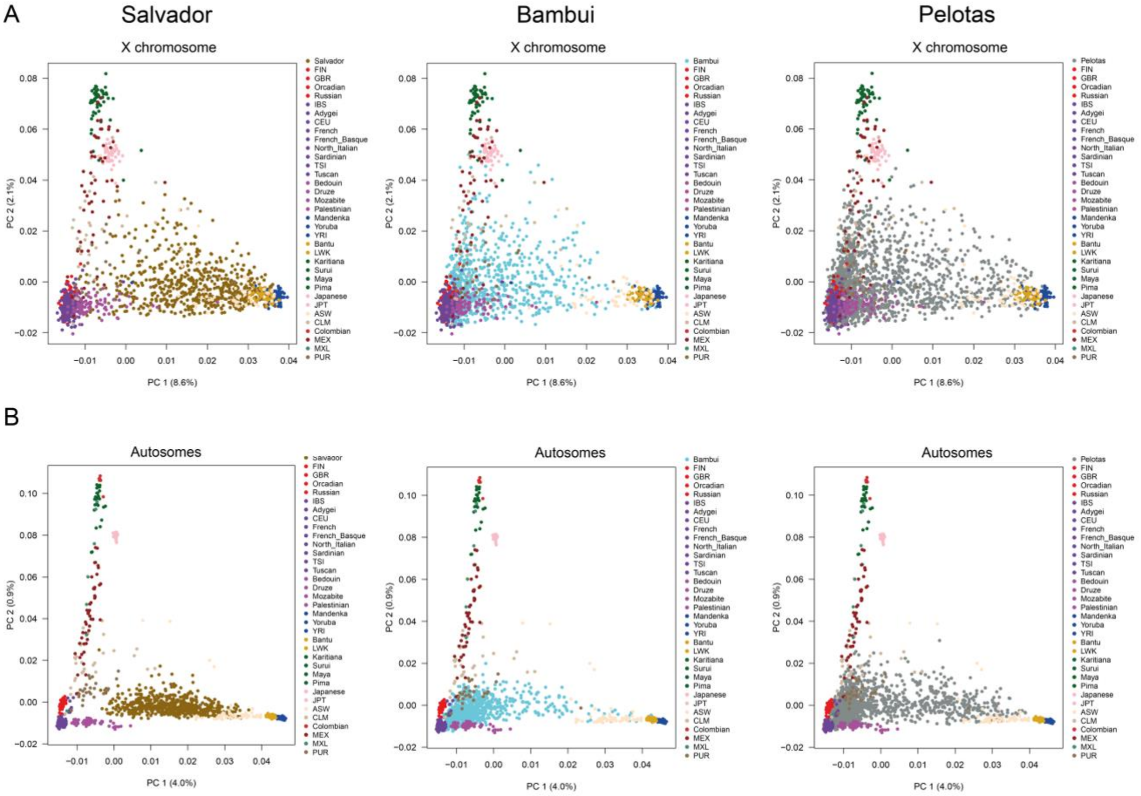


Figure S11. Principal Component Analysis of three Brazilian cohorts. (A) using X-chromosome SNPs and (B) autosomal markers for the same female individuals. Population acronyms are the same than in Figure 1.

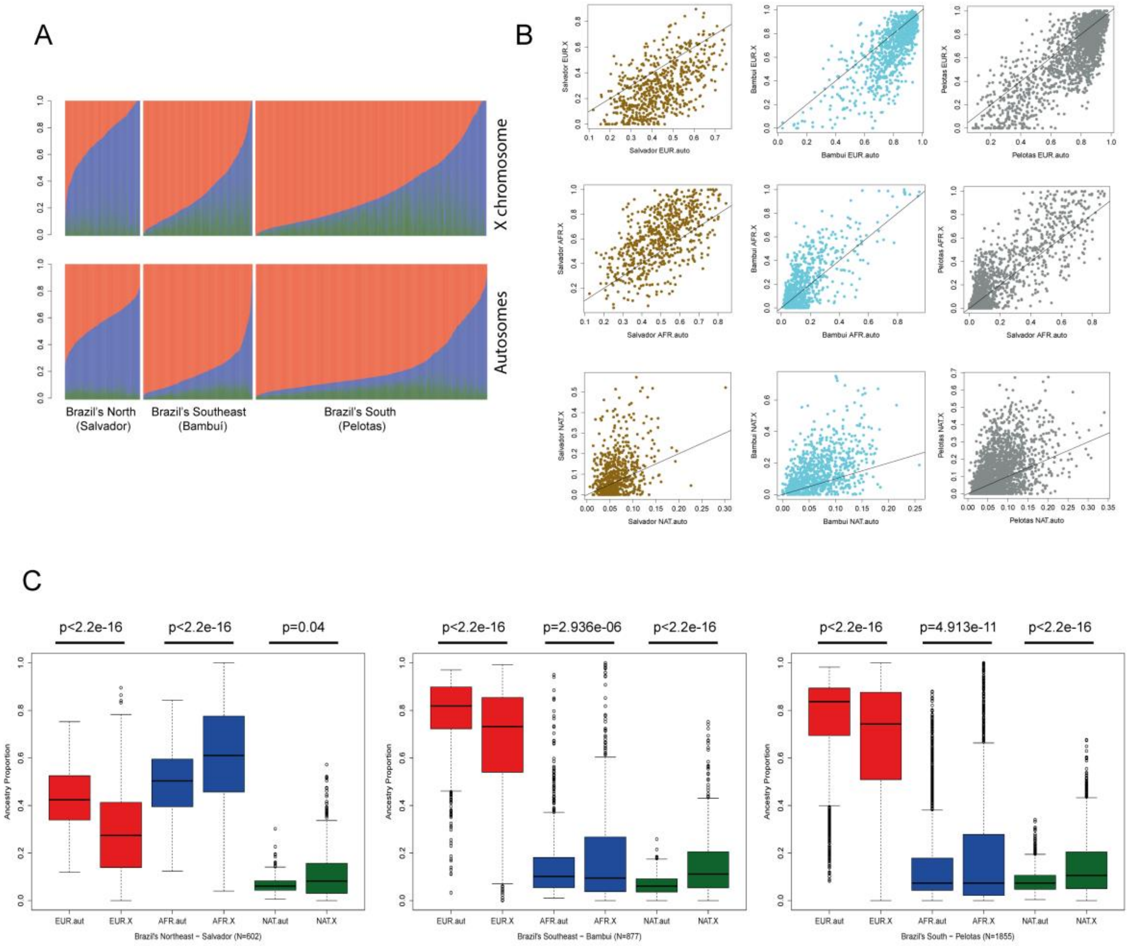


Figure S12. ADMIXTURE analysis on the X-chromosome and autosomal SNPs the on same females from the Brazilian EPIGEN populations, using the same set of parental populations. (A) Clusters obtained for K=3 (unsupervised mode) (B) Scatterplot of inferred autosomal continental ancestry (horizontal axis) vs. inferred X-chromosome continental ancestry for each individual analysed. (C) Boxplot of the distribution of continental ancestry for autosomes and X-chromosome data (p-value obtained by Wilcoxon Signed Rank Test on top). Res: European, Blue: African, Green: Native American ancestries.

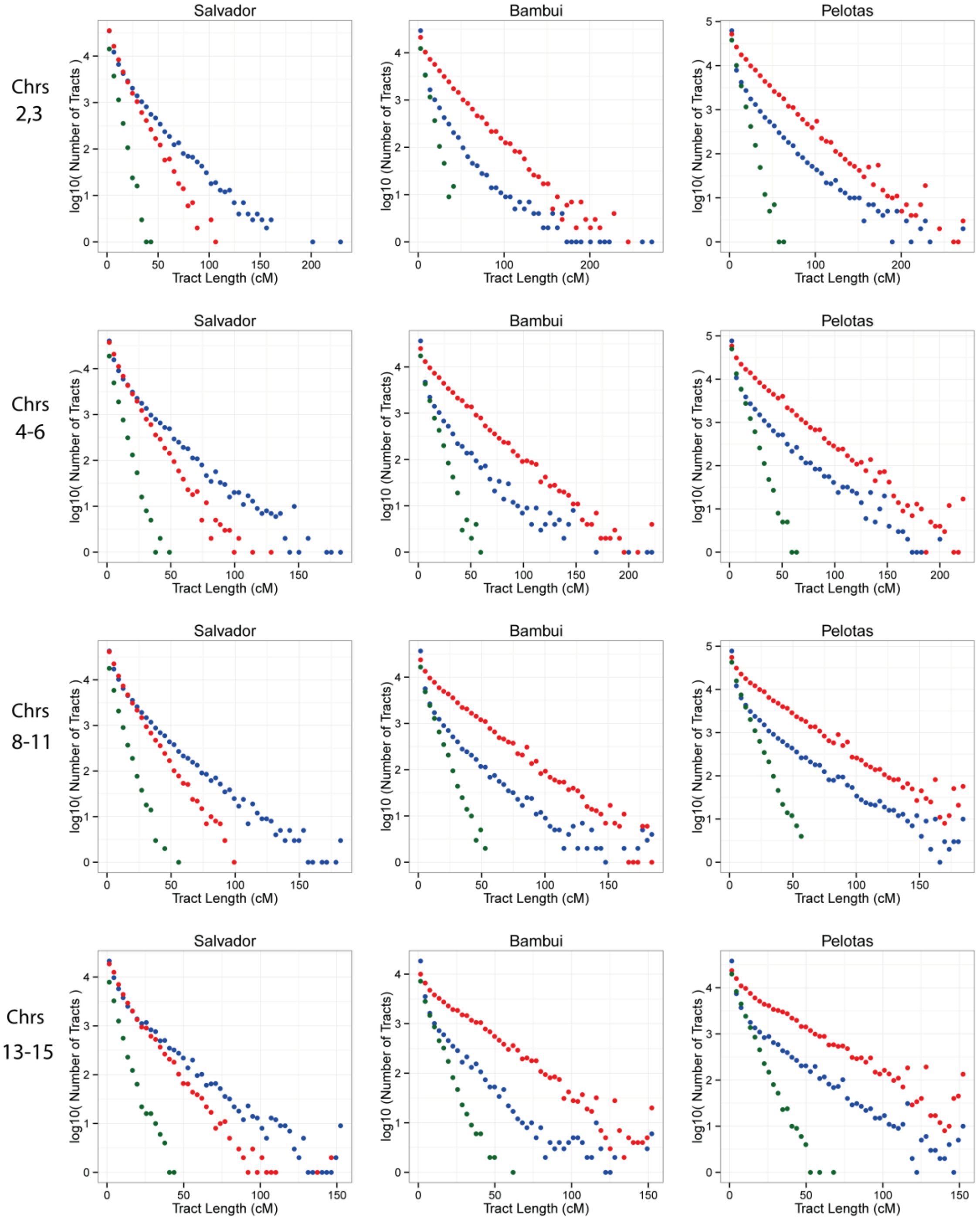


Figure S13. The distribution of lengths of chromosomal segments of continuous specific ancestry (CSSA) across the genome calculated for Salvador, Bambuí and Pelotas. CSSA lengths are organized in 50 equally spaced bins per population. We represented different sets of chromosomes with similar length. Green: Native American, Blue: African, Red: European ancestries.

