

# Land, Language, and Loci: mtDNA in Native Americans and the Genetic History of Peru

Cecil M. Lewis, Jr.,<sup>1</sup> Raúl Y. Tito,<sup>2</sup> Beatriz Lizárraga,<sup>2</sup> and Anne C. Stone<sup>1\*</sup>

<sup>1</sup>*Department of Anthropology, University of New Mexico, Albuquerque, New Mexico 87131*

<sup>2</sup>*Centro de Investigación de Bioquímica y Nutrición, Universidad Nacional Mayor de San Marcos, Lima, Peru*

**KEY WORDS** mitochondrial DNA; genetic distance; American Indian

**ABSTRACT** Despite a long history of complex societies and despite extensive present-day linguistic and ethnic diversity, relatively few populations in Peru have been sampled for population genetic investigations. In order to address questions about the relationships between South American populations and about the extent of correlation between genetic distance, language, and geography in the region, mitochondrial DNA (mtDNA) hypervariable region I sequences and mtDNA haplogroup markers were examined in 33 individuals from the Department of Ancash, Peru. These sequences

were compared to those from 19 American Indian populations using diversity estimates, AMOVA tests, mismatch distributions, a multidimensional scaling plot, and regressions. The results show correlations between genetics, linguistics, and geographical affinities, with stronger correlations between genetics and language. Additionally, the results suggest a pattern of differential gene flow and drift in western vs. eastern South America, supporting previous mtDNA and Y chromosome investigations. *Am J Phys Anthropol* 127:351–360, 2004. © 2004 Wiley-Liss, Inc.

Mitochondrial DNA (mtDNA) data has been used extensively to understand the population history of humans (e.g., Vigilant et al., 1991; Kittles et al., 1999; Oota et al., 2002), and, in the Americas, was the first line of molecular genetic evidence to support the hypothesis that the original migration into the New World was both Asiatic and from a relatively small group of hunter-gatherers (Wallace et al., 1985). Recent studies determined that modern Native American populations are represented by at least four major mtDNA haplogroups (Schurr et al., 1990; Torroni et al., 1992, 1993a, 1994; Horai et al., 1993). There is also evidence for a fifth haplogroup, X, which is less prevalent (Baillet et al., 1994; Brown et al., 1998; Mahli and Smith, 2002).

The four major haplogroups were thought by some to represent separate founding populations (Torroni et al., 1992, 1993b; Horai et al., 1993; Shields et al., 1993), but others argued that they represent the variation within a single founding population (Merriwether et al., 1995; Kolman et al., 1996; Bonatto and Salzano, 1997; Stone and Stoneking, 1998). Moreover, the number of migrations and the patterning of settlement in South America have also been important issues, but have yet to be resolved. Moraga et al. (2000) used mtDNA to argue that South America was populated by a single migration, while others proposed two migrations (Lalueza et al., 1997). Genetic studies of both classical and Y-chromosome markers found complicated patterns of genetic variation in South America. These patterns may be the result of small population sizes and thus high genetic drift in parts of South America (partic-

ularly among groups in Amazonia) as well as other complex demographic processes (O'Rourke and Suarez, 1985; Salzano and Callegari-Jacques, 1988; Black, 1991; Cavalli-Sforza et al., 1994; Luiselli et al., 2000; Tarazona-Santos et al., 2001; Fagundes et al., 2002). Tarazona-Santos et al. (2001) proposed a model of occupation involving differential gene flow and drift in eastern and western South America based on Y-chromosome data, although this model may be complicated by the difficulties of sampling (Rothhammer and Moraga, 2001). The major evidence for the model of Tarazona-Santos et al. (2001) is the high level of genetic diversity yet low level of genetic divergence in Andean populations relative to eastern South American populations. This model is also supported by mtDNA data (Fuselli et al., 2003).

In other molecular genetic studies, anthropological researchers attempted to correlate DNA signatures to mating practices (Dipierrri et al., 1998), environmental and cultural diversity (Tarazona-

Grant sponsor: University of New Mexico Research Allocation Grant.

\*Correspondence to: Anne C. Stone, Department of Anthropology, PO Box 872402, Arizona State University, Tempe, AZ 85287-2402. E-mail: acstone@asu.edu

Received 4 August 2003; accepted 30 December 2003.

DOI 10.1002/ajpa.20102  
Published online 6 December 2004 in Wiley InterScience (www.interscience.wiley.com).

Santos et al., 2001), mortuary and spatial patterns (Stone, 1996), and linguistics (Greenberg et al., 1986; Cavalli-Sforza et al., 1988; Sokal, 1988; Poloni et al., 1997). At a regional level, genetic differences between populations arise through processes of genetic drift and gene flow which are influenced by population size, by linguistic and cultural differences, and by geographic distances and barriers. In particular, debate has centered on the relative importance of geography and language for influencing genetic differentiation between populations in different regions of the world (e.g., Barbujani et al., 1997; Poloni et al., 1997; Zerjal et al., 2001).

To date, the investigation of the genetic history and diversity of Peru has been limited primarily to classical genetic markers and to loci that affect high-altitude adaptation (Modiano et al., 1972; Garruto and Hoff, 1976; Salzano and Callegari-Jacques, 1988; Salzano et al., 1997; Rupert et al., 1999; Tarazona-Santos et al., 2001). The topographical and ecological diversity, and the linguistic diversity, of native groups in Peru provide an excellent arena to examine the relative importance of these factors in the patterning of genetic diversity. Today, approximately 50 native languages falling into at least 18 distinct language families are spoken in Peru (Tamayo, 1994; Pozzi-Escot, 1998). Of these, Quechua is spoken by the largest number of people. Several dialects of Quechua (many of which are not mutually intelligible) exist throughout the country, with a total of approximately 3.5 million speakers. Campbell (1997) divided the Quechuan family into two main groups: Central Quechua (including the Departments of Ancash, Huánuco, Junín, Pasco, and parts HVI of the Department of Lima), and Peripheral Quechua (all others).

