Document Title:	The Enhancement of the Native American CODIS STR Database for Use in Forensic Casework
Author(s):	Sreetharan Kanthaswamy, Kelly L. McCulloh, Jillian Ng, Robert F. Oldt, Jessica A. Weise, Joy Viray, Bruce Budowle, David Glenn Smith
Document Number:	253079
Date Received:	July 2019
Award Number:	2014-DN-BX-K024

Special Report title: The Enhancement of the Native American CODIS STR Database for Use in Forensic Casework

Award Number: National Institute of Justice (NIJ) grant number 2014-DN-BX-K024

Author: Sreetharan Kanthaswamy, Kelly L. McCulloh¬, Jillian Ng, Robert F. Oldt, Jessica A. Weise, Joy Viray, Bruce Budowle, David Glenn Smith,

Purpose of the project

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 gentoyped 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 population (F_{ST} or θ = 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 NRC[2] will be needed to adjust for population structure in forensic cases, including paternity testing, involving Native American individuals. After peer-review, STR data fro this study, including full profiles (genotype information) and allele frequencies, will be submitted to the National Institute of Standards and Technology (NIST) to enable CODIS laboratories to use this information in forensic investigations involving these tribal communities.

Project design and methods

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, South San Francisco, 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.

The STR database (including full profiles (genotype information) and allele frequencies) produced in this study will be submitted to a peer-reviewed journal prior to being uploaded onto the National Institute of Standards and Technology (NIST) forensic human database website (http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm). This will facilitate the accessibility of this data to CODIS laboratories that are performing forensic investigations involving Native American communities.

Project subjects

Over 6800 tribal DNA are samples currently archived and available at the Department of Anthropology at UC Davis, the 418 samples 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 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, and Baja California and Mexico were included in this study.

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.

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.

Sample Amplification

DNA samples were diluted to 1.0ng/µL and amplified along with the 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.

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 also was 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 also among the geographic regions these tribes represent. The statistical significance of the pairwise F_{ST} computations was determined with a probability distribution constructed from permutation tests (N = 1000) with Bonferroni corrections for multiple comparisons. Mann-Whitney U tests were performed to determine if population-specific estimates of diversity and F_{IS} differed significantly across populations and from the overall average. The Hardy-Weinberg Exact Test in the program GENEPOP 4.2 was used to determine if any of the tribal samples showed detectable deviations from expectations of equibrium [13, 14]. CONVERT v1.31 [15] was used to compute private allele frequencies (or alleles restricted to one group) at each locus within each geographically separate sample. Because differences in sample size can affect allele representation and estimates of genetic variation (particularly due to the presence or absence of rare alleles), each of the genetic parameters was recalculated using 1,000 iterations of 21 randomly selected individuals from each tribe (Table 2) normalized to match that of the Chippewa tribe (N = 21).

Project findings

Table 2 presents the estimates of allele numbers (Na), and observed (OH) and expected (EH) heterozygosities across the geographic regions for all 21 autosomal STR loci. The tribal samples averaged between

6 (Eskimo - Arctic) and 8 (Miwok - CA/Great Basin, Cherokee - Southeast, Cora - Mexico, and Apache (San Carlos Apache Reservation) and Yavapai - Southwest) alleles per locus. Estimates of allele numbers, both rare and common, based on 21 random individuals from each tribe suggest an influence of sample size; the difference between Na based on total sample and the sample of 21 is greatest for those tribes with the largest sample size (i.e., Cora - Mexico, and Apache and Yavapai - Southwest). The values of OH and EH in Table 2 did not appear to be influenced by sample size. OH values range from 0.68 (Eskimo - Arctic) to 0.78 (Miwok - CA/Great Basin) while EH values range from 0.69 (Eskimo - Arctic) to 0.77 (Cherokee - Southeast). Several private alleles among the tribes were identified with the Cherokee (Southeast) sample having the most (10), followed by Chippewa (Midwest - 5), Apache (Southwest - 5), Cora (Mexico - 5), Miwok (CA/Great Basin - 5), Yavapai (Southwest - 4), Huichol (Mexico - 1) and Seri (Mexico - 1) (Table 3). Frequencies of private alleles ranged from 0.006 to 0.005 (Table 3).

