Am. J. Hum. Genet. 61:1015–1035, 1997

Human Genetic Affinities for Y-Chromosome P49a,f/*Taq*I Haplotypes Show Strong Correspondence with Linguistics

E. S. Poloni,[1,2] O. Semino,[3] G. Passarino,[3,4] A. S. Santachiara-Benerecetti,[3] I. Dupanloup,[1,2] A. Langaney,[1,2] and L. Excoffier[1,2]

[1]Département d'Anthropologie et Ecologie, Université de Genève, Carouge, Switzerland; [2]Laboratoire d'Anthropologie Biologique, Musée de l'Homme, Paris; [3]Università di Pavia, Dipartimento di Genetica e Microbiologia 'A. Buzzati Traverso,' Pavia, Italy; and [4]Università della Calabria, Dipartimento di Biologia Cellulare, Rende, Italy

Summary

Numerous population samples from around the world have been tested for Y chromosome–specific p49a,f/*Taq*I restriction polymorphisms. Here we review the literature as well as unpublished data on Y-chromosome p49a,f/*Taq*I haplotypes and provide a new nomenclature unifying the notations used by different laboratories. We use this large data set to study worldwide genetic variability of human populations for this paternally transmitted chromosome segment. We observe, for the Y chromosome, an important level of population genetic structure among human populations (F_{ST} = .230, P < .001), mainly due to genetic differences among distinct linguistic groups of populations (F_{CT} = .246, P < .001). A multivariate analysis based on genetic distances between populations shows that human population structure inferred from the Y chromosome corresponds broadly to language families (r = .567, P < .001), in agreement with autosomal and mitochondrial data. Times of divergence of linguistic families, estimated from their internal level of genetic differentiation, are fairly concordant with current archaeological and linguistic hypotheses. Variability of the p49a,f/*Taq*I polymorphic marker is also significantly correlated with the geographic location of the populations (r = .613, P < .001), reflecting the fact that distinct linguistic groups generally also occupy distinct geographic areas. Comparison of Y-chromosome and mtDNA RFLPs in a restricted set of populations shows a globally high level of congruence, but it also allows identification of unequal maternal and paternal contributions to the gene pool of several populations.

Received March 4, 1997; accepted September 10, 1997; electronically published October 29, 1997.

Address for correspondence and reprints: Dr. Estella Poloni, Laboratoire de Génétique et Biométrie, Département d'Anthropologie et Ecologie, Université de Genève, Case Postale 511, 1211 Genève 24, Switzerland. E-mail: estella.poloni@anthro.unige.ch

0002-9297/97/6105-0006$02.00

Introduction

The Y chromosome is currently subject to numerous studies aiming at assessment of the extent of human genetic variability that can be attributed to men. This interest is motivated by the fact that, for most of its length, the Y chromosome can be seen as a single uniparentally transmitted linkage group, allowing deduction of the paternal counterpart to the history of maternal lineages, which is revealed by mtDNA studies (Maynard Smith 1990; Pääbo 1995).

It is well known that, as revealed by RFLPs and DNA sequences, the Y chromosome does not exhibit as much polymorphism as the autosomes or the X chromosome (Jakubiczka et al. 1989; Malaspina et al. 1990; Spurdle and Jenkins 1992*a,* 1992*c;* Dorit et al. 1995; Hammer 1995). Whether it is necessary to invoke episodes of selective sweeps to account for this relative lack of variability is a matter of discussion (Dorit et al. 1995; Hammer 1995; Whitfield et al. 1995*b;* Goldstein et al. 1996; Underhill et al. 1996; Tavaré et al. 1997). This reduced level of polymorphism has also triggered the search for population specific Y-chromosome haplotypes that could act as markers of population founding events. These studies include simple repeat sequences (Roewer et al. 1992; Santos et al. 1993; Gomolka et al. 1994; Muller et al. 1994; Pena et al. 1995), an *Alu* insertion (Persichetti et al. 1992; Hammer 1994; Spurdle et al. 1994*a*), the sequencing of large portions of the Y chromosome, Y chromosome–specific sequence-tagged sites (Ellis et al. 1990; Seielstad et al. 1994; Dorit et al. 1995; Hammer 1995; Whitfield et al. 1995*a,* 1995*b;* Underhill et al. 1996), and combinations of several polymorphisms to construct compound haplotypes (Oakey and Tyler-Smith 1990; Persichetti et al. 1992; Jobling 1994; Mathias et al. 1994; Ciminelli et al. 1995; Hammer and Horai 1995; Jobling and Tyler-Smith 1995; Jobling et al. 1996; Ruiz Linares et al. 1996; Spurdle and Jenkins 1996; G. Passarino, unpublished data). The variability of these markers at the worldwide scale has recently started to be documented (e.g., see Deka et al. 1996; Santos et al. 1996; Hammer et al. 1997; Karafet et al. 1997; Zerjal et al. 1997), promising new insights into the history of our species.

The human Y chromosome–specific p49a,f/*Taq*I polymorphism, described >10 years ago (Ngo et al. 1986), does not fall into the low-variability category of conventional Y-specific RFLPs, since >100 distinct haplotypes have already been identified (appendix A). The p49a and p49f probes, 6 kb distant from each other, are subclones of cosmid 49 (Bishop et al. 1983) that hybridize to the DYS1 locus on the Y-chromosome long arm (Ngo et al. 1986). The DYS1 locus has recently been shown to correspond to the DAZ gene cluster (Saxena et al. 1996), located on the human Y-chromosome AZF (*A Z*oospermia *F*actor) region. The DAZ gene (*D*eleted in *A Z*oospermia), encoding an RNA-binding protein that is most probably involved in spermatogenesis, would have been transposed to the Y chromosome from a homologous gene on chromosome 3 (DAZH) prior to the divergence of orangutans from the human lineage, giving rise to the Y chromosome–specific DAZ gene (Saxena et al. 1996). Part of the genomic sequence of DAZ is organized into nine tandem repeats of highly homologous 2.4-kb segments, and several copies of the DAZ gene could be present in the AZF region of the Y chromosome (Saxena et al. 1996). The sequences of the probes p49f and p49a have been shown to match, respectively, the first and the fourth of the nine DAZ 2.4-kb tandem repeats (Saxena et al. 1996, fig. 5).

After *Taq*I digestion, probes p49f and p49a detect, in combination, 18 restriction fragments of different sizes, named "A"–"R," by size order, with fragment A being the largest and fragment R being the smallest. These fragments are male specific and correspond to the DAZ gene cluster, except for fragments K and L, which correspond to DAZH on chromosome 3 (Saxena et al. 1996, fig. 7). At least five fragments (A, C, D, F, and I) are considered polymorphic because they can be either present, absent, or, in the case of A, D, F and I, variable in size. For the A, D, and F series, the co-occurrence of several "allelic" fragments of distinct sizes is observed in a number of haplotypes (e.g., A3/A2 in haplotype 29; see appendix A). Also, fragments generally considered as constant bands (such as B, G, and H) are missing in several haplotypes, and additional new fragments are occasionally observed. Different mutation mechanisms had been proposed to explain this polymorphism (Ngo et al. 1986; Torroni et al. 1990*b;* Spurdle and Jenkins 1992*b;* Santachiara-Benerecetti et al. 1993) before the description of the DAZ gene cluster. It is now likely that the marker p49a,f/*Taq*I reveals a polymorphism specific to the array of 2.4-kb tandem repeats of DAZ, a polymorphism that could be due to different mechanisms: (1) nucleotide substitutions in the sequence of some 2.4-kb tandem repeats could account for the distinct fragment series (e.g., A, B, C, etc.) observed in p49a,f/*Taq*I electrophoretic profiles; (2) variation of the number of 2.4-kb tandem repeats within a DAZ unit could lead either to the missing fragment phenotypes (e.g., A0, C0, G0, etc.) or to additional fragment phenotypes (e.g., +3.5 kb, etc.); and/or (3) divergence of the sequences of the same repeats on distinct copies of the DAZ gene could account for the observation of different alleles of homologous fragments (e.g., A3/A2, D2/D1, etc.) in the same individual. Checking the validity of these complex mutation mechanisms is clearly beyond the scope of the present paper, and we will not attempt to do it here. However, in order to reflect the uncertainties of the evolutionary relationships between the haplotypes, we hereafter will use the terminology "band group" to refer to the distinct fragment series (e.g., A, C, D, etc.), and we will give the same weight to any mutation affecting a haplotype (see below).

