A Formal Test of Linguistic and Genetic Coevolution in Native Central and South America

K.L. Hunley,[1]* G.S. Cabana,[2] D.A. Merriwether,[3] and J.C. Long[4]

[1]*Department of Anthropology, University of New Mexico, Albuquerque, NM 87106*
[2]*Department of Anthropology, University of Tennessee, Knoxville, TN 37996*
[3]*Department of Anthropology, Binghamton University, Binghamton, NY 13902*
[4]*Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109*

KEY WORDS mtDNA; model fitting; Native American language classifications

ABSTRACT This paper investigates a mechanism of linguistic and genetic coevolution in Native Central and South America. This mechanism proposes that a process of population fissions, expansions into new territories, and isolation of ancestral and descendant groups will produce congruent language and gene trees. To evaluate this population fissions mechanism, we collected published mtDNA sequences for 1,381 individuals from 17 Native Central and South American populations. We then tested the hypothesis that three well-known language classifications also represented the genetic structure of these populations. We rejected the hypothesis for each language classification. Our tests revealed linguistic and genetic correspondence in several shallow branches common to each classification, but no linguistic and genetic correspondence in the deeper branches contained in two of the language classifications. We discuss the possible causes for the lack of congruence between linguistic and genetic structure in the region, and describe alternative mechanisms of linguistic and genetic correspondence and their predictions. Am J Phys Anthropol 132:622–631, 2007. ©2007 Wiley-Liss, Inc.

In Chapter 13 of *The Origin of Species*, Darwin (1859) proposed that "…a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world…". He reasoned that if two groups descended from a common ancestor, their languages likely descended from the language of that ancestor. Geneticists have since proposed a mechanism to explain in more detail how such linguistic and genetic correspondences might form. This mechanism follows directly from Darwin's idea: populations are related to each other through a series of fissions that occurred at times of range expansion and colonization of new regions (Cavalli-Sforza et al., 1988). These fissions produce a branching or treelike pattern of population relationships. Linguistic and genetic evolution occur independently on the branches of this population tree. Accordingly, this population fissions mechanism predicts perfectly congruent linguistic and genetic trees.

In 1988, Cavalli-Sforza et al. informally tested this prediction by comparing a tree of linguistic relationships proposed by Joseph Greenberg (described in Ruhlen, 1991) to a gene tree constructed from protein and enzyme data collected from 42 globally distributed human populations (Cavalli-Sforza et al., 1988). In this and subsequent articles, Cavalli-Sforza et al. concluded that the gene and language trees were fundamentally congruent (Cavalli-Sforza et al., 1992; Cavalli-Sforza, 1997). Subsequent studies called these findings into question on both theoretical and methodological grounds. From a theoretical perspective, anthropologists argued that human groups regularly exchange both individuals and components of their languages. Such exchange is potentially inconsistent with treelike population structure. In a related vein, researchers noted that modes of linguistic and genetic transmission differ: languages, unlike genes, are transmitted both vertically from parents to offspring and horizontally between unrelated individuals (Boas, 1911; Spuhler, 1979; Renfrew, 1992; Chen et al., 1995b; Cavalli-Sforza, 1997). Thus languages and genes may move independently between human groups, thereby preventing or obscuring linguistic and genetic correspondence. From a methodological perspective, several studies noted that the Cavalli-Sforza et al. method of visually comparing trees did not test the prediction of linguistic and genetic treeness, nor did they formally test the prediction that the Greenberg language and population gene trees were congruent (Bateman et al., 1990; Hunley and Long, 2005).

Also at issue is the fact that Cavalli-Sforza et al. considered only a single language classification that is rejected by most linguists. It is rejected because most linguists believe that relationships among long-diverged languages have not been or cannot be established (discussed in Campbell, 1997).

This paper has two goals. First, we wish to more formally test the population fissions mechanism in a geographic region whose native people have been well characterized linguistically and genetically. Second, we wish to examine the issue of linguistic and genetic correspondence for language classifications that propose both recent and deep linguistic divergences. Central and South America represents an ideal location to achieve these goals for

*Correspondence to: Keith L. Hunley, PhD, Department of Anthropology, MSC01-1040, Anthropology, 1 University of New Mexico Albuquerque, NM 87131, USA. E-mail: khunley@unm.edu

Received 17 April 2006; accepted 23 October 2006

DOI 10.1002/ajpa.20542
Published online 4 January 2007 in Wiley InterScience (www.interscience.wiley.com).

TABLE 1. Population size, nucleotide diversity, and location

Population	N	π	Latitude	Longitude
Ache[a]	63	0.001	24.8S	56.2W
Gavião[b]	28	0.012	10.2S	61.1W
Surui[c]	24	0.005	10.2S	61.1W
Zoro[b]	29	0.011	10.3S	60.3W
Huetar[d]	27	0.011	9.7N	84.5W
Kuna[e]	63	0.009	9.5N	82.1W
Ngobe[f]	46	0.013	9.2N	82.2W
Embera[g]	44	0.019	7.9N	78.6W
Wounan[g]	31	0.021	4.4N	78.7W
Arequipa[h]	22	0.013	12.0S	74.0W
Tacajaya[h]	50	0.016	14.0S	72.0W
Cayapa[i]	30	0.019	1.2N	78.5W
Mapuche[j]	39	0.018	41.3S	69.3W
Pehuenche[k]	24	0.015	37.4S	71.2W
WaiWai[c]	26	0.018	1.0N	59.0W
Xavante[b]	25	0.008	13.3S	51.7W
Yanomamo[l]	810	0.014	2.5N	64.0W
Total	*1,381*			

[a] Schmitt et al., 2004.
[b] Ward et al., 1996.
[c] Bonatto and Salzano, 1997.
[d] Santos et al., 1994.
[e] Batista et al., 1995.
[f] Kolman et al., 1995.
[g] Kolman and Bermingham, 1997.
[h] Fuselli et al., 2003.
[i] Rickards et al., 1999.
[j] Ginther et al., 1993.
[k] Moraga et al., 2000.
[l] Hunley, 2002.

two reasons. First, numerous genetic studies have been carried out in native groups in these areas. Mitochondrial DNA (mtDNA) variation in particular has been well characterized in Native Central and South American populations. Second, the indigenous languages of the regions have received considerable attention from linguists. Their research has produced alternative Native American language classifications that tend to agree about recently diverged linguistic groups, but disagree about the deeper evolutionary relationships between these groups (Loukotka, 1968; Ruhlen, 1991; Campbell, 1997).

