Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire

Daniel N. Harris[a,b,c,1], Wei Song[a,b,c,1], Amol C. Shetty[a,b,1], Kelly S. Levano[d], Omar Cáceres[d], Carlos Padilla[d], Víctor Borda[d,e], David Tarazona[d], Omar Trujillo[f], Cesar Sanchez[d], Michael D. Kessler[a,b,c], Marco Galarza[d], Silvia Capristano[d], Harrison Montejo[d], Pedro O. Flores-Villanueva[d], Eduardo Tarazona-Santos[e], Timothy D. O'Connor[a,b,c,2], and Heinner Guio[d,2]

[a]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; [b]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [c]Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [d]Laboratorio de Biotecnología y Biología Molecular, Instituto Nacional de Salud, Lima 11, Perú; [e]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 6627 Belo Horizonte, Brazil; and [f]Centro Nacional de Salud Intercultural, Instituto Nacional de Salud, Lima 11, Perú

Edited by Tony Tosi, Kent State University, Kent, OH, and accepted by Editorial Board Member C. O. Lovejoy June 1, 2018 (received for review November 30, 2017)

Native Americans from the Amazon, Andes, and coastal geographic regions of South America have a rich cultural heritage but are genetically understudied, therefore leading to gaps in our knowledge of their genomic architecture and demographic history. In this study, we sequence 150 genomes to high coverage combined with an additional 130 genotype array samples from Native American and mestizo populations in Peru. The majority of our samples possess greater than 90% Native American ancestry, which makes this the most extensive Native American sequencing project to date. Demographic modeling reveals that the peopling of Peru began ∼12,000 y ago, consistent with the hypothesis of the rapid peopling of the Americas and Peruvian archeological data. We find that the Native American populations possess distinct ancestral divisions, whereas the mestizo groups were admixtures of multiple Native American communities that occurred before and during the Inca Empire and Spanish rule. In addition, the mestizo communities also show Spanish introgression largely following Peruvian Independence, nearly 300 y after Spain conquered Peru. Further, we estimate migration events between Peruvian populations from all three geographic regions with the majority of between-region migration moving from the high Andes to the low-altitude Amazon and coast. As such, we present a detailed model of the evolutionary dynamics which impacted the genomes of modern-day Peruvians and a Native American ancestry dataset that will serve as a beneficial resource to addressing the underrepresentation of Native American ancestry in sequencing studies.

Native American demography | identity by descent | population history | gene flow | fine-scale structure

Native American ancestry is underrepresented by recent whole-genome studies (1–4). As a result, there are numerous questions that remain in both Native American genetic architecture of disease and their early history. One such question pertains to the peopling of the Americas, which began when the Native American ancestral population (5, 6) diverged from East Asians ∼23,000 y ago (ya) (7–9). Following the separation from East Asian populations, Native Americans entered a period of isolation, potentially in Beringia (8, 10), for as long as 10,000 y (3, 7, 11). Thereafter, they migrated into the New World, likely through a coastal route (8, 12, 13), and quickly populated both North and South America. This rapid peopling of the Americas is supported by Monte Verde, one of the oldest archeological sites in the Americas at ∼14,000 ya, being in southern Chile (14) and the divergence between Central American and South American populations being ∼12,000–13,000 ya (3, 7). Therefore, it took Native Americans only 1,000–2,000 y after the isolation period to populate the majority of the Americas. While it is widely supported that this was a rapid process, questions still remain regarding the peopling of South America (11).

Peruvian populations have a rich cultural heritage that derives from thousands of years of New World prehistory (15). The Amazon, Andes, and coast populations in South America likely descend from one major population movement from Central America, ∼12,000–15,000 ya (3, 7, 14, 16, 17). However, the route Native American ancestors took once entering South America is unknown. Genetic differentiation between east and west in South American populations indicates that the Andes Mountains were crucial to this process (18, 19). Furthermore, by combining archaeological, anthropological, and genetic data, Rothhammer and Dillehay (17) hypothesize three major migration routes which involve a trifurcation beginning shortly after Panama and separating into the Andes, Amazon, and coastal regions. However, as suggested by Skoglund and Reich (11), it is also possible that the initial split occurred around both sides of

Significance

Through the Peruvian Genome Project we generate and analyze the genomes of 280 individuals where the majority have >90% Native American ancestry and explore questions at the interface of evolutionary genetics, history, anthropology, and medicine. This is the most extensive sampling of high-coverage Native American and mestizo whole genomes to date. We estimate an initial peopling of Peru was rapid and began by 12,000 y ago. In addition, the mestizo populations exhibit admixture between Native American groups prior to their Spanish admixture and was likely influenced by the Inca Empire and Spanish conquest. Our results address important Native American population history questions and establish a dataset beneficial to address the underrepresentation of Native American ancestry in sequencing studies.

Author contributions: O.C., C.P., T.D.O., and H.G. designed research; D.N.H., W.S., A.C.S., M.D.K., and T.D.O. performed research; D.N.H., W.S., A.C.S., V.B., M.D.K., and E.T.-S. analyzed data; K.S.L., O.C., and C.P. generated genotype data; V.B. and E.T.-S. performed sample quality control; K.S.L., O.C., C.P., D.T., O.T., C.S., M.G., S.C., H.M., P.O.F.-V., and H.G. collected samples; and D.N.H., W.S., A.C.S., K.S.L., O.C., V.B., D.T., O.T., C.S., M.D.K., M.G., S.C., H.M., P.O.F.-V., E.T.-S., T.D.O., and H.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. T.T. is a guest editor invited by the Editorial Board.

[1]D.N.H., W.S., and A.C.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: timothydoconnor@gmail.com or heinnerguio@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720798115/-/DCSupplemental.

Published online June 26, 2018.

the Andes, with the Andes likely then being colonized from the coastal side.

Following the peopling of the Americas, there is a complex pattern of gene flow among Native American populations in both North and South America, even between different language groups and across great geographic distances (19–23). In South America, the Andes again appear to have a dominant role in shaping the migration patterns between the three regions. There is some evidence of migration between the Andes and Amazon (23); however, there is greater migration within the Andes than within the Amazon (21, 23, 24). There is also evidence of migration between the coast, Andes, and Amazon geographic regions (25). Beginning in the 1930s, populations moved from rural areas into cosmopolitan cities (26), where it appears that individuals in the cosmopolitan cities have higher rates of gene flow both between Native American and Old World sources than between Native American groups (19, 22). However, there is a need to further refine these migration histories and estimate the timing for these population-defining events. Since Peru contains all three regions within its borders and its population has predominantly Native American ancestry (25, 27), we are able to construct models to address these questions using genomic information.

Recent migration was also likely affected by major pre-Columbian civilizations, including the Inca Empire, the latest in a long line of historical groups (28). The populations of this region underwent drastic demographic changes that derive from experiences such as forced migration (28–32) and major population size reduction due to the Spanish conquest, during the 16th century, which introduced mass pandemics to the region (3, 4, 33–36). These forced population movements introduced by the Inca Empire (28, 31, 37) and the Spanish colonial rule (31, 32) likely created cosmopolitan communities with individuals from different Native American ancestries. Therefore, the Inca Empire and Spanish conquest likely had a profound impact on gene flow patterns in Peruvians, as previously hypothesized (25). Furthermore, due to the Spanish conquest, admixture occurred between Native Americans and individuals with European and African ancestry, which resulted in modern mestizo (i.e., predominantly admixed ancestry) populations in addition to Native American ones (25). These events, along with the original peopling of the region, created a dynamic pattern of evolutionary history that can now be investigated by genomics (1–4).

To reconstruct the genetic history of Peruvian populations and address the peopling process of the three geographic regions and their migration dynamics at different time points, we analyze high-coverage whole-genome sequence data (n = 150) and genotype array data (n = 130) to create a geographically diverse dataset as part of the Peruvian Genome Project (Fig. 1 and *SI Appendix*, Table S1). Using this data, we show that the Peruvian area was originally peopled ~11,684–12,915 ya and that modern mestizo populations originated from admixture of multiple Native American sources before their African and European admixture. We demonstrate that the majority of migration between the geographic regions is from the high-altitude Andes to the low-altitude regions in Peru, which suggests high-altitude adaption or Andean empire dynamics were crucial in influencing migration dynamics in the region. These findings establish detailed models of mestizo and Native American evolution in the Peruvian region and provide insights into the sociopolitical impacts on genetic variation in the populations of Peru. By providing the largest high-coverage genomic dataset of Native American haplotypes to date, this work lays the foundation for understanding the evolutionary history of the Peruvian region and