Pairwise F_{ST}, as well as population-specific F_{ST}, and average F_{IS} are shown in Table 4; all pairwise F_{ST} p-values were statistically significant at the 0.05 level. Pairwise F_{ST} values from Table 4 suggest that differentiation among Native American tribes ranged from 0.006 (between Apache and Yavapai - Southwest) to 0.113 (between Eskimo - Arctic and Seri - Mexico). In addition to exhibiting the greatest levels of differentiation with each other, the Eskimo (Arctic) and Seri (Mexico) populations also exhibited the greatest differences from most of the study samples, with mean pairwise F_{ST} values of 0.073 and 0.070, respectively. The Arctic sample also showed genetic differences from other geographic samples that were correlated with geographic distance. Differentiation within the Continental US did not appear to be correlated with their geographical distances. Within Mexico, the mean pairwise F_{ST} among the Cora, Huichol, and Seri was approximately 0.05 with Cora and Huichol exhibiting the least differences (0.02) and Seri appearing to be the most genetically isolated. When the Cochimi tribe from Baja California was compared with the other samples from Mexico, a range of pairwise F_{ST} from 0.02 (Cochimi-Cora) to 0.068 (Cochimi-Seri) was observed. It appears that geographic and genetic distances between Mexico and the other study samples are correlated.

F_{IS} values (Table 4) were highest for the Cora tribe from Mexico (F_{IS} = 0.04), followed by the Cherokee tribe (Southeast) and Eskimo (Arctic) samples (F_{IS} = 0.034 and 0.017, respectively). The other tribes exhibited either low (nearing zero) levels of F_{IS} values or none at all (negative values).

Implications for criminal justice policy and practice in the United States

The 21 autosomal STR loci in the Globalfiler® PCR Amplification Kit were evaluated for genetic diversity, differentiation, and structure within and between the seven geographically representative Native American tribal samples. Larger tribal sample sizes tended to be more optimal than smaller ones for finding the most alleles or for computing genetic diversity estimates; for instance the decline in Na when samples of size 21 were analyzed is greatest for the largest sample sizes (i.e., Cora - Mexico, and Apache and Yavapai - Southwest). The same average number of 8 alleles per locus was observed in this study as in the Budowle et al. studies [3, 16] which also used Apache and Eskimo samples albeit with much greater sample numbers. In spite of having screened many more individuals from the Apache, Athabaskan, Inupiat, and Yupik tribes, i.e. at least twice as many used here, Budowle et al. [3] reported slightly lower OH (0.70) as well as EH (0.71) in the Apache tribe and comparable OH and EH estimates among the Alaskan tribes; average OH = 0.70 and average EH = 0.71.

Most of the STR alleles observed in this study are worldwide in distribution. Private STR alleles with a maximum frequency of 5% have been estimated in the present study. While no private allele with a frequency above 0.13 has been found [17], with the exception of a nine repeat allele (9RA) in D9S1120 which occurs at a high average frequency of 0.36 among Native American samples [18-20], the determination of population specific private alleles in this study, ranging from 1 (in the Seri and Huichol tribes of Mexico, respectively) to 10 (in the Cherokee from the Southeast) could further assist forensic investigators given their potential to differentiate tribal samples and to find perpetrators of specific tribal origin. However, a caveat of relying on private alleles for forensic purposes runs the risk of attributing alleles alike whether they stem from convergence (homoplasy) or from common descent. Also, inferring these alleles as private may be an artifact of limited sample size.

