Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 4670–4673, May 1993
Population Biology

Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily

GUIDO BARBUJANI* AND ANDREA PILASTRO

Dipartimento di Biologia, Universitá di Padova via Trieste 75, 35121 Padova, Italy

Communicated by Robert R. Sokal, February 16, 1993 (received for review December 7, 1992)

ABSTRACT **Contemporary patterns of allele frequencies allow inferences on past evolutionary processes. L. L. Cavalli-Sforza [(1988) *Munibe* 6, 129–137] and C. Renfrew [(1991) *Cambridge Archaeol. J.* 1, 3–23] proposed that neolithic farmers from the Near East propagated a group of related ancestral languages, from which three or four linguistic families developed. Here we show that genetic variation among Indo-European, Elamo-Dravidian, and Altaic speakers (grouped by some linguists in the Nostratic macrofamily) supports this hypothesis, whereas the evidence on Afro-Asiatic speakers is ambiguous. Gene-frequency clines within these linguistic families suggest that language diffusion was largely associated with population movements rather than with purely cultural transmission. Archeological, linguistic, and genetic evidence can be reconciled by envisaging a process of population growth and multidirectional dispersal from the Near East as the main factor shaping genetic and linguistic diversity in Eurasia and perhaps in North Africa.**

Affinities among the vocabularies and the morphologies of many Eurasian and African languages have led to the hypothesis that they derive from a common linguistic ancestor, Nostratic (1, 2). Schematically, language resemblance on such a large scale may be due to either cultural exchange between sedentary populations or a demographic process whereby the speakers of a language move into different areas (3). Both phenomena have had an influence on the distribution of contemporary languages, but their relative importance in specific cases is not always established.

Genetic information may allow us to discriminate between the results of cultural and demographic processes, for the latter leave long-lasting marks on allele-frequency distributions (4). If the same phenomena caused biologic and linguistic evolution, analysis of gene frequencies in groups defined by linguistic criteria can reveal otherwise elusive patterns (5), as is the case for subSaharan Africa (6). On the contrary, had languages spread mainly by cultural means, genetic variation within linguistic groups should only show the consequences of isolation by distance and of barriers to gene flow, if any.

The main demographic process associated with cultural change, demic diffusion, is the expansion into additional territories of a population whose size is increasing (4). Clines are its expected genetic consequence (7, 8). The frequencies of several alleles are clinally distributed in Europe and the Near East (4, 9) and correlate with the estimated dates for the onset of farming (4, 10–12). Genetic variation in Europe is therefore interpreted as largely reflecting a population expansion starting in the neolithic, 7000–8000 B.C., and permitted by an advanced subsistence technology, farming (7).

THE NOSTRATIC DEMIC DIFFUSION MODEL

Similarities between patterns of genetic and linguistic variation (13, 14) suggest that neolithic farmers also introduced languages into Europe (15) and possibly elsewhere. Renfrew (16) proposed that Nostratic was spoken by populations of the Near East more than 10,000 years ago. The ability to produce food increased the population densities (17). Populations then expanded outward in four major waves, with each wave propagating farming along with a protolanguage from which Indo-European, Elamo-Dravidian, Afro-Asiatic, and Altaic later developed (16) (Fig. 1). We refer to this hypothesis as the Nostratic demic diffusion (NDD) model, and we call these language families the NDD families.

Successive events must have blurred the genetic patterns thus determined. But if the demic diffusion model is correct, two biological consequences are to be expected. (*i*) Genetic variation among populations should be larger in the NDD than in other linguistic families. Indeed, other evolutionary pressures being equal, the former received from immigrant farmers different proportions of novel alleles, depending on their location along the routes of dispersal. Greater genetic diversity among population units should have resulted (20, 21). (*ii*) In the NDD families, one should observe clines radiating away from the Near East, analogous to those that allowed identification of demic diffusion in Europe (4).

MATERIALS AND METHODS

To test the NDD model, information on aboriginal populations from North and East Africa was incorporated into a data base of Eurasian allele frequencies (22), bringing its size from 960 to 3441 records. Each sample was assigned to the Near East, regardless of the language spoken (the geographical limits of this area were defined according to ref. 16), or to one of the following linguistic groups (18)—NDD groups: Indo-European of Europe, Indo-European of Asia and Elamo-Dravidian, Afro-Asiatic, and Altaic. Other families were Uralic, Caucasian, Sino-Tibetan, and Austric (Fig. 1). Recent immigrants and cases of ambiguous linguistic classification were excluded.

