

# GM Haplotype Diversity of 82 Populations Over the World Suggests a Centrifugal Model of Human Migrations

Jean-Michel Dugoujon,<sup>1</sup> Serge Hazout,<sup>2</sup> France Loirat,<sup>2</sup> Bruno Mourrieras,<sup>2</sup> Brigitte Crouau-Roy,<sup>3</sup> and Alicia Sanchez-Mazas<sup>4\*</sup>

<sup>1</sup>Laboratory of Anthropobiology, Anthropology Center, UMR 8555 CNRS, 31000 Toulouse, France

<sup>2</sup>Team of Genomic and Molecular Bioinformatics, INSERM U436, University of Paris 7, 75251 Paris, France

<sup>3</sup>Evolution and Biological Diversity, Paul Sabatier University, 31062 Toulouse, France

<sup>4</sup>Laboratory of Genetics and Biometry, Department of Anthropology and Ecology, University of Geneva, 1227 Geneva, Switzerland

**KEY WORDS** GM polymorphism; immunoglobulins; human genetic diversity; human evolution

**ABSTRACT** This study investigates the GM genetic relationships of 82 human populations, among which 10 represent original data, within and among the main broad geographic areas of the world. Different approaches are used: multidimensional scaling analysis and test for isolation by distance, to assess the correlation between genetic variation and spatial distributions; analysis of variance, to investigate the genetic structure at different hierarchical levels of population subdivision; genetic similarity map (geographic map distorted by available genetic information), to identify regions of high and low genetic variation; and minimal spanning network, to point out possible migration routes across continental areas. The results show that the GM polymorphism is characterized by one of the highest amounts of genetic variation observed so far among populations of different continents

( $F_{ct} = 0.3915$ ,  $P < 0.0001$ ). GM diversity can be explained by a model of isolation by distance (IBD) at most continental levels, with a particularly significant fit to IBD for the Middle East and Europe. Five peripheral regions of the world (Europe, west and south sub-Saharan Africa, Southeast Asia, and America) exhibit a low level of genetic diversity both within and among populations. By contrast, East and North African, Southwest Asian, and Northeast Asian populations are highly diverse and interconnected genetically by large genetic distances. Therefore, the observed GM variation can be explained by a “centrifugal model” of modern humans peopling history, involving ancient dispersals across a large intercontinental area spanning from East Africa to Northeast Asia, followed by recent migrations in peripheral geographic regions. *Am J Phys Anthropol* 125:175–192, 2004. © 2004 Wiley-Liss, Inc.

The analysis of human polymorphisms at broad geographic levels is very powerful to investigate the origins and history of modern human populations. Much has been done in recent years to improve our knowledge of human peopling history by the analysis of DNA markers, mostly used in phylogenetic reconstructions (e.g., Cruciani et al., 2002; Ingman et al., 2000; Quintana-Murci et al., 1999; Underhill et al., 2000, 2001; Wells et al., 2001). In addition, relevant results have been obtained over many decades by the study of classical polymorphisms (synthesized by Cavalli-Sforza et al., 1994). Although the analysis of classical markers does not allow us to estimate the precise molecular diversity of human populations, gene frequencies are known for hundreds of populations worldwide, and continue to provide highly suitable data to explore spatial patterns of gene frequencies throughout the world (e.g., Bosch et al., 1997; Chen et al., 1995; Dugoujon et al., 1995; Sanchez-Mazas et al., 2001; Weber et al., 2000).

Two major kinds of statistical analyses may be used to investigate the genetic diversity of human populations. Firstly, spatial methods like principal components analysis (Jolliffe, 1986), multidimensional scaling analysis (Kruskal, 1964), or spatial autocorrelation (Sokal and Oden, 1978) are useful to explore the patterns of genetic variation in relation to geography, showing that the latter is most often a

Grant sponsor: Centre d'Anthropologie; Grant sponsor: UMR 8555; Grant sponsor: CNRS; Grant sponsor: Swiss National Fund for Scientific Research; Grant number: 3100-039847.93.

\*Correspondence to: Alicia Sanchez-Mazas, Laboratory of Genetics and Biometry, Department of Anthropology and Ecology, University of Geneva, 12 rue Gustave-Revilliod, 1227 Geneva, Switzerland. E-mail: alicia.sanchez-mazas@anthro.unige.ch

Received 30 January 2003; accepted 2 July 2003.

DOI 10.1002/ajpa.10405

Published online 26 March 2004 in Wiley InterScience (www.interscience.wiley.com).

good predictor of human population relationships (Cavalli-Sforza et al., 1994). Secondly, analyses of genetic structure, including hierarchical analysis of variance (Cockerham, 1969, 1973; Long, 1986; Excoffier et al., 1992), wotbling (Barbujani et al., 1989) or the algorithm of Monmonier (1973), are used to identify genetically homogeneous clusters and/or to find and test genetic boundaries (i.e., zones of limited gene flow). The methods mentioned above are complementary approaches using different statistics. Genetic distances like those of Prevosti et al. (1975), Nei (1972), or  $F_{st}$ -derived measures (e.g., the distance of Reynolds et al., 1983) are used for spatial analyses involving comparisons with geographic distances. On the other hand, hierarchical analyses often use fixation indexes like the well-known  $F_{st}$ , although this measure is sometimes confusing.

The  $F_{st}$  "fixation index" was first introduced by Wright (1965) to quantify the decrease in heterozygosity in a subdivided population. When mating is not random in a population but depends on a population structure, each subdivision (or deme) undergoes genetic drift, and thus shows a tendency toward allele *fixation*. A measure of population subdivision, here equivalent to deme differentiation within such a population, is then provided by the "fixation index"  $F_{st}$ .  $F_{st}$  may also be interpreted as the level of inbreeding within the demes relative to the total population, and hence the notations  $s$  (for subdivision) and  $t$  (for total population). In addition, for a review  $F_{st}$  can be defined as the standardized variance of allele frequencies among demes (Excoffier, 2001).

When a higher level of subdivision is considered, that is, several groups (or clusters) of populations, one may be interested in knowing whether such a group structure is significant. The question then is: what proportion of the total variance of allele frequencies among populations could be explained, respectively, by the group structure, and by the population (or subdivision) structure within the groups? The  $F_{st}$ , which is here defined as the allele frequency variation among subdivisions (or populations) relative to the total population, is then subdivided into a part due to differences among clusters (or groups), relative to the total population (and summarized by the  $F_{ct}$  index) and a part due to differences among subdivisions (or populations) within clusters (or groups), and summarized by the  $F_{sc}$  index. The remaining variation is due to differences among individuals *within* populations (Excoffier et al., 1992; Schneider et al., 2000; Weir, 1996; Weir and Cockerham, 1984). These three proportions of variance (among continental groups, among populations within continental groups, and within populations) were initially estimated for the human species by Lewontin (1972) on a set of protein data and, later on, by Barbujani et al. (1997) and Jorde et al. (2000) on DNA polymorphisms. They are still very useful both to understand the global pattern of population differentiations and to

investigate possible discrepancies between different polymorphisms.

GM allotypes carried by IgG immunoglobulins are informative serological markers whose transmission as haplotypes may provide more information than single polymorphisms considered as independent markers (reviewed by Lefranc and Lefranc, 1990). Four systems of allotypic antigens have been described so far: GM, AM, EM, and KM (Grubb, 1956; Kunkel et al., 1969; Ropartz et al., 1961; van Loghem et al., 1984; Vyas and Fudenberg, 1969). The GM system with 18 allotypes shows the greatest degree of polymorphism, and is therefore most often studied in population genetics. GM allotypes are located on the  $\gamma$  constant domains of the heavy chains of the three IgG subclasses IgG1: G1M(1, 2, 3, 17), IgG2: G2M(23), and IgG3: G3M(5, 6, 10, 11, 13, 14, 15, 16, 21, 24, 26, 27, 28), and many of them have been defined at the DNA level (Dard et al., 2001). The genes coding for these domains are closely linked on chromosome 14 (14q 32.3) (Flanagan and Rabbitts, 1982; Milstein et al., 1984). GM allotypes are thus inherited in fixed combinations, or haplotypes. As GM haplotype frequencies are very heterogeneous among ethnic groups (Sanchez-Mazas, 1990; Steinberg and Cook, 1981), they are very useful in assessing genetic relationships among populations from different geographic areas.

The aim of this study is to explore the genetic diversity of GM haplotype frequencies observed in 82 world populations in relation to their spatial distribution. In addition to data previously published in the literature, this investigation includes 10 original population samples from sub-Saharan Africa (6), North Africa (1), the Arabian Peninsula (1), and New Guinea (2), which are tested for the whole set of GM allotypes. The GM data of these 82 populations are analyzed by applying several complementary methods. First, a multidimensional scaling analysis (MDS) is carried out to explore the overall genetic diversity pattern of GM haplotype frequencies. Population genetic structure is then investigated by a hierarchical analysis of variance (ANOVA) among 10 population groups representing different broad continental areas (sub-Saharan Africa, North Africa, Europe, Southwest Asia, Northeast Asia, Southeast Asia, Oceania, the circum-Arctic area, North and Central America, and South America). We also estimate the correlation between genetic and geographic distances, both globally and at continental levels, to assess the role of isolation by distance (IBD) in shaping GM diversity patterns. Finally, we construct a genetic similarity map (GS map), which is a distorted geographic map reflecting genetic proximities among populations (Hazout et al., 1993; Mourrieras et al., 1997), and a minimum-spanning network (MS network) superimposed on the genetic similarity map. These latter analyses are designed to identify genetically homogeneous regions, and to see whether and how they are interconnected. Overall, our aim is to go thoroughly into

the interpretation of GM diversity observed worldwide by using in part original data, and several complementary analytical methods allowing us to draw inferences about past migration events.

## SUBJECTS AND METHODS

### Population samples

In total, 82 populations corresponding to 30,469 subjects were used and represent a very large set of populations from different geographic origins, namely: 12 populations from Europe, 17 from Africa, 28 from Asia and the Near East, 7 from Oceania and 18 from America (Table 1 and Fig. 1). Most data (72 samples) are taken from the Gene[VA] database (Sanchez-Mazas, personal database), where population data originally taken from the literature and personal communications were systematically checked for the reliability of their haplotype frequencies (Sanchez-Mazas, 1990). The remaining populations (10 samples) were previously tested for GM in Toulouse (personal unpublished data) and the results are presented for the first time within the scope of the present study (Tables 2 and 3). These new data represent populations from New Guinea (Papuan-speakers from southern and northern fringes), Algeria (Kabiles), north Yemen (Yemenites), Djibuti (Issa), Ethiopia (Amhara-Tigré), Mali (Bwa), the Ivory Coast (Baoule), the Central African Republic (Aka Pygmies), and Madagascar (Malagasy). Among the other population samples (previously published data), only samples including a minimum of 100 individuals were chosen, with the exception of two populations (93 Negritos from Mindanao, and 97 Basques from Guipuzcoa). Population samples were also chosen on the basis of data quality and GM typing: all populations were tested for G1M(1, 2, 3, 17) and G3M(5, 6, 10, 11, 13, 14, 15, 16, 21, 24). Overall, 9 common haplotypes were considered: GM 1, 17; 21, GM 1, 2, 17; 21, GM 3; 5\*, GM 1, 17; 10, 11, 13, 15, 16, GM 1, 17; 5\*, GM 1, 17; 10, 11, 13, 15, GM 1, 3; 5\*, GM 1, 17; 5, 6, 10, 11, 14, and GM 1, 17; 5, 6, 11, 24, as well as three uncommon haplotypes: GM 3; 5\*, 15, 24, GM 1, 17; 5, 14, and GM 1, 17; -. Within one additional category ("others"), we included some very rare haplotypes (frequencies between 0.001–0.026, concerning 15 populations). The data for G2M(23) were taken into account for the three Oceanian populations (New Guinean southern and northern fringes and Australian Aborigines), in order to discriminate the peculiar GM 1, 17; 23; 5\* haplotype, often observed in that area, from the more common GM 1, 17; -; 5\*.

### Statistical analyses

**Estimation of haplotype frequencies and test for Hardy-Weinberg equilibrium.** GM haplotype frequencies were estimated from the phenotypic distributions of the 10 new populations tested by a standard maximum likelihood procedure using the program Genef2, which is Lathrop's modification

of Yasuda's (1968) algorithm. The hypothesis of Hardy-Weinberg equilibrium was tested in these populations by a classical goodness-of-fit chi-square test (Sokal and Rohlf, 1994).