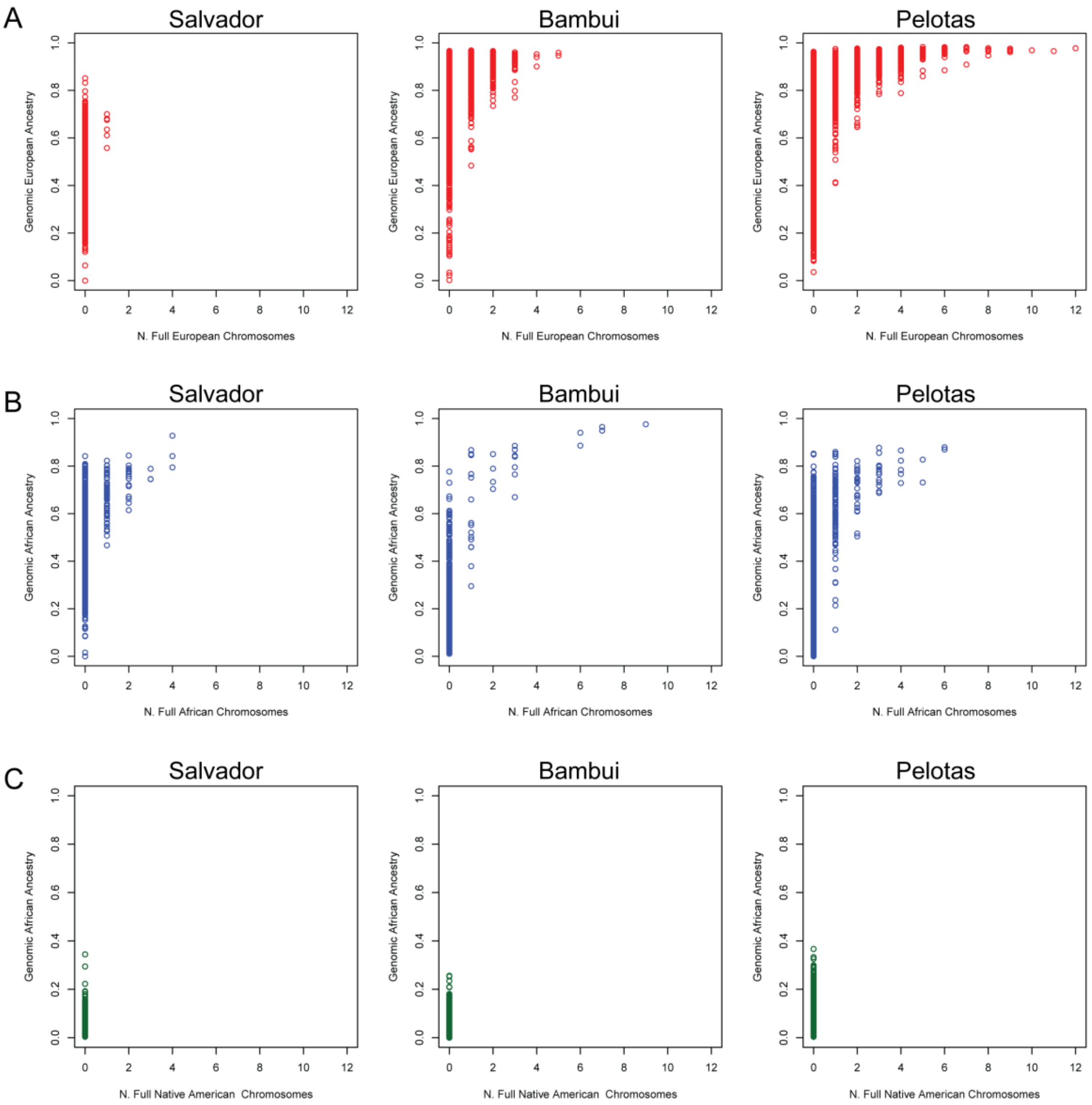


Figure S14. Observed number of full chromosomes from a unique European (A), African (B) and Native American (C) ancestry (horizontal axis)and total individual genomic ancestry (vertical axis).

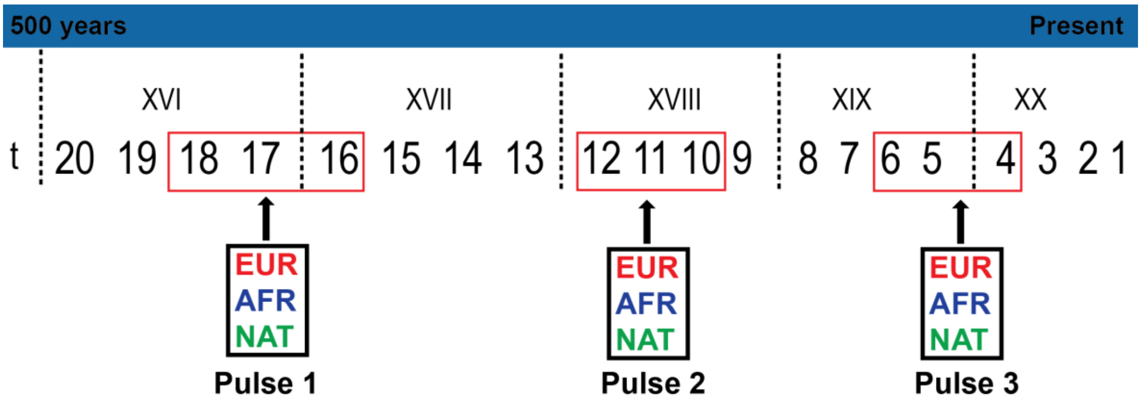


Figure S15. Admixture dynamics model for Brazil. Pulses of early (1), intermediate (2) and recent (3) continental admixture along the last five centuries (roman numbers) considered in the admixture dynamics model. **t** corresponds to the number of past generations and each generation corresponds to 25 years.

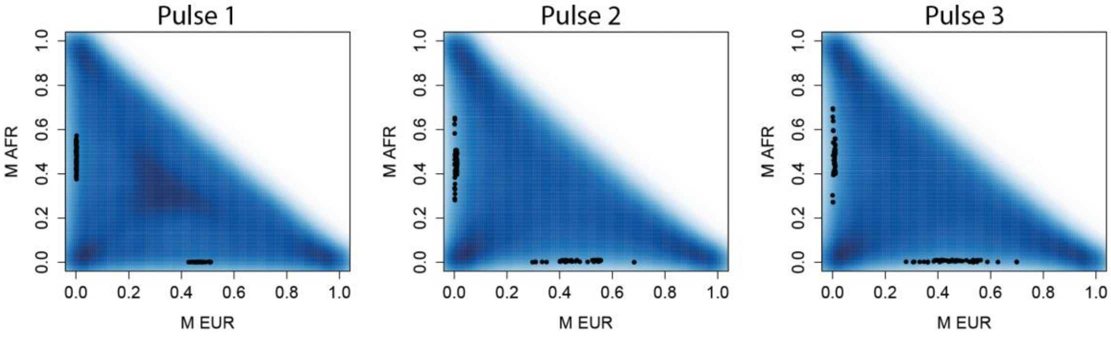
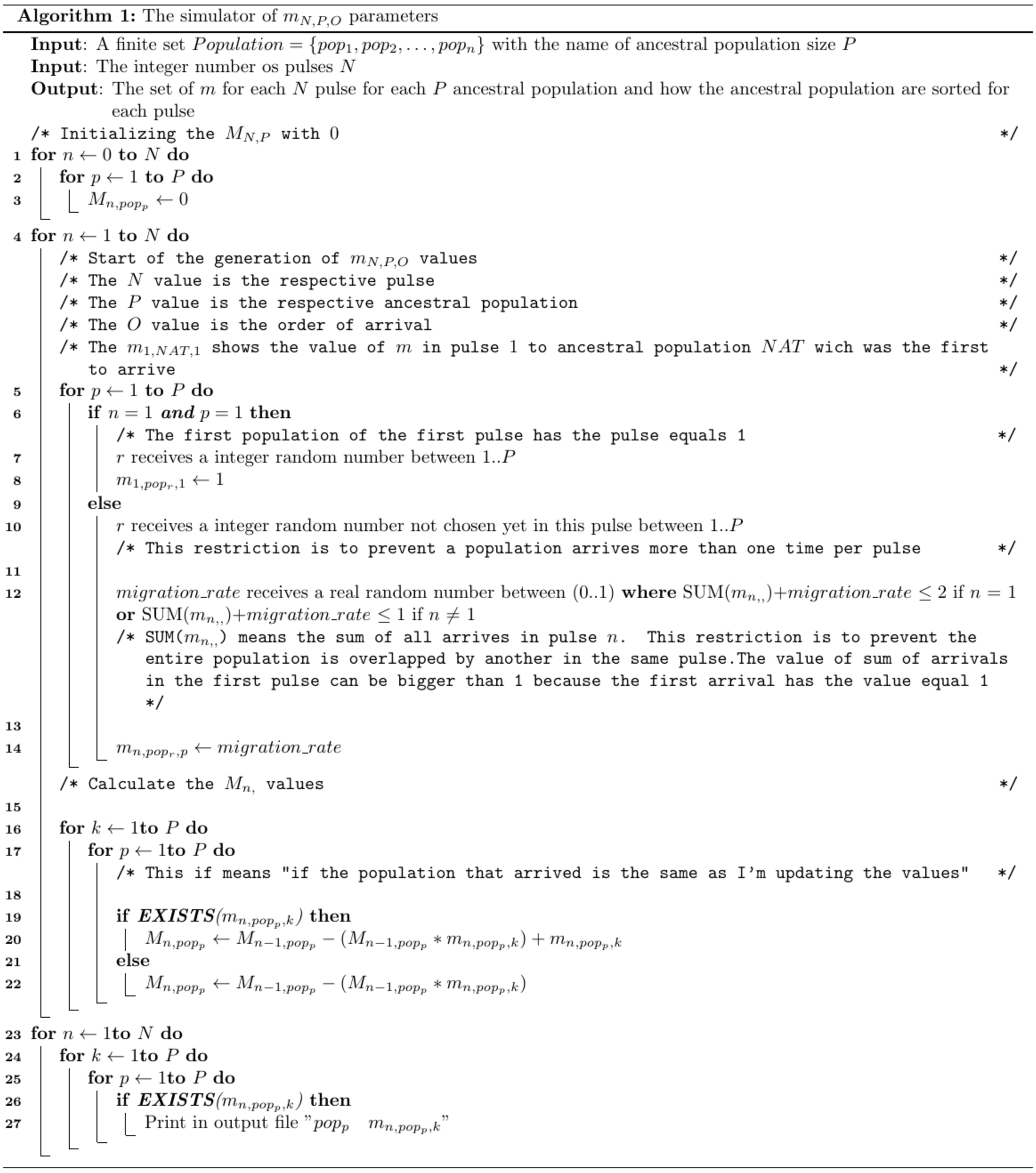