This paper focuses on mtDNA hypervariable region 1 (HV I) sequences and mtDNA haplogroup marker data from individuals from the Department of Ancash, Peru to begin to examine the influence of language and geography on genetic diversity in the Andes, as well as to investigate the general patterns of diversity in South America.

## MATERIALS AND METHODS

### Sampling

Blood samples were collected with informed consent from 33 unrelated individuals currently living in Lima who originated from 19 villages (most were from the three villages of Cabana, Chacas, and Independencia) in the highlands in the department of Ancash (Fig. 1). Sequences from 19 other populations examined in this study were obtained from published sources (Ward et al., 1991, 1993; Ward, 1996; Ginther et al., 1993; Shields et al., 1993; Torroni et al., 1993a; Santos et al., 1994; Batista et al., 1995; Kolman et al., 1995; Rickards et al., 1999; Kittles et al., 1999; Fuselli et al., 2003). Populations included in these analyses were those in which the total number of sequences available equaled 20 or



Fig. 1. Map showing location of Department of Ancash in Peru.

more and were free of major sequencing errors. The geographical locations of these populations are shown in Figure 2. The San Martin de Pangoa population from Fuselli et al. (2003) was not used, since that sample is comprised of a mix of Quechua and Nomatsiguenga speakers.

### DNA extraction and mtDNA analysis

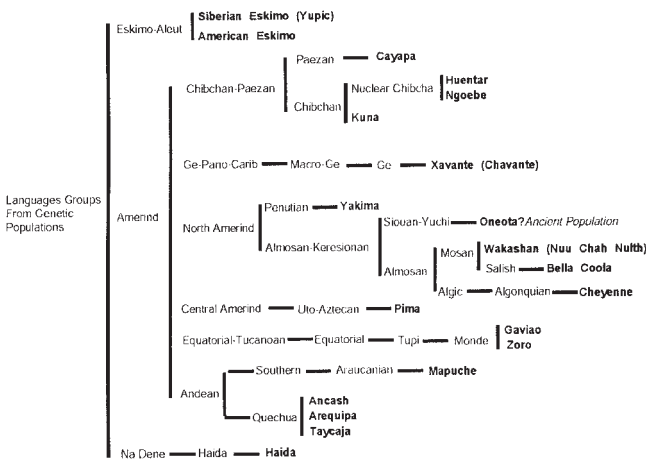
DNA was obtained from blood samples using a standard phenol/chloroform extraction in the Laboratorio de Biología Molecular at the Universidad Nacional Mayor de San Marcos. A portion of the purified DNA was sent to the Laboratory of Molecular Anthropology at the University of New Mexico. Markers characteristic of the four major Native American mtDNA haplotypes (A–D) were amplified using PCR and typed using the appropriate restriction enzyme (*Hae*III, *Hinc*II, or *Alu*I) or by size (9 base-pair deletion), as described previously (Stone and Stoneking, 1998). The mtDNA HV I region was amplified using PCR primers L15996 and H16401 and sequenced in both directions using the BigDye protocol from Applied Biosystems™ on an ABI 377.

### ANALYTICAL METHODS

The 9-bp deletion and three restriction sites were recorded as present or absent. These data were then



**Fig. 2.** Map of Americas, showing location of Native American populations included in study. Circles delineate geographical groupings. Line between Cheyenne and Pima distinguishes one of nine-group analyses (9a), which joined these populations, and 10-group analysis that kept them separated. Another nine-group analysis was conducted based on grouping Haida with Bella Coola, Yakima, and Wakashan, and also grouping together Pima and Cheyenne (9b).



**Fig. 3.** Language tree based on Greenberg (1987).

used to categorize each individual into one of the four major Native American haplogroups. Haplogroup diversity within populations was estimated using the formula

$$h = \left( \frac{n}{n-1} \right) \left( 1 - \sum_i p_i^2 \right)$$

where  $p_i$  is the sample frequency of the  $i$ th haplotype, and  $n$  is the number of individuals in the sample (Nei, 1987).

Ancash sequence data were edited using DNASTAR software (DNASTAR, Inc.). Nucleotide diversity ( $\pi$ )

for the Ancash and other populations was estimated by the formula

$$\pi = \frac{\sum_{i=j}^k \sum_{j<i} p_i p_j \hat{d}_{ij}}{L}$$

(Nei, 1987), while haplotype diversity was estimated in the same manner as haplogroup diversity.

The distance matrix used for the ordination methods was generated from a population pairwise  $\Phi_{st}$  analysis, using the Tamura-Nei (Tamura and Nei, 1993) distance method in Arlequin 2.0 (Schneider, 2000) with 1,000,000 permutations. In the final analysis, an  $\alpha$  parameter of 0.26 was used, as suggested by Meyer et al. (1999), but  $\alpha$  values ranging from 0.20–0.47 were also explored, all of which gave the same results. The multidimensional scaling (MDS) analysis was conducted in Statistica 6.0 (StatSoft, 2001).

The population pairwise  $\Phi_{st}$  values calculated using Arlequin 2.0 (Schneider, 2000) for the MDS plot were also used in a regression comparing genetic distance to geographical and linguistic distance. Geographic distances were calculated using ArcView GIS 3.1 (Environmental Systems Research Institute, Inc., 1998). Linguistic distance was calculated based on an ordinal scale using the language tree in Figure 3. If the differences were at the scale of Amerind to Eskimo-Aleut or Na Dene, a distance of 10 was given. If the differences were at the scale of Chibchan-Paezan to Ge-Pano-Carib, North Amerind, Central Amerind, Equatorial-Tucanoan, or Andean, a distance of 5 was given. If the difference was within one of these language groupings, a distance of 1, 2, or 3 was given, based on their proximity in the tree. The American Eskimo and Yupic pair was given a linguistic difference of 1. Because linguistic families do not easily translate into numerical representations of distance, we attempted to be conservative in these estimates.

Using distance data for regressions violates the assumption of independence (Koenig, 1999). To evaluate the degree to which the distance data might bias the regression toward significance, both the geographic and linguistic distances were randomly paired to the genetic distances 100 times without replacement to get the distribution of  $P$ -values. This is consistent with the Mantel test randomization procedure.

