HHS Public Access

Author manuscript

Am J Biol Anthropol. Author manuscript; available in PMC 2024 July 30.

Published in final edited form as:

Am J Biol Anthropol. 2023 April ; 180(4): 703–714. doi:10.1002/ajpa.24673.

Spatiotemporal fluctuations of population structure in the Americas revealed by a meta-analysis of the first decade of archaeogenomes

Andre Luiz Campelo dos Santos[1,2], **Henry Socrates Lavalle Sullasi**[2], **Omer Gokcumen**[3], **John Lindo**[4], **Michael DeGiorgio**[1]

[1]Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, Florida, USA

[2]Department of Archaeology, Federal University of Pernambuco, Recife, Pernambuco, Brazil

[3]Department of Biological Sciences, University at Buffalo, Buffalo, New York, USA

[4]Department of Anthropology, Emory University, Atlanta, Georgia, USA

Abstract

Objectives: Since 2010, genome-wide data from hundreds of ancient Native Americans have contributed to the understanding of Americas' prehistory. However, these samples have never been studied as a single dataset, and distinct relationships among themselves and with present-day populations may have never come to light. Here, we reassess genomic diversity and population structure of 223 ancient Native Americans published between 2010 and 2019.

Materials and Methods: The genomic data from ancient Americas was merged with a worldwide reference panel of 278 present-day genomes from the Simons Genome Diversity Project and then analyzed through ADMIXTURE, *D*-statistics, PCA, t-SNE, and UMAP.

Results: We find largely similar population structures in ancient and present-day Americas. However, the population structure of contemporary Native Americans, traced here to at least 10,000 years before present, is noticeably less diverse than their ancient counterparts, a possible outcome of the European contact. Additionally, in the past there were greater levels of population

Correspondence: Andre Luiz Campelo dos Santos and Michael DeGiorgio, Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. acampelodossanto@fau.edu and mdegiorg@fau.edu, John Lindo, Department of Anthropology, Emory University, Atlanta, GA, USA. john.lindo@emory.edu.

AUTHOR CONTRIBUTIONS

Andre Luiz Campelo dos Santos: Conceptualization (equal); investigation (equal); methodology (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Henry Socrates Lavalle Sullasi:** Conceptualization (equal); supervision (equal). **Omer Gokcumen:** Conceptualization (equal); investigation (equal); methodology (equal); supervision (equal); visualization (equal); writing – review and editing (equal). **John Lindo:** Conceptualization (equal); investigation (equal); methodology (equal); writing – review & editing (equal). **Michael DeGiorgio:** Conceptualization (equal); investigation (equal); methodology (equal); supervision (equal); visualization (equal); writing – original draft (equal); writing – review & editing (equal).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

structure in North than in South America, except for ancient Brazil, which harbors comparatively high degrees of structure. Moreover, we find a component of genetic ancestry in the ancient dataset that is closely related to that of present-day Oceanic populations but does not correspond to the previously reported Australasian signal. Lastly, we report an expansion of the Ancient Beringian ancestry, previously reported for only one sample.

Discussion: Overall, our findings support a complex scenario for the settlement of the Americas, accommodating the occurrence of founder effects and the emergence of ancestral mixing events at the regional level.

Keywords

ancestry; ancient genomics; meta-analysis; native Americans; population structure

1 | INTRODUCTION

In 2010, the near-complete genome of an ancient human individual was first published. The genomic sequence was obtained from the DNA present in an ~4000-year-old permafrost-preserved tuft of hair excavated from culturally-deposited sediments at Qeqertasussuk, Greenland (Rasmussen et al., 2010). This achievement inaugurated a prolific era of ancient humans' genome-wide studies that have been conducted not only in the Americas, but also around the world (Nielsen et al., 2017). In the former, however, these studies have played a leading role in the most current discussions about the history of the continents (Waters, 2019).

While it is unanimously accepted that the Americas were the last continents populated by humans, the timing of this first arrival, their places of origin, the routes they took to enter and subsequently explore the continents (how many migratory movements took place), the speed of human dispersal at different times and regions, and how they settled in its most diverse and extreme environments are intriguing questions that have been the target of major (and sometimes heated) debates (Guidon et al., 1996; Meltzer et al., 1994; Moreno-Mayar, Vinner, et al., 2018; Waters, 2019). Addressing these questions is key to understanding the ancestral history of prehistoric and contemporary Native American populations (Skoglund et al., 2015; Skoglund & Reich, 2016). To better study such aspects, since 2010, after years of ancient DNA studies limited to the analysis of mitochondrial and eventually Y-chromosome DNA, genome-wide data from hundreds of ancient humans have been generated and published (Flegontov et al., 2019; Lindo et al., 2016; Lindo et al., 2017; Lindo, Haas, et al., 2018; Moreno-Mayar, Potter, et al., 2018; Moreno-Mayar, Vinner, et al., 2018; Posth et al., 2018; Raghavan et al., 2015; Rasmussen et al., 2010; Rasmussen et al., 2014; Rasmussen et al., 2015; Scheib et al., 2018). This deluge of data has been fueled by recent advances and technical breakthroughs in DNA sequencing technologies that occurred in the last 10–15 years (Rasmussen et al., 2010; Waters, 2019).

Prior to this work, however, these ancient individuals have not been studied as a single set of samples. As a result, important information regarding possible genomic affinities and relationships among them (and with present-day populations) and a more comprehensive assessment of the genomic diversity in the ancient Americas may have never come to light.

To holistically evaluate variation in the ancient Americas, we conduct a reassessment of the genomic diversity and structure of the ancient individuals of the Americas published between 2010 and 2019. We combine ancient and extant genome-wide data into a single dataset and employ a suite of established quantitative and qualitative techniques to investigate potential genomic relationships among past and present-day populations. Our investigation provides additional evidence for the current hypothesis on the settlement of the Americas, while revealing new aspects of its genomic history.

2 | MATERIALS AND METHODS

2.1 | Dataset preparation and description

We compiled previously published genome-wide data from 223 ancient human individuals unearthed in eight countries and one territory of the Americas (Figure 1a and Supporting Information S1). Their estimated chronologies range from the Pleistocene (>11,700 years before present) to post-European contact periods (approximately the last 500 years). However, 72 individuals have not yet been directly dated (Flegontov et al., 2019; Lindo et al., 2016; Lindo et al., 2017; Lindo, Haas, et al., 2018; Moreno-Mayar, Potter, et al., 2018; Moreno-Mayar, Vinner, et al., 2018; Posth et al., 2018; Raghavan et al., 2015; Rasmussen et al., 2010; Rasmussen et al., 2014; Rasmussen et al., 2015; Scheib et al., 2018) (Data S1).

We generated a single variant call format (VCF) file with genotype data from all the ancient individuals. Initially, this VCF file contained almost 17 million variant loci. However, a diverse range of genomic coverages can be observed across individuals in the dataset: some low-coverage genomes harbored only a few hundred loci with data, whereas high-coverage ones carried more than seven million variant loci (Figure 1b). To guard against potential biases caused by post-mortem DNA damage patterns characteristic of ancient individuals (Dabney et al., 2013; Ginolhac et al., 2011) and to prevent erroneous inferences (Axelsson et al., 2008), we removed variant loci for which C → T and G → A transitions (in a REF → ALT scheme) were observed. We removed mitochondrial and sex-chromosomal loci using BCFtools (Danecek et al., 2021), thus proceeding with only autosomal data. At the end of these steps, more than seven million variant loci were retained in the VCF file (Figure 1c).

