Address for correspondence and reprints: Dr. Guido Barbujani, Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, I-44100 Ferrara, Italia. E-mail: bjg@dns.unife.it

0002-9297/2000/6603-0040$02.00

Am. J. Hum. Genet. 66:1177–1179, 2000

Reconstruction of Prehistory on the Basis of Genetic Data

To the Editor:

In their letter, Torroni et al. (2000) express a radical disagreement with the assumptions, methods, and conclusions of Simoni et al.'s (2000) article. We think that their many criticisms can be reduced to four points:

1. Haplogroups have been incorrectly defined, and therefore the spatial autocorrelation analysis (SAAP) of their frequencies is flawed;
2. Aside from these errors, the frequencies of haplogroup J and of superhaplogroup JT do not match previous reports;
3. Only 22 polymorphic sites have been considered, and therefore the results of AIDA are flawed;
4. Meaningful patterns of mtDNA diversity can only be identified by the analysis of the distributions of recent mutations.

Point 2 is correct. In the article by Simoni et al. (2000), the column with the frequencies of haplogroup J is wrong, and the frequencies of several haplogroups in Galicia and Spain have been put in each other's places. We apologize to the readers for these errors. However, the correct data (see the erratum published in this issue of the *Journal*) were used in all the analyses, including SAAP, and therefore the autocorrelation results in table 5 in the article by Simoni et al. (2000) are correct. Before we consider the other points, it is important to exactly define the subject of this discussion.

The general question being asked in our study and in similar studies is: What combination of evolutionary factors is most likely to account for the current levels and patterns of genetic diversity? To answer this question, one has to study as many loci as possible and has to study them by using the same statistical methods, so that the results will be comparable. The methods of SAAP and AIDA are especially suitable, because they have long been used to summarize both protein (Sokal and Menozzi 1982; Sokal et al. 1989; O'Rourke et al. 1992; Epperson and Li 1996; Crawford et al. 1997) and DNA (Barbujani et al. 1995; Chikhi et al. 1998; Casalotti et al. 1999; Krings et al. 1999; Rickards et al. 1999) diversity.

Point 1: haplogroup definition and SAAP.—AIDA can be directly applied to any set of DNA data, whereas SAAP processes frequencies and therefore requires prior definition of the entities whose frequencies will be analyzed. AIDA found very little spatial structuring of mtDNA. To confirm this result, we reanalyzed the data by using SAAP, and hence we had to identify evolutionarily meaningful clusters of hypervariable region 1 (HVR-1) haplotypes, or haplogroups.

The categories that we used for that purpose—and that Torroni et al. (2000) question—were proposed by Richards et al. (1998) in a paper cosigned by two other authors of Torroni et al.'s letter. The classification of mitochondrial haplotypes is no easy task; there is an unresolved uncertainty about the best way to cluster and interpret mitochondrial data. Analysis, at the nucleotide level, of the whole mitochondrial genome will be a suitable approach only in the not-so-near future. Indeed, as we stated in the "Database" section of the article by Simoni et al. (2000), only three European samples have been typed at the RFLP level. There is no current alternative to the study of HVR-1 sequences, if one wants to understand whether mitochondrial variation shows any structuring in Europe. On the basis of 22 polymorphic sites, Richards et al. (1998) identified what they consider to be monophyletic clades in the HVR-1 phylogeny, and we chose to use those sites to define haplogroups. Of course, the frequencies of the haplogroups defined in this way do not perfectly overlap with those which are based on RFLPs (table 1 in Torroni et al. 2000). We have since discovered that they even differ between table 2 and figure 2 of Richards et al. (1998), because site 16189 is mentioned as being part of the "X motif" only in the former, and we trusted the latter. That is not our fault.

In quantitative terms, a nonparametric discriminant analysis that we ran on worldwide data shows that 15.3% of suitable mitochondrial data are assigned to different haplogroups that are based on RFLPs or on HVR-1 sequences, with variable levels of disagreement for the different haplogroups—for example, 7.1% for haplogroup J (for more details, see Simoni 2000); that 7.1% of uncertainty accounts, for example, for most of the persisting differences between the haplogroup J frequencies that we considered and the frequencies presented in Torroni et al.'s (2000) table 2.