The p49a,f/*Taq*I polymorphic system has already been tested on >70 samples worldwide. This important data collection thus provides a unique opportunity to investigate Y-chromosome variation at both large and fine scales. In this study, we review all data available in the literature on this polymorphic system, and we also incorporate unpublished data from several populations. Using a new nomenclature unifying the notations elaborated by different authors, we analyze the nature and the extent of human population structure inferred from this paternally transmitted chromosome segment and discuss the relative importance of geographic and linguistic factors in shaping the observed variability. A comparison of the genetic variability revealed by mtDNA RFLP data with that of the Y chromosome–specific p49a,f/*Taq*I polymorphism is also presented for 19 population samples analyzed for both polymorphisms. The differences in the pattern of genetic diversity inferred from these maternally and paternally transmitted markers are discussed.

Material and Methods

*p49a,f/*Taq*I Haplotype Nomenclature*

Y-chromosome p49a,f/*Taq*I haplotypes, presented in appendix A, were defined as the combination of alternative restriction fragments at eight polymorphic band groups (A–D and F–I), plus one additional band group containing information such as the absence of a very common band or the presence of a new rare restriction fragment. A total of seven possible distinct states were considered for this last group.

Samples Choice

We selected samples from two distinct sources: 45 samples were selected from a total of 60 published samples, on the basis of sample size ($n \geq 20$) and ethnic homogeneity; 13 additional samples taken from various Mediterranean populations have also been included,

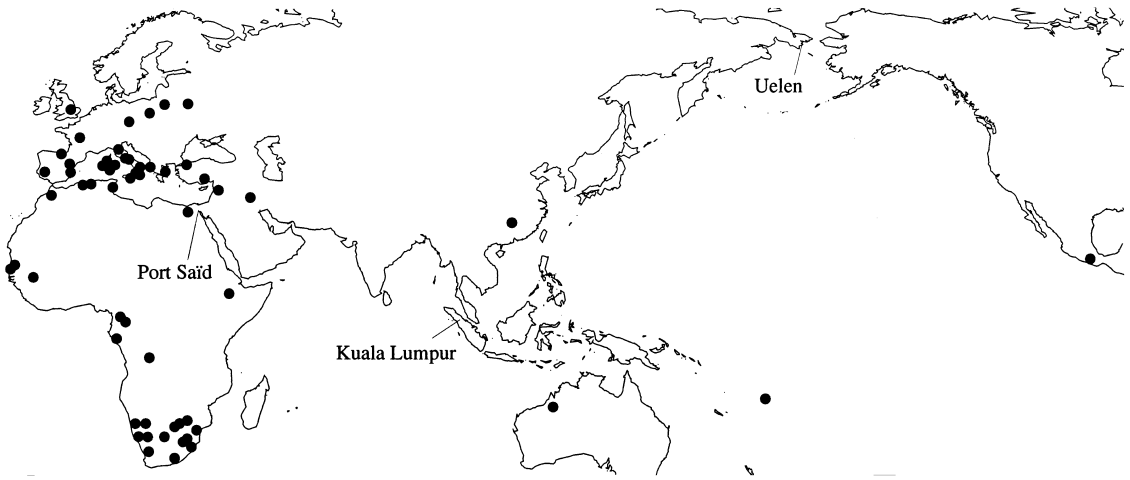

Figure 1 Geographic location of 58 samples used in analysis of Y-chromosome p49a,f/*Taq*I variability. Port Said, Kuala Lumpur, and Uelen are the three intercontinental gateways used to compute geographic distances between populations from different continents (see text).

even though the data are still unpublished (A. S. Santachiara-Benerecetti, unpublished data). We thus gathered a total of 58 samples (appendix B and fig. 1), representing the analysis of 3,767 chromosomes for the p49a,f/*Taq*I polymorphism.

Statistical Analyses

Population genetic–structure indexes were estimated by use of the analysis-of-molecular-variance approach (AMOVA) (Excoffier et al. 1992), which leads to *F*-statistics indexes averaged over all band groups (Michalakis and Excoffier 1996). In this analysis, both haplotype frequencies and molecular differences between the haplotypes are taken into account. Because of our ignorance concerning the exact mutation mechanism leading to the observed polymorphism, we consider all allelic variants of a given band group as equally distant from each other. The number of band-group differences is then taken as a genetic distance between pairs of haplotypes. Giving the same weight to different types of allelic differences is certainly a conservative assumption with regard to the detection of genetic structure. Variance components due to different sources of variation (between individuals within demes, between demes within populations, and between populations) were estimated, and their significance was tested by use of a nonparametric permutational procedure (Excoffier et al. 1992). Genetic distances between populations were obtained as coancestry coefficients, or linearized pairwise F_{ST} values (Reynolds et al. 1983). The significance of the genetic distance between any pair of populations was tested by use of a resampling permutational procedure (Excoffier et al. 1992). The genetic distances were considered significant if their associated probability was <5%. The matrix of genetic distances between 58 population samples was used as input for a multidimensional scaling analysis (Kruskal 1964).

Possible factors of population genetic differentiation were studied by means of correlation and partial correlation coefficients computed between genetic-, geographic-, and linguistic-distance matrices. The geographic-distance matrix was build up as a matrix of the logarithm of great-circle distances between pairs of populations, on the basis of sample geographic coordinates (fig. 1 and appendix B). In order to make these distances more realistic, the path between any two populations located on different continents was forced to pass through the following three gateways: Port Said (Egypt), between sub-Saharan Africa and Eurasia; Kuala Lumpur (Indonesia), between mainland and insular Southeast Asia; and, finally, Uelen (Siberia), between Asia and America (see fig. 1). The matrix of pairwise linguistic distances among populations was built up according to the method described by Excoffier et al. (1991) and the language classification, by Ruhlen (1987), of worldwide linguistic families. Linguistic distances between pairs of populations were defined as simple dissimilarity indexes. The way in which these dissimilarity indexes have been constructed assumes that the classification of linguists reflects the historical fissions from common ancestral languages within families and that the differences between language families are much larger than the differences between languages belonging to the same family. In more detail, they were computed as follows: two populations within the same language family are set to a distance of 3 if they belong to different subfamilies; their distance is decreased by 1 for each shared level of classification—up to three shared levels, where their dis-

tance is set to 0. The linguistic distance was not refined any further at the intrafamily level, for two reasons: first, linguistic families are not all characterized with the same level of precision by the classification at hand, and, second, the linguistic classification of the available populations was possible only up to a certain degree. Finally, because the evolutionary distances between language families are still largely unknown but assumed to be important, a dissimilarity index of 8 was arbitrarily assigned to any pair of populations belonging to different language families. The linguistic assignation of the 58 samples into 10 distinct linguistic families is given in appendix B. In the Results section, we discuss the effect of alternative weighting schemes. The genetic-, geographic-, and linguistic-distance matrices were then used as input for pairwise and multiple Mantel tests (Mantel 1967; Smouse et al. 1986).

Male-mediated genetic diversity was compared with female-mediated genetic diversity in a subset of 19 samples previously tested for mtDNA RFLPs by use of five enzymes (*Ava*II, *Hpa*I, *Hae*II, *Msp*I, and *Bam*HI) available from the literature. Some individuals may have been tested for both Y-chromosome and mtDNA markers in the same population, but this information was not available to us. Levels of population genetic structure estimated for both Y-chromosome and mtDNA polymorphic systems were compared. Patterns of female- and male-population genetic differentiation were analyzed by means of principal-coordinates analyses using the statistical package NTSYS, version 1.8 (Rohlf 1993). The strength of the association of populations' genetic affinities observed through Y-chromosome and mtDNA data was tested with correlation and partial correlation coefficients.

Results

Haplotype Geographic Distribution

After the initial numbering of 16 p49a,f/*Taq*I haplotypes presented in the study by Ngo et al. (1986), several studies of p49a,f/*Taq*I variability have each provided their own subsequent nomenclature for newly observed haplotypes, but no standard haplotype notation has been agreed on yet. We have identified at least nine cases in which different authors have attributed a different number to the same haplotype. This confusing situation has prompted us to revise all published haplotypes and their nomenclature and to provide a new numbering system, which is based on the dates of the initial descriptions of the haplotypes. The new nomenclature is presented in appendix A and lists the 144 distinct haplotypes identified, to date, at a worldwide scale.

A subset of 126 haplotypes were observed in the 58 samples used in this analysis. Among these, only 43 were observed at a frequency of $\geqslant 5\%$ in at least one sample. Only a few haplotypes (4, 5, 7, 8, 11, 12, and 15) were observed at frequencies $\geqslant 5\%$ in several samples of at least two continental regions. Globally, we observe that the p49a,f/*Taq*I polymorphism displays a diversity pattern in which a few haplotypes are found at high frequencies in populations from a given linguistic family, along with a number of rare haplotypes often not seen anywhere else.