**GM haplotype diversity.** In a preliminary step of data analysis, we made up a clustering of populations by using a conventional UPGMA algorithm (Sneath and Sokal, 1973), based on a Prevosti (Prevosti et al., 1975; Wright, 1978) genetic distance matrix obtained from the GM haplotype distributions (results not shown). This hierarchical structure was taken into account to arrange linearly the 82 populations in Figure 2, where genetically close populations are plotted in adjacent or nearby positions, and where the detailed GM haplotype frequencies are shown. In Figure 2, we also indicated the gene diversity of each population, estimated by

$$h = \frac{2n}{2n-1} \left(1 - \sum_{i=1}^k x_i^2\right),$$

where  $n$  denotes the number of individuals in the sample and  $x_i$  the  $i$ th gene frequency over a total of  $k$  haplotypes (Nei, 1987).

**Analysis of population genetic structure.** Pair-wise coancestry coefficients defined as linearized Fsts (Reynolds et al., 1983) were computed from GM haplotype frequencies among the 82 populations by using the program package Arlequin (Schneider et al., 2000). The resulting genetic distance matrix was then used for a multidimensional scaling analysis (MDS) (Kruskal, 1964), using the software NTSys (Rohlf, 2000). An analysis of genetic variance (ANOVA) was performed with Arlequin to assess whether the worldwide GM diversity was significantly structured by geography. As shown in Table 1, an a priori geographic structure was defined by grouping the populations into 10 geographic clusters inspired from those described by Cavalli-Sforza et al. (1994, p. 126): Europe, North Africa, sub-Saharan Africa, Southwest Asia (corresponding to the Near East and India), Northeast Asia (East Asian populations located between latitudes 30° and 60° north), Southeast Asia (East Asian populations located south of latitude 30° north), Oceania, the circum-Arctic area (Siberian and Eskimo), North and Central America, and South America. A hierarchical analysis of variance was performed, where the total genetic variability was subdivided into several variance components, i.e., among groups (Fct), among populations within groups (Fsc), and within populations (Excoffier, 2001). As explained above, the usual Fst index represents the total variation among populations, independent of a group structure. These indexes were tested for significance by permuting haplotypes, individuals, or populations among individuals, populations, or population groups (Excoffier et al., 1992; Schneider et al., 2000). After each permutation round, the indexes were recomputed in order to get their empirical distribution after a high number (here 10,000) of permutations. We also es-

TABLE 1. Populations studied

Label	Sample size	Population name	Geographic location	Geographic coordinates	Reference
Group 1: sub-Saharan Africa					
1:	144	Issa	Djibuti	11°3N 43°1E	Present study
2:	175	Amahra & Tigre	Ethiopia, Gondar	12°4N 37°3E	Present study
3:	140	Sidamo	Ethiopia	5°4N 37°3E	Steinberg (1973)
4:	600	Bwa	Mali	15°N 5°W	Present study
5:	198	Baoule	Ivory Coast	7°N 5°W	Present study
6:	556	Mandenka	Senegal, Kedougou	12°4N 12°1W	Blanc et al. (1990)
7:	255	Sara Majingay	Chad, Ndila	8°N 17°E	Hiernaux (1976)
8:	900	Aka Pygmies	Central African Republic	4°2N 18°4E	Present study
9:	162	Babinga Pygmies	Central African Republic	4°2N 18°4E	Cavalli-Sforza et al. (1969)
10:	189	Mlozi	Western Zambia	15°1S 23°1E	Jenkins et al. (1970)
11:	214	Xhosa	South Africa	33°6S 25°4E	Jenkins et al. (1970)
12:	394	!Kung	Botswana, Dobe	18°6S 21°2E	Steinberg et al. (1975)
13:	231	Malagasy	Madagascar	19°S 46°E	Present study
Group 2: North Africa					
14:	227	Kabile	Algeria, Kabily	35°N 3°E	Present study
15:	107	Berbers	Tunisia, Douiret-Chenini	32°6N 10°1E	Chaabani et al. (1984)
16:	845	Twareg	Sahara	20°N 5°E	Lefèvre-Wittier (1982)
17:	251	Arabs	Tunisia, Monastir	35°5N 10°6E	Helal et al. (1981)
Group 3: Europe					
18:	2314	French	France, all regions	46°N 2°E	Blanc and Ducos (1986)
19:	146	Corsicans	Corsica	41°6N 8°4E	Blanc and Ducos (1986)
20:	419	Portuguese	Portugal, all regions	39°3N 8°1W	Pereira and Manso (1975)
21:	97	Basques	Spain, Guipuzcoa	43°N 1°5W	Calderón et al. (1998)
22:	869	Croatians	Pag, Olib, Silba Islands	44°N 15°E	Borot et al. (1991), Dugoujon et al. (1989)
23:	184	Hungarians	Hungary, Budapest	47°3N 19°E	Schanfield et al. (1975a)
24:	684	Czech	Central Bohemia	49°3N 14°4E	Schanfield et al. (1975b)
25:	407	Scots	Orkney Islands	59°2N 2°6W	Welch et al. (1973)
26:	119	Finns	Aland Islands	60°N 19°6E	Steinberg et al. (1974)
27:	100	Finns	Ristiina	61°3N 27°1E	Steinberg et al. (1974)
28:	199	Rumanians	Rumania, Bucarest	44°3N 26°1E	Johnson et al. (1977)
29:	441	Sardinians	Northern Sardinia	40°4N 8°5E	Piazza et al. (1976)
Group 4: Southwest Asia					
30:	205	Yemenites	North Yemen	15°N 44°E	Present study
31:	303	Guilanians	Iran, Caspian Sea	37°2N 49°4E	Van Loghem et al. (1977)
32:	195	Sunni	Lebanon, Beyruth	33°5N 35°3E	Lefranc et al. (1978)
33:	119	Hindus	India, Andhra Pradesh	17°2N 78°3E	Van Loghem et al. (1985)
34:	203	Koya Dora	India, Andhra Pradesh	17°4N 80°6E	Van Loghem et al. (1985)
35:	163	Hindus	India, Delhi	28°4N 77°1E	Schanfield and Kirk (1981)
36:	137	Naicker	India, Madras	13°N 80°2E	Schanfield and Kirk (1981)
Group 5: Northeast Asia (> 30°N)					
37:	137	Buriat	Russia, North Baikal	51°6N 107°4E	Matsumoto et al. (1984)
38:	106	Mongolians	Mongolia, Huhehote	40°5N 111°4E	Zhao and Lee (1989)
39:	170	Tibetans	Western Tibet	32°N 90°E	Matsumoto (1984)
40:	343	Japanese	Japan, Osaka	34°4N 135°3E	Matsumoto and Takatsuki (1968)
41:	195	Han	China, Beijing	39°6N 116°3E	Matsumoto et al. (1986)
42:	135	Han	China, Hefei	31°6N 117°2E	Matsumoto et al. (1986)
43:	103	Hui	China, Changji	44°N 87°2E	Zhao and Lee (1989)
Group 6: Southeast Asia (<30°N)					
44:	114	Miao	China, Taijiang	26°4N 108°2E	Zhao and Lee (1989)
45:	119	Shui	China, Sandu	25°6N 107°6E	Zhao and Lee (1989)
46:	93	Negritos	Philippines, Mindanao	8°6N 125°3E	Matsumoto et al. (1979)
47:	127	Negritos	Philippines, Luzon	15°3N 120°4E	Omoto et al. (1978)
48:	1467	Balinese	Indonesia, Bali	8°4S 115°1E	Blanc and Breguet (1985)
49:	152	Han	China, Guangzhou	23°1N 113°2E	Matsumoto et al. (1986)
50:	175	Senoi	Malaysia, Perak	4°4N 101°E	Steinberg and Lie-Injo (1972)
51:	293	Filipino	Philippines, Samar	11°1N 125°E	Yogore and Schanfield (1981)

(continued)

TABLE 1. (Continued)

Label	Sample size	Population name	Geographic location	Geographic coordinates	Reference
Group 7: Oceania					
52:	277	Papuan-speakers	New Guinea, Southern fringe	9°S 140°E	Present study
53:	423	Papuan-speakers	New Guinea, Northern fringe	2°S 140°E	Present study
54:	386	Tolai	New Britain	4°1S 152°1E	Curtain et al. (1971)
55:	185	Enga (Papuan-speakers)	New Guinea, Highlands	5°3S 140°3E	Curtain et al. (1971)
56:	254	Micronesians	Caroline Islands, Kusaie	5°5N 162°E	Steinberg and Morton (1973)
57:	271	Fijians	Fiji, Viti Levu	18°1S 178°3E	Schanfield (1971)
58:	113	Australian Aborigines	Australia, Mornington Island	16°4S 139°1E	Curtain et al. (1972)
Group 8: Circum-Arctic					
59:	403	Chukchi	North East Kamchatka	59°2N 163°1E	Sukernik and Osipova (1982)
60:	640	Selkup	North West Siberia	64°N 82°E	Sukernik et al. (1992)
61:	731	Forest Nentsi	Russia, Ob and Taz	64°6N 77°5E	Sukernik et al. (1992)
62:	171	Evens	Russia, Beryozovka	67°4N 155°5E	Posukh et al. (1990)
63:	204	Evens	Russia, Seyban-Kujhel	65°2N 130°E	Posukh et al. (1990)
64:	100	Yupik	Alaska, New Chaplino	64°3N 172°2W	Sukernik and Osipova (1982)
65:	283	Inuit	Groenland, Angmassalik	65°4N 38°W	Nielsen et al. (1971)
66:	365	Inuit	Melville Peninsula	69°1N 83°6W	McAlpine et al. (1974)
67:	692	Yupik	South West Alaska	60°5N 161°5W	Petersen et al. (1991)
Group 9: North and Central America					
68:	136	Navajo	Keams Canyon	35°5N 110°1W	Williams et al. (1985)
69:	100	Ojibwa	Canada, Ontario	51°5N 94°W	Szathmary et al. (1974)
70:	401	Cree	Canada plains	50°N 70°W	Callegari-Jacques et al. (1993)
71:	476	Pima	USA, Arizona	33°N 111°4W	Williams et al. (1985)
72:	173	Huasteco	Mexico	19°3N 99°1W	Steinberg et al. (1967)
73:	143	Mixteco	Mexico	17°N 97°W	Dugoujon et al. (1995)
74:	482	Guaymi	Panama	8°6N 79°3W	Gershowitz and Neel (1978)
Group 10: South America					
75:	165	Wayana	French Guyana	2°5N 52°3W	Daveau et al. (1975)
76:	294	Wayampi	French Guyana	3°1N 52°2W	Dugoujon et al. (1994)
77:	363	Baniwa	North West Brazil	1°3N 68°1W	Salzano et al. (1986)
78:	3447	Yanomama	Venezuela	5°4N 67°3W	Gershowitz and Neel (1978)
79:	440	Cayapo	Brazil	10°S 53°W	Salzano et al. (1973)
80:	464	Xavantes	Brazil	13°S 46°W	Shreffler and Steinberg (1967)
81:	759	Quechua	Peru	12°S 77°W	Quilici (1975)
82:	507	Macushi	Brazil, Guyana	5°1N 60°4W	Neel et al. (1977)

timated the average population gene diversity ( $H$ ) for each geographic group taken separately.

**Test for isolation by distance (IBD).** According to Rousset (1997), one way to evaluate the fit of a set of genetic data to a model of isolation by distance is by performing a regression analysis between the logarithm of the geographic distance and the  $F_{st}/(1 - F_{st})$  coefficient between populations. Here, we used a Mantel test running 10,000 permutations (Mantel, 1967) to assess the correlation between the two matrices (software NTSys). Geographic distances were computed from latitude and longitude coordinates on the basis of the arc length of a sphere, and transformed to natural logarithms (Ray, 2002).

