Original Paper

Eur J Hum Genet 1995;3:256–263

Giorgio Bertorelle[a]
Jaume Bertranpetit[b]
Francesc Calafell[b]
Ivane S. Nasidze[c]
Guido Barbujani[d]

[a] Dipartimento di Biologia, Università di Padova, Italia;
[b] Departament de Biologia Animal, Universitiat de Barcelona, España;
[c] Department of Nucleic Acids, Institute of Biochemistry, Georgian Academy of Sciences, Tbilisi, Republic of Georgia, and
[d] Dipartimento di Scienze Statistiche, Università di Bologna, Italia

Do Basque- and Caucasian-Speaking Populations Share Non-Indo-European Ancestors?

Key Words
Evolution
Genetic polymorphisms
Language
Basque speakers
Caucasian speakers
Genetic distance

Abstract

Genetic evidence is consistent with the view that the Indo-European languages were propagated in Europe by the diffusion of early farmers. The existence of phylogenetic relationships between European populations speaking other languages has been proposed on linguistic and archaeological grounds, and is here tested by analyzing allele frequencies at ten polymorphic protein and blood group loci. Genetic distances between speakers of Basque and Caucasian languages are compared with those between controls, i.e. contiguous populations speaking Indo-European and Altaic. Although some statistical tests show an excess of genetic similarity between Basque and South Caucasian speakers, most results do not support their common origin. If the Basques and the Caucasian-speaking populations share common ancestors, recent evolutionary phenomena must have caused divergence between them, so that their gene frequencies do not appear more similar now than those of random pairs of populations separated by the same geographic distance.

Introduction

Europeans generally speak languages of the Indo-European family. Exceptions include the Basques and most inhabitants of the Caucasus, the latter speaking Caucasian languages [1]. Both are genetically distinct from most of their geographical neighbors [2–5], and both are considered to have evolved under demographic pressures other than those prevailing in the rest of Europe. Basically, it seems that the Basque country [6] and the Caucasus [7] have only marginally been affected by the demic expansion that propagated in Europe

Received: October 24, 1994
Revision received· May 16, 1995
Accepted: May 19, 1995

Giorgio Bertorelle
Department of Biology
University of Padova
I-35121 Padova (Italy)

1018–4813/95/
0034–0256$8.00/0

farming technologies [8], genes [8–10], and possibly languages [5, 11, 12] of Near Eastern origin.

A certain degree of linguistic resemblance between Basque and Caucasian was first described in the 1920s [13], and proposed again more recently [14, 15]. There is no consensus among linguists on whether all Caucasian languages, or only some of them, may be related to Basque, nor is it clear if Caucasian and Basque may belong to an even larger linguistic unit, including other groups of Asia and America, defined as Dene-Caucasian [16]. Various independent kinds of evidence, however, hint at the existence of evolutionary relationships between Basques and Caucasians (this term will hereafter indicate the inhabitants of the Caucasus area, and not the so-called Caucasian race, i.e. whites).

A model accounting for these relationships could be as follows. According to the hypothesis of demic diffusion [8, 9, 17], neolithic farmers dispersed in Europe from the Near East, about 10,000 years ago. They partly replaced old mesolithic occupants, and partly intermixed with them [8, 11], establishing large-scale gradients of allele frequencies [8, 18]. Populations living in the western Pyrenees may have resisted this [6], as well as subsequent [19] immigration waves introducing genetic and linguistic novelty, and may have retained specific genetic features. Indeed, Basques show the lowest frequency of the B and Rhesus-negative blood groups in Europe [2], and differ in many gene frequencies from their spatial neighbors [3, 20, 21]. Present-day Basques, therefore, could be in phylogenetic continuity with the pre-Indo-European inhabitants of Europe.

As for the Caucasus, few allele frequencies form gradients [7; Nasidze, unpubl. results], which does not correspond to the likely consequences of demic diffusion as originally envisaged by Ammerman and Cavalli-Sforza [9]. Presumably, the Caucasus mountains did not prove a favorable environment for the establishment of agriculture, and therefore this area has been overlooked, totally or partly, by dispersing neolithic farmers [22]. If so, the alleles propagated in Europe by the expanding neolithic groups should be rare among Caucasians as well as among Basques.