Figure S16. Exploring all the M space from a uniform m over the three pulses. The EUR and AFR **M** (cumulative population mean proportion of ancestry) space generated from uniform values of **m** (proportions of immigrants per pulse) over the **m** interval [0,1], as described in the text and over the three admixture pulses. This result suggests that the space of **M** values is adequately explored.



```
Algorithm 1: The simulator of $m_{N,P,O}$ parameters
Input: A finite set $Population = \{pop_1, pop_2, \ldots, pop_n\}$ with the name of ancestral population size $P$
Input: The integer number os pulses $N$
Output: The set of $m$ for each $N$ pulse for each $P$ ancestral population and how the ancestral population are sorted for
        each pulse
/* Initializing the $M_{N,P}$ with 0                                                    */
for $n \leftarrow 0$ to $N$ do
    for $p \leftarrow 1$ to $P$ do
        $M_{n,pop_p} \leftarrow 0$
for $n \leftarrow 1$ to $N$ do
    /* Start of the generation of $m_{N,P,O}$ values                                    */
    /* The $N$ value is the respective pulse                                             */
    /* The $P$ value is the respective ancestral population                              */
    /* The $O$ value is the order of arrival                                             */
    /* The $m_{1,NAT,1}$ shows the value of $m$ in pulse 1 to ancestral population $NAT$ wich was the first
       to arrive                                                                         */
    for $p \leftarrow 1$ to $P$ do
        if $n = 1$ and $p = 1$ then
            /* The first population of the first pulse has the pulse equals 1           */
            $r$ receives a integer random number between $1..P$
            $m_{1,pop_r,1} \leftarrow 1$
        else
            $r$ receives a integer random number not chosen yet in this pulse between $1..P$
            /* This restriction is to prevent a population arrives more than one time per pulse    */

            $migration\_rate$ receives a real random number between $(0..1)$ where SUM$(m_{n,,})$+$migration\_rate \leq 2$ if $n = 1$
            or SUM$(m_{n,,})$+$migration\_rate \leq 1$ if $n \neq 1$
            /* SUM$(m_{n,,})$ means the sum of all arrives in pulse $n$.  This restriction is to prevent the
               entire population is overlapped by another in the same pulse.The value of sum of arrivals
               in the first pulse can be bigger than 1 because the first arrival has the value equal 1
               */

            $m_{n,pop_r,p} \leftarrow migration\_rate$
    /* Calculate the $M_{n,}$ values                                                     */

    for $k \leftarrow$ 1to $P$ do
        for $p \leftarrow$ 1to $P$ do
            /* This if means "if the population that arrived is the same as I'm updating the values"   */

            if EXISTS$(m_{n,pop_p,k})$ then
                $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k}) + m_{n,pop_p,k}$
            else
                $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k})$
for $n \leftarrow$ 1to $N$ do
    for $k \leftarrow$ 1to $P$ do
        for $p \leftarrow$ 1to $P$ do
            if EXISTS$(m_{n,pop_p,k})$ then
                Print in output file "$pop_p$   $m_{n,pop_p,k}$"
```

Figure S17. Pseudocode to generate the distribution of the parameters of the demographic model of admixture used for the admixture dynamics inferences. The parameters are ***m*** (proportions of immigrants) and **M** (proportion of ancestry) conditioned on the observed continental admixture.

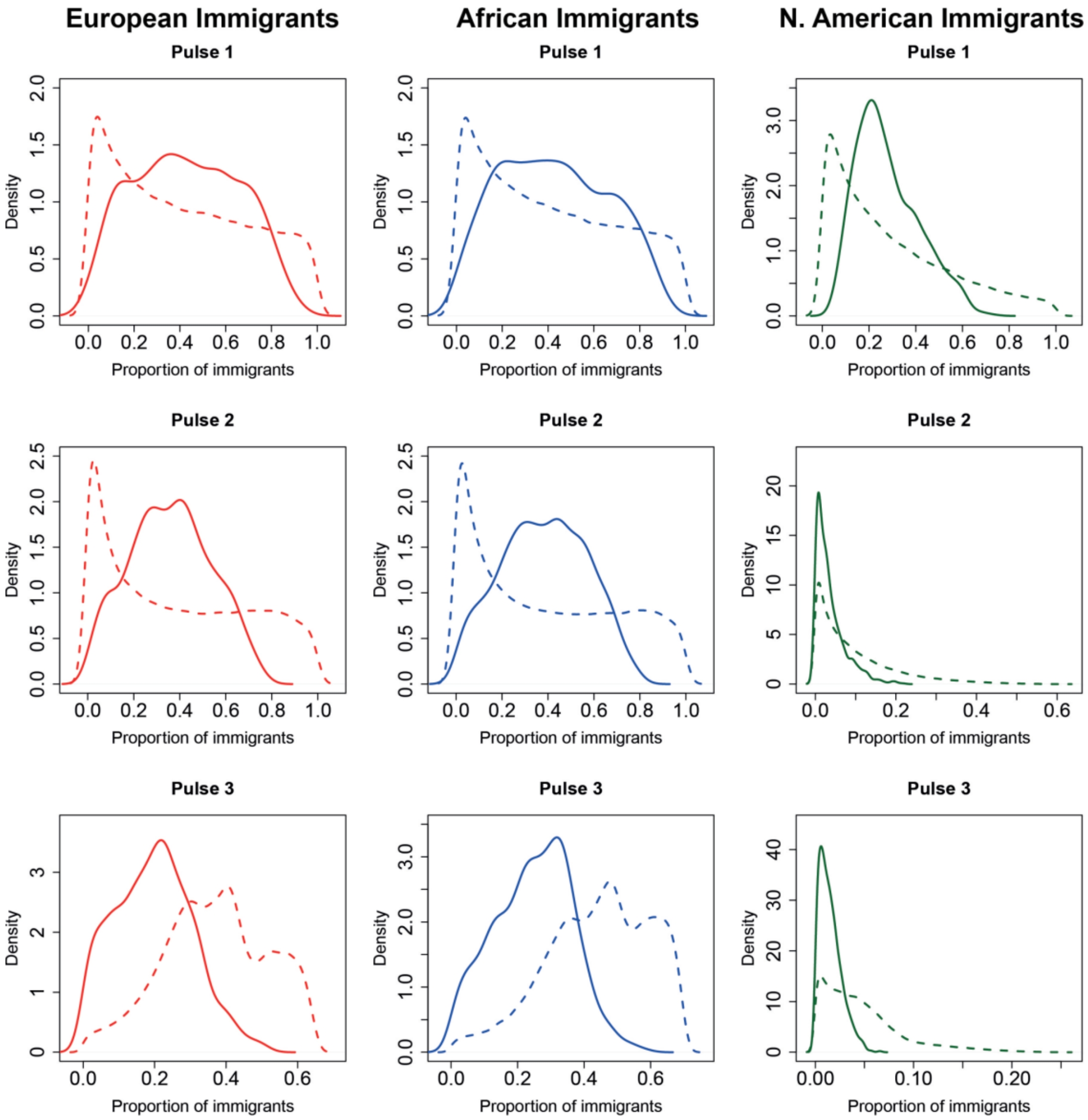


Figure S18. Posterior probability distributions of the 9 $\mathbf{m}_{n,P}$ (Admixture parameters) for Salvador (Northeastern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters $\mathbf{m}_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to $\mathbf{m}_{Europeans,P}$, blue lines ($\mathbf{m}_{African,P}$) and green lines ($\mathbf{m}_{N. American,P}$).

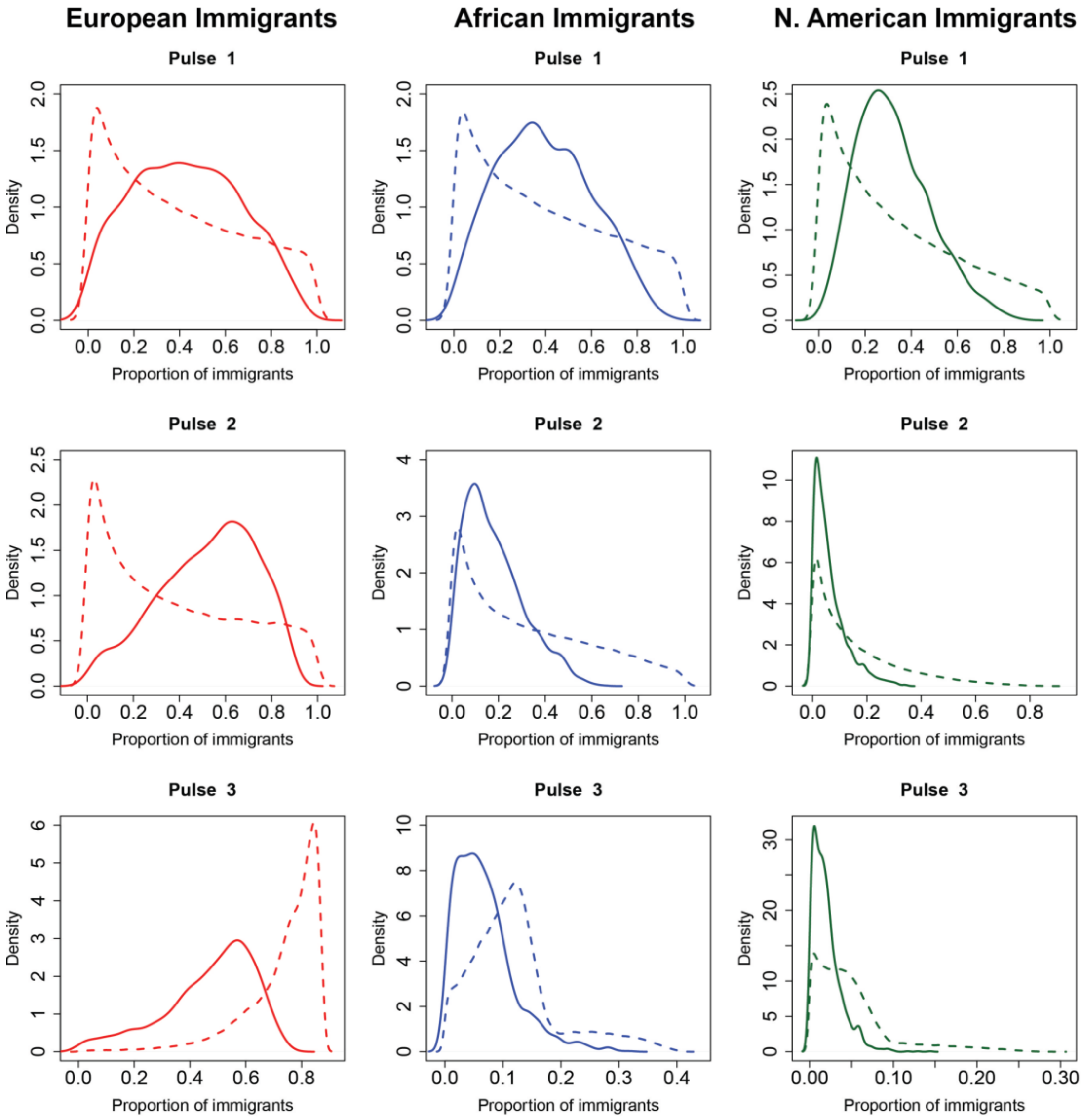


Figure S19. Posterior probability distributions of the 9 $m_{n,P}$ (Admixture parameters) for Bambuí (Southeastern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters $\mathbf{m}_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to $\mathbf{m}_{Europeans,P}$, blue lines ($\mathbf{m}_{African,P}$) and green lines ($\mathbf{m}_{N. American,P}$).

