Origins and genetic legacies of the Caribbean Taino

Schroeder, Hannes; Sikora, Martin; Gopalakrishnan, Shyam; Cassidy, Lara M.; Delserc, Pierpaolo Maisano; Velasco, Marcela Sandoval; Schraiber, Joshua G.; Rasmussen, Simon; Homburger, Julian R.; Avila-Arcos, Maria C.; Allentoft, Morten E.; Moreno-Mayar, J. Victor; Renaud, Gabriel; Gomez-Carballa, Alberto; Laffoon, Jason E.; Hopkins, Rachel J. A.; Higham, Thomas F. G.; Carr, Robert S.; Schaffer, William C.; Day, Jane S.; Hoogland, Menno; Salas, Antonio; Bustamante, Carlos D.; Nielsen, Rasmus; Bradley, Daniel G.; Hofman, Corinne L.; Willerslev, Eske

Published in: Proceedings of the National Academy of Sciences of the United States of America

Link to article, DOI: 10.1073/pnas.1716839115

Publication date: 2018

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):

DTU Library
Technical Information Center of Denmark

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
The Caribbean was one of the last parts of the Americas to be settled by humans, but how and when the islands were first occupied remains a matter of debate. Ancient DNA can help answering these questions, but the work has been hampered by poor DNA preservation. We report the genome sequence of a 1,000-year-old Lucayan Taino individual recovered from the site of Preacher’s Cave in the Bahamas. We sequenced her genome to 12.4-fold coverage and show that she is genetically most closely related to present-day Arawakan speakers from northern South America, suggesting that the ancestors of the Lucayans originated there. Further, we find no evidence for recent inbreeding or isolation in the ancient genome, suggesting that the Lucayans had a relatively large effective population size. Finally, we show that the native American components in some present-day Caribbean genomes are closely related to the ancient Taino, demonstrating an element of continuity between precontact populations and present-day Latino populations in the Caribbean.

ancient DNA studies in the Caribbean have been hampered by poor preservation (8), and the few studies that exist are limited to mitochondrial DNA and, therefore, lack in resolution (9–11).

Results and Discussion

Here, we report the genome sequence of a Lucayan Taino individual who lived in the Bahamas ~500 y before European contact. We sequenced the genome to an average depth of 12.4-fold, using ancient DNA studies in the Caribbean have been hampered by poor preservation (8), and the few studies that exist are limited to mitochondrial DNA and, therefore, lack in resolution (9–11).

Significance

Ancient DNA has revolutionized the field of archaeology, but in the Caribbean and other tropical regions of the world, the work has been hampered by poor DNA preservation. We present an ancient human genome from the Caribbean and use it to shed light on the early peopling of the islands. We demonstrate that the ancestors of the so-called “Taino” who inhabited large parts of the Caribbean in pre-Columbian times originated in northern South America, and we find evidence that they had a comparatively large effective population size. We also show that the native components in some modern Caribbean genomes are closely related to the ancient Taino, suggesting that indigenous ancestry in the region has survived through the present day.
whole-genome enrichment and high-throughput sequencing. The sequence was obtained from a tooth excavated at the site of Preacher’s Cave, which is located on the island of Eleuthera in the Bahamas (SI Appendix, Fig. S1). The tooth was directly dated to 1,882 ± 29 14C yr BP (cal AD 776–992) (SI Appendix, section 3), and strontium isotope analysis suggests that the individual grew up locally in the Bahamas (SI Appendix, section 4). All DNA libraries displayed features typical for ancient DNA, including short average fragment lengths, characteristic fragmentation patterns, and an increased frequency of apparent C-to-T substitutions at the 5′ end of DNA molecules (SI Appendix, section 8). Contamination was estimated to be around 0.1−1.2% (SI Appendix, section 9), which is within the normal range observed for other ancient genomes (13–15) and unlikely to affect downstream analyses (16).

Chromosomal Sex and Mitochondrial DNA. We determined the sex of the individual to be female, based on the number of reads mapping to the X and Y chromosomes, respectively (SI Appendix, section 10). The mitochondrial genome was sequenced to an average depth of ~167× and was placed at the root of Native American haplogroup B2 (SI Appendix, section 11). As one of the founding lineages of the Americas, B2 has a pan-American distribution among present-day Native Americans (17), although our analysis suggests that it occurs at higher frequency among South Americans (SI Appendix, Fig. S9). A close search of the literature on modern published mtDNAs from the Caribbean (7, 18–23) revealed no matches or closely related sequences (SI Appendix, section 11). Generally speaking, the B2 lineage appears to be quite rare in Caribbean populations today and, interestingly, it has not been previously detected in ancient populations from the region (9–11). It is possible, therefore, that haplotype B2 was relatively rare in the Caribbean in the past. Alternatively, it may have been lost during the dramatic population declines experienced by Caribbean populations after 1492 (5).

