Legal Medicine 23 (2016) 49–54

Contents lists available at ScienceDirect

Legal Medicine

journal homepage: www.elsevier.com/locate/legalmed

The genetic structure of native Americans in North America based on the Globalfiler® STRs

Kelly L. McCulloh [a], Jillian Ng [b], Robert F. Oldt [b], Jessica A. Weise [a], Joy Viray [c], Bruce Budowle [d,e], David Glenn Smith [a,b], Sreetharan Kanthaswamy [a,b,f,*]

[a] Forensic Science Graduate Program, University of California, One Shields Avenue, Davis, CA 95616, USA
[b] Molecular Anthropology Laboratory, Department of Anthropology, University of California, One Shields Avenue, Davis, CA 95616, USA
[c] Sacramento County District Attorney's Crime Laboratory, 4800 Broadway Suite 200, Sacramento, CA 95820, USA
[d] University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia
[f] California National Primate Research Center, University of California, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 15 July 2016
Received in revised form 11 September 2016
Accepted 23 September 2016
Available online 23 September 2016

Keywords:
Forensic science
Native Americans
North America
Geography
Short tandem repeats
Population genetics

ABSTRACT

Current forensic STR databases, such as CODIS, lack population genetic data on Native American populations. Information from a geographically diverse array of tribes is necessary to provide improved statistical estimates of the strength of associations with DNA evidence. The Globalfiler® STR markers were used to characterize the genetic structure of ten tribal populations from seven geographic regions in North America, including those not presently represented in forensic databases. Samples from the Arctic region, Baja California, California/Great Basin, the Southeast, Mexico, the Midwest, and the Southwest were analyzed for allele frequencies, observed and expected heterozygosities, and F-statistics. The tribal samples exhibited an F_{ST} or θ value above the conservative 0.03 estimate recommended by the National Research Council (NRC) for calculating random match probabilities among Native Americans. The greater differentiation among tribal populations computed here ($\theta = 0.04$) warrants the inclusion of additional regional Native American samples into STR databases.

1. Introduction

Population structure can be used to quantify genetic differentiation among subpopulations relative to the total population, and is expressed as F_{ST} [1] or theta (θ) [2]. F_{ST} determinations are necessary for calculating random match probabilities in forensic casework, as they provide investigators population genetic information to estimate match probabilities of a forensic sample to a known source. The National Research Council (NRC) [2] recommends that a correction factor value of F_{ST} or $\theta = 0.01$ be used for general United States populations while a value of 0.03 be used for smaller and more isolated populations, such as Native Americans, where subdivision is more prevalent when determining genetic variation among populations.

Consistent with the NRC's recommendation, Budowle et al. [3] found that Native Americans exhibited the highest differentiation compared to Caucasian, Hispanic, African American, and Asian populations, with an F_{ST} estimate of 0.0282. While Caucasian Americans showed little or no genetic subdivision, the estimates of F_{ST} between Navajos and Apaches was 50 times that among African-Americans, 14 times that among Hispanic-Americans, and only 0.13 times of the value of the estimate for Asian-Americans [3]. This observation is especially significant because Navajo and Apache are closely related genetically. These tribes share a relatively recent common ancestry, which undoubtedly contributed to their F_{ST} value, even though both tribes have been highly admixed with different populations, including unrelated Native American tribes, for at least 500 years [4].

Furthermore, based on a study of 678 autosomal STR loci genotyped across 422 individuals from 29 Native American populations in North America, Central America, and South America [5], Native American tribes, including Chipewyan, Cree, Ojibwa (North America), Cabecar, Guaymi, Kaqchikel, Maya, Mixe, Mixtec, Pima, Zapotec (Central America), Arhuaco, Aymara, Embera, Huilliche, Inga, Kogi, Quechua, Waunana, Wayuu, Zenu (western South America), and Ache, Guarani, Kaingang, Karitiana, Piapoco, Surui, Ticuna [Arara], and Ticuna [Tarapaca] (eastern South America), showed greater differentiation than any other comparably sized

* Corresponding author at: School of Mathematics and Natural Sciences, ASU at the West Campus, 4701 W Thunderbird Rd., Glendale, AZ 85306-4908, USA.
E-mail address: sree.kanthaswamy@asu.edu (S. Kanthaswamy).

http://dx.doi.org/10.1016/j.legalmed.2016.09.007
1344-6223/

population (F_{ST} or $\theta = 0.08$). Therefore, the F_{ST} estimate from Wang et al. [5] suggests a higher F_{ST} than the 0.03 value currently recommended by the NRC [2] will be needed to adjust for population structure in forensic cases, including paternity testing, involving Native American individuals. To establish an informative Native American population database, a more detailed examination is necessary to determine whether significant differentiation exists to warrant the creation of additional Native American datasets. Given that the CODIS Native American STR database lacks tribes that are genetically similar to the vast majority of tribes living today and that geography is responsible for 60% of genetic differentiation [6], it is necessary to generate information for a more geographically diverse representation of additional tribes representing a greater number of geographic populations to better characterize genetic variation among Native Americans [4].