The estimated overall F_{ST} value for the total population of Native Americans ($F_{ST} = 0.04$) is greater than reported by Budowle et al. [3] as would be expected with increased sample populations studied. The present study shows that Native American populations exhibit greater inter-population differentiation than estimated by Budowle et al. [3]. The geographic (as well as language) affiliations of this study sample are more heterogeneous than the samples in the Budowle et al. [3] study which may have attributed to the higher F_{ST} values generated here. Wang et al.'s [5] study based on STRs (albeit NOT the CODIS STRs) computed F_{ST} values for the Americas that far exceeded the value obtained herein, especially for Central and South American populations (F_{ST} = 0.06 to 0.15), which were not considered in the present study. However, they also observed a value of F_{ST} of 0.03 among the North American tribes of Chipewyan, Cree and Ojibwa. While the North American F_{ST} estimate reported by Wang et al. [5] is more consistent with that of Budowle et al. [3] than with the present study, the three tribes in their study all derive from the same geographic region and belong to the same language group [5]. The present study does not support the NRC's recommendations [2] for using a correction factor of θ of only 0.03 for calculating match probabilities in small isolated populations, such as the Native Americans. In fact, the present results show that a more stringent F_{ST} or θ value of at least 0.04 be used for computing match probabilities in Native Americans.

The correlation between genetic and geographic distances is especially pronounced between the Arctic population and all other regional populations. This correlation is probably confounded by other factors, such as pronounced language differences that accompany the geographic difference. Dissimilarities in languages either play a partial role in producing genetic barriers or otherwise co-occur with factors that impede gene flow [5]. The higher F_{ST} values of the Arctic region for average and across all pairwise comparisons reflect the population's relative geographic isolation (as well as language isolation) from the other populations. A Mann-Whitney U treatment of the heterozygosity and F_{IS} estimates revealed significantly ($p < 0.05$) lower observed and expected heterozygosities of the Arctic population (OH = 0.68 and EH = 0.69) in relation to the other populations (OH = 0.73 and EH = 0.73) and averaged across all other populations. The higher F_{IS} value as compared to the total population average can be attributed to a lack of migration and an increase of non-random mating that also stems from isolation. The Arctic population's low amount of variation based on OH and EH estimates is consistent with the population's mtDNA variation. The Arctic population is almost exclusively mtDNA haplogroup A (average haplogroup A frequency = 0.97), which could be due to its descent from one distinct Beringia ancestor and the theory that only the well-suited populations remained in Alaska [21].

In contrast to the Arctic population, other Native American populations have a wider range of mtDNA haplogroups (predominantly A, B, C and D) with a few tribes having higher frequencies of haplogroups X and L [9] and an average F_{ST} value of 0.05, which is higher than all other sample comparions if the Arctic sample was not included. Most X haplotypes are indigenous while the L haplotype is an introduct via African ancestry [9]. Much of the genetic diversity found in the Mexico population is thought to be from the influx of genes from primarily Aztec and Spanish conquerers [9]. In Mexico, the Seri, Cora, and Huichol tribes, especially the Seri, are more isolated from the rest of the Mexican tribes since they live in inaccessible places, preserve their customs, and only reproduced amongst themselves [9]. The geographic isolation of these tribes also may have contributed to the differences in haplogroup frequencies and the relatively high tribe specific F_{ST} values of 0.03, 0.05, and 0.07, respectively for the Cora, Huichol, and Seri. Moreover, the high F_{IS} value in Mexico can be attributed to the geographic isolation and inbreeding across its tribal groups.

The lower differentation (pairwise $F_{ST} = 0.02$) between Cochimi (Baja CA) and the Miwok (CA/Great Basin) compared to the differentiation between the former and Mexico ($F_{ST} = 0.04$) is consistent with the theory that coastal migration brought populations to the Baja peninsula [22]. The pairwise F_{ST} values between Baja CA and the rest of the populations (mean pairwise $F_{ST} < 0.05$) also suggest that Baja CA is not significantly differentiated from the rest of North America. The Yuman-speaking tribes of Baja California (including the Cochimi, as well as the Cucupa, Kiliwa, Kumiai, and Pai Pai which were not analyzed here) were moved to their current location from their homeland in Mexico Proper, and are closely related to the Yuman-speaking tribes of the American Southwest (e.g. Hualapai and Yavapai), which can explain the lack of differentiation among those regions.