Genome-Wide Affinities. To assess the genome-wide affinities of the ancient Taino, we computed outgroup f3-statistics of the form f3(Yoruba; Taino, X), where X is one of 50 Native American groups from a previously published dataset (24) that we used as reference. Due to high levels of recent European and African admixture in many Native Americans, those genomic segments were excluded before analysis (SI Appendix, section 12). We find that the ancient Taino is most closely related to the Palikur and other Arawakan speakers from the Amazon and Orinoco basins (Fig. L4). We observe similar affinities using D-statistics (Fig. 1B), principal component analysis (SI Appendix, Fig. S10), and a neighbor-joining tree based on pairwise FST distances, which places the Taino on the same branch as other Arawakan speakers (Fig. 1C). These results are further supported by ADMIXTURE (25) results, which show that the Taino has ancestry proportions similar to those of the Palikur and other Arawakan speakers (SI Appendix, Fig. S11).

To further explore the ancestry of the ancient Taino, we used the haplotype-based approach implemented in ChromoPainter (26). By leveraging linkage information, haplotype-based approaches are more powerful in detecting fine-scale structure than those using unlinked loci. To avoid the confounding effects of missing data, we ran ChromoPainter (26) on the unmasked dataset. As expected, we observe the highest levels of shared haplotypes between the Taino and Arawakan speakers, which strikingly provide all of the top hits in the analysis, as shown in Fig. 1D. Interestingly, this includes admixed groups, such as the Wayuu, who were not picked up in the SNP-based analyses, probably as a result of additional gene flow from the Isthmo-Colombian area, which can be seen as the light blue component in the ADMIXTURE result (SI Appendix, Fig. S11).

We also specifically looked for traces of Australasian ancestry in the ancient Taino genome, since previous studies (27) have found surprising affinities between some Amazonian populations (e.g., Surui) and populations from Melanesia, Australia, and the

Fig. 1. The genetic origins of the Taino. The individual from Preacher’s Cave is most closely related to Arawakan and Cariban speakers from the Amazon and Orinoco basins. (A) Heat map of outgroup f3-statistics testing (Yoruba; Taino, X) where X is one of 50 Native American populations (24). Warmer colors indicate higher levels of allele sharing. (B) We computed D-statistics of the form D (Yoruba, Taino; Palikur, X) to test if any other group is more closely related to the Taino than the Palikur. Thick and thin whiskers represent 1 and 3 SEs, respectively. (C) Neighbor-joining tree based on FST distances. (D) Expected total length (cm) of shared haplotypes between the Taino and 50 Native American groups based on ChromoPainter analysis (26). To avoid the confounding effects of missing data, we ran ChromoPainter (26) on the unmasked dataset. Horizontal bars mark mean values ± SD. For language classification, see SI Appendix, section 1.
and demonstrate
Previous studies (5, 16) show a model that is a good fit to the data in the sense that

the Clovis genome (13), against a backdrop of 53 modern Native American and Siberian genomes (15, 31, 32). As previously observed (29, 30), all Native American genomes, including the Taino, show clear evidence for having undergone one or more ancestral population bottlenecks, as indicated by the excess of shorter (<2 Mb) ROH. This is consistent with the proposed occurrence of an extreme founder event on entry to the continent, followed by successive bottlenecks (33, 34). Interestingly, the Clovis genome (13) (~12,600 BP) appears to provide a snapshot of one such early bottleneck. The individual does not share the same excess of shorter runs seen in modern Native Americans, but instead exhibits inflated ROH coverage between 2 and 8 Mb. The relatively low level of shorter ROH could argue against an extremely long or intense Beringian Incubation Model, which states that the people who eventually colonized the Americas descended from a small population that spent up to 15,000 y isolated on the Bering Land Bridge before entering the Americas (35).