We report in appendix C the mean frequencies of the most common haplotypes in the major linguistic groups. Haplotypes 5 and 11 are the most ubiquitous haplotypes, since they are observed in almost all the samples analyzed. Haplotype distribution in Niger-Congo samples is characterized by the presence of haplotype 4 at very high frequencies (>50% in all samples but the Xhosa), as well as by the presence of haplotype 5 at substantial frequencies, together with a few low-frequency haplotypes. In the Afro-Asiatic and Indo-European groups, it is common to find a few haplotypes (5, 7, 8, 11, 12, or 15) with frequencies of ~10%–20%, along with several other low-frequency haplotypes. Indo-European and Afro-Asiatic samples also tend to show a larger number of distinct haplotypes, as compared with Niger-Congo samples, and this difference does not seem to be due to unequal sample sizes. Haplotypes 7 and 8 are especially common among populations from the Mediterranean basin. Haplotype 15 is very common among European populations. The haplotypic distribution of the two Khoisan-speaking samples, Sekele San and Tsumkwe San, is characterized by a restricted number of frequent haplotypes, some of which (haplotypes 26, 50, and 51) are rarely observed in other groups. Inferences on haplotype distribution in Asia, Oceania, and America are impeded by the small number of samples in these regions. More details on the distribution of frequent p49a,f/*Taq*I haplotypes in specific geographic areas can be found in several other reports (Torroni et al. 1990*b,* 1994; Persichetti et al. 1992; Spurdle and Jenkins 1992*b,* 1996; Santachiara-Benerecetti et al. 1993; Semino et al. 1996; G. Passarino, unpublished data).

Tests of Selective Neutrality

The hypothesis of selective neutrality and population equilibrium under the infinite alleles model was rejected by the Ewens-Watterson neutrality test (Watterson 1978) in 13 (22%) of 58 samples (appendix B). These test rejections are always associated with a very low level of gene diversity due to the presence of one high-frequency haplotype. Interestingly, 8 of the 13 significant samples belong to the Niger-Congo linguistic group. They are characterized by a very high frequency of haplotype 4, associated with low frequencies of other haplotypes. Thus, almost half (8 of 18) of the Niger-Congo samples

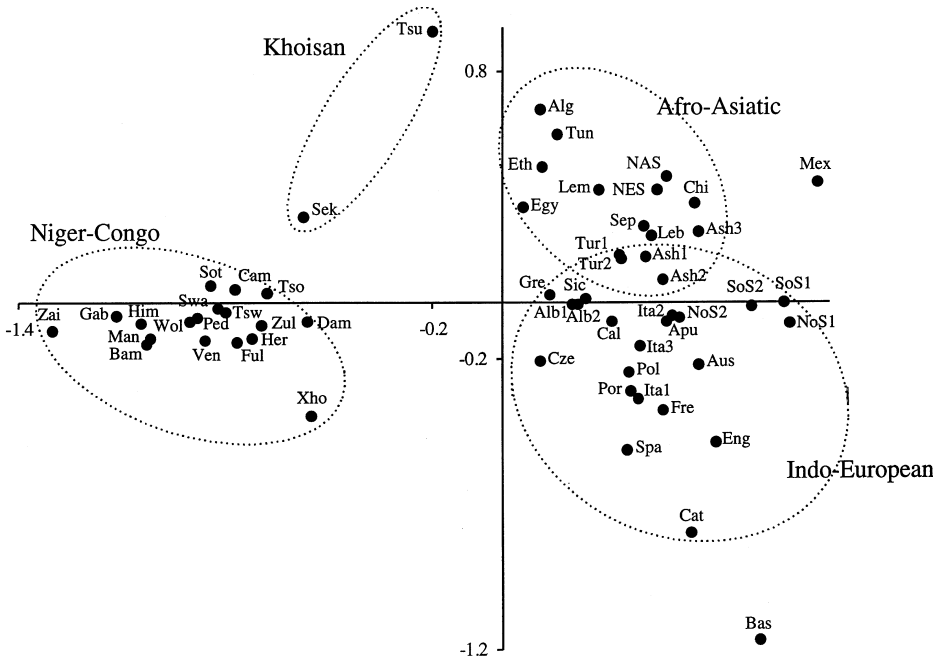

Figure 2 Multidimensional scaling analysis of 58 samples tested for Y chromosome–specific p49a,f/*Taq*I polymorphism. Genetic and linguistic distances are significantly correlated ($r = .567$, $P < .001$).

present a frequency distribution statistically different from that expected under the hypothesis of neutrality and population equilibrium. Note that the Khoisan-speaking Dama sample, for which selective neutrality was rejected, also presents a haplotypic configuration similar to that observed in Niger-Congo samples.

Genetic Structure

A multidimensional scaling analysis revealing broad genetic affinities between populations is shown in figure 2. Three main clusters of genetically closely related populations are observed: a Niger-Congo group, an Afro-Asiatic group, and an Indo-European group. As previously observed for a smaller number of samples (Excoffier et al. 1996), Niger-Congo populations show a low level of differentiation ($F_{ST} = .04$; table 1), and 54% of the pairwise genetic distances within this cluster are found to be nonsignificant. The three Khoisan-speaking population samples are found to be statistically different from each other, and they do not really cluster together in the multivariate space. The two Khoisan populations other than the Dama are also found to be statistically different from all other populations. The Dama sample is rather integrated within the Niger-Congo cluster, from which it is not very differentiated: 83% of the pairwise distances between the Dama and Niger-Congo samples are found to be nonsignificant. Contrastingly, the Lemba sample, belonging to the Niger-Congo language family, is found to be statistically different from all other Niger-Congo samples and appears within the Afro-Asiatic cluster. The Xhosa sample is located somewhat outside the tight Niger-Congo cluster on figure 2, but it is found to be not different from the Zulu, the Fulani from Senegal, and the Dama. On the basis of these observations, the other statistical analyses were performed by consideration of the Dama sample as being included within the Niger-Congo group and the Lemba sample as being included within the Afro-Asiatic group.

Non–sub-Saharan African samples are all grouped on the right-hand portion of figure 2, together with the southern-African Lemba sample and the Ethiopian Amharic sample. Ethiopians are not statistically differentiated from the Egyptian and Tunisian samples, in agreement with their linguistic affiliation with the Afro-Asiatic family. Afro-Asiatic and Indo-European samples differentiate along the second axis of the multivariate analysis. The Sephardim Jews, the Ashkenazim Jews, the Turks, and the Lebanese samples are genetically located at the intersection of these two linguistic groups, the Ashkenazim samples being somewhat closer to Indo-Europeans. Italian samples occupy a central position among Indo-European populations but also display close genetic affinities with some Afro-Asiatic samples. Spanish Basques, Czechoslovaks, Chinese, and Mexican Indians are statistically different from all other samples. Australians and Polynesians, who are located within the Indo-European cluster on figure 2, are, however, different from almost all other samples, the exceptions being

Table 1

Indexes of Population Structure Revealed by Y-Chromosome p49a,f/*Taq*I Polymorphism

Standardized Variance[a]	No. of Samples	No. of Haplotypes	F_{ST}	
Among populations:				
Global analysis	58	126	.230***	
Khoisan	2	11	.129**	
Afro-Asiatic	9	62	.085***	
Indo-European	22	75	.071**	
Niger-Congo	18	34	.039***	
			F_{CT}	F_{SC}
Among groups and among populations within groups:				
Niger-Congo vs. Khoisan vs. Afro-Asiatic vs. Indo-European		109	.246***	.070***
Niger-Congo vs. Khoisan vs. Afro-Asiatic		77	.332***	.067***
Niger-Congo vs. Afro-Asiatic		74	.355***	.065***
Niger-Congo vs. Indo-European		88	.320***	.065***
Niger-Congo vs. Khoisan		38	.247*	.044***
Indo-European vs. Khoisan		81	.219***	.073***
Afro-Asiatic vs. Khoisan		65	.214***	.088***
Afro-Asiatic vs. Indo-European		97	.083***	.075***

[a] Khoisan includes the Sekele San and Tsumkwe San samples; Afro-Asiatic includes all Afro-Asiatic–speaking population samples and the Lemba sample; and Niger-Congo includes the Dama sample and all Niger-Congo–speaking population samples but not the Lemba sample.

* $P < .01$.