Analyses of molecular variance approaches (AMOVAs) were conducted in Arlequin 2.0 (Schneider et al., 2000). An AMOVA estimate was performed by comparing variation within populations, among populations within language groups, and among language groups. Language groups were largely based on classifications by Greenberg (1987) (Fig. 3). Greenberg categorized Ancash (from the Huaylas area) as a dialect of Quechua. This classification is in agreement with other research (Campbell, 1997).

Three other AMOVA estimates were calculated from geographical groupings. Since there were nine language groups, initially nine different geographic groups were delineated so as to make them comparable. The nine-group analysis was done in two different ways (see 9a and 9b in Fig. 2). Other AMOVA estimates were calculated for 10 geographic groups, to evaluate the persistence of the patterns observed.

Mismatch distributions were calculated using Arlequin 2.0 (Schneider et al., 2000) for all populations as well as for groups based on language affiliation. The prehistoric Norris Farms Oneota were not included when grouping the Almosan populations, since their affiliation with this group is arguable. Mismatch distributions present the frequency or count of nucleotide difference between each pair of individuals in a group, and display the frequency in a histogram or scatterplot. It is argued that histograms with multimodal distributions are characteristic of populations in equilibrium, and that unimodal distributions are characteristic of populations that have undergone population growth or allele fixation (Rogers and Harpending, 1992; Rogers, 1995; Rogers et al., 1996). This assumes that the sample of the population was representative. Small sample sizes can give a false pattern that a population is multimodal. This can be a significant problem for calculating mismatch distributions for Native American mtDNA haplogroups. Since Native American mtDNA is defined by four major haplogroups, peaks at high values on the X-axis would signify the difference between these haplogroups, i.e., the intermatch between A, B, C, and D. Thus, examining haplogroups separately may be more appropriate. However, this can significantly reduce the sample size used to make the mismatch distribution. There are two solutions to these problems. First, mismatch distributions can be calculated that include all haplogroups, but it must be remembered to note that the expected peaks at high nucleotide differences are the intermatch distributions of the haplogroups. Second, those haplogroups that are abundant in each population can be examined separately in the hopes that they are representative. Both approaches were used in this investigation.

## RESULTS

All 33 individuals sampled had mtDNA belonging to one of the four major Native American haplogroups, as indicated both by the pattern of the four characteristic markers and by characteristic polymorphisms in the HVI region. Haplogroups B (52%) and D (21%) were most common in the Ancash, although haplogroups A (9.1%) and C (18%) were also present (Table 1). The Ancash haplogroup diversity estimate ( $h$ ) was 0.67.

Forty-three polymorphic sites were identified in the Ancash mtDNA HV I sequences when compared to the reference sequence (Anderson et al., 1981; Andrews et al., 1999) (Table 2). Ancash haplotype diversity ( $h = 0.966$ ) was one of the highest of the

populations included in this analysis (Table 3). The Mapuche, Arequipa, Tayacaja, Pima, Norris Farms Oneota, Cheyenne, Nuu Chah Nulth, Bella Coola, and American Eskimo also had high levels of haplotype diversity ( $>0.9$ ). The estimate of Ancash nucleotide diversity ( $\pi = 0.018$ ) was also high, although three populations, the Cheyenne, Pima, and Cayapa, had higher values. Of the 27 haplotypes found in the Ancash population, 19 were not reported previously (70%). This frequency, although high, is not uncommon in Native American populations. Similar or higher percentages were observed in all studied Andean populations, such as the Arequipa (67%), Tayacaja (71%), and Mapuche (79%), as well as North American populations of the Cheyenne (79%), Eskimo (70%), and Pima (86%). Additionally, none of the population data have statistically significant Tajima D-values.

The MDS plot in this analysis displays the genetic distances between populations in three-dimensional space. The plot had a very low stress value of 0.0580 (Fig. 4), indicating that these data fit well in three dimensions. Although the graphical representations do cluster populations of similar linguistic heritage, the distributions can also be interpreted as geographical, with northern populations tending toward the left side of Figure 4 and southern populations tending toward the right. The Ancash, Arequipa, Tayacaja, and Mapuche are in close proximity to one another in the MDS plot. However, the Pima and Yakima are also close to these Andean populations.

The regression comparing population pairwise  $\Phi_{st}$  estimates to geographic and linguistic distance was significant ( $P < 0.001$ ) in both cases (Fig. 5). When the population pairwise  $\Phi_{st}$  was compared to geographic and linguistic distance, the  $r^2$  values equaled 0.1128 and 0.3051, respectively. After 100 randomizations, no significant  $P$ -values were found at an alpha of 0.05, suggesting that the original  $P$ -value was valid.

All AMOVA estimates found the largest percentage of variation present within populations ( $\sim 74\%$ ) rather than among populations and among groups in all analyses. The percentage of variation among language groups was 19.37%, and the variation among populations within language groups was 5.89% (Table 4). The variation observed among geographic regions was 11.40% and 11.52% when grouped into nine areas (9a and 9b, respectively), and 14.68% when grouped into 10 areas; the variation among populations within geographic regions was 13.29% and 13.32% when grouped into nine areas (9a and 9b, respectively), and 10.05% when grouped into 10 areas.