At the other end of the spectrum, the Taino genome displays some of the lowest levels of longer (>8 Mb) ROH of any Native American genome (Fig. 2). This argues against a history of recent isolation or inbreeding in the Lucayan population and suggests that the Lucayans had a relatively large effective population size. Based on the distribution of longer ROH (≥1.6 Mb), we estimate an effective size of around 1,600 individuals, which is considerably higher than our estimates for some present-day South American populations, such as the Karitiana and Surui (SI Appendix, Table S13). However, the island of Eleuthera measures only around 518 km², and it is difficult to imagine how this community was able to sustain such a relatively large effective size without outside contact. Current thinking suggests that Caribbean communities were highly mobile and maintained pan-regional networks that extended far beyond the local scale (36, 37). Our results are consistent with this view. Evidently, these networks did not only involve the exchange of goods and ideas, as evidenced by archaeology, but also of genes. With the arrival of Europeans, however, these networks were disrupted, which may have contributed to the catastrophic population declines suffered by Caribbean communities soon after contact (38).

Genetic Legacies. Previous studies (5–7) have shown that the amount of Native American ancestry in modern Caribbean populations varies widely across the region. While some retain substantial amounts of Native American ancestry (only the Lucayan and a small number of other native American groups), others are largely composed of African and/or European ancestry (5–7). Puerto Ricans, for example, harbor between 10 and 15% Native American ancestry; however, it is unclear to what extent this component reflects Taino ancestry. To address this issue, we added 104 modern Puerto Rican genomes from the 1000 Genomes Project (39) to our dataset and used the clustering algorithm ADMIXTURE (25) to estimate the composition of genetic ancestry in each individual (SI Appendix, Fig. S13). Due to the high levels of African and European ancestry in modern Puerto Ricans, the native components are difficult to discern; however, when we compare only the estimated ancestry clusters that reflect non-African/European ancestries, there are clear similarities between Puerto Ricans, Arawakan speakers, and the ancient Taino (SI Appendix, Figs. S14 and S15).

To explore these relationships further, we then masked segments of African and European ancestry in the Puerto Rican genomes (SI Appendix, section 12) and computed a set of out-group f-statistics to assess the amount of shared drift between Puerto Ricans, other present-day Native Americans, and the ancient Taino. The results are shown in Fig. 3A and demonstrate that Puerto Ricans share more drift with the Taino than any other native American group in our dataset. To formally test this relationship, we then computed two sets of D-statistics of the form D(YRI, Taino; PUR, X) and D(YRI, X; Taino, PUR), where X is the test population. Results are consistent with Puerto Ricans and the ancient Taino forming a clade without any significant gene-flow postdivergence (SI Appendix, Fig. S16). To test whether other present-day Latino populations in the Caribbean share the same affinities with the ancient Taino, we repeated the analyses with SNP array data for a more diverse set of Caribbean populations from Haiti, Cuba, and the Dominican Republic (5); however, due to the low amounts of Native American ancestry in these populations, we were unable to replicate the results.

Finally, we tried to fit both the ancient Taino and masked Puerto Ricans on a previously defined admixture graph (24). Fig. 3B shows a model that is a good fit to the data in the sense that none of the predicted f-statistics are more than three SEs from what is observed (max |Z| = 2.6). In this model, the ancient Taino and masked Puerto Ricans form a clade that branches off the main South American lineage. By contrast, a model where Puerto Ricans are added as direct descendants of the Taino does not fit the data (SI Appendix, Fig. S18). To determine if patterns of allele frequencies in modern Puerto Ricans and the ancient Taino individual are compatible with direct ancestry we then

Andaman Islands. Using D-statistics of the form D(Yoruba, X; Mixe, Taino) computed with the Affymetrix Human Origins SNP array data (28), we do not detect the same excess affinity in the ancient Taino (SI Appendix, Fig. S12), suggesting either that the signal was somehow lost in the Taino or that it entered Amazonian populations after the divergence from the Taino within the last 3,000 y.