From these data, measures of genetic variance, F_{st} (23), were computed at 15 loci. Statistical independence was approximated by excluding from calculation one polymorphic allele for each locus. It has been estimated that 10^4 generations are needed for F_{st} to reach its equilibrium value in human populations (24); this statistic can therefore be applied to analyze processes that occurred $\approx$400 generations ago. The F_{st} values at different loci were jointly compared by the nonparametric Kruskal–Wallis test (25), NDD groups versus other families and Near Eastern populations.

Geographic and genetic distances (26) were computed between the Near East and each population in the eight language groups. Spearman's correlation coefficients (25) were then calculated between arrays of genetic and geographic distances separately for each locus and each language

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: NDD, Nostratic demic diffusion.

*To whom reprint requests should be sent at present address: Department of Statistical Sciences, University of Bologna, Bologna, Italy.

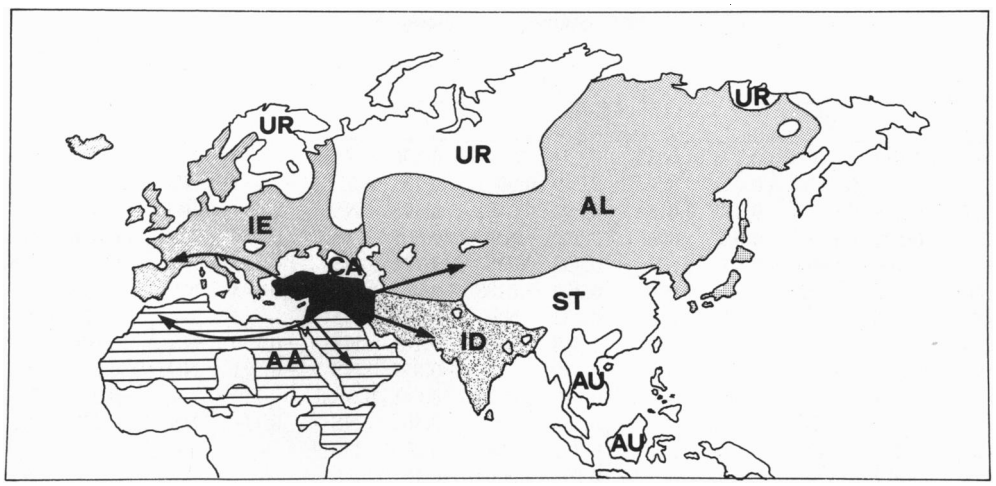

FIG. 1. Language families in Eurasia and North Africa (18). Region of origin of agriculture is black; arrows indicate proposed (16) routes of farming diffusion. Different shades of gray indicate Nostratic languages supposed (16) to have spread by NDD: IE, Indo-European of Europe; ID, Indo-European of Asia and Elamo-Dravidian (clumped together under the assumption that the spread of the former languages around 3000 B.C. involved negligible population replacement) (19); AA, Afro-Asiatic; AL, Altaic (not including populations that adopted Turkic in the early second millennium A.D.) (19). Unshaded areas: other families, either Nostratic languages supposedly (16, 19) unaffected by demic diffusion, or nonNostratic languages. UR, Uralic; CA, Caucasian; ST, Sino-Tibetan; AU, Austric.

group. The null hypothesis here is that genetic variation is random with respect to the Near East; the alternative hypothesis is that it is consistent with gene flow from there. In this context, random means that only the short-range decline of genetic similarity predicted by isolation by distance (27, 28) should be apparent, with no large scale clines. Since these hypotheses were tested independently in eight language groups, the significance threshold for the overall comparison of genetic and geographic distances was $P = 0.05/8 = 0.0063$.

RESULTS

Based on the F_{st} values calculated for 17 alleles (Table 1), the populations of the NDD groups appear significantly more differentiated than the others ($H = 4.25$; $P = 0.039$). The values of F_{st} are not correlated ($P = 0.304$) with the size of the area occupied by the respective language family. Hence, although spatial distance certainly affects genetic variances, in this case F_{st} does not seem to simply reflect the different geographic extension of the different language families.