**Genetic similarity map (GS map).** Recently, Mourrieras et al. (1997) developed an approach that combines genetic diversity between populations and geographic factors. The method leads to the construction of a genetic similarity map, which corresponds to a distorted geographic map that integrates the spatial proximity of populations into the analy-

sis of genetic diversity. The principle of the method is to move the set of points (initially representing the geographic locations of populations) iteratively to obtain the best bidimensional representation of “normalized genetic distances” (the genetic distances to be compared with the geographic distances must be multiplied by a constant), while keeping the geographic constraints (i.e., the map outlines). The genetic distances used here are Prevosti’s genetic distances (Prevosti et al., 1975; Wright, 1978). The method allows one to draw the successive distortions of the continental outlines until reaching a final configuration, the genetic similarity map (GS map). The GS map is interpreted according to the type of observed geographic map distortions. Several cases can be distinguished: 1) a stretching of a geographic region associated with an alignment of populations may indicate a genetic cline that corresponds to a large variation of a given haplotype frequency, 2) an inflating of a region in various directions specifies a large genetic diversity of the populations located in this region, resulting from large variations of sev-

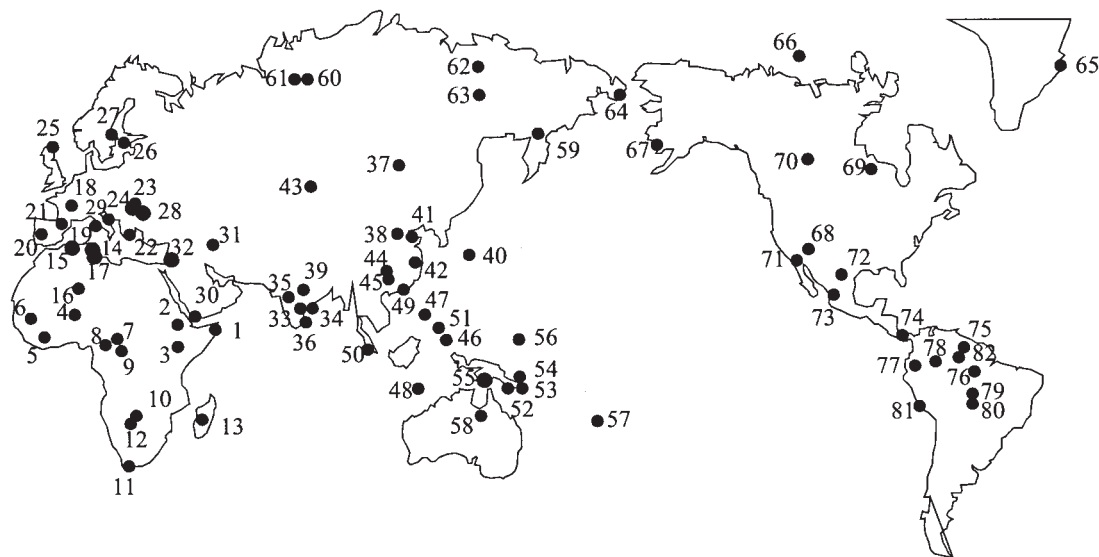


Fig. 1. Geographical location of 82 populations studied (see Table 1 for references).

eral haplotype frequencies, and 3) a size reduction of a geographic region is a sign of weak variation in haplotype frequencies, i.e., of low genetic diversity.

**Minimal spanning network (MS network).** A minimal spanning network<sup>1</sup> consists of building a tree (i.e., a graph without cycles) by joining the minimal distances (Jain and Dubes, 1988). A superimposition of an MS network on the genetic similarity map makes it possible to visualize the main geographic axes linking genetically close populations. As the GS map aims at bringing the populations closer together when they are genetically close while conserving the continental outlines, the minimal spanning network allows one to point out possible routes of population migrations over the world.

## RESULTS

### GM haplotype frequencies and Hardy-Weinberg equilibrium

GM phenotypes and estimated haplotype frequencies of the 10 original population samples are presented in Tables 2 and 3, respectively. The fit to Hardy-Weinberg equilibrium is given by the results of the chi-square test (Table 3, bottom). It is significant at the 1% level for the Aka ( $P = 0.008$ ), the Malagasy ( $P = 0.0004$ ), and the south and north Papuans ( $P = 0.0001$  and  $P < 0.0001$ , respectively). In Aka, Malagasy and southern Papuans, the high chi-square value is due to very rare phenotypes (expected numbers  $< 5$ ), and Hardy-Weinberg equilibrium cannot be rejected on this ground. In northern Papuans, Hardy-Weinberg deviation is due to discrepancies between observed and expected numbers in genotypes involv-

ing haplotypes GM 1, 17; 5\* and GM 1, 3; 5\*. This could be an effect of admixture with Austronesians (see below).

### GM haplotype diversity on different continents

Figure 2 shows the detailed GM haplotype frequencies and gene diversities ( $h$ ) of the 82 populations considered in this study. The graph reveals large frequency variations among populations from different continents, a virtual genetic homogeneity within each continent, and a continuous genetic variation across contiguous continental areas. East Africans (Issa, Sidamo, and Amhara-Tigré) exhibit the highest level of internal genetic diversity, followed by several North African (Twareg), Southwest Asian (Hindus and Yemenite), and Northeast Asian and Siberian (Hui, Selkup, Mongolian, northern Han, southern Han, Japanese, Forest Nentzi, and Buriat) populations, as indicated both by their heterogeneous GM genetic profiles and their high gene diversities ( $h \sim 0.65-0.75$ , Fig. 2, right). By contrast, Amerindian populations, West and South Africans, and many Southeast Asian populations are among the less diverse ( $h < 0.3-0.4$ ).

### Global GM genetic differentiations

Figure 3A,B shows the results of the multidimensional scaling analysis (MDS) carried out for the 82 populations. The final stress of 0.158 is "good" (Rohlf, 2000), meaning that the display of the population set (as a point set) in the 3 dimensions used here closely describes the total genetic variation. The 3-dimensional configuration has been plotted on two 2-dimensional planes (A and B), which point out four main axes of genetic differentiation, along which a large part of the populations are aligned. Figure 3A, where the first and second dimensions

<sup>1</sup>"Network" is used in place of "tree" to underline the fact that it has no root.

TABLE 2. GM phenotypes

Allotypes tested	1 2 3 17 5 6 10 11 13						1 2 3 17 5 6 10 11 13			1 2 3 17 5 6 10 11 13		
	14 15 16 21 24						South Papuan	North Papuan	Amhara Tigre	14 15 16 21 23 24 28		
Phenotypes	Aka	Bwa	Baoule	Malagasy	Algerians	Yemenites	Phenotypes				Phenotypes	Issa
3; 5*	—	—	—	—	68	52	3;-; 5*	—	—	1	3;-; 5*	2
							3; 23; 5*	—	—	11	3; 23; 5*	9
1 3; 5*	—	—	—	26	—	—	1 3;-; 5*	1	27	—	—	—
							1 3; 23; 5*	1	—	1	—	—
1 3 17; 5*	—	—	—	53	44	49	1 3 17;-; 5*	—	11	5	1 3 17;-; 5*	9
							1 3 17; 23; 5*	9	14	41	1 3 17; 23; 5*	13
1 3 17; 5* 6	—	—	—	9	2	—	1 3 17;-; 5* 6	—	—	1	1 3 17;-; 5* 6	3
							1 3 17; 23; 5* 6	—	—	3	1 3 17; 23; 5* 6	9
1 3 17 5* 6 24	—	—	—	20	1	5	1 3 17;-; 5* 6 24	—	—	3	1 3 17;-; 5* 6 24	1
							1 3 17; 23; 5* 6 24	—	—	4	—	—
1 3 17; 5* 21	—	—	—	2	55	36	1 3 17;-; 5* 21	2	73	3	1 3 17;-; 5* 21 28	6
							1 3 17; 23; 5* 21	—	4	6	1 3 17; 23; 5* 21	7
											28	—
1 3 17; 5* 15	—	—	—	6	—	—	1 3 17; 23; 5* 15	—	—	2	1 3 17;-; 5* 15	1
							1 3 17;-; 5* 15 16	—	—	2	1 3 17;-; 5* 15 16	2
1 3 17; 5* 15 16	—	—	—	—	4	6	1 3 17; 23; 5* 15 16	—	—	2	—	—
							1 17;-; 5*	—	1	26	1 17;-; 5*	10
1 17; 5*	589	278	63	45	12	18	1 17; 23; 5*	161	24	—	—	—
							1 2 17;-; 21	3	18	1	1 2 17;-; 21 28	2
1 2 17; 21	—	—	—	—	2	1	—	—	—	—	1 2 17; 23; 21 28	1
							1 2 17;-; 5* 21	1	—	3	1 2 17;-; 5* 6	1
1 2 17; 5* 21	—	—	—	—	4	—	1 2 17; 23; 5* 21	6	1	—	—	—
							1 2 3 17;-; 5* 21	—	2	—	—	—
1 2 3 17; 5*	—	—	—	—	—	2	1 2 3 17; 23; 5* 21	—	—	3	1 2 3 17; 23; 5* 21	1
1 2 3 17; 5* 21	—	—	—	—	2	3	—	—	—	—	28	—
							1 17;-; 21	22	199	5	1 17;-; 21 28	12
1 17; 21	—	—	—	1	15	7	1 17; 23; 21	—	2	—	—	—
							1 17;-; 5* 21	—	5	21	1 17;-; 5* 21, 28	18
1 17; 5* 21	—	3	—	1	18	7	1 17; 23; 5* 21	71	38	—	—	—
							1 17;-; 5* 6	—	—	8	1 17;-; 5* 6	14
1 17; 5* 6	2	24	6	17	—	—	—	—	—	—	1 17;-; 5* 6 28	1
							1 17;-; 5* 6 21 28	—	—	—	1 17;-; 5* 6 21 28	2
1 17; 5* 6 15	—	6	4	—	—	—	1 17;-; 5* 6 15 16	—	—	2	1 17;-; 5* 6 15 16	2
							1 17;-; 5 6 10 11 14	—	—	5	1 17;-; 5 6 10 11 14	1
1 17; 5 6 10 11 14	—	3	—	6	—	—	1 17;-; 5 6 10 11 14 21	—	—	1	1 17;-; 5 6 10 11 14	8
1 17; 5 6 10 11 14 21	—	—	—	1	—	—	—	—	—	—	21 28	—
							1 17;-; 5* 6 24	—	—	4	1 17;-; 5* 6 24	4
1 17; 5* 6 24	248	116	66	32	—	10	1 17;-; 5 6 11 24	—	—	1	—	—
							—	—	—	—	—	—
1 17; 5 6 11 24	29	6	14	5	—	—	—	—	—	—	—	—
							—	—	—	—	—	—
1 17; 5 6 11 21 24	—	2	—	—	—	—	—	—	—	—	—	—
							—	—	—	—	—	—
1 17; 5 6 11 14 24	—	—	—	2	—	—	—	—	—	—	—	—
							—	—	—	—	—	—
1 17; 5 6 10 11 14 24	2	8	3	—	—	—	—	—	—	—	1 17; 23; 5 6 10 11	1
							1 17;-; 5 6 10 11 13 15 16	—	—	1	14 24 28	—
1 17; 5 6 10 11 13 15 24	1	29	13	1	—	—	—	—	—	—	—	—
							24	—	—	—	—	—
1 17; 10 11 13 15	2	13	2	—	—	—	—	—	—	—	—	—
							1 17;-; 10 11 13 15 16 21	—	—	3	1 17;-; 10 11 13 15	1
1 17; 10 11 13 15 21	—	3	—	—	—	—	—	—	—	—	16 21 28	—
							—	—	—	—	—	—
1 17; 10 11 13 15 16	—	—	—	—	—	1	—	—	—	—	—	—
							—	—	—	—	—	—
1 17; 10 11 13 15 16 21	—	—	—	—	—	1	—	—	—	—	—	—
							1 17;-; 5* 15	—	4	—	—	—
1 17; 5* 15	27	109	27	4	—	—	1 17;-; 5* 15 16	—	—	2	—	—
							—	—	—	—	—	—
1 17; 5* 15 16	—	—	—	—	—	7	—	—	—	4	1 17;-; 5* 15 16	2
							—	—	—	—	—	—
Total	900	600	198	231	227	205	Total	277	423	175	Total	144

<sup>1</sup> 5\* = 5 10 11 13 14  
;-; indicates the absence of the corresponding phenotype.

are plotted, shows three such axes, roughly corresponding to Europeans, Southwest Asians, and North Africans for the first axis (right bottom), Southeast Asians (with some Oceanians) for the second axis (right top), and Northeast Asians, Amerindians, and circum-Arctic populations (again with some Oceanians) for the third axis (left). Figure 3B, where the first and third dimensions are plotted, displays a fourth axis (right bottom) corresponding to sub-Saharan Africans (the case of population 52 is discussed below). These trends are mostly associated to the four haplotypes GM 3; 5\*, GM 1, 3; 5\*, GM 1, 17; 21, and GM 1, 17; 5\*, respectively, whose frequency variations are particularly high between

broad geographic areas. Other common haplotypes are GM 1, 17; 10, 11, 13, 15, 16 and GM 1, 2, 17; 21, in Northeast Asia; GM 1, 17; 10, 11, 13, 15, particularly frequent in sub-Saharan Khoisan populations; and GM 1, 17; 5, 6, 11, 24, in most sub-Saharan Africans (Sanchez-Mazas, 1990; Steinberg and Cook, 1981).