In the mismatch analyses using all haplogroups, the American Eskimo distribution was slightly bimodal. A pronounced bimodal distribution was observed in the populations of Cheyenne, Nuu Chah Nulth, Ngoebe, Kuna, Gaviao, Huentar, Norris Farms Oneota, Ancash, Tayacaja, Arequipa, Zoro,

TABLE 1. Native American haplogroup frequencies and diversity estimates

Population	mtDNA lineage cluster (%)						
	n	A	B	C	D	N	h
Columbia							
Embera	22	73.0	23.0	0.0	0.0	5.0	0.43
Ingano	27	15.0	44.0	37.0	0.0	4.0	0.67
Piaroa	10	50.0	0.0	10.0	40.0	0.0	0.64
Ticuna	54	13.0	15.0	39.0	33.0	0.0	0.71
Wayuu	40	25.0	35.0	38.0	0.0	3.0	0.69
Zenu	37	19.0	41.0	30.0	5.0	5.0	0.72
Ecuador							
Cayapa	30	33.3	20.0	16.7	30.0	0.0	0.76
Venezuela							
Makiritare	10	20.0	0.0	70.0	10.0	0.0	0.51
Brazil							
Kraho	14	28.6	57.1	14.3	0.0	0.0	0.62
Macushi	10	10.0	20.0	30.0	40.0	0.0	0.78
Marubo	10	10.0	0.0	60.0	30.0	0.0	0.60
Ticuna	28	17.9	0.0	32.1	50.0	0.0	0.64
Wapishana	12	0.0	25	8.3	66.7	0.0	0.53
Yanomama	24	0.0	16.7	54.2	29.2	0.0	0.62
Peru							
Ancash	33	9.1	51.5	18.2	21.2	0.0	0.67
Arequipa	22	9.0	68.0	14.0	9.0	0.0	0.53
Tayacaja	61	21.0	33.0	13.0	30.0	3.0	0.75
Other Quechua	19	26.3	36.8	5.3	31.6	0.0	0.73
Bolivia							
Aymara	33	0.0	93.9	3.0	3.0	0.0	0.12
Quechua	32	15.6	75.0	9.4	0.0	0.0	0.42
Chimane	41	39.0	53.7	4.9	0.0	2.4	0.57
Mosetén	20	40.0	55.0	0.0	0.0	5.0	0.56
Ignaciano	22	18.2	36.4	40.9	0.0	4.5	0.70
Trinitario	35	14.3	40.0	37.1	2.9	5.7	0.70
Movima	22	9.1	9.1	63.6	18.2	0.0	0.57
Yuacaré	28	39.3	32.1	21.4	3.6	3.6	0.72
Chile							
Atacamenos	63	14.3	71.4	9.5	4.8	0.0	0.47
Atacamenos	50	12.0	72.0	10.0	6.0	0.0	0.46
Aymara	172	6.4	67.4	12.2	14.0	0.0	0.51
Huilliches	38	5.3	28.9	18.4	47.4	0.0	0.67
Huilliches	80	3.75	28.75	18.75	48.75	0.0	0.65
Mapuche	111	0.0	7.2	44.1	48.7	0.0	0.57
Peneunche	105	2.8	10.5	41.0	45.7	0.0	0.62
Peneunche	100	2.0	9.0	37.0	52.0	0.0	0.59
Yaghan	21	0.0	0.0	43.0	47.7	0.0	0.62
Argentina							
Choroti	20	15.0	40.0	30.0	15.0	0.0	0.74
Fueguian	45	0.0	0.0	42.0	56.0	2.0	0.52
Mapuche	58	5.3	31.0	20.6	29.3	10.3	0.78
Mapuche	97	8.4	33.7	22.1	28.4	7.4	0.75
Mataco (Chaco)	28	10.7	35.7	0.0	53.6	0.0	0.60
Mataco	72	5.6	62.4	2.8	26.4	2.8	0.54
Mataco (Formosa)	44	9.1	54.5	20.5	15.9	0.0	0.64
Pilagá (Formosa)	41	4.9	36.6	26.8	29.3	2.4	0.72
Quebrada de Humahuaca	46	10.9	67.4	17.4	4.3	0	0.51
San Salvador de Jujuy	19	15.8	57.9	15.8	10.5	0	0.64
Tehuelche	29	0.0	20.7	24.1	55.2	0	0.62
Toba	8	0.0	33.3	11.1	55.6	0	0.65
Toba (Chaco)	30	13.3	46.7	6.7	26.7	6.7	0.71
Toba (Formosa)	26	26.9	34.6	3.8	34.6	0	0.71

Cayapa, Yakima, and in the language groupings of the Almosan, Chibchan Paezan, Tupi Monde, and Andean. Multimodal patterns were observed for the Pima, Yupic, Mapuche, Bella Coola, Xavante, Haida, and the language group of Eskimo-Aleut. Within all these multimodal groups except the Haida, the presence of a peak at zero nucleotide differences is found. It is possible that this peak is a product of sampling closely related individuals or is a product of drift. If this peak is disregarded, then all of these populations except the Haida have a bimodal distribution

when including all haplogroups. All the mismatch distributions constructed from one abundant haplogroup and excluding the peak at zero nucleotide differences had a peak between 1–3 differences. Figure 6 shows both mismatch distributions from the Ancash population.

**DISCUSSION**

The Ancash haplogroup frequencies and haplogroup diversity estimates are similar to those of other South American groups in the region, suggest-

ing no major discontinuities. The sequence data for the Ancash have some of the highest amounts of variation and a high number of unique haplotypes, signifying no major episodes of drift or sweeps to fixation in recent history. The high mtDNA diversity of the Ancash matches expectations from Tarazona-Santos et al. (2001) and Fuselli et al. (2003) for

Andean populations. All the Andean populations had high levels of diversity and a close genetic relationship to each other.

All the analyses support some congruence between genetic distance and linguistic and/or geographic distance. The MDS plot fits largely with what is known about the linguistic and geographical