This dataset was then expanded with the addition of a reference panel composed of genotypes at variant loci from 278 extant human individuals from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016) (Table S1). The merged dataset was further pruned by removing multiallelic sites and indels, monomorphic sites, and loci in strong linkage disequilibrium ($r^2 > 0.1$ in windows of 50 SNPs) with BCFtools (Danecek et al., 2021). After these last filtering steps, a total of 284,010 SNPs and 500 ancient and extant individuals remained for downstream analyses (Figure 1c)—sample CK-07, from Canada, harbored no data after filtering.

Finally, using the software PLINK v.1.90 (Chang et al., 2015), we converted the VCF file containing the remaining 500 individuals to a PED “12” coded file, which was used for the two analyses to assess population structure. The first analysis employed the soft-clustering algorithm implemented in ADMIXTURE (Alexander et al., 2009) and the second was linear dimensionality reduction in the form of principal component analysis (PCA) and nonlinear

dimensionality reduction in the forms of t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP).

2.2 | Soft clustering with ADMIXTURE

We reassessed the population structure of the ancient individuals of the Americas with the software ADMIXTURE v.1.3.0 (Alexander et al., 2009), which assigns individuals to K clusters with some probability, with each cluster taken to represent some ancestral component in the dataset. The number of ancestral components (K), however, is user-defined, and so we considered a wide range of values from $K = 2$ to 20 for the first round of ADMIXTURE analyses applied to the whole dataset. The best run (i.e., the optimal value for K, which can be interpreted as the number of ancestral components that best explains the data) was chosen based on the smallest 10-fold cross-validation error after 100 iterations using the "-cv = 10" and "-C 100" options, respectively. Two additional rounds of ADMIXTURE analyses were conducted: one after removing low-coverage ancient individuals, and another after the removal of present-day individuals with African ancestry. The rationale behind these filtering procedures is presented in the Supporting Information S1 (*Initial soft clustering analysis with ADMIXTURE*). Based on the results of the first round of ADMIXTURE analyses (Supporting Information S1), we limited the range of values for K between three and 10 in subsequent rounds. To visualize the best ADMIXTURE run (as is common in the form of bar plots), we used the R package POPHELPER v.2.3.1 (Francis, 2017).

2.3 | Linear dimensionality reduction with principal component analysis

To investigate broad genomic affinities among the ancient and present-day individuals, we conducted PCA on our dataset using the "smartpca" program from the EIGENSOFT v7.2.1 package (Patterson et al., 2006). Principal components (PCs) were calculated using the present-day populations with the "poplistname" and "autoshrink: YES" options. Ancient data, characterized by a large proportion of missing sites, were then projected onto the computed PCs with the "lsqproject: YES" option. We also used the "nomoutevec: 50" option to retain and output the first 50 PCs. No outliers were excluded for this analysis. Using the Python package "seaborn" (Waskom, 2021), we produced scatterplot visualizations of the PCA results of the first four PCs, though the full set of PCs were later fed as downstream input to t-SNE and UMAP.

2.4 | Nonlinear dimensionality reduction with t-SNE and UMAP

The previously computed PCs were target of further nonlinear dimensionality reduction with t-SNE and UMAP. In these two methods, all previously-computed PCs can be used as input, thus preserving the greater overall dissimilarity of high-dimensional data in a two- or three-dimensional embedded space (Diaz-Papkovich et al., 2021; Li et al., 2017). We applied both methods to obtain embeddings of the top 10, 30, and 50 PCs into a two-dimensional space. This complementary analysis to PCA has the potential for more informative visualization of the genetic structure of sampled individuals, as visualizations resulting directly from PCA only account for the variance explained by two or three PCs at a time. To apply both methods to our dataset, we used the t-SNE and UMAP implementations present in the "scikit-learn" and "umap-learn" Python libraries, respectively (McInnes et al.,

2018; Pedregosa et al., 2011), and visualized results through scatterplots produced using "seaborn" (Waskom, 2021).

3 | RESULTS

3.1 | Population structure

In the final round of the ADMIXTURE analyses, the optimal *K* value for our dataset was four (Table S2). We find that the prevalent ancestral component in present-day Americas, Cluster4 (green), is also predominant in the ancient samples, which is expected (Figure 2). Cluster2 (light blue) and Cluster1 (dark blue), respectively the second and third most frequently observed components in the ancient individuals, are shared worldwide. The diminutive Cluster3 (red), is restricted to present-day Oceania, South Asia, and a few individuals of East Asia.

A component in ancient Americas that, in present-day populations, is restricted to Oceania and South and East Asia is an intriguing finding as it is reminiscent of the previously-reported Australasian signal only observed for an ancient genome from Lagoa Santa (Sumidouro5) and in the present-day Surui, both sampled in Brazil (Moreno-Mayar, Vinner, et al., 2018; Skoglund et al., 2015). However, though none of these individuals present this Cluster3 component, we elected to examine its occurrence more deeply.

Because the presence of Cluster3 in ancient Americas could be a genomic artifact (e.g., as in Figures S3 and S4), we first correlated the proportion of Cluster3 with genomic coverage across 13 ancient individuals (318, 406, 413, 468, 532, B-04, I9054_d, LU-02, LU-03, SN-12, SN-20, SN-43, and SN-55). We specifically analyzed these 13 ancient samples because they display a higher proportion of Cluster3 component than most of the present-day individuals—the exceptions are Papuans, Bougainvilleans, and Indigenous Australians. Conversely to the African component observed in Figure S3, we find no correlation between the proportion of Cluster3 component and the coverage of the 13 genomes (Figure S5), suggesting that this component's presence in ancient Americas is not a genomic artifact.

We next attempted to identify the previously-reported Australasian signal in these individuals through *D*-statistic tests of the form *D* (Yoruba, Simons; Mixe, Ancient) (Moreno-Mayar, Vinner, et al., 2018; Skoglund et al., 2015), in which "Simons" is a non-African and non-American population from the SGDP (Mallick et al., 2016)—more details on these tests are presented in the Supporting Information (*D-statistic analysis*). We could not find excess affinity between a present-day population and the 13 ancient samples in comparison to the Mixe ($Z > 3$) (Figure 3 and Figure S6).

Ten of these thirteen samples were unearthed in North America's Pacific coast (Lindo et al., 2016; Scheib et al., 2018), whereas samples LU-02 and LU-03 were found in Ontario, Central Canada (Scheib et al., 2018). Only I9054_d was excavated in the Atlantic coast (Posth et al., 2018). These observations imply that, though the presence of Cluster3 in ancient Americas remains an intriguing finding, it likely entered the continents following general dispersal through Beringia.

Though Cluster2 is shared worldwide (Figure 2), the finding that ancient samples of the Americas harbor a relevant proportion of this genomic ancestry is also intriguing. We thus elected to further investigate Cluster2 in the ancient samples, as we have previously carried out for the African genomic ancestry and Cluster3. We find a statistically significant negative correlation between coverage and Cluster2 (Figure S7), that is, the lower the coverage, the higher the proportion of Cluster2. This result likely indicates that the presence of Cluster2 in the ancient dataset may also be a genomic artifact. On the other hand, we also find relevant proportions of this genomic ancestry in present-day populations from Central Asia and Siberia, which suggest that this component may have entered the Americas through the Beringia in ancient times.