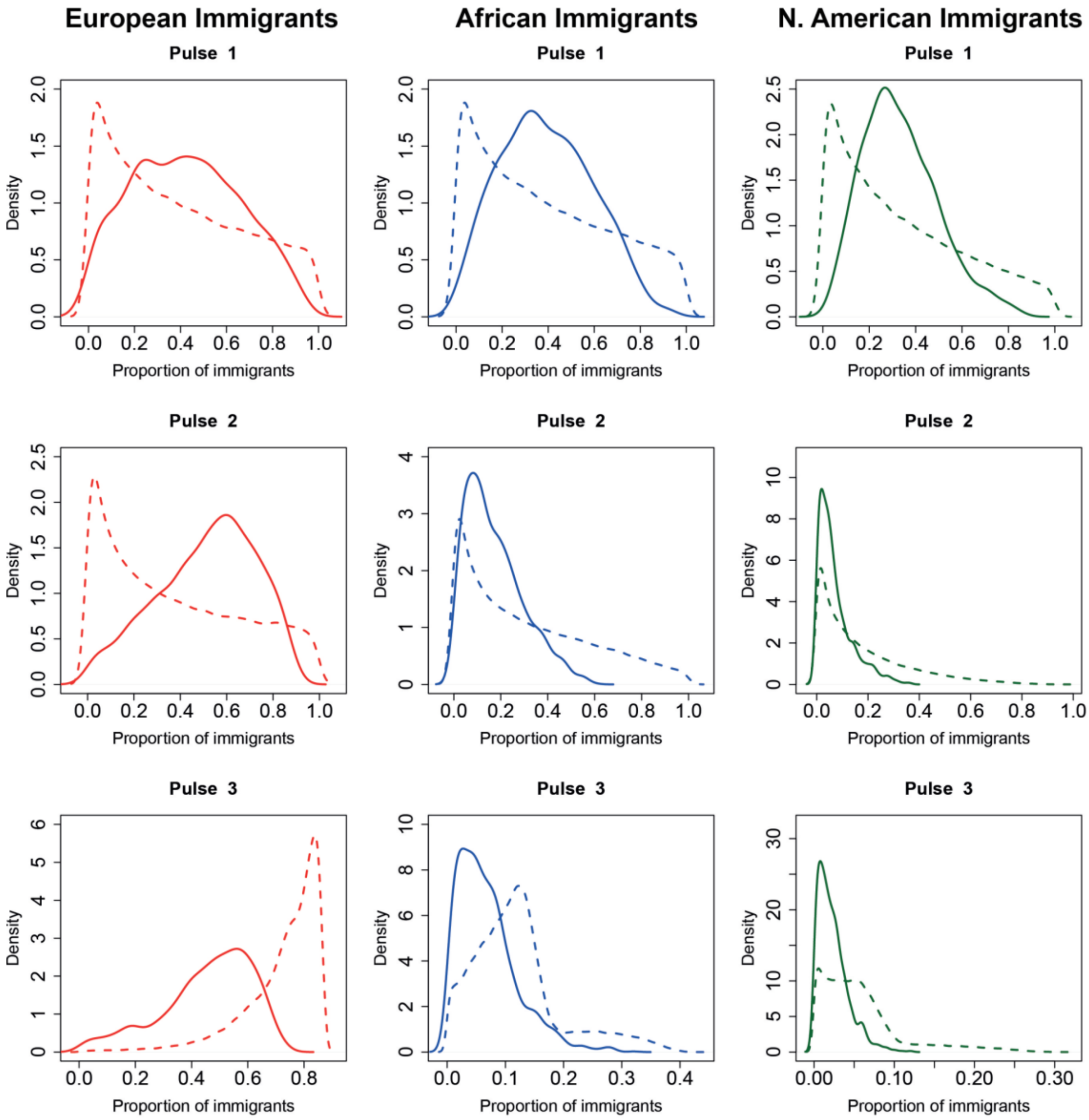


Figure S20. Posterior probability distributions of the 9 $m_{n,P}$ (Admixture parameters) for Pelotas (Southtern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters $\mathbf{m}_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to $\mathbf{m}_{Europeans,P}$, blue lines ($\mathbf{m}_{African,P}$) and green lines ($\mathbf{m}_{N. American,P}$).

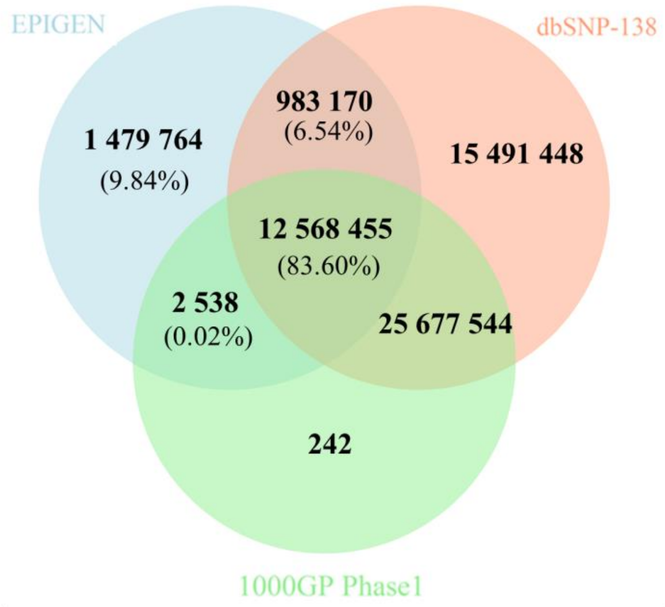


Figure S21. Venn diagram of the distribution of the 15,033,927 SNPs identified in the 30 Brazilian genomes and the intersection with the databases dbSNP-138 and 1000 Genomes Phase 1 SNPs. Percentages refer to the EPIGEN SNPs.

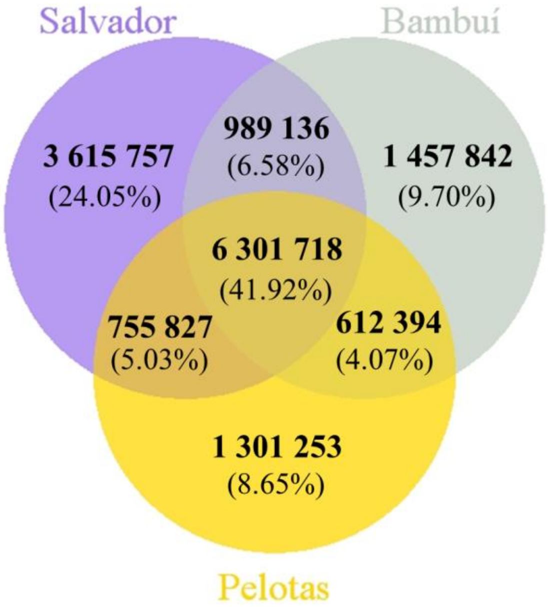


Figure S22. Distribution of the 15,033,927 SNPs identified in the 30 Brazilian genomes among the three studied Brazilian populations.

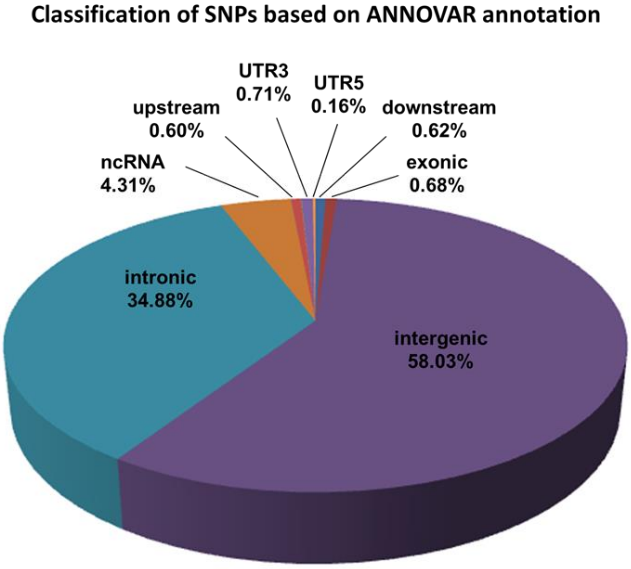


Figure S23. Distribution of biallelic SNPs based on their functional annotation by ANNOVAR using RefSeq database. The (upstream, downstream) category (0.02%) does not appear in the graphic. The (upstream, downstream) variants are located both downstream and upstream region (possibly for 2 different genes).

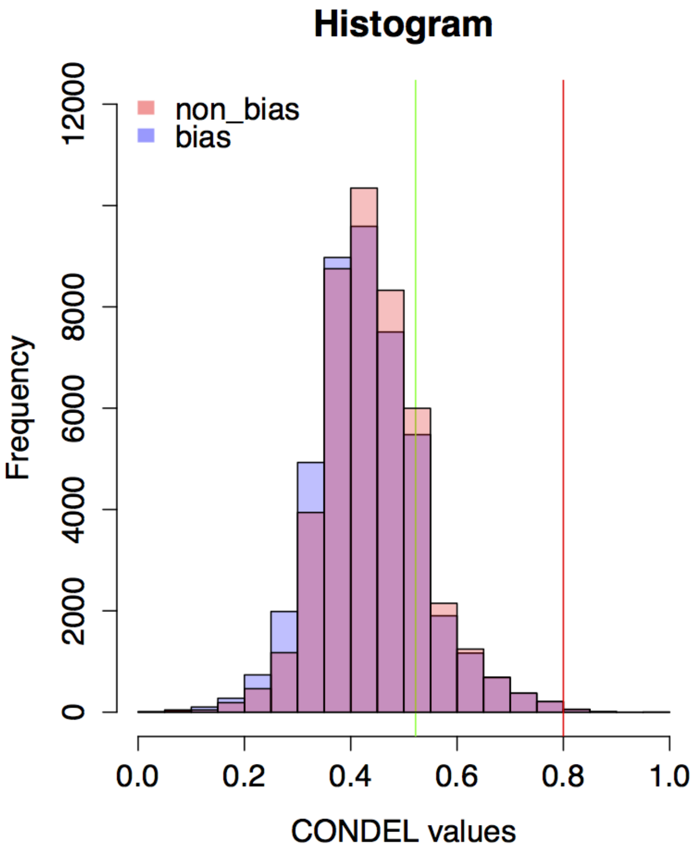


Figure S24. Condel scores distribution of autosomal non-synonymous SNPs with bias and with bias correction. The cutoff of 0.52 for deleterious variants is showed by the green line and the cutoff of 0.80 for very deleterious variants is showed by the red line.