Correlations of genetic and geographic distances are positive and significant at 9 of 15 loci tested for Indo-European of Europe, 8 of 15 for Indo-European of Asia and Elamo-Dravidian, 5 of 14 for Afro-Asiatic, and 8 of 15 for Altaic (Table 2). Overall, correlations are significant in all NDD groups, as assessed by Fisher's method for combining probabilities (25) ($P << 0.001$ for all groups except Afro-Asiatic, where $P < 0.001$). Equally significant results were obtained by using a different genetic distance measure (29). A typical distribution of allele frequencies is plotted in Fig. 2. At the glyoxalase locus, approximately longitudinal clines are evident for populations speaking Indo-European, Elamo-Dravidian, and Altaic languages, but not for Afro-Asiatic speakers.

In the other families, 8 of 53 correlations are individually significant. Fisher's test shows significant overall departure from chance expectations for Austric ($P < 0.001$) but not for Uralic, Caucasian, and Sino-Tibetan.

DISCUSSION

Genetic variation appears significantly larger in the NDD groups than in the other families or in the Near East. Clines compatible with diffusion of alleles from the Near East are numerous and highly significant among Indo-European, Elamo-Dravidian, Afro-Asiatic, and Altaic speakers. Three control groups show gene frequency patterns that are random with respect to the Near East. However, in a fourth control group, Austric, there are clines resembling those caused by the spread of alleles of Near Eastern origin. Their overall significance is the same as that calculated for Afro-Asiatic speakers. In synthesis, genetic variation in Eurasia and North

Table 1. Standardized allele-frequency variance, F_{st} (×10,000), in eight language groups and in the Near East

	NDD				Other families				
Allele	IE	ID	AA	AL	Near East	UR	CA	ST	AU
ABOA	94	166	145	129	161	261	51	107	219
ABOB	132	210	187	118	93	312	124	78	186
RHD	75	856	542	1150	71	1823	174	884	673
MNM	113	226	572	165	168	408	46	510	418
Duffy-a	122	1290	647	304	169	797	9	1497	1522
KEL	101	402	275	251	111	184	43	265	75
HP1	96	333	464	161	74	149	26	281	411
GC1	179	163	595	82	144	219	2	121	117
GLO1	134	343	929	893	187	121	99	58	338
ESD1	87	353	116	548	38	78	147	158	1023
PGM1	251	145	465	356	231	803	107	176	374
ACPA	133	262	639	292	302	679	120	152	249
ACPB	152	265	639	343	208	789	140	142	237
AK1	53	189	220	258	139	58	65	92	376
ADA1	80	147	185	172	68	252	40	195	199
PGDA	256	182	256	202	138	59	56	498	473
PGDC	78	170	254	194	138	59	56	507	1585

Linguistic codes are as abbreviated in Fig. 1. GPT was not analyzed because of lack of data in several families. The data base comprises the following 15 loci and 3441 samples whose average size is >300 (numbers of localities in the whole sample; numbers for Indo-European of Europe, Indo-European of Asia and Elamo-Dravidian, Afro-Asiatic, Altaic, and the Near East, respectively, are in parentheses): *ABO* (299; 109, 39, 33, 33, and 14), *RH* (231; 75, 35, 31, 25, 14), *MN* (281; 89, 41, 37, 33, 16), Duffy (178; 60, 23, 24, 9, 10), *KEL* (236; 100, 21, 30, 12, 12), *HP* (368; 152, 48, 30, 41, 15), *GC* (233; 93, 31, 16, 22, 8), *GLO* (171; 69, 20, 9, 15, 7), *ESD* (216; 84, 48, 5, 18, 6), *PGM* (261; 92, 45, 12, 39, 10), *ACP* (240; 92, 38, 20, 28, 10), *AK* (212; 75, 45, 19, 13, 6), *ADA* (165; 68, 25, 6, 15, 11), *PGD* (237; 73, 38, 20, 30, 9), *GPT* (113; 46, 10, 4, 18, 1).

Table 2. Spearman correlation coefficients of genetic and geographic distances (r), populations of eight language groups versus the Near East, and their one-tailed significance (P)