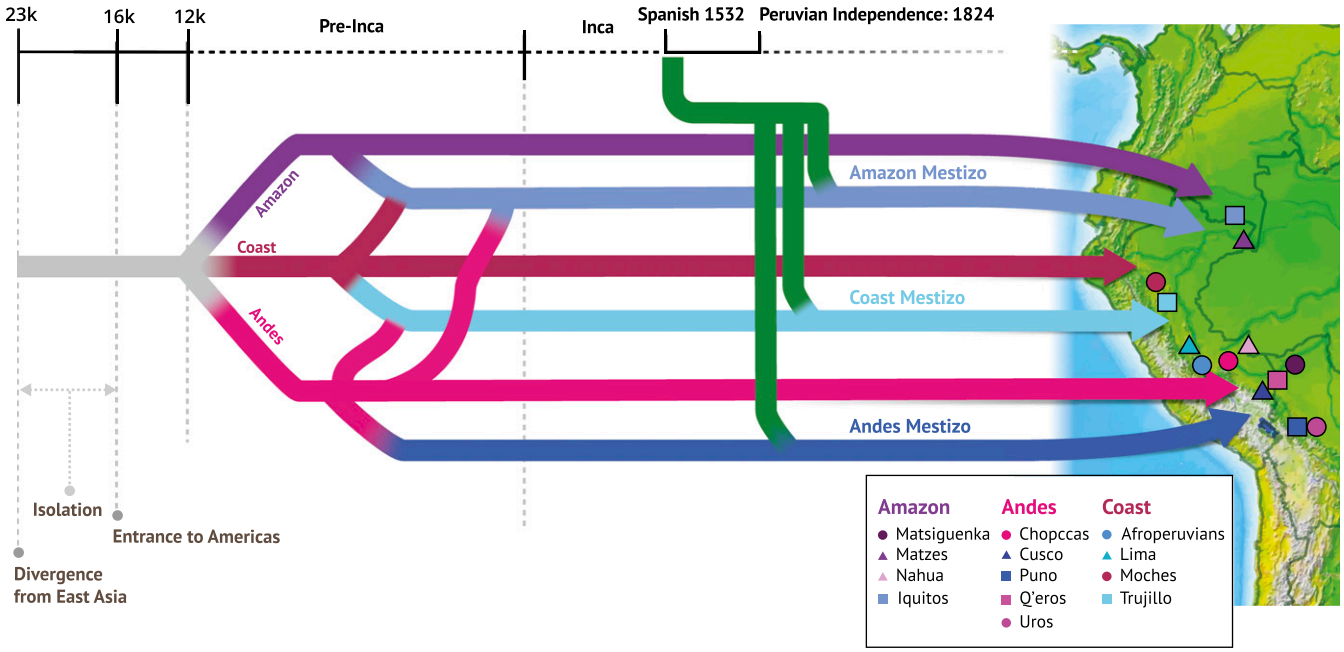


Fig. 1. Population history of Peru. No later than 23,000 ya the Native American ancestral population diverged from East Asia and entered a period of isolation (7). Then, ~16,000 ya the ancestral population began to populate the Americas (7). We inferred that the studied populations of the three geographic regions in Peru (Amazon, Andes, and coast) diverged from each other ~12,000 ya. Before the Inca Empire, the Native American populations admixed with each other to form separate populations that would go on to form the mestizo populations. Migration did occur between all regions; however, we only represent the predominant asymmetrical patterns that shaped modern-day populations (Andes to other regions and coast to Amazon) and represent continuous migration starting during the pre-Inca timeframe and ending at Peruvian independence. The Spanish conquerors arrived in Peru in 1532; however, the majority of Spanish admixture did not occur until after Peruvian independence in 1824. At this time, the Spanish admixture was only with the prior admixed populations of different Native American ancestries to form the modern mestizo populations, while the Native American populations remained essentially isolated. The map at the right of the figure gives the sampling location of these populations and how they correspond to the three major areas indicated in the timeline. Purple shades represent Native American identifying populations and blue shades represent mestizo groups. The shapes on the map have no meaning.

the genomic medical needs for Native American ancestry populations worldwide (4, 38, 39).

Results

Ancestry of Peruvian Populations. We studied 13 Peruvian populations that consist of either self-identifying Native American or mestizo individuals from the Amazon, Andes, and coast (Fig. 1 and *SI Appendix*, Table S1). To perform a broad characterization of our samples' ancestry, we combined them with the Human Genome Diversity Panel genotyped on the Human Origins Array and the 1000 Genomes Project Phase 3 to include other sources of Native American and global genetic variation (2, 40) (*SI Appendix*, Table S2). We find seven ADMIXTURE clusters, including three Old World continental sources and four Native American groups, which represent Amazonian, Andean, Central American, and coastal ancestries (Fig. 2*A* and *SI Appendix*, Figs. S1 and S2). Our samples are enriched for Native American ancestry, as identified by a prior study of Peruvian populations (25), because all sequenced individuals have Native American mitochondrial haplotypes (*SI Appendix*, Table S3) and 103 (59 sequenced) of our samples have ≥99% Native American ancestry estimates (Fig. 2*A*). The high frequency of Native American mitochondrial haplotypes suggests that European males were the primary source of European admixture with Native Americans, as previously found (23, 24, 41, 42). The only Peruvian populations that have a proportion of the Central American component are in the Amazon (Fig. 2*A*). This is supported by Homburger et al. (4), who also found Central American admixture in other Amazonian populations and could represent ancient shared ancestry or a recent migration between Central America and the Amazon.

The majority of Old World ancestry is European and is mostly seen in mestizo populations, which is consistent with prior studied Peruvian groups (25). We find little African ancestry in these populations, except in Afroperuvians and 10 individuals from Trujillo (Fig. 2*A*). To investigate the Old World sources of European and African ancestry in Peru, we performed ancestry-specific principal component analysis (ASPCA), which performs a PCA within distinct continental ancestral origins (43, 44) (*SI Appendix*, Table S2). As expected for South America based on our understanding of colonial history (4), the European and African ancestry component of Peruvian genomes predominantly comes from Spanish and West African populations respectively (*SI Appendix*, Fig. S3).

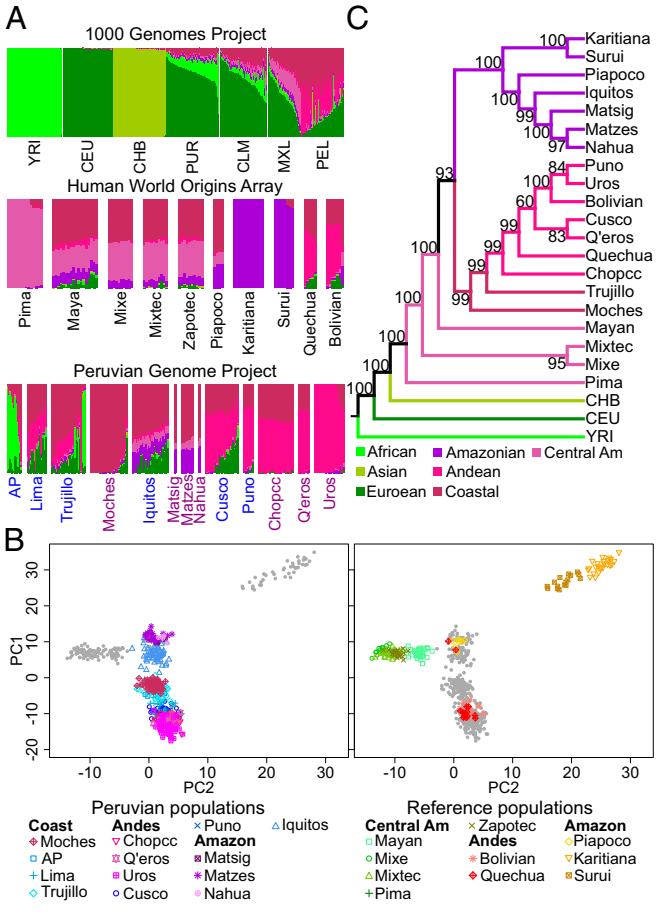


Fig. 2. Peruvian population structure. (*A*) Admixture analysis (K = 7) of 1000 Genomes populations, Native American populations from the Human Genome Diversity Panel (HGDP), and all populations from the Peruvian Genome Project. The Peruvian Genome Project population labels are colored based on the self-identifying label of Native American (Purple) or mestizo (blue). The legend depicts our interpretation of the ancestry represented by each cluster. (*B*) ASPCA of combined Peruvian Genome Project samples with the HGDP. Samples are filtered by their Native American ancestral proportions: ≥50%. Each point represents one haplotype. (*C*) A tree, computed by TreeMix (47), using 1000 Genomes, HGDP, and Peruvian Genome Project samples is shown. Values represent the percent of bootstraps ($n = 500$) in which each node was formed. AP, Chopcc, Matsig, and Central Am are short for Afroperuvians, Chopccas, Matsiguenka, and Central America, respectively. The legend below the tree corresponds to the ancestry represented by each branch of the tree and our interpretation of each ancestral component in the ADMIXTURE analysis.

Biogeography in Peru. To investigate the fine-scale population structure within Peru based on the four Native American ancestry components (Fig. 2*A*) and independent of European and African admixture, we use ASPCA to analyze only the Native American ancestral regions of each individual's genome (*SI Appendix*, Table S2). The Native American ASPCA recapitulates the corresponding geographic locations of samples within Peru (*SI Appendix*, Fig. S4), as well as the locations of samples within Central and South America (Fig. 2*B*), which is consistent with previous studies (4, 42). Therefore, there is strong biogeographic signal within the genetic variation of Native American populations as it is possible to determine the geographic region of these populations based on their genetics alone (45, 46). Further, the mestizo populations from the Amazon, Andes, and coast cluster closest to the corresponding Native American populations from their same geographic region, which resembles the same biogeography identified in the Native Americans (Fig. 2*B* and *SI Appendix*, Fig. S4). Importantly, these ASPCAs do not cluster whole-genome sequencing (WGS) and array samples based on technology (Fig. 2*B* and *SI Appendix*, Figs. S3 and S4), which means we do not expect that any of our results are biased based on differences between WGS and array technologies.