TABLE 2. Ancash mtDNA HV I sequence variation<sup>1</sup>

			0000001111	1111222222	2222222222	2222233333	333
			2578891246	7789001112	3345667889	9999912223	567
			4055621958	4692793473	5926064040	1236894576	721
Haplogroup	Haplotype	n	TTTCTTCGGC	CCTCATGCTC	ACCCCGAAC	CCACTGTTTCG	TTA
A	<b>An1</b>	1	?G...CT...	.....T	.....T	.....A....	.C.
	<b>An2</b>	1	?.....T...	.....T	....T...T	.....A....	.C.
	<b>An3</b>	1	.....T...	.....A..T	.....T	.T...A....	.C.
B	<b>An4</b>	2	.....	..C....C.	.....	T.....	...
	<b>An5</b>	1	G?C.....	..C....C.	.....A...	.....	...
	<b>An6</b>	1	.....	..C....C.	.....G..	.....	...
	<b>An7</b>	1	G..G.....	..C....C.	.....	.....	...
	<b>An8</b>	4	.?.....	..C....C.	.....	.....	...
	<b>An9</b>	1	?.....	..C....C.	.....G.	.....	..?
	<b>An10</b>	1	.....	..C....C.	G.....	...T.....	..G
	<b>An11</b>	1	...C.....	..C.G...C.	.....	.....A	...
	<b>An12</b>	1	.....	..C....C.	.T.....	.....	C..
	<b>An13</b>	1	.....	..C....C.	.....A...	.....	..?
	<b>An14</b>	1	.....	..C....C.	..T.....	.....C...	...
<b>An15</b>	1	.....	..C....TC.	.....	..G.....	...	
<b>An16</b>	1	.....	..C....C.	..T.....	.....	...	
C	<b>An17</b>	1	G.....	...T....T	....T....	....C..CT.	...
	<b>An18</b>	1	.....A..	.....T	.....	....C..C..	.C.
	<b>An19</b>	1	...C.....	.T.T....T	.....	....C..CT.	...
	<b>An20</b>	3	.....	.....T	.....	....C..CT.	...
D	<b>An21</b>	1	.....	T.....T	.....	....C..	.C.
	<b>An22</b>	1	.....	..C....T	.....	.T....C..	.C.
	<b>An23</b>	1	.....	.....C...T	.....	....C..	.C.
	<b>An24</b>	1	??.....T	.....T	.....T	....C..	.C.
	<b>An25</b>	1	.....	.....T	.....	....C..	.C.
	<b>An26</b>	1	.....A.	....C....	.....	....C..	.C.
	<b>An27</b>	1	.....	.....C....	.....	....C..	.C.

<sup>1</sup>Unique haplotypes in bold. Haplotype is defined as unique when not observed in other populations included in this study. The numbering (plus 16000) is as given in Anderson et al. (1981).

TABLE 3. Native American mtDNA HV I haplotype diversity estimates and estimates of Tajima's  $D^1$

Population	n	No. of haplotypes	Frequency of unique haplotypes	Gene diversity (h)	Nucleotide diversity ( $\pi$ )	Tajima's D
American Eskimo	49	27	0.704	0.914	0.010	-1.458
Yupic Eskimo	77	12	0.417	0.719	0.010	-0.596
Haida	41	10	0.200	0.709	0.008	-1.205
Bella Coola	40	11	0.364	0.904	0.015	0.064
Cheyenne	31	24	0.792	0.965	0.023	-0.744
Oneota	51	24	0.583	0.910	0.016	-0.783
Wakashan	63	27	0.667	0.952	0.017	-0.011
Yakima	42	20	0.700	0.893	0.015	-1.116
Pima	41	21	0.857	0.922	0.021	-1.090
Huetar	27	7	0.571	0.709	0.011	0.413
Kuna	63	7	0.286	0.592	0.011	1.519
Ngoebe	46	8	0.625	0.773	0.014	1.684
Cayapa	30	8	0.625	0.837	0.020	1.155
Xavante	25	4	0.250	0.677	0.009	0.439
Gaviao	27	7	0.429	0.866	0.014	0.107
Zoro	28	8	0.375	0.773	0.013	-0.140
Ancash	33	27	0.704	0.966	0.018	-1.267
Arequipa	22	18	0.667	0.978	0.017	-1.101
Tayacaja	61	42	0.714	0.968	0.025	-1.440
Mapuche	39	14	0.786	0.924	0.017	0.254

<sup>1</sup> Frequency of unique haplotypes was estimated by dividing number of unique haplotypes by total number of haplotypes.

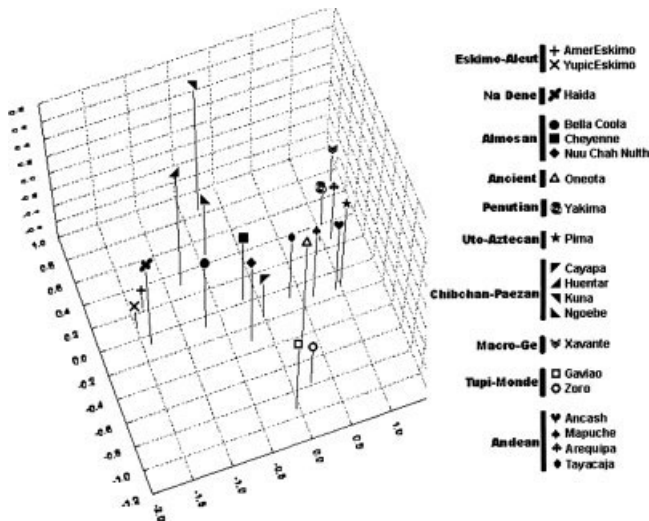


Fig. 4. MDS plot (stress = 0.0580) showing relationships of Native American mtDNA HV I data, using a Tamura-Nei distance method with gamma correction model of evolution ( $\alpha = 0.26$ ).

distance between these populations. Two notable exceptions are the placement of the Pima and Yakima near the Andean populations in the MDS plot. Although these exceptions could be a product of population events or a stress outlier in the MDS calculations, it also should be noted that the Pima and Yakima were the only representatives of their language group, suggesting that more samples from these two language families may be needed to refine the resolution of the MDS plot. The regression comparing the population pairwise  $\Phi_{st}$  to both geographic and linguistic evidence supports a strong correlation in both instances. In the AMOVA analyses, the higher percentage of variation represented within populations than among populations and groups is typical for human populations. The esti-

