From: Prof. Carlos D. Bustamante, Ph.D.
Date: October 10, 2018

Executive Summary. *We find strong evidence that a DNA sample of primarily European descent also contains Native American ancestry from an ancestor in the sample's pedigree 6-10 generations ago. We find little or no evidence of African ancestry in this sample.*

Request.
To analyze genetic data from an individual of European descent and determine if there is reliable evidence of Native American and/or African ancestry. The identity of the sample donor, Elizabeth Warren, was not known to the analyst during the time the work was performed.

Background.
The individual's DNA was previously sent to a DNA testing laboratory for genotyping, with a DNA microarray. The resulting DNA genotypes were sent to Dr. Bustamante for analysis. An expert in genetic ancestry analysis carried out the computational work described below under Dr. Bustamante's supervision.

Methodology.
Analysis was performed to scan the human genome to identify individual chromosomal segments with European, African, East Asian, and Native American ancestry, using the RFMix computer program, which was developed by us (Maples et al., 2013) and is one of the leading methods for ancestry analysis. The ancestry analysis used reference samples from various regional populations used in human genetics (see below). Because available samples do not provide complete coverage of all Native American groups, some segments with Native American ancestry may be missed. In addition, it is not possible to reliably associate smaller segments having Native American ancestry with any specific tribe or group.

Samples.
The individual's sample contained information on 764,958 sites of genetic variation across the human chromosomes, of which 660,173 overlapped with sites in the reference set used for ancestry analysis. Our population reference set consisted of 148 individuals (a continental reference panel of 37 individuals from across Europe, 37 from Nigeria with Sub-Saharan African ancestry, 37 from across the Americas with Native American ancestry, and 37 individuals from China). To determine whether the Native American ancestry results in the sample were unusually high relative to other individuals of European ancestry, analysis was also performed on 185 individuals from two reference sets from the 1000 Genomes Project— Americans of predominantly European ancestry from Utah (n = 99 individuals) and British individuals of European ancestry from Great Britain (n = 86 individuals).

Results. The results were as follows:
(1) The great majority of the individual's identifiable ancestry is European: 95% of high confidence segments (defined as those segments with at least 99% posterior probability of assignment) were identified by RFMix as being of European origin. This is likely an underestimate as many of the segments not classified as high-confidence are also likely to be European in origin. The analysis also identified 5 genetic segments as Native American in origin at high confidence, defined at the 99% posterior probability value. We performed several additional analyses to confirm the presence of Native American ancestry and to estimate the position of the ancestor in the individual's pedigree.

(2) The largest segment identified as having Native American ancestry is on chromosome 10. This segment is 13.4 centiMorgans in genetic length, and spans approximately 4,700,000 DNA bases. Based on a principal components analysis (Novembre et al., 2008), this segment is clearly distinct from segments of European ancestry (nominal p-value 7.4×10^{-7}, corrected p-value of 2.6×10^{-4}) and is strongly associated with Native American ancestry.

(3) The total length of the 5 genetic segments identified as having Native American ancestry is 25.6 centiMorgans, and they span approximately 12,300,000 DNA bases. The average segment length is 5.8 centiMorgans. The total and average segment size suggest (via the method of moments) an unadmixed Native American ancestor in the pedigree at approximately 8 generations before the sample, although the actual number could be somewhat lower or higher (Gravel, 2012 and Huff et al., 2011).

(4) The sample was compared to the results of the 185 reference individuals with European ancestry, from Great Britain and Utah.

- The segment on chromosome 10 observed in the individual is larger than any of the segments identified as having Native American ancestry in any of the 185 reference individuals.
- The total length of Native American segments observed in the individual is greater than the average value for the reference individuals — by 12.4-fold (corresponding to 12.7 standard deviations) for the individuals from Great Britain and 10.5-fold (corresponding to 4.9 standard deviations) for the individuals from Utah.

(5) The sample also contained smaller segments that could not be reliably assigned to any specific ancestry group (at 99% posterior probability). The total length of these unassigned segments was 366 centiMorgans, and they span 267,650,000 DNA bases.