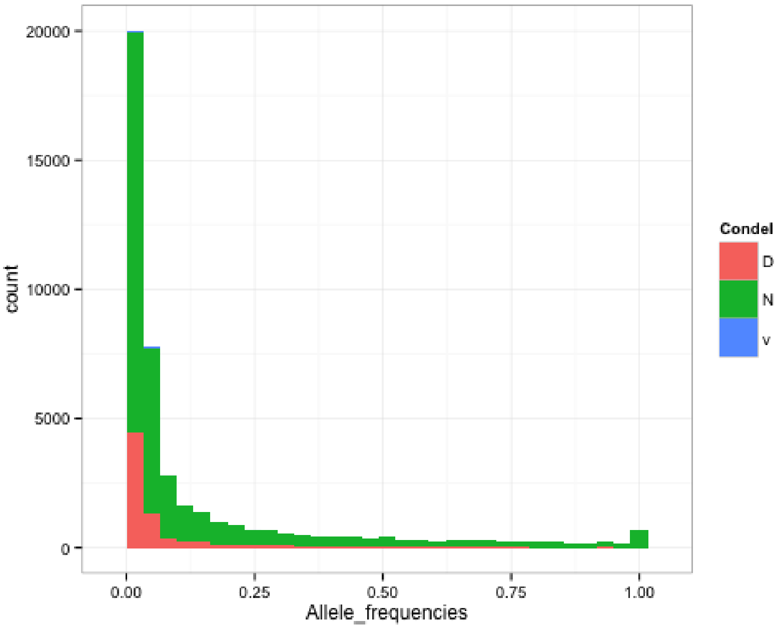


Figure S25. Allele frequency spectrum of autosomal non-synonymous SNPs before correcting the bias reported by Simon et al.[30], stratified by deleterious (D), normal (N), and very deleterious (V) predictions.

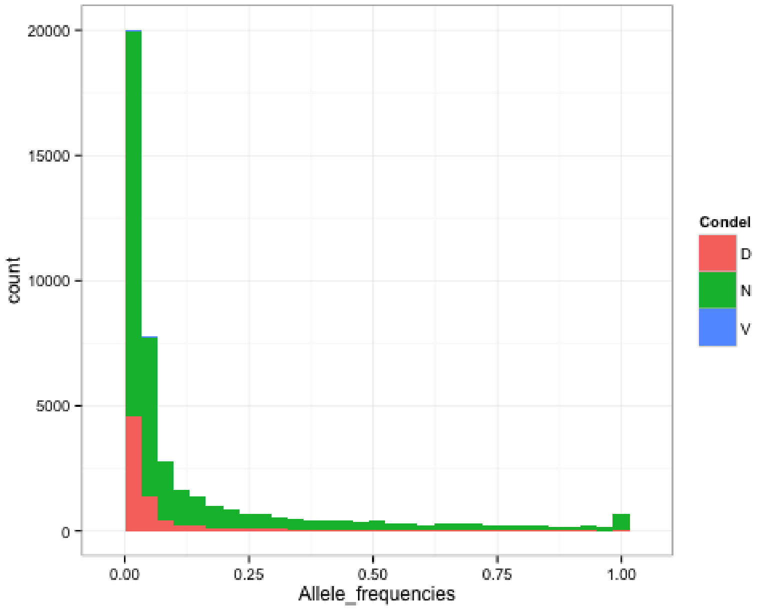


Figure S26. Allele frequency spectrum of autosomal non-synonymous SNPs after correcting the bias reported by Simon et al.[30], stratified by deleterious (D), normal (N), and very deleterious (V) predictions.

11. TABLES

Table S1. Data cleaning summary for the 2.5M and 5M datasets. Filter 1 removes SNPs with zeroed chromosomes and Filter 2 removes repeated SNPs.

Datasets	Initial SNPs	Excluded SNPs		Final SNPs
		Filter 1	Filter 2	
2.5M	2,361,093	6,926	5,570	2,348,597
5M	4,252,517	8,654	5,832	4,238,031

Table S2. Dataset separation and 2.5M-5M consensus.

Datasets	Total SNPs	Autosomal	Mit/X/Y SNPs	Samples
2.5M	2,348,597	2,293,235	55,362	6,504
5M	4,238,031	4,123,873	114,158	270
Consensus		**2,256,647**	**49,709**	**6,774**

Table S3. Quality Control summary for consensus autosomal SNPs. Filter 4 is the PLINK geno filter and Filter 5 is the inconsistent-SNPs-to-be-removed list.

Datasets	Initial SNPs	Excluded SNPs		Cohort Merge	Final SNPs
		Filter 4	Filter 5		
Bambuí	2,256,647	1,469	21,513		2,233,665
Pelotas	2,256,647	135	21,527		2,234,985
Salvador	2,256,647	365	21,507		2,234,775
Total	2,256,647	1,969	21,527	2,256,636	2,235,109

Table S4. Quality Control summary for the consensus autosomal dataset samples. Filter 1 is the PLINK mind filter, Filter 2 is sample duplicates, and Filter 3 is the sex check filter.

Datasets	Initial samples	Excluded Samples			Final samples
		Filter 1	Filter 2	Filter 3	
Bambuí	1,502	46	14	0	1,442
Pelotas	3,858	81	40	1	3,736
Salvador	1,414	87	14	4	1,309
TOTAL	6,774	214	68	5	**6,487**

Table S5. Quality Control summary for consensus Mitochondrial, X- and Y- chromosome samples. Individuals were excluded based on the --mind filter of the PLINK software.

Datasets	**X-chromosomal Samples**			**Y-chromosomal Samples**			**Mitochondrial Samples**		
	Initial	Excluded	**Final**	Initial	Excluded	**Final**	Initial	Excluded	**Final**
Bambuí	1,442	1	1,441	564	2	562	1,442	0	1,442
Pelotas	3,735	0	3,735	1,880	7	1,873	3,736	1	3,735
Salvador	1,309	0	1,309	707	0	707	1,309	1	1,308

Table S6. Quality Control summary for consensus Mitochondrial, X- and Y- chromosome SNPs. SNPs were excluded based on the --geno filter of the PLINK software.

Datasets	**X-chromosomal SNPs**			**Y-chromosomal SNPs**			**Mitochondrial SNPs**		
	Initial	Excluded	**Final**	Initial	Excluded	**Final**	Initial	Excluded	**Final**
Bambuí	46,945	45	**46,900**	2,153	38	**2,115**	220	7	**213**
Pelotas	46,945	43	**46,902**	2,153	9	**2,144**	220	2	**218**
Salvador	46,945	39	**46,906**	2,153	17	**2,136**	220	4	**216**

Table S7. Data summary for the HapMap (phase II+III) frozen datasets.

HapMap Populations*	**N of individuals**	**N of Autossomal SNPS**	**N of ChrX SNPS**	**N of ChrY SNPS**	**N of mtDNA SNPS**
ASW	83	1,506,278	54,720	384	71
CEU	174	3,907,239	122,601	722	212
CHB	86	3,928,480	122,933	716	207
CHD	85	1,265,389	40,409	354	44
GIH	88	1,362,120	45,322	376	59
JPT	89	3,928,521	122,979	716	207
LWK	90	1,475,622	53,704	367	71
MEX	77	1,363,399	46,475	357	34
MKK	171	1,483,727	53,486	348	77
TSI	88	1,374,150	45,376	335	60
YRI	176	3,860,794	122,642	710	210

*ASW, African ancestry in Southwest USA; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria.

Table S8. Summary of HGDP frozen datasets, divided by population (644,246 autosomal SNPs, 16,471 X-chromosome SNPs, 25 Y-chromosome SNPs and 163 mitochondrial-SNPs).

HGDP Populations	Geographic Origin	N of individuals
Adygei	Russia Caucasus	17
Balochi	Pakistan	25
Bantu	Kenya/South Africa	20
Bedouin	Israel (Negev)	48
Biaka_Pygmies	Central African Republic	32
Brahui	Pakistan	25
Burusho	Pakistan	25
Cambodians	Cambodia	11
Colombians	Colombia	15
Daí	China	10
Daur	China	9
Druze	Israel (Carmel)	47
French_Basque	France	24
French	France	29
Han	China	44
Hazara	Pakistan	24
Hezhen	China	9
Japanese	Japan	29
Kalash	Pakistan	25
Karitiana	Brazil	22
Lahu	China	10
Makrani	Pakistan	25
Mandenka	Senegal	27
Maya	Mexico	25
Mbuti_Pygmeu	Democratic Republic of Congo	15
Miaozu	China	10
Mongola	China	10
Mozabite	Algeria (Mzab)	30
NAN_Melanesian	Bougainville	19
Naxi	China	9
North_Italian	Italy (Bergamo)	13
Orcadian	Orkney Island	16
Oroqen	China	10
Palestinian	Israel (Central)	51
Papuan	New Guinea	17
Pathan	Pakistan	23
Pima	Mexico	25
Russian	Russia	25
San	Namibia	6
Sardinian	Italy	28
She	China	10
Sindhi	Pakistan	25
Surui	Brazil	21
Tujia	China	10
Tuscan	Italy	8
Tu	China	10
Uygur_China	China	10
Xibo	China	9
Yakut	Siberia	25
Yizu	China	10
Yoruba	Nigeria	21

Table S9. Summary of individuals and populations for 2,132,104 autosomal SNPs in the 1000 Genomes Project phase I frozen datasets.

1000 Genomes Populations*	N of individuals
ASW	61
CEU	85
CHB	97
CHS	100
CLM	60
FIN	93
GBR	89
IBS	14
JPT	89
LWK	97
MXL	66
PUR	55
TSI	98
YRI	88

*ASW, Americans of African Ancestry in SW USA ; CEU, Utah Residents (CEPH) with Northern and Western European ancestry; CHB, Han Chinese in Bejing, China; CHS, Southern Han Chinese ; CLM, Colombians from Medellin, Colombia ; FIN, Finnish in Finland ; GBR, British in England and Scotland ; IBS, Iberian population in Spain ; JPT, Japanese in Tokyo, Japan ; LWK, Luhya in Webuye, Kenya; MXL, Mexican Ancestry from Los Angeles USA ; PUR, Puerto Ricans from Puerto Rico ; TSI, Toscani in Italia ; YRI, Yoruba in Ibadan, Nigeira.

Table S10. Number of SNPs per chromosome and populations present in the phased 1000 Genomes phase I frozen datasets.

Chr	Number of SNPs*
Chr1	2,980,130
Chr2	3,277,861
Chr3	2,739,531
Chr4	2,712,965
Chr5	2,509,110
Chr6	2,404,770
Chr7	2,196,168
Chr8	2,164,645
Chr9	1,638,291
Chr10	1,866,772
Chr11	1,877,176
Chr12	1,811,857
Chr13	1,361,289
Chr14	1,245,407
Chr15	1,120,852
Chr16	1,199,899
Chr17	1,035,965
Chr18	1,079,340
Chr19	807,096
Chr20	847,692
Chr21	512,682
Chr22	489,301

* ASW (N=61), CEU (N=85), CHB (N=97), CHS (N=100), CLM (N=60), FIN (N=93), GBR (N=89), IBS (N=14), JPT (N=89), LWK (N=97), MXL (N=66), PUR (N=55), TSI (N=98), YRI (N=88)

Table S11. Number of samples per population of the Original Dataset in the integrated autosomal dataset (Original Dataset in the Main Text).

Populations	N	Dataset
Adygei	17	HGDP
Ashanincas	44	LDGH
ASW	97	HapMap/1000G
Bambuí	**1,442**	**EPIGEN**
Bantu	20	HGDP
Bedouin	48	HGDP
CEU	173	HapMap/1000G
CLM	60	1000G
Colombians	15	HGDP
Druze	47	HGDP
FIN	93	1000G
French	29	HGDP
French_Basque	24	HGDP
GBR	89	1000G
IBS	14	1000G
Japanese	29	HGDP
JPT	100	HapMap/1000G
Karitiana	22	HGDP
LWK	100	HapMap/1000G
Mandenka	27	HGDP
Maya	25	HGDP
MEX/MXL	97	HapMap/1000G
Mozabite	30	HGDP
North_Italian	13	HGDP
Orcadian	16	HGDP
Palestinian	51	HGDP
Pelotas	**3,736**	**EPIGEN**
Pima	25	HGDP
PUR	55	1000G
Russian	25	HGDP
Salvador	**1,309**	**EPIGEN**
Sardinian	28	HGDP
Shimaa	45	LDGH
Surui	21	HGDP
TSI	98	HapMap/1000G
Tuscan	8	HGDP
Yoruba	21	HGDP
YRI	174	HapMap/1000G
TOTAL	8,267	-

Table S12. Number of SNPs per chromosome in the integrated original autosomal dataset.