Having further explored individual ancestral components in the ancient dataset, we are now able to shed some light on the present-day samples of the Americas. We find highly homogenous ancestral component structure in these populations, as opposed to a previous study that reported evidence of genomic substructure in other present-day populations of the Americas (Moreno-Estrada et al., 2014). To assess signals of substructure in our Native American dataset, we performed further ADMIXTURE analyses restricted to either Native-American populations or SNPs defining the prevalent genomic ancestry in the Americas (see Assessment of genomic substructure in ancient and present-day Americas subsection in the Supporting Information S1).

First, we restricted the ADMIXTURE analysis to ancient and present-day individuals of the Americas. We find that the optimal number of clusters (K) for this Native American dataset is one (second column of Table S3). This result indicates that putative signs of substructure in our dataset appear to be nuanced in light of the overall pattern of ancestral component structure that can be observed in the Americas.

Then, as a second test, we filtered the dataset used in the final round of ADMIXTURE runs (Figure 2), removing SNPs that did not define Cluster4 (green color; the prevalent genomic ancestry in the Americas) in the first round of ADMIXTURE runs (Figure S1; see Supporting Information). We find that the overall ancestral component structure observed in ancient and present-day Americas becomes even more homogenous (Figure S9), especially when ancient South America is compared with present-day Native American populations, which is consistent with the results presented in the previous paragraph that indicate little to no substructure in the Americas (see second column of Table S3). Moreover, the proportions of a putative West Eurasian ancestry in the ancient dataset considerably decreased (Figure S9) when compared with our original result (Figure 2).

Furthermore, to test whether our findings would remain consistent if we have considered a less restrictive scenario of linkage disequilibrium, we generated an alternate dataset using $r^2 > 0.4$ as the linkage disequilibrium pruning threshold and performed a set of ADMIXTURE analyses, as in our original dataset (see Supporting Information S1). While there are no major discrepancies in the component structure of the alternate dataset (Figure S11) in comparison with our original results (Figure S1), we find that the proportion of Cluster2 (predominant in West Eurasia) in the ancient samples is higher in the alternate (Figure S11; $r^2 > 0.4$) than in our original dataset (Figure S1; $r^2 > 0.1$).

We then proceeded with the ADMIXTURE analysis restricted to ancient and present-day individuals of the Americas (see Supporting Information S1). Consistent with the result obtained for the original dataset, we find that the optimal number of clusters (K) for the alternate Native American dataset is also one (fourth column of Table S3). Finally, replicating the second test performed with the original dataset, we retained only Cluster4-defining SNPs in the alternate dataset (green color in Figure S11; see Supporting Information), removed the 41 low-coverage ancient genomes (Table S4) and African samples, and then proceeded with a final ADMIXTURE analysis. We find, however, an alternate scenario in which a putative West Eurasia genomic ancestry (Cluster2) is also prevalent in the Americas, especially in the past (Figure S12). This result is therefore inconsistent with our previous findings (Figure S9) and is explained by the use of a less restrictive threshold for linkage disequilibrium pruning, causing the retention of SNPs that are also associated with West Eurasian haplotypes.

3.2 | Assessing diversity

To quantitatively assess the genomic diversity of the ancient individuals while accounting for the number of missing genotypes in each sample, we estimated observed genomic heterozygosity with the aid of the software BCFtools (Danecek et al., 2021). Heterozygosity was calculated for the 182 ancient individuals analyzed in the last round of ADMIXTURE runs. We find, however, that heterozygosity positively correlates with coverage in the ancient dataset ($p < 0.001$) (Figure S13), which means that high-coverage ancient genomes will tend to show higher heterozygosity, regardless of actual diversity. That being the case, given the range of sequencing coverages of the samples in our dataset, heterozygosity will not be a reliable measure for exploring patterns of diversity.

Therefore, we elected to investigate patterns of diversity in the ancestral component structure of the ancient individuals as a proxy for genomic diversity. Specifically, assuming that the number of estimated ancestral components in the ADMIXTURE analysis is K, we compute diversity of these ancestral components in each individual using the Gini impurity (Breiman et al., 1984), often employed as a measure of entropy or diversity (Yuan et al., 2021), as

$$\text{Gini impurity} = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k is the proportion of component k, $k = 1,2,\ldots,K$, estimated by ADMIXTURE for the individual. The Gini impurity is also termed the Gini diversity index (Breiman et al., 1984) and the Gini-Simpson index of diversity (Caso & Angeles Gil, 1988). This measure has a maximum value of $(K-1)/K$, when the cluster proportions in an individual are identical across all clusters ($p_1 = p_2 = \ldots = p_K$) and a minimum value of zero, when the proportion for one cluster is one and the remaining $K-1$ clusters are proportion zero (Breiman et al., 1984). Hence, assuming four clusters as in Figure 2, individuals with the highest diversity have values of 3/4. A Gini impurity was calculated for each of the 182 ancient individuals analyzed in the last round of ADMIXTURE runs.

Correlating the Gini impurity across ancestral component proportions with coverage, we find that the Gini impurity is a reliable proxy for assessing diversity in a given genomic cohort regardless of the genomic coverages of the samples (Figure 4a). When comparing different regions of the Americas using this measure, we find that the distribution of ancestral component diversity is shifted toward higher values for ancient individuals unearthed in North America compared to those found in South America (Figure 4b). However, when stratifying by country we find that ancient Brazil harbors similar degrees of diversity as Canada and the USA (Figure 4c). Though, in contemporary times, mean ancestral component diversity harbored by the Indigenous populations of the Americas, as a whole, is near its minimum of zero (Figure 2).

Moreover, we find that ancestral component diversity decreased in the first few millennia following the initial settlement of the Americas (Figure 4d), which is an expected observation as overall genomic diversity is believed to decrease over time due to genetic drift (Allendorf, 1986; Star & Spencer, 2013). However, we see an increase in ancestral component diversity initiating approximately 7000 years before present and eventually reaching similar levels as in the first migrations into the Americas (Figure 4d). In addition, diversity was maintained at a steady level over the last 2000 years, until around European contact (Figure 4d and Figure S14). These are surprising findings, though we are not directly measuring genomic diversity, and instead measuring diversity of ancestral components or population structure over time.

3.3 | Genomic relationships

The separation of individuals resulting from PCA shows that the ancient individuals mainly cluster with present-day samples from the Americas and Central Asia/Siberia (Figure 5a). However, we notice a trend in PC1 in which ancient genomes with higher coverage (Box-Cox-transformed values) fall closer to present-day individuals from the Americas (Figure 5a and Figure S15 a and b). This positioning of low-coverage genomes in PC1 appears to be significantly affected by their higher proportions of homozygous genotypes (Figure S13), which might be driving patterns of population differentiation, that is, F_{ST} (Meirmans & Hedrick, 2011), between high- and low-coverage ancient genomes. While we are not directly measuring F_{ST} here, it has been demonstrated that the fraction of the total variance explained by PC1 is equal to F_{ST} (McVean, 2009). Analogously, Peter (2022) highlights that distance between individuals in a PCA plot is proportionate to the f_2-statistic. We thus hypothesize that the genotypes estimated for the low-coverage ancient genomes might be proportionally more discrepant from their true values than those from the high-coverage ancient samples, which would result in a high f_2 distance between low- and high-coverage ancient individuals. Moreover, it has also been demonstrated that the distance between samples in the PCA plot will always be an underestimate of the full f_2 distance (Peter, 2022), resulting in two populations appearing to be closer to each other than they really are, which appears to be the case for the low-coverage ancient genomes in relation to the Eurasian samples.