The current 13 CODIS loci are CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11 [7]. This study included eight additional autosomal loci (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, D22S1045, and SE33), which are included in the Globalfiler® PCR Amplification Kit (Applied Biosystems, Foster City, CA), and are included in the expanded CODIS core loci [8]. Profiling new Native American samples with these 21 loci will expand the existing pool of genetic profiles in the DNA database and provide more information on allele frequencies and population substructures. In addition, this study focused on the effect of geographic location on population structure and differentiation and quantified such variation. The STR typing of the geographically representative North American tribes from the Arctic region, Baja California, California/Great Basin, the Southeast, Mexico, the Midwest, and the Southwest establishes a more complete Native American database that can directly assist in forensic investigations as well as provide more reliable estimates of allele frequencies and genetic variation within and among the tribes.

2. Materials and methods

The Department of Anthropology Laboratory at UC Davis houses one of the largest databanks of geographically and linguistically representative full blood Native North American samples. Of the 3327 tribal DNA samples currently archived and available at the Department of Anthropology at UC Davis, the 418 samples from random individuals analyzed here were the only ones that met the quantification requirements for STR analysis. Prior approval from the UC Davis IRB (ID 430207-2) was obtained for the use of these samples for this study. The list of 418 tribal samples included in the study, as well as their geographic origins and mtDNA haplogroup distributions are shown in Table 1. In North America, haplogroup frequencies exhibit regional continuity that can be helpful in understanding relationships among the populations in those areas [9]. The geographical regions of the Native American tribes used in this study were based on Driver [10] and Lorenz and Smith [11]. Samples from the Southwest, Southeast, Midwest/Great Plains and Arctic region as well as samples from California/Great Basin, Baja California, and Mexico were included in this study.

2.1. Sample extraction

Samples consisting of serum, buffy coat, blood, or purified DNA were originally stored at −20 °C but have recently been maintained at 4 °C. DNA samples were extracted from serum, buffy coat, and blood samples using the QIAamp DNA Blood Mini Kit (QIAGEN, Redwood City, CA) following the manufacturer's protocol.

2.2. Sample quantification

DNA samples were quantified using the Quantifiler® Duo Quantification Kit and the 7500 Fast Real-time PCR system (Applied Biosystems). The quantification standards and DNA samples were both run in duplicate following the manufacturer's protocol.

2.3. Sample amplification

DNA samples were diluted to 1.0 ng/μL and amplified along with the National Institute of Standards and Technology (NIST) Standard Reference Material (SRM) 2391c reference DNA sample using the Globalfiler® PCR Amplification Kit (Applied Biosystems) according to the manufacturer's protocols. Amplified samples were diluted in Hi-Di Formamide (Applied Biosystems) and run on a 3130*xl* Genetic Analyzer with POP-4 polymer (Applied Biosystems) following manufacturer recommended parameters. The GeneScan™ 600 LIZ® Size Standard (Applied Biosystems) was used as the internal sizing standard and the Globalfiler® Allelic Ladder (Applied Biosystems) was used for sizing the alleles. Alleles were called using GeneMapperID-X v.1.4 (Applied Biosystems) with the Local Southern sizing method.

2.4. Statistical or other methods of data analyses

The extent of genetic variation within and among tribal samples, number of alleles, and observed and expected heterozygosity for each autosomal locus in each geographic region were calculated using Arlequin v3.5.1.2 [12]. Arlequin was also used to calculate the following F-statistics: F_{ST} – the proportion of genetic variance in a population that is due to differences among subdivisions within that population; F_{IS} – inbreeding coefficient, F_{IT}: total inbreeding coefficient, and pairwise F_{ST} – to assess the degree of differentiation between pairs of tribal samples which provides an insight into the historical connections among tribal samples and among the geographic regions these tribes represent. The statisti-

Table 1
The seven geographic samples represented by 10 tribes, their sample sizes (N), and mtDNA haplogroup frequencies. Tribes in the southwest US region of North America, such as Apache and Yavapai, have a high frequency of haplogroup B, a moderate frequency of haplogroup C, and low frequencies of haplogroups A, D, and X [11], while a few tribes in the northern half of Mexico, such as Huichol, and Cora, have lower frequencies of A, suggesting gene flow between the North American Southwest and Mexico [24].

Geographic region	Tribe	N	A	B	C	D	X	Refs.
Arctic	Eskimo	44	0.97	0	0	0.03	0	[11]
Baja CA	Cochimi	25	0.08	0.46	0.46	0	0	[11]
CA/Great Basin	Miwok	33	0.12	0.41	0.06	0.41	0	[11]
Southeast	Cherokee	34	0	0.31	0.31	0	0.38	[11]
Mexico	Cora	64	0.31	0.51	0.14	0.04	0	[24]
	Huichol	30	0.31	0.53	0.16	0	0	[24]
	Seri	29	0	0.13	0.86	0	0	[25]
Midwest	Chippewa	21	0.48	0.11	0.19	0	0.21	[11]
Southwest	Apache	88	0.62	0.17	0.14	0.07	0	[11]
	Yavapai	50	0	0.86	0.03	0.03	0.08	[11]