Am. J. Hum. Genet. 61:1011–1014, 1997

INVITED EDITORIAL
DNA Variation and Language Affinities

Guido Barbujani

Department of Biology, University of Ferrara, Ferrara

Populations tend to diverge genetically because of genetic drift, but their differences are reduced by the exchange of individuals or gametes—in brief, by gene flow. The evolutionary weight of drift depends on a property of each single population—its effective size. Conversely, the rate—and, therefore, the impact—of gene flow depends on the relationships between populations. Geographers say that everything is related to everything else but that close objects are more closely related than distant objects. That is the case for human populations too. Migrating to nearby localities is easier—and, therefore, more likely—than traveling very far; as a consequence, if one compares several pairs of populations, their allele frequencies will tend to be similar at short spatial distances. However, as the comparisons involve samples that are more and more distant, the respective levels of gene flow will be lower, and so will be their genetic similarity. Beyond a certain distance, pairs of allele frequencies will not be correlated (Kimura and Weiss 1964; Morton et al. 1968).

Within limited areas, geographic distance is therefore the main factor limiting gene flow, but, on a broader scale, additional factors are important, especially barriers. Mountains, seas, and deserts are examples of geographic barriers. But other, more elusive obstacles also profoundly affect the patterns of human genetic variation: cultural barriers. Of these, language boundaries are reasonably stable and easy to locate in space, and hence they are liable to study by quantitative approaches; but social or religious barriers may also exert similar effects.

Humans do not tend to easily cross language boundaries when choosing a partner. As a consequence, populations separated by such barriers are somewhat isolated from each other. The genetic consequences may be substantial. In Europe, for example, linguistic boundaries show increased rates of of allele-frequency change (Sokal et al. 1988; Barbujani and Sokal 1990; Calafell and Bertranpetit 1994), and it is well known that several inheritable diseases differ, in their incidence, between geographically close but linguistically distant populations (e.g., see de la Chapelle 1993). But languages have an even greater evolutionary significance, because linguistic affinities are also clues to population history (Renfrew 1991; Guglielmino et al. 1995). As Sokal (1988) wrote, a common language frequently reflects a common origin, and a related language indicates a common origin too, but farther back in time. Population admixture and linguistic assimilation should have weakened the correspondence between patterns of genetic and linguistic diversity. The fact that such patterns are, on the contrary, well correlated at the allele-frequency level (Sokal 1988; Cavalli-Sforza et al. 1988; Sokal et al. 1992) suggests that parallel linguistic and allele-frequency change were not the exception, but the rule.

Before the advent of molecular techniques in population genetics, some of us rather naively expected that DNA data would show even clearer correlations with language (Barbujani 1991). By now, although a few genetically differentiated linguistic isolates have actually been found (Lahermo et al. 1996; Stenico et al. 1996), it is clear that things are not that simple. At the DNA level, the human species is not subdivided into distinct groups more than it is at the protein level (Lewontin 1972; Barbujani et al. 1997). Of course, differences between populations can be demonstrated by use of molecular data, but, with a few exceptions (e.g., see Torroni et al. 1992), the patterns described so far do not seem to parallel the distribution of languages, especially for mitochondrial polymorphisms (e.g., see Ward et al. 1993; Watson et al. 1996; Bonatto and Salzano 1997). Among the explanations proposed is the different time scale of the evolutionary processes affectingDNA sequences vis-à-vis allele frequencies (Sajantila et al. 1995; Stenico et al. 1996). The former evolve essentially by the slow accumulation of mutations, whereas the frequencies of allelic variants, no matter whether they are estimated at the DNA level or at the protein level, fluctuate rapidly because of drift, probably paralleling linguistic change, which is also deeply affected by population contacts and isolation (Ruhlen 1992).

Received September 19, 1997; accepted September 22, 1997; electronically published October 29, 1997.

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

This article represents the opinion of the author and has not been peer reviewed.

© 1997 by The American Society of Human Genetics. All rights reserved.

0002-9297/00/6105-0005$02.00

However, a paper appearing in the current issue of the *Journal* (Poloni et al. 1997) shows that, if one looks closely enough, a good degree of congruence with language may be found even for DNA-sequence data. This finding is reassuring, because, after all, our population's history has been one; we may try to reconstruct it from archaeological, osteological, linguistic, or genetic sources, but we should eventually come up with a coherent set of inferences. Poloni et al. have collected a large set of population data on RFLP polymorphisms of the Y chromosome (p49a,f/*Taq*I) and of mtDNA, including 19 population samples that had been typed for both markers. By using a combination of recent nonparametric statistical methods and classic population-genetics theory, they have shown that linguistically related populations of Europe and Africa are also genetically close. In addition, they have estimated the most likely divergence time for each language family, on the basis of the levels of genetic differentiation among its samples.