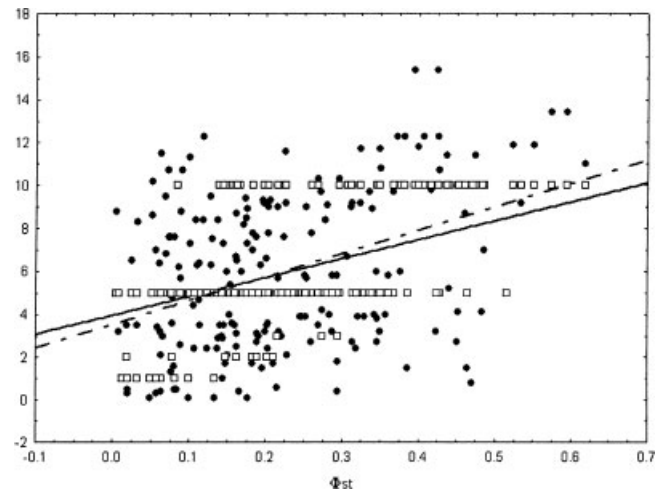


Fig. 5. Regressions comparing population pairwise  $\Phi_{st}$  to geographic distance and  $\Phi_{st}$  to linguistic distance. Regression of geographic distance to  $\Phi_{st}$  is represented by solid circles and solid line in units of 1,000 km, with  $r^2 = 0.1128$ ,  $P < 0.001$ , and  $y = 3.967 + 8.791 \cdot x$ . Regression of linguistic distance to  $\Phi_{st}$  is represented by squares with dashed line in units based on ordinal scale from 1-10, with  $r^2 = 0.3051$ ,  $P < 0.001$  and  $y = 3.561 + 10.922 \cdot x$ .

TABLE 4. AMOVA percentages of variation

	Percentage of variation among groups	Percentage of variation within groups
Nine language groups	19.37	5.89
Nine geographic areas (9a)	11.40	13.29
Nine geographic areas (9b)	11.52	13.32
Ten geographic areas	14.68	10.05

mates suggest that language differentiation often corresponds with population divergence. Variation observed in populations within geographic regions and among geographic regions is relatively equal, even when the populations are placed into 10 differ-

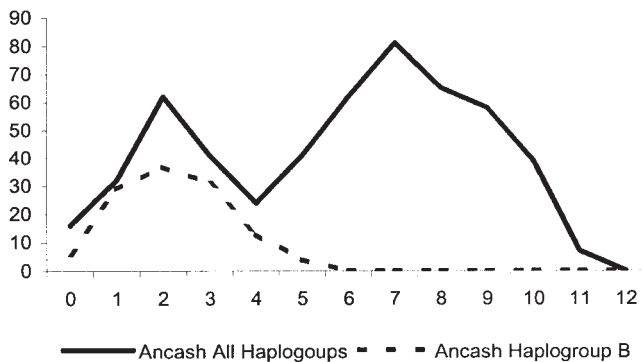


Fig. 6. Ancash mismatch distributions.

ent groups. These estimates are dependant on how the geographical and language groups are separated, but attempts were made to do so conservatively. These AMOVA analyses support language, more than geography, as an indicator of genetic affinities.

The mismatch distributions in this analysis support patterns observed in previous studies (Bonatto and Salzano, 1997; Stone and Stoneking, 1998). All populations have a bimodal pattern, with one peak at approximately two differences and the other at approximately seven differences when comparing all haplogroups, and a peak at around two differences when examining only one abundant haplogroup. In mismatch distributions, a multimodal pattern is argued to be evidence of a population in equilibrium, while a unimodal pattern is argued to be evidence of a population that underwent expansion (Rogers and Harpending, 1992; Rogers, 1995; Rogers et al., 1996). In Native American populations, the first peak likely represents the population expansion in the Americas (Bonatto and Salzano, 1997; Stone and Stoneking, 1998), and the latter is a product of the intermatch between haplogroups A, B, C, and D, which reflects the divergence of these haplogroups prior to the time of entry into the Americas.

The mtDNA sequence data from the Ancash and other South American populations have a distribution that is similar to Y-chromosome results by Tarazona-Santos et al. (2001) and mtDNA results by Fuselli et al., (2003). That is, the four Andean populations examined in this analysis have a great deal of genetic diversity and a very close genetic distance compared to other South American populations. These Ancash sequence data represent several villages in the Department of Ancash, which may explain the high levels of diversity in this instance. Also, in comparison to the Amazonian populations, Andean populations have maintained relatively large population sizes, allowing the retention of more diversity compared to smaller populations that are more subject to drift. However, the origin of the Ancash sample and the history of large population sizes in the Andes do not explain the degree of genetic similarity these Andean populations appear to

have to one another, given that they are separated by several thousand kilometers. Although the Andean mtDNA sample includes only four populations, these results tentatively corroborate the hypothesis of differential genetic drift as well as gene flow in eastern and western South America, specifically with regard to the Andes.

The genetic data support a correlation between genetic distance and language more than geography. The current pattern of genetic relatedness in the Andes could be a product of colonial Europe and Inca influences, or a more general pattern of Andean life throughout its prehistory. Numerous cultural traditions such as the Chimu, Chavin, Moche, Wari, Tiwanaku, and Nasca have had significant effects on Andean society. The genetic data may suggest that these cultural traditions did not spread primarily from the enculturation of indigenous/endogamous populations, but rather were significantly influenced by population migration. The degree to which Andean cultural systems influenced the observed genetic pattern in Peru remains to be seen. Additional data are required from other modern and ancient Andean populations in order to evaluate the consistency of the observed pattern and to develop a time frame for population events.

Several language groups were represented by a single population, and some of the groups themselves are controversial. For instance, the placement of the Haida in the Na Dene language group is hotly debated (Krauss, 1973; Cook and Rice, 1989; Pinnow, 1990). Linguists have also warned that a language phylogeny can be obscured within a matter of 8–10 or fewer millennia, which suggests that only closely related language groups should be examined phyletically (Kaufman and Golla, 2000). This admonition from linguists emphasizes the need for more sampling of Native American populations within language groups on both continents.

There is no single catalyst for genetic population divergence. Geographical, linguistic, environmental, cultural, biological, and historical phenomena all influence a population's genetic pattern. Yet it is possible to evaluate the degrees of influence of these mechanisms. When looking at the mtDNA HV I region, the data support a level of congruence of genetic distance to both geographical and linguistic distance, with evidence pointing to linguistic distance as a more important indicator. In addition, the comparison of the mtDNA HV I data from South American populations shows similar patterns of differential gene flow and drift in eastern vs. western South America, as observed in previous Y-chromosome and mtDNA studies.