To address our first goal of more formally testing the population fissions model, we apply a novel maximum likelihood tree-fitting procedure to genetic data collected from 17 Native Central and South American populations. This tree-fitting procedure allows us to test the hypothesis that each language classification also represents the genetic structure of the Native American populations. With respect to the second goal, this model-fitting procedure also allows us to determine whether two Native American linguistic classifications that propose deeper linguistic relationships correspond to the genetic data better or worse than a language classification that contains only relatively shallow relationships.

MATERIALS AND METHODS

Sample

We assembled the first hypervariable region sequences of 1,381 individuals for 17 Native Southern Central and South American populations from published sources (Table 1). No biological samples were handled. The population samples include mtDNA sequences from at least 20 individuals. These data include only sequences belonging to founding Native American haplogroups A, B, C, and D. The sequences were aligned and edited to 401 nucleotides, covering the reference nucleotide positions (np) 16,000–16,400 (Andrews et al., 1999). The location of each population is depicted in Figure 1.

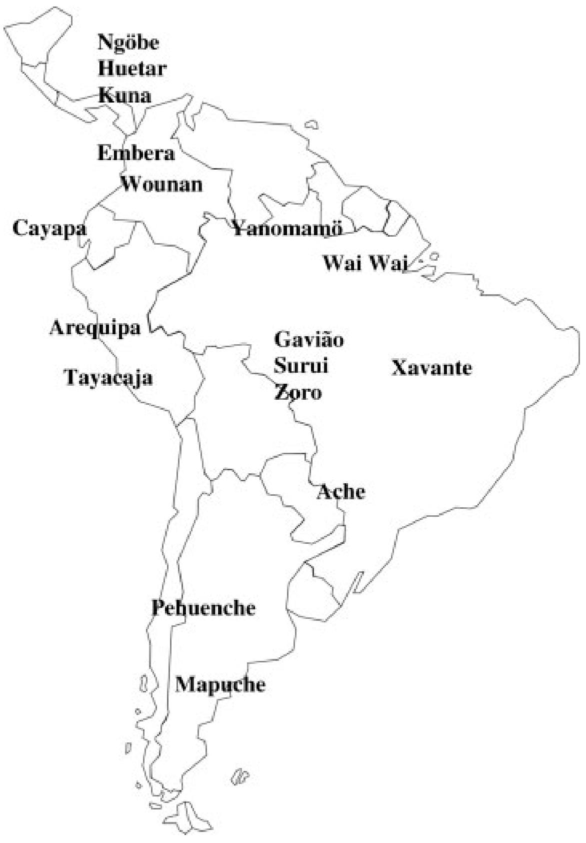

Fig. 1. Population locations.

Language classifications

We chose three distinctive Native American languages classifications from Loukotka (1968), Greenberg (described in Ruhlen, 1991), and Campbell (1987), hereafter referred to as respectively as the LLC, GLC, and CLC. The left-hand portions of Figure 2C–E depict the linguistic relationships between the 17 populations for each language classification. We note at the outset that the LLC and GLC are rejected by most linguists (e.g., see Campbell, 1997). Nonetheless, we chose these three classifications because they are similar and dissimilar to one another in important ways.

They are similar in that all three classifications share the same four shallow (distal) language groups (Chibchan, Choco, Quechuan, and Tupi), though they differ somewhat about which languages belong in each group, e.g., only the LLC places the Cayapa within the Chibchan language group, while only the GLC places the Yanomamö in the Chibchan language group. They also differ somewhat about the branching pattern between languages within the groups. We refer to this shared shallow structure as *external* structure.

The language classifications differ in that the LLC and GLC propose deeper internal relationships between these four shared groups and other languages. For example, Loukotka proposes three deeper groups (Andean,