A necessary assumption for those calculations is that all populations in a language family separated at once and never exchanged migrants afterward. Another assumption is that genetic differences accumulated owing to drift, which means quickly in small populations and slowly in large ones; therefore, to transform genetic variances into separation times, average population sizes had to be figured out, for each language family. Statements of this kind may seem unrealistic, but they provide the necessary starting point for comparisons that could not take place otherwise. If those assumptions are not far from true, which can be proved only by other studies of comparable scope, the estimated dates indicate that the demographic phenomena accompanying the spread of some language families of the Old World also left a significant molecular mark on our genome. The genetic differences among samples of the Afro-Asiatic family of northern Africa and western Asia and among members of the Indo-European family point to a split occurring sometime 9,000–7,000 years ago, which is in excellent agreement with dates based on archaeology (Renfrew 1991) and with comparative studies of languages and protein markers (Barbujani and Pilastro 1993). The Niger-Congo-Kordofanian family, including all Bantu languages, seems to have dispersed later, some 4,000 years ago, which, again, is in agreement with linguistic hypotheses (reviewed by Ruhlen 1991).

Less consistent with previous knowledge is the separation of Khoisan-speaking populations some 1,400 years ago. This family includes San (Bushmen) and other populations of southern Africa, who are regarded by many anthropologists and linguists as having been differentiated very anciently. Perhaps the Khoisan-speakers, now essentially small bands of hunter-gatherers, were much more numerous in the past. If so, the time necessary for them to reach the levels of genetic differentiation described in this study would increase relative to the time deduced by Poloni et al., who assumed constant population sizes within each language family.

In the study by Poloni et al., both the estimates of genetic diversity inferred from the Y chromosome and those inferred from mtDNA show similar modes of variation. This result is by no means trivial. Although it is easier to consider gene flow as a property of whole populations, in contemporary societies females and males seem to have different tendencies to migrate (Roberts 1988). Whether these differences existed in the past and have produced significant genetic effects is a matter of debate (Cavalli-Sforza and Minch 1997), but, in the cases in which comparative genetic analyses have been possible, the results for the two sexes did not seem to overlap. In Finland, the observed levels of sequence diversity suggest that the population underwent drastic reductions in size, but probably not at the same moment for females and males (Sajantila et al. 1996). The mitochondrial sequences of Basques resemble those of most other Europeans (Bertranpetit et al. 1995), but their Y chromosomes show unusual frequencies of two RFLP alleles (Semino et al. 1996). Poloni et al. performed what seems to be the first large-scale comparison of the patterns of variation for maternally and paternally transmitted components of the genome. Most of their samples come from Africa and Europe, and so general conclusions seem premature. Nevertheless, if these results are confirmed in other regions of the world, it will be clear that the migrational behavior of males and females has not differed much, on a global scale, during our recent evolutionary history. With only a few apparent exceptions, females and males have moved together along similar routes, which has resulted in similar levels and patterns of DNA differentiation. An intriguing observation is that the partial correlations with language are stronger for the Y chromosome than for mtDNA. On the contrary, it is generally believed that it is the mother who transmits the language to the child—and whose genes, therefore, should more closely match linguistic variation. Poloni et al. suggest that, when women were incorporated into a group speaking a different language, they passed to the future generations, along with their own genes, their husbands' language.

Since the global patterns of variation of female- and male-transmitted alleles seem to rather faithfully reflect patterns of language diversity, can we treat linguistic groups as evolutionary units and expect that what is true for one sample will also be true for its linguistic relatives? This study suggests that, by and large, that may be the case, at least in the sense that, with respect to the predictable relationships between populations, languages may be as informative as geography. But Poloni et al. are very cautious in drawing inferences of that

kind, and rightly so. Every region of the world has its own demographic history. Chance must have been important, both by affecting the evolutionary phenomena that we are trying to reconstruct and by introducing inaccuracies into the data on which we base our reconstructions. Therefore, the consistency of linguistic and genetic information seems more a hypothesis to test each time than a sensible assumption; in each case, one first has to establish that some overlapping exists, and then the problem becomes to quantify and interpret it.