**Runs of Homozygosity.** Next, we analyzed the ancient genome for runs of homozygosity (ROH) to investigate the demographic history of the Taino (SI Appendix, section 14). ROH can inform about past demography: short ROH being indicative of ancient restrictions in effective population size, while longer ROH reflect recent episodes of isolation and/or inbreeding (29, 30). Fig. 2 plots the ROH distributions for the ancient Taino genome and the Clovis genome (13), against a backdrop of 53 modern Native American and Siberian genomes (15, 31, 32). As previously observed (29, 30), all Native American genomes, including the Taino, show clear evidence for having undergone one or more ancestral population bottlenecks, as indicated by the excess of shorter (<2 Mb) ROH. This is consistent with the proposed occurrence of an extreme founder event on entry to the continent, followed by successive bottlenecks (33, 34). Interestingly, the Clovis genome (13) (~12,600 BP) appears to provide a snapshot of one such early bottleneck. The individual does not share the same excess of shorter runs seen in modern Native Americans, but instead exhibits inflated ROH coverage between 2 and 8 Mb. The relatively low level of shorter ROH could argue against an extremely long or intense Beringian Incubation Model, which states that the people who eventually colonized the Americas descended from a small population that spent up to 15,000 y isolated on the Bering Land Bridge before entering the Americas (35).

At the other end of the spectrum, the Taino genome displays some of the lowest levels of longer (>8 Mb) ROH of any Native American genome (Fig. 2). This argues against a history of recent isolation or inbreeding in the Lucayan population and suggests that the Lucayans had a relatively large effective population size. Based on the distribution of longer ROH (≥1.6 Mb), we estimate an effective size of around 1,600 individuals, which is considerably higher than our estimates for some present-day South American populations, such as the Karitiana and Surui (SI Appendix, Table S13). However, the island of Eleuthera measures only around 518 km², and it is difficult to imagine how this community was able to sustain such a relatively large effective size without outside contact. Current thinking suggests that Caribbean communities were highly mobile and maintained pan-regional networks that extended far beyond the local scale (36, 37). Our results are consistent with this view. Evidently, these networks did not only involve the exchange of goods and ideas, as evidenced by archaeology, but also of genes. With the arrival of Europeans, however, these networks were disrupted, which may have contributed to the catastrophic population declines suffered by Caribbean communities soon after contact (38).

**Genetic Legacies.** Previous studies (5–7) have shown that the amount of Native American ancestry in modern Caribbean populations varies widely across the region. While some retain substantial amounts of Native American ancestry (only the Lucayan and a small number of other native American groups), others are largely composed of African and/or European ancestry (5–7). Puerto Ricans, for example, harbor between 10 and 15% Native American ancestry; however, it is unclear to what extent this component reflects Taino ancestry. To address this issue, we added 104 modern Puerto Rican genomes from the 1000 Genomes Project (39) to our dataset and used the clustering algorithm ADMIXTURE (25) to estimate the composition of genetic ancestry in each individual (SI Appendix, Fig. S13). Due to the high levels of African and European ancestry in modern Puerto Ricans, the native components are difficult to discern; however, when we compare only the estimated ancestry clusters that reflect non-African/European ancestries, there are clear similarities between Puerto Ricans, Arawakan speakers, and the ancient Taino (SI Appendix, Figs. S14 and S15).

To explore these relationships further, we then masked segments of African and European ancestry in the Puerto Rican genomes (SI Appendix, section 12) and computed a set of out-group f-statistics to assess the amount of shared drift between Puerto Ricans, other present-day Native Americans, and the ancient Taino. The results are shown in Fig. 3A and demonstrate that Puerto Ricans share more drift with the Taino than any other native American group in our dataset. To formally test this relationship, we then computed two sets of D-statistics of the form D(YRI, Taino; PUR, X) and D(YRI, X; Taino, PUR), where X is the test population. Results are consistent with Puerto Ricans and the ancient Taino forming a clade without any significant gene-flow postdivergence (SI Appendix, Fig. S16). To test whether other present-day Latino populations in the Caribbean share the same affinities with the ancient Taino, we repeated the analyses with SNP array data for a more diverse set of Caribbean populations from Haiti, Cuba, and the Dominican Republic (5); however, due to the low amounts of Native American ancestry in these populations, we were unable to replicate the results.

Finally, we tried to fit both the ancient Taino and masked Puerto Ricans on a previously defined admixture graph (24). Fig. 3B shows a model that is a good fit to the data in the sense that none of the predicted f-statistics are more than three SEs from what is observed (max |Z| = 2.6). In this model, the ancient Taino and masked Puerto Ricans form a clade that branches off the main South American lineage. By contrast, a model where Puerto Ricans are added as direct descendants of the Taino does not fit the data (SI Appendix, Fig. S18). To determine if patterns of allele frequencies in modern Puerto Ricans and the ancient Taino individual are compatible with direct ancestry we then
used a recently developed likelihood ratio test (40). While the test rejects the hypothesis of direct ancestry, it also shows that the ancient Taino only recently diverged from the ancestors of modern Puerto Ricans (SI Appendix, Table S15). This result is mirrored in the ChromoPainter analysis (26), which shows that Puerto Ricans share large parts of their genomes with the ancient Taino, despite significant European and African admixture (SI Appendix, Fig. S19).