The Southwest (Apache and Yavapai) exhibited the lowest amount of differentiation ($F_{ST} = 0.02$) with the Midwest (Chippewa), which suggests that a rate of high gene flow between the Southwest and Midwest (Chippewa) populations existed historically. MtDNA haplogroup A-D and X frequencies observed in the Southwest, Mexico, and North America also are consistent with high levels of gene flow among those regions [23]. Even though in this study the Southwest was slightly differentiated from the CA/Great Basin and Baja CA (range $F_{ST} = 0.02$ to 0.04), mtDNA haplogroup B, which is predominant in the Southwest, also is prevalent in the CA/Great Basin and northern Mexico. That this haplogroup is rarely seen in Mesoamerica and occurs in much lower levels elsewhere in North America [23] further supports the importance of the Southwest as a point of divergence during the migration out of Alaska [23].

Fladmark [22] theorized that only Pacific coastal and coastal interior migration along river systems occurred during the glacial period, therefore populations such as the Arctic, Baja CA, Mexico, and Southwest were most likely

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

to have been established first. Once these populations were established and differentiated, the glaciers receded and the mid-continental populations were able to migrate out of Alaska and into North America [21]. Since mid-continental migration of the Midwest and Southeast populations occurred more recently than the Pacific coastal and coastal interior migrations, less differentiation is expected (F_{ST} = 0.02) between those populations when compared to the other populations. The Arctic had the least amount of differentiation from the Midwest and Southeast (pairwise F_{ST} = 0.05 and 0.06, respectively) compared to the other populations, suggesting those two populations were the last to diverge from the Arctic. The divergence model predicts that populations located close together will be more genetically similar than distant populations [24]. Divergence models assume that every time a divergence event occurs, the new population occupies an area adjacent to the parent population, such as between the Midwest, and Southeast. The Southeast population was least differentiated from the Midwest (pairwise F_{ST} = 0.02), and CA/Great Basin populations (pairwise F_{ST} = 0.01), suggesting a migration out of the Northwest rather than from the west, as Fladmark [22] proposed.

In humans, there is a clear pattern of isolation by distance that can be explained by a migration model. Geographic isolation is primarily a result of sequential eastward colonization out of Africa by modern human populations (also known as a “serial founder effect” as in Ramachandran et al. [25]), including the populations used in this study. Similar to the divergence model, migration models, such as the stepping-stone, are based on the idea that migration only occurs between adjacent populations [24], such as between the Arctic and the Northwest, or between the Northwest and Midwest. If migration proceeded after glacial recession and originated from Alaska and the Northwest, then increased levels of gene flow in the Midwest, as populations migrated through Northwest to the Midwest to the Southeast, would be expected. The F_{IS} value for the Midwest was –0.02, indicating a lack of inbreeding, possibly due to migration through the Midwest into the Southeast after the glacial recession. Migration through the Midwest would bring in excess gene flow and would increase the amount of heterozygosity seen in that population. Alongside Mexico, the Southeast exhibited a high F_{IS} value (0.03), suggesting that population migration ended once the Atlantic Ocean was reached.

The low level of differentiation between the Midwest and the Southeast is consistent with the theory that migration into the mid-continental region of North America only took place when the glaciers retreated and at a more recent time than the Pacific coastal and coastal interior migration into South America [22]. Although not isolated by geography, these mid-continental populations were probably isolated from the Arctic, Baja CA, Southwest, and Mexico populations by time. It is clear that the Arctic population is more differentiated than the rest of the Native American populations in North America.

Since the CODIS Native American STR database contains only tribes from the Arctic and Subarctic regions and does not include the vast majority of other geographically diverse tribes, it is necessary to expand the database to include unique genetic populations. Adding individuals from the Southwest, Baja CA, Mexico, and a mid-continental group to the Arctic group already included in the STR database will improve representation of Native Americans, allow higher confidence in robust statistical estimates, and possibly provide inferrance on ethnicity from an unknown sample.