Founding of Peru. There is a need for better characterization of the peopling of the Peruvian region (11). As such, we perform a tree-based similarity analysis (47) of individuals showing ≥99% Native American ancestry as determined by ADMIXTURE analysis, which reveals clusters within South America that are consistent with our ASPCA and ADMIXTURE results (Fig. 2). We find that the Amazonian cluster is sister to the coastal and Andean populations (Fig. 2*C*). This suggests an initial peopling of South America by a split migration around both sides of the Andes Mountains (Occidental and Oriental Cordillera), as previously suggested by others (18, 48). This was likely followed by subsequent splits within the coastal and Andean lineage. Interestingly, the Moches and Trujillo do not share a common ancestor independent from the Andean populations (Fig. 2*C*). This is surprising due to the geographic proximity of the Moches and Trujillo (sample sites differ by only 1.2 km) (*SI Appendix*, Table S1). However, the Moches best represent an unadmixed coastal Native American ancestry, whereas the Trujillo have

substantial genetic contributions from both the Andean and coastal Native American populations, (Fig. 2*A*), which likely explains why the two coastal groups do not form a monophyletic clade that is separate from the Andes.

To further compare the evolutionary dynamics of Native American populations from these three regions, we construct pairwise population demographic models, which include effective population sizes and population divergence times but no population growth or migration, using only WGS data (49) (*SI Appendix*, Figs. S5 and S6). We estimate divergence time between regions to be ∼11,684–12,915 ya without gene flow, although we demonstrate through simulation that modeling without both gene flow and population size changes still accurately determines divergence times while slightly overestimating effective population sizes (*SI Appendix*, Fig. S5). The inferred divergence time between the two Andean populations is 10,453 ya and is statistically distinct from the other population split times (*SI Appendix*, Fig. S6). All between-region divergence times are not statistically significantly different from each other, when the same Andes population is used in the model (*SI Appendix*, Fig. S6). Therefore, we cannot further support the model presented by the tree analysis; however, we can state that the peopling process of the three regions began by ∼12,000 ya.

Native American Admixture in Mestizo Populations. Following the peopling of the three regions, the Native American populations likely remained relatively isolated as the Native American groups tightly cluster within the Peruvian population only Native American ASPCA and the figure is consistent with geographic location. In contrast, the mestizo populations are found intermediate between the Native American groups, which is indicative of admixture from multiple Native American populations (*SI Appendix*, Fig. S4).

To test this more formally, we performed a haplotype-based supervised admixture analysis to further describe this fine-scale structure (50, 51) (*SI Appendix*, Fig. S7 and Table S4). Results from GLOBETROTTER (50) illustrate not only the admixture events between New and Old World sources but also the admixture amid multiple Native American populations (Fig. 3*A*). Each mestizo group exhibits contributions of Native American ancestry from populations in different geographic regions. For example, the Cusco population has greater than 50% Andean ancestry represented by the Chopccas, Q'eros, and Uros sources along with smaller proportions (<10%) of coastal ancestry represented by Moches and Amazonian ancestry represented by Nahua, Matsiguenka, and Matzes. Further, the genomic contributions from different Native American sources are correlated to the geographic regions of mestizo genomes. For example, genomic contributions from high-altitude populations (Andes) increase gradually from Afroperuvians, Trujillo, Lima, and Iquitos in the lowland regions to Cusco and Puno in the Andes. Similarly, the coastal contribution from Moches decreases from Trujillo to Iquitos, Cusco, and Puno, which are far from the coast. Iquitos contains the largest Amazon contribution, from both non-Peruvian and Peruvian Amazon sources with Amazonian contribution lower in coastal and Andean mestizo groups. These results support our hypothesis that mestizo groups have admixture between multiple Peruvian Native American sources. However, this raises an additional question as to when the admixture between different Native American populations occurred in Peruvian history.

To estimate the time of admixture between different Native American populations, we evaluate the genetic distance between each pair of Peruvian individuals using pairwise identity-by-descent (IBD) analysis (52–55). Since generation time for the most recent common ancestor between two individuals has an inverse relationship to the length of IBD segments shared between their genomes (52, 54), we use the shared IBD segments to infer the individual relatedness at different time periods in

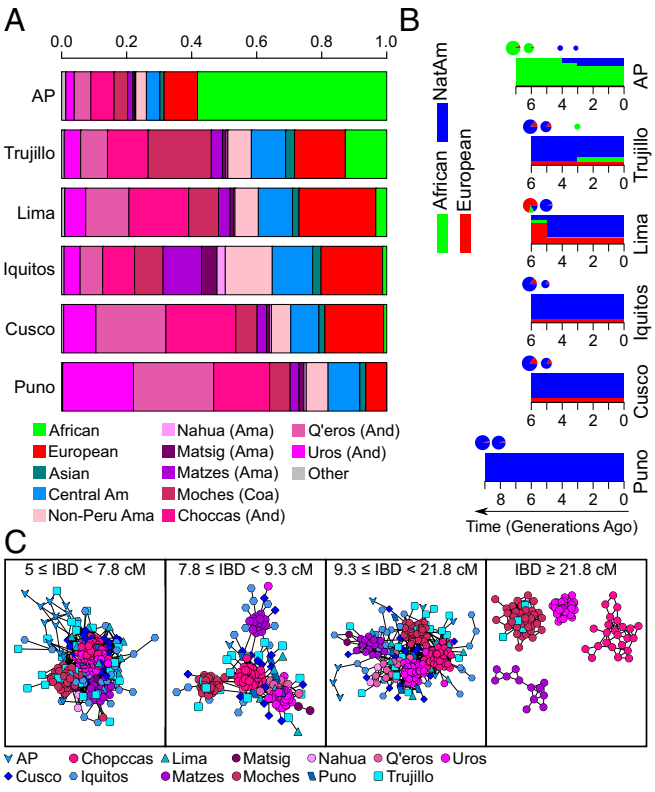


Fig. 3. Admixture among Peruvian populations. (*A*) Colors represent contributions from donor populations into the genomes of Peruvian mestizo groups, as estimated by CHROMOPAINTER and GLOBETROTTER. The label within parentheses for each Peruvian Native American source population corresponds to their geographic region where Ama, And, and Coa represent Amazon, Andes, and coast, respectively. (*B*) Admixture time and proportion for the best fit three-way ancestry (AP, Trujillo and Lima) and two-way ancestry (Iquitos, Cusco, and Puno) TRACT models [European, African, and Native American (NatAm) ancestries] for six mestizo populations. (*C*) Network of individuals from Peruvian Native American and mestizo groups according to their shared IBD length. Each node is an individual and the length of an edge equals to (1/total shared IBD). IBD segments with different lengths are summed according to different thresholds representing different times in the past (52), with 7.8 cM, 9.3 cM, and 21.8 cM roughly representing the start of the Inca Empire, the Spanish conquest and occupation, and Peruvian independence. IBD networks are generated by Cytoscape (98) and only the major clusters in the network are shown for different cutoffs of segment length. AP, Central Am, and Matsig are short for Afroperuvians, Central American, and Matsiguenka, respectively. The header of each IBD network specifies the length of IBD segments used in each network.

Peruvian history (52). We focus our analysis on the transition between pre-Inca civilizations, the Inca Empire, and Spanish colonial rule based on enrichment of certain lengths of IBD segments, and further combine with their common ancestries across the genome from these time periods (Fig. 3*C* and *SI Appendix*, Fig. S8). Before the Inca Empire (IBD segments of 5–7.8 cM, about AD 1116–1438), all Andes populations cluster together (Fig. 3*C* and *SI Appendix*, Table S5). During the Inca Empire (IBD segments of 7.8–9.3 cM, about AD 1438–1532), we identify clear separations between four Native American populations. The Chopccas are at the center of this pattern and maintain connections directly or through mestizo groups to all other Peruvian populations (Fig. 3*C*).