Conclusion

Our study provides a glimpse of the initial peopling of the Caribbean from an ancient genome perspective. Specifically, we were able to show that the Lucayan Taino were genetically most closely related to present-day Arawakan speakers from northern South America, suggesting that their ancestors originated there. However, we note that this does not preclude the possibility of other/earlier migrations to the Caribbean that originated elsewhere, and more data will need to be collected to address this issue. Further, we find no evidence for recent isolation or inbreeding in the ancient genome, suggesting that the Lucayans had a comparatively large effective population size despite their island location. This is consistent with archaeological evidence, which suggests that indigenous Caribbean communities were highly mobile and maintained complex regional networks of interaction and exchange that extended far beyond the local scale. Lastly, we find that the native component in present-day Puerto Rican genomes is closely related to the ancient Taino, demonstrating an element of continuity between precontact populations and present-day Latino populations in the Caribbean despite the disruptive effects of European colonization.

Materials and Methods

Samples. The samples for this study were excavated at the site of Preacher’s Cave, which is located on the northern part of the island of Eleuthera in the Bahamas (SI Appendix, section 2). During excavations in 2007 (SI Appendix, section 2), a total of six Lucayan primary burials were discovered within the cave, three of which were well preserved (12). The three burials belonged to two adult males and one female, aged 20–35 y at the time of death (12). For the present study, we sampled five of the burials for isotopic and ancient DNA analysis, and in the absence of petrous bones, we opted for teeth.

Radiocarbon Dating. Radiocarbon dating was performed at the Oxford Radiocarbon Accelerator Unit (SI Appendix, section 3). The standard method for radiocarbon dating is measuring the amount of $^{14}$C in collagen from bone or dentine. However, in tropical environments, collagen is often only poorly preserved or not at all (41), and since part of the dentine was used for DNA analysis, the remaining sample was very small. Consequently, we turned to the enamel fraction. Chemical pretreatment was done as described in ref. 42 to remove labile carbonates on crystal surfaces and grain boundaries. While the procedure is still far from being standardized, it is thought to provide a reliable terminus ante quem (SI Appendix, section 3).

Isotope Analyses. We conducted multiple isotope (Sr, C, O) analyses to determine whether the individuals buried in the cave were of local or nonlocal origin. The logic behind this approach is that it cannot be reasonably argued that an individual’s ancestors were local and, thus, “representative” of a particular population if the individual was in fact not local, but a first-generation migrant, especially if the results indicate long-distance migration. This is an especially important consideration for the ancient Antilles where high rates of migration have been documented for various time periods (43). The analytical procedure is described in SI Appendix, section 4.

DNA Extraction and Library Preparation. DNA was extracted from ~100 mg of starting material (SI Appendix, section 5). Thirty microliters of each DNA extract was then built into DNA libraries using Illumina specific adapters (44). Ten microliters of the DNA libraries were then amplified and indexed in 50–μL PCRs using sample-specific barcodes, as described in ref. 45. The optimal number of PCR cycles was determined by qPCR. The amplified libraries were purified using AMPure XP beads (Beckman Coulter), quantified on a 2200 TapeStation (Agilent Technologies), pooled in equimolar amounts, and sequenced on an Illumina HiSeq 2500 run in SR mode. The results of the screening run are shown in SI Appendix, Table S4. As expected, all of the samples yielded extremely low endogenous DNA contents, except one (PCS37), which turned out to be exceptionally well preserved (SI Appendix, Table S4).

Whole-Genome Capture and Deep-Sequencing. Following the initial screening run, we built three more libraries for PCS37 and enriched them using the MyBaits Human Whole Genome Capture Kit (MYcroarray), following the manufacturer’s instructions (46). The method makes use of biotinylated RNA probes transcribed from genomic DNA libraries to capture the human DNA in the library. The captured libraries were amplified for 10–12 cycles using primers ISS and ISS (44), purified using AMPure XP beads (Beckman Coulter), quantified, and sequenced as above. After capture, the endogenous fraction increased from 13% to around 35%, albeit with some loss in complexity. We

Fig. 3. The genetic legacy of the Taino. (A) Heat map showing the amount of allele sharing between the Native American component in present-day Puerto Ricans, Native Americans, and the Taino. Warmer colors indicate higher levels of allele sharing. (B) Model of Native American population history that fits the patterns of observed allele frequencies in our dataset (max$Z = 2.6$). The Taino and masked Puerto Ricans form a clade that branches off the South American lineage. Branches are colored by language family (SI Appendix, section 1). Drift values are shown in units proportional to $F_{ST} \times 1,000$. 