** $P < .005$.

*** $P < .001$.

Table 2

Correlation and Partial Correlation Coefficients, between Genetic, Geographic, and Linguistic Distances, Computed on the Basis of Y-Chromosome p49a,f/*Taq*I Data

Distances Considered	Correlation Coefficient (r)	Proportion of Genetic Variance Explained (%)
Genetic and geographic	.613***	37.6
Genetic and linguistic	.567***	32.1
Genetic, geographic, and linguistic		44.1
	Partial Correlation Coefficient[a]	
Genetic and geographic (linguistic kept constant)	.419***	
Genetic and linguistic (geographic kept constant)	.323***	

NOTE.—The Dama and Lemba samples were not included in this analysis, because of the observed discrepancy between their linguistic affiliation and their genetic affinities.

[a] Between the Y-chromosome genetic distances and one of the two predictor variables (geography or linguistics), with the second variable kept constant.

*** $P < .001$.

nonsignificant genetic distances between (1) Australians and a northern Sardinian sample, (2) Polynesians and an Italian sample, and (3) Polynesians and the Portuguese sample.

The global F_{ST} value computed for the 126 p49a,f/*Taq*I distinct haplotypes found in the set of 58 samples is .230 (table 1), indicating a considerable degree of population differentiation for this Y chromosome–specific marker. Levels of population structure within the three clusters defined in figure 2 are also reported in table 1, as F_{ST} indexes. The Khoisan group shows the largest level of population differentiation ($F_{ST} = .129$), followed by the Afro-Asiatic ($F_{ST} = .085$) and Indo-European ($F_{ST} = .071$) groups. As mentioned earlier, a low but still significant level of structuration is observed in the Niger-Congo group ($F_{ST} = .039$). The extent of differentiation between groups of populations is reported in the second part of table 1, as F_{CT} indexes. A considerable degree of genetic structure is observed within the African continent, because of a strong genetic differentiation between the Niger-Congo, the Afro-Asiatic, and the Khoisan groups ($F_{CT} = .332$). In agreement with figure 2, a lower level of population structuration is observed between Indo-European and Afro-Asiatic populations ($F_{CT} = .083$).

Linguistics and Geography

Correlation and partial correlation coefficients between geographic, linguistic, and genetic distances are reported in table 2. All correlation coefficients observed

are high and significant ($P < .001$), suggesting that both geographic and historical factors have participated to the genetic differentiation of these populations. Geographic proximity and linguistic classification, considered as two predictor variables, are themselves highly correlated ($r = .588$, $P < .001$), reflecting the fact that linguistic differentiation follows a strong pattern of geographic structuration.

As shown in table 2, ~38% of the genetic variance is explained by geography, and ~44% is explained when one adds linguistics. Conversely, linguistics alone explains only ~32% of the genetic variance, and hence the geographic location of the populations contributes marginally more than linguistics to the prediction of genetic distances. Nonetheless, the two significant partial correlation coefficients reported in table 2 indicate that both predictor variables influence the observed genetic variability in different ways. However, one has to keep in mind that 56% of the observed genetic variability is due to unknown factors, other than geography and linguistics.

Because of the present lack of quantitative tools to accurately estimate linguistic relationships, our estimates of linguistic distances are certainly a crude approximation of evolutionary relationships between languages. Since the scale of values chosen to reflect linguistic differences (0, 1, 2, 3, and 8) was based on an a priori decision, we investigated the effect of varying the value of the dissimilarity index assigned to pairs of populations belonging to distinct linguistic families. We found that the correlation between genetics and linguistics increased when more weight was given to differences between language families. However, the correlation quickly reached an asymptote for values >8. Indeed, when interfamily distances of 4, 6, 8, 10, and 16 were used, the correlation coefficient, *r,* between genetic distances and linguistic distances was found to be equal to .517, .559, .567, .570, and .571, respectively (all values significant [$P < .001$]). Thus, we can conclude that the relationship between language and genetics would not be drastically affected by alternative weights given to language-family differences.

Comparison with mtDNA

The genetic diversity patterns of a subset of 19 populations studied for both Y-chromosome p49a,f/*Taq*I haplotypes and mtDNA low-resolution RFLPs were compared. The number of individuals tested for Y-chromosome ($n = 1,428$) and mtDNA markers ($n = 1,478$) is very similar, and the number of distinct haplotypes found in the two data sets is identical (91 haplotypes). We observe a globally higher level of population structuration for mtDNA ($F_{ST} = .271$, $P < .001$) than for the Y chromosome ($F_{ST} = .191$, $P < .001$), but the difference between the two statistics could not be tested. The results of two separate principal-coordinates analyses are shown on figure 3. Two separate clusters are clearly visible: a Niger-Congo cluster and another cluster, including mostly Indo-European samples.

Comparably low levels of genetic variability are found, for both systems, within the group of Indo-European samples ($F_{ST_Y} = .055$, $P < .001$, $F_{ST_{mtDNA}} = .024$, $P < .001$). Contrastingly, in the Niger-Congo cluster, we observe a much lower level of population structure for the Y chromosome ($F_{ST_Y} = .023$, $P < .01$) than for mtDNA ($F_{ST_{mtDNA}} = .163$, $P < .001$). This had first been interpreted as a drastic reduction in genetic variability for the Y chromosome among Niger-Congo populations (Excoffier et al. 1996). However, we now find that the large level of mtDNA diversity among Niger-Congo populations can be attributed entirely to the Herero sample (fig. 3); the F_{ST} value computed among Niger-Congo mtDNA samples drops from .163 to .027 ($P < .001$) when the Herero sample is removed from the analysis and thus becomes very similar to that computed for Y-chromosome data for the same populations ($F_{ST} = .030$, $P < .02$). We can therefore conclude that, barring the Herero, the level of genetic differentiation is virtually identical for the Y chromosome and mtDNA in sampled Niger-Congo populations.

For the polymorphisms tested, Y-chromosome and mtDNA genetic distances are highly and significantly correlated (table 3) ($r = .529$, $P < .001$) for the 19 samples considered here. We note that this correlation is reduced to a larger extent when linguistics is taken into account than when geography is controlled. In keeping with the larger correlation of Y-chromosome polymorphism with linguistics ($r = .699$) than with geography ($r = .529$), for the 19 samples considered, this suggests that the clustering observed in figure 3*a* is mainly due to linguistic differences between populations. Another interesting feature emerging from table 3 is the lack of significant correlation between mitochondrial polymorphism and geography or linguistics once Y-chromosome polymorphism is taken into account. Since the correlation between these two predictor variables and Y-chromosome polymorphism is preserved when mitochondrial polymorphism is controlled, it suggests that female-mediated genetic diversity can be better predicted from the knowledge of male-mediated diversity than vice versa.

Discussion

The pattern of populations' genetic affinities inferred from the Y-chromosome p49a,f/*Taq*I polymorphism is very similar to that inferred from conventional genetic markers (Cavalli-Sforza et al. 1988; Nei and Roychoudhury 1993; Sanchez-Mazas et al. 1994), nuclear DNA (Tiercy et al. 1992; Bowcock et al. 1994), and mtDNA (Excoffier et al. 1996), in the sense that population clus-