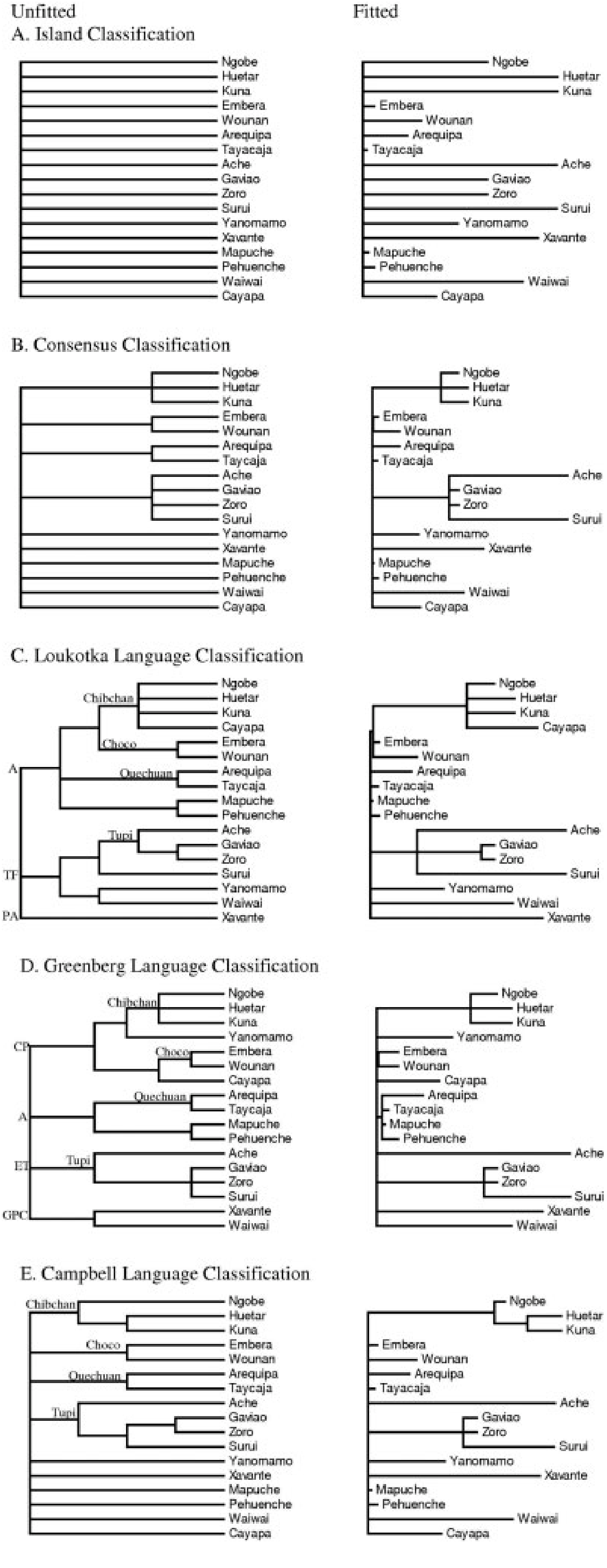

Fig. 2. Selected model classifications drawn as trees (left). Fitted language classifications are plotted to the right of each unfitted classification. Abbreviations: A, Andean; CP, Chibchan-Paez; ET, Equatorial Tucanoan; GPC, Ge Pano Carib; PA, Paleo-American; TF, Tropical Forest.

Tropical Forest, and Paleo-American), while Greenberg proposes four deeper groups (Chibchan-Paez, Andean, Equatorial-Tucanoan, and Ge-Pano-Carib). We refer to this deeper structure as *internal* structure. The CLC lacks this internal structure, because like most linguists, Campbell believes the data and methods employed by Loukotka and Greenberg are insufficient to establish these deeper relationships. The differences between these language classifications provide the opportunity to formally test for the first time whether this deeper structure enhances or detracts from linguistic and genetic correspondence. We return to the issue of linguistic reconstruction in the discussion.

We note that Loukotka does not specify the evolutionary relationships between his three internal clusters. Similarly, Campbell does not specify the relationships between the four external groups and other populations. We therefore chose the simplest topology to join the branches of the LLC and CLC to a common ancestor. This topology assumes a single origin for Native American languages but makes no further assumptions about the relationships between proposed internal and external groupings.

To distinguish how shared and unique portions of each of these three language classifications affect linguistic and genetic correspondence, we next constructed four additional classifications. The first classification, termed the island classification (IC), contains no structure (Fig. 2A, left portion). It assumes independent evolution among all populations. The second classification, termed the consensus classification (CC), adds to the IC only the structure shared by all three language classifications (Fig. 2B, left portion). It represents a strict consensus of the three language classifications.

The remaining two classifications add the unique external linguistic structure (e.g., unique branching pattern within the four shared groups) of the LLC and GLC to the CC. We refer to these as the external branch-only versions of the LLC and GLC. We next employed the statistical analyses described below to test the hypothesis that each language classification represents the genetic structure of the Central and South American populations.

Statistical analyses

Model and estimation. We first constructed a matrix of the average number of nucleotide substitutions between each pair of mtDNA sequences. This measure incorporates the differences in mutations between mtDNA sequences as well as frequency differences, making it a sensitive measure of evolutionary change (Hudson et al., 1992). The diagonal elements of the matrix, $\hat{d}_{ii}$, pertain to averages for pairs drawn from the same population. The off-diagonal elements, $\hat{d}_{ij}$, pertain to averages for pairs drawn from two different populations. Pairwise differences lead naturally to net number of nucleotide substitutions, $\hat{d}_{A_{ij}} = \hat{d}_{ij} - (\hat{d}_{ii} + \hat{d}_{jj})/2$, as a measure of genetic distance between populations (Nei, 1987).

There is an expected form of the pairwise difference matrix for any given hierarchical structure, e.g., language classification. Cavalli-Sforza and Piazza (1975) showed that the expected form of the matrix can be expressed in the form of a linear combination of fixed symmetric $s \times s$ matrices composed of zeros and ones, where s is the number of population samples. These matrices are defined by the nodes of a given classification. The number of such matrices is equal to the number of nodes hypothesized to have equal pairwise difference values.

We used a system of equations developed by Anderson (1973) to obtain estimates of the elements of the expected form of the pairwise difference matrix for each classification. The results of this procedure are approximate maximum likelihood solutions. The method is given in more detail by Cavalli-Sforza and Piazza (1975), Urbanek et al. (1996), and Long and Kittles (2003).

Hypothesis testing. The likelihood ratio test proposed by Cavalli-Sforza and Piazza (1975) provides a way to assess the goodness-of-fit of each classification described earlier to the pairwise difference matrix. The lack of fit of the expected form of the pairwise difference matrix for a given classification to the observed pairwise difference matrix is measured by a likelihood ratio statistic, Λ. Under the assumption of a large number of independently evolving sites, Λ is distributed as a χ^2 random variable, with degrees of freedom equal to $s(s + 1)/2$ minus the number of parameters specified by the fitted classification. The expected value of Λ is equal to the degrees of freedom when the classification fits perfectly. However, the likelihood ratio test is generally too liberal, i.e. rejects the null hypothesis too easily, because the assumption of a large number of independently evolving sites is violated by mtDNA sequences. Nevertheless, Λ is a useful gauge for rank ordering the fit of different classifications applied to the same data.

Visual inspection of fit. To examine the specific ways in which the linguistic classifications and genetic data departed, we plotted the realized and expected net nucleotide distances against one another for each classification. If a model classification fits the genetic data, the scattergram of realized and expected genetic distances will assume a symmetric cigar-shaped distribution and the expected genetic distances for each population will be overestimated approximately as frequently as they are underestimated.

Additional tests. There are at least four reasons why a given model classification might fit poorly. First, one or a few misplaced populations might disrupt the fit even though the majority of a particular classification fits well. Second, linguistic evolution may not be treelike. Third, population genetic evolution may not be treelike. Fourth, a given model classification may incorrectly represent evolutionary relationships among populations.

To distinguish between these alternatives, we applied several additional tests. First, to be certain that the fit of a classification was not disrupted by one or a few misplaced populations, we repeated the analysis for each classification 17 times, each time leaving out a different sample from the analysis. If one or a few populations caused the lack of fit, Λ would decrease dramatically when they were left out.