The 10 populations typed in the present study are genetically close to other populations from the same geographic region (Fig. 3):

Issa (population 1) and Amhara-Tigré (population 2) are very similar to Sidamo (population 3), where GM 1, 17; ±23; 21 (0.243 for Issa, and 0.128 for

TABLE 3. GM haplotype frequencies

Allotypes tested	1 2 3 17 5 6 10 11 13 14 15 16 21 24							1 2 3 17 5 6 10 11 13 14 15 16 21 23 24 (28) <sup>3</sup>					
	Population N (individuals)	Aka 900	Bwa 600	Baoule 198	Malagasy 231	Algerians 227	Yemenites 205	Population N (individuals)	South Papuans 277	North Papuans 423	Amhara Tigre 175	Population N (individuals)	Issa 144
1 17; 21	—	0.0067	—	0.013	0.2311	0.1438	1 17;-; 21	0.2017	0.5633	0.1284	1 17;-; 21 28	0.2344	
1 2 17; 21	—	—	—	—	0.0178	0.0098	1 17; 23; 21	—	0.0032	—	1 17; 23; 21 28	0.0084	
3; 5* <sup>1</sup>	—	—	—	—	0.5375	0.5	1 2 17;-; 21	0.0182	0.0252	0.0202	1 2 17;-; 21 28	0.0176	
1 17; 10 11 13 15 16	—	—	—	—	0.0088	0.039	3;-; 5*	—	—	0.057	3;-; 5*	0.1095	
1 3; 5*	—	—	—	0.3074	—	—	3; 23; 5*	—	—	0.2198	3; 23; 5*	0.1474	
1 17; 5*	0.8083	0.6733	0.5682	0.4026	0.1982	0.2659	1 17;-; 10 11 13 15 16	—	—	0.04	1 17;-; 10 11 13 15 16	0.0243	
1 17; 10 11 13 15	0.0178	0.1442	0.1212	0.0238	—	—	1 3;-; 5*	0.0215	0.1787	—	1 3;-; 5*	—	
1 17; 5 6 10 11 14	0.0022	0.0366	0.0328	0.0784	0.0044	—	1 3; 23; 5*	0.0056	0.0081	0.0118	1 3; 23; 5*	—	
1 17; 5 6 11 24	0.1717	0.1392	0.2778	0.1407	0.0022	0.0366	1 17;-; 5*	0.0065	0.0099	0.4	1 17;-; 5*	0.2695	
1 17;-	—	—	—	—	—	—	1 17; 23; 5*	0.694	0.0982	—	1 17;-; 5* 28	0.0045	
1 2 17; 5*	—	—	—	—	—	—	1 17;-; 10 11 13 15	—	—	0.0114	1 17; 23; 5*	—	
1 17; 5 14	—	—	—	0.0341	—	—	1 17;-; 5 6 10 11 14	—	—	0.0714	1 17;-; 10 11 13 15	0.0035	
Chi-square	17.38	17.14	4.92	43.58	13.47	31.01	1 17;-; 5 6 11 24	—	—	0.04	1 17;-; 5 6 10 11 14	0.121	
Df <sup>2</sup>	6	10	6	17	20	17	1 17;-;	0.0525	0.1134	—	13 14	0.039	
P-value	0.0080	0.0713	0.5541	0.0004	0.8563	0.0199	1 2 17;-; 5*	—	—	—	1 17;-; 5 6 11 24	0.0174	
							1 17;-; 5 14	—	—	—	1 17;-;	—	
											1 2 17;-; 5*	0.0035	
											1 17;-; 5 14	—	
											Chi-square	62.44	
											Df	55	
											P-value	0.2289	

<sup>1</sup> 5\* = 5 10 11 13 14

<sup>2</sup> Df: degrees of freedom

<sup>3</sup> Issa tested for GM 28.

Amhara-Tigré, respectively) and GM 3; ±23; 5\* (0.260 and 0.277, respectively) exhibit high frequencies, in addition to GM 1, 17; -23; 5\* (0.274 and 0.4, respectively). These populations reveal a high internal genetic diversity ( $h = 0.788$  and  $0.742$ , respectively). Also, their genetic profile is intermediate between those observed in western and southern sub-Saharan Africans, on one hand, and in North Africans and Southwest Asians, on the other hand.

Aka Pygmies (population 8) are very close to Babinga Pygmies (population 9), and to Sara Majingay (population 7, Nilo-Saharan speakers). This tribe is also the most homogeneous among all tested sub-Saharan African populations, with a particularly high frequency of haplotype GM 1, 17; 5\* (0.808) and a low gene diversity ( $h = 0.317$ ).

The western Africans Baoule (population 5) and Bwa (population 4) are very close to Mandenka from Eastern Senegal (population 6) and Mlozi from Zambia (population 10), with high GM 1, 17; 5\* (0.568 and 0.673, respectively) and GM 1, 17; 5, 6, 11, 24 (0.278 and 0.139, respectively) frequencies.

The Malagasy (population 13) exhibit a very peculiar genetic profile, with both frequent GM 1, 17; 5\* (0.403) and GM 1, 17; 5, 6, 11, 24 (0.141), as in sub-Saharan Africans, and GM 1, 3; 5\* (0.307), as in Southeast Asians. This result confirms that the peopling of Madagascar by Austronesians from Southeast Asia highly contributed to the present genetic profile of Malagasy (Soodyall et al., 1996), although the genetic contribution of sub-Saharan Africans (as far as the GM 1, 17; 5\* haplotype frequency is considered to indicate to such a contribution) appears to be predominant here. Admixture may be partly responsible for the high gene diversity observed in this

population ( $h = 0.719$ ), as well as for its deviation from Hardy-Weinberg equilibrium, as previously observed for HLA (Renquin et al., 2001).

The populations from north Yemen (population 30) and Algeria (population 14) are close to other North Africans and Southwest Asians. All these populations are genetically intermediate between sub-Saharan Africans and Europeans, and generally exhibit high gene diversities ( $\sim 0.5-0.7$ ).

In New Guinea, we analyzed two different areas of the highlands: the southern fringe, and the contiguous northern fringe of the Schrader Mountains. The distribution of GM haplotypes shows that northern (population 53) and southern (population 52) populations are strongly differentiated from each other. Northern Papuans exhibit high GM 1, 17; ±23; 21 (0.566) and GM 1, 3; ±23; 5\* (0.187) frequencies, which make them genetically close to some Oceanians of Austronesian origin (e.g. Fijians, population 57). Southern Papuans exhibit a high GM 1, 17; 23; 5\* (0.694) frequency, which gives them a very peculiar genetic profile and an extreme differentiation<sup>2</sup> compared to other Oceanians. The results probably reflect the old colonization of the Highlands (Stoneking et al., 1990), with a rapid

<sup>2</sup>Due to the lack of G2M 23 typing in most populations, Gm 1, 17; 23; 5\* and Gm 1, 17; -; 5\* were not differentiated for the multidimensional scaling analysis. As a consequence, southern Papuans (population 52) appear to be close to sub-Saharan Africans, but they are in fact more differentiated. Actually, G2M 23 is not observed in sub-Saharan African populations tested thus far for this allotype (in the Sidamo sample; Steinberg, 1973; and tested by more than 2,000 individuals from three ethnic in groups from West Africa analyzed by Sanchez-Mazas and Dugoujon, unpublished results).

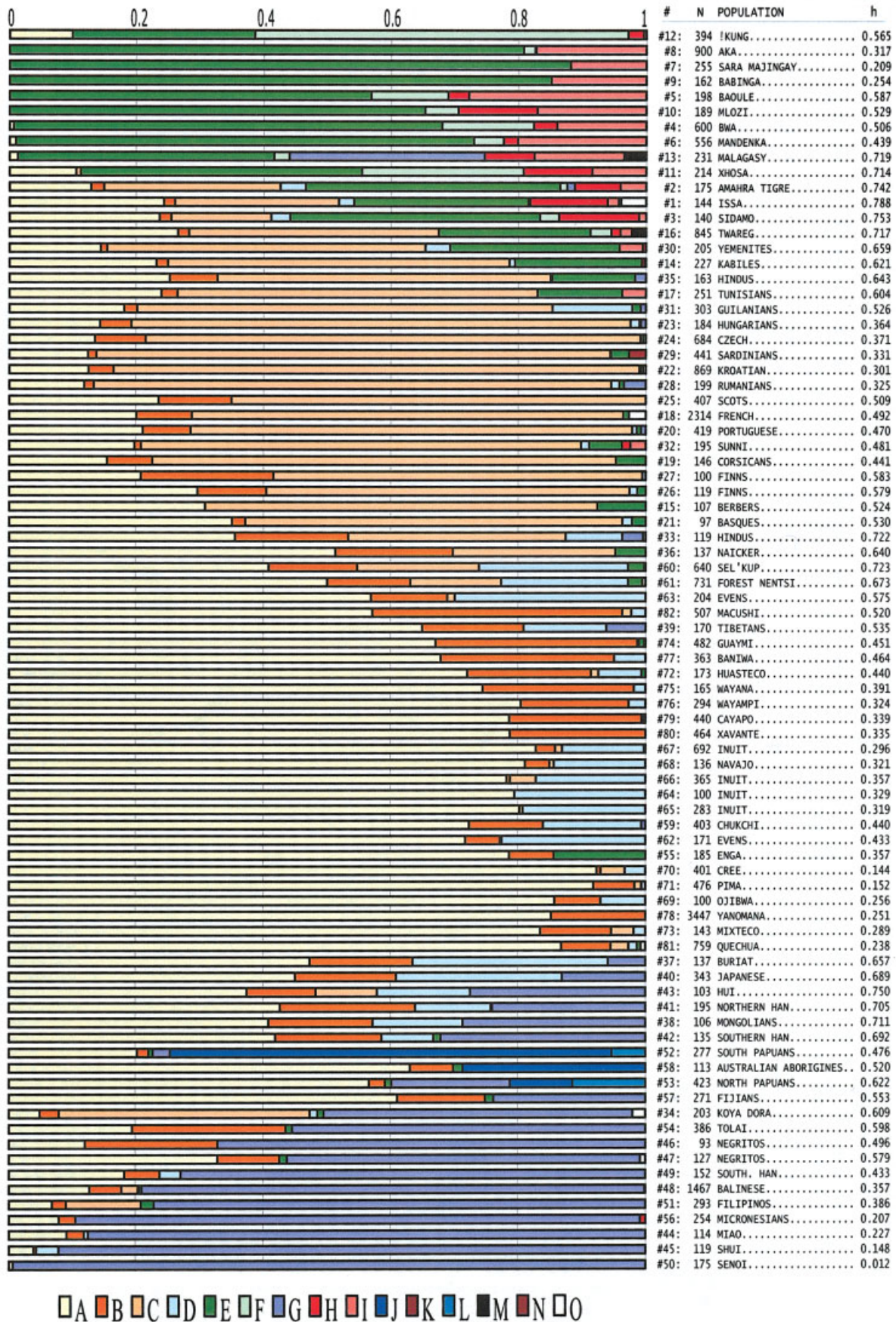


Fig. 2. (Legend page 184.)

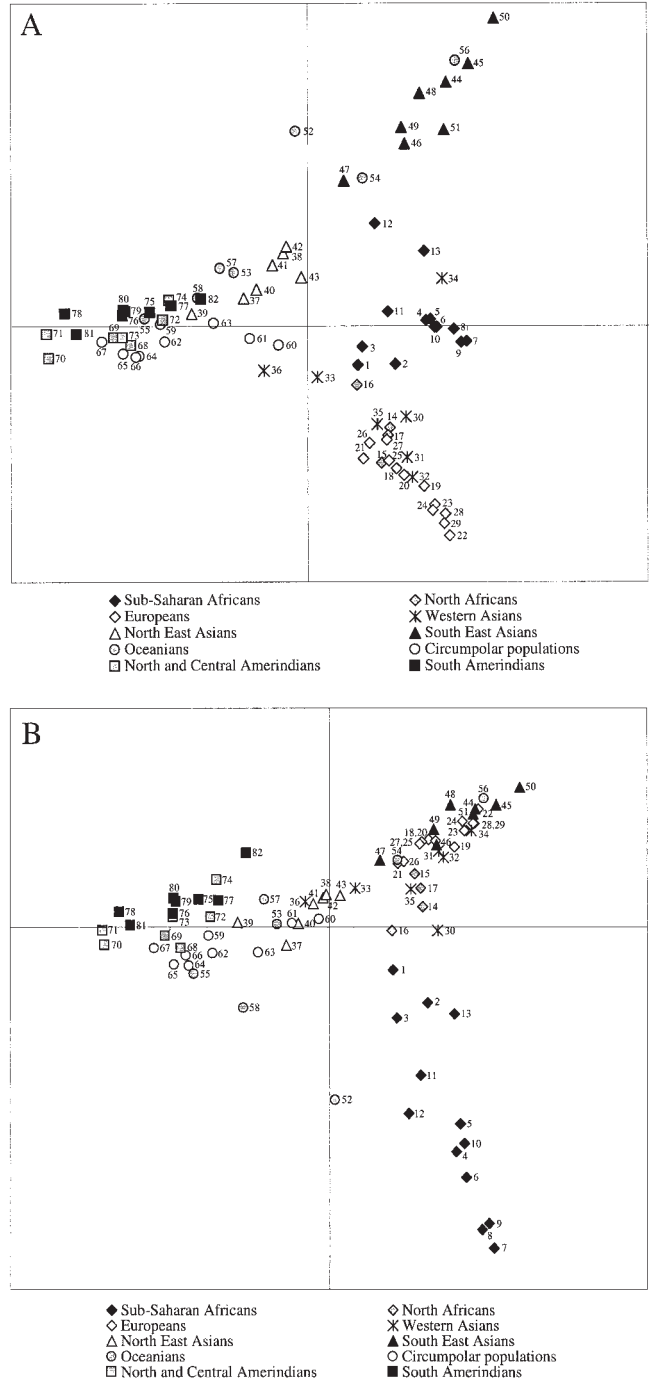
genetic drift in some populations (e.g., southern Papuans), but also possible contacts with Austrone- sians (e.g., for northern Papuans). It is striking that the uncommon GM 1, 17; - haplotype is frequent in Papuans, with particularly high values (up to 0.23) in some villages (detail not shown). Also, Oceanian populations as a whole show very heterogeneous genetic profiles and do not cluster together (Fig. 3A,B).