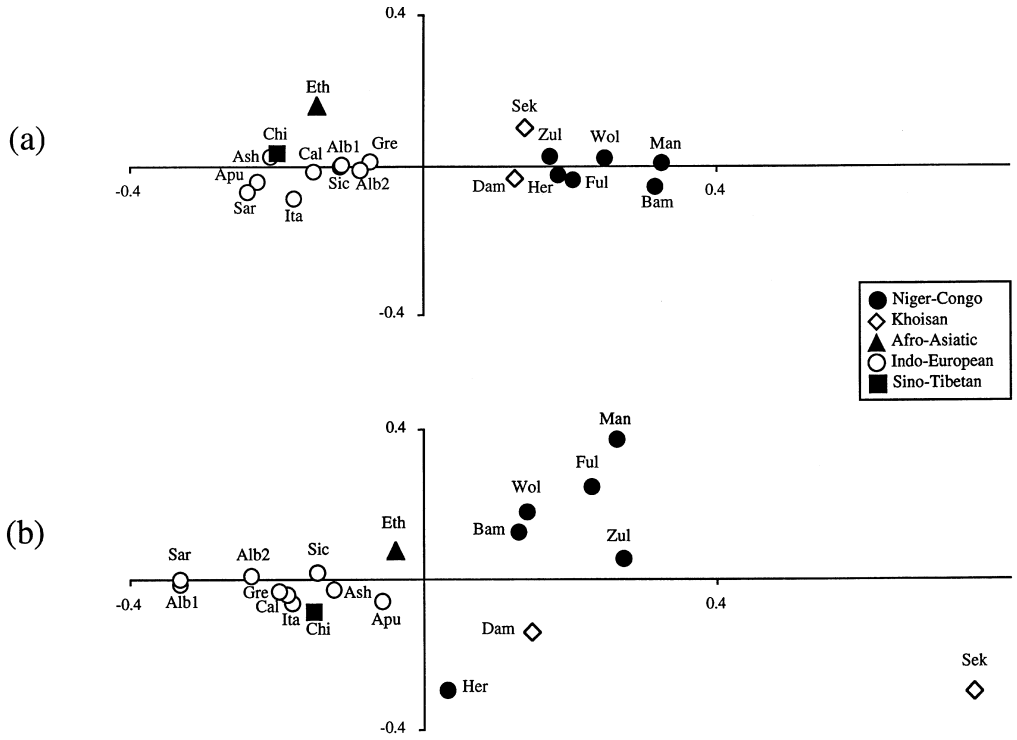

Figure 3 Principal-coordinates analysis of 19 samples tested for (*a*) Y chromosome–specific p49a,f/*Taq*I polymorphism (first axis—73% of total variance; second axis—6% of total variance) and (*b*) mtDNA RFLPs for enzymes *Ava*II, *Hpa*I, *Hae*II, *Msp*I, and *Bam*HI (first axis—56% of total variance; second axis—21% of total variance). Alb1 = Albanians 1 (De Benedictis et al. 1994); Alb2 = Albanians 2 (Albanians in Calabria) (Torroni et al. 1990*a*); Apu = Apulians (De Benedictis et al. 1989*b*); Ash = Ashkenazim (Ritte et al. 1993*b*); Bam = Bamileke (Scozzari et al. 1994); Cal = Calabrians (De Benedictis et al. 1989*a*); Chi = Chinese (Ballinger et al. 1992); Dam = Dama (Soodyall and Jenkins 1993); Eth = Ethiopians (G. Passarino, unpublished data); Ful = Fulani (Scozzari et al. 1988); Gre = Greeks (Astrinidis and Kouvatsi 1994); Her = Herero (Soodyall and Jenkins 1993); Ita = Italians (Brega et al. 1986); Man = Mandenka (Graven et al. 1995); Sar = Sardinians (Brega et al. 1986); Sek = Sekele San (Soodyall and Jenkins 1992); Sic = Sicilians (Semino et al. 1989); Wol = Wolof (Scozzari et al. 1988); and Zul = Zulu (Johnson et al. 1983).

ters defined on the basis of genetic information broadly correspond to linguistic and regional groups of populations. This correspondence has three main implications. It first suggests that there has not been any strong selective sweep in the recent history of Y-chromosome lineages, in agreement with both the study by Goldstein et al. (1996) on Y-chromosome microsatellite variation and the quite old Y-chromosome coalescence times (>150,000 years) estimated by Tavaré et al. (1997). Second, the overall pattern of differentiation and gene flow seem to have been fairly similar for men and women in the recent evolution of modern humans, even though some interesting exceptions can be postulated, as discussed below. Finally, the p49a,f/*Taq*I polymorphism appears to be a very useful marker in anthropology, despite the uncertainties concerning its exact mutation scheme (Spurdle et al. 1994*a*). Even though regions containing repeated sequences are difficult to analyze, they recently have been shown to contain a lot of information concerning the genetic history of human populations (Armour et al. 1996). In view of the high variability found for the p49a,f/*Taq*I marker, both a finer characterization of the repeated structure of the DAZ gene and its study at the sequence level would certainly provide a very powerful tool to better depict the evolution and spread of modern men.

Genetic Structure Inferred from the Y Chromosome

A considerable extent of local differentiation in human populations is found for the Y chromosome (F_{ST} = .230), mainly due to broad differences between clusters of populations corresponding to language families (fig. 2). The level of population structure estimated for this Y chromosome–specific marker is larger than F_{ST} values inferred from autosomal markers and mitochondrial polymorphisms (range .09–.14, as reviewed by Relethford 1995). This larger value is expected because of the haploid mode of transmission of the Y chromosome, leading to a smaller effective population size. For instance, the equilibrium value of F_{ST} under the island model is $F_{ST} = 1/(1 + Nm)$ for Y chromosome or

Table 3

Correlation and Partial Correlation Coefficients, between Genetic Distances, Computed on the Basis of Y-Chromosome P49a,f/*Taq*I Data and mtDNA RFLP Data, Considering the Association with Geographic and Linguistic Distances

Distances Considered	Correlation Coefficient	Partial Correlation Coefficient
Y chromosome and mtDNA	.529***	
Y chromosome and mtDNA (geographic kept constant)		.400***
Ychromosome and mtDNA (linguistic kept constant)		.281***
Y chromosome and geographic	.529***	
Y chromosome and geographic (mtDNA kept constant)		.400***
Y chromosome and linguistic	.699***	
Y chromosome and linguistic (mtDNA kept constant)		.588***
mtDNA data and geographic	.417**	
mtDNA and geographic (Y chromosome kept constant)		.191 (NSa)
mtDNA and linguistic	.510***	
mtDNA and linguistic (Y chromosome ept constant)		.231 (NSa)

[a] NS = not significant at the 5% level.
** $P < .01$.
*** $P < .001$.

mtDNA, whereas it is $F_{ST} = 1/(1 + 4Nm)$ for nuclear loci, where N is the effective diploid population size and m is the migration rate among populations. Even though this model is quite irrelevant to the case of human populations, one can see how a larger fixation index is expected for haploid markers. If one agrees with a mean nuclear F_{ST} value of .1 (Relethford 1995), the same pattern of migration would lead to a value of .307 for Y-chromosome markers, a value indeed very close to that found by Hammer et al. (1997) ($F_{ST} = .302$) for the Y-chromosome YAP haplotype polymorphism. The present value of $F_{ST} = .230$ for p49a,f/*Taq*I haplotypes is indeed lower than both that for YAP haplotypes and that expected under the island model, but this discrepancy can be due to factors others than the fact that the island model does not correctly depict the pattern of migrations among populations: first, the populations analyzed here have been mainly sampled in Africa and around the Mediterranean area, therefore not completely covering the worldwide diversity; second, recurrent mutations may have acted on the p49a,f/*Taq*I polymorphism to lower the global level of diversity among distant populations. However, these homoplasic events may not have been common enough to erase the structure of populations, in contrast with some Y chromosome–specific microsatellite markers (Deka et al. 1996).

A hierarchical analysis of variance reveals that ~25% of the total diversity is due to differences among language families (table 1), whereas only 5% of the variance is due to differences among populations within language families, which agrees with the population clustering pattern observed in figure 2. It thus confirms that the high F_{ST} level (.230) discussed above reflects mainly the large differences between the clusters of populations defined in figure 2. If the extent of molecular differences between haplotypes is not taken into account and genetic structure is estimated from mere haplotype frequencies, we then obtain a lower F_{ST} value, .167 ($P < .001$). This suggests that not only genetic drift but also mutations in this particular segment of the Y chromosome have played an important role in differentiating human populations. Because of the low level of genetic diversity

found within each linguistic group, it appears that mutations are more important in distinguishing between linguistic groups than in differentiating populations belonging to the same language family.

Differentiation Within and Between Language Families

Our present analysis supports the existence of distinct genetic entities correlated with linguistic families in sub-Saharan Africa (fig. 2), in keeping with results obtained with classical markers (Excoffier et al. 1987, 1991). This view is reflected by the sharp differentiation observed among African populations belonging either to the Niger-Congo, the Afro-Asiatic, or the Khoisan language families. The Niger-Congo group, extending geographically from western to southern Africa, shows the lowest level of internal genetic diversity, a result that is compatible with the hypothesis of a recent dispersal of Niger-Congo populations (Greenberg 1963; Phillipson 1977; Ehret 1984; Vansina 1984). This recent expansion event could also explain the high proportion of neutrality-test rejections among Niger-Congo samples, since sudden expansions also lead to highly unbalanced haplotype frequency distributions (Maruyama and Fuerst 1985; Watterson 1986). An approximation of the time of divergence of Niger-Congo populations can be derived from the observed F_{ST} value of .039, by use of the well-known relationship $F_{ST} = 1 - /(1 - N)^t$. This equation relates the level of population differentiation to the effective population size, N, and the divergence time, t, and assumes that population differentiation is due only to genetic drift, which seems reasonable if divergence time is small. If we use a male effective population size of 5,000 individuals (Hammer 1995; Goldstein et al. 1996; Tavaré et al. 1997), we find an estimated t = 199 generations, or ~4,000 years if 20 years/generation is assumed. This time is certainly underestimated, since the equation above was derived by assuming that there has been no gene flow between populations since their simultaneous divergence. However, this divergence time for Niger-Congo populations is in very good agreement with archaeological and linguistic estimations for the expansion of Bantu speakers from the Niger-Congo border (Greenberg 1963; Phillipson 1977; Ehret 1984; Vansina 1984). It could imply that the large migration that spread Bantu speakers to central and southern Africa would have been accompanied by a similar expansion toward western Africa. An alternative hypothesis required to explain this low level of diversity would invoke an episode of selective sweep that had been restricted to Niger-Congo Y chromosomes and that had favored haplotype 4. Because we observe comparably low levels of variability, among Niger-Congo populations, for mtDNA but also for other nuclear markers (Excoffier et al. 1991), a recent geographic range expansion seems to be a more plausible explanation.