This trend becomes clearer after further nonlinear dimensionality reduction with UMAP and t-SNE (Figure 5b,c, and Figure S15 c–f). However, there are some exceptions to this pattern,

with many ancient individuals with relatively high coverage falling closer to present-day Central Asian/Siberian individuals (Figure 5a). In the UMAP and t-SNE results, specifically, we can see three ancient individuals falling with Central Asians/Siberians (Figure 5b,c): I0719 from the Aleutian Islands (Flegontov et al., 2019), USR1 unearthed in Alaska (Moreno-Mayar, Potter, et al., 2018), and Saqqaq found in Greenland (Rasmussen et al., 2010) (Figure 5d).

The fact that three ancient individuals cluster with present-day samples from Central Asia/Siberia could represent an expansion of the previously-reported Ancient Beringian ancestry (Moreno-Mayar, Potter, et al., 2018). Among the ancient individuals of the Americas, this putative distinct ancestry would be restricted to the Arctic portion of North America. Because USR1 is the oldest individual among them, it is then possible to propose a west-to-east directionality to the Ancient Beringian ancestry's dispersion (Figure 5e), consistent with previously-proposed hypotheses on ancient migrations in northern North America (Skoglund & Mathieson, 2018; Willerslev & Meltzer, 2021). Such dispersion should have occurred at least 4000 before present (Saqqaq's age; Rasmussen et al., 2010), putatively made possible by a scenario of less glacial coverage in North America, but still lower sea levels worldwide.

4 | DISCUSSION

Taken together, our analyses point to a Beringian origin for the first settlers of the Americas, as expected. Though we find a putative "Australasian" (Oceanic- and South Asian-specific) ancestry component in some ancient individuals from ADMIXTURE analysis (Figure 2 and Figure S5), it does not mean a direct migration from these continents. Also, because this component is not found in the ancient Sumidouro5 and present-day Surui from Brazil (Figure 2), it cannot be translated as the previously-reported Australasian signal (Moreno-Mayar, Vinner, et al., 2018; Skoglund et al., 2015)—which suggests that the component followed the general migratory events into the Americas (from Beringia).

We also find that the population structure identified in present-day Americas can be traced to at least 10,000 years before present, as indicated by the structure of Sumidouro5 and Sumidouro6, from Lagoa Santa in Brazil (Moreno-Mayar, Vinner, et al., 2018), which are highly similar to present-day Quechua, Zapotecs and Mixtecs (Figure 2). On the other hand, the fact that present-day Native American populations harbor lower degrees of ancestral component diversity than their ancient counterparts (Figure 2) might be one of the many direct outcomes of European contact (Lindo, Rogers, et al., 2018). Moreover, we also observed that the ancient individuals of North America had higher degrees of ancestral component diversity than those from South America (Figure 4b). This finding might indicate one of the following scenarios (or a mix of them): (1) over time, there was a continual flow of migrations from Eastern Asia into the Americas in the past; or (2) founder effects occurred at some point during the settlement of the American continents, leading to losses of distinct ancestral components as small human groups migrated in a North–South direction. Model formulations such as the serial founder model (DeGiorgio et al., 2009; Ramachandran et al., 2005) have been shown to be able to give expected patterns from scenario 2, whereas similar patterns from scenario 1 have been shown to give rise due to ancestral mixing of population without need for a series of founding events (Pickrell & Reich, 2014). However,

a mixture of both scenarios is likely the most reasonable explanation for the observed trends (Figure 4b–d), and diversity in ancient Brazil illustrates this belief. In particular, though founder effects appear to have occurred as humans migrated from North to South America, previously reported ancestral mixing events emerging in eastern South America (dos Santos et al., 2022; Lindo et al., 2022), may have produced similar levels of diversity as those in North America (Figure 4c). In addition, Figure 4d depicts an increase of diversity happening in the Americas some millennia after the initial settlement of the continents, suggesting the occurrence of new ancestral migrations into the Americas at approximately 7000 years before present (likely followed by mixing events).

Lastly, our study presents an overview of the patterns of genomic coverage achieved in the first decade of archaeogenomics. Based on these patterns, we can hypothesize that the archeological sites located closer to the Pacific coast seem to provide optimal environmental conditions for DNA preservation, regardless of the estimated age of the archeological samples (Figure S16). On the other hand, sites located in intertropical zones (usually highly humid and warm) appear to not allow for much DNA persistence—exceptions being a few areas that seem to present specific environmental conditions that favor DNA preservation, for example, the Andes and Sumidouro Cave (Figure 1b). However, it is important to note that these observations reflect the broad patterns of genomic coverage surveyed in this study, which could indeed suggest trends in DNA preservation and quality, though these aspects have not been directly assessed here.

Overall, our results provide a holistic examination of population structure and ancient genomic diversity in the Americas while also contributing novel insights on the Beringian origin of the first settlers of the continents.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We express our gratitude to the National Council for Scientific and Technological Development of Brazil (CNPq), the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) and the Pernambuco State Foundation for Science and Technology (FACEPE), from the State of Pernambuco, Brazil, for providing funding support to Andre Luiz Campelo dos Santos in the form of a Doctoral Scholarship (process number 140239/2016-2), a PDSE Sandwich Doctorate Scholarship (process number 88881.189760/2018-01) and a BFP Postdoctoral Fellowship (process number BFP-0191-7.04/20), respectively, throughout the duration of this research. This work was also supported by National Science Foundation grants BCS-2001063, DEB-1949268 and DBI-2130666 to Michael DeGiorgio, and by National Institutes of Health grant R35GM128590 to Michael DeGiorgio.