For that purpose, specific statistical methods need be applied. After a period in which the simple novelty of the molecular data available seemed to justify adventurous evolutionary conclusions, we are now entering a phase of deeper reflection. The markers available and the populations studied are many, and contradictory results are emerging and will continue to emerge. Analysis of DNA data by quantitative methods is indispensable, but the traditional tools, developed for the analysis of allele frequencies, are not automatically suitable. A shortcut proposed by some is to treat molecular information in terms of frequencies of DNA lineages or variants. This approach has some advantages, but the sample sizes currently available, sometimes as few as just 10 individuals, represent a problem, because allele frequencies are necessarily estimated with large standard errors. In addition, sequence differences between individuals, which contain evolutionarily relevant information, are disregarded in this way. Conversely, the method by which Poloni et al. have estimated genetic variances—AMOVA (analysis of molecular variance) (Excoffier et al. 1992)—exploits all the information available. Along with improved tree-building algorithms (Bandelt 1994) and the techniques for quantifying spatial diversity—that is, autocorrelation indices for DNA analysis (AIDA) (Bertorelle and Barbujani 1995)—AMOVA has solid theoretical bases, takes sequence diversity into consideration, and relies on relatively robust assumptions about the evolutionary mechanisms supposed to have affected the populations. The importance of these methods is that they offer, at least in principle, a general, quantitative framework for evolutionary inferences based on DNA information. Whether the differences between two populations are "large" is an issue that can lead to endless discussion, but AMOVA, AIDA, and similar methods permit one to objectively test hypotheses. If population genetics will become more quantitative, as is to be hoped, then in the near future many of us will disagree on levels of significance rather than on broad—and hard-to-compare–evolutionary interpretations. That may prove to be no small step forward.

Acknowledgments

I thank Italo Barrai and Giorgio Bertorelle for critical reading of a draft manuscript. This paper is dedicated to Julie and Robert Sokal, with gratitude.

References

Bandelt HJ (1994) Phylogenetic networks. Verh Naturwiss Ver Hamburg 34:51–71

Barbujani G (1991) What do languages tell us about human microevolution? Trends Ecol Evol 6:151–155

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

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87:1816–1819

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–4519

Bertorelle G, Barbujani G (1995) Analysis of DNA diversity by spatial autocorrelation. Genetics 140:811–819

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59:63–81

Bonatto SL, Salzano FM (1997) A single and early migration for the peopling of the Americas supported by mitochondrial sequence data. Proc Natl Acad Sci USA 94:1866–1871

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

Cavalli-Sforza LL, Minch E (1997) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 61:247–251

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

de la Chapelle A (1993) Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 30:857–865

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Guglielmino CR, Viganotti C, Hewlett B, Cavalli-Sforza LL (1995) Cultural variation in Africa: role of mechanisms of transmission and adaptation. Proc Natl Acad Sci. USA 92: 7585–7589

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

Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L, Savontaus M-L (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am J Hum Genet 58:1309–1322

Morton NE, Miki C, Yee S (1968) Bioassay of population structure under isolation by distance. Am J Hum Genet 20: 411–419

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L (1997) Human genetic affinities for Y chromosome p49a,f/*Taq*I hapotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035 (in this issue)

Renfrew C (1991) Before Babel: speculations on the origins of linguistic diversity. Camb Archaeol J 1:3–23

Roberts D (1988) Migration in the recent past: societies with records. In: Mascie-Taylor CGN, Lasker TW (eds) Biological aspects of human migration. Cambridge University Press, Cambridge

Ruhlen M (1991) A guide to the world's languages. Vol. 1: Classification. Edward Arnold, London

Ruhlen M (1992) An overview of genetic classification. In: Hawkins JA,Gell-Mann M (eds) The evolution of human languages. Addison-Wesley, Redwood City, CA, pp 159–189

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, et al (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5:42–52

Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, Pääbo S (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc Natl Acad Sci USA 93:12035–12039

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS (1996) A view of the Neolithic demic diffusion in Europe through two Y chromosome–specific markers. Am J Hum Genet 59:964–968

Sokal RR (1988) Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 85:1722–1726

Sokal RR, Oden NL, Thomson BA (1988) Genetic changes across language boundaries in Europe. Am J Phys Anthropol 76:337–361

Sokal RR, Oden NL, Thomson BA (1992) Origins of the Indo-Europeans: genetic evidence. Proc Natl Acad Sci USA 89: 7669–7673

Stenico M, Nigro L, Bertorelle G, Calafell F, Capitanio M, Corrain C, Barbujani G (1996) High mitochondrial sequence diversity in linguistic isolates of the Alps. Am J Hum Genet 59:1363–1375

Torroni A, Schurr TG, Yang CC, Szathmary EJE, Williams RC, Schanfield MS, Troup GA, et al (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Na-Dene populations were founded by two independent migrations. Genetics 130:153–162

Ward RH, Redd A, Valencia D, Frazier B, Pääbo S (1993) Genetic and linguistic differentiation in the Americas. Proc Natl Acad Sci USA 90:10663–10667

Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Pääbo S (1996) mtDNA sequence diversity in Africa. Am J Hum Genet 59:437–444