#### ACKNOWLEDGMENTS

We thank the people who participated in this study. We also thank Jada Benn, GianCarlo Iannaccone, Hsiuman Lin, Paul López, Kenneth Nystrom, Erica Tyler, Alicia Wilbur, and Gregory Zaro for

helpful discussions. We also thank Gregory Zaro for translation assistance.

### LITERATURE CITED

- Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465.
- Andrews R, Kubacka I, Chinnery P, Lightowlers R, Turnbull D, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23:147.
- Bailliet G, Rothhammer F, Carnese FR, Bravi CM, Bianchi NO. 1994. Founder mitochondrial haplotypes in Amerindian populations. *Am J Hum Genet* 55:27–33.
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza. 1997. An apportionment of human DNA diversity. *Proc Natl Acad Sci USA* 94:4516–4519.
- Batista O, Kolman CJ, Bermingham E. 1995. Mitochondrial DNA diversity in the Kuna Amerinds of Panama. *Hum Mol Genet* 4:921–929.
- Black FL. 1991. Reasons for failure of genetic classifications of South Amerind populations. *Hum Biol* 63:763–774.
- Bonatto SL, Salzano FM. 1997. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proc Natl Acad Sci USA* 94:1866–1871.
- Brown MD, Hosseini SH, Torroni A, Bandelt H-J, Allen JC, Schurr TG, Scozzari R, Cruciani F, Wallace DC. 1998. MtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? *Am J Hum Genet* 63:1852–1861.
- Campbell L. 1997. *American Indian languages: the historical linguistics of Native America*. Oxford: Oxford University Press.
- Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological and linguistic data. *Proc Natl Acad Sci USA* 85:6002–6006.
- Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The history and geography of human genes*. Princeton: Princeton University Press.
- Cook E, Rice K. 1989. Introduction. In: Cook E, Rice K, editors. *Athapaskan linguistics. Current perspectives on a language family*. Berlin: Mouton de Gruyter. p 1–61.
- Dipierrri JE, Alfaro E, Martínez-Marignac VL, Bailliet G, Bravi CM, Cejas S, Bianchi NO. 1998. Paternal directional mating in two Amerindian subpopulations located at different altitudes in northwestern Argentina. *Hum Biol* 70:1001–1010.
- Environmental Systems Research Institute, Inc. 1998. *ArcView GIS 3.1*. Redlands, CA; Environmental Systems Research Institute, Inc.
- Fagundes NJR, Bonatto SL, Callegari-Jacques SM, Salzano FM. 2002. Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol* 117:68–78.
- Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean Highlanders. *Mol Biol Evol* 20:1682–1691.
- Garruto RM, Hoff CJ. 1976. Genetic history and affinities. In: Baker P, Little M, editors. *Man in the Andes: a multidisciplinary study of high-altitude Quechua*. Stroudsburg, PA: Dowden, Hutchinson and Ross, Inc. p 98–114.
- Ginther C, Corach D, Penacino GA, Rey JA, Carnese FR, Hutz MH, Anderson A, Just J, Salzano FM, King M-C. 1993. Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. In: Pena S, Chakraborty R, Epplen J, Jeffreys A, editors. *DNA fingerprinting: state of the science*. Basel: Birkhäuser Verlag. p 211–219.
- Greenberg JH. 1987. *Language in the Americas*. Stanford: Stanford University Press.
- Greenberg JH, Turner CG, Zegura SL. 1986. The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. *Curr Anthropol* 27:477–497.
- Horai S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, Tajima K. 1993. Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Mol Biol Evol* 10:23–47.
- Kaufman T, Golla V. 2000. Language groupings in the New World: their reliability and usability in cross-disciplinary studies. In: Renfrew C, editor. *America past, America present: genes and language in the Americas and beyond*. Cambridge: MacDonald Institute for Archaeological Research.
- Kittles RA, Bergen AW, Urbanek M, Virkkunen M, Linnoila M, Goldman D, Long JC. 1999. Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am J Phys Anthropol* 108:381–399.
- Koenig WD. 1999. Spatial autocorrelation of ecological phenomena. *Trends Ecol Evol* 14:22–25.
- Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F. 1995. Reduced mtDNA diversity in the Ngöbé Amerinds of Panama. *Genetics* 140:275–283.
- Kolman CJ, Sambuughin N, Bermingham E. 1996. Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics* 142:1321–1334.
- Krauss M. 1973. Na-Dene. In: Sebeok T, editor. *Linguistics in North America II*. Mouton: The Hague. p 903–978.
- Lalueza C, Pérez-Pérez A, Prats E, Cornudella L, Turbón D. 1997. Lack of founding Amerindian mitochondrial DNA lineages in extinct aborigines from Tierra del Fuego-Patagonia. *Hum Mol Genet* 6:41–46.
- Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D. 2000. Genetic structure of Quechua-speakers of the central Andes and geographic patterns of gene frequencies in South Amerindian populations. *Am J Phys Anthropol* 113:5–17.
- Malhi RS, Smith D. 2002. Brief communication: haplogroup X confirmed in prehistoric North America. *Am J Phys Anthropol* 119:84–86.
- Merriwether DA, Ferrell RE, Rothhammer F. 1995. MtDNA D-loop 6-bp deletion found in Chilean Aymara: not a unique marker for Chibcha-speaking Amerindians. *Am J Hum Genet* 56:812–813.
- Meyer S, Weiss G, von Haeseler A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 152:1103–1110.
- Modiano G, Bernini L, Carter ND, Santachiara Benerecetti SA, Detter JC, Baur EW, Paolucci AM, Gigliani F, Morpurgo G, Santolamazza C, Scozzari R, Terrenato L, Meera Khan P, Nijenhuis LE, Kanashiro VK. 1972. A survey of several red cell and serum genetic markers in a Peruvian population. *Am J Hum Genet* 24:111–123.
- Moraga ML, Rocco P, Miquel JF, Nervi F, Llop E, Chakraborty R, Rothhammer F, Carvallo P. 2000. Mitochondrial DNA polymorphisms in Chilean aboriginal populations: implications for the peopling of the southern cone of the continent. *Am J Phys Anthropol* 113:19–29.
- Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.
- Oota H, Kitano T, Jin F, Yuasa I, Wang L, Ueda S, Saitou N, Stoneking M. 2002. Extreme mtDNA homogeneity in continental Asian populations. *Am J Phys Anthropol* 118:146–153.
- O'Rourke DH, Suarez BK. 1985. Patterns and correlates of genetic variation in South Amerindians. *Ann Hum Biol* 13:13–31.
- Pinnow H. 1990. *Die Na-Dene-Sprachen im Lichte der Greenberg-Klassifikation*. Nortorf: Volkerkundliche Arbeitsgemeinschaft.
- Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L. 1997. Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 61:1015–1035.
- Pozzi-Escot I. 1998. *El multilingüismo en el Peru*. Cuzco: Biblioteca de la Tradición Oral Andina.
- Rickards O, Martínez-Labarga C, Lum JK, De Stefano GF, Cann RL. 1999. MtDNA history of the Cayapa Amerinds of Ecuador:

- detection of additional founding lineages for the Native American populations. *Am J Hum Genet* 65:519–530.
- Rogers A, Harpending H. 1992. Population growth makes waves in the distribution of pairwise differences. *Mol Biol Evol* 9:552–569.
- Rogers AB, Fraley AE, Bamshad MJ, Watkins WS, Jorde LB. 1996. Mitochondrial mismatch analysis is insensitive to the mutational process. *Mol Biol Evol* 13:895–902.
- Rogers AR. 1995. Genetic evidence for a pleistocene population explosion. *Evolution* 49:608–615.
- Rothhammer F, Moraga M. 2001. Patterns of Y-chromosome variation in South Amerindians. *Am J Hum Genet* 69:904.
- Rupert JL, Devine DV, Monsalve MV, Hochachka PW. 1999. Angiotensin-converting enzyme ACE alleles in the Quechua, a high altitude South American native population. *Ann Hum Biol* 26:375–380.
- Salzano FM, Callegari-Jacques SM. 1988. South American Indians: a case study in evolution. Oxford: Clarendon Press.
- Salzano FM, Bonatto SL, Callegari-Jacques SM. 1997. Genetic variability in Andean and non-Andean populations and its interpretations. In: Barton S, Rothhammer F, Schull W, editors. Patterns of morbidity in Andean aboriginal populations: 8000 years of evolution. Santiago de Chile: AmpHora Editores. p 14–31.
- Santos M, Ward RH, Barrantes R. 1994. MtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. *Hum Biol* 66: 963–977.
- Schneider S, Roessli D, Excoffier L. 2000. Arlequin version 2.000: a software for population genetics data analysis. Geneva: Genetics and Biometry Laboratory, University of Geneva.
- Schurr TG, Ballinger SW, Gan Y-Y, Hodge JA, Merriwether DA, Lawrence DN, Knowler WC, Weiss KM, Wallace DC. 1990. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary lineages. *Am J Hum Genet* 46:613–623.
- Shields GF, Schmiechen AM, Frazier BL, Redd A, Voevoda MI, Reed JK, Ward RH. 1993. MtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *Am J Hum Genet* 53:549–562.
- Sokal RR. 1988. Genetic, geographical, and linguistic distances in Europe. *Proc Natl Acad Sci USA* 85:1722–1726.
- StatSoft, Inc. 2001. Statistica data analysis software system.
- Stone AC. 1996. Genetic and mortuary analyses of a prehistoric Native American community. Ph.D. dissertation. University Park: Pennsylvania State University.
- Stone AC, Stoneking M. 1998. MtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. *Am J Hum Genet* 62:1153–1170.
- Tamayo JLP. 1994. Heterogeneidad étnico-lingüística del Perú: magistri et doctores. San Marcos: Universidad Nacional Mayor de San Marcos. p 24–29.
- Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.
- Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Martinez Labarga C, Rickards O, Tyler-Smith C, Pena SDJ, Santos FR. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. *Am J Hum Genet* 68: 1485–1496.
- Torrioni A, Schurr TG, Yang C-C, Szathmary E, Williams RC, Schanfield MS, Troup GA, Knowler WC, Lawrence DN, Weiss KM, Wallace DC. 1992. Native American mitochondrial DNA analysis indicates that the Amerindian and the Nadene populations were founded by two independent migrations. *Genetics* 130:153–162.
- Torrioni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC. 1993a. Asian affinities and continental radiation of the four founding Native Americans mtDNAs. *Am J Hum Genet* 53:563–590.
- Torrioni A, Sukernik RI, Schurr TG, Starikovskaya YB, Cabell MF, Crawford MH, Comuzzie AG, Wallace DC. 1993b. MtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am J Hum Genet* 53:591–608.
- Torrioni A, Neel JV, Barrantes R, Schurr TG, Wallace DC. 1994. Mitochondrial DNA “clock” for the Amerinds and its implications for timing their entry into North America. *Proc Natl Acad Sci USA* 91:1158–1162.
- Vigilant L, Stoneking M, et al. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507.
- Wallace DC, Garrison K, Knowler WC. 1985. Dramatic founder effects in Amerindian mitochondrial DNAs. *Am J Phys Anthropol* 68:149–155.
- Ward RH. 1996. Linguistic divergence and genetic evolution: a molecular perspective from the New World. In: Boyce A, Mascie-Taylor C, editors. Molecular biology and human diversity. Cambridge: Cambridge University Press, pp. 205–224.
- Ward RH, Frazier B, Dew-Jager K, Pääbo S. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci USA* 88:8720–8724.
- Ward RH, Redd A, Valencia D, Frazier B, Pääbo S. 1993. Genetic and linguistic differentiation in the Americas. *Proc Natl Acad Sci USA* 90:10663–10667.
- Zerjal T, Beckman L, Beckman G, Mikelsaar A-V, Krumina A, Kucinskis V, Hurles ME, and Tyler-Smith C. 2001. Geographic, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in northern European populations. *Mol Biol Evol* 18:1077–1087.