Chromosome	N SNPs	Chromosome	N SNPs
Chr1	25,504	Chr12	16,246
Chr2	27,078	Chr13	12,418
Chr3	22,858	Chr14	11,235
Chr4	19,766	Chr15	10,646
Chr5	21,049	Chr16	10,583
Chr6	21,189	Chr17	9,139
Chr7	18,118	Chr18	10,495
Chr8	19,194	Chr19	5,998
Chr9	16,546	Chr20	9,110
Chr10	17,917	Chr21	5,175
Chr11	16,469	Chr22	5,057
TOTAL		331,790	

Table S13. Number of relatedness samples excluded from each EPIGEN cohort and non-related remaining samples. (Dataset U).

Cohort	N excluded samples	N non-related samples
Salvador	63	1,246
Bambuí	516	926
Pelotas	83	3,653
Total	662	5,825

Table S14. Number of females per population of the Original Dataset in the integrated X-chromosome dataset.

Populations	N	Data Base
Adygei	10	HGDP
ASW	53	HapMap/1000G
Bambuí	**877**	**EPIGEN**
Bantu	1	HGDP
Bedouin	20	HGDP
CEU	92	HapMap/1000G
CLM	31	1000G
Colombians	8	HGDP
Druze	33	HGDP
FIN	58	1000G
French	17	HGDP
French_Basque	8	HGDP
GBR	48	1000G
IBS	7	1000G
Japanese	7	HGDP
JPT	46	HapMap/1000G
Karitiana	14	HGDP
LWK	50	HapMap/1000G
Mandenka	8	HGDP
Maya	23	HGDP
MEX/MXL	54	HapMap/1000G
Mozabite	10	HGDP
North_Italian	5	HGDP
Orcadian	9	HGDP
Palestinian	34	HGDP
Pelotas	**1,855**	**EPIGEN**
Pima	11	HGDP
PUR	27	1000G
Russian	9	HGDP
Salvador	**602**	**EPIGEN**
Sardinian	12	HGDP
Surui	10	HGDP
TSI	48	HapMap/1000G
Tuscan	2	HGDP
Yoruba	12	HGDP
YRI	81	HapMap/1000G
TOTAL	4,192	-

Table S15. Number of SNPs per chromosome shared between populations used in local ancestry analyses.

Chromosome	N of common SNPs
1	160,082
2	170,715
3	144,131
4	134,702
5	128,184
6	125,346
7	113,418
8	111,173
9	91,189
10	104,935
11	101,906
12	98,591
13	73,697
14	67,464
15	63,634
16	66,998
17	57,352
18	61,054
19	40,491
20	50,165
21	28,214
22	28,927

Table S16. Genetic differentiation (F_{ST}) matrix between ADMIXTURE ancestry clusters obtained with K=8.

K=8	🟪	🟩	🟥	🟥	🟦	🟩	🟨	🟦
🟪								
🟩	0.174							
🟥	0.03	0.158						
🟥	0.118	0.129	0.11					
🟦	0.029	0.167	0.029	0.115				
🟩	0.215	0.141	0.202	0.173	0.21			
🟨	0.141	0.222	0.142	0.161	0.137	0.261		
🟦	0.144	0.224	0.146	0.163	0.141	0.263	0.019	
🟪	0.031	0.172	0.042	0.114	0.039	0.215	0.131	0.135

Table S17. Mean sub-continental proportions for the Mustard (East-associated, EAFR) and Blue (West Africa – associated, WAFR) ancestry clusters of the 3 EPIGEN populations and the Afro-American population ASW, Colombians (CLM), Mexicans (MEX) and Puerto Ricans (PUR) from HapMap.

Mean	**Bambuí**	**Pelotas**	**Salvador**	**ASW**	**CLM**	**MEX**	**PUR**
Blue	0.095	0.087	0.378	0.632	0.052	0.030	0.094
Mustard	0.053	0.068	0.126	0.130	0.030	0.019	0.029
Ratio Blue/Mustard	1.79	1.30	3.00	4.85	1.74	1.60	3.22

Table S18. Mean Contributions and sex-bias of Europeans (EUR), Africans (AFR) and Native Americans (NAT) ancestry for X-chromosome and autosomal data.

Parental Contributions	**Salvador females**	**Bambuí females**	**Pelotas females**
EUR Autosomal	0.43	0.78	0.76
EUR X-chromosome	0.29	0.67	0.67
Mean bias*	**0.15**	**0.11**	**0.09**
AFR Autosomal	0.50	0.15	0.16
AFR X-chromosome	0.60	0.18	0.19
Mean bias*	**-0.10**	**-0.03**	**-0.03**
NAT Autosomal	0.07	0.07	0.08
NAT X-chromosome	0.11	0.15	0.14
Mean bias*	**-0.04**	**-0.08**	**-0.06**

* The mean of the differences between autosomal minus X-chromosome ancestry

Table S19. Absolut numbers and frequencies of all mitochondrial haplogroups and sub-haplogroups inferred by HaploGrep.

Absolut Numbers/Frequencies				
mt-haplogroup	Salvador	Bambui	Pelotas	Total
A	41 / 0.0313	0 / 0	154 / 0.0412	195 / 0.0301
A2a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
A7	1 / 0.0008	0 / 0	11 / 0.0029	12 / 0.0019
B2	43 / 0.0329	223 / 0.1546	287 / 0.0768	553 / 0.0853
B2b	22 / 0.0168	33 / 0.0229	33 / 0.0088	88 / 0.0136
B4a	0 / 0	0 / 0	20 / 0.0054	20 / 0.0031
B4b	2 / 0.0015	0 / 0	0 / 0	2 / 0.0003
B5	0 / 0	5 / 0.0035	2 / 0.0005	7 / 0.0011
C	43 / 0.0329	78 / 0.0541	133 / 0.0356	254 / 0.0392
C1a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
C1b	0 / 0	35 / 0.0243	0 / 0	35 / 0.0054
C1c	6 / 0.0046	21 / 0.0146	35 / 0.0094	62 / 0.0096
C1d	10 / 0.0076	27 / 0.0187	35 / 0.0094	72 / 0.0111
C4b	2 / 0.0015	0 / 0	3 / 0.0008	5 / 0.0008
C7a	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
D1j	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
D4	16 / 0.0122	8 / 0.0055	79 / 0.0212	103 / 0.0159
D4g	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
H	0 / 0	12 / 0.0083	0 / 0	12 / 0.0019
H1	10 / 0.0076	42 / 0.0291	259 / 0.0693	311 / 0.048
H11	0 / 0	0 / 0	5 / 0.0013	5 / 0.0008
H13	1 / 0.0008	0 / 0	14 / 0.0037	15 / 0.0023
H15	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H17	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H1a	0 / 0	14 / 0.0097	21 / 0.0056	35 / 0.0054
H1b	1 / 0.0008	0 / 0	15 / 0.004	16 / 0.0025
H1c	2 / 0.0015	6 / 0.0042	92 / 0.0246	100 / 0.0154
H1h	4 / 0.0031	0 / 0	2 / 0.0005	6 / 0.0009
H1n	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H2a	10 / 0.0076	15 / 0.0104	145 / 0.0388	170 / 0.0262
H2c	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3	5 / 0.0038	8 / 0.0055	102 / 0.0273	115 / 0.0177
H30	0 / 0	0 / 0	26 / 0.007	26 / 0.004
H3g	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3h	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3u	0 / 0	0 / 0	3 / 0.0008	3 / 0.0005
H3x	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H4	0 / 0	0 / 0	18 / 0.0048	18 / 0.0028
H4a	0 / 0	0 / 0	44 / 0.0118	44 / 0.0068
H5a	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
H6	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H6a	1 / 0.0008	2 / 0.0014	11 / 0.0029	14 / 0.0022

H7a	1 / 0.0008	0 / 0	1 / 0.0003	2 / 0.0003
H7d	0 / 0	0 / 0	12 / 0.0032	12 / 0.0019
H45	0 / 0	1 / 0.0007	0 / 0	1 / 0.0002
H60	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
HV	1 / 0.0008	2 / 0.0014	32 / 0.0086	35 / 0.0054
HV0	4 / 0.0031	19 / 0.0132	119 / 0.0319	142 / 0.0219
HV5	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
I	1 / 0.0008	0 / 0	3 / 0.0008	4 / 0.0006
I1a	0 / 0	17 / 0.0118	1 / 0.0003	18 / 0.0028
I2	4 / 0.0031	0 / 0	14 / 0.0037	18 / 0.0028
I5a	1 / 0.0008	7 / 0.0049	0 / 0	8 / 0.0012
J1	4 / 0.0031	19 / 0.0132	27 / 0.0072	50 / 0.0077
J1b	0 / 0	5 / 0.0035	9 / 0.0024	14 / 0.0022
J1c	0 / 0	14 / 0.0097	60 / 0.0161	74 / 0.0114
J2	4 / 0.0031	3 / 0.0021	23 / 0.0062	30 / 0.0046
J2a	1 / 0.0008	8 / 0.0055	22 / 0.0059	31 / 0.0048
K	0 / 0	1 / 0.0007	0 / 0	1 / 0.0002
K1	1 / 0.0008	19 / 0.0132	39 / 0.0104	59 / 0.0091
K1a	3 / 0.0023	0 / 0	55 / 0.0147	58 / 0.0089
K1b	0 / 0	0 / 0	5 / 0.0013	5 / 0.0008
K1c	0 / 0	0 / 0	18 / 0.0048	18 / 0.0028
K2b	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
L0a	88 / 0.0673	67 / 0.0465	85 / 0.0228	240 / 0.037
L0b	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
L0d	3 / 0.0023	0 / 0	33 / 0.0088	36 / 0.0056
L1	1 / 0.0008	0 / 0	0 / 0	1 / 0.0002
L1b	91 / 0.0696	56 / 0.0388	57 / 0.0153	204 / 0.0315
L1c	128 / 0.0979	101 / 0.07	147 / 0.0394	376 / 0.058
L2a	232 / 0.1774	50 / 0.0347	191 / 0.0511	473 / 0.0729
L2b	6 / 0.0046	0 / 0	1 / 0.0003	7 / 0.0011
L2c	29 / 0.0222	3 / 0.0021	17 / 0.0046	49 / 0.0076
L2d	7 / 0.0054	0 / 0	3 / 0.0008	10 / 0.0015
L2e	1 / 0.0008	1 / 0.0007	0 / 0	2 / 0.0003
L3	99 / 0.0757	33 / 0.0229	147 / 0.0394	279 / 0.043
L3b	62 / 0.0474	37 / 0.0257	22 / 0.0059	121 / 0.0187
L3c	67 / 0.0512	8 / 0.0055	29 / 0.0078	104 / 0.016
L3d	64 / 0.0489	12 / 0.0083	37 / 0.0099	113 / 0.0174
L3e	101 / 0.0772	65 / 0.0451	124 / 0.0332	290 / 0.0447
L3f	36 / 0.0275	21 / 0.0146	37 / 0.0099	94 / 0.0145
L3h	6 / 0.0046	2 / 0.0014	8 / 0.0021	16 / 0.0025
L3i	1 / 0.0008	0 / 0	1 / 0.0003	2 / 0.0003
L3k	5 / 0.0038	0 / 0	0 / 0	5 / 0.0008
L3x	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
L4b	6 / 0.0046	0 / 0	9 / 0.0024	15 / 0.0023
L5a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
M	0 / 0	2 / 0.0014	3 / 0.0008	5 / 0.0008
M1	0 / 0	2 / 0.0014	66 / 0.0177	68 / 0.0105
M5a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002