Latitude

Longitude

Chorotega

Teribe

Huetar

Mixtec

Purepecha

Yaqui

Ojibwa

Aymara

Karitiana

Quechua

Chile

Diaguita

Yoruba

Ojiwa

Santo Domingo

Colima

Valle del Cauca

Canoe

Coclé

Huila

San Luis

Gualábaco

Tambogrande

Buenos Aires

Useda

Madrid

Buenavista

Manati

Montego Bay

San Juan

Barcelona

Chiriquí

Pedro de Heredia

Acapulco

Arecibo

Guanabacoa

San Pedro de Macoris

Panama City

Valle del Cauca

San Juan

Barcelona

Chiriquí

Pedro de Heredia

Acapulco

Arecibo

Guanabacoa

San Pedro de Macoris

Panama City
then sequenced the ancient genome to an average depth of 12.4-fold using a combination of shotgun and captured libraries.

Mapping. Basecalling was done with CASAVA-1.8.2. Only reads with correct indexes were kept. FASTQ files were filtered using AdapterRemoval (47) to remove adapter sequences, low quality stretches, and ambiguous bases at the ends of reads. The minimum length allowed after trimming was 25 nucleotides. Trimmed and filtered reads were mapped to GRCh37/hg19 (build 37.1) using bwa-0.7.5 (48), with the seed disabled to allow for better sensitivity (49) and filtering for reads with a minimum mapping quality of 30. The mitochondrial sequence in the reference was replaced by the revised Cambridge Reference Sequence (50). Clonal reads were removed using samtools-1.2.1 (51) rmdup function, and bam files from different sequencing runs were merged using samtools-1.2.1 (51) merge.

Genotype Calling. Diploid genotypes were called using samtools-1.2.1 (51) mpileup function (-C 50 option) and bcf2tsv-1.2.4 call with the consensus caller enabled. Genotype calls were filtered for a minimum depth of one-third and a maximum depth of two times the average depth of coverage (12.4-fold). Subsequently, clustered variants were filtered out by removing SNPs/InDels that were called within 5 base pairs of each other. Variants were also filtered for a Phred posterior probability of less than 30, strand bias, or end distance bias $p < 10^{-4}$. Overall and type-specific error rates were estimated using ANGSD (52) (SI Appendix, section 7).

Genetic Affinities. To explore the genetic affinities of the Taino individual, we merged the called ancient genome with a previously published SNP-chip dataset (24), which includes 493 individuals from 50 Native American populations genotyped at 346,465 SNPs (SI Appendix, Table 57) for which segments of European and African ancestry had been masked. In addition, we included 21 Yoruba, 34 Han Chinese, and 28 French individuals from HGDP (53). Merging was done using PLINK 1.9 (54). While transition SNPs are sensitive to postmortem damage, we opted to include all SNPs, since only ~1% of the sites included in the reference panel involve transversions, corresponding to 4,856 sites. Genotypes where the ancient genome had a different allele to the ones observed in the reference panel were set to missing, resulting in a final dataset of 325,139 SNPs.

Outgroup f3- and D-statistics. Outgroup f3- and D-statistics were computed using AdmixTools (55). To estimate the amount of shared drift between the ancient Taino and present-day Native Americans, we computed outgroup f3-statistics of the form f3(Yoruba; Taino; X), where X is one of 50 Native American populations in our dataset (24) (Fig. 1A). We then computed a set of allele frequency-based D-statistics of the form D(Yoruba; Taino; Palikur), where X is the test population, to test whether any other other Native American group in our dataset is more closely related to the Taino than the Palikur (Fig. 1B).

F2 Tree. We used EIGENSOFT (102) to calculate pairwise F2 distances based on allele frequencies in our dataset. The distance matrix was then used to build a neighbor-joining tree using the APE package in R (56). The phylogenetic tree in Fig. 1C was rendered using FigTree v1.4.2 (tree.bio.ed.ac.uk/software/figtree).