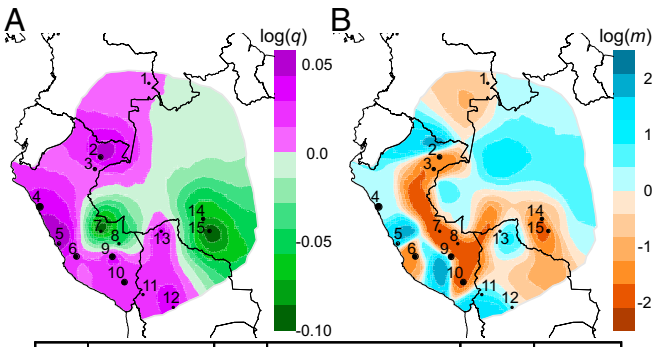
During the time period associated with the Viceroyalty of Peru (IBD segments of 9.3–21.8 cM, about AD 1532–1810), the Native American populations are still tightly connected, but the Chopccas are no longer the major intermediate connector. This

is consistent with the historical observation that the location of cultural dominance shifted away from the Andes to the coast and major cities (31). In addition, mestizo individuals share IBD segments with multiple Native American populations. Therefore, these patterns during and after the time periods of the Inca Empire are consistent with our GLOBETROTTER (50) results (Fig. 3*C* and *SI Appendix*, Fig. S7) and provide further support that mestizo populations share ancestry from multiple Native American groups. After Peruvian independence (IBD ≥ 21.8 cM, about AD 1810 AD to the present), there were fewer mestizo individuals sharing IBD segments with Native American groups (Fig. 3*C*), therefore, indicating that Native American and mestizo populations became isolated during this time. These IBD results show that the admixture between Native American populations in mestizo groups began before the arrival of the Spanish in AD 1532.

Timing European Admixture in Mestizo Populations. In light of these IBD results, we hypothesize that individuals who migrated as a result of dynamics within the Inca Empire and Spanish colonial rule were more likely to later admix with Spanish-ancestry individuals. To further investigate this hypothesis we use TRACTS (56) to estimate the major admixture events between European, African, and Native American ancestries to occur between ∼AD 1836 and 1866 (Fig. 3*B* and *SI Appendix*, Fig. S9). This suggests that the majority of admixture between the Spanish and Native Americans did not occur until ∼300 y after Spain conquered Peru, which is consistent with what others have found for South America (4) and may correspond to the sociopolitical shifts of the Peruvian war of independence that occurred between AD 1810 and 1824. Taken together with our IBD results, the admixture between Native American groups likely occurred before Peruvian Independence, and that later admixture between Europeans and these same admixed Native American populations led to the modern mestizo groups.

We further test these observations using the ancestry-specific IBD (ASIBD) method, which we calculate for all Peruvian samples by intersecting traditional IBD calls with local ancestry inferences (*SI Appendix*, Fig. S10). To remove the influences from recent shared European and African ancestries, we focus exclusively on Native American genomic components. The ASIBD analysis reveals that mestizo groups are more likely to share IBD segments with multiple Native American populations from different geographic regions across all IBD segment lengths we tested ($P < 0.001$; *SI Appendix*, Table S6). We also observe a more recent gene flow into Peruvian populations from Central America, as these two groups share predominantly large ASIBD segments, which is also consistent with our ADMIXTURE results (Fig. 2*A* and *SI Appendix*, Fig. S10).

Genetic Diversity and Clinical Implications for Native American Ancestry Populations. All four Native American populations (Chopccas, Matzes, Moches, and Uros) have small effective population sizes (*SI Appendix*, Fig. S6) in comparison with other Native American ancestry populations (3) and outbred Old World populations (57, 58). This difference could be due to our ability to more robustly estimate demography using only the noncoding regions of the genome obtained by our WGS approach. However, these estimates are supported by the observation that Native American populations have a larger proportion of their genome in runs of homozygosity, as well as decreased heterozygosity counts, compared with mestizo populations (*SI Appendix*, *Methods* and Figs. S11 and S12). We also estimated diversity across Peru (59) and found that Native American populations are in geographic regions with low estimated effective genetic diversity relative to mestizo groups, which is consistent with prior observations of other urban populations (19). Lima (population size of 9,886,647) and Iquitos (population size of 437,376) (60), two of the largest cities in Peru, have the greatest measures of genetic diversity (Fig. 4*A*). Further, when the Moches



Deme	Population	Deme	Population	Deme	Population
1	Piapoco	6	Chopccas	11,12,13	Bolivian
2	Iquitos	7	Nahua	14	Karitiana
3	Matzes	8	Matsiguenka	15	Surui
4	Moches, Trujillo	9	Cusco, Q'eros, Quechua		
5	Lima	10	Puno, Uros		

Fig. 4. Peruvian demography. (*A*) Effective diversity rates in Peru. Green represents areas of low diversity and purple high diversity. (*B*) Contemporary model of effective migration rates in Peru. Brown represents areas of low migration and blue areas of high migration. The legend details which populations are grouped into the different numbered demes on the map.

are removed from this analysis, the Trujillo join Lima and Iquitos as the most diverse populations in Peru (*SI Appendix*, Fig. S13). This signal is robust to the exclusion of known low-diversity populations (i.e., Surui and Karitiana) (61), to the removal of European admixed individuals, and to single population inclusion/exclusion (*SI Appendix*, Fig. S13).

The low genetic diversity estimates of Native Americans suggest that there may be an enrichment for rare diseases in Native American ancestry communities (62–64) living in small populations, as is already observed for isolated groups of different ancestries worldwide (65). Further, the previously mentioned admixture dynamics in mestizo populations lead us to suspect that they will have greater European-specific clinical variation. Due to the underrepresentation of Native American ancestry in genomic databases, we hypothesize that Native American communities may have an increase in recessive disease alleles that are unobserved in current clinical databases (38, 66). In fact, consistent with this hypothesis, we observe fewer ClinVar (67) variants in our Native American populations than mestizo and Old World populations (*SI Appendix*, *Methods* and Fig. S14). Mestizo populations potentially inherited risk factors from multiple Native American sources, which further represents the importance of our efforts to better characterize native Peruvian genomic diversity and disease susceptibility.

Migration Dynamics in Peru. These low genetic diversity estimates for Native American populations (*SI Appendix*, Figs. S11–S13) and their minimal connectedness in recent IBD networks (Fig. 3*C*) also suggest that Native American populations are isolated and therefore receive minimal external gene flow. This isolation is further supported by our migration pattern estimates, which show low migration for Native American populations (Fig. 4*B*), and agrees with prior observations of low genetic diversity (Fig. 4*A* and *SI Appendix*, Figs. S11 and S12). Further, these results are robust to European admixture and individual populations' exclusions, with two exceptions. First, the Moches exist in an area of high migration, but this changes to an area of low migration when the Trujillo samples are removed from the model (*SI Appendix*, Fig. S15). Therefore, this supports the idea that the Trujillo received more gene flow from other populations and the Moches have not, which is consistent with our conclusion of low migration and minimal external gene flow for Native American populations (Fig. 2 *A* and *C*). Second, the removal of Lima does represent a

noteworthy complication to this pattern, as a slight migration barrier appears between the Andes and the coast (*SI Appendix*, Fig. S15). This suggests that Lima is a crucial intermediate for the migration seen between the coast and the Andes.

As our migration analysis is unable to indicate directionality of migration (59), we developed an alternative approach based on fine-scale ancestry estimates among the mestizo populations to give the ratio of genomic contributions originating from another geographic region into the local region vs. the opposite (Fig. 3*A* and *SI Appendix*, Table S7). We find the indication of asymmetrical migration from the Andes to the coast where the proportion of Andes ancestry which exists in coastal mestizo individuals is 4.6 times more than the coastal ancestry that exists in Andes mestizo individuals ($P = 0.038$). Other migrations are from the Andes to the Amazon (ratio = 6.3, $P = 0.001$), and from the coast to the Amazon (ratio = 2.2, $P = 0.002$). These results suggest that the major direction of migration, at least in terms of mestizo individuals, is in descent of the Andes Mountains toward the cities in the lower-altitude Amazon and coast.

Discussion

The Peruvian Genome Project presents the largest reference panel of high-quality Native American WGS data to date and can serve as a resource for genetic and genomic analysis of populations worldwide with Native American ancestry, such as Latino groups. As a result, we are able to model aspects of South American demographic history, especially in Peru. We address the founding of the Amazon, Andes, and coast regions in Peru and the migration dynamics of populations between and within the geographic regions. Through tree analysis we suggest the initial divergence was a split around the Andes followed by the Andes and coast divergence, which supports the question stated by Skoglund and Reich (11) and does not support the trifurcation hypothesis of Rothhammer and Dillehay (17). However, our divergence time estimates between all geographic regions are not statistically different (*SI Appendix*, Fig. S6) and so we cannot confirm the peopling model suggested through tree analysis (Fig. 2*C*). Instead, the divergence times are consistent with Rothhammer and Dillehay (17), which indicates a trifurcation at the initial founding of South America. Therefore, we propose this tree model as support for one potential scenario of the peopling process; however, we cannot reject the trifurcation hypothesis. We can state that the average divergence time between the regions is ∼12,000 ya, which indicates the peopling process began by this time.

In addition, this implies that the three regions were peopled rapidly, as suggested for other Native American populations (3, 68). We find that the divergence between these populations is similar to prior estimates of divergence times between Central American and South American populations (12,219 ya), Central American and Caribbean populations (11,727 ya) (3), and North and South American populations (∼13,000 ya) (7). Furthermore, this time range also corresponds to the oldest archeological sites in Peru as Guitarrero Cave in the Andes, Amatope on the coast, and Monte Alegre in the Brazilian Amazon are ∼11,000–12,000 y old (15). These estimates for the peopling process predate the appearance of agriculture in the Amazon by roughly 6,000 y and in the coast by roughly 4,000 y (15), whereas the divergence between the Chopccas and Uros predates the appearance of agriculture in the Andes by ∼400 y (15). Therefore, it is likely that agriculture was developed after each major geographic region of Peru was populated for multiple generations. However, with these results, we cannot conclude whether agriculture was independently developed in each geographic region or spread from one region to another. In combination with archeological and other population genetics studies, we are confident in the overall time frame suggested by our demographic model; however, we need denser sampling of these populations and methods that are sensitive enough to distinguish small differences in divergence times to further refine this model.