This study supports the theory of migration across the Bering Strait into Alaska, and down into South America [22], and how historic events including migration have impacted extant Native American population structure. F_{ST} values among the study populations are greater than the conservative estimate of $F_{ST} = 0.03$ [2]; the present examination of geographically representative tribal samples suggests that a greater F_{ST} value of 0.04 be used in random match calculations involving Native Americans. Expanding the study to include samples from Central and South America may increase the F_{ST} estimate [5]. The differentiation seen among the tribes in the mid-continental group is likely due to later migration, after the recession of the glaciers, compared to other groups. Differentiation between groups appears to be related to geographic distance between the groups, as seen most clearly in the Arctic and Mexico populations. Groups isolated by geography, such as the Arctic Eskimo and the Seri from Mexico, had the highest differentiation, while groups that have recently migrated out of the Northwest report low F_{ST} values. These F_{ST} values can help forensic investigators obtain random match probabilites or make inferences of ethnic orgin in casework samples. While the sampling from each tribe used in this study can be considered low (N = 21- 88), the use of a mimimum allele frequency estimates [26] would help compensate for limited sampling of infrequent alleles. Furthermore, if allele frequency estimates are not used because of the sample sizes, then the new F_{ST} value of 0.04 is recommended to compensate for effects of genetic differentiation across North American tribal populations when performing statistical calculations.

References

1. Wright S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution. 1965;19(3):395-420.
2. National Research Council. The Evaluation of Forensic DNA Evidence. Washington DC: The National Academies Press, 1996.
3. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci. [Research Support, U.S. Gov't, P.H.S.]. 2001 May;46(3):453-89.
4. Kanthaswamy S, Smith DG. Genetic and ethnohistoric evidence suggest current Native American population datasets in the FBI's CODIS database are not sufficiently representative. Forensic Sci Int Genet. 2014 Nov;13:e13-5.
5. Wang S, Lewis CM, Jr., Jakobsson M, Ramachandran S, Ray N, Bedoya G, et al. Genetic Variation and Population Structure in Native Americans. PLoS Genet. 2007;3(11):e185.
6. Eller E. Population substructure and isolation by distance in three continental regions. Am J Phys Anthropol. 1999 Feb;108(2):147-59.
7. Budowle B, Moretti T, Niezgoda S, Brown B. CODIS and PCR-based short tandem repeat loci: law enforcement tools. Proceedings of the Second European Symposium on Human Identification, 1998; Madison, WI. Madison, WI: Promega Corporation; 1998. p. 73-88.
8. Hares DR. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet. 2015 Jul;17:33-4.
9. Penaloza-Espinosa RI, Arenas-Aranda D, Cerda-Flores RM, Buentello-Malo L, Gonzalez-Valencia G, Torres J, et al. Characterization of mtDNA haplogroups in 14 Mexican indigenous populations. Hum Biol. 2007 Jun;79(3):313-20.
10. Driver HE. Indians of North America, Second Edition. Chicago: University of Chicago Press, 1969.
11. Lorenz JG, Smith DG. Distribution of four founding mtDNA haplogroups among Native North Americans. American Journal of Physical Anthropology. 1996;101(3):307-23.
12. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1(1):47-50.
13. Raymond M, Rousset F. GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. The Journal of heredity. 1995 May 1, 1995;86(3):248-9.
14. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour. 2008 Jan;8(1):103-6.
15. Glaubitz JC. convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Molecular Ecology Notes. 2004;4(2):309-10.
16. Budowle B, Chidambaram A, Strickland L, Beheim CW, Taft GM, Chakraborty R. Population studies on three Native Alaska population groups using STR loci. Forensic Science International. 2002 Sep 10;129(1):51-7.
17. Phillips C, Rodriguez A, Mosquera-Miguel A, Fondevila M, Porras-Hurtado L, Rondon F, et al. D9S1120, a simple STR with a common Native American-specific allele: forensic optimization, locus characterization and allele frequency studies. Forensic Sci Int Genet. 2008 Dec;3(1):7-13.
18. Schroeder KB, Schurr TG, Long JC, Rosenberg NA, Crawford MH, Tarskaia LA, et al. A private allele ubiquitous in the Americas. Biol Lett. 2007 Apr 22;3(2):218-23.
19. Rangel-Villalobos H, Sanchez-Gutierrez VM, Botello-Ruiz M, Salazar-Flores J, Martinez-Cortes G, Munoz-Valle JF, et al. Evaluation of forensic and anthropological potential of D9S1120 in Mestizos and Amerindian populations from Mexico. Croat Med J. 2012 Oct;53(5):423-31.
20. Yuasa I, Irizawa Y, Nishimukai H, Fukumori Y, Umetsu K, Nakayashiki N, et al. A hypervariable STR polymorphism in the complement factor I (CFI) gene: Asian-specific alleles. Int J Legal Med. 2011 Jan;125(1):121-5.
21. Aigner JS. Early Holocene Evidence for the Aleut Maritime Adaptation. Arctic Anthropology. 1976;13(2):32-45.
22. Fladmark KR. Routes: Alternate Migration Corridors for Early Man in North America. American Antiquity. 1979;44(1):55-69.