Divergence times can also be computed for other linguistic groups. Samples of Khoisan speakers show a very high level of internal genetic diversity, a result that holds, whether or not the Dama are considered as belonging to this group (F_{ST} = .144 or F_{ST} = .129, respectively). Since the two San populations of hunter-gatherers have a much smaller population size, genetic drift certainly had a much stronger impact on the differentiation process. Thus, if an effective population size of 500 individuals is assumed, only 70 generations (1,400 years) are necessary to produce an F_{ST} value of .129. As the Khoisan effective population size is not known with any precision, this recent divergence could even be overestimated. The F_{ST} value of .085 in the Afro-Asiatic group leads to a slightly larger estimate of divergence time, ~8,900 years, in accordance with some estimates put forward on linguistic and archaeological grounds (Renfrew 1991). For the sampled Indo-European populations, an observed F_{ST} of .071 leads to an estimate of 368 generations (7,400 years) of divergence time, in good agreement with the attested spread of Neolithic farmers in Europe, by demic diffusion from the Middle East (Ammerman and Cavalli-Sforza 1984; Renfrew 1987, 1991).

As shown on figure 2, there are indications of relatively close genetic affinities between Afro-Asiatic and Indo-European groups of populations, despite the linguistic barrier. Populations from the Oriental part of the Mediterranean basin, as well as Ashkenazim samples, show genetic affinities with both the Afro-Asiatic and the Indo-European linguistic groups, which may be due to their intermediate geographic location. Alternatively, this relationship could be due to a recent common origin of the two language families, which have been recently grouped into the Nostratic superfamily, thought to have originated in the Middle-East (Kaiser and Shevoroshkin 1988; Dolgopolsky 1989; Bomhard and Kerns 1994). Note that the Spanish Basques, who do not belong to the Indo-European group, appear also genetically quite distinct from European populations, with regard to their male-transmitted genes. This is in keeping with observations based on classical markers (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994), but it strongly contrasts with results based on the sequencing of mtDNA (Bertranpetit et al. 1995), which would indicate a larger extent of female-mediated versus male-mediated gene flow into the Basque population.

Contrasting Male and Female Patterns of Gene Flow

Although the overall pattern of population differentiation globally appears to be very similar for male- and female-transmitted markers (fig. 3), some populations

clearly show different affinities for their maternal and paternal genetic components, as already noticed for Ethiopian Jews (Zoosmann-Diskin et al. 1991), Arab tribal groups in the Sinai Peninsula (Salem et al. 1996), Finns (Zerjal et al. 1997), and Basques (as discussed above). For instance, as mentioned above, the Khoisan-speaking Dama show very strong genetic affinities with Niger-Congo populations, for p49a,f/*Taq*I markers, a result that is in agreement with their supposed Bantu origin (Spurdle and Jenkins 1992*b*). They bear, however, obvious Khoisan features with regard to their mtDNA RFLPs, such as haplotypes 3 and 5 and their derivatives. This clearly suggests that this population has been founded by males of Bantu origin while retaining Khoisan features by the incorporation of female Khoisan speakers. According to the first axis of the two-principal-coordinate analyses in figure 3, the Ethiopian sample is more differentiated, from other African samples, for the Y chromosome than for mtDNA. This holds true when additional microsatellite markers on the Y chromosome are analyzed, and it has been attributed to unequal patterns of male and female gene flow between the Middle-East and Ethiopia (G. Passarino, unpublished data). The Niger-Congo–speaking Lemba sample is extremely divergent from other sub-Saharan Africans and shows close affinities with Afro-Asiatic populations (fig. 2), which is in agreement with the postulated Semitic origin of this population (Spurdle and Jenkins 1992*b,* 1996). Finally, the Herero are not significantly different from Niger-Congo populations, with regard to Y-chromosome polymorphism, whereas they present large differences with regard to their mitochondrial genome (fig. 3). The presence of the typical Khoisan mitochondrial RFLP haplotypes 3 and 21 in the Herero suggests that they have certainly incorporated Khoisan genes through female gene flow.

Even though the global pattern of populations' genetic differentiations is mostly similar for mtDNA and the Y chromosome, differential gene flow through men and women, resulting in unequal contributions to population gene pools, as exemplified above, may indeed have been a fairly common phenomenon throughout the history of our species. The relationships of Y-chromosome and mtDNA genetic distances with geography and linguistics in 19 populations presented in table 3 suggest that the differentiation of cultures has left a stronger mark in the male-specific component of the human genome than in its female counterpart. The fact that the Y chromosome explains mtDNA relationships with geography and language but that mtDNA does not explain Y-chromosome relationships with these factors (see table 3) is compatible with the idea (*a*) that, in expanding populations, men and women would have traveled together and that local women (but not men) would have been recruited during and after the migrations or (*b*) that it would have been easier for women to cross cultural boundaries. As a consequence, the female-specific diversity of our genome would fit less well with geography and linguistics than would our male-specific component. Additional comparative studies between mitochondrial and Y-chromosome polymorphisms are clearly necessary to confirm the present interpretation and to see whether it can be extended to groups other than Niger-Congo and Indo-Europeans, whence most of our samples came. If that were to prove to be the case, then the common belief that we speak our mother's tongue should be revised in favor of the concept of a "father's tongue."

Am. J. Hum. Genet. 61:1015–1035, 1997

Appendix A

The nomenclature and reference sources for the haplotypes used in the present study are provided in table A1.

Table A1

Haplotype Nomenclature Adopted in Present Study

Haplotype	No. of Allelic Variants					Missing (−)/ Additional (+) Fragment(s)	Original Number	Original Reference(s)
	A	C	D	F	I			
Other studies:								
1	0	0	0	1	1		I	Ngo et al. (1986)
2	0	0	1	1	1		II	Ngo et al. (1986)
3	1	0	0	1	0		III	Ngo et al. (1986)
4	1	0	0	1	1		IV	Ngo et al. (1986)
5	2	0	0	1	1		V	Ngo et al. (1986)
6	2	0	1	0	1		VI	Ngo et al. (1986)
7	2	0	1	1	0		VII	Ngo et al. (1986)
8	2	0	1	1	1		VIII	Ngo et al. (1986)
9	2	1	0	1	1		IX	Ngo et al. (1986)
10	3	0	0	1	0		X	Ngo et al. (1986)
11	3	0	0	1	1		XI	Ngo et al. (1986)
12	3	0	1	1	0		XII	Ngo et al. (1986)
13	3	0	1	1	1		XIII	Ngo et al. (1986)
14	3	1	1	1	1		XIV	Ngo et al. (1986)
15	3	1	2	1	1		XV	Ngo et al. (1986)
16	4	0	1	1	0		XVI	Ngo et al. (1986)
17[a]	2	0	b	1	1		XVII	Breuil et al. (1987)
18	4	0	1	1	1		XVIII	Lucotte et al. (1989)
19	0	0	1	0	0		XIX	Lucotte et al. (1989)
20	3	1	0	1	0		XX	Lucotte et al. (1989)
21	3	1	1	1	0		XXI	Lucotte et al. (1989)
22	2	1	1	1	0		XXII	Lucotte et al. (1989)
23	2	1	2	1	1		XXIII	Lucotte et al. (1989)
24	2	1	1	1	1		XXIV	Lucotte et al. (1989)
25	0	0	0	0	1		XL(27)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
26	2	0	0	0	1		XLI (33)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
27	1	0	0	0	1		XXIII(34)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
28	1	0	0	1	0	G −	XXXVIII (41)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
29	3/2	1	2	1	1		XXXIV(50)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
30	2	0	0	1	0		XXVII(56)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
31	4/3	1	2	1	1		XXXVII(60)	Torroni et al. (1990*b*) (Spurdle and Jenkins [1992*b*])
32	0	0	0	1	0		XXII	Torroni et al. (1990*b*)
33[b]	2	0	1	0	1	5.2 +	XXV	Torroni et al. (1990*b*)