Answering the criticisms raised by Torroni et al. (2000), which we do not feel are justified, would entail reclassification of just a few sequences, <20 for haplogroups X and U4, of a total of >800 distinct sequences. After they have been reallocated, the SAAP coefficients do not change, up to the second decimal place. Table 1

Table 1

SAAP of Superhaplogroup JT Frequencies

Upper Limits for Distance Classes (Pairs of Populations)	*II*[a]
500 (31)	−.06
1,000 (92)	−.12
1,500 (98)	.07
2,000 (96)	.01
2,500 (106)	−.16
3,000 (97)	.04
3,800 (79)	.02
5,310 (31)	−.04

[a] The overall probability of the correlogram is not significant.

shows that nothing much happens after modification of the frequencies of JT either. Only for haplogroup H do the criteria that we borrowed from Richards et al. (1998) result in serious ambiguities, and, therefore, we decided to classify several sequences in a residual group, which we called "others." At any rate, haplogroup definition influences only SAAP. Had data been seriously misclassified, the patterns described by AIDA and SAAP would have been discordant, which was not the case (tables 4 and 5 in Simoni et al. 2000).

In synthesis, the discrepancies between the haplogroup assignments in the article by Simoni et al. (2000) and in the letter by Torroni et al. (2000) are a consequence of the fact that we consistently used one set of criteria, which were based on HVR-1 sequence motifs (Richards et al. 1998), whereas they used a cocktail of criteria, which were based on unreleased material, HVR-1, and RFLPs; sadly, the recipe of that cocktail has not yet been disclosed to the public.

Point 3: AIDA.—Contrary to what Torroni et al. (2000) appear to believe, reducing the number of sites considered by AIDA does not reduce the probability of identifying a pattern. The AIDA coefficient *II* can be regarded as the increase in the average probability (across sites) of observing the same nucleotide in two DNA sequences sampled at a given geographic distance, with respect to two random sequences (see Barbujani 2000). The values of *II* depend essentially on the genetic variance among populations, F_{ST}. Like F_{ST} (Hartl and Clark 1997, p. 195), *II* is strongly influenced by variation at sites where the alternative nucleotides have intermediate frequencies, whereas poorly polymorphic sites will have a weaker effect. That is also intuitive; if one considers a long DNA segment in which substitutions are rare, even identical sequences will be only slightly more similar than average. Therefore, marginally significant patterns are more likely to be identified by a *further reducing* (and not an increasing, as Torroni et al. 2000 suggest) of the number of sites in the analysis, keeping the most variable sites. By analyzing only the 22 nucleotide positions that Richards et al. (1998) consider to be the most informative, therefore, we were enhancing the sensitivity of the method. Even then, a significant spatial structure could be detected only in southern Europe.

Point 4. Ancient haplogroups.—There is no reason why molecules or groups of molecules that originated through mutations that occurred a long time ago should show insignificant spatial patterns. Actually, the opposite is more likely. The European clines were first described in the distributions of very ancient variants, such as the molecules responsible for the AB0, Rh, and MN blood-group specificities (Menozzi et al. 1978). The alleles of some HLA loci, whose polymorphism probably dates back to several million years ago (Ayala et al. 1995), show very clear gradients in Europe (Sokal and Menozzi 1982).

There is a persisting confusion in this area. It seems necessary to repeat that ages of molecules are not ages of populations (Pamilo and Nei 1988; Templeton 1993; Donnelly 1996). The phylogenies of different molecules are notoriously different, and therefore they cannot possibly represent population histories (Langaney et al. 1992; Hartl and Clark 1997, p. 361). DNA mutates and recombines; populations disperse, fluctuate in size, mix, split, or become extinct. When a population expanded, all its alleles did (Excoffier and Schneider 1999), because prehistorical candidate migrants were not selected on the basis of their haplogroup. Two molecules may well have originated, say, 40,000 and 20,000 years ago, and yet both may owe their distributions to phenomena that occurred, say, 10,000 years ago. This may also explain why autocorrelation patterns at different loci are so similar in Europe.