N	0 / 0	188 / 0.1304	2 / 0.0005	190 / 0.0293
N14	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
N15	3 / 0.0023	8 / 0.0055	9 / 0.0024	20 / 0.0031
N1a	0 / 0	0 / 0	35 / 0.0094	35 / 0.0054
N1b	2 / 0.0015	2 / 0.0014	2 / 0.0005	6 / 0.0009
N2	1 / 0.0008	1 / 0.0007	106 / 0.0284	108 / 0.0167
P7	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
T	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
T1	0 / 0	7 / 0.0049	27 / 0.0072	34 / 0.0052
T1a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
T2	2 / 0.0015	13 / 0.009	146 / 0.0391	161 / 0.0248
T2b	1 / 0.0008	0 / 0	25 / 0.0067	26 / 0.004
T2f	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U	4 / 0.0031	0 / 0	1 / 0.0003	5 / 0.0008
U2	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U2d	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
U2e	2 / 0.0015	3 / 0.0021	10 / 0.0027	15 / 0.0023
U3a	0 / 0	8 / 0.0055	6 / 0.0016	14 / 0.0022
U4	2 / 0.0015	21 / 0.0146	25 / 0.0067	48 / 0.0074
U4b	0 / 0	39 / 0.027	2 / 0.0005	41 / 0.0063
U5	0 / 0	0 / 0	29 / 0.0078	29 / 0.0045
U5a	3 / 0.0023	13 / 0.009	38 / 0.0102	54 / 0.0083
U5b	1 / 0.0008	6 / 0.0042	120 / 0.0321	127 / 0.0196
U6	7 / 0.0054	20 / 0.0139	54 / 0.0145	81 / 0.0125
U6a	1 / 0.0008	0 / 0	10 / 0.0027	11 / 0.0017
U6b	0 / 0	0 / 0	3 / 0.0008	3 / 0.0005
U7	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U8a	0 / 0	0 / 0	4 / 0.0011	4 / 0.0006
V1	1 / 0.0008	0 / 0	7 / 0.0019	8 / 0.0012
V2	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
V7	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
V7a	0 / 0	0 / 0	13 / 0.0035	13 / 0.002
W3a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
Y	0 / 0	3 / 0.0021	0 / 0	3 / 0.0005
TOTAL	1308 / 1	1442 / 1	3735 / 1	6485 / 1

Table S20. Population genetics indices based on the haplogroup and subhaplogroup distribution in the three Brazilian EPIGEN cohorts.

Mitochondrial DNA	**Salvador**	**Bambuí**	**Pelotas**
n. individuals	1,308	1,442	3,735
n. inferred different haplogrups[1]	62	59	111
Gene diversity (SD)[1]	0.926 (0.003)	0.938 (0.003)	0.969 (0.001)
Admixture estimates			
African	78.9%	31.7%	25.4%
European	6.8%	38.2%	53.1%
Native American	14.2%	29.9%	21.5%

Y-chromosome	**Salvador**	**Bambuí**	**Pelotas**
n. individuals	707	562	1,873
n. inferred different haplogrups[2]	51	43	60
Gene diversity (SD)[2]	0.881(0.009)	0.814(0.016)	0.868(0.007)
Admixture estimates			
African	28%	12.5%	11%
European	70%	87.0%	87.6%
Native American	1.8%	0.5%	1.4%

[1] Based on Table S19.

[2] Expected haplogroups/sub-haplogroups heterozygosity based on frequencies of Table S23. SD: standard deviation.

TableS21. Absolut numbers and frequencies of continental biogeographic assignments of mt-haplogroups.

Ancestry	**Absolut Numbers/ Frequencies**				
	mt-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
Native American	A	42/0.2258	0/0.0000	167/0.2080	209
	B	67/0.3602	261/0.6042	342/0.4259	670
	C	61/0.3280	163/0.3773	208/0.2590	432
	D	16/0.0860	8/0.0185	86/0.1071	110
	Total	**186**	**432**	**803**	**1421**
European	H	35/0.4217	102/0.2948	787/0.4484	924
	HV	5/0.0602	21/0.0607	152/0.0866	178
	I	6/0.0723	24/0.0694	18/0.0103	48
	J	9/0.1084	49/0.1416	141/0.0803	199
	K	4/0.0482	20/0.0578	119/0.0678	143
	T	3/0.0361	20/0.0578	203/0.1157	226
	U	20/0.2410	110/0.3179	307/0.1749	437
	V	1/0.0120	0/0.0000	28/0.0160	29
	Total	**83**	**346**	**1755**	**2184**
Asian	M	0/0.0000	4/0.0194	70/0.3084	74
	N	6/1.0000	199/0.9660	155/0.6828	360
	P	0/0.0000	0/0.0000	1/0.0044	1
	Y	0/0.0000	3/0.0146	0/0.0000	3
	W	0/0.0000	0/0.0000	1/0.0044	1
	Total	**6**	**206**	**227**	**439**
African	L0a	88/0.0852	67/0.1463	85/0.0895	240
	L0b	0/0.0000	0/0.0000	1/0.0011	1
	L0d	3/0.0029	0/0.0000	33/0.0347	36
	L1	1/0.0010	0/0.0000	0/0.0000	1
	L1b	91/0.0881	56/0.1223	57/0.0600	204
	L1c	128/0.1239	101/0.2205	147/0.1547	376
	L2a	232/0.2246	50/0.1092	191/0.2011	473
	L2b	6/0.0058	0/0.0000	1/0.0011	7
	L2c	29/0.0281	3/0.0066	17/0.0179	49
	L2d	7/0.0068	0/0.0000	3/0.0032	10
	L2e	1/0.0010	1/0.0022	0/0.0000	2
	L3	99/0.0958	33/0.0721	147/0.1547	279
	L3b	62/0.0600	37/0.0808	22/0.0232	121
	L3c	67/0.0649	8/0.0175	29/0.0305	104
	L3d	64/0.0620	12/0.0262	37//0.0389	113
	L3e	101/0.0978	65/0.1419	124/0.1305	290
	L3f	36/0.0348	21/0.0459	37/0.0389	94
	L3h	6/0.0058	2/0.0044	8/0.0084	16
	L3i	1/0.0010	0/0.0000	1/0.0011	2
	L3k	5/0.0048	0/0.0000	0/0.0000	5
	L3x	0/0.0000	2/0.0044	0/0.0000	2
	L4b	6/0.0058	0/0.0000	9/0.0095	15
	L5a	0/0.0000	0/0.0000	1/0.0011	1
	Total	**1033**	**458**	**950**	**2441**

Table S22. Genetic differentiation (F_{ST}) between the three EPIGEN cohorts estimated from mt-DNA (upper matrix) and Y-chromosome haplogrups (lower matrix)[1].

	Salvador	Bambuí	Pelotas
Salvador		0.0344[2]	0.0236[2]
Bambuí	0.1394[2]		0.0188[2]
Pelotas	0.0079[2]	0.1339[2]	

[1] F_{ST} are estimated by Arlequin based on haplotype frequencies Tables S19 and S23 assuming the infinite allele model.

[2] $P < 10^{-5}$ based on a randomization test of individuals among populations (5,000 replicates of the test).

Table S23. Absolut numbers and frequencies of all Y chromosome sub-haplogroups.

	Absolut Numbers / Frequencies			
Y-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
A3b2*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
B2a1a2a2*	3 / 0.0042	1 / 0.0018	8 / 0.0043	12 / 0.0038
B2b*	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
B2b1*	0 / 0	1 / 0.0018	2 / 0.0011	3 / 0.001
D2*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
DE*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
E1a*	3 / 0.0042	0 / 0	7 / 0.0037	10 / 0.0032
E1a1	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
E1b1a1*	1 / 0.0014	0 / 0	1 / 0.0005	2 / 0.0006
E1b1a1a1a	0 / 0	0 / 0	5 / 0.0027	5 / 0.0016
E1b1a1a1f*	17 / 0.024	1 / 0.0018	4 / 0.0021	22 / 0.007
E1b1a1a1f1a*	2 / 0.0028	1 / 0.0018	0 / 0	3 / 0.001
E1b1a1a1f1a1*	72 / 0.1018	8 / 0.0142	32 / 0.0171	112 / 0.0356
E1b1a1a1g1*	45 / 0.0636	11 / 0.0196	38 / 0.0203	94 / 0.0299
E1b1a1a1g1a*	24 / 0.0339	4 / 0.0071	13 / 0.0069	41 / 0.013
E1b1b*	1 / 0.0014	1 / 0.0018	6 / 0.0032	8 / 0.0025
E1b1b1a*	11 / 0.0156	9 / 0.016	12 / 0.0064	32 / 0.0102
E1b1b1a2*	12 / 0.017	17 / 0.0302	47 / 0.0251	76 / 0.0242
E1b1b1a3b	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
E1b1b1b*	1 / 0.0014	0 / 0	4 / 0.0021	5 / 0.0016
E1b1b1b1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
E1b1b1b1b	33 / 0.0467	41 / 0.073	82 / 0.0438	156 / 0.0496
E1b1b1c*	3 / 0.0042	5 / 0.0089	8 / 0.0043	16 / 0.0051
E1b1b1c1* or E1b1b1c1a*	3 / 0.0042	9 / 0.016	9 / 0.0048	21 / 0.0067
E2b*	0 / 0	1 / 0.0018	0 / 0	1 / 0.0003
E2b1*	1 / 0.0014	1 / 0.0018	5 / 0.0027	7 / 0.0022
G1* or G1a*	1 / 0.0014	2 / 0.0036	4 / 0.0021	7 / 0.0022