23. Snow MH, Durand KR, Smith DG. Ancestral Puebloan mtDNA in Context of the Greater Southwest. Journal of archaeological science. 2010 Jul 1;37(7):1635-45.
24. Nielsen R, Slatkin M. An Introduction to Population Genetics: Theory and Applications. Sunderland, Massachusetts: Sinauer Associates, Inc., 2013.
25. Budowle B, Monson KL, Chakraborty R. Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. Int J Legal Med. 1996;108(4):173-6.
26. Ng J, McCulloh KL, Oldt RF, Weise JA, Viray J, Budowle B, et al. Native American Population Data Based on the Globalfiler® autosomal STR loci. Forensic Sci Int Genet. submitted.
27. Malhi RS, Mortensen HM, Eshleman JA, Kemp BM, Lorenz JG, Kaestle FA, et al. Native American mtDNA prehistory in the American Southwest. Am J Phys Anthropol. 2003 Feb;120(2):108-24.

Tables

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 (23).

Geographic Region	Tribe	N	A	B	C	D	X	Ref
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	(23)
	Huichol	30	0.31	0.53	0.16	0	0	(23)
	Seri	29	0	0.13	0.86	0	0	(27)
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)

Table 2. Allele number (Na), observed (OH) and expected (EH) heterozygosities for each tribe and geographic sample. Estimates based on 21 randomly chosen samples parenthesized show that sample size has not affected the analyses significantly. * indicates tribal populations that conformed with HWE at $p < 0.01$ when all samples were included in the analyses. None of these populations deviated from HWE at $p < 0.01$ when 21 random samples from each population were analyzed.

Geographic Region	Tribe	N	Na	OH	EH
Arctic	Eskimo	44	6 (6)	0.68 (0.67)	0.69 (0.71)
Baja CA	Cochimi	25	7 (7)	0.75 (0.74)	0.75 (0.75)
CA/Great Basin	Miwok	33	8 (7)	0.78 (0.76)	0.76 (0.76)
Southeast	Cherokee	34	8 (8)	0.74 (0.75)	0.77 (0.77)
Mexico	Cora*	64	8 (6)	0.70 (0.68)	0.73 (0.72)
Mexico	Huichol*	30	6 (6)	0.70 (0.69)	0.70 (0.71)
Mexico	Seri*	29	6 (5)	0.66 (0.67)	0.64 (0.64)
Midwest	Chippewa*	21	7 (7)	0.77 (0.77)	0.76 (0.76)
Southwest	Apache	88	8 (6)	0.73 (0.69)	0.73 (0.72)
Southwest	Yavapai*	50	8 (7)	0.74 (0.72)	0.73 (0.71)
Average estimates		*41.8 (21)*	*7.2 (6.5)*	*0.73 (0.71)*	*0.73 (0.73)*

Table 3. Private alleles observed in this study: Midwest (5), CA/Great Basin (5), Mexico (7), Southwest (9), and Southeast (10).

