Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean

Hannes Schroeder,a,b,1,2 Maria C. Ávila-Arcos,a,c,1 Anna-Sapfo Malaspinas,a G. David Poznikd, Marcela Sandoval-Velasco,a Meredith L. Carpenter,a,b José Victor Moreno-Mayord, Martin Sikora,a,c Philip L. F. Johnso, Morten Erik Allentoft,b José Alfredo Samaniego,a Bay H. Haversib, Michael W. Deeg, Thomas W. Stafford Jr., Antonio Salasa, Ludovic Orlando,a Eske Willerslev,a Carlos D. Bustamantea, and M. Thomas P. Gilberdb,2

aCentre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen, Denmark; bFaculty of Archaeology, Leiden University, 2300 Leiden, The Netherlands; cDepartment of Genetics, Stanford University, Stanford, CA 94305; dProgram in Biomedical Informatics and Department of Statistics, Stanford University, Stanford, CA 94305; eDepartment of Biology, Emory University, Atlanta, GA 30322; fSt. Maarten Archaeological Center, Philipsburg, Saint Martin; gResearch Laboratory for Archaeology and the History of Art, University of Oxford, OX1 3QY Oxford, United Kingdom; hAMSS 14C Dating Centre, Department of Physics and Astronomy, Aarhus University, 8000 Aarhus, Denmark; and iUnidade de Xenética, Departamento de Anatomía Patológica e Ciencias Forenses, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, 15872 Galicia, Spain

Edited by Rick A. Kittles, University of Arizona, Tucson, AZ, and accepted by the Editorial Board February 2, 2015 (received for review November 17, 2014)

Between 1500 and 1850, more than 12 million enslaved Africans were transported to the New World. The vast majority were shipped from West and West-Central Africa, but their precise origins are largely unknown. We used genome-wide data to trace the origins of three enslaved Africans whose remains were recovered on the Caribbean island of Saint Martin. We trace their origins to distinct subcontinental source populations within Africa, including Bantu-speaking groups from northern Cameroon and non-Bantu speakers living in present-day Nigeria and Ghana. To our knowledge, these findings provide the first direct evidence for the ethnic origins of enslaved Africans, at a time for which historical records are scarce, and demonstrate that genomic data provide another type of record that can shed new light on long-standing historical questions.

Results

Initial shotgun sequencing revealed that the DNA in the samples was very poorly preserved (SI Appendix, Section 8). This result was expected because DNA preservation in the Caribbean is known to be poor (7). The fraction of nonredundant reads mapping to the human reference genome varied between 0.3% and 7.6%, and the sequences showed all features typical of ancient DNA, including short average read lengths (~67 bp), characteristic fragmentation patterns, and an increased frequency of apparent C-to-T substitutions toward the 5′ ends of molecules (see Table 1 and SI Appendix, Fig. S7). To increase sequencing efficiency and lower costs, we enriched the ancient

Significance

The transatlantic slave trade resulted in the forced movement of over 12 million Africans to the Americas. Although many coastal shipping points are known, they do not necessarily reflect the slaves’ actual ethnic or geographic origins. We obtained genome-wide data from 17th-century remains of three enslaved individuals who died on the Caribbean island of Saint Martin and use them to identify their genetic origins in Africa, with far greater precision than previously thought possible. The study demonstrates that genomic data can be used to trace the genetic ancestry of long-dead individuals, a finding that has important implications for archeology, especially in cases where historical information is missing.
DNA libraries using two recently developed whole-genome capture methods (8, 9). Following enrichment, we generated between 0.1- and 0.5-fold genome-wide coverage for the three individuals, which proved to be sufficient to infer their likely origins within Africa.

Although the presence of characteristic damage patterns and short average read lengths suggest that authentic ancient molecules were sequenced, it is possible that some degree of modern contamination could be present in the data. Therefore, we used a previously published Bayesian method (10) to detect contamination and found very low (i.e., <1%) levels overall (see Table 1 and SI Appendix, Section 9). We then merged our ancient sequence data with genotype data from the Human Genome Diversity Cell Line Panel (HGDP) reference panel (11) and used principal component analysis (PCA) (12) to confirm the individuals’ African ancestry (for more details, see SI Appendix, Section 12). Because of low depth, we randomly sampled a single allele from both ancient and modern individuals, as done in ref. 13. As expected, all three individuals fell within African variation as defined by PC1 and PC2 (SI Appendix, Fig. S17), indicating that the signal was not driven by modern DNA contamination.