ADMIXTURE Analysis. We ran ADMIXTURE analysis (25) both on the masked and unmasked datasets using default parameters for $K = 2$ to $K = 14$ and diploid genotype calls for both the ancient genome and the modern reference populations. For the masked dataset, we removed individuals with more than 60% missing genotypes and any variants with call rates of less than 40%, resulting in a final dataset of 466 individuals typed at 346,418 SNPs. The unmasked dataset includes 1,112 individuals typed at 346,465 SNPs. For both datasets, we ran 100 replicates for each $K$ and picked the one with the highest log-likelihood as result for that $K$ (SI Appendix, Figs. S11 and S13). SI Appendix, Fig. S15 shows the ancestry proportions estimated for the unmasked dataset after removing the first three ancestry components (corresponding to African, Western European, and East Asian ancestries) and normalizing the remaining ancestry clusters such that they sum to 1. SI Appendix, Fig. S15 displays the estimated ancestry proportions averaged by population/language group.

ChromoPainter Analysis. To avoid the confounding effects of missing data, ChromoPainter (26) was run on the unmasked dataset. The dataset was split by chromosomes and phased using SHAPEIT 2.x (57). For the estimation of haplotypes, the 1000 Genomes Project Phase3 dataset was used as a reference panel including 2,504 individuals from 26 populations. Hap files were converted into ChromoPainter (26) format using the “impute2chromopainter.pl” script, while recombination maps were produced with “convertcrefile.pl” (both scripts are available for download on the ChromoPainter website).

 Runs of Homozygosity. For the ROH analysis, we merged the ancient Taino genome with 109 other modern Native American, Siberian, and European genomes (96-98). The ancient Clovis genome (13) was also included. The dataset was then filtered for missingness and minor allele frequency, retaining only transversions, resulting in a final dataset of 583,623 SNPs. ROH were estimated using PLINK 1.9 (110), as described in SI Appendix, section 14.

Australasian Ancestry. To test whether the Taino genome harbored any traces of Australasian ancestry, we merged the ancient genome with the Human Origins dataset (28), which contains several other ancient genomes as well as 2,345 contemporary human populations typed at ~600,000 SNPs on the Affymetrix Human Origins array. We then computed three sets of D-statistics of the form D(Yoruba; X; Mixe, Suriu/Taino/Clovis), where X is one of a subset of 59 populations in the Human Origins dataset (28), including Australians, Onge, and Papuans. We find that the Taino and the Clovis genome do not share the same excess affinity with Australasians as the Surui (SI Appendix, Fig. S12).

Genetic Legacies. To explore the relationship between the ancient Taino and modern Caribbean populations, we added 104 modern Puerto Rican genomes from the 1000 Genomes Project (39) to our dataset and performed ADMIXTURE analysis (25) as described above (SI Appendix, Fig. S13). Outgroup f3- and D-statistics were computed using AdmixTools (55), but due to the high levels of European and African ancestry in the Puerto Rican genomes, those segments were masked before analysis (SI Appendix, section 12). The direct ancestry test was also performed on the masked data (SI Appendix, section 16), and the admixture graphs (Fig. S8 and SI Appendix, Fig. S15) were fitted by ogPGraph from the AdmixTools package (55) (SI Appendix, section 17). The ChromoPainter (26) analysis was run on the unmasked dataset (SI Appendix, Fig. S19).

ACKNOWLEDGMENTS. We thank the staff at the Danish National High-Throughput Sequencing Centre for technical support, Juan Carlos Martinez-Cruzado and Tom Gilbert for input and helpful discussions, and Jorge Estevez from the Union Higuayagua for providing guidance and sharing his insights into the story of his people. We acknowledge the Irish Centre for High-End Computing for access to computational facilities and support, and we are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (BioInfo Genotoul) for providing additional computing resources. The research was funded by the European Research Council through the Seventh Framework Programme under Grant Agreement 319209 (ERC Synergy Project NEXUS1492), the Lundbeck Foundation, the Danish National Research Foundation, and the KU2016 initiative, with additional support from the HERA Joint Research Programme “Uses of the Past” (Citigen), and the European Union’s Horizon 2020 research and innovation programme under Grant Agreement 649307. L.M.C. was funded by the Irish Research Council Government Scholarship Scheme GO/3/2013/1219. J.G.S. received funding by US National Institutes of Health Grant R35-GM124745.