Following the peopling of Peru, we find a complex history of admixture between Native American populations from multiple geographic regions (Figs. 2*B* and 3 *A* and *C*). This likely began before the Inca Empire due to Native American and mestizo groups sharing IBD segments that correspond to the time before the Inca Empire. However, the Inca Empire likely influenced this pattern due to their policy of forced migrations, known as "mitma" (*mitmay* in Quechua) (28, 31, 37), which moved large numbers of individuals to incorporate them into the Inca Empire. We can clearly see the influence of the Inca through IBD sharing where the center of dominance in Peru is in the Andes during the Inca Empire (Fig. 3*C*). A similar policy of large-scale consolidation of multiple Native American populations was continued during Spanish rule through their program of *reducciones*, or reductions (31, 32), which is consistent with the hypothesis that the Inca and Spanish had a profound impact on Peruvian demography (25). The result of these movements of people created early New World cosmopolitan communities with genetic diversity from the Andes, Amazon, and coast regions as is evidenced by mestizo populations' ancestry proportions (Fig. 3*A*). Following Peruvian independence, these cosmopolitan populations were those same ones that predominantly admixed with the Spanish (Fig. 3*B*). Therefore, this supports our model that the Inca Empire and Spanish colonial rule created these diverse populations as a result of admixture between multiple Native American ancestries, which would then go on to become the modern mestizo populations by admixing with the Spanish after Peruvian independence. Further, it is interesting that this admixture began before the urbanization of Peru (26) because others suspected the urbanization process would greatly impact the ancestry patterns in these urban centers (25). It is possible that urbanization also impacted the genetic ancestry of the mestizo populations in this study; however, our IBD analysis did not have sufficient sampling to infer genetic relatedness dynamics in the last two generations (*SI Appendix*, Fig. S8).

These IBD results represent recent migration patterns within Peru centered on the Inca Empire and Spanish conquerors. However, we also present an overall topography of migration in Peru that is representative of all time ranges since the initial peopling of the region. This migration topography shows that there is a corridor of high migration connecting the city centers in the Amazon, Andes, and coast (Fig. 4*B*). We also note that there is asymmetrical migration between these regions with the majority of migration being in descent of the Andes toward cities in the Amazon and coast. Therefore, we hypothesize that this migration pattern may be due to selection pressures for alleles that assist in high-altitude adaptation creating a disadvantage to new migrants (69–71). However, this also aligns with Spanish efforts to assimilate Native American populations (31, 32) and the complex history of Native American admixture in mestizo populations (Fig. 3 *A* and *C*), which could be a result of Inca Empire policies (28, 31, 37). Therefore, future research is required to determine if this migration pattern began during Native American empires or during Spanish rule compared with being constant since the initial peopling of the region and explore the possibility that this pattern is a result of a combination between the two forces.

The Native American populations seem to be relatively isolated in Peru, especially following Peruvian independence (Figs. 2*B*, 3*C*, and 4*B*), which is consistent with prior analyses of Native American populations in South America (9, 18). However, the Peruvian mestizo populations exhibit strong signatures of migration, which disagrees with notions of isolation and is partially consistent with the observation of increased levels of gene flow in urban Brazilian cities (19). Fuselli et al. (19) attribute this migration to the influence of the European conquest, whereas we note that a combination of the Inca Empire and dynamics before

the Inca also impacted these migration dynamics in addition to the European conquest (Fig. 3*C*). Furthermore, some analyses indicate that not all gene flow was within one single geographic region but rather between the Andes and Amazon, or all three geographic regions (23, 25). We demonstrate that, in fact, migration occurred between all three geographic regions, though in asymmetrical ways from the Andes to other parts of Peru as well as from the coast to the Amazon (*SI Appendix*, Table S7). Therefore, we can confirm prior analyses which indicate a greater degree of isolation within Native American populations, while also providing additional support to the hypothesis of gene flow between populations, even across different geographic regions.

These presented evolutionary dynamics of Peruvian people provide insights into the genetic public health of Native American and mestizo populations. Due to the overall low genetic diversity of both Native American and mestizo populations (*SI Appendix*, Figs. S11 and S12), it is likely that there is an enrichment of rare genetic diseases unique to these groups. While this is a pattern observed worldwide and has been extensively documented in isolated Old World populations (65), the diseases that may increase in frequency as the product of these phenomena of isolation are likely to be population-specific and should be systematically documented and addressed by the Peruvian health system based on other Latin American experiences (72) and can add to the growing number of studies finding new insights into biology through analysis of isolated populations. Further, the finding that few ClinVar (67) variants are present in Native American ancestry populations also stresses that these populations are underrepresented in genome studies and demands more sequencing efforts. Additional studies will lead to an increase in our understanding of Native American genetic variation and the sequences we present will serve as a reference for future studies. However, we also find strong population structure between the different regions in Peru (Fig. 2 *B* and *C*), and as a result of the rapid peopling of the New World we hypothesize larger differences in the distribution of pathogenic variants from one group of Native American ancestry to another (3). This means that future studies must sequence from different regions of the Americas as the Native American ancestry populations from other North and South American countries are expected to be less closely related to these samples. The Peruvian samples we present here likely capture most of the Native American common genetic variation (2); however, large sample sizes in other geographic regions are required to discover rare genetic variation for any given region, which is crucial for understanding genetic causes of diseases, both complex and Mendelian (73, 74).

In conclusion, we present here the largest collection of high-coverage Native American sequences to date, which makes important strides in addressing the genetic underrepresentation of these populations. In addition, these data help address important questions in Native American evolutionary history as we find that populations from the Andes, Amazon, and coast diverged rapidly, ~12,000 ya. Following the initial peopling of each region, we demonstrate that the majority of the migration in Peru is in descent of the Andes toward the Amazon and coast. As part of this migration dynamic, we demonstrate that mestizo populations have a complex history of Native American ancestry which occurred before their admixture with European ancestry populations, following Peruvian independence. The understanding of this complex evolutionary history, in addition to the low genetic diversity of these populations, is crucial for bringing the era of personalized medicine to Native American ancestry populations.

Methods

Sample Collection. The protocol for this study was approved by The Research and Ethical Committee (OI-003-11 and OI-087-13) of the Instituto Nacional de Salud del Perú. The participants in this study were selected to represent different Peruvian Native American and Mestizo populations. We obtained informed consent first from the Native American or Mestizo community and then from each study participant. Native American population cohort participants were recruited from the Matzes, Uros, Afroperuvians, Chopccas, Moches, Q'eros, Nahuas, and Matsiguenka populations. We applied three criteria to optimize individuals to best represent the Native American populations: (*i*) the place of birth of the participant and that of his or her parents and grandparents, (*ii*) their last names (only those corresponding to the region), and (*iii*) age (eldest to mitigate effects of the last generation). Participants of the mestizo population cohorts were recruited from the cities Iquitos, Puno, Cusco, Trujillo, and Lima and were randomly selected. The Afroperuvians were sampled as a Native American population; however, for all analyses we treated them as a mestizo group due to their expected admixture between multiple ancestries.

Genomic Data Preparation. One hundred fifty Native American and mestizo Peruvian individuals were sequenced to an average of 35× coverage on the Illumina HiSeq X 10 platform by New York Genome Center (NYGC). Following sequencing, the raw reads were aligned to hg19 with bwa mem (75), duplicates marked with Picard MarkDuplicates (76), and variants called using GATK in each individual's genome by NYGC (77). We calculated the coverage of chr1–22 for each Peruvian WGS sample with the SAMtools version 1.7 depth command (78, 79), restricting to the specified coordinates for chr1–22 in the hg19 fasta reference file (2), and calculated the average number of times each base on chr1–22 was represented in each sample. We then used GATK UnifiedGenotyper (80–82) to jointly identify the set of all biallelic single nucleotide variants (SNVs) that were independently discovered in each individual's nuclear and mitochondrial genome. All SNVs flagged as LowQual or a quality score <20 were removed (83, 84). We used the webserver of Haplogrep (85–87) to determine the mitochondrial haplogroup of all 150 Peruvian WGS samples based on the unified mitochondrial variant database. KING was used to identify related individuals in our WGS data and we removed the smallest set of individuals to create a final WGS dataset with no pairs having a kinship coefficient ≥0.044, which pruned first to third degree relations (88).