D-Statistic Test. We then tested the relationships between each sample (henceforth referred to as STM1, STM2, and STM3) and 11 populations from across the world for which whole genome data were available (15), using a D-test of the form \( D(\text{chimpanzee, STM}; \text{Yoruba, X}) \), where X stands for a population other than Yoruba (for more details, see SI Appendix, Section 13). We found that the STMs were significantly more closely related to the Yoruba than to any non-African population (Fig. 1A), again confirming their African origins. Within Africa, the STMs appeared significantly more closely related to the Yoruba than to hunter-gatherer populations (San, Mbuti Pygmies). This was not surprising, as the San and the Mbuti were not represented in the transatlantic slave trade. For the Mandenka and Dinka, the D-test results were not significant, suggesting that these populations are equally closely related to the STMs as are the Yoruba. The lack of rejection for the Dinka was surprising, as this population—from southern Sudan—is not known to have been involved in the Atlantic slave trade (16).

Principal Component Analysis. To refine our assignment within Africa, we compared our samples to another reference panel, consisting of genotype data from 11 West African populations (Fig. 1B) (17). We intersected 294,651 sites from this reference panel with our sequence data (SI Appendix, Section 14) and conducted PCA to determine whether the individuals showed close affinity to a particular population within the panel. For each of the three individuals, we merged the sequence data with the reference panel genotypes and calculated PC1 and PC2 based on the overlapping sites (SI Appendix, Fig. S18). We then combined the three analyses using Procrustes transformation, as done in ref. 13. Interestingly, the samples clustered with different populations: Bantu-speaking groups in the case of STM1 (specifically, Bamoun) and non-Bantu–speaking groups for STM2 and STM3 (Fig. 1C). We observed similar patterns using the probabilistic model of population splits and divergence implemented in TreeMix (18) (SI Appendix, Fig. S20).
To further explore the genetic ancestry of—

ADMITXURE Analysis. To further explore the genetic ancestry of the STMs, we used the maximum-likelihood-based clustering algorithm ADMIXTURE (19). When assuming three ancestral populations (K = 3), the clusters in the reference panel mirror the grouping of individuals in the space defined by PC1 and PC2: a cluster predominating in Bantu-speaking populations, a cluster for non-Bantu West African populations, and a third restricted mostly to Kuba, Mada, and Bulala (Fig. 1D). The distribution of these components in our samples indicates that STM1 has a higher proportion of Bantu-specific ancestry, whereas STM2 and STM3 carry higher proportions of the component prevalent among the non-Bantu–speaking Yoruba, Brong, and Igbo. Notably, STM2 also shows a slightly higher proportion of the component prevalent among the Kuba, Mada, and Bulala, perhaps suggesting closer affinity with Chadic or Sudanic speakers (Fig. 1D).

Uniparental Markers. Furthermore, we determined the individuals’ mitochondrial DNA (mtDNA) haplogroups and the Y-chromosome haplogroup of STM1 (SI Appendix, Section 11). The mtDNAs were assigned to haplogroups L3b1a, L3d1b2, and L2a1f, respectively. Tracing these lineages to particular regions in Africa is challenging because of their pan-continental distribution, which is the result of thousands of years of population movements (e.g., the Bantu migrations) and continued gene flow (20–22). Nevertheless, we note that haplogroup L3b1a is one of the most common lineages found in the Lake Chad Basin (23). This finding is noteworthy, because the Y-chromosome lineage of this individual (STM1) was identified as belonging to haplogroup R1b1c-V88, which—although quite rare in Africa on the whole—occurs at extremely high frequency in the Lake Chad Basin, rising to 95% in one population of northern Cameroon (24).