Locus	Size	Tribe (Geographic Region)	Frequency
vWA	21	Chippewa (Midwest)	0.024
CSF1PO	12.1	Apache (Southwest)	0.006
TPOX	6	Cherokee (Southeast)	0.030
TPOX	7	Cherokee (Southeast)	0.015
D21S11	24.2	Miwok (CA/Great Basin)	0.015
D21S11	27	Miwok (CA/Great Basin)	0.046
D21S11	29.2	Apache (Southwest)	0.011
D21S11	35.2	Yavapai (Southwest)	0.010
D18S51	9	Cherokee (Southeast)	0.030
D18S51	10	Cherokee (Southeast)	0.015
D18S51	11.2	Apache (Southwest)	0.006
D18S51	13.2	Cherokee (Southeast)	0.015
D18S51	23	Huichol (Mexico)	0.017
D2S441	12.3	Cherokee (Southeast)	0.015
D19S433	11	Yavapai (Southwest)	0.010
D19S433	17	Miwok (CA/Great Basin)	0.015
TH01	10.3	Cora (Mexico)	0.016
FGA	17	Cora (Mexico)	0.008
FGA	22.2	Miwok (CA/Great Basin)	0.030
FGA	26.2	Chippewa (Midwest)	0.024
FGA	29	Cora (Mexico)	0.008
D22S1045	10	Cherokee (Southeast)	0.015
D22S1045	12	Chippewa (Midwest)	0.024
D7S820	15	Cherokee (Southeast)	0.015
SE33	11	Cora (Mexico)	0.008
SE33	12	Yavapai (Southwest)	0.010
SE33	13.2	Yavapai (Southwest)	0.020
SE33	15.2	Apache (Southwest)	0.017
SE33	24	Cherokee (Southeast)	0.016
SE33	30	Apache (Southwest)	0.006
D10S1248	10	Chippewa (Midwest)	0.024
D1S1656	10	Cherokee (Southeast)	0.030
D1S1656	14.3	Chippewa (Midwest)	0.024
D1S1656	19	Seri (Mexico)	0.035
D12S391	17.3	Miwok (CA/Great Basin)	0.046
D12S391	19.3	Cora (Mexico)	0.008

Table 4. Pairwise and population specific F_{ST} and F_{IS} based on the 22 autosomal STR loci in the seven geographic samples. Estimates based on 21 randomly chosen samples are above the diagonal. The overall F-statistics for all populations are F_{IS} = 0.006 (0.014), F_{ST} = 0.039 (0.041), and F_{IT} = 0.045 (0.056), where parenthesized values are estimates based on the 21 random samples.

Tribe (Geographic Region)	Eskimo (Arctic)	Cochimi (Baja California)	Miwok (CA/Great Basin)	Cherokee (Southeast)	Cora (Mexico)	Huichol (Mexico)	Seri (Mexico)	Chippewa (Midwest)	Apache (Southwest)	Yavapai (Southwest)	F_{ST}	F_{IS}
Eskimo		0.057	0.066	0.052	0.074	0.083	0.090	0.029	0.045	0.051	0.074 (0.061)	0.017 (0.07)
Cochimi	0.073		0.019	0.017	0.027	0.038	0.067	0.024	0.027	0.028	0.034 (0.034)	0.002 (0.02)
Miwok	0.076	0.018		0.015	0.036	0.039	0.087	0.026	0.034	0.051	0.040 (0.041)	-0.029 (0)
Cherokee	0.064	0.016	0.012		0.036	0.042	0.076	0.019	0.032	0.032	0.035 (0.036)	0.034 (0.03)
Cora	0.072	0.020	0.029	0.026		0.019	0.052	0.029	0.025	0.030	0.032 (0.036)	0.040 (0.04)
Huichol	0.101	0.038	0.046	0.043	0.020		0.068	0.040	0.036	0.046	0.048 (0.046)	0.008 (0.02)
Seri	0.113	0.068	0.087	0.079	0.050	0.067		0.055	0.050	0.052	0.070 (0.066)	-0.043 (-0.05)
Chippewa	0.046	0.022	0.026	0.018	0.022	0.039	0.061		0.018	0.021	0.029 (0.029)	-0.020 (-0.02)
Apache	0.061	0.023	0.029	0.026	0.022	0.036	0.057	0.016		0.012	0.031 (0.031)	0.006 (0.04)
Yavapai	0.058	0.022	0.037	0.027	0.023	0.044	0.052	0.014	0.006		0.032 (0.036)	-0.011 (-0.02)

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.