Linguistic relationships, if confirmed, would further imply that Caucasians and Basques not only lack linguistic elements typical of contemporary European populations, but also had a cultural relationship, albeit distant. Should their gene frequencies also show some level of resemblance, their *common ancestry* could be reasonably envisaged. Caucasians and Basques could thus be regarded as descendants of a Mesolithic population, which was restricted in mountain regions by the diffusion of other, technologically superior, groups. Testing this model against the available genetic evidence is the purpose of this paper.

Comparing Gene Frequencies in the Geographic Space

Allele frequency differences between groups depend not only on the populations' histories, but also on geographical factors, like physical barriers, environmental gradients, and distance itself. Under drift and short-range gene flow (isolation by distance), the genetic relationships between distant localities cannot be predicted in general, but neighboring localities tend to resemble each other genetically [23], and allele frequencies are spatially autocorrelated. Autocorrelation of data may affect statistical tests to the point that genetic differences among groups may appear to be significant when they are not [24]. Consequently, the effects of geography must somehow be taken into account when testing hypotheses on historical population processes.

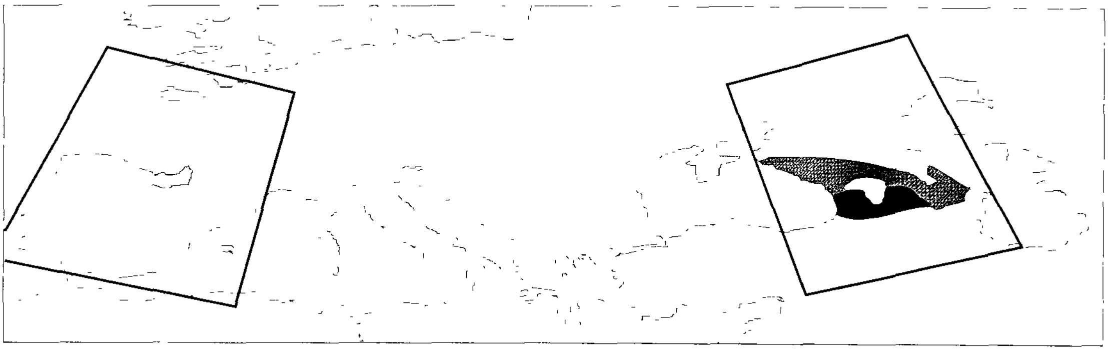


Fig. 1. A map of Europe and Asia. In the areas indicated by boxes, all populations speaking languages other than Basque and Caucasian were included in the Peri-Basque and Peri-Caucasian groups. In the eastern box, the regions where northern and southern Caucasian languages are spoken are shaded in grey and black, respectively.

In a study of Jewish populations, Livshits et al. [25] solved this problem by pairwise comparisons of Jewish and non-Jewish samples from the same country; all their comparisons therefore took place at distance 0. In our case this was clearly impossible. The Legendre et al. [24] approximate analysis of variance for autocorrelated data did not seem suitable either, as here we were asking whether two populations are more *similar*, rather than more different, than expected. We then tried to eliminate the effects of geography by comparing measures of genetic relatedness between Basques and Caucasians with values estimated from available controls.

To be suitable, pairs of control populations had to lie at approximately the same geographical distance as Basques and Caucasians, so as to be subject to approximately equal effects of spatial distance. In addition, they had to be located in such a way as to minimize selective effects that could result in an increase or decrease of genetic similarity. Only the most obvious factors of adaptive significance for humans are known, but latitude is regarded as important [26]. Therefore, we defined two rectangular areas of equal size at the same latitude around the Basque- and Caucasian-speaking regions; genetic distances were calculated between pairs of Indo-European- or Altaic-speaking populations (control groups) of these areas.

Materials and Methods

The Data

We analyzed allele frequencies at ten loci. The data came from three data bases, including allele frequencies of Eurasia [described in ref. 27, and updated by incorporating data from ref. 28], of the Iberian peninsula [3, 4], and of the Caucasus [7]. Samples were attributed to one of the following population groups (fig. 1): Basques (B), i.e. Basque-speaking populations of Spain and France; Caucasians (C), i.e. populations living in the Caucasus and speaking Caucasian languages; Peri-Basques (PB); i.e. speakers of Indo-European languages living in the region between 36 and 50°N and between 9.5°W and 4.5°E; Peri-Caucasians (PC), i.e. populations in the area between 36 and 50°N and between 37 and 51°E, speaking Altaic and Indo-European languages.