	IE		ID		AA		AL		UR		CA		ST		AU	
	r	P	r	P	r	P	r	P	r	P	r	P	r	P	r	P
ABO	0.472	0.0000	−0.038	NS	0.420	0.0088	−0.053	NS	0.196	NS	0.139	NS	−0.305	NS	−0.183	NS
RH	0.303	0.0045	0.462	0.0036	−0.002	NS	0.749	0.0001	0.778	0.0139	−0.251	NS	−0.236	NS	0.401	NS
MN	0.211	0.0238	0.478	0.0012	0.348	0.0185	0.294	0.0481	0.392	NS	0.647	0.0436	−0.362	NS	0.173	NS
Duffy	0.053	NS	0.599	0.0025	−0.012	NS	0.300	NS	0.565	0.0144	0.154	NS	0.678	0.0123	0.240	NS
KEL	−0.085	NS	0.760	0.0004	−0.053	NS	0.750	0.0064	0.121	NS	0.577	NS	0.343	NS	0.259	NS
HP	0.397	0.0000	0.527	0.0002	0.589	0.0007	0.300	0.0288	−0.277	NS	−0.143	NS	−0.594	NS	−0.288	NS
GC	−0.155	NS	−0.246	NS	0.485	0.0301	−0.233	NS	0.245	NS	—	—	0.126	NS	0.286	NS
GLO	0.401	0.0004	0.558	0.0075	0.150	NS	0.888	0.0005	−0.143	NS	−0.250	NS	−0.006	NS	0.273	NS
ESD	0.059	NS	0.353	0.0077	0.600	NS	0.389	NS	−0.857	NS	−0.571	NS	0.258	NS	0.585	0.0117
PGM	0.649	0.0000	−0.069	NS	0.018	NS	0.448	0.0029	0.641	0.0168	—	—	−0.285	NS	0.477	0.0187
ACP	0.052	NS	0.133	NS	0.171	NS	0.134	NS	0.400	NS	0.119	NS	−0.239	NS	0.082	NS
AK	0.012	NS	0.260	0.0424	−0.309	NS	0.763	0.0041	−0.070	NS	—	—	0.511	NS	0.448	NS
ADA	0.405	0.0000	0.032	NS	−0.058	NS	0.895	0.0004	−0.152	NS	—	—	0.200	NS	−0.099	NS
PGD	0.497	0.0000	−0.085	NS	0.387	0.0457	−0.167	NS	−0.333	NS	−0.500	NS	−0.791	NS	0.550	0.0021
GPT	0.314	0.0176	0.479	NS	—	—	0.191	NS	—	—	—	—	0.357	NS	0.305	NS

Linguistic codes are as abbreviated in Fig. 1. Geographic distances were calculated from the appropriate locality among the four places representing our best estimates of the origins of farming—namely, Catal Hüyük (Turkey) for Indo-European of Europe; Jericho (Jordan) for Afro-Asiatic; Ali Kosh (Iraq) for Indo-European of Asia and Elamo-Dravidian, Caucasian, Sino-Tibetan and Austric; and Jeitun (Turkmenia) for Altaic and Uralic (16). NS, not significant.

Africa corresponds broadly, but not totally, to a model of demic diffusion from the Near East, predicting clinal variation in the areas where languages of the Nostratic macrofamily are spoken.

The longitudinal gradients observed for Austric cannot be due to demic diffusion from the Near East, because in two of three cases (for *PGM* and *PGD*) they do not encompass Iran and the Indian subcontinent, where Indo-European and Elamo-Dravidian are spoken (Table 2). Other evolutionary processes must therefore have determined geographical structuring of allele frequencies in this region (see ref. 30) as well as in others, where, on the contrary, genetic distances from the Near East decrease as spatial distances increase (8 and 1 correlation coefficients are lower than −0.3 in the other families and among the NDD groups, respectively; Table 2).

Although historically implausible, the spread of Near Eastern genes in the Austric-speaking area is then statistically consistent with the observed correlations of genetic and geographic distances. This raises the question whether some

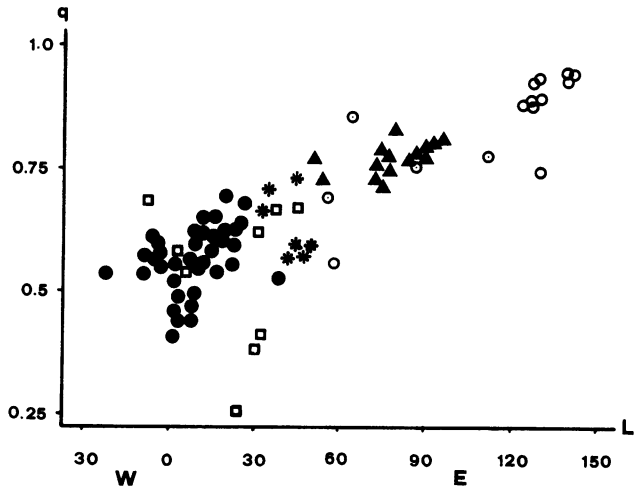

FIG. 2. Frequencies of the *GLO2* allele (q) in the Near East (*) and among Afro-Asiatic speakers (□), Indo-European speakers of Europe (●), Indo-European speakers of Asia and Elamo-Dravidian speakers (▲), and Altaic speakers (○) plotted against longitude (L). For the sake of clarity, part of the data points for Indo-Europeans, between 0° and 15° longitude East, have been omitted.

gradients in the NDD groups may also reflect processes other than neolithic demic diffusion of Nostratic speakers.