Next, to assess possibilities 2–4 mentioned earlier, we applied the neighbor joining (NJ) algorithm (Saitou and Nei, 1987) to identify a tree topology optimized to the genetic data. We then used the maximum likelihood method to fit this topology to the genetic data, applied the likelihood ratio test for treeness, and examined plots of the realized versus expected genetic distances. The rationale underlying this approach is that if the NJ tree fits and the other classifications do not, the genetic data are treelike, but none of the other model classifications capture this treelike structure, i.e. the language classifications are incorrect, or linguistic evolution has not been treelike (Hunley and Long, 2005).

RESULTS

Sequence polymorphisms

The 1,381 mtDNA sequences contain 103 variable sites and 274 haplotypes. Of these 103 polymorphisms, 41 were population-specific. Only three polymorphisms were specific to deeper branches proposed by one or more language classification: two to the Choco branch contained in all three language classifications, and one to the Gavião-Zoro branch of the LLC and CLC. Table 1 reports population sample size, estimated nucleotide diversity ($\hat{\pi}$), and geographic coordinates. Nucleotide diversity varies among the 17 populations. For example, nucleotide diversity is low in several populations, most notably the Ache, wherein the sample of 63 individuals contains 56 identical sequences (Schmitt et al., 2004). Nucleotide diversity is highest in the Wounan followed closely by the Embera and Cayapa, populations all located in the northernmost portion of the South American continent. The magnitude and range of variation across the 17 populations is comparable to that observed in Native North Americans (Hunley and Long, 2005), though the magnitude of variation is low compared to other world regions (e.g., Merriwether et al., 1991).

Statistical analyses

Model fitting and test of treeness. The right-hand portions of Figure 2 show the results of fitting the IC, the CC, and the three language classifications to the genetic data. Because the unfitted classifications on the left of Figure 2 do not make predictions about branch lengths, these lengths are arbitrary. The fitted classifications on the right on the other hand contain branch lengths estimated from the model-fitting procedure. For this reason, the unfitted and fitted classifications look quite different.

Most notable are the many zero-length branches in the fitted classifications. When we first fit the classifications, as a result of a lack of treelike structure, these branches were negative. Because negative branches are inconsistent with treeness, we removed them by combining them with the next deepest associated node in the classification. We then reapplied the model-fitting procedure. This procedure was repeated until all negative branches were eliminated from a given model classification. The outcome of this process is the zero-length branches in the fitted classifications.

Zero-length branches of the LLC include the internal Andean and Tropical Forest branches and the external Choco and Quechuan branches. Zero-length branches of the GLC include the internal Chibchan-Paez and Equatorial Tucanoan branches, the external Quechuan branch, and the Yanomamö branch of the Chibchan group. For the CLC, the Choco and Quechuan branches had zero length.

More important than the branch lengths of the fitted classifications is their fit. Table 2 lists the Λ values for all classifications, numbered 1–8. The high Λ values relative to the degrees of freedom indicate that all classifications fit poorly. We emphasize that the poor fit of these classifications was not the result of one or a few misplaced populations. When we repeated the analysis for each classification, each time leaving out a different population from the analysis, the likelihood ratio statistic substantially exceeded its degrees of freedom in each analysis (results not shown). We conclude that the poor fit of each classification is systemic and not caused by a

TABLE 2. Fit of model classifications

No.	Model classification	Λ	DF	Λ rank	R^2	R^2 rank
1	Island (IC)	2,280	140	8	0.48	8
2	Consensus (CC)	1,854	137	5	0.49	7
3	Loukotka language (LLC)	1,856	136	6	0.5	6
4	Loukotka language EBO[a]	1,835	136	4	0.57	5
5	Greenberg language (GLC)	1,885	137	7	0.6	3
6	Greenberg language EBO	1,776	136	3	0.59	4
7	Campbell language (CLC)	1,767	136	2	0.64	2
8	Neighbor joining (NJ)	1,440	136	1	0.71	1

[a] EBO, external branch only.

small number of outlier populations. However, despite the fact that all classifications fit poorly, their relative fit varies substantially. Several aspects of this variation in fit provide insights into linguistic and genetic correspondence in the region.

First, the IC contains no structure. It is the worst fitting classification. The CC, which contains only the shared features of each language classification, fits substantially better than the IC. The superior fit of the CC relative to the IC indicates that the genetic data do in fact contain some structure.

Second, the CLC, which contains no internal connections between the four shared language groups, fits better than all other language classifications and the CC. At the same time, the external branch-only versions of the LLC and GLC, which also contain no internal structure, also fit better than the CC. These results indicate that the unique external features of the language classifications improve their correspondence with the genetic data relative to the CC.

Third, when the deeper internal linguistic structure is added to the external branch-only versions of the LLC and GLC, their fit decreases (Table 2). The complete version of the LLC fits only slightly worse than the external branch-only version of the LLC, while the complete version of the GLC fits substantially worse than its external branch-only version. In addition, both of these complete language classifications fit worse than the CC. These results emphasize the important point that the improved fit of the languages classifications relative to the CC comes exclusively from their external linguistic structure.

Plots of realized and expected genetic distances. To identify specific ways that the language classifications and genetic data depart, we examined the plots of realized versus expected genetic distances (Fig. 3). The squared correlation coefficients for each plot are listed in Table 2. The plots indicate that the language classifications either consistently over- or underestimate genetic distances for each population. For example, the LLC (Fig. 3A) underestimates genetic distances for the Cayapa. At the same time, the LLC tends to underestimate lower genetic distances for the Yanomamö and overestimate higher genetic distances. Similar trends occur for most populations for each language classification (data not shown). In addition, all plots also indicate that correlations are stronger at lower genetic distances than they are at higher genetic distances.

Figure 3A,B plots the genetic distances for the full and external branch-only versions of the LLC. The trend of over and underestimation exists in both plots. Interestingly, Figure 3B indicates that the Cayapa fit better when the internal structure is removed from the LLC.

A similar trend of over- and underestimation is observed for the GLC (Fig. 3C,D). Results of the statistical analyses in Table 2 indicate that the external branch-only version of the GLC (Fig. 3D) fits substantially better than the complete GLC. The improved fit is not obvious in the comparison of Figure 3C and D. This may be because the improved fit is not caused by any one population, but instead by small improvements in agreement between realized and expected genetic distances for one or more nodes in the external branch-only version. Note that in the complete GLC (Fig. 3C), the Yanomamö is not placed in the Chibchan group, because the Yanomamö branch was negative.