Conclusion. While the vast majority of the individual's ancestry is European, the results strongly support the existence of an unadmixed Native American ancestor in the individual's pedigree, likely in the range of 6-10 generations ago.

Background Information

The human genome consists of 23 chromosomes containing sequences of 3,234,830,000 DNA bases, with genetic length of 3,595 centiMorgans. Because individuals receive one set from each parent, they carry 46 chromosomes, with a total of 6,469,660,000 DNA bases and a total genetic length of 7,190 centiMorgans. (CentiMorgans are a unit of genetic recombination, which is particularly relevant for ancestry analysis.)

Genetic methods for ancestry determination look at the small fraction of these DNA bases that commonly vary between individuals. A subset of these variable bases (called genetic markers) had been genotyped for the test sample. The markers from the test sample were then analyzed using two classes of methods: global ancestry methods and local ancestry methods.

Global ancestry methods treat each marker as independent, ignoring correlations between neighboring markers on the DNA strand. Using the observed frequency of each marker across regional reference populations (e.g. European, African, Native American), global ancestry methods determine the likelihood for each marker in the sample originating from each population. A probabilistic model is then applied to estimate the proportion of markers in the test sample originating from each regional ancestry. Global ancestry methods have the advantage of considering all genotyped markers at once in the model (more data), but the disadvantage of ignoring local correlations and patterns between neighboring markers (since each marker is treated as independent). Global ancestry methods thus may fail to identify contributions to ancestry, evident in chromosomal segments (Alexander et al., 2011).

Local ancestry methods consider short sequences of neighboring markers; each segment has fewer markers but has greater statistical power to look at ancestry patterns, by searching for similar sequences of markers amongst the reference populations. Using local ancestry methods one can identify the ancestral population (European, African, Native American) from which each piece of each chromosome derived. Our method (RFMix) uses random forests, a highly accurate modern machine learning algorithm, to identify and match these sequence patterns (Maples et al., 2013). To analyze the longest Native American ancestry segment in the sample, we used an independent machine learning technique called principal component analysis (Novembre et al., 2008).

For maximal accuracy, we use reference populations that have been fully sequenced (complete genomes) rather than references that had been genotyped at only a subset of sites. These samples come from the 1000 Genomes research project, which sequenced full genomes from individuals around the world (1000 Genomes Project Consortium, 2012). For Native American references, we used samples within the 1000 Genomes project of Native American ancestry; these samples come from Mexico, Peru, and Colombia. (It is not possible to use Native American reference sequences from inside the United States, since Native American groups within the US have not chosen to participate in recent population genetics studies.) The 1000 Genomes reference samples come from Nigerian Yoruba individuals (for Sub-Saharan Africa), Finnish, Tuscan Italian, and Spanish individuals (for Europe), and northern Chinese individuals for East Asia. (The latter reference was used to test for East Asian regional ancestry, since that can otherwise be mis-assigned as Native American). In our analysis, an individual with 100% ancestry assigned to a single population (e.g., European or African) is defined as an “unadmixed”.

We also compared the ancestry segments seen in the sample to 185 reference individuals from Europe (Great Britain) as well as American individuals from Utah. A few of the Utah individuals have a small amount of Native American ancestry, and for this reason the standard deviation of Native American ancestry in the Utah individuals is somewhat higher than in the British samples.

Biography

Dr. Carlos D. Bustamante is an internationally recognized leader in the application of data science and genomics technology to problems in medicine, agriculture, and biology. He received his Ph.D. in Biology and MS in Statistics from Harvard University (2001), was on the faculty at Cornell University (2002-9), and was named a MacArthur Fellow in 2010. He is currently Professor of Biomedical Data Science, Genetics, and (by courtesy) Biology at Stanford University. Dr. Bustamante has a passion for building new academic units, non-profits, and companies to solve pressing scientific challenges. He is Founding Director of the Stanford Center for Computational, Evolutionary, and Human Genomics (CEHG) and Inaugural Chair of the Department of Biomedical Data Science. He is the Owner and President of CDB Consulting, LTD. and also a Director at Eden Roc Biotech, founder of Arc-Bio (formerly IdentifyGenomics and BigData Bio), and an SAB member of Imprimed, Etalon DX, and Digitalis Ventures among others.