Funding information

Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 140239/2016-2; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 88881.189760/2018-01; National Institutes of Health, Grant/Award Number: R35GM128590; Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, Grant/Award Number: BFP-0191-7.04/20; National Science Foundation, Grant/Award Numbers: BCS-2001063, DBI-2130666, DEB-1949268

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

Alexander DH, Novembre J, & Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664. 10.1101/gr.094052.109 [PubMed: 19648217]

Allendorf FW (1986). Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology, 5(2), 181–190. 10.1002/zoo.1430050212

Axelsson E, Willerslev E, Gilbert MTP, & Nielsen R (2008). The effect of ancient DNA damage on inferences of demographic histories. Molecular Biology and Evolution, 25(10), 2181–2187. 10.1093/molbev/msn163 [PubMed: 18653730]

Breiman L, Friedman JH, Olshen RA, & Stone CJ (1984). Classification and regression trees. Routledge https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-richard-olshen-charles-stone

Caso C, & Angeles Gil M (1988). The Gini-Simpson index of diversity: Estimation in the stratified sampling. Communications in Statistics: Theory and Methods, 17(9), 2981–2995. 10.1080/03610928808829784

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, & Lee JJ (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience, 4(1), 7. 10.1186/s13742-015-0047-8 [PubMed: 25722852]

Dabney J, Meyer M, & Pääbo S (2013). Ancient DNA damage. Cold Spring Harbor Perspectives in Biology, 5(7), a012567. 10.1101/cshperspect.a012567 [PubMed: 23729639]

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, & Li H (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), 1–4. 10.1093/gigascience/giab008

DeGiorgio M, Jakobsson M, & Rosenberg NA (2009). Out of Africa: Modern human origins special feature: Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proceedings of the National Academy of Sciences of the United States of America, 106(38), 16057–16062. 10.1073/pnas.0903341106 [PubMed: 19706453]

Diaz-Papkovich A, Anderson-Trocmé L, & Gravel S (2021). A review of UMAP in population genetics. Journal of Human Genetics, 66(1), 85–91. 10.1038/s10038-020-00851-4 [PubMed: 33057159]

dos Santos ALC, Owings A, Sullasi HSL, Gokcumen O, DeGiorgio M, & Lindo J (2022). Genomic evidence for ancient human migration routes along South America's Atlantic coast. Proceedings of the Royal Society B: Biological Sciences, 289(1986), 20221078. 10.1098/rspb.2022.1078

Flegontov P, Altınışık NE, Changmai P, Rohland N, Mallick S, Adamski N, … Schiffels S (2019). Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. Nature, 570(7760), 236–240. 10.1038/s41586-019-1251-y [PubMed: 31168094]

Francis RM (2017). Pophelper: An R package and web app to analyse and visualize population structure. Molecular Ecology Resources, 17(1), 27–32. 10.1111/1755-0998.12509 [PubMed: 26850166]

Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, & Orlando L (2011). mapDamage: Testing for damage patterns in ancient DNA sequences. Bioinformatics (Oxford, England), 27(15), 2153–2155. 10.1093/bioinformatics/btr347 [PubMed: 21659319]

Guidon N, Pessis A-M, Parenti F, Fontugue M, & Guérin C (1996). Nature and age of the deposits in Pedra Furada, Brazil: Reply to Meltzer, Adovasio & Dillehay. Antiquity, 70(268), 408. 10.1017/s000359800083356

Li W, Cerise JE, Yang Y, & Han H (2017). Application of t-SNE to human genetic data. Journal of Bioinformatics and Computational Biology, 15(4), 1750017. 10.1142/S0219720017500172 [PubMed: 28718343]

Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, … Malhi RS (2017). Ancient individuals from the north American northwest coast reveal 10,000 years of regional genetic continuity. Proceedings of the National Academy of Sciences of the United States of America, 114(16), 4093–4098. 10.1073/pnas.620410114 [PubMed: 28377518]

Lindo J, De La Rosa R, dos Santos ALC, Sans M, DeGiorgio M, & Figueiro G (2022). The genomic prehistory of the indigenous peoples of Uruguay. PNAS Nexus, 1(2), 1–7. 10.1093/pnasnexus/pgac047

Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, … Di Rienzo A (2018). The genetic prehistory of the Andean highlands 7000 years BP though European contact. Science. Advances, 4(11), eaau4921. 10.1126/sciadv.aau4921

Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, … Malhi RS (2016). A time transect of exomes from a native American population before and after European contact. Nature Communications, 7, 13175. 10.1038/ncomms13175

Lindo J, Rogers M, Mallott EK, Petzelt B, Mitchell J, Archer D, …DeGiorgio M (2018). Patterns of genetic coding variation in a native American population before and after European contact. The American Journal of Human Genetics, 102(5), 806–815. 10.1016/j.ajhg.2018.03.008 [PubMed: 29706345]

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, …Reich D (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature, 538(7624), 201–206. 10.1038/nature18964 [PubMed: 27654912]

McInnes L, Healy J, Saul N, & Grossberger L (2018). UMAP: Uniform Manifold Approximation and Projection. The Journal of Open Source Software, 3(29), 861. 10.21105/joss.00861

McVean G (2009). A genealogical interpretation of principal components analysis. PLoS Genetics, 5(10), e1000686. 10.1371/journal.pgen.1000686 [PubMed: 19834557]

Meirmans PG, & Hedrick PW (2011). Assessing population structure: Fst and related measures. Molecular Ecology Resources, 11(1), 5–18. 10.1111/j.1755-0998.2010.02927.x [PubMed: 21429096]

Meltzer DJ, Adovasio JM, & Dillehay TD (1994). On a Pleistocene human occupation at Pedra Furada, Brazil. Antiquity, 68(261), 695–714. 10.1017/s000359800047414

Moreno-Estrada A, Gignoux CR, Fernández-Lopez JC, Zakharia F, Sikora M, Contreras AV, … Bustamante CD (2014). Human genetics. The genetics of Mexico recapitulates native American substructure and affects biomedical traits. Science (New York, N.Y.), 344(6189), 1280–1285. 10.1126/science.1251688 [PubMed: 24926019]

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, … Willerslev E (2018). Terminal Pleistocene Alaskan genome reveals first founding population of native Americans. Nature, 553(7687), 203–207. 10.1038/nature25173 [PubMed: 29323294]

Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, … Willerslev E (2018). Early human dispersals within the Americas. Science (New York, N.Y.), 362(6419), eaav2621. 10.1126/science.aav2621

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, & Willerslev E (2017). Tracing the peopling of the world through genomics. Nature, 541(7637), 302–310. 10.1038/nature21347 [PubMed: 28102248]

Patterson N, Price AL, & Reich D (2006). Population structure and eigenanalysis. PLoS Genetics, 2(12), e190. 10.1371/journal.pgen.0020190 [PubMed: 17194218]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 2825–2830 https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Peter BM (2022). A geometric relationship of F2, F3 and F4-statistics with principal component analysis. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 377(1852), 20200413. 10.1098/rstb.2020.0413 [PubMed: 35430884]

Pickrell JK, & Reich D (2014). Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics: TIG, 30(9), 377–389. 10.1016/j.tig.2014.07.007 [PubMed: 25168683]

Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, … Reich D (2018). Reconstructing the deep population history of central and South America. Cell, 175(5), 1185–1197. e22. 10.1016/j.cell.2018.10.027 [PubMed: 30415837]

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, … Willerslev E (2015). Genomic evidence for the Pleistocene and recent population history of native Americans. Science (New York, N.Y.), 349(6250), aab3884. 10.1126/science.aab3884

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, & Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America, 102(44), 15942–15947. 10.1073/pnas.0507611102 [PubMed: 16243969]

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr., … Willerslev E (2014). The genome of a late Pleistocene human from a Clovis burial site in western Montana. Nature, 506(7487), 225–229. 10.1038/nature13025 [PubMed: 24522598]

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, … Willerslev E (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature, 463(7282), 757–762. 10.1038/nature08835 [PubMed: 20148029]

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, … Willerslev E (2015). The ancestry and affiliations of Kennewick man. Nature, 523(7561), 455–458. 10.1038/nature14625 [PubMed: 26087396]

Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, … Kivisild T (2018). Ancient human parallel lineages within North America contributed to a coastal expansion. Science (New York, N.Y.), 360(6392), 1024–1027. 10.1126/science.aar6851 [PubMed: 29853687]

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, … Reich D (2015). Genetic evidence for two founding populations of the Americas. Nature, 525(7567), 104–108. 10.1038/nature14895 [PubMed: 26196601]

Skoglund P, & Mathieson I (2018). Ancient genomics of modern humans: The first decade. Annual Review of Genomics and Human Genetics, 19, 381–404. 10.1146/annurev-genom-083117-021749

Skoglund P, & Reich D (2016). A genomic view of the peopling of the Americas. Current Opinion in Genetics & Development, 41, 27–35. 10.1016/j.gde.2016.06.016 [PubMed: 27507099]

Star B, & Spencer HG (2013). Effects of genetic drift and gene flow on the selective maintenance of genetic variation. Genetics, 194(1), 235–244. 10.1534/genetics.113.149781 [PubMed: 23457235]

Waskom M (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. 10.21105/joss.03021

Waters MR (2019). Late Pleistocene exploration and settlement of the Americas by modern humans. Science (New York, N.Y.), 365(6449), 1–9. 10.1126/science.aat5447

Willerslev E, & Meltzer DJ (2021). Peopling of the Americas as inferred from ancient genomics. Nature, 594(7863), 356–364. 10.1038/s41586-021-03499-y [PubMed: 34135521]

Yuan Y, Wu L, & Zhang X (2021). Gini-impurity index analysis. IEEE Transactions on Information Forensics and Security, 16, 3154–3169. 10.1109/tifs.2021.3076932

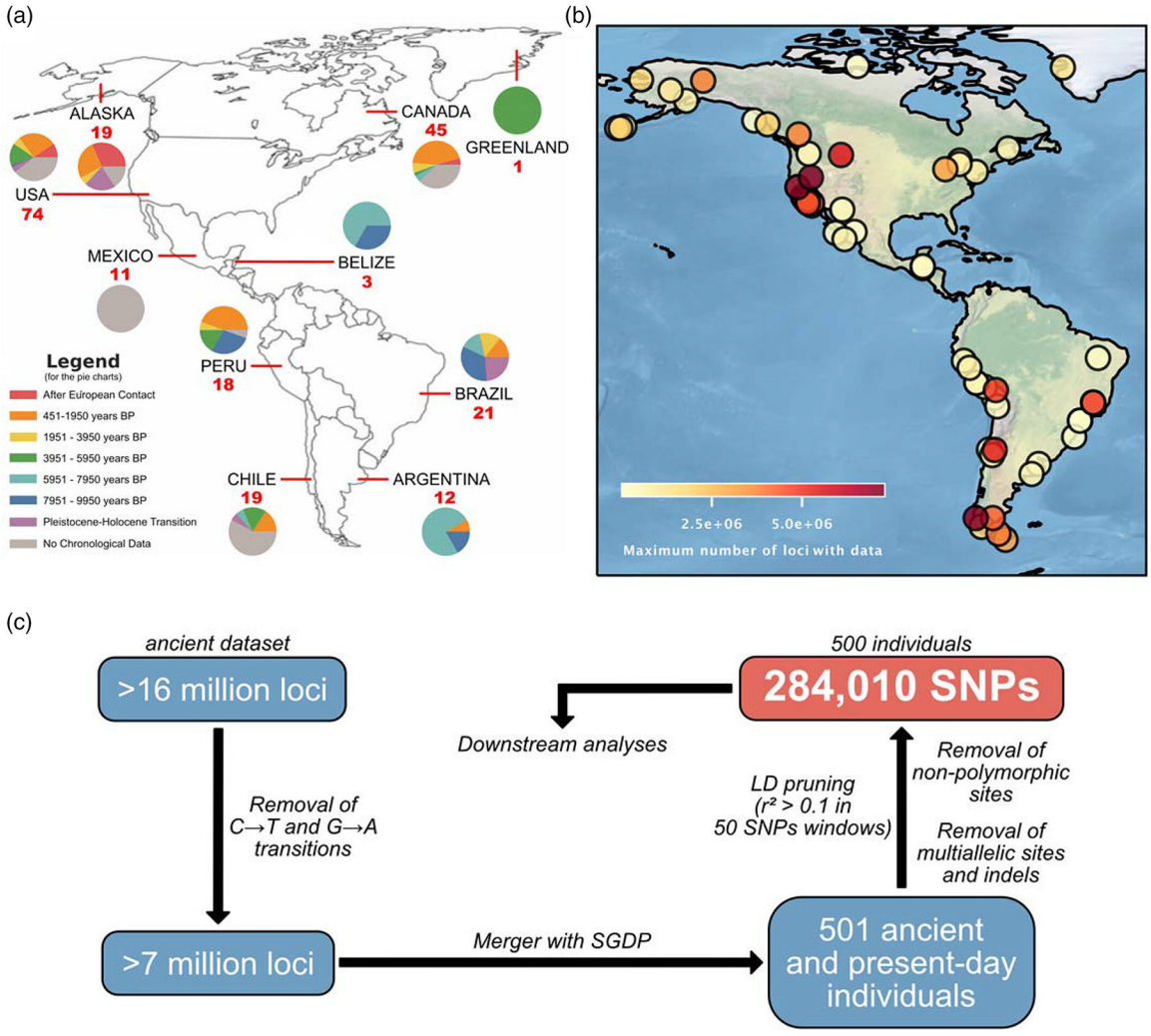

FIGURE 1.

Ancient samples information and dataset preparation. (a) Number of analyzed ancient samples by country and estimated age. (b) Heatmap of the maximum number of genomic loci assayed by each archeological site. (c) Flowchart detailing dataset preparation steps for downstream analyses applied in this article