Figure 3E plots the genetic distances for the CLC. The plot looks similar to others in Figure 3. This result indicates that the superior fit of the CLC relative to other language classifications may also be caused by small improvements in fit associated with one or more of the nodes in the CLC. Nonetheless, the CLC also consistently over- and underestimates genetic distances for most populations. For example, the CLC also consistently underestimates genetic distances for the Cayapa.

NJ tree. The NJ tree fits the genetic data substantially better than all other classifications (Table 2). This result indicates that the structure exists in the genetic data that is not captured by the language classifications. At first glance, the distribution of points in the scatter plot for the NJ tree (Fig. 3F) seems fairly wide. Closer inspection indicates that this wide dispersion is caused in large part by the Yanomamo. However, despite this wide spread, the Yanomamo tend to be more evenly distributed on either side of the line across the full range of genetic distances. With the exception of the Yanomamo, data points for other populations are actually less dispersed for the NJ tree compared to other plots. For example, the Cayapa are less dispersed in the NJ plot than in any of the language classification plots. These results visually confirm the superior fit of the NJ tree. Despite this superior fit, we emphasize that the likelihood ratio statistic for the NJ tree is still high relative to its degrees of freedom, indicating that even a tree optimized to the genetic data does not fit.

The NJ tree (Fig. 4) also reveals several interesting patterns that are consistent with the results reported earlier. First, populations contained within internal branches of the LLC and GLC do not cluster together in the NJ. Second, the NJ tree contains portions of 2 of the 4 groups common to the language classifications. These two groups are the Chibchan and the Tupi. Third, the NJ tree groups the Cayapa and Chibchan populations together as well as the Yanomamö and Wai Wai. The LLC also contains these two groupings. However, the topologies of the NJ tree and LLC are otherwise quite distinct. These results highlight the fact that the changes in topology in one portion of a tree may affect genetic and linguistic correspondence in other portions of that tree.

DISCUSSION

Despite the fact that individuals and their languages may move independently between populations, many studies nevertheless have identified a relationship between patterns of linguistic and genetic diversity. One way this relationship may arise is through congruent linguistic and genetic evolution in relatively isolated populations. This population fissions mechanism of lin-

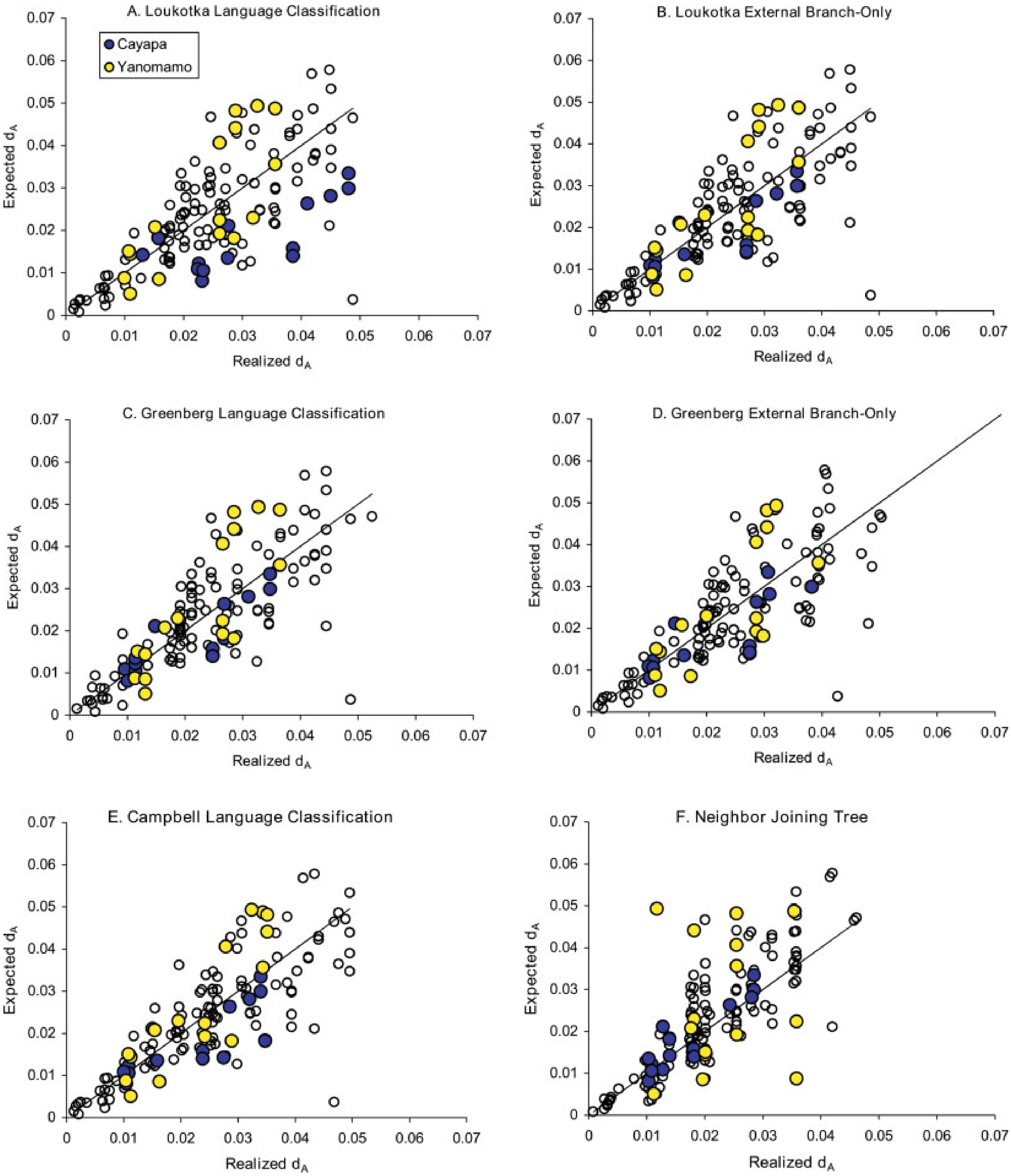

Fig. 3. Plots of realized versus expected d_A genetic distances for selected model classifications. The Cayapa and Yanomamö are highlighted in each plot. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

guistic and genetic correspondence was previously evaluated in a world wide sample of populations (Cavalli-Sforza et al., 1988, 1992). In this study, we adopted a more rigorous approach to tree comparisons (Urbanek et al., 1996; Long and Kittles, 2003; Hunley and Long, 2005). First, we proposed various classifications as a priori hypotheses for South American mtDNA genetic structure and then fit these classifications to the genetic data. We then tested for treeness using the method developed by Cavalli-Sforza and Piazza (1975). The null hypothesis