(*continued*)

Table A1 (continued)

Haplotype	No. of Allelic Variants					Missing (−)/ Additional (+) Fragment(s)	Original Number	Original Reference(s)
	A	C	D	F	I			
34[b]	2	0	1	1	0	2.5 +	XXVI	Torroni et al. (1990*b*)
35	3	1	0	1	1		XXVIII	Torroni et al. (1990*b*)
36	3	1	2	1	0		XXIX	Torroni et al. (1990*b*)
37	3	1	2	0	1		XXXI	Torroni et al. (1990*b*)
38	3	0	1	0	0		XXXII	Torroni et al. (1990*b*)
39[c]	3	0	0	1	0	BHPR −	XXXIII	Torroni et al. (1990*b*)
40	3/2	1	2/1	1	1		XXXV	Torroni et al. (1990*b*)
41	5/2	1	0	1	1		XXXVI	Torroni et al. (1990*b*)
42[b]	1	0	0	0	1	3.5 +	XXXIX	Torroni et al. (1990*b*)
43	0	0	1	1	0		XXI	Torroni et al. (1990*b*)
44	0	0	0	0	1	O −	XX	Torroni et al. (1990*b*)
45	2	0	1	1	0	G −	VII-G	Lucotte et al. (1990)
46	3	1	2	1	1	B −	XV-B	Lucotte et al. (1990)
47	3/2	1	0	1	1		25	Spurdle and Jenkins (1992*b*)
48	3	0	3	1	0		26	Spurdle and Jenkins (1992*b*)
49	3	1	0	1	0	B −	28 (XXX)	Spurdle and Jenkins (1992*b*) (Torroni et al. [1990*b*])
50	3	0	0	0	0		29	Spurdle and Jenkins (1992*b*)
51	3/2	0	0	1	1		30	Spurdle and Jenkins (1992*b*)
52	3/2	0	0	1	0		31	Spurdle and Jenkins (1992*b*)
53	3	0	0	0	1		32	Spurdle and Jenkins (1992*b*)
54	3/2	0	0	0	0		35	Spurdle and Jenkins (1992*b*)
55	4	0	0	1	1		36	Spurdle and Jenkins (1992*b*)
56	4/3	0	2	1	1		37	Spurdle and Jenkins (1992*b*)
57[c]	3/2	0	0	1	0	BH −	38	Spurdle and Jenkins (1992*b*)
58	0	0	0	0	0		39	Spurdle and Jenkins (1992*b*)
59	4/3	0	0	1	1		40	Spurdle and Jenkins (1992*b*)
60	5	0	1	1	1		42	Spurdle and Jenkins (1992*b*)
61	4	0	0	1	0		43	Spurdle and Jenkins (1992*b*)
62	4/3	0	0	0	1		44	Spurdle and Jenkins (1992*b*)
63	4/3/2	1	2/1	1	1		45	Spurdle and Jenkins (1992*b*)
64	5/3	0	0	1	1		46	Spurdle and Jenkins (1992*b*)
65	5/3	1	2	1	1		47	Spurdle and Jenkins (1992*b*)
66	3	0	1	1	0	B −	48	Spurdle and Jenkins (1992*b*)
67	0	0	1	1	0	BG −	49	Spurdle and Jenkins (1992*b*)
68	2	0	1	1	0	B −	51 (XLV)	Spurdle and Jenkins (1992*b*) (Santachiara-Benerecetti et al. [1992])
69	3	0	2	1	1		52	Spurdle and Jenkins (1992*b*)
70	2	0	1	0	0	B −	53	Spurdle and Jenkins (1992*b*)
71	4	0	0	0	1		54	Spurdle and Jenkins (1992*b*)
72	2	0	0	0	0		55	Spurdle and Jenkins (1992*b*)
73	3	1	0	0	1		57 (XLVII)	Spurdle and Jenkins (1992*b*) (Santachiara-Benerecetti et al. [1992])
74	0	0	2	0	1		58	Spurdle and Jenkins (1992*b*)
75	4	0	0	0	0		59	Spurdle and Jenkins (1992*b*)
76	4	1	2	1	1		61	Spurdle and Jenkins (1992*b*)
77[b]	0	0	1	2/1	1		62	Spurdle and Jenkins (1992*b*)
78	3	0	2	1	0		New-a	Persichetti et al. (1992)
79	2	0	2	0	1		New-d	Persichetti et al. (1992)
80	4/3	1	0	1	1		New-f (48)	Persichetti et al. (1992) (Santachiara-Benerecetti et al. [1993])
81	1	0	1	1	1		42	Santachiara-Benerecetti et al. (1993)
82	1	0	1	1	0		43	Santachiara-Benerecetti et al. (1993)
83	2	0	2/1	0	1		46	Santachiara-Benerecetti et al. (1993)
84	5	0	0	1	1		50	Santachiara-Benerecetti et al. (1993)
85	0	0	1	0	1		51	Santachiara-Benerecetti et al. (1993)

(*continued*)

Table A1 (continued)

Haplotype	No. of Allelic Variants					Missing (−)/ Additional (+) Fragment(s)	Original Number	Original Reference(s)
	A	C	D	F	I			
86[c]	3/2	0	0	1	1	R −	53	Santachiara-Benerecetti et al. (1993)
87	3/2	1	2	1	0		55	Santachiara-Benerecetti et al. (1993)
88[c]	6/3	0	0	1	0	BHPR −	62	Santachiara-Benerecetti et al. (1993)
89	2	0	1	0	0		44	Santachiara-Benerecetti et al. (1993)
90	3	0	1	0	1		70	Ritte et al. (1993*a*)
91	4	0	1	0	1		73	Ritte et al. (1993*a*)
92	5	0	1	0	1		74	Ritte et al. (1993*a*)
93	1	0	1	0	1		65	Ritte et al. (1993*a*)
94	2	0	1	1	1	B −	66	Ritte et al. (1993*a*)
95	2	0	2	1	1		68	Ritte et al. (1993*a*)
96	2	1	1	0	1		69	Ritte et al. (1993*a*)
97	3	0	2/1	1	1		71	Ritte et al. (1993*a*)
98	3	1	1	0	1		72	Ritte et al. (1993*a*)
99	3/2	0	1	1	1		64	Ritte et al. (1993*a*)
100	4/2	0	1	0	1		75	Ritte et al. (1993*a*)
101	4/3	1	0	0	1		76	Ritte et al. (1993*a*)
102	4/3	1	1	1	1		77	Ritte et al. (1993*a*)
103[b]	1	0	0	1	1	3.7 +		Lucotte et al. (1994)
104	3/2	0	2	1	1		64	Torroni et al. (1994)
105	6	0	1	1	1		65	Torroni et al. (1994)
106	5	0	3/1	1	1		66	Torroni et al. (1994)
107	5/4	0	0	1	1		67	Torroni et al. (1994)
108	6/2	0	1	1	0	B −	68	Torroni et al. (1994)
109	3	0	2/1	1	0		New-1	Jobling (1994)
110	4/2	0	2	1	1		New-2	Jobling (1994)
111	<3	0	3/1	1	1		New-3	Jobling (1994)
112	4	0	1	1	1	B −	New-4	Jobling (1994)
113[b]	3	0	0	1	1	2.7 +	65	Spurdle et al. (1994*b*)
114	3	0	1	1	2		66	Spurdle et al. (1994*b*)
115	4	0	1	1	2		67	Spurdle et al. (1994*b*)
116	2	0	0	1	0	G −		G. Passarino (unpublished data)
117	3/2	0	0	0	1			G. Passarino (unpublished data)
118	4/3	0	2	0	1			O. Semino (unpublished data)
119	4	0	1	1	0	B −		A. S. Santachiara-Benerecetti (unpublished data)
120	5	0	0	1	0			A. S. Santachiara-Benerecetti (unpublished data)
121	2	0	3/1	0	1			A. S. Santachiara-Benerecetti (unpublished data)
122	<3	0	3/1	1	0			A. S. Santachiara-Benerecetti (unpublished data)
123	3/1	0	1	1	1			A. S. Santachiara-Benerecetti (unpublished data)
124	4/2	0	1	1	1			A. S. Santachiara-Benerecetti (unpublished data)
125	6/5	0	0	1	0			A. S. Santachiara-Benerecetti (unpublished data)
126	6	0	0	1	1		116	Excoffier et al. (1996)
127[b]	2	0	0	1	0	2.5 +	117	Excoffier et al. (1996)
128	3	0	0	1	0	G −	118	Excoffier et al. (1996)
								Sample(s) Where First Observed
Present study:								
129	3/2	1	2	0	1			Northern Sardinians 2
130	2/1	0	0	1	0			Northern Sardinians 2
131	4/2	0	0	1	0			Southern Sardinians 2
132	5/3	0	2	1	1			Calabrians
133	2/1	0	1	1	1			Calabrians
134	4/3/2	0	2	1	1			Sicilians
135	2	0	1	1	2			Sicilians
136	3/2	0	1	1	0			Turks from Konya
137	4/3/2	0	1	1	0			Turks from Konya, Albanians in Calabria