Overall, and with the exception of Oceanians, our results indicate a close relationship between genetic and geographic differentiations for the GM polymorphism. This relationship was further assessed by two different approaches, analysis of variance (ANOVA) and test for isolation by distance (IBD). According to the first approach (Table 4, "global structure"), we note a very high level of differentiation among the 10 geographic groups defined in Table 1 ( $F_{st} = 0.3915, P < 0.0001$ ), although the greatest part of the variability is found *within* populations, as for other genetic systems (Barbujani et al., 1997; Lewontin, 1972). This result was strongly suggested by the MDS (Fig. 3A,B). The tests for IBD (Table 5, "global structure") also indicate that GM distributions fit a model of isolation by distance at the global scale ( $r = 0.316, P = 0.0001$ ), which underscores the genetic continuity between adjacent geographic areas, despite a high differentiation among continental groups.

**GM Genetic diversity within continental areas**

In Table 4 ("continental structures") are also shown the  $F_{st}$  values among populations, for each geographic group. Among the 10 regions initially considered, Oceania (group 7) is by far the most heterogeneous ( $F_{st} = 0.338$ ), followed by sub-Saharan Africa (group 1,  $F_{st} = 0.169$ ) and Southwest Asia (group 4,  $F_{st} = 0.138$ ). However, after collapsing Northeast Asian and Southeast Asian groups within a single East Asian cluster (group 5 + 6), the resulting  $F_{st}$  (0.269) is the highest after Oceania. This suggests a high genetic diversity among East Asian populations, possibly related to the apparent northeast vs. southeast differentiation (Fig. 3A), as formerly observed by Schanfield and Gershowitz (1973), Matsumoto (1988), and Zhao and Lee (1989). (Poloni et al., to appear; Sanchez-Mazas et al., to appear). Moreover, if we split the sub-Saharan African group into Niger-Congo (roughly corresponding to west and south sub-Saharan Africans) and Afro-Asiatic (here, equivalent to East Africans) popula-

**Fig. 2.** GM haplotype frequencies observed in 82 populations (see Table 1 for references). Population names are preceded by sample labels (#) and sample sizes, and followed by estimated heterozygosities. Order of populations was determined in part on basis of their clustering in an UPGMA tree (not shown). Haplotypes are: A: 1, 17; 21; B: 1, 2, 17; 21; C: 3; 5\*; D: 1, 17; 10, 11, 13, 15, 16; E: 1, 17; 5\*; F: 1, 17; 10, 11, 13, 15; G: 1, 3; 5\*; H: 1, 17; 5, 6, 10, 11, 14; I: 1, 17; 5, 6, 11, 24; J: 1, 17; 23; 5\*; K: 3; 5\*; L: 1, 17; -, M: 1, 17; 5, 14; N: 1, 2, 17; 5\*; O: others.



**Fig. 3.** Multidimensional scaling analysis for 82 populations of world, based on coancestry coefficients computed from GM haplotype frequencies. **A:** Axes I (horizontal) and II (vertical). **B:** Axes I (horizontal) and III (vertical). Stress = 0.158 (good). Population numbers correspond to those listed in Table 1.

tions (groups 1a and 1b in Table 4), the resulting low  $F_{st}$  observed in Niger-Congo is low (0.055), even lower than that observed in Southeast Asians (group 6,  $F_{st} = 0.060$ ) for a close number of populations ( $n = 7$  and  $n = 8$ , respectively). This suggests that Niger-Congo populations exhibit very homogeneous genetic profiles. Aboriginal populations from Amer-

TABLE 4. GM genetic structure of 82 populations arranged into 10 geographic groups (see Table 1)

Global structure (n = 82)	Percent of variation	
Among populations (Fst)	45.90	
Among geographic groups (Fct)	39.15***	
Among populations within geographic groups (Fsc)	6.75***	
Within populations	54.10***	
Continental structures	Fst (%)	H (s.d.)
1. Sub-Saharan Africa (n = 13) <sup>1</sup>	16.95***	0.462 (0.20)
1a. Niger-Congo (n = 7)	5.47***	0.477 (0.16)
1b. Afro-Asiatic (n = 3)	1.70***	0.758 (0.02)
2. North Africa (n = 4)	2.62***	0.651 (0.05)
3. Europe (n = 12)	2.39***	0.503 (0.13)
4. Southwest Asia (n = 7)	13.81***	0.625 (0.08)
5. North East Asia (n = 7)	3.71***	0.675 (0.07)
6. South East Asia (n = 8)	5.96***	0.406 (0.21)
5 + 6. East Asia (n = 15)	26.90***	
7. Oceania (n = 7)	33.77***	0.510 (0.17)
8. Circum-Arctic area (n = 9)	8.10***	0.453 (0.24)
9. North and Central America (n = 7)	10.64***	0.358 (0.11)
10. South America (n = 8)	6.09***	0.357 (0.10)
9 + 10. America (n = 15)	6.63***	

\*\*\*  $P < 0.00001$ ; H: average within-population heterozygosity; s.d.: standard deviation.

<sup>1</sup> Malagasy excluded: Fst = 0.169.

ica are also very homogeneous for GM, even after collapsing the North and Central American group with the South American group (group 9 + 10, Fst = 0.066). Finally, Europe, with 12 populations considered, is particularly homogeneous (group 3, Fst = 0.024). This is observed despite the inclusion of Basques (Calderon et al., 1998; Esteban et al., 1998), Finns (Steinberg et al., 1974), and Hungarians (Schanfield et al., 1975), which are non-Indo-European populations, and of the geographically isolated Sardinians (Piazza et al., 1976) and Corsicans (Calafell et al., 1996).

Table 4 (“continental structures”) also shows the average within-population gene diversity (H) estimated in each geographic group. Sub-Saharan populations speaking Afro-Asiatic languages (i.e., East Africans) exhibit by far the most heterogeneous internal diversity ( $H = 0.758 \pm 0.05$ ), followed by Northeast Asians ( $H = 0.675 \pm 0.14$ ), North Africans ( $0.651 \pm 0.1$ ), and Southwest Asians ( $H = 0.625 \pm 0.16$ ). By contrast, Amerindians are very homogeneous ( $H = 0.358 \pm 0.2$  and  $H = 0.357 \pm 0.2$  for North-Central and South Amerindians, respectively), followed by Southeast Asians ( $H = 0.477 \pm 0.42$ ), circum-Arctic populations ( $H = 0.453 \pm 0.48$ ), Niger-Congo-speakers (or West and South Africans,  $H = 0.477 \pm 0.32$ ), and Europeans ( $H = 0.503 \pm 0.26$ ).

As our MDS analysis revealed four main axes along which populations of a particular geographic area were differentiated (Fig. 3A,B), we also tested whether those differentiations could be explained by an IBD model within each continental area (Table 5, “continental structures”). Only broad groups (sub-Saharan Africa, Europe, East

TABLE 5. Test for isolation by distance

Global structure	r <sup>1</sup>	P-value <sup>2</sup>
All populations (n = 82)	0.316	0.0001
Continental structures <sup>3</sup>	r <sup>1</sup>	P-value <sup>2</sup>
Sub-Saharan Africa (group 1, n = 13)	0.187	0.0372
Africa (groups 1 and 2, n = 17)	0.177	0.0162
Europe (group 3, n = 12)	0.306	0.0199
Europe and southwest Asia (groups 3 and 4, n = 19)	0.490	0.0001
East Asia (groups 5 and 6, n = 15)	0.355	0.0019
South East Asia and Oceania (groups 6 and 7, n = 15)	0.247	0.0158
America (groups 9 and 10, n = 15)	0.149	0.0779
Circum-Arctic area (group 8, n = 9)	0.446	0.0059

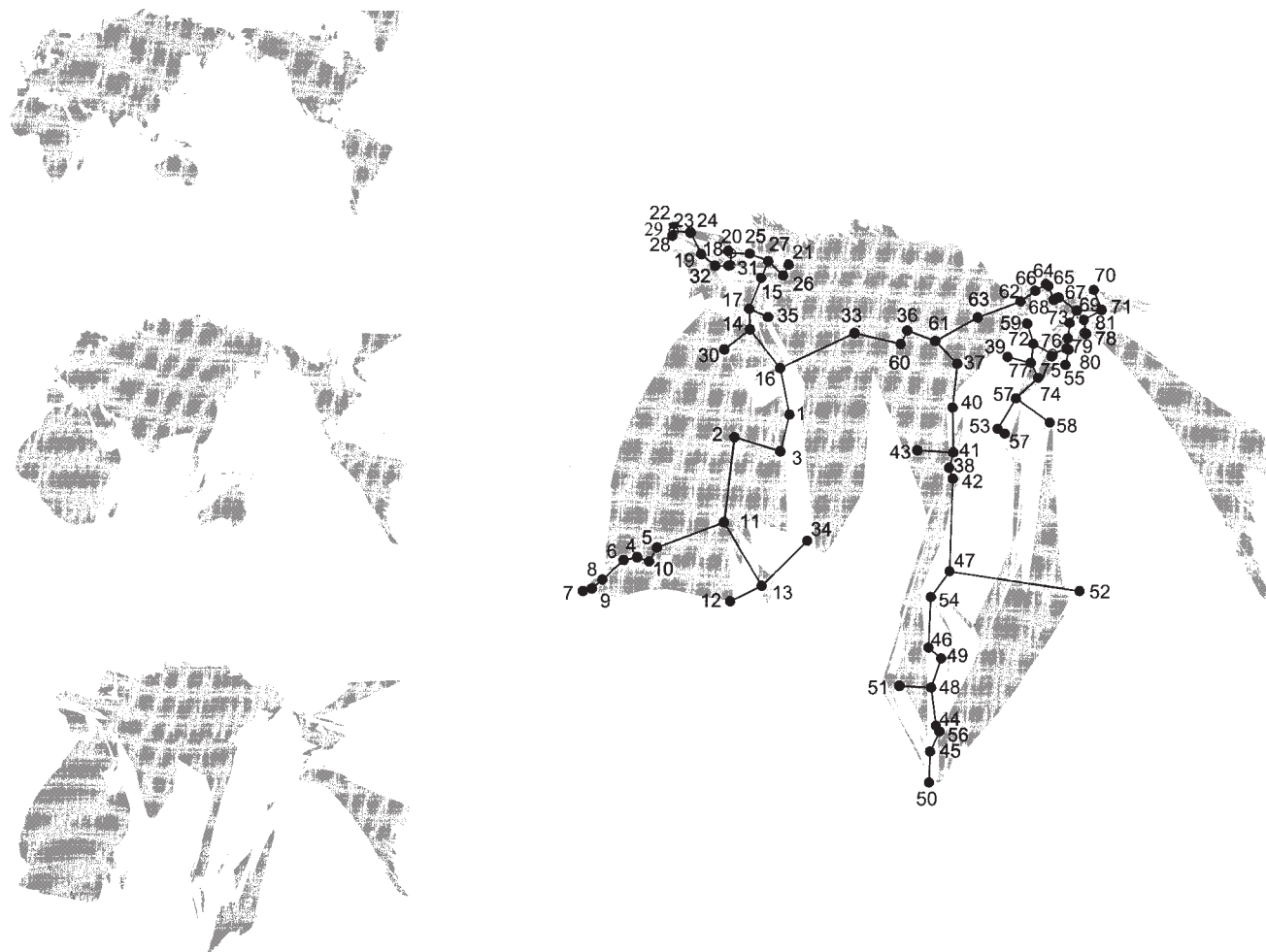
<sup>1</sup> Correlation coefficient between the logarithm of geographic distance and Fst/(1-Fst).

<sup>2</sup> Assessed by Mantel test.