In the statistical analysis, the Caucasian group was split into two linguistic subgroups, North (NC) and South Caucasian (SC), which are generally regarded as

Table 1. Loci studied and number of samples considered

Locus	Chromosome	NC	SC	PC	B	PB
ABO	9q34	45	28	36	36	17
ACP1	2p23	16	15	18	18	19
DUFFY	11p12	18	3	7	10	9
ESD	13q14.11	17	15	13	18	17
GC	4q12	18	11	12	21	17
GLO1	6p21.2	18	11	12	18	17
HP	16q22.1	19	16	19	18	21
KEL	7q33-q35	21	3	12	10	13
MN	4q28-q31	55	30	32	15	3
RH	1p36.2-p34	52	28	33	20	8

loosely related subdivisions of a unique linguistic family (table 1) [1]. The NC samples included speakers of ten languages, namely Abaza, Abkhaz, Adigh, Andi, Avar, Chechen, Ingush, Kabardian, Lak, and Lezgi. The SC language family comprises a unique language, Georgian [also referred to as Kartvelian in ref. 7].

Statistical Methods

Two approaches were applied, namely comparison of genetic distances between groups, Basques-Caucasians versus controls, and evaluation of evolutionary trees.

The null hypothesis was that genetic differences between populations speaking Basque and Caucasian are the same as between control populations at the same geographic distance. The alternative hypothesis, common ancestry, was that Basques and Caucasians are genetically closer than controls. Under the alternative hypothesis, Basques and Caucasians should also form a cluster in the evolutionary tree.

Average allele frequencies were calculated for each of the five groups at the ten loci considered. Prevosti's genetic distance measures were computed between pairs of population groups [29]. SEs of genetic distances were estimated by bootstrapping [30], as suggested by Sanchez et al. [31]. Pseudogenetic distances were calculated by sampling with replacement from the ten loci; the procedure was repeated 10,000 times. A null distribution of the pseudogenetic distances was thus obtained, based on random assemblages of loci; SEs of the estimates were evaluated on the basis of that distribution.

In addition, the number of times the pseudodistances between B and SC, and between B and NC, were greater than the pseudodistances between controls were counted in the 10,000 iterations. These values are an empirical estimate of the probability of the null hypothesis, namely that Basques and Caucasian speakers are evolutionarily independent.

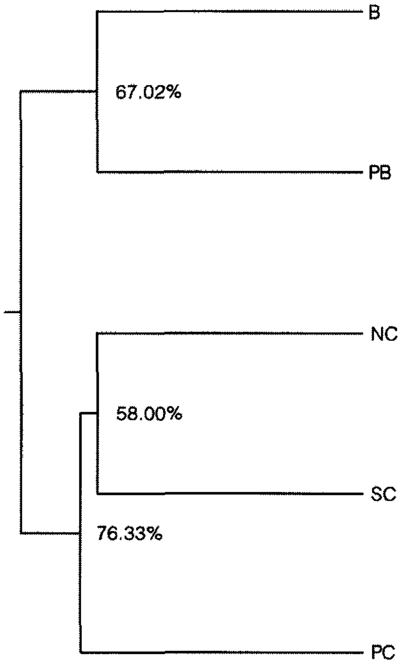
Table 2. Mean genetic distances between groups (below diagonal) and SEs based on 10,000 bootstrap iterations (above diagonal)

	B	PB	NC	SC	PC
B		0.0104	0.0199	0.0183	0.0197
PB	0.0787		0.010	0.0096	0.0096
NC	0.1202	0.0895		0.0126	0.0133
SC	0.1143	0.0812	0.0767		0.0124
PC	0.1149	0.0864	0.0776	0.0831	

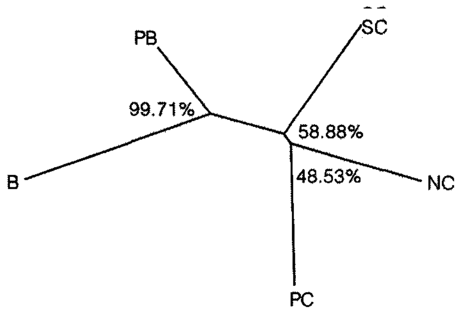
Evolutionary trees (or dendrograms) were constructed on the basis of the genetic distances calculated as described above. A rooted UPGMA tree [32] and unrooted neighbor-joining trees [33] were used. Once again, the significance of the branching was assessed by bootstrapping across loci.