Discussion

Taken together, the genetic data suggest that STM1 may have originated among Bantu-speaking groups in northern Cameroon, whereas STM2 and STM3 more likely originated among non-Bantu speakers living in present-day Nigeria and Ghana. To our knowledge, these findings provide the first direct evidence for the ethnic origins of enslaved Africans, with the important caveat that the modern reference populations might not be the same as the historical populations who lived in the same locations at the time of the Atlantic slave trade. Nevertheless, the data suggest that the Africans who reached Saint Martin were drawn from the historical populations who lived in the same locations at the time of death. The most striking feature of the skeletons recovered during construction work in the Zoutsteeg area of Philipsburg, the capital of the Caribbean island of Saint Martin in 2010. Skeletal analysis suggested that the individuals—two males and one female—were of African ancestry and that they had been aged between 25 and 40 y at the time of death. The most striking feature of the skeletons was that all three had culturally modified teeth (SI Appendix, Section 1). Similar types of dental modification are known to have been practiced by different groups in Africa but a look at the ethnographic literature (e.g., refs. 31 and 40) suggests that they cannot be used to infer points of origin or specific “tribal” affiliations.

Table 1. Modeled radiocarbon dates and sequencing results for the Zoutsteeg Three

<table>
<thead>
<tr>
<th>Sample</th>
<th>Modeled 14C age*</th>
<th>Sex</th>
<th>Nuclear coverage</th>
<th>Contamination (%)</th>
<th>Damage (%)</th>
<th>Mt coverage</th>
<th>Mt haplogroup</th>
<th>Y haplogroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>STM1</td>
<td>A.D. 1660–1688</td>
<td>M</td>
<td>0.3x</td>
<td>0.63</td>
<td>16.8</td>
<td>641x</td>
<td>L3b1a</td>
<td>R1b1c-V88</td>
</tr>
<tr>
<td>STM2</td>
<td>A.D. 1660–1688</td>
<td>M</td>
<td>0.1x</td>
<td>0.22</td>
<td>23.2</td>
<td>543x</td>
<td>L3d1b</td>
<td>—</td>
</tr>
<tr>
<td>STM3</td>
<td>A.D. 1660–1688</td>
<td>F</td>
<td>0.5x</td>
<td>0.15</td>
<td>14.9</td>
<td>651x</td>
<td>L2a1f</td>
<td>—</td>
</tr>
</tbody>
</table>

*Modeled age range of the samples based on Bayesian analysis of individual calibrated radiocarbon dates.

DNA Extraction and Library Preparation. DNA was extracted from tooth roots using a silica-based method (33) and eluted in 60 μL EB. Thirty microliters of extract were then built into Illumina libraries using the NEBNext DNA Sample Prep Master Mix Set 2 (New England Biolabs) and Illumina-specific adapters (34) following the manufacturer’s instructions, with some minor changes to the protocol (SI Appendix, Section 3). The remaining 30 μL of DNA extract were built into Illumina libraries using a single-stranded library preparation protocol, as described in ref. 35 but without first removing deoxyuracils. Both sets of libraries were amplified and indexed in 50-μL PCR reactions, purified, quantified, and pooled for sequencing (for more details, see SI Appendix, Section 3).

Whole-Genome Capture. We used two whole-genome capture methods to enrich two sets of aDNA libraries in their human DNA content. Both methods make use of biotinylated RNA probes transcribed from genomic DNA libraries to capture the human DNA in the aDNA libraries. The first method, which we refer to as...
WISC (Whole Genome In-Solution Capture), was carried out as described in ref. 8, using home-made biotinylated RNA probes. For the second capture experiment, we used the MyBait Human Whole Genome Capture Kit (MYcroarray), following the manufacturer’s instructions (9). Following the capture experiments, the enriched libraries were amplified again, purified, quantified, and sequenced on an Illumina HiSeq 2000 (for more details, see SI Appendix, Section 4).

Data Processing. We used AdapterRemoval (36) to trim adapter sequences and to remove adapter dimers and low-quality reads (SI Appendix, Section 5). Filtered fastq files were mapped to the human reference genome version hg18 and hg19, but replacing the mitochondria with the revised Cambridge Reference Sequence (37). Mapping was done using BWA v0.7.5a-r405 (38), keeping only reads with mapping quality 30 and above. Duplicate reads were removed using SAMtools’ (39) rmdup function. BAM files from different runs were merged using SAMtools merge. Sequencing error rates were estimated to be on the order of 0.3% (SI Appendix, Section 6). MapDamage2 (40) was used to rescale the quality of bases that had a mismatch error.