For Afro-Asiatic, the answer is not obvious. There, five gradients are consistent with demic diffusion from the Near East, but their overall significance is not higher than among Austric speakers. The area of origin of Afro-Asiatic languages is discussed among linguists (ref. 16 suggests that, contrary to what is stated by the NDD model, they spread from Africa into Asia), and the sampling localities available for this study were irregularly distributed. In conclusion, despite the significant departure from null expectations, genetic variation among Afro-Asiatic speakers may not necessarily reflect neolithic demic diffusion; other processes, perhaps partly overlapping with it, may have played a greater role.

On the other hand, although it is impossible to rule out that individual gradients may have a different origin, their number and significance among Indo-European, Elamo-Dravidian, and Altaic speakers is very high and much higher than in the Austric-speaking area. Computer simulations show that clines rapidly disappear if the expanding and recipient populations are not very different genetically (31); even under the hypothesis of NDD, therefore, gradients were expected only for a small fraction of the genes studied (as discussed in ref. 11). The occurrence of clines at so many unlinked loci, and their East–West orientation, do not suggest that climatic selection played any significant role. Genetic drift and short-range dispersal—i.e., isolation by distance—may account for genetic relatedness between near localities, but they certainly cannot explain regular patterns over thousands of kilometers (20, 27, 28, 32, 33). Large-scale population movements from the Near East are therefore the most likely explanation for the clines observed in these families.

Because these clines extend longitudinally, the genetic data are certainly consistent with other centers of origin of agriculture, provided they lie approximately at the same latitude as the Near East. A similar objection could be raised about the timing of the demic diffusion process, which cannot be inferred from allele-frequency data. However, the archeological evidence strongly suggests that farming spread from localities situated between Anatolia and Turkmenia and that this spread took place in the Neolithic (16, 34).

Disaggregation of a large body of genetic data by linguistic criteria has made evident clines (for *PGM*, *AK*, *ADA*, and *GPT*) that were not recognized when Eurasian populations

were jointly analyzed regardless of language (35, 36). Clines are the main evidence supporting spread of agriculture in Europe by population dispersal rather than by cultural transmission (4, 7, 8, 10). The present study provides the same type of evidence for most of Eurasia and possibly for North Africa. In addition, because the clines expected under the hypothesis of neolithic demic diffusion occur within the Nostratic macrofamily (and are seldom detected in other linguistic groups or when gene frequencies are analyzed regardless of language), farming and at least three families of Nostratic seem to have spread together. This strongly suggests that both came from the Near East, as proposed by Cavalli-Sforza (37) and Renfrew (16).

No historic process documented after the neolithic seems to have been associated with population growth to the degree sufficient to cause such a strong patterning of genetic variation (19). One could envisage the possibility that the gradients result from founder effects (38, 39) occurring earlier, during the initial colonization of Eurasia by *Homo sapiens sapiens*, also starting from the Near East (34, 40). However, the correspondence between Nostratic language families and regions of clinal variation would also imply that these families originated in the early Paleolithic (19). Few linguists would be ready to root current linguistic differences in so distant times (18).

Be that as it may, levels and patterns of genetic variation among Indo-European, Elamo-Dravidian, Altaic, and (to a lesser, but still significant, extent) Afro-Asiatic speakers are consistent with diffusion of alleles of Near Eastern origin. Genetic and linguistic variation in most of Eurasia (and, less likely, North Africa) may reflect the same generating process: languages spread as people moved.

We thank Silvia De Domenico and Gerard Whitehead for their help, and Luca Cavalli-Sforza, Bambos Kyriacou, Alex Peixoto, Colin Renfrew, Robert Sokal, and Michael Turelli for many suggestions and critical reading of previous versions of the manuscript. This study was supported by funds from the Italian Ministry for University and Scientific Research.