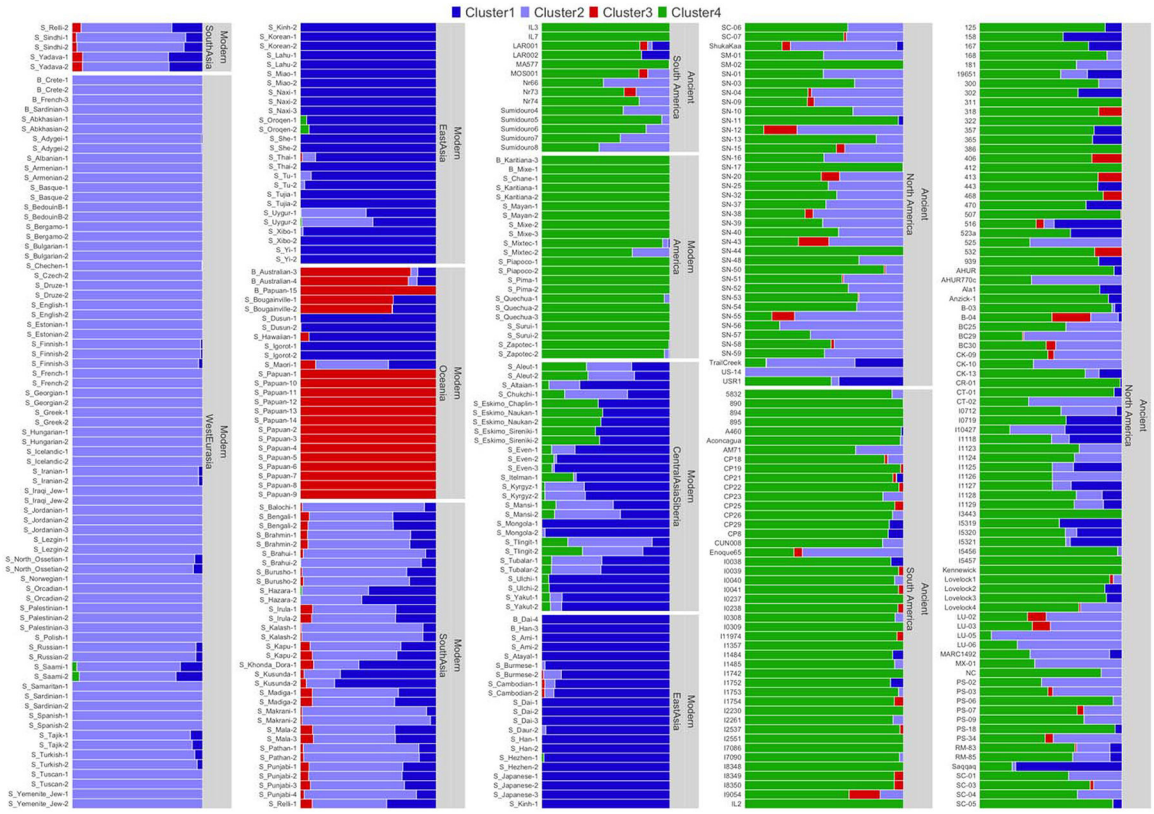

FIGURE 2.

ADMIXTURE analysis for $K = 4$ clusters. The predominant cluster (Cluster4 in green) in the ancient samples is also the representative ancestry of present-day native Americans. Cluster2 (light blue) and Cluster1 (dark blue), respectively the second and third most frequently observed components in the ancient individuals, are shared worldwide. Horizontal bars represent individuals, bar color denotes a distinct ancestral component or cluster, and bar width corresponds to the proportion of a specific component comprising a given individual

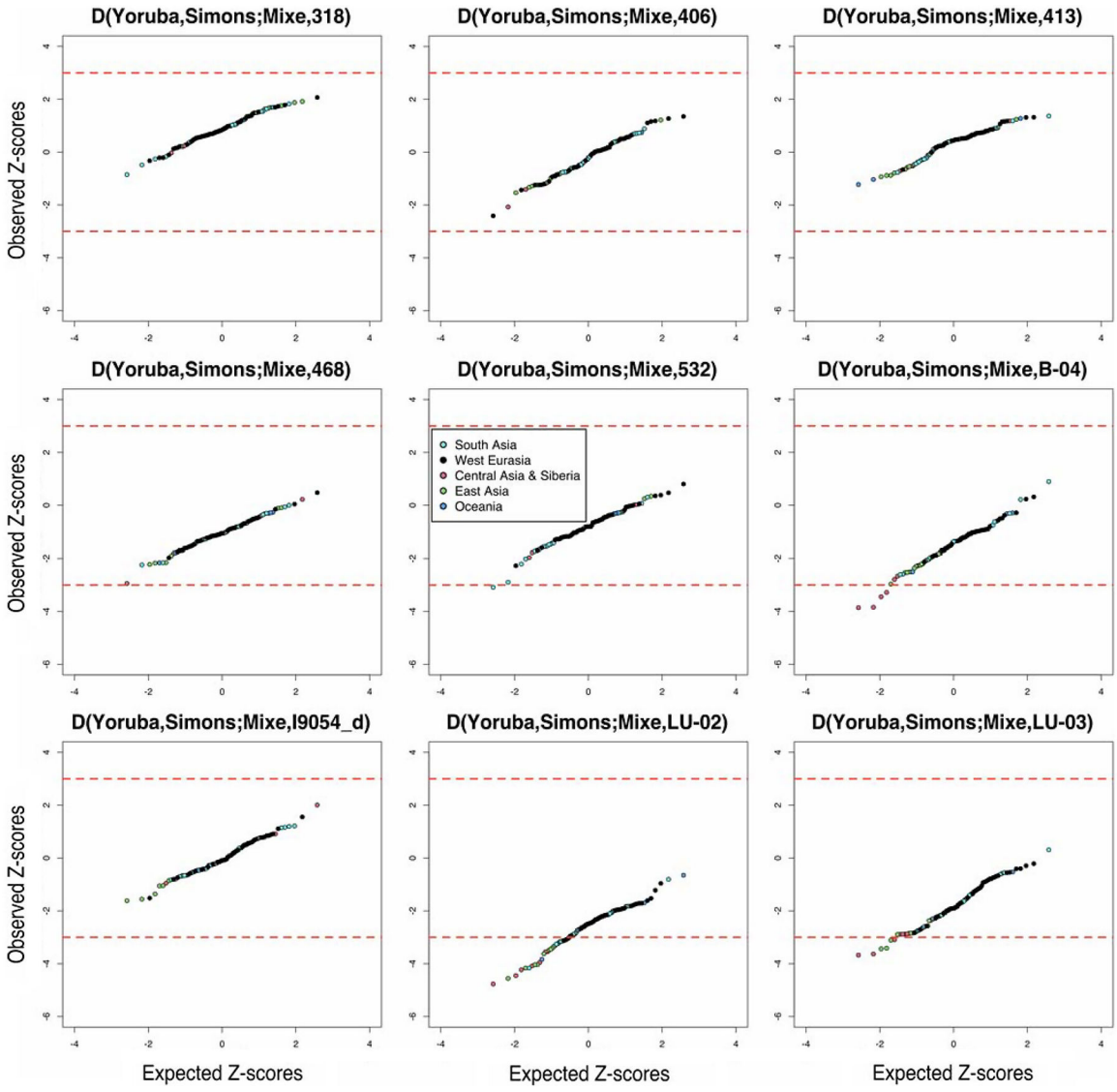

FIGURE 3.

Quantile–quantile plots of the *Z*-scores for the *D*-statistic tests of 9 of the 13 ancient individuals tested. Each point represents a distinct population from the SGDP. Horizontal dashed red lines represent significance thresholds. We see no excess affinity between the present-day populations and ancient samples in comparison to the Mixe ($Z > 3$). See Figure S6 for the remaining four ancient individuals tested