This and other points in the letter by Torroni et al. (2000), in fact, raise another important issue. Is mtDNA HVR-1, after all, one of many thousands of genetic markers, albeit a highly polymorphic and extensively studied one? If so, there is no doubt that, once HVR-1 data have been cleaned up and, as far as possible, ambiguities have been removed, they *can and must* be treated like any other set of genetic data and be analyzed by the standard population-genetics methods. It is unclear whether Torroni et al. (2000) believe, instead, that the nature of HVR-1 variation is such that only nonconventional, specific numeric methods apply to it. To be accepted, such a view, which we do not share, must be stated explicitly and justified. For the time being, we maintain that repeatable, quantitative procedures should be applied to any genetic polymorphism. Synthetic statistical indices should be calculated and compared with the predictions of models based on external evidence. One way or another, probabilities or likelihoods should

be estimated. We see no other way to establish the relative merits of alternative hypotheses. Only in this way can the patterns of polymorphism, shown by various loci, be compared.

In particular, the comparison of genetic variation across loci is indispensable for the study of past migrations and expansions, because individuals (i.e., entire genomes), not single genes, migrate (Hartl and Clark 1997, pp. 189–197). Parallel analyses of the extensive available data, whether protein (Menozzi et al. 1978; Sokal et al. 1989; Cavalli-Sforza et al. 1994) or DNA (Chikhi et al. 1998; Casalotti et al. 1999; Simoni et al. 2000) polymorphisms, show that broad clines encompassing much of Europe are the rule, with mtDNA representing the most conspicuous (if partial) exception. Those patterns can only be due to a large-scale directional expansion, which archaeological evidence suggests took place either during the initial Paleolithic colonization of Europe or during the Neolithic demic diffusion.

LUCIA SIMONI,[1,2] FRANCESC CALAFELL,[2] DAVIDE PETTENER,[1] JAUME BERTRANPETIT,[2] AND GUIDO BARBUJANI[3]

[1]*Dipartimento di Biologia Evoluzionistica e Sperimentale, Università di Bologna, Bologna;* [2]*Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona; and* [3]*Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy*

References

Ayala FJ, Escalante A, O'Huigin C, Klein J (1995) Molecular genetics of speciation and human origins. In: Fitch WM, Ayala FJ (eds) Tempo and mode in evolution. National Academy Press, Washington, DC, pp 187–211

Barbujani G (2000) Geographical patterns: how to identify them, and why. Hum Biol 72:133–153

Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. Proc Natl Acad Sci USA 92:9171–9175

Casalotti R, Simoni L, Belledi M, Barbujani G (1999) Y-chromosome polymorphism and the origins of the European gene pool. Proc R Soc Lond Bio Sci 266:1959–1965

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G (1998) Clines of nuclear DNA markers suggest a recent, Neolithic ancestry of the European gene pool. Proc Natl Acad Sci USA 95:9053–9058

Crawford MH, Williams JT, Duggirala R (1997) Genetic structure of the indigenous populations of Siberia. Am J Phys Anthropol 104:177–192

Donnelly P (1996) Interpreting genetic variability: the effects of shared evolutionary history. In: Weiss K (ed) Variation in the human genome. John Wiley, Chichester, UK, pp 25–50

Epperson BK, Li T (1996) Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proc Natl Acad Sci USA 93:10528–10532

Excoffier L, Schneider S (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc Natl Acad Sci USA 96:10597–10602

Hartl DL, Clark AG (1997) Principles of population genetics, 3d ed. Sinauer Associates, Sunderland, MA

Krings M, Salem AH, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, et al (1999) mtDNA analysis of Nile River valley populations: a genetic corridor or a barrier to migration? Am J Hum Genet 64:1166–1176

Langaney A, Roessli D, van Blyenburgh NH, Dard P (1992) Do most human populations descend from evolutionary trees? Hum Evol 2:47–61

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

O'Rourke DH, Mobarry A, Suarez BK (1992) Patterns of genetic variation in Native America. Hum Biol 64:417–434

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62:241–260

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

Simoni L (2000) Pattern di diversità mitocondriale in Eurasia. Tesi di dottorato in Scienze antropologiche, XII ciclo, Università di Bologna, Bologna

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262–278

Sokal RR, Harding RM, Oden NL (1989) Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267–294

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

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. Am Anthropol 95:51–72

Torroni A, Richards M, Macaulay V, Forster P, Villems R, Nørby S, Savontaus M-L, et al (2000) mtDNA haplogroups and frequency patterns in Europe. Am J Hum Genet 66: 000–000 (in this issue)

Address for correspondence and reprints: Dr. Guido Barbujani, Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, I-44100 Ferrara, Italia. E-mail: bjg@dns.unife.it

0002-9297/2000/6603-0041$02.00