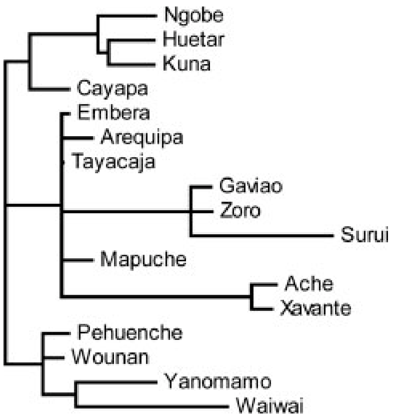

Fig. 4. Unrooted fitted NJ tree.

for each test was that the pattern of relationships among languages was the same as the pattern of relationships among gene pools. The null hypothesis was rejected for each classification.

Nonetheless, we found that all language classifications fit significantly better than the IC, an island classification that lacked structure. In addition, versions of the language classifications that lacked internal structure fit better than the CC, a strict consensus of the three languages classifications. In fact, the CLC, which contains no internal structure, was the best fitting language classification. In contrast, versions of the language classifications that contained internal structure fit worse than the CC. We conclude from these results that the deeper internal linguistic structure of the LLC and GLC is inconsistent with the genetic structure of Central and South American populations.

One might object that the Greenberg and Loukotka classifications were unlikely to fit because they are wrong, i.e., they represent incorrect language classifications. One might also object that we did not really test the LLC and CLC, because neither Loukotka nor Campbell comments on the internal relationships that connect their proposed groups to a common ancestor. Our answers to these objections are twofold. First, we chose the simplest topologies to join the branches of the LLC and CLC to a common ancestor (Fig. 2). These simple topologies make only one assumption: Native America languages are related by descent. These topologies fit poorly. Second, it is possible that other deeper branching arrangements will fit the genetic data better. However, no topology will fit substantially better than the NJ tree. While the NJ tree shares several external linguistic features of the language classifications, it does not contain any of the deeper internal structure of the LLC, GLC, or any alternative Native American language classification (e.g., those described in Campbell, 1997). We therefore conclude that no internal linguistic structure is likely to fit the genetic data.

Importantly, this result tells us nothing about the existence of deeper linguistic structure in Latin America. Our results simply reject the hypothesis that the proposed linguistic structures are also the genetic structure of the selected populations. Interestingly, one recent study argues that deep linguistic relationships may be identified using grammatical features rather than vocabulary (Dunn et al., 2005). This novel approach may permit the construction of deeper topologies in various world regions and even at larger geographic scales. In the event such classifications are constructed, the methods employed in this study may be used to address the population fissions mechanism of linguistic and genetic coevolution in these regions.

So why is the fit of all classifications poor? The poor fit is not the result one or a few misplaced populations. All classifications still fit poorly when individual populations were removed one at a time. The poor fit is not the result of a lack of information in the genetic data. All classifications fit better than an IC, which contains no structure.

European contact might contribute to the poor fit of the various classifications. Contact resulted in the eradication of many South American populations, and undoubtedly reduced the amount of genetic variation in others (Crawford, 1998; Salzano and Callegari-Jacques, 1988). If contact also increased interactions between populations, it may have erased previously existing linguistic and genetic structure. Ethnographic literature documents contact-mediated interactions of this sort. To cite one such example, Christian missions facilitated migration between linguistically distinct Yanomamö and Makiritare villages in Amazonia (Chagnon et al., 1970). On the other hand, some contact-mediated events may not have substantially affected interactions between groups, but may have instead only reduced variation within populations. Because such events would not affect variation between populations, they would not have the disrupted treelike structure. Instead they would have only decreased variation within populations, and hence increased branch lengths within the trees. At this point, we unfortunately lack the data to assess the importance of contact-mediated events of this latter sort.

Yet another possible reason for the poor fit of the language classifications is that linguistic evolution is inherently non-treelike (e.g., see Bateman et al., 1990 and associated comments). For example, linguistic exogamy occurs in South America today and may have been common in the past (e.g., Epps, 2003). Linguistic exogamy may impede or prevent treelike linguistic evolution (Nettle, 1999). Languages and patterns of linguistic diversity also change within and between groups as sociocultural organization changes in expanding, contracting, and moving populations (Gumperz, 1962; Thomason and Kaufman, 1988; Nichols, 1997; Nettle, 1999). Sociocultural variation is indeed marked in Central and South America (Steward and Faron, 1959). As a result, linguistic evolution may be treelike to some extent at some places and times, but it is unlikely to be treelike across geographically and temporally fluid sociocultural boundaries.

Certainly one major reason the language classifications fit poorly is that population genetic evolution has not been treelike. This lack of genetic treeness is reflected in the poor fit of the NJ tree. Treeness may never have existed in the region, or it may have been disrupted by indigenous population processes, including contractions, expansions, fusions, fissions, genetic exchange, and movements (Cavalli-Sforza et al., 1992; Barbujani et al., 1994; Cavalli-Sforza, 1997; Hunley and Long, 2005). Again we suspect that sociocultural variation and change significantly influenced these population processes.

Other genetic data

Also of interest is whether similar results are likely to apply to Y chromosome and autosomal genetic markers.

If patrilocality predominates in Native Latin American groups, Y chromosome variation in particular may be more structured, and one or more of the linguistic classifications may more closely fit patterns of Y chromosome variation. However, one relevant study suggests that patilocality and matrocality are equally frequent in Native South America (Burton et al., 1996). Consistent with this finding, Mesa et al. (2000) identified similar G_{ST} values for mtDNA and Y chromosome markers in a sample of Native South American populations. Analyses of more Y chromosome and autosomal data are required to more fully address the issue of linguistic and genetic correspondence in this and other regions.