<sup>3</sup> Group numbers correspond to those defined in Tables 1 and 4.

Asia, America, and the circum-Arctic region) were considered, in order to avoid groups containing low numbers of populations. In addition, we tested IBD after collapsing some groups showing genetic continuity in the MDS analysis (sub-Saharan Africa collapsed with North Africa, Europe collapsed with Southwest Asia, and Southeast Asia collapsed with Oceania).

Significant correlation coefficients were obtained for all groups (Table 5, “continental structures”) except America ( $r = 0.149$ ,  $P = 0.078$ ), suggesting that an IBD model, compatible with GM data at a global scale, is verified at most continental levels. However, the results are very heterogeneous among different groups. Highly significant correlations are found for Europe and Southwest Asia taken together (groups 3 and 4,  $r = 0.49$ ,  $P = 0.0001$ ), Northeast and Southeast Asia taken together (groups 5 and 6, “East Asia,”  $r = 0.355$ ,  $P = 0.002$ ), and the circum-Arctic area (group 8,  $r = 0.446$ ,  $P = 0.0059$ ). The high correlation found for groups 3 and 4 favors the hypothesis of a European differentiation by IBD from the Near East. The direction of this differentiation is suggested by the fact that Southwest Asians are more heterogeneous genetically than Europeans, and are closer to the center of the MDS where all groups are interconnected (Fig. 3). The high correlation found for groups 5 and 6 suggests that geographic differentiation is generally a good predictor of genetic relationships in East Asia, mostly along a north-south direction, and despite a significant contribution of linguistic differentiations (Poloni et al., 2004; Sanchez-Mazas et al., 2004). Finally, the result obtained for the circum-Arctic area (group 8) indicates genetic continuity around the North Pole, which may be related, in part, to a long-range migration of Eskimo-Aleut-speakers from Siberia to Greenland via the Bering Strait (Langaney, 1978).



**Fig. 4.** Distortion of an initial global geographic map (left, top) into a final genetic similarity map (GS map, right) by successive steps, according to method described in text. Two intermediate steps (left, center, and bottom) are shown for better comprehension of final geographic deformation. A minimal spanning network is superimposed on final GS map of 82 populations (see Table 1 for population numbers).

#### Genetic similarity map and minimal spanning network

As explained above, the genetic similarity map (GS map) is obtained by distorting the geographic map until the differences between genetic distances and geometric distances (those observed in the distorted map) are the smallest. In Figure 4, four maps are selected to show the successive distortions. The three small maps (left, top to bottom) and the final map (right) correspond, respectively, to steps 0 (initial), 5, 10, and 192 (final) of the distortion. We observe that during intermediate steps 5 and 10, the main distortions are: 1) a bulge in Africa, 2) a high reduction of size of America and of Europe (but at a lower level), and 3) an accentuated stretching of Southeast Asia. The larger map shows the final GS map of the 82 populations. The high proportion (87%) of explained distance variability (the relative variation of the sum of the squares of distance differences) indicates that a large concordance exists between

genetic and geographic distances, as suggested before. The final GS map is also incremented by a minimal spanning network (Fig. 4). This network mostly reveals three zones of high genetic homogeneity, i.e., Europe (left top), America (in addition to Eskimo populations from all regions, right top), and west and south sub-Saharan Africa (left bottom), each region being characterized by very short connections among populations. To a lesser extent, Southeast Asian populations exhibit a similar pattern (right bottom). Those regions are further linked to each other via long branches spanning throughout East (and partly North) Africa, India, and Northeast Asia (via some Siberian populations). Europeans coalesce to the network through populations from the Near East and North Africa, west and south sub-Saharan Africans through East and North Africans, and Amerindians through Northeast Asians. Oceanian populations are connected both to Southeast Asian and to Amerindian populations.

## DISCUSSION

Despite the very large amount of GM genetic data collected during several decades (Steinberg and Cook, 1981), and the numerous analyses of these data carried out at worldwide and continental levels (Blanc et al., 1990; Callegari-Jacques et al., 1993; Dugoujon et al., 1995; Excoffier et al., 1987, 1991; Matsumoto, 1988; Mourrieras et al., 1997; Sanchez-Mazas, 1990; Sanchez-Mazas and Langaney, 1988; Sanchez-Mazas and Pellegrini, 1990; Schanfield, 1992; Weber et al., 2000; Zhao and Lee, 1989), several questions related to this system remained unanswered, due in great part to the heterogeneity of the data record. In particular, the GM polymorphism is known to exhibit a very high level of heterogeneity among the main broad continental areas, but the genetic profiles of populations located at geographically intermediate regions are inadequately known. In this paper, we improve the global GM data record by 10 new population samples, 4 of them being located in intermediate areas between Africa, Southwest Asia, and Europe (Issa, Amhara-Tigré, Yemenite, and Kabyle). Three additional populations represent West (Bwa and Baoule) and Central (Aka) African populations, for which data were also scarce in former studies. The last three populations are located in regions where successive settlements of very distinct origins and time ranges are known to have occurred (Malagasy, and the northern and southern fringes of New Guinea). As GM diversity is highly heterogeneous among broad continental areas, it can be useful to highlight the extent of gene flow in such regions.

### Genetic diversity in central and peripheral geographic areas

Based on our important final data set, this study reveals contrasting levels of genetic diversity between some of the different geographic regions that we considered (Fig. 2 and Table 4). East African GM genetic profiles were only known for Sidamo thus far (Steinberg, 1973). By adding new population samples from East Africa (Issa from Djibuti, and Amhara-Tigré from Ethiopia), we confirm that a very high level of internal genetic diversity characterizes the populations of this region (Fig. 2 and Table 4,  $H = 0.758 \pm 0.04$ ). Issa exhibit the highest level of genetic variation ( $h = 0.788$ ) among the 82 worldwide populations included in this study, and Sidamo immediately follow ( $h = 0.753$ ). Several haplotypes are found with intermediate frequencies in these populations, which is not a usual pattern for the GM polymorphism. Moreover, the most frequent haplotypes observed in East Africans are those observed at very high frequencies in sub-Saharan Africa (GM 1, 17; 5\*), Europe (GM 3; 5\* and GM 3; 23; 5\*), and Northeast Asia and America (GM 1, 17; 21). Close genetic profiles in which almost all haplotypes present in the world are represented are also observed in North Africa and Southwest Asia, as em-

phasized by our new data for Kabyles and Yemenites. As the populations of these three regions exhibit the majority of GM haplotypes found all over the world, their genetic profile may be close to that of a highly polymorphic ancestral population from which present human populations may have derived. Unless this hypothetical ancestral population moved from its original homeland in the course of human evolution, these three regions are thus likely to represent putative geographic origins of human migrations, as suggested, at least for East Africa, from the study of both classical (Excoffier et al., 1987; Sanchez-Mazas and Langaney, 1988) and DNA (Quintana-Murci et al., 1999; Semino et al., 2002; Watson et al., 1997) polymorphisms.

Another region that exhibits high levels of within-population genetic diversity is Northeast Asia (Table 4,  $H = 0.625 \pm 0.16$ ). The gene diversity observed in some populations of this region (as well as in some populations included in the circum-Arctic group, such as Selkup) is close to, and sometimes even higher than that found in East Africa (e.g.,  $h = 0.750$  in Hui,  $h = 0.711$  in Mongolians), due to the presence of several frequent haplotypes (GM 1, 17; 21, GM 1, 2, 17; 21, GM 1, 17; 10, 11, 13, 15, 16, and GM 1, 3; 5\*; Fig. 2). Genetic studies support the hypothesis of demic diffusion (population migration) from the Middle East to Northeast Asia, possibly corresponding to the expansion of both Neolithic farmers and speakers of the Altaic linguistic family (Barbujani and Pilastro, 1993; Barbujani et al., 1994; Renfrew, 1991). Moreover, East Asia as a whole embraces several centers of origin of plant domestication (summarized in Cavalli-Sforza et al., 1994, p. 202–208) which, due to demographic increase, could represent putative demic diffusion centers leading to possible contacts with neighboring regions (Bellwood, 2001). The high level of genetic diversity currently observed for GM within Northeast Asian populations suggests that historical events leading to intensive genetic exchanges among neighboring populations could have been important in the past.

According to the MS network (Fig. 4), relatively long branches (i.e., large genetic distances) connect, successively, East and North Africa (populations 1–3 and 16), Southwest Asia, and Northeast Asia extended to some circum-Arctic populations (populations 33, 36–38, 40–43, and 60–63). As stated before, populations of these regions exhibit high levels of gene diversity. On the other hand, several genetically homogeneous clusters (west and south sub-Saharan Africans, populations 4–11; Europeans, populations 18–29; Southeast Asians, populations 44–51; and Amerindians, populations 68–82) emerge at side positions of this network. These groups are the most homogeneous at the level of both within- and between-population genetic diversity (Table 4, where “Niger-Congo” roughly correspond to west and south sub-Saharan Africans). This global structure, genetically heterogeneous in

central regions and genetically homogeneous at the periphery, is also visible in Figure 3A,B, where four branches emerge from a relatively heterogeneous central cluster. Actually, these four groups correspond to four geographically “peripheral” regions of the world.

### Role of human migration

Overall, these results may be seen as signatures of ancient population differentiations occurring along central areas of the Old World (the long branches of the network in Fig. 4), followed by more recent differentiation events occurring in peripheral regions (the homogeneous clusters at side positions of the network in Fig. 4). Such a “centrifugal” mode of dispersion is generally in agreement with models of modern human expansions proposed from the study of GM and other polymorphisms. Indeed, the genetic homogeneity observed in Europe and its continuity with Southwest Asia, characterized by a highly significant isolation-by-distance pattern of migration (Table 5), may be a result of the expansion of Neolithic farmers from Southwest Asia to Northwest Europe in the Holocene period. This was suggested, initially, by the study of classical polymorphisms (Ammerman and Cavalli-Sforza, 1984), and is still sustained by of DNA polymorphisms, more recent analyses (Barbujani and Bertorelle, 2001; Barbujani et al., 1998; Chikhi et al., 1998, 2002), despite heated controversy (Cavalli-Sforza and Minch, 1997; Richards et al., 1996; Semino et al., 2000; Simoni et al., 2000). More generally, gene frequency variation of classical polymorphisms sustains the demic diffusion model for 3 of 4 linguistic families of the “Nostratic” phylum (Indo-European, Elamo-Dravidian, and Altaic, but not Afro-Asiatic), involving population movements from Southwest Asia towards Northwest Europe, India, and Northeast Asia, respectively (Barbujani and Pilastro, 1993; Barbujani et al., 1994). For Africa, several alternative hypotheses have been put forward to explain the present genetic diversity, most often related to the first emergence of modern humans (Cruciani et al., 2002; Hammer et al., 1998, 2001; Quintana-Murci et al., 1999; Satta and Takahata, 2002; Semino et al., 2002; Templeton, 2002). However, a previous reconstruction of African history based on “classical” polymorphisms (such as GM) showed that African genetic differentiations are compatible with linguistic hypotheses suggesting the greatest antiquity for the Afro-Asiatic phylum (although Nilo-Saharan is proposed, by some linguists, as the most ancient before Afro-Asiatic; Blench, 1993). Archaeological and linguistic evidence also indicates that the first Niger-Congo differentiations would have occurred around 8,000 years BP, with a main Bantu expansion starting between 5,000–3,000 years BP (Excoffier et al., 1987). Thus, Africa has been the scene of recent population movements, probably from eastern to western and southern regions, and this is compatible with our centrifugal migration model. Finally,

the genetic homogeneity seen in Amerindians could also be attributed to recent migrations, although there is still much debate about the time of the first peopling of the Americas, the number of migration waves crossing Beringia, and the place of the Amerindian homeland in East Asia (e.g., Bonatto and Salzano, 1997; Karafet et al., 1997, 1999; Lell et al., 2002; Merriwether et al., 1995; Schanfield, 1992; Silva et al., 2002; Tarazona-Santos and Santos, 2002; Williams et al., 1985). We must note here that the GM haplotypic repertoire is considerably reduced in Amerindians, compared to other populations, probably as a result of one or several bottlenecks (Fig. 2). Consequently, the ability of the GM system to describe population differentiations is very low in America. This is also partly the case for Oceania, where several populations (e.g., Papuans Enga and Australian Aborigines) are close to Northeast Asians and Amerindians both in the MDS (Fig. 3) and the MS network (Fig. 4), a result that could be attributed to genetic convergence. But contrary to Amerindians, Oceanian populations are highly differentiated among them, e.g., the Papuans from the northern and southern fringes of New Guinea tested in our study. The most frequent haplotypes observed in these populations are different (GM 1, 17; 21 for the former, and GM 1, 17; 5\* for the latter), and may thus indicate independent differentiations, at different times, from highly distinct genetic profiles. In addition, northern Papuans exhibit a high frequency of haplotype GM 1, 3; 5\*, which is very frequent in Southeast Asia and the Pacific. It may reveal gene flow from Austronesians, in agreement with other genetic evidence (Kayser et al., 2000).