Results

Estimates of genetic distances between groups are reported in table 2, along with their errors. The B group shows the highest distances from the NC, SC, and PC groups. In 10,000 bootstrap iterations, the distance be-



2



3

Fig. 2. UPGMA tree summarizing the genetic relationships among the five population groups. Figures at the nodes of the tree indicate the frequency at which each clustering was observed in 10,000 bootstrap realizations of the tree.

Fig. 3. Unrooted neigbor-joining tree summarizing the genetic relationships among five population groups. For percent values, see legend to figure 2.

tween B and NC was higher than between controls in 9,902 cases, while it was less than expected (suggesting common origin) in less than 1% of cases (98/10,000). Similarly, the difference between B and SC was lower than between controls in only 3.82% of the iterations. Both results are fully compatible with the null hypothesis of no special genealogical relationships between the speakers of Basque and Caucasian languages.

The ten loci were also analyzed separately (data not given), using a different strategy. Genetic distances were computed and compared between individual populations, rather than between average frequencies within languages. At three loci, ACP1, GLO1, and KEL, the observed distances between both NC and SC speakers on the one hand, and B speakers on the other, were less than between PB and PC. The differences between the B and SC groups (but not B and NC) were smaller than expected for two more markers, ABO and HP. At the other five loci, however, the C and B groups differed more than control groups (PB and PC). Apparently, the latter loci are those most affecting the overall genetic distances, and determining the results described above.

The UPGMA dendrogram summarizing the overall levels of genetic relatedness between groups showed mainly an effect of spatial distances, with B and PB populations clustering in one major branch of the tree, and NC, SC, and PC populations in the other (fig. 2). These relationships seem statistically robust, as the B-PB and the NC-SC-PC clusterings are observed in the vast majority (67 and 76%, respectively) of the bootstrapped trees. They are confirmed by the neighbor-joining tree (fig. 3), where the clustering of B and PB is observed in more than 99% of the bootstrap iterations. For a comparison, note that clusters B-NC and B-SC are observed in only 0.03 and 0.01% of the iterations, respectively.

Discussion

Allele frequencies of Basque and Caucasian speakers tend to fall at the extremes of the European range [see ref. 28 and 34]. As a consequence, only for a few loci do the differences between C and B correspond to the differences observed between other populations the same distance in space. For three markers (specifically, ACP1, GLO1, and KEL), speakers of Basque and Caucasian are genetically closer than expected, but in other cases they differ more than control populations.

The available genetic distance matrices do not provide an unequivocal picture. Some sets of data agree with the view that modern speakers of Basque and Caucasian have biological ancestors in common, but the overall measures of Prevosti's distances do not. Statistical tests based on bootstrapping strongly support the view that the B, SC, and NC populations bear no special evolutionary relationship. All in all, therefore, comparisons of observed and expected genetic distances do not support common ancestry.

Our tests might have been exceedingly conservative. Indo-European speakers may have shared common ancestors, living less than 10,000 years ago, in the Neolithic era [5, 35]. If so, Indo-European speakers of different geographical areas, like those in the PB and PC samples, could appear more similar genetically than Basques and Caucasians not because the latter evolved independently, but simply because Basques and Caucasians, even if they had a common ancestor, separated earlier. There is no way to test this hypothesis using allele frequencies; only archaeological data can either confirm it, or lead to its rejection

In the neighbor-joining tree (fig. 3), it is surprising to see that the branch connecting the western (B and PB) and the eastern (NC, SC, and PC) populations is very short, shorter indeed than the branches departing from it. This should suggest a certain caution, as it is unlikely that all these groups are tightly related evolutionarily. When heterogeneous populations that live in distant locations do not appear differentiated, there is reason to suspect some degree of evolutionary convergence, due either to the common environmental conditions, or, more likely, to chance. Certainly, because of their small sizes, these populations must have been exposed to major random fluctuations in allele frequencies. It may then be that the genetic relationships described in this paper between the western (B and PB) and the eastern populations result mostly from phenomena that did not lead to a linear accumulation of genetic diversity with time.

In general, therefore, these results suggest that Basques and Caucasians do not share remote biological ancestors (not more than random pairs of populations at the same spatial distance, at any rate). Alternatively, traces of their common origin have been obliterated by factors acting after their separation.