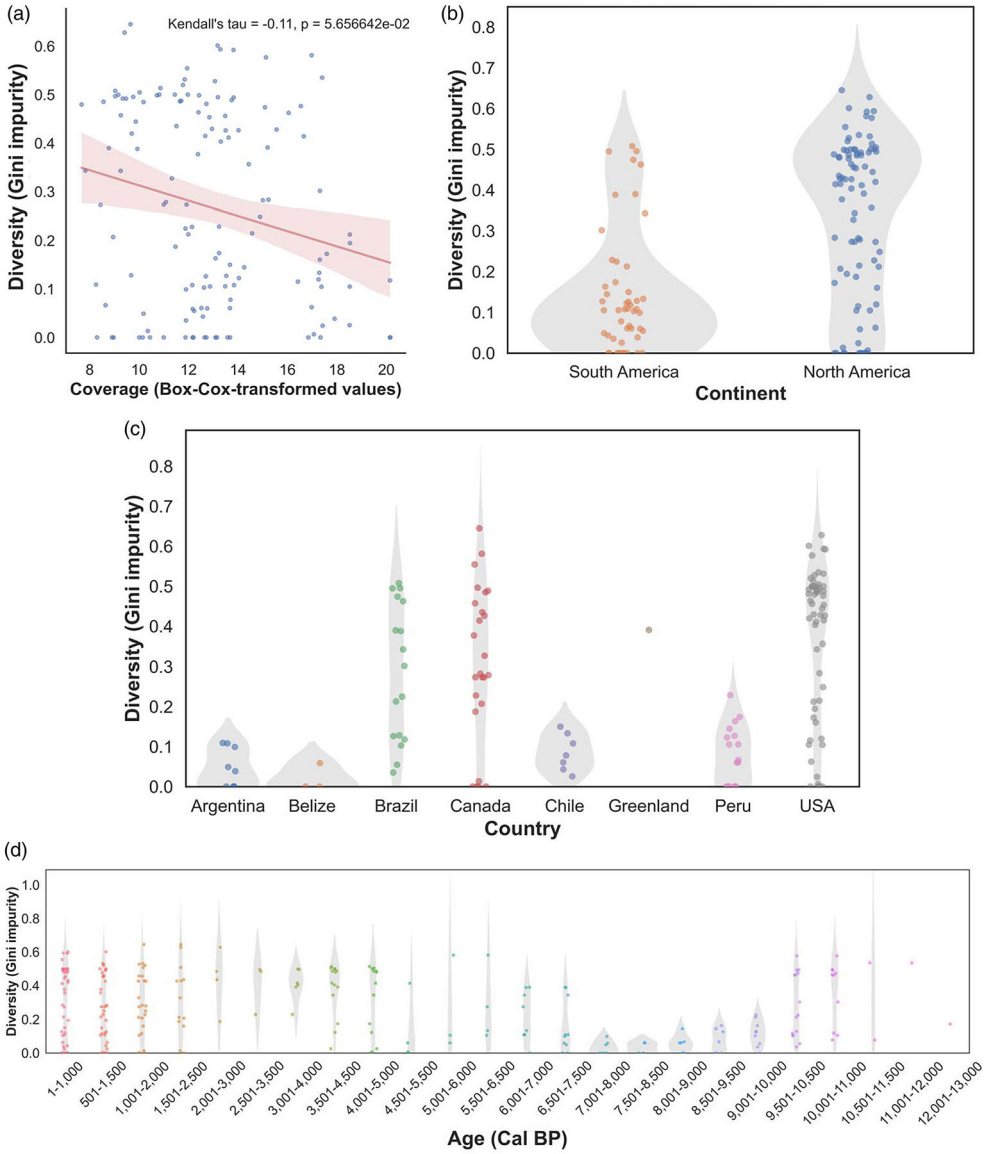

FIGURE 4.

Population structure diversity of the ancient individuals of the Americas across space and time. (a) Scatter plot depicting that ancestral component diversity in a cohort is unaffected by their genomic coverages. (b–d) violin plots showing the distribution of ancestral component diversity across different geographies within the Americas, across various countries of the Americas, and over time, respectively. Cal BP means calibrated age before present. These plots (a–d) only include the 137 ancient samples with estimated age that were analyzed in the final ADMIXTURE runs

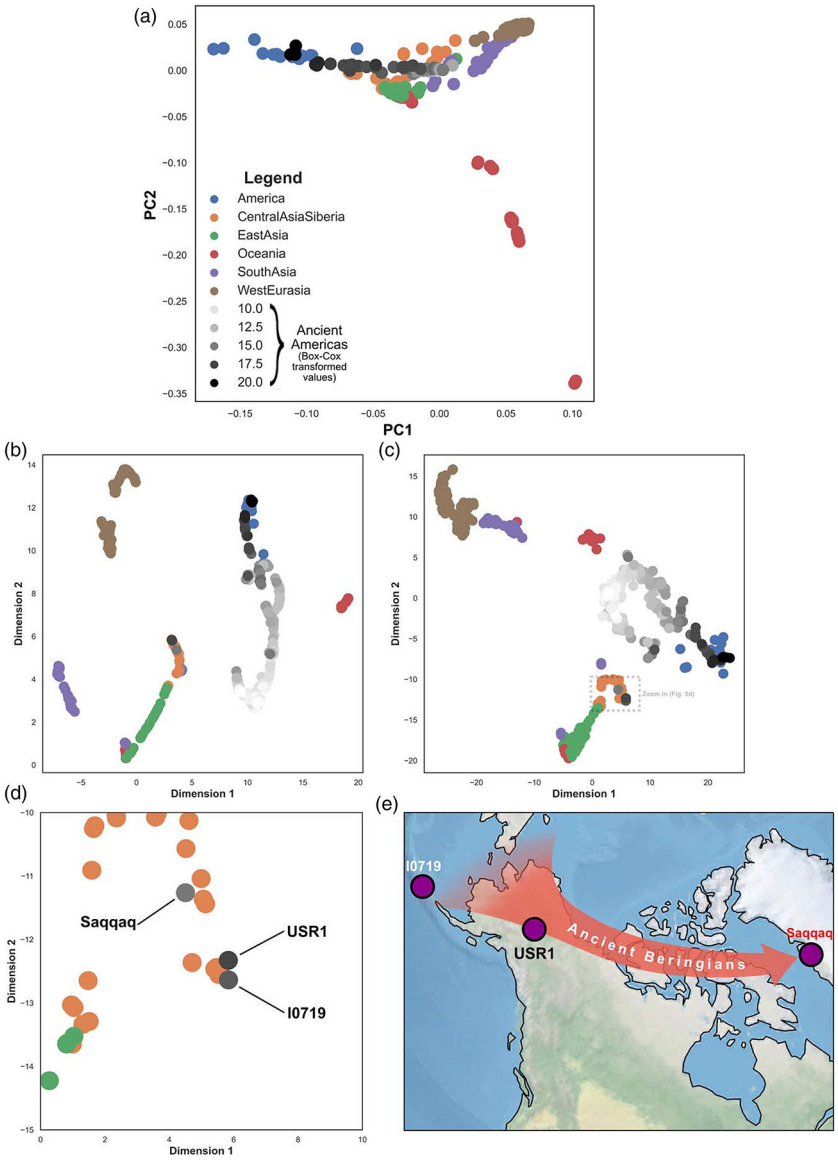

FIGURE 5.

Genomic relationships of the ancient individuals of the Americas. (a) PCA results of PC1 versus PC2 showing that the ancient individuals fall between present-day populations from the Americas and Central Asia/Siberia. The gray scale is proportionate to the level of genomic coverage in the ancient dataset. (b and c) Further dimensionality reduction with UMAP and t-SNE, respectively, using 10 PCs as input. Ancient individuals with lower coverage tend to fall at some distance from present-day individuals from the Americas, with a few exceptions.(d) Zoomed in t-SNE results showing specific ancient individuals of relatively high coverage clustering with present-day individuals from Central Asia/Siberia, putatively representing an ancient Beringian ancestry. (e) Hypothesis for dispersion of ancient Beringians in North America.