An additional 130 Native American and mestizo Peruvian individuals were genotyped on a 2.5M Illumina chip. To create a combined Peruvian WGS and array dataset with 280 Native American and mestizo Peruvians, the intersect of the two datasets was filtered to remove all A-T and G-C sites to prevent analyzing two different DNA strands in the array vs. WGS data. We also removed any triallelic sites created by combining the two datasets that were not due to a strand flip in either dataset. The array-WGS dataset was filtered to remove sites with ≥10% missingness or ≤2 minor allele count (83, 84). We then ran KING (88), with default settings, to determine additional relatedness in the merged dataset and to determine the optimal number of individuals with no pair of ≥0.044 kinship coefficient, which consisted of 227 individuals. In the case of pairs of related WGS and array samples, the WGS sample was included instead of the array sample. In addition, the WGS individuals were selected from the duplicate samples (i.e., both array and WGS data for the same sample). A final combined dataset of 4,694 samples was created, which contained 227 unrelated samples from Peru, 2,504 samples from 1000 Genomes Project (2), and 1,963 Human Genome Diversity Panel (HGDP) samples genotyped on the Human Origin array (40) by taking the intersection of all three datasets. There were 183,579 markers remaining after removing all A-T, G-C, and triallelic sites from the combined dataset (83, 84). We performed standard ADMIXTURE (89, 90), phasing, and local ancestry analyses as described in *SI Appendix, Methods*.

ASPCA. We used the approach developed by Browning et al. (43) to perform the ASPCA. PCAdmix (91) was used to prune the markers according to their allele frequency and linkage disequilibrium with default thresholds and ASPCA was then performed based on the remaining 100,202 markers. To reduce ASPCA's error from samples with small portions of their genome in a single ancestry class, we only included samples with ≥30% European ancestry, ≥10% African ancestry, and ≥50% Native American ancestry for the European, African, and Native American ASPCA, respectively.

CHROMOPAINTER and GLOBETROTTER. CHROMOPAINTER (51) explores ancestry using SNV data of haplotypes sampled from multiple populations and reconstructs recipient individual's genome as serials of genetic fragments from potential donor individuals. The donor group contained 47 populations, from five ancestries: (*i*) European, (*ii*) African, (*iii*) Native American (NatAm3), (*iv*) Asian, and (*v*) Oceanic, as listed in *SI Appendix*, Table S2. Each selected donor population had more than two samples to avoid spurious estimation of ancestry (except for Peruvian Matsiguenka and Nahua populations because these populations only have two samples and represent variation we are keenly interested in). All six Peruvian mestizo populations are treated as recipient populations. Due to the computational complexity,

we estimated the parameters "recombination scaling constant" and "per site mutation rate" using five randomly selected chromosomes (1, 2, 6, 10, 16) with 10 iterations of expectation-maximization algorithm, as suggested by Montinaro et al. (92). The two estimated parameters were 225.08 and 0.00073, respectively. We "painted" both the donor and recipient individuals' genomes using the combination of fragments from all donor chromosomes. The companion program GLOBETROTTER (50) was then used to estimate the ancestral contributions from different donor populations into the recipient populations, with DNA information on multiple sampled groups (as summarized by CHROMOPAINTER). In this way, we could achieve finer description of population structure.

Directional Migration for Mestizo Populations. To infer the direction of migration inside Peru, we defined a parameter, ratio $= A(b)/B(a)$, where A and B are target mestizo populations from two different geographic regions and a and b correspond to the source Native American ancestry from these two regions, based on the fine contributions from different source populations estimated by GLOBETROTTER. The three geographic regions are Amazon (target mestizo: Iquitos; source native: Matsiguenka, Matzes, and Nahua), coast (target mestizo: Trujillo and Lima; source native: Moches), and Andes (target mestizo: Cusco and Puno; source native: Chopccas, Q'eros, and Uros). $A(b)$ means the average proportion of an individual's genome originated from region b which then existed in individuals living in region A [for example, Coast(Andes) means for those Trujillo and Lima mestizo individuals in the coast, what portion of their genomes is from Chopccas, Q'eros, and Uros in the Andes, according to the GLOBETROTTER results]. Similarly, Andes (coast) means for those Cusco and Puno mestizo individuals in Andes, what portion of their genomes is from Moches in the coast. The ratio of Coast (Andes)/Andes(coast) then represents the direction of migration from Andes to the coast if the ratio is >1.0. We first calculated the original ratio for a pair of regions A and B and then mixed the mestizo individuals from these two regions by randomly permuting the regional labels of them and each time a shuffled ratio was calculated by recomputing the population admixture proportion with the new labels. Empirical P values were then calculated based on comparing the original with the 1,000 permutations.

ASIBD. For the phased dataset of Peru with HGDP genotyped on the Human Origins Array, including related individuals, genomic segments that are IBD between individuals were estimated between all pairs of samples on haplotypes with Beagle 4.0, with the following settings: ibd = true, overlap = 101, and ibdtrim = 7 (55). Based on the length of standard IBD segments, an approximation can be used to infer the past generation time (52): E(generation ago) $\simeq 3/2l$, where l is the length of IBD segment in morgans. Our IBD network analysis are based on all IBD segments that are longer than 5 cM to avoid possible background noise.

For all time frames we had sufficient signal to identify patterns. *SI Appendix*, Fig. S8 gives the distribution of the number of segments, with larger segments being fewer but clearer in terms of the signals they represent (i.e., if you have a large segment it is less likely to be a false positive and represent its signal well). For example, during the Inca Empire we use segments of length 7.8–9.3 cM, which constitute 13% of all segments identified greater than 5 cM.

In addition to standard IBD, we also defined the ASIBD using the local ancestry segments information from RFMix. If two haplotypes from different individuals share IBD segments and part of the shared segments come from the same ancestry, then the overlapping part between shared IBD and common ancestral segments is defined as ASIBD. In our IBD analysis, to preserve the diversity of Peruvian samples, we kept 259 individuals from Peru with ≥50% of Native American component in their genomes, including those related ones. Because the mestizo samples had shorter Native American component in their genomes than the Native American samples, the significance of their Native American IBD might be underestimated. To avoid this, we calculated the total lengths of common Native American segments among all pairs of haplotypes (between different individuals) and used the shortest length as a cap threshold for all other pairs, which was about 700.4 cM between the haplotypes from one Afroperuvian sample and one Cusco sample. For each pair of haplotypes, their common Native American segments were randomly picked without replacement until the accumulated length reached this threshold value. These selected Native American segments were then combined with their shared IBD segments to estimate the Native American-specific IBD between the two haplotypes. We could then have equal comparison for all individual pairs with different Native American proportions in their genomes, by using the value of Native American segment length per shared IBD length (NatAm/IBD). However, different from the similar approach using the whole distribution of IBD segments (93), ASIBD combines both tracts from same ancestry and shared IBD from recent common ancestor for each pair of individuals. In this case accurate time estimation is not available yet based on the total length of ASIBD segments, although longer ASIBD segments still represent more recent common ancestors compared with shorter ones.

We statistically evaluated the hypothesis that mestizo individuals were more likely to be admixed from multiple Native American populations. We accomplished this by defining four possible population categories, including three Native American based on their geographic origin: Amazon (represented by Matzes, Matsiguenka, and Nahua), coast (represented by Moches), and Andes (represented by Chopccas, Q'eros, and Uros), along with one mestizo group (represented by Afroperuvians, Iquitos, Cusco, Puno, Lima, and Trujillo). Based on the Native American-specific IBD network, we defined the connection pattern $N_1(i) - M(l) - N_2(k)$ as two distinct individuals, i and k, who come from two different Native American geographic regions, N_1 and N_2, where $N_1 \neq N_2$, and connect to each other through a mestizo individual, l. For example, Coast (i)-mestizo(l)-Andes(k), i, l, k are the sample IDs. Similarly, $N_1(i) - N_2(j) - N_3(k)$ represents the pattern of two distinct individuals, i and k, who come from two different Native American geographic regions, N_1 and N_2, connect to each other through an individual j, who comes from either the same [if $N_1 = N_2 \neq N_3$, e.g., Uros(i)-Chopccas(j)-Moches(k), as both the Uros and Chopccas come from the Andes] or a different Native American label [if $N_1 \neq N_2 \neq N_3$, e.g., Moches(i)-Chopccas(j)-Matzes(k), where here N_1 is the Coast region, N_2 is the Andes region and N_3 is the Amazon region]. The ratio of number of *NMN* connections to number of *NNN* connections represents the frequency of the mestizo individuals being mixed from different Native American groups. To test for significance, we then randomly permuted the labels for all individuals in an IBD network and each time a shuffled ratio was calculated by recomputing the numbers of two patterns with the new labels. Empirical P values are based on 1,000 permutations.

TRACTS. TRACTS (56) was used to estimate the admixture time and proportion in six mestizo populations from Peru admixed by European, African, and Native American ancestries, according to the local ancestry inference by RFmix. Ancestral segments shorter than 11.7 cM were not used for model optimization because their numbers might not be accurately estimated, as suggested by Baharian et al. (52). For each mestizo population, five models were optimized, including two two-ancestry and three three-ancestry models. Then, the best-fit models were chosen according to their Bayesian information criterion (BIC) scores.

For two-ancestry admixture, each model assumes Native American and European components (since the African component is very small in Puno, Iquitos, and Cusco). The first model, p_1p_2, represents the admixture between different ancestries p_1 and p_2 in a single migration event. The second model, $p_1p_2\_p_1x$, represents the second incoming migration event from ancestry p_1 in addition to the initial admixture between p_1 and p_2, where x indicates no migration from the other ancestry p_2. p_1 and p_2 could be either one of Native American and European ancestries.