References

1. 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." Nature 491.7422 (2012): 56.
2. Alexander, David H., and Kenneth Lange. "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation." BMC bioinformatics 12, no. 1 (2011): 246.
3. Gravel, Simon. "Population genetics models of local ancestry." Genetics (2012): genetics-112.
4. Huff, Chad, David Witherspoon, Tatum Simonson, Jinchuan Xing, Scott Watkins, Yuhua Zhang, Therese Tuohy et al. "Maximum-likelihood estimation of recent shared ancestry (ERSA)." Genome research (2011): gr-115972.
5. Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics, 93(2), 278-288.
6. Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap et al. "Genes mirror geography within Europe." Nature 456, no. 7218 (2008): 98.

Appendix

***Figure 1**: Principal component analysis of the segment from chromosome 10 (genotype)*

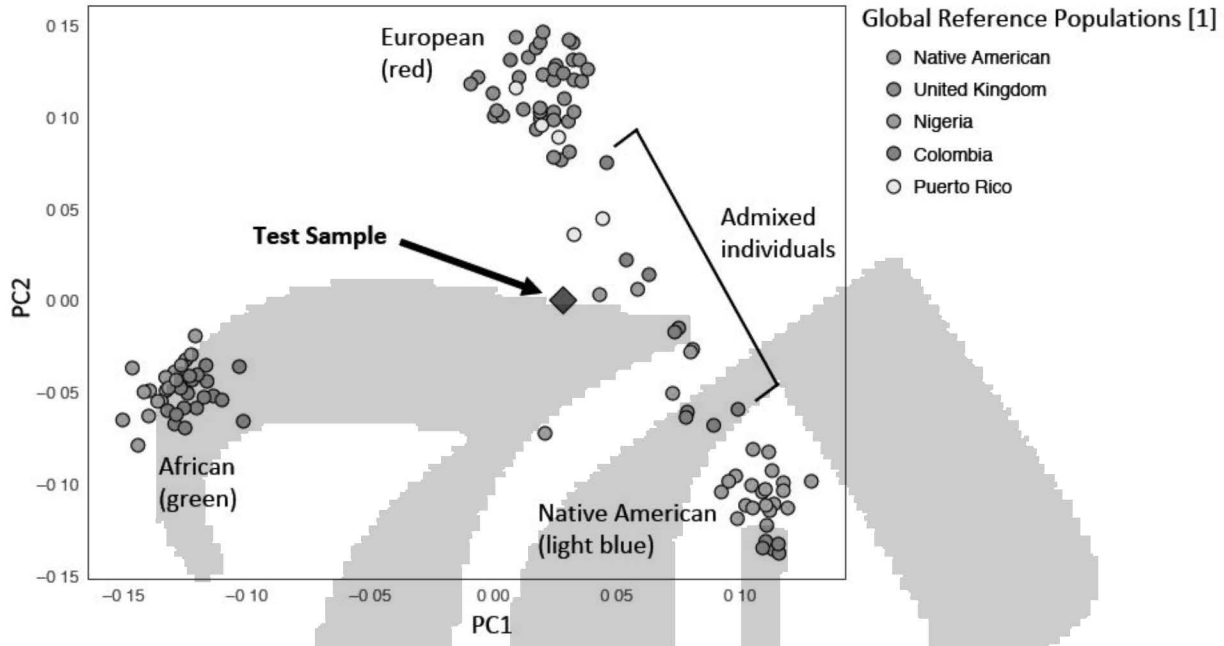

Figure 1: Test sample has a segment on chromosome 10 with one haplotype (one parental ancestry) of Native American background and one of European background. For this reason it lies halfway between the European and Native American clusters. Other individuals having dual Native American – European heritage in this particular chromosomal segment (admixed individuals) are seen to fall nearby (for example some individuals from Colombia and some from Puerto Rico).