If one Caucasian group has tighter evolutionary ties with Basques, it is probably South Caucasian, for which genetic distances at five loci agree with the predictions of a model of common ancestry. This would also be in agreement with the linguistic evidence [17]. Is there any good reason to maintain that the Basque and the South Caucasian populations may be evolutionarily related? Genetic population characteristics tend to be conserved if gene flow is limited and populations are large [36]. In the conditions prevailing in the Caucasus and in the Basque-speaking area, gene flow was presumably restricted (as suggested by simulation studies [37]) and populations could not be large [3, 7]. Random fluctuations of allele frequencies across generations are then to be expected. In the long run (archaeologists place the expansion of Neolithic farmers between 8,000 and 3,000 BC [11], this may have concealed the genetic affinities, if any, resulting from a common early evolutionary history. If a population ancestral to both Basque and South Caucasian speakers existed, the allele frequency similarity observed for ABO, ACP, GLO, HP and KEL would not result from random convergence, but would be a reminder of their ancient genealogical ties.

At the allele frequency level it is not possible to speculate any further in the absence of reliable estimates of Nm, the product of population size and immigration rate, for the two compared groups [38]. Only DNA data (frequencies of RFLP morphs, sequences, and mismatch distributions) [39] may help us reconstruct the evolutionary processes on this time scale. Indeed, many of the correspondences observed between patterns of linguistic and genetic diversity seem due to comparatively recent demographic processes, occurring in the last 2 millennia; hypotheses on earlier phenomena could simply be untestable using allele frequencies [40]. Although this is not strictly true for all studies [5, 41], allele frequency data are certainly less stable than molecular data through long evolutionary periods, such as the one considered here.

Should the linguistic relationships between Basques and Caucasians be confirmed, the view whereby the traces of their parallel evolution have been obliterated at the allele-frequency level would be more plausible, and an investigation of mtDNA diversity should be carried out. At present, however, this hypothesis does not find support in our analysis of gene frequencies.

Acknowledgments

We are grateful to Colin Renfrew for calling our attention to the linguistic relationships between Caucasians and Basques, and for illuminating discussion. Italo Scardovi listened patiently to our problems and helped us out of a statistical blind alley. Ana Aguirre provided copies of several papers, Andrea Pilastro and Gerard Whitehead helped with some preliminary calculations. We thank them all. This study was supported by grants from the Italian Ministry for the University (MURST).

References

1 Ruhlen M: A Guide to the World's Languages, ed 2. London, Edward Arnold, vol 1: Classification, 1991.

2 Piazza A, Cappello N, Olivetti E, Rendine S: The Basques in Europe: A genetic analysis. Munibe 1988;6: 169–177.

3 Bertranpetit J, Cavalli-Sforza LL: A genetic reconstruction of the history of the population of the Iberian peninsula. Ann Hum Genet 1991;55: 51–67.

4 Aguirre AI, Vicario A, Mazon LI, Estomba A, Martinez de Pancorbo M, Arrieta Picò V, Perez Elortondo FP, Lostao CM: Are the Basques a single and a unique population? Am J Hum Genet 1991;49:450–458.

5 Barbujani G, Pilastro A: Genetic evidence on origin and dispersal of populations speaking languages of the Nostratic macrofamily. Proc Natl Acad Sci USA 1993;90:4670–4673.

6 Cavalli-Sforza LL: The Basque population and ancient migrations in Europe. Munibe 1988;6:129–137.

7 Barbujani G, Nasidze IS, Whitehead GN: Genetic diversity in the Caucasus. Hum Biol 1994;66:639–668.

8 Menozzi P, Piazza A, Cavalli-Sforza LL: Synthetic maps of human gene frequencies in Europeans. Science 1978;201:786–792.

9 Ammerman AJ, Cavalli-Sforza LL: The Neolithic Transition and the Genetics of Populations in Europe. Princeton, Princeton University Press, 1984.

10 Sokal RR, Oden NL, Wilson C: Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 1991;351:143–145.

11 Renfrew C: Archaeology and Language: The Puzzle of Indo-European Origins. London, Cape, 1987.

12 Starostin SA: A statistic evaluation of the time-depth and subgrouping of the Nostratic macrofamily; in Dawkins R (ed): Evolution: From Molecules to Culture. Cold Spring Harbor, Cold Spring Harbor Laboratory, 1990, p 33.