For three-ancestry admixture models we assumed European, African, and Native American components (Fig. 3*B*). The first model, $p_1p_2x\_xxp_3$, represents the initial admixture event between ancestries p_1 and p_2, then followed by the second incoming migration event from ancestry p_3. The second model, $p_1p_2x\_xxp_3\_xxp_3$, represents a third migration event from ancestry p_3 again into the admixed population, in addition to the first p_1p_2 and the second p_3 admixture events. Similarly, the third model, $p_1p_2x\_xxp_3\_p_1xx$, represents the third migration event from ancestry p_1 again in addition to the first and second admixture events. p_1, p_2 and p_3 could be one of European, African, and Native American ancestries and are different from each other. Again, we used BIC to select the best model, while penalizing for overparametrization.

Tree Analysis. All individuals identified as being ≥99% Native American ancestry, through ADMIXTURE analysis (Fig. 2*A*), and the YRI, CEU, and CHB from the 1000 Genomes Project were extracted from the final combined dataset and applied the same filters as in the kinship analysis. We then constructed a tree using TreeMix v1.12 (47) over 500 bootstraps, with the YRI set as the root of the tree and the cluster set to 1 SNV. We used PHYLIP consense tree v3.68 (94) to calculate the consensus of the 500 bootstraps and then used MEGA v7.0.14 (95) to plot the consensus tree.

δaδi Modeling. We calculated the site frequency spectrum, using δaδi (49), to estimate a simple demographic history of our Native American populations similar to the model implemented by Gravel et al. (3), which models a population split and modern effective population sizes and does not model growth or migration. We removed the influence of European admixture in our model by selecting individuals from the Chopccas, Matzes, Moches, and

Uros populations with ≥99% Native American ancestry as calculated by ADMIXTURE analysis (Fig. 2*A*), so that we only analyzed Native American demography. Using WGS data only, we removed gene regions from the genomes (chr1–22) of these individuals by excluding all positions within ±10,000 bp from all genes in RefSeq downloaded from the UCSC Genome Browser on February 16, 2016 (96). We further removed sites with ≥10% missing genotype values and that lacked a high-quality human ancestor (2) trinucleotide (including the variant and ±1 bp), which resulted in a total of 2,891,734 SNVs. We then jointly called all positions in the autosome, excluding prior identified variant positions, in all individuals used in δaδi analysis with GATK joint genotyper (80–82). The same WGS and gene region filters as before were applied and we then added these discovered high-quality invariant sites to the total variant sites to yield the entire sequenceable size of our dataset to be 1,025,346,588 bp. We then performed all pairwise two-population models among the Chopccas, Matzes, Moches, and Uros calculating the following parameters: (*i*) N_A = ancestral effective population size change, (*ii*, *iii*) N_1, N_2 = effective population size of population 1 and 2, (*iv*) T_A = time of ancestral effective population size change, and (*v*) T_S = time of divergence between population 1 and population 2 (*SI Appendix*, Fig. S6). We down-sampled the Matzes by two haplotypes and the other three populations by four haplotypes due to some samples having missing genotypes. Therefore, by downsampling, we increased the number of segregating sites for δaδi analysis as described by Gutenkunst et al. (49). We corrected the site frequency spectrum for sites that contain multiple mutations with the following files provided in the δaδi installation (49): "tri_freq.dat", "Q.HwangGreen.human.dat", and "fux_table.dat" with the δaδi function: dadi.Spectrum.from_data_dict_corrected (49, 97). We used a grid size of [30, 40, 50] and started the guess of each parameter at [0.05, 1.2, 1.5, 0.01, 0.029] with an upper bound of [10, 500, 500, 1, 1] and a lower bound of [1e-2, 1e-2, 1e-2, 0, 0]. We generated a perturbed set of parameters with the dadi.misc.perturb_params function which included the upper and lower bound of our parameters, the guess of each parameter, and a fold of 1. We used the dadi.Inference.optimize_log to optimize each model with 1 million maximum iterations, the guess of each parameter, upper and lower bound of each parameter, corrected site frequency spectrum, the verbose option set to the number of total parameters, and our specified model function which we generated with the dadi.Numerics.make_extrap_log_func function. Each model was run 1,000 times and the maximum log likelihood was selected to best represent the model.

We calculated 95% CIs by performing 500 bootstraps and removing the upper and lower 2.5% of the inferred values. We first formed 1-MB segments of sequenceable bases across each chromosome (final segment is often truncated as it was not 1 MB long). We then formed each bootstrap by randomly selecting, with replacement, these 1-MB genomic segments to equal the total sequenceable size of our dataset. Then each δaδi model was run on the bootstraps with the same method as the initial inferred value analysis. We used a human generation time of 30 y and a mutation rate of 1.44×10^{-8}, as calculated for Native American ancestry by Gravel et al. (3), to convert estimated values into chronological time and effective population sizes. We evaluated the accuracy of our demographic model through simulations, as described in *SI Appendix*, *Methods*.

EEMS Migration and Diversity Modeling. To create a contemporary model of migration and diversity in Peru, we used estimated effective migration surfaces (EEMS) to model the effective migration and diversity rates (59) of all nonrelated Peruvian Genome Project samples and HGDP samples genotyped on the Human Origins Array (40) from Peru, Bolivia, Colombia, and Brazil with <1% African ancestry. We also ran EEMS on individuals with ≥99% Native American ancestry and removed Lima from the original EEMS analysis to reveal any effects European admixture had on our model. We attempted to run EEMS modeling on all individuals with <99% Native American ancestry; however, the resulting migration topographies in each model did not converge even though the Markov chain Monte Carlo (MCMC) chain did. This is likely due to the filter's leaving a small number of individuals and demes which created too many local maxima within the data so it is not possible to discern the optimal <99% Native American ancestry model. In addition, to further test the robustness of our model we removed the following subsets of populations from the African ancestry filtered dataset: (*i*) Trujillo, (*ii*) Moches, (*iii*) Puno, (*iv*) Uros, (*v*) Cusco, (*vi*) Quechua, (*vii*) Cusco and Quechua, (*viii*) Q'eros, (*ix*) Karitiana, (*x*) Surui, and (*xi*) Karitiana and Surui. In all EEMS runs we removed SNVs with ≥10% missingness (83, 84). We optimized EEMS parameters by adjusting the qEffctProposalS2, qSeedsProposalS2, mEffctProposalS2, mSeedsProposalS2, mrateMuProposalS2, and negBiProb such that the acceptance proportions for all parameters, except degrees of freedom, were within 10–40%, as suggested by Petkova et al. (59). All models were tested with 300 demes and the MCMC chain was run for 15 million iterations with a 14 million iteration burn-in for nine independent runs to ensure the MCMC chain converged to the optimal log posterior. We then plotted the migration rates, diversity rates, MCMC chain log posterior, dissimilarities within sampled demes, dissimilarities between pairs of sampled demes in relation to fitted dissimilarities, and geographic distance between demes with the EEMS distributed R scripts (59). We ensured that at least three of the nine MCMC chains converged and that the dissimilarities within and between demes met the accepted distributions as specified by Petkova et al. (59). We then replotted all migration and diversity plots for each model with only the MCMC chains that converged to the maximum log posterior probability.

Data Availability. Access to the data is available upon request of the authors through a Control Access Committee of the Peruvian NIH, please contact the corresponding authors for the access directions.

ACKNOWLEDGMENTS. We thank Susan O'Connor, Claire Fraser, Lance Nickel, Dr. Cesar Cabezas, and Ruth Shady Solis for their constructive comments and perspectives. We thank all those who facilitated the recruitment of participants, including the Direcciones Regionales de Salud from Loreto, Puno, Cusco, La Libertad, Huancavelica, Ica, Piura, Ancash, Arequipa, Ayacucho, Tacna, Ucayali, San Martin, Amazonas; Universidad Andina Nestor Caceres Velasquez Facultad de Ciencias de la Salud, Universidad Nacional Jorge Basadre Grohmann, Universidad Nacional Mayor de San Marcos, Universidad Nacional San Agustín, Universidad Nacional de San Cristóbal de Huamanga, Universidad Nacional Santiago Antúnez de Mayolo, and Universidad Nacional de Trujillo; and all participants in this study. This work was supported by the Center for Health Related Informatics and Bioimaging at the University of Maryland School of Medicine (D.N.H., M.D.K., A.C.S., and T.D.O.), the Institute for Genome Sciences and Program in Personalized Genomic Medicine at the University of Maryland School of Medicine (T.D.O.), and the Instituto Nacional de Salud, Lima, Perú (K.S.L., O.C., C.P., D.T., V.B., O.T., C.S., M.G., H.M., P.O.F.-V., and H.G.).