According to the hypotheses presented above, most genetic differentiations suggested by the present study would not date back to remote periods of modern humans history (i.e., the first dispersal of *Homo sapiens sapiens*), but to more recent times corresponding to the expansion of the main linguistic families, and/or the dispersal of Neolithic farmers, around, or since 10,000 years ago. However, the fact that populations of the different continental regions are so strongly differentiated among them (see below) for the GM polymorphism suggests that genetic traces of ancient population divergence corresponding to old dispersals of hunter-gatherers groups would not have disappeared completely.

### Contrasting levels of population structure among genetic systems

One relevant result of the present study is indeed the very high level of genetic variation observed among populations (Table 4,  $F_{st} = 0.459^{***}$ ), with a main proportion of this variation due to differences among continental groups (Table 4,  $F_{ct} = 0.3915^{***}$ ). To our knowledge, such values are not reached by any other polymorphism. Lewontin (1972) reported a variability of 15.6% among populations, of whom 6.3% represented the variance among continental groups. Further studies based on

protein and DNA polymorphisms gave values of about 15% among populations, and 10% among continental groups, respectively (Barbujani et al., 1997). Also,  $F_{st}$  values around 20% were found for some classical polymorphisms, such as RH (Flint et al., 1999) and the Y-chromosome (Jorde et al., 2000; Poloni et al., 1997). Appreciably higher levels of variation among populations were only observed for mtDNA (~28% for HVS1 and ~31% for HVS2; Jorde et al., 2000) and the FY blood group polymorphism (~40%; Flint et al., 1999). Interestingly, the mtDNA polymorphism exhibits a departure from neutrality in some population groups such as Europeans and Orientals (Excoffier, 1990; Mishmar et al., 2003), and FY is believed to be subject to directional selection, as FY\*O homozygotes appear to be protected against malaria in sub-Saharan Africa (Hamblin and Di Rienzo, 2000). Here, the very high  $F_{st}$  and  $F_{ct}$  values observed for the GM polymorphism, besides the simple explanation of ancient differentiations of small populations, could be explained by at least two additional hypotheses: natural selection, and incomplete assessment of GM genetic variability.

**Influence of natural selection.** Natural selection has sometimes been invoked to explain GM variation. Piazza et al. (1976) underlined contrasted GM frequencies between lowland and highland Sardinia (in particular, for allotype G2M(23)) and related them to a possible protective effect against malaria, which was endemic in Sardinia in the past. A similar hypothesis was recently proposed to explain GM diversity in Taiwan (Schanfield et al., 2002). However, in general, GM and disease association studies failed to demonstrate strong directional selective effects, or gave discordant results (Pandey et al., 2001; Propert, 1995). Other studies focused on the *maintenance* of GM diversity by natural selection (e.g., balancing selection), but significant results were found for the KM polymorphism of the light  $\kappa$  chains rather than for GM (Black and Pandey, 1997; Propert and Balkau, 1986). To check our data for a possible departure of GM haplotype frequency distributions from neutrality, we applied the Ewens-Watterson test (Ewens, 1972, 1979; Watterson, 1978, 1986) to the 82 populations of the present study. Of 79 tests done,<sup>3</sup> only 6 gave a significant result ( $P < 0.05$  for populations 37, 38, 40, 41, 43, and 60): in all cases, towards an excess of heterozygotes. Moreover, no departure from neutrality was observed after Bonferroni's correction for multiple tests (results not shown). Therefore, we cannot globally reject the hypothesis of selective neutrality for the GM system on statistical grounds. If some selection nevertheless played a role, the main effect would have been to enhance the magni-

tude of intercontinental differentiation without altering the global pattern of genetic relationships.

**Incomplete assessment of GM genetic variability.** Another hypothesis to explain the very high level of GM diversity observed among populations is to invoke an incomplete assessment of GM molecular variability. Indeed, GM allotypes detected by immunological methods represent only a reduced part of the true GM diversity. DNA RFLP analyses and DNA sequencing of the IGCH loci involved in GM variation indicate that many different DNA haplotypes are hindered within some "classical" GM haplotypes (Dard et al., 1996, 2001). A high-resolution molecular typing of these loci would thus probably lead to the observation of more heterogeneous GM genetic profiles within populations, and perhaps of smoother frequency changes among populations, as it occurs for HLA when considering either serologically-typed specificities or DNA-typed alleles (personal results). Also, even by using the conventional definition of GM variability, heterozygotes for common and uncommon GM haplotypes (e.g., GM 1, 17; 5\*/GM 1, 17; 5, 14) may be confounded with homozygotes for the predominant haplotype (e.g., GM 1, 17; 5\*/GM 1, 17; 5\*), and the estimation of GM frequencies may be biased towards an excess of the most frequent haplotype in each continental region. This may mostly increase the level of GM variability observed among geographic regions ( $F_{ct}$ ) at the expense of that observed within populations, but not at the expense of that observed among populations within geographic regions ( $F_{sc}$ ), as the latter share the same common haplotypes. Interestingly, this is what we observe in our data (Table 4), as  $F_{sc}$  is about the same compared to the studies of Lewontin (1972) and Barbujani et al. (1997), but  $F_{ct}$  is increased and the proportion of variation within populations is decreased.

## CONCLUSIONS

In this study, we showed that the diversity of the GM system can be satisfactorily explained by geographic differentiations on the world scale. A particularity of this system is also that it exhibits the highest level of genetic variability ever observed among populations of different worldwide geographic regions. Although either natural selection or an incomplete assessment of GM genetic variability due to serological typing (or both) may have enhanced the magnitude of estimated intercontinental genetic variation, the global GM diversity can basically be explained by an interesting pattern of modern human migrations that we call "the centrifugal model." This model states that ancient population differentiations occurred between East Africa and East Asia and were followed by recent migrations in peripheral geographic areas. It is sustained by previous works relating the present genetic diversity to linguistic and archaeological hypotheses of human migrations occurring during the last 10,000 years

<sup>3</sup>The tests could not be performed on three populations (18, 48, and 78) because sample sizes were too high.

towards Africa (e.g., Bantu), Europe (Indo-Europeans), and East Asia/Oceania (e.g., Altaic and Austronesians). However, to bring more evidence to this hypothesis, DNA studies should be done to perform a more thorough evaluation of GM molecular variability within and between populations. Moreover, as the present results are based on the study of a single polymorphism which does not represent the human genome as a whole, they should be linked to results obtained for other polymorphisms, using equivalent data sets and similar approaches.

### ACKNOWLEDGMENTS

We are particularly grateful to E. Guitard and M.T. S en egas for their help in the determination of GM allotypes. The authors also thank M. Blanc, A. Chaventr e, G. Bellis, P. Lef evre Witier, M. Benabadi, G.F. de Stefano, G. Larrouy, R. Cabannes, G. Jaeger, M.J. Palisson, A. Sevin, F.X. Soloarivony, P. Richard, and K. Bathia, who provided us with serum samples for GM typing or generous help and technical assistance. We sincerely acknowledge three anonymous reviewers who gave us very detailed comments on the manuscript.

### LITERATURE CITED

- Ammerman AJ, Cavalli-Sforza LL. 1984. The Neolithic transition and the genetics of populations in Europe. Princeton, NJ: Princeton University Press.
- Barbujani G, Bertorelle G. 2001. Genetics and the population history of Europe. *Proc Natl Acad Sci USA* 98:22–25.
- Barbujani G, Pilastro A. 1993. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostrotrac macrofamily. *Proc Natl Acad Sci USA* 87:1816–1819.
- Barbujani G, Oden NL, Sokal RR. 1989. Detecting regions of abrupt change in maps of biological variables. *Syst Zool* 38:376–389.
- Barbujani G, Pilastro A, De Domenico S, Renfrew C. 1994. Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonization. *Am J Phys Anthropol* 95:137–154.
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997. An appointment of human DNA diversity. *Proc Natl Acad Sci USA* 94:4516–4519.
- Barbujani G, Bertorelle G, Chikhi L. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* 62:488–492.
- Bellwood P. 2001. Early agriculturalist population diasporas? Farming, languages, and genes. *Annu Rev Anthropol* 30:181–207.
- Black FL, Pandey JP. 1997. Evidence for balancing of KM, but not GM, alleles by heterotic advantage in South Amerinds. *Hum Genet* 100:240–244.
- Blanc M, Sanchez-Mazas A, Hubert van Blyenburgh N, Sevin A, Pison G, Langaney A. 1990. Inter-ethnic genetic differentiation: Gm polymorphism in eastern Senegal. *Am J Hum Genet* 46:383–392.
- Blench R. 1993. Recent developments in African language classification and their implications for prehistory. In: Shaw T, Sinclair P, Andah B, Okpoko A, editors. *The archaeology of Africa. Food, metals and towns*. London: Routledge. p 126–138.
- Bonatto SL, Salzano FM. 1997. Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am J Hum Genet* 61:1413–1423.
- Bosch E, Calafell F, Perez-Lezaun A, Comas D, Mateu E, Bertranpetit J. 1997. Population history of North Africa: evidence from classical genetic markers. *Hum Biol* 69:295–311.
- Calafell F, Bertranpetit J, Rendine S, Capello N, Mercier P, Amoros JP, Piazza A. 1996. Population history of Corsica: a linguistic and genetic analysis. *Ann Hum Biol* 23:237–251.
- Calderon R, Vidales C, Pena JA, Perez-Miranda A, Dugoujon JM. 1998. Immunoglobulin allotypes (GM and KM) in Basques from Spain: approach to the origin of the Basque population. *Hum Biol* 70:667–698.
- Callegari-Jacques SM, Salzano FM, Constans J, Mauri eres P. 1993. Gm haplotype distribution in Amerindians: relationship with geography and language. *Am J Phys Anthropol* 90:427–444.
- Cavalli-Sforza LL, Minch E. 1997. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 61:247–254.
- Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The history and geography of human genes*. Princeton, NJ: Princeton University Press.
- Chen J, Sokal RR, Ruhlen M. 1995. Worldwide analysis of genetic and linguistic relationships of human populations. *Hum Biol* 67:595–612.
- Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. 1998. Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool. *Proc Natl Acad Sci USA* 95:9053–9058.
- Chikhi L, Nichols RA, Barbujani G, Beaumont MA. 2002. Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci USA* 99:11008–11013.
- Cockerham CC. 1969. Variance of gene frequencies. *Evolution* 23:72–83.
- Cockerham CC. 1973. Analysis of gene frequencies. *Genetics* 74:679–700.
- Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70:1197–1214.
- Dard P, Sanchez-Mazas A, Dugoujon J-M, De Lange G, Langaney A, Lefranc M-P, Lefranc G. 1996. DNA analysis of the immunoglobulin IGHG loci in a Mandenka population from eastern Senegal: correlation with Gm haplotypes and hypotheses for the evolution of the Ig CH region. *Hum Genet* 98:36–47.
- Dard P, Lefranc M-P, Osipova L, Sanchez-Mazas A. 2001. DNA sequence variability of IGHG3 alleles associated to the main G3m haplotypes in human populations. *Eur J Hum Genet* 9:765–772.
- Dugoujon JM, Mourrieras B, Senegas MT, Guitard E, Sevin A, Bois E, Hazout S. 1995. Human genetic diversity (immunoglobulin GM allotypes), linguistic data, and migrations of Amerindian tribes. *Hum Biol* 67:231–249.
- Esteban E, Dugoujon JM, Guitard E, S en egas MT, Manzano C, De la R ua C, Valveny N, Moral P. 1998. Genetic diversity in northern Spain (Basque country and Cantabria): GM and KM variations related to demographic histories. *Eur J Hum Genet* 6:315–324.
- Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87–112.
- Ewens WJ. 1979. Testing the generalized neutrality hypothesis. *Theor Popul Biol* 15:205–216.
- Excoffier L. 1990. Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *J Mol Evol* 30:125–139.
- Excoffier L. 2001. Analysis of population subdivision. In: Balding DJ, Bishop M, Cannings C, editors. *Handbook of statistical genetics*. Chichester, UK: John Wiley & Sons. p 271–307.
- Excoffier L, Pellegrini B, Sanchez-Mazas A, Simon C, Langaney A. 1987. Genetics and history of sub-Saharan Africa. *Yrbk Phys Anthropol* 30:151–194.
- 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.
- Excoffier L, Smouse P, Quattro J-M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplo-