13 Trombetti A: Le Origini della Lingua Basca. Bologna, Accademia delle Scienze, 1926.

14 Lafon R: Concordances morphologiques entre le basque et les langues caucasiques. Word 1951;7:227–244 and 1952;8:80–94.

15 Ruhlen M: An overview of genetic classification; in Hawkins JA, Gell-Mann M (eds): The Evolution of Human Languages. Redwood City, Addison-Wesley, 1992, pp 159–189.

16 Shevoroshkin V, Manaster-Ramer A: Some recent work on the remote relations of languages; in Lamb S, Douglas Mitchell E (eds): Sprung from a Common Source: Investigations into the Prehistory of Languages. Stanford, Stanford University Press, 1991.

17 Cavalli-Sforza LL, Menozzi P, Piazza A: Demic expansions and human evolution. Science 1993;259:639–646.

18 Sokal RR, Menozzi P: Spatial autocorrelation of HLA frequencies in Europe supports demic diffusion of early farmers. Am Naturalist 1982; 119:1–17.

19 Collins R: The Basques. New York, Blackwell, 1986.

20 Calafell F, Bertranpetit J: Mountains and genes: A population history of the Pyrenees. Hum Biol 1994; 66:823–842.

21 Calafell F, Bertranpetit J: Principal component analysis of gene frequencies and the origin of Basques. Am J Phys Anthropol 1994;93:201–215.

22 Renfrew C: World languages and human dispersals: A minimalist view; in Hall JA, Jarvie IC (eds): Transition to Modernity. Cambridge, Cambridge University Press, 1992, pp 11–68.

23 Kimura M, Weiss GH: The stepping-stone model of population structure and the decrease of genetic correlation with distance. Genetics 1964;49:561–576.

24 Legendre P, Oden NL, Sokal RR, Vaudor A, Kim J: Approximate analysis of variance of spatially autocorrelated regional data. J Classif 1990;7:53–75.

25 Livshits G, Sokal RR, Kobyliansky E: Genetic affinities of Jewish populations. Am J Hum Genet 1991;49: 131–146.

26 Piazza A, Menozzi P, Cavalli-Sforza LL: Synthetic gene frequency maps of man and selective effects of climate. Proc Natl Acad Sci USA 1981;78:2638–2642.

27 Barbujani G, Pilastro A, De Domenico S, Renfrew C: Genetic variation in North Africa and Eurasia: Neolithic demic diffusion versus paleolithic colonisation. Am J Phys Anthropol 1994;95:137–154.

28 Roychoudhury AK, Nei M: Human Polymorphic Genes. Oxford, Oxford University Press, 1988.

29 Prevosti A, Ocana J, Alonso G: Distances between populations of *Drosophila suboscura* based on chromosome arrangement frequencies. Theor Appl Genet 1975;45:231–241.

30 Efron B: The jackknife, the bootstrap, and other resampling plans. Philadelphia, Society for Industrial and Applied Mathematics, 1982.

31 Sanchez A, Ocaña J, Utzet F: Sampling theory, estimation and significance testing for Prevosti's estimate of genetic distance. Biometrics, in press.

32 Sneath PHA, Sokal RR: Numerical Taxonomy. San Francisco, Freeman, 1973.

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

34 Cavalli-Sforza LL, Menozzi P, Piazza A: History and Geography of Human Genes. Princeton, Princeton University Press, 1994.

35 Barbujani G, Sokal RR, Oden NL: Indo-European origins: A computer-simulation test of five hypotheses. Am J Phys Anthropol 1995;96:109–132.

36 Wijsman EM, Cavalli-Sforza LL: Migration and genetic population structure with special reference to humans. Annu Rev Ecol Syst 1984; 15:279–301.

37 Calafell F, Bertranpetit J: The genetic history of the Iberian peninsula: A simulation. Curr Anthropol 1993; 34:735–745.

38 Slatkin M: Population structure and evolutionary progress. Genome 1989;31:196–202.

39 Harpending HC, Sherry ST, Rogers AR, Stoneking M: The genetic structure of ancient human populations. Curr Anthropol 1993;34:483–496.

40 Nocentini A: Power and limits of the genetic classification of languages. Mankind Q 1993;33:265–281.

41 Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J: Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 1988;85:6002–6006.