Reference Sequences. Reference sequences were obtained from the Ensembl Genome Browser (41) for Homo sapiens (GRCh37/hg19) and Rattus norvegicus (Rnor37). Thechimp (42) was used to identify chimpanzee-specific SNPs and INDELs.

mtDNA and Y-Chromosome Haplogroups. mtDNA haplogroups were determined by recovering reads mapping to the revised Cambridge Reference Sequence (37) from the BAM files and generating a consensus sequence and list of indels using SAMTools/oro and rmdup. Indels and hotspot mutations were excluded from analysis. Haplogroups were determined using HaploGrep (41). The maximum parsimony tree (SI Appendix, Fig. S12) was built using mtPhyl (eltsov.org). The Y-chromosome haplogroup for ST1 was determined by assembling a panel of phylogenetically informative SNPs with emphasis on those lineages previously reported to occur at appreciable frequencies within Africa. For more details, see SI Appendix, Section 11.

Principal Component Analysis. For each file containing the genotypes of the sample and reference panels, we ran smartpcg (EIGENSOFT v4.0) (42) to perform PCA. Eigenvectors were plotted independently for each dataset using RStudio (www.rstudio.com). To visualize the three samples in a single PCA plot we used Procrustes transformation as done in ref. 13. We transformed the first two PCs calculated for each intersected dataset to match the reference-only PC1 and PC2. When transforming the PCs, the ancient individual was removed from the ancient samples and the transformed coordinates were overlaid on the reference-only PC1 and PC2 plot (Fig. 1C). For more details, see SI Appendix, Section 14.

**TreeMix Analysis.** We used the probabilistic model of population splits and divergence implemented in TreeMix (18) to infer ancestry graphs (SI Appendix, Fig. S20). As input we used estimated allele frequencies for our samples and the populations in our reference panel (17). For each dataset we ran 100 bootstraps with random seeds and with the -noss and -global flags to disable sample size correction, and perform a round of global rearrangements of the graph, respectively. Additionally, the number of SNPs per block was calculated for each dataset to allow ~1,000 blocks. Finally, the root of the tree was set to Xhosa. For more details, see SI Appendix, Section 15.

**ADMixTURE Analysis.** We used the maximum-likelihood-based clustering algorithm ADMIXTURE (19) to estimate the genetic structure in our merged dataset (for more details, see SI Appendix, Section 16). We first estimated the cross-validation error with the -cv flag for K values between 1 and 6. This analysis revealed that the CV error increased with K, probably reflecting the very low Fst between the populations in the reference panel. For K = 4 to K = 6 we ran 100 replicates using a random seed and kept the Q (ancestral cluster proportions) and P (inferred ancestral cluster allele frequencies) matrices from the run with the best log likelihood. We used the P matrix from each K to estimate the most likely cluster proportions in the ancient samples as was done in ref. 43. SI Appendix, Fig. S21 shows the converged runs from K = 2 to K = 6 for ST1M, ST2M, and ST3M and 11 sub-Saharan populations in our reference panel (17).

**Acknowledgments.** We thank the staff at the Danish National High-throughput Sequencing Centre for technical support; P. Campos, M. Rasmussen, T. Korneliussen, F. Sánchez-Quinto, P. Skoglund, A. Moreno-Estrada, F. Mendez, and P. Underhill for their various input and helpful discussions; and J.S. Handler for reading earlier drafts of the paper and providing valuable comments. The Center for GeoGeneetics is funded by the Danish National Research Foundation (DNRF94). This work was supported in part by Marie Curie Fellowships from the Directorate General for Research and Innovation of the European Commission FP7/2007-2013/236435 and 317184 (to H.S. and M.E.A.); the European Research Council (to H.S. and M.E.A.); National Science Foundation Grants DMS-1201234 and DGE-1147470 (to M.C.A.A. and G.D.P.); Swiss National Science Foundation Fellowship PBSPS1-143529 (to A.S.M.); National Institute of Health Grants N3A 5F32HG007342 and K99 GM104158 (to M.L.C. and P.L.F.J.); Leverhulme Early Career Fellowship ECF-2012-123 (to M.W.D.); Ministerio de Ciencia e Innovación Grant SAF2011-26983 (to A.S.); the “Plan Galego IDT” EM 2012/045 (to A.S.); “Sistema Universitario de Investigación y Estudios Doctorales” (to P.L.F.J.); and Lundbeck Foundation Grant R52-A5602 (to M.T.P.G.); and Danish Council for Independent Research Grant 10-081390 (to M.T.P.G.).