1. Abecasis GR, et al.; 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
2. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
3. Gravel S, et al.; 1000 Genomes Project (2013) Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet* 9:e1004023.
4. Homburger JR, et al. (2015) Genomic insights into the ancestry and demographic history of South America. *PLoS Genet* 11:e1005602.
5. Skoglund P, et al. (2015) Genetic evidence for two founding populations of the Americas. *Nature* 525:104–108.
6. Raghavan M, et al. (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505:87–91.
7. Raghavan M, et al. (2015) POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349:aab3884.
8. Fagundes NJ, et al. (2008) Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet* 82:583–592.
9. Llamas B, et al. (2016) Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv* 2:e1501385.
10. Tamm E, et al. (2007) Beringian standstill and spread of Native American founders. *PLoS One* 2:e829.
11. Skoglund P, Reich D (2016) A genomic view of the peopling of the Americas. *Curr Opin Genet Dev* 41:27–35.
12. Pedersen MW, et al. (2016) Postglacial viability and colonization in North America's ice-free corridor. *Nature* 537:45–49.
13. Wang S, et al. (2007) Genetic variation and population structure in Native Americans. *PLoS Genet* 3:e185.
14. Dillehay TD, et al. (2008) Monte Verde: Seaweed, food, medicine, and the peopling of South America. *Science* 320:784–786.
15. Scliar MO, et al. (2014) Bayesian inferences suggest that Amazon Yunga Natives diverged from Andeans less than 5000 ybp: Implications for South American prehistory. *BMC Evol Biol* 14:174.
16. Lewis CM, Jr, et al. (2007) Mitochondrial DNA and the peopling of South America. *Hum Biol* 79:159–178.
17. Rothhammer F, Dillehay TD (2009) The late Pleistocene colonization of South America: An interdisciplinary perspective. *Ann Hum Genet* 73:540–549.
18. Reich D, et al. (2012) Reconstructing Native American population history. *Nature* 488:370–374.
19. Fuselli S, et al. (2003) Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. *Mol Biol Evol* 20:1682–1691.
20. Hunley K, Long JC (2005) Gene flow across linguistic boundaries in Native North American populations. *Proc Natl Acad Sci USA* 102:1312–1317.
21. Tarazona-Santos E, et al. (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *Am J Hum Genet* 68:1485–1496.

22. Moreno-Estrada A, et al. (2014) Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344:1280–1285.
23. Rodriguez-Delfin LA, Rubin-de-Celis VE, Zago MA (2001) Genetic diversity in an Andean population from Peru and regional migration patterns of Amerindians in South America: Data from Y chromosome and mitochondrial DNA. *Hum Hered* 51:97–106.
24. Lewis CM, Jr, Tito RY, Lizárraga B, Stone AC (2005) Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru. *Am J Phys Anthropol* 127:351–360.
25. Sandoval JR, et al. (2013) Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. *J Hum Genet* 58:627–634.
26. Dufour DL, Piperata BA (2004) Rural-to-urban migration in Latin America: An update and thoughts on the model. *Am J Hum Biol* 16:395–404.
27. Pereira L, et al. (2012) Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population. *PLoS One* 7:e41200.
28. D'Altroy TN (2014) *The Incas* (Wiley, New York).
29. De Carvalho Teixeira De Freitas LC (2009) *Who Were the Inca?* (Biblioteca24horas, São Paulo, Brazil).
30. Mannheim B (1991) *The Language of the Inka Since the European Invasion* (Univ of Texas Press, Austin, TX), 1st Ed.
31. Mumford JR (2012) *Vertical Empire: The General Resettlement of Indians in the Colonial Andes* (Duke Univ Press, Durham, NC).
32. Stern SJ (1993) *Peru's Indian Peoples and the Challenge of Spanish Conquest: Huamanga to 1640* (Univ of Wisconsin Press, Madison, WI).
33. Coe M, Snow D, Benson E (1986) *Atlas of Ancient America* (Facts on File, New York).
34. Lovell WG (1992) "Heavy shadows and black night": Disease and depopulation in colonial Spanish America. *Ann Assoc Am Geogr* 82:426–443.
35. Dobyns HF (1966) An appraisal of techniques with a new hemispheric estimate. *Curr Anthropol* 7:395–416.
36. McEwan GF (2008) *The Incas: New Perspectives* (Norton, New York).
37. Espinoza WS (1997) Las llactas en el Imperio de los Incas. *Actas y trabajos del XI Congreso Peruano del Hombre y la Cultura Andina*, eds Olazábal HA, Palomino LG (Universidad Nacional Hermilio Valdizán de Huánuco, Huánuco, Peru). Spanish.
38. Kessler MD, et al.; Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (2016) Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nat Commun* 7:12521.
39. Petrovski S, Goldstein DB (2016) Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol* 17:157.
40. Lazaridis I, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.
41. Bryc K, et al. (2010) Colloquium paper: Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci USA* 107:8954–8961.
42. Wang S, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4:e1000037.
43. Browning SR, et al. (2016) Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *G3 (Bethesda)* 6:1525–1534.
44. Moreno-Estrada A, et al. (2013) Reconstructing the population genetic history of the Caribbean. *PLoS Genet* 9:e1003925.
45. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.
46. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 40:646–649.
47. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8:e1002967.
48. Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D (2000) Genetic structure of Quechua-speakers of the Central Andes and geographic patterns of gene frequencies in South Amerindian populations. *Am J Phys Anthropol* 113:5–17.
49. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5:e1000695.
50. Hellenthal G, et al. (2014) A genetic atlas of human admixture history. *Science* 343: 747–751.
51. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8:e1002453.
52. Baharian S, et al. (2016) The Great Migration and African-American genomic diversity. *PLoS Genet* 12:e1006059.
53. Palamara PF, Pe'er I (2013) Inference of historical migration rates via haplotype sharing. *Bioinformatics* 29:i180–i188.
54. Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91:809–822.
55. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084–1097.
56. Gravel S (2012) Population genetics models of local ancestry. *Genetics* 191:607–619.
57. Fu W, et al.; NHLBI Exome Sequencing Project (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220, and erratum (2013) 495:270.
58. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69.
59. Petkova D, Novembre J, Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* 48:94–100.
60. INEI (2012) Perú: Estimaciones y Proyecciones de Población Total y por Sexo de las Ciudades Pricipales, 2000-2015 (Dirección Técnica de Demografía e Indicadores Sociales del Instituto Nacional de Estadística e Informática, Lima, Peru), p 17. Available at proyectos.inei.gob.pe/web/biblioineipub/bancopub/Est/Lib1020/index.html. Accessed September 13, 2017. Spanish.
61. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
62. Narasimhan VM, et al. (2016) Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 352:474–477.
63. Strauss KA, Puffenberger EG (2009) Genetics, medicine, and the plain people. *Annu Rev Genomics Hum Genet* 10:513–536.
64. Morton DH, et al. (2003) Pediatric medicine and the genetic disorders of the Amish and Mennonite people of Pennsylvania. *Am J Med Genet C Semin Med Genet* 121C: 5–17.
65. Kristiansson K, Naukkarinen J, Peltonen L (2008) Isolated populations and complex disease gene identification. *Genome Biol* 9:109.
66. Kessler MD, O'Connor TD (2017) Accurate and equitable medical genomic analysis requires an understanding of demography and its influence on sample size and ratio. *Genome Biol* 18:42.
67. Landrum MJ, et al. (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42:D980–D985.
68. Bodner M, et al. (2012) Rapid coastal spread of First Americans: Novel insights from South America's Southern Cone mitochondrial genomes. *Genome Res* 22:811–820.
69. Bigham AW (2016) Genetics of human origin and evolution: High-altitude adaptations. *Curr Opin Genet Dev* 41:8–13.
70. Bigham AW, et al. (2013) Andean and Tibetan patterns of adaptation to high altitude. *Am J Hum Biol* 25:190–197.
71. Bigham AW, et al. (2009) Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum Genomics* 4:79–90.
72. Castilla EE, Schuler-Faccini L (2014) From rumors to genetic isolates. *Genet Mol Biol* 37(Suppl):186–193.
73. Walter K, et al.; UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526:82–90.
74. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11:415–425.
75. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2.
76. Broad Institute (2018) Picard. Available at broadinstitute.github.io/picard/. Accessed March 13, 2018.
77. New York Genome Center (2017) Whole Genome Sequencing – Germline (New York Genome Center, New York). Available at www.nygenome.org/wp-content/uploads/2018/01/Whole-Genome-Sequencing-Germline.pdf. Accessed March 13, 2018.
78. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.
79. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
80. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11.10.1–11.10.33.
81. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
82. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303.
83. Chang CC, et al. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4:7.
84. Purcell S, Chang C (2017) PLINK 1.9. beta. Available at www.cog-genomics.org/plink/1.9/. Accessed March 13, 2018.
85. Weissensteiner H, et al. (2016) HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* 44:W58–W63.
86. Kloss-Brandstätter A, et al. (2011) HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* 32:25–32.
87. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30:E386–E394.
88. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.
89. Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246.
90. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664.
91. 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* 84:343–364.
92. Montinaro F, et al. (2015) Unravelling the hidden ancestry of American admixed populations. *Nat Commun* 6:6596.
93. Han E, et al. (2017) Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun* 8:14238.
94. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package). (Dept of Genome Sciences, Univ of Washington, Seattle), Version 3.6.
95. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874.
96. O'Leary NA, et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733–D745.
97. Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* 24: 1792–1800.
98. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.