G2a*	5 / 0.0071	3 / 0.0053	15 / 0.008	23 / 0.0073
G2a1c*	22 / 0.0311	8 / 0.0142	49 / 0.0262	79 / 0.0251
G2a1c1a	5 / 0.0071	6 / 0.0107	4 / 0.0021	15 / 0.0048
G2a1c2a1	3 / 0.0042	2 / 0.0036	3 / 0.0016	8 / 0.0025
G2a1c2b1a	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
I1*	15 / 0.0212	33 / 0.0587	89 / 0.0475	137 / 0.0436
I1a1c1	4 / 0.0057	2 / 0.0036	8 / 0.0043	14 / 0.0045
I2*	3 / 0.0042	3 / 0.0053	18 / 0.0096	24 / 0.0076
I2a1a1*	8 / 0.0113	8 / 0.0142	34 / 0.0182	50 / 0.0159
I2a2a*	16 / 0.0226	17 / 0.0302	56 / 0.0299	89 / 0.0283
I2a2b	1 / 0.0014	0 / 0	4 / 0.0021	5 / 0.0016
J1*	9 / 0.0127	8 / 0.0142	58 / 0.031	75 / 0.0239
J2*	21 / 0.0297	11 / 0.0196	78 / 0.0416	110 / 0.035
J2a1b2*	8 / 0.0113	13 / 0.0231	26 / 0.0139	47 / 0.015
J2a1b2a1*	5 / 0.0071	6 / 0.0107	9 / 0.0048	20 / 0.0064
J2b*	11 / 0.0156	4 / 0.0071	35 / 0.0187	50 / 0.0159
J2b1	2 / 0.0028	0 / 0	1 / 0.0005	3 / 0.001
L1* or L1b*	0 / 0	0 / 0	3 / 0.0016	3 / 0.001
L1b1	2 / 0.0028	4 / 0.0071	2 / 0.0011	8 / 0.0025
N1b1a*	2 / 0.0028	1 / 0.0018	3 / 0.0016	6 / 0.0019
O1b1a1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
Q1a2*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
Q1a2a1*	12 / 0.017	3 / 0.0053	24 / 0.0128	39 / 0.0124
Q1a4	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
Q1b1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
R1a1a*	0 / 0	0 / 0	3 / 0.0016	3 / 0.001
R1a1a1a*	7 / 0.0099	6 / 0.0107	66 / 0.0352	79 / 0.0251
R1b*	8 / 0.0113	0 / 0	0 / 0	8 / 0.0025
R1b1a2a*	14 / 0.0198	0 / 0	0 / 0	14 / 0.0045
R1b1a2a1*	217 / 0.3069	0 / 0	632 / 0.3374	849 / 0.2702
R1b1a2a1a*	14 / 0.0198	16 / 0.0285	37 / 0.0198	67 / 0.0213
R1b1a2a1a2b1	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
R1b1a2a1a2b2*	7 / 0.0099	7 / 0.0125	29 / 0.0155	43 / 0.0137
R1b1a2a1a2b2a1*	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
R1b1a2a1b1a1a1*	5 / 0.0071	8 / 0.0142	36 / 0.0192	49 / 0.0156
R1b1a2a1b2c*	10 / 0.0141	14 / 0.0249	45 / 0.024	69 / 0.022
R1b1a2a1b2c1a*	0 / 0	1 / 0.0018	8 / 0.0043	9 / 0.0029
R1b1a2a1b3*	25 / 0.0354	31 / 0.0552	112 / 0.0598	168 / 0.0535
R2a*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
Root	0 / 0	231 / 0.411	47 / 0.0251	278 / 0.0885

T*	0 / 0	1 / 0.0018	0 / 0	1 / 0.0003
T1*	8 / 0.0113	2 / 0.0036	37 / 0.0198	47 / 0.015
T1b*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
TOTAL	707 / 1	562 / 1	1873 / 1	3142 / 1

Table S24. Continental biogeographic assignment distribution of Y chromosome haplogroups.

Ancestry	**Absolut Numbers/Frequencies**				
	Y-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
Native American	Q	13/1.0000	3/1.0000	26/1.0000	42
	Total	**13**	**3**	**26**	**42**
European	G	36/0.0783	21/0.0469	77/0.0495	134
	I	47/0.1022	63/0.1406	209/0.1343	319
	J	56/0.1217	42/0.0938	207/0.1330	305
	L	2/0.0043	4/0.0089	5/0.0032	11
	N	2/0.0043	1/0.0022	3/0.0019	6
	R	308/0.6696	83/0.1853	971/0.6240	1362
	T	9/0.0196	3/0.0067	37/0.0238	49
	Root	0/0.0000	231/0.5156	47/0.0302	278
	Total	**460**	**448**	**1556**	**2464**
Asian	D	1/1.0000	0/0.0000	1/0.5000	2
	O	0/0.0000	0/0.0000	1/0.5000	1
	Total	**1**	**0**	**2**	**3**
African	A	1/0.0043	0/0.0000	0/0.0000	1
	B	3/0.129	2/0.0180	12/0.0415	17
	E	229/0.9828	109/0.9820	277/0.9585	615
	Total	**233**	**111**	**289**	**633**

Table S25. GWAS hits for SNPs differentiated between Blue (West Africa, non-Bantu-associated) and mustard (East Africa/Bantu associated) ADMIXTURE clusters (K=9).

Disease / Trait	**N.SNPs**	**SNP list (38)**	**F_{ST}[1]**
Cognitive performance	3	rs2807580	0.0941
		rs2229741	0.0707
		rs4751674	0.0703
Crohn's disease	3	rs7702331	0.0750
		rs7517847	0.0603
		rs6556412	0.0599
Inflammatory bowel disease	3	rs477515	0.1261
		rs2382817	0.0683
		rs7517847	0.0603
Multiple sclerosis	2	rs12466022	0.0688
		rs533259	0.0688
Obesity related	2	rs7964120	0.1322
		rs7784447	0.0957
Emphysema-related traits	1	rs641525	0.1469
Epstein-Barr virus immune response	1	rs477515	0.1261
Liver enzyme levels	1	rs4547811	0.1108
Schizophrenia	1	rs1635	0.0862
Myopia (pathological)	1	rs4142248	0.0825
Alzheimer's disease	1	rs610932	0.0822
F-cell distribution	1	rs7565301	0.0738
Amyotrophic lateral sclerosis	1	rs2819332	0.0726
Menopause	1	rs11889862	0.0725
Eosinophil counts	1	rs4143832	0.0719
Obsessive-compulsive disorder	1	rs9652236	0.0717
HIV related	1	rs1020064	0.0716
Sphingolipid levels	1	rs1000778	0.0689
IgE levels in asthmatics	1	rs10404342	0.0673
Economic and political preferences	1	rs210648	0.0667
Bladder cancer	1	rs2294008	0.0660
Duodenal ulcer	1	rs2294008	0.0660
Nasopharyngeal carcinoma	1	rs6774494	0.0660
Non-alcoholic fatty liver disease histology	1	rs887304	0.0658
Prostate cancer	1	rs4242382	0.0652
Resp.to irinotecan/platinum-based chemo. lung cancer	1	rs344924	0.0647
Sudden cardiac arrest	1	rs5762311	0.0637
Type 1 diabetes	1	rs1004446	0.0634
Bipolar disorder	1	rs7250872	0.0626
Response to gemcitabine in pancreatic cancer	1	rs1901440	0.0625
Mean platelet volume	1	rs12526480	0.0625
Pancreatic cancer	1	rs10088262	0.0620
Breast size	1	rs7104745	0.0612

Bold indicates unique entries and underline indicate co-occurrence in OMIM disease results.

[1] The list is sorted by decreasing F_{ST}.

Table S26 - Summary of the data after EPIGEN QC analysis.

	EPIGEN – 30 Brazilians
Coverage	42.7x
% Called genome fraction	93
% mapped reads	87.73
% Array agreement Omni2.5	99.27
Ts/Tv	2.04
% Array agreement HumanOmni5	99.53
Total SNPs	**15,033,927**
Average of Indels/lenght	714,436 / (20-300)

Table S27. Definitions of functional categories of ANNOVAR.

Functional category	**Definition**
Exonic	variant overlaps a coding exon, excluding the 5'UTR and 3'UTR
Synonymous	a single nucleotide change that does not cause an amino acid change
Non-synonymous	a single nucleotide change that cause an amino acid change
Stopgain	a SNV that lead to the immediate creation of stop codon at the variant site. This class is not included in the Non-synonymous class.
Stoploss	a SNV that lead to the immediate elimination of stop codon at the variant site. This class is not included in the Non-synonymous class.
Unknown	unknown function (due to various errors in the gene structure definition in the database file)
Splicing	variant is within 2-bp of a splicing junction
ncRNA	variant overlaps a transcript without coding annotation in the gene definition
UTR5	variant overlaps a 5' untranslated region
UTR3	variant overlaps a 3' untranslated region
Intronic	variant overlaps an intron
Upstream	variant overlaps 1-kb region upstream of transcription start site
Downstream	variant overlaps 1-kb region downstream of transcription end site
Intergenic	variant is in intergenic region

**Adapted from ANNOVAR website (http://www.openbioinformatics.org/annovar/annovar_gene.html)*

Table S28. Proportion of synonymous and non-synonymous exonic SNPs in the 30 Brazilian genomes and in similar studies.

Study	# Samples	Coverage	% of Synonymous	% of Non-synonymous
EPIGEN – current study	30	42.7x	49.91	47.88
1000 Genomes Project et al.[44]	1,092	~50x*	44.59	50.63
Lachance et al.[57]	15	~60x	~43.21	~45.69
Shen et al. [58]	44	65.8x	45.80	52.50

* coverage of exomes.

Table S29. Exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes, based on RefSeq database.

Exonic	Number of SNPs on the 30 samples	% of SNPs
Non-synonymous	50518	49.91
Synonymous	48464	47.88
Stopgain	563	0.56
Stoploss	45	0.05
Unknown	1621	1.60
Total	101211	100

Table S30. Exonic SNPs classified by VEP (Ensembl) in the 30 Brazilian genomes.

Exonic	Number of SNPs on the 30 samples	% of SNPs
Missense	58142	53.52
Synonymous	49419	45.49
Stop_gained	890	0.82
Stop_lost	177	0.16
Coding sequence	6	0.01
Total	108634	100

Table S31. Exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes, based on the Ensembl transcripts database.

Exonic	**Number of SNPs on the 30 samples**	**% of SNPs**
Non-synonymous	57066	51.68
Synonymous	50516	45.75
Stopgain	841	0.76
Stoploss	137	0.12
Unknown	1857	1.68
Total	110417	100

Table S32. CONDEL scores for the derived/non-reference and derived/reference SNPs from 30 genomes as a function of allele frequency classes.

Allele frequency classes*	**# Variants analyzed by Condel**	**% Variants analyzed**	**Average Condel score**	**# Variants analyzed by Condel**	**% Variants analyzed**	**Average Condel score**	**Bias[1]**
	Derived/non-reference SNPs			**Derived/reference SNPs**			
0 – 0.10	30171	79.794	0.452	1361	21.433	0.351	0.101
0.11 – 0.20	3055	8.080	0.430	573	9.024	0.357	0.074
0.21 – 0.30	1629	4.308	0.426	527	8.299	0.349	0.078
0.31 – 0.40	1005	2.658	0.424	404	6.362	0.346	0.078
0.41 – 0.50	740	1.957	0.419	489	7.701	0.347	0.073
0.51 – 0.60	447	1.182	0.420	496	7.811	0.342	0.078
0.61 – 0.70	298	0.788	0.424	526	8.283	0.350	0.074
0.71 – 0.80	234	0.619	0.410	476	7.496	0.347	0.063
0.81 – 0.90	137	0.362	0.411	465	7.323	0.348	0.063
0.91 – 1.0	95	0.251	0.403	1033	16.268	0.340	0.063
Total	37811	100	-	6350	100	0.347	-

* In EPIGEN individuals

[1] Bias = CONDEL score$_{\text{derived/non-reference}}$ – CONDEL score$_{\text{derived/reference}}$.