- types: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
- Flanagan JG, Rabbitts TH. 1982. Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. *Nature* 300:709–713.
- Flint J, Bond J, Rees DC, Boyce AJ, Roberts-Thomson JM, Excoffier L, Clegg JB, Beaumont MA, Nichols RA, Harding RM. 1999. Minisatellite mutational processes reduce  $F_{st}$  estimates. *Hum Genet* 105:567–576.
- Grubb R. 1956. Agglutination of erythrocytes coated with “incomplete” anti-Rh by certain rheumatoid arthritic sera and some other sera. The existence of human serum groups. *Acta Pathol Microbiol Scand* 39:195–197.
- Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66:1669–1679.
- Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL. 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* 15:427–441.
- Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18:1189–203.
- Hazout S, Guasp G, Loirat F, Maurieres P, Larrouy G, Dugoujon JM. 1993. A new approach for interpreting the genetic diversity in space: “mobile site method.” Application to Gm haplotype distribution of twenty-seven Amerindian tribes from North and Central America. *Ann Hum Genet* 57:221–237.
- Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708–713.
- Jain AK, Dubes RC. 1988. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall.
- Jolliffe IT. 1986. Principal component analysis. New York: Springer Verlag.
- Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data. *Am J Hum Genet* 66:979–988.
- Karafet T, Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, Romero F, Long JC, Harihara S, Jin F, Dashnyam B, Gerelsaikhan T, Omoto K, Hammer MF. 1997. Y chromosome markers and trans-Bering Strait dispersals. *Am J Phys Anthropol* 102:301–314.
- Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF. 1999. Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *Am J Hum Genet* 64:817–831.
- Kayser M, Brauer S, Weiss G, Underhill PA, Roewer L, Schiefenhovel W, Stoneking M. 2000. Melanesian origin of Polynesian Y chromosomes. *Curr Biol* 10:1237–1246.
- Kruskal JB. 1964. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:28–42.
- Kunkel HG, Smith WK, Joslin FG, Natvig JB, Litwin SD. 1969. Immunogenetic study of Am(1), the first allotype of human IgA. *Nature* 223:1247–1248.
- Langaney A. 1978. De la Sibérie au Groenland de l’est: immunologie et histoire d’une dispersion. In: Actes du XLIIème Congrès International des Américanistes, Congrès du Centenaire. Paris, 2–9 Sept 1976, vol 5. p 39–46.
- Lefranc MP, Lefranc G. 1990. Molecular genetics of immunoglobulin allotype expression. In: Shakib F, editor. The human IgG subclasses: molecular analysis of structure, function and regulation. Oxford: Pergamon Press. p 43–78.
- Lell JT, Sukernik RI, Starikovskaya YB, Su B, Jin L, Schurr TG, Underhill PA, Wallace DC. 2002. The dual origin and Siberian affinities of Native American Y chromosomes. *Am J Hum Genet* 70:192–206.
- Lewontin R. 1972. The appointment of human diversity. *Evol Biol* 6:381–398.
- Long JC. 1986. The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright’s F-statistics. *Genetics* 112:629–647.
- Mantel G. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res* 27:209–220.
- Matsumoto H. 1988. Characteristic of Mongoloid and neighboring populations based on the genetic markers of immunoglobulins. *Hum Genet* 80:207–218.
- Merriwether DA, Rothhammer F, Ferrell RE. 1995. Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. *Am J Phys Anthropol* 98:411–430.
- Milstein C, Deverson EV, Rabbitts TH. 1984. The sequence of the human immunoglobulin  $\mu$ - $\delta$  intron reveals possible vestigial switch segments. *Nucleic Acids Res* 12:6523–6535.
- Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC. 2003. Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 100:171–176.
- Monmonier MS. 1973. Maximum-difference barriers: an alternative numerical regionalization method. *Geogr Anal* 3:245–261.
- Mourrieras B, Dugoujon JM, Buffat L, Hazout S. 1997. Assessment of genetic diversity in space by superimposition of a distorted geographic map with a spatial population clustering. Application to GM haplotypes of Native Amerindian tribes. *Ann Hum Genet* 61:37–47.
- Nei M. 1972. Genetic distance between populations. *Am Nat* 106:283–292.
- Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
- Pandey JP, Cooper GS, Treadwell EL, Gilkeson GS, St Clair EW, Dooley MA. 2001. Immunoglobulin GM and KM allotypes in systemic lupus erythematosus. *Exp Clin Immunogenet* 18:117–122.
- Piazza A, van Loghem E, de Lange G, Curtioni ES, Ulizzi L, Terrenato L. 1976. Immunoglobulin allotypes in Sardinia. *Am J Hum Genet* 28:77–86.
- Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L. 1997. Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 61:1015–1035.
- Poloni ES, Sanchez-Mazas A, Jacques G, Sagart L. 2004. Comparing linguistic and genetic affinities among East Asian populations: A study of the RH and GM polymorphisms. In: Sagart L, Blench R, Sanchez-Mazas A, editors. The peopling of East Asia: putting together archaeology, linguistics and genetics. London: Routledge Curzon. In press.
- Prevosti A, Ocaña J, Alonso G. 1975. Distances between populations of *Drosophila subobscura* based on chromosome arrangement frequencies. *Theor Appl Genet* 45:231–241.
- Proper DN. 1995. Immunoglobulin allotypes and RFLPs in disease association. *Exp Clin Immunogenet* 12:198–205.
- Proper DN, Balkau BJ. 1986. Selective forces and the maintenance of immunoglobulin polymorphisms. *Hum Biol* 58:79–84.
- Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet* 23:437–441.
- Ray N. 2002. GEODIST. Laboratory of Genetics and Biometry, University of Geneva.
- Renfrew C. 1991. Before Babel: speculations on the origins of linguistic diversity. *Cambridge Archaeol J* 1:3–23.
- Renquin J, Sanchez-Mazas A, Hallé L, Rivalland S, Jaeger G, Mbayo K, Bianchi F, Kaplan C. 2001. HLA class II polymorphism in Aka Pygmies and Bantu Congolese and a reassessment of HLA-DRB1 African diversity. *Tissue Antigens* 58:211–222.
- Reynolds J, Weir BS, Cockerham CC. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105:767–779.
- Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes

- B. 1996. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59:185–203.
- Rohlf FJ. 2000. NTSYS-PC, numerical taxonomy and multivariate analysis system. Setauket: Exeter Software, Applied Biostatistics, Inc.
- Ropartz C, Lenoir J, Rivat L. 1961. A new inheritable property of human sera: the Inv factor. *Nature* 189:586–587.
- Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* 145:1219–1228.
- Sanchez-Mazas A. 1990. Polymorphisme des systèmes immunologiques Rhésus, GM et HLA et histoire du peuplement humain. Ph.D. thesis, University of Geneva.
- Sanchez-Mazas A, Langaney A. 1988. Common genetic pools between human populations. *Hum Genet* 78:161–166.
- Sanchez-Mazas A, Pellegrini B. 1990. Polymorphismes Rhésus, Gm et HLA et histoire de l'Homme moderne. *Bull Mem Soc Anthropol Paris* 2:57–76.
- Sanchez-Mazas A, Butler-Brunner E, Butler R, Calderon R, Chaventre A, Dugoujon JM, Hammond M, Lefranc G, Matsumoto H, Osipova L, Politis C, Pullmann R, Langaney A. 2001. A worldwide analysis of AG molecular diversity inferred from serology. *Hum Biol* 73:637–659.
- Sanchez-Mazas A, Poloni ES, Jacques G, Sagart L. 2004. HLA genetic diversity and linguistic variation in East Asia. In: Sagart L, Blench R, Sanchez-Mazas A, editors. *The peopling of East Asia: putting together archaeology, linguistics and genetics*. London: Routledge Curzon. In press.
- Satta Y, Takahata N. 2002. Out of Africa with regional interbreeding? Modern human origins. *Bioessays* 24:871–875.
- Schanfield MS. 1992. Immunoglobulin allotypes (GM and KM) indicate multiple founding populations of Native Americans: evidence of at least four migrations to the New World. *Hum Biol* 64:381–402.
- Schanfield MS, Gershowitz H. 1973. Nonrandom distribution of Gm haplotypes in East Asia. *Am J Hum Genet* 25:567–574.
- Schanfield MS, Gergely J, Fudenberg HH. 1975. Immunoglobulin allotypes of European populations. I. Gm and Km(Inv) allotypic markers in Hungarians. *Hum Hered* 25:370–377.
- Schanfield MS, Ohkura K, Lin M, Shyu R, Gershowitz H. 2002. Immunoglobulin allotypes among Taiwan aborigines: evidence of malarial selection could affect studies of population affinity. *Hum Biol* 74:363–379.
- Schneider S, Roessler D, Excoffier L. 2000. Arlequin: a software for population genetics data analysis. Geneva: University of Geneva.
- Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* 290:1155–1159.
- Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* 70:265–268.
- Silva WA Jr, Bonatto SL, Holanda AJ, Ribeiro-Dos-Santos AK, Paixao BM, Goldman GH, Abe-Sandes K, Rodriguez-Delfin L, Barbosa M, Paco-Larson ML, Petzl-Erler ML, Valente V, Santos SE, Zago MA. 2002. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *Am J Hum Genet* 71:187–192.
- 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.
- Sneath PHA, Sokal RR. 1973. *Numerical taxonomy*. San Francisco: W.H. Freeman.
- Sokal RR, Oden NL. 1978. Spatial autocorrelation in biology. *Biol J Linn Soc* 10:199–249.
- Sokal RR, Rohlf FJ. 1994. *Biometry*. New York: W.H. Freeman and Co.
- Soodyall H, Jenkins T, Hewitt R, Krause A, Stoneking M. 1996. The peopling of Madagascar. In: Boyce AJ, Mascie-Taylor CGN, editors. *Molecular biology and human diversity*. Cambridge: Cambridge University Press. p 156–170.
- Steinberg AG. 1973. Gm and Inv allotypes of some Sidamo Ethiopians. *Am J Phys Anthropol* 39:403–408.
- Steinberg AG, Cook CE. 1981. The distribution of the human immunoglobulin allotypes. Oxford: Oxford University Press.
- Steinberg AG, Tilikainen A, Eskola MR, Eriksson AW. 1974. Gammaglobulin allotypes in Finnish Lapps, Finns, Aland islanders, Maris (Cheremis), and Greenland Eskimos. *Am J Hum Genet* 26:223–243.
- Stoneking M, Jorde LB, Bhatia K, Wilson AC. 1990. Geographic variation in human mitochondrial DNA from Papua New Guinea. *Genetics* 124:717–733.
- Tarazona-Santos E, Santos FR. 2002. The peopling of the Americas: a second major migration? *Am J Hum Genet* 70:1377–1381.
- Templeton A. 2002. Out of Africa again and again. *Nature* 416:45–51.
- Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ. 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet* 26:358–361.
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65:43–62.
- van Loghem E, Aalberse RC, Matsumoto H. 1984. A genetic marker of human IgE heavy chains, Em(1). *Vox Sang* 46:195–206.
- Vyas GN, Fudenberg HH. 1969. Am(1), the first genetic marker of human immunoglobulin A. *Proc Natl Acad Sci USA* 64:1211–1216.
- Watson E, Forster P, Richards M, Bandelt HJ. 1997. Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* 61:691–704.
- Watterson GA. 1978. The homozygosity test of neutrality. *Genetics* 88:405–417.
- Watterson GA. 1986. The homozygosity test after a change in population size. *Genetics* 112:899–907.
- Weber W, Nash DJ, Motulsky AG, Henneberg M, Crawford MH, Martin SK, Goldsmid JM, Spedini G, Glidewell S, Schanfield MS. 2000. Phylogenetic relationships of human populations in sub-Saharan Africa. *Hum Biol* 72:753–772.
- Weir BS. 1996. *Genetic data analysis II*. Sunderland, MA: Sinauer Associates.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.
- Wells RS, Yuldashева N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, Balakrishnan K, Read M, Pearson NM, Zerjal T, Webster MT, Zholoshvili I, Jamarjashvili E, Gambarov S, Nikbin B, Dostiev A, Aknazarov O, Zalloua P, Tsoy I, Kitaev M, Mirrakhimov M, Chariev A, Bodmer WF. 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA* 98:10244–10249.
- Williams RC, Steinberg AG, Gershowitz H, Bennett PH, Knowler WC, Pettitt DJ, Butler W, Baird R, Dowla RL, Burch TA, Morse HG, Smith CG. 1985. Gm allotypes in Native Americans: evidence for three distinct migrations across the Bering Land Bridge. *Am J Phys Anthropol* 66:1–19.
- Wright S. 1965. The interpretation of population structure by F-statistics with special regards to systems of mating. *Evolution* 19:395–420.
- Wright S. 1978. *Evolution and the genetics of populations, volume 4. Variability within and among natural populations*. Chicago: University of Chicago Press.
- Yasuda N. 1968. Estimation of the inbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. *Biometrics* 24:915–935.
- Zhao T, Lee TD. 1989. Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. *Hum Genet* 83:101–110.