***Figure 2**: Native American ancestry specific principal component analysis (haplotype)*

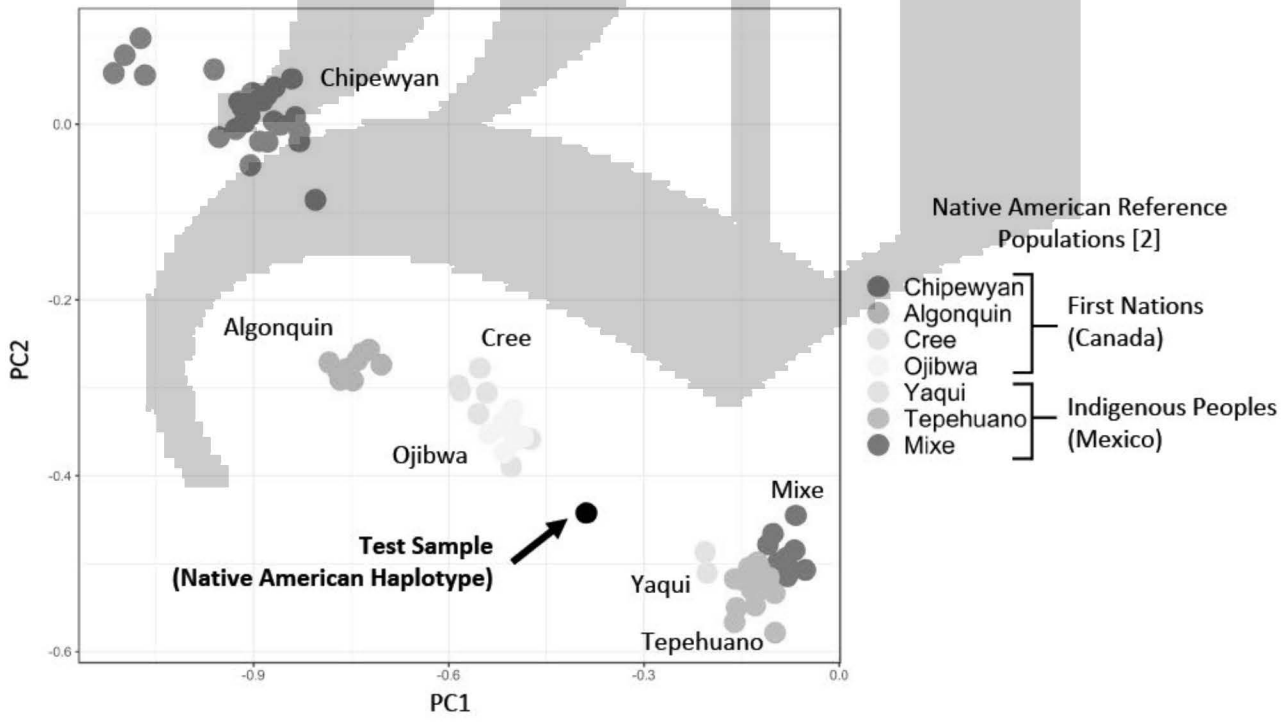

Figure 2: In this ancestry specific analysis [3], the Native American haplotype of the test sample is displayed against a panel of Native American references from across North America. The test sample's Native American ancestry falls between First Nations (Canada) and Indigenous Mexican peoples, as would be expected for Native American ancestry deriving from the lower 48 states of the United States.

[1] 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." Nature 491, no. 7422 (2012): 56.

[2] Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, Maria V. Parra et al. "Reconstructing native American population history." Nature 488, no. 7411 (2012): 370.

[3] Moreno-Estrada, Andrés, Simon Gravel, Fouad Zakharia, Jacob L. McCauley, Jake K. Byrnes, Christopher R. Gignoux, Patricia A. Ortiz-Tello, Ricardo J. Martínez, Dale J. Hedges, Richard W. Morris, Celeste Eng, Karla Sandoval, Suehelay Acevedo-Acevedo, Paul J. Norman, Zulay Layrisse, Peter Parham, Juan Carlos Martínez-Cruzado, Esteban González Burchard, Michael L. Cuccaro, Eden R. Martin , Carlos D. Bustamante. 2013. “Reconstructing the population genetic history of the Caribbean.” PLoS Genetics 9, no. 11 (2013): e1003925.