Alternative mechanisms of linguistic and genetic correspondence

Other mechanisms may produce linguistic and genetic correspondence. For example, a second mechanism is analogous to the Wright/Malecot model of isolation by distance (Wright, 1943; Malecot, 1948). This mechanism applies to populations spread evenly over a continuum, without sharp divisions between groups. Individuals within the same neighborhood are likely to have recent common ancestors, and their language characteristics are likely to be retained from those ancestors. By contrast, for individuals located in different neighborhoods, their common ancestors are likely to have lived in the more remote past, and their speech will have drifted further from the speech of their ancestors. The degree to which individuals in different neighborhoods share either language or genes is a function of the geographic distance between neighborhoods (Morton, 1973; Cavalli-Sforza and Wang, 1986; Cavalli-Sforza et al., 1992). This "diffusion" mechanism predicts that linguistic and genetic distances will be correlated, and that this correlation will disappear when geographic distance is controlled.

Tests of these predictions provide equivocal support for this diffusion mechanism. For example, studies frequently identify moderate correlations between linguistic, genetic, and geographic distances in various world regions (Barrantes et al., 1990; Chen et al., 1995a; Nettle and Harriss, 2003). On the other hand, studies often identify only relatively weak correlations between these distances in the same and other world regions (Spuhler, 1972; Chakraborty, 1976; Smouse and Long, 1992; Nettle and Harriss, 2003). In fact, one study identified a negative correlation between these distances in South America (Fagundes et al., 2002). These weak and negative correlations may in part result from inadequate measures of linguistic distance. They may also result from the failure to account for linguistic and sociocultural heterogeneity. Indeed, correlations between linguistic, genetic, and geographic distances tend to be stronger within more culturally and linguistically homogeneous Native Central and South American groups (Roisenberg and Morton, 1970; Lalouel and Morton, 1973; Spielman et al., 1974; Sokal et al., 1986; Barbujani et al., 1989; Barrantes et al., 1990; Rothhammer, 1990; Demarchi and Marcellino, 1998; Mesa et al., 2000; Fuselli et al., 2003).

We did in fact examine the diffusion mechanism for the 17 populations included in this study. For this examination, we estimated linguistic distances as the number of nodes separating each population for each language classification (Excoffier et al., 1991). Great-circle geographic distances were estimated from geographic coordinates provided in Table 1. We identified the strongest genetic–linguistic distance correlation for the Loukotka classification. However, the squared correlation coefficient for this comparison was only 0.053 ($P < 0.05$). Importantly, though the squared correlation coefficient dropped to 0.040 when geographic distance was held constant, the correlation remained statistically significant at the 0.05 level. This latter result represents a rejection of the diffusion mechanism and suggests gene–language correspondence that is independent of geography.

We suspect that one reason this correlation is weak is that the method of linguistic distance estimation is crude. Unfortunately, the data and methods required to estimate linguistic distances more accurately are currently unavailable. However, Dunn et al. (2005) provide a method for linguistic distance estimation that might aid in examining the diffusion mechanism at different geographic scales. In this vein, a future goal of our research is to involve linguistics in the construction and tests of mechanisms of coevolution (see Hunley et al., in press).

Finally, linguistic and genetic correspondence may be produced through a combination of shared ancestry and shared patterns of linguistic and genetic exchange. In this case, neither linguistic nor genetic data will be tree-like, but, nonetheless, trees constructed from these data may produce similar topologies. For example, in this study, both the language and NJ classifications contain Chibchan and Tupi groups. This common structure may be the result of shared ancestry and genetic exchange among Chibchan and Tupi speaking groups. Sharing of this sort may be common in other world regions and at differing geographic and sociopolitical scales. We suggest that one fruitful approach for future research in this area will be to propose and test these and related models for the formation of linguistic and genetic correspondences within more geographically, ecologically, and socioculturally homogeneous regions.

LITERATURE CITED

Anderson T. 1973. Assymptotically efficient estimation of covariance matrices with linear structure. Ann Stat 1:135–141.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Barbujani G, Oden NL, Sokal RR. 1989. Detecting regions of abrupt change in maps of biological variables. Syst Zool 38:376–389.

Barbujani G, Whitehead GN, Bertorelle G, Nasidze IS. 1994. Testing hypotheses on processes of genetic and linguistic change in the Caucasus. Hum Biol 66:843–864.

Barrantes R, Smouse PE, Mohrenweiser HW, Gershowitz H, Azofeifa J, Arias TD, Neel JV. 1990. Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am J Hum Genet 46:63–84.

Bateman R, Goddard I, O'Grady R, Funk VA, Moori R, Kress WJ, Cannell P. 1990. Speaking of forked tongues: the feasibility of reconciling human phylogeny and the history of languages. Curr Anthropol 31:1–24.

Batista O, Kolman CJ, Bermingham E. 1995. Mitochondrial DNA diversity in the Kuna Amerinds of Panama. Hum Mol Genet 4:921–929.

Boas F. 1911. Handbook of American Indian languages, Part 1. Washington. D.C.: Government Printing Office.

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.

Burton M, Moore C, Whiting J, Romney A. 1996. Regions based on social structure. Curr Anthropol 37:87–123.

Campbell L. 1997. American Indian languages: the historical linguistics of Native America. New York: Oxford University Press.

Cavalli-Sforza LL. 1997. Genes, peoples, and languages. Proc Natl Acad Sci USA 94:7719–7724.

Cavalli-Sforza LL, Minch E, Mountain JL. 1992. Coevolution of genes and languages revisited. Proc Natl Acad Sci USA 89: 5620–5624.

Cavalli-Sforza LL, Piazza A. 1975. Analysis of evolution: evolutionary rates, independence and treeness. Theor Popul Biol 8: 127–165.

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, Wang WS-Y. 1986. Spatial distance and lexical replacement. Language 62:38–55.

Chagnon N, Neel J, Weitkamp L, Gershowitz H, Ayres M. 1970. The influence of cultural factors on the demography and pattern of gene flow from the Makiritare to the Yanomama Indians. Am J Phys Anthropol 32:339–349.

Chakraborty R. 1976. Cultural, language and geographical correlates of genetic variability in Andean highland Indians. Nature 264:350–352.