1. Dolgopolsky, A. B. (1987) *Mediterr. Lang. Rev.* **3**, 7–31.
2. Kaiser, M. & Shevoroshkin, V. (1988) *Annu. Rev. Anthropol.* **17**, 309–329.
3. Renfrew, C. (1989) *Trans. Philol. Soc.* **87**, 103–155.
4. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. (1978) *Science* **201**, 786–792.
5. Barbujani, G. (1991) *Trends Ecol. Evol.* **6**, 151–156.
6. Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C. & Langaney, A. (1987) *Yearb. Phys. Anthropol.* **30**, 151–194.
7. Ammerman, A. J. & Cavalli-Sforza, L. L. (1984) *The Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ. Press, Princeton, NJ).
8. Sgaramella-Zonta, L. & Cavalli-Sforza, L. L. (1973) in *Genetic Structure of Populations*, ed. Morton, N. E. (Univ. of Hawaii Press, Honolulu), pp. 128–135.
9. Sokal, R. R., Harding, R. M. & Oden, N. L. (1989) *Am. J. Phys. Anthropol.* **80**, 267–294.
10. Sokal, R. R. & Menozzi, P. (1982) *Am. Nat.* **119**, 1–17.
11. Sokal, R. R., Oden, N. L. & Wilson, C. (1991) *Nature (London)* **351**, 143–145.
12. Dennell, R. W. (1973) *Econ. Bot.* **27**, 329–331.
13. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6002–6006.
14. Cavalli-Sforza, L. L., Minch, E. & Mountain, J. L. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5620–5624.
15. Renfrew, C. (1987) *Archaeology and Language* (Jonathan Cape, London).
16. Renfrew, C. (1991) *Cambridge Archaeol. J.* **1**, 3–23.
17. Hassan, F. A. (1973) *Curr. Anthropol.* **14**, 535–543.
18. Ruhlen, M. (1987) *A Guide to the World's Languages* (Edward Arnold, London), Vol. 1.
19. Renfrew, C. (1992) in *Transition to Modernity*, eds. Hall, J. A. & Jarvie, I. C. (Cambridge Univ. Press, Cambridge, U.K.), pp. 11–68.
20. Slatkin, M. (1985) *Annu. Rev. Ecol. Syst.* **16**, 393–430.
21. Sokal, R. R. (1991) *Hum. Biol.* **63**, 589–606.
22. Barbujani, G., Jacquez, G. M. & Ligi, L. (1990) *Am. J. Hum. Genet.* **47**, 867–875.
23. Cavalli-Sforza, L. L. (1966) *Proc. R. Soc. London Ser. B* **164**, 362–379.
24. Takahata, N. (1991) *Genetics* **129**, 585–595.
25. Sokal, R. R. & Rohlf, F. J. (1981) *Biometry* (Freeman, San Francisco), 2nd Ed.
26. Cavalli-Sforza, L. L. & Edwards, A. J. F. (1967) *Am. J. Hum. Genet.* **19**, 223–257.
27. Kimura, M. & Weiss, G. H. (1964) *Genetics* **49**, 561–576.
28. Morton, N. E. (1975) *Theor. Pop. Biol.* **7**, 246–255.
29. Nei, M. (1972) *Am. Nat.* **106**, 283–291.
30. Bellwood, P. (1985) *Prehistory of the Indo-Malaysian Archipelago* (Academic, New York).
31. Sokal, R. R. & Jacquez, G. M. (1991) *Evolution* **45**, 152–168.
32. Wijsman, E. M. & Cavalli-Sforza, L. L. (1984) *Annu. Rev. Ecol. Syst.* **15**, 279–301.
33. Barbujani, G. (1987) *Genetics* **117**, 777–782.
34. Renfrew, C. (1992) *Man New Ser.* **27**, 445–478.
35. Barbujani, G. (1987) *Ann. Hum. Genet.* **51**, 345–353.
36. Farabegoli, A. & Barbujani, G. (1990) *Hum. Hered.* **40**, 313–321.
37. Cavalli-Sforza, L. L. (1988) *Munibe* **6**, 129–137.
38. Easteal, S. (1988) *Evol. Biol.* **23**, 49–84.
39. Boileau, M. G., Hebert, P. D. N. & Schwartz, S. S. (1992) *J. Evol. Biol.* **5**, 25–39.
40. Stringer, C. B. & Andrews, P. (1988) *Science* **239**, 1263–1268.