Chen J, Sokal RR, Ruhlen M. 1995a. Worldwide analysis of genetic and linguistic relationships of human populations. Hum Biol 67:595–612.

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. 1995b. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133–149.

Crawford MH. 1998. The origins of Native Americans: Evidence from anthropological genetics. Cambridge: Cambridge University Press.

Darwin C. 1859. The origin of species. London: John Murray.

Demarchi DA, Marcellino AJ. 1998. Dermatoglyphic relationships among South Amerindian populations. Hum Biol 70: 579–596.

Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC. 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309:2072–2075.

Epps P. 2003. Evidentiality as an areal feature: evidence from Hup. Studies in Language 29:617–650.

Excoffier L, Harding RM, Sokal RR, Pellegrini B, Sanchez-Mazas A. 1991. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum Biol 63:273–307.

Fagundes NJ, 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.

Ginther C, Corach D, Penacino GA, Rey JA, Carnese FR, Hutz MH, Anderson A, Just J, Salzano FM, King MC. 1993. Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. Exs 67:211–219.

Gumperz JJ. 1962. Types of linguistic communities. Anthropol Linguist 4:28–36.

Hudson RR, Boos DD, Kaplan NL. 1992. A statistical test for detecting geographic subdivision. Mol Biol Evol 9:138–151.

Hunley KL. 2002. The anthropological utility of genetic data in small-scale populations: migration rates and patterns among the Yanomamö. PhD thesis, University of Michigan, Ann Arbor.

Hunley K, Dunn M, Lindström E, Reesink G, Terrill A, Norton H, Scheinfeldt L, Friedlaender F, Merriwether D, Koki G, Friedlaender J. 2007. Inferring prehistory from genetic, linguistic, and geographic variation. In: Friedlaender J, editor. Genetics, linguistics, and culture history in the Southwest Pacific. Cambridge: Cambridge University Press. p 313–330.

Hunley K, Long JC. 2005. Gene flow across linguistic boundaries in Native North American populations. Proc Natl Acad Sci USA 102:1312–1317.

Lalouel JM, Morton NE. 1973. Bioassay of kinship in a South American Indian population. Am J Hum Genet 25:62–73.

Long J, Kittles R. 2003. Human genetic diversity and the nonexistence of biological races. Hum Biol 75:449–471.

Loukotka C. 1968. Classification of South American Indian languages. Los Angeles: University of California.

Kolman CJ, Bermingham E. 1997. Mitochondrial and nuclear DNA diversity in the Choco and Chibcha Amerinds of Panama. Genetics 147:1289–1302.

Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F. 1995. Reduced mtDNA diversity in the Ngöbe Amerinds of Panama. Genetics 140:275–283.

Malecot G. 1948. Les Mathématiques de l'hérédité. Paris: Masson et Cie.

Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST, Wallace DC. 1991. The structure of human mitochondrial DNA variation. J Mol Evol 33:543–555.

Mesa NR, Mondragon MC, Soto ID, Parra MV, Duque C, Ortiz-Barrientos D, Garcia LF, Velez ID, Bravo ML, Munera JG, Bedoya G, Bortolini MC, Ruiz-Linares A. 2000. Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: Pre- and post-Columbian patterns of gene flow in South America. Am J Hum Genet 67:1277–1286.

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.

Morton N. 1973. Population structure of Micronesia. In: Crawford M, Workman P, editors. Methods and theories of anthropological genetics. Albuquerque: University of New Mexico Press. p 333–366.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nettle D. 1999. Linguistic diversity. Oxford: Oxford University Press.

Nettle D, Harriss L. 2003. Genetic and linguistic affinities between human populations in Eurasia and West Africa. Hum Biol 75:331–344.

Nichols J. 1997. Modeling ancient population structures and movements in linguistics. Annu Rev Anthropol 26:359–384.

Renfrew C. 1992. Archaeology, genetics, and linguistic diversity. Man 27:445–478.

Rickards O, Martinez-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.

Roisenberg I, Morton N. 1970. Population structure of blood groups in Central and South American Indians. Am J Phys Anthropol 32:373–376.

Rothhammer F. 1990. Ethnogenesis and affinities to other South American aboriginal populations. In: Schull WJ, Rothhammer F, Barton SA, editors. The Aymara: strategies in human adaptation to rigorous environments. Dordrecht: Kluwer. p 203–210.

Ruhlen M. 1991. A guide to the world's languages. Stanford: Stanford University Press.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.

Salzano F, Callegari-Jacques SM. 1988. South American Indians: A case study in evolution. Oxford: Clarendon Press.

Santos M, Ward RH, Barrantes R. 1994. mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Hum Biol 66:963–977.

Schmitt R, Bonatto SL, Freitas LB, Muschner VC, Hill K, Hurtado AM, Salzano FM. 2004. Extremely limited mitochondrial

DNA variability among the Ache Natives of Paraguay. Ann Hum Biol 31:87–94.

Smouse P, Long J. 1992. Matrix correlation analysis in anthropology and genetics. Yrbk Phys Anthropol 35:187–213.

Sokal R, Smouse P, Neel J. 1986. The genetic structure of a tribal population, the Yanomama Indians. XV. Patterns inferred by autocorrelation analysis. Genetics 114:259–287.

Spielman RS, Migliazza EC, Neel JV. 1974. Regional linguistic and genetic differences among Yanomama Indians. Science 184:637–644.

Spuhler J. 1972. Genetic, linguistic and geographical distances in Native North America. In: Weiner J, Huizinga J, editors. The assessment of population affinities in man. Oxford: Clarendon Press.

Spuhler JN. 1979. Genetic distance, trees, and maps of North American Indians. In: Laughlin WS, Harper AB, editors. The first Americans: origins, affinities, and adaptations. New York: Gustav Fischer. p 135–183.

Steward J, Faron L. 1959. Native peoples of South America. New York: McGraw Hill.

Thomason SG, Kaufman T. 1988. Language contact, creolization, and genetic linguistics. Berkley: University of California Press.

Urbanek M, Goldman D, Long JC. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. Mol Biol Evol 13:943–953.

Ward R, Salzano F, Bonatto S, Santos R. 1996. Mitochondrial DNA polymorphism in three Brazilian Indian tribes. Am J Human Biology 8:317–323.

Wright S. 1943. Isolation by distance. Genetics 28:114–138.