(*continued*)

Table A1 (continued)

Haplotype	No. of Allelic Variants					Missing (−)/ Additional (+) Fragment(s)	Sample(s) Where First Observed
	A	C	D	F	I		
138	0	1	1	1	1		Turks from Konya, Tunisians
139	3	0	0	1	0	B −	Greeks
140	3	0	1	1	0	BG −	Albanians
141	4	1	0	1	1		Albanians
142	3/2	1	0	1	0	B −	Albanians in Calabria
143[c]	0	0	0	0	1	P −	Albanians in Calabria
144[b]	3	0	1	1	0	2.5 +	Lebanese

[a] Fragment Db (size 7.6 kb), identified in a sample of 21 Baruya men from Papua New Guinea, all of whom were monomorphic for haplotype 17, has not yet been observed in other tested samples or individuals.

[b] The additional 5.2-kb fragment (haplotype 33) was considered to be the same as the 5.9-kb fragment F2 (haplotype 77); the additional 2.5-kb fragment (haplotypes 34, 127, and 144) was considered to be the same as the additional 2.7-kb fragment (haplotype 113); and the additional 3.5-kb fragment (haplotype 42) was considered to be the same as the additional 3.7-kb fragment (haplotype 103). For further information, see the original references.

[c] Observing concomitant absence of bands B, H, P, and R (haplotype 39, in one ltalian individual), Torroni et al. (1990*b*) suggested that the loss of these four fragments was due to a single mutational event (a deletion). Haplotype 39 was observed in an additional six individuals (from Czechoslovakians, Turks, Greeks, and Albanians), and the same pattern of absent bands was observed in association with haplotype 88 (one Czechoslovakian individual). Spurdle and Jenkins (1992*b*) observed the absence of bands B and H in haplotype 57 (two Tsumkwe San individuals). The electrophoretic resolution that these authors used did not allow them to identify fragments beneath the O band, but they note that, in at least one case, they were able to observe the absence of band H, associated with the absence of band P; the present nomenclature does not reflect this latter case. Since all haplotypes identified by Spurdle and Jenkins (1992*b*) would have to be checked for the presence or absence of fragments beneath the O band, we tentatively considered that they all bear bands P and R. Furthermore, the absence of band R, associated with the presence of bands B, H and P, was observed in haplotype 86 (one Czechoslovakian individual), and the absence of band P, associated with the presence of bands B, H, and R, was observed in haplotype 143 (one Albanaian individual). We thus tentatively considered the absence of the B, H, P, and R fragments as being four independent mutational events.

Appendix B

Background information and references sources for the 58 samples used in the present study are provided in table B1.

Table B1

Description of 58 Samples Used in Global Analysis of Y-Chromosome p49a,f/*Taq*I Variability

Sample	Code	No. of Chromosomes Analyzed	No. of Distinct Haplotypes Observed	Gene Diversity[a]	Linguistic Affiliation	Geographic Location	Reference
Other studies:							
Wolof (Senegal)	Wol	41	7	.51	Niger-Congo	16°01′ N, 16°30′ W	Torroni et al. (1990*b*)
Fulani (Senegal)	Ful	22	8	.67*	Niger-Congo	16°01′ N, 16°30′ W	Torroni et al. (1990*b*)
Bamileke (Cameroon)	Bam	22	4	.32*	Niger-Congo	3°51′ N, 11°31′ E	Torroni et al. (1990*b*)
Italians 1	Ita1	125	25	.88	Indo-European	45°28′ N, 9°12′ E	Torroni et al. (1990*b*)
Australians	Aus	59	8	.82	Australian	17°19′ S, 123°38′ E	Lucotte et al. (1991)
Zulu	Zul	53	8	.60	Niger-Congo	31°35′ S, 28°47′ E	Spurdle and Jenkins (1992*b*)
Xhosa	Xho	23	7	.74	Niger-Congo	33°58′ S, 25°36′ E	Spurdle and Jenkins (1992*b*)
Swazi	Swa	33	7	.55	Niger-Congo	26°20′ S, 31°08′ E	Spurdle and Jenkins (1992*b*)
Sotho	Sot	48	7	.58	Niger-Congo	29°19′ S, 27°29′ E	Spurdle and Jenkins (1992*b*)

(*continued*)

Table B1 (continued)

Sample	Code	No. of Chromosomes Analyzed	No. of Distinct Haplotypes Observed	Gene Diversity[a]	Linguistic Affiliation	Geographic Location	Reference
Pedi	Ped	53	8	.54	Niger-Congo	23°54′ S, 29°23′ E	Spurdle and Jenkins (1992*b*)
Tswana	Tsw	41	9	.59*	Niger-Congo	24°59′ S, 25°19′ E	Spurdle and Jenkins (1992*b*)
Tsonga	Tso	31	9	.66*	Niger-Congo	27°22′ S, 29°54′ E	Spurdle and Jenkins (1992*b*)
Venda	Ven	28	5	.60	Niger-Congo	23°54′ S, 29°23′ E	Spurdle and Jenkins (1992*b*)
Lemba	Lem	49	9	.80	Niger-Congo	23°54′ S, 29°23′ E	Spurdle and Jenkins (1992*b*)
Himba	Him	37	5	.25**	Niger-Congo	21°28′ S, 15°56′ E	Spurdle and Jenkins (1992*b*)
Herero	Her	46	10	.47**	Niger-Congo	22°30′ S, 18°58′ E	Spurdle and Jenkins (1992*b*)
Dama	Dam	24	9	.64**	Khoisan	24°50′ S, 17°00′ E	Spurdle and Jenkins (1992*b*)
Sekele San	Sek	56	10	.81	Khoisan	19°07′ S, 13°39′ E	Spurdle and Jenkins (1992*b*)
Tsumkwe San	Tsu	23	7	.78	Khoisan	19°13′ S, 17°42′ E	Spurdle and Jenkins (1992*b*)
Egyptians	Egy	34	15	.85	Afro-Asiatic	31°03′ N, 31°23′ E	Persichetti et al. (1992)
Italians 2	Ita2	20	10	.86	Indo-European	41°53′ N, 12°30′ E	Persichetti et al. (1992)
Northern Sardinians 1	NoS1	23	9	.69*	Indo-European	40°43′ N, 8°34′ E	Persichetti et al. (1992)
Southern Sardinians 1	SoS1	25	6	.79	Indo-European	39°13′ N, 9°08′ E	Persichetti et al. (1992)
English	Eng	21	14	.89	Indo-European	51°30′ N, 0°10′ W	Persichetti et al. (1992)
French	Fre	196	16	.79	Indo-European	48°52′ N, 2°20′ E	Lucotte and David (1992)
Italians 3	Ita3	46	12	.86	Indo-European	45°28′ N, 9°12′ E	Lucotte and David (1992)
Spanish	Spa	31	9	.73	Indo-European	41°25′ N, 2°10′ E	Lucotte and David (1992)
Portuguese	Por	26	10	.77	Indo-European	38°44′ N, 9°08′ W	Lucotte and David (1992)
Sephardim	Sep	83	17	.88	Afro-Asiatic	36°50′ N, 3°00′ E	Santachiara-Benerecetti et al. (1993)
Ashkenazim 1	Ash1	83	16	.85	Indo-European	53°51′ N, 27°30′ E	Santachiara-Benerecetti et al. (1993)
Czechoslovaks	Cze	105	27	.91	Indo-European	50°06′ N, 14°26′ E	Santachiara-Benerecetti et al. (1993)
Ashkenazim 2	Ash2	159	15	.84	Indo-European	52°15′ N, 21°00′ E	Lucotte et al. (1993)
Ashkenazim 3	Ash3	35	13	.81	Indo-European	50°03′ N, 19°55′ E	Ritte et al. (1993*a*)
Northern African Sephardim	NAS	94	35	.92	Afro-Asiatic	34°02′ N, 6°51′ W	Ritte et al. (1993*a*)
Near Eastern Sephardim	NES	78	22	.91	Afro-Asiatic	33°20′ N, 44°26′ E	Ritte et al. (1993*a*)
People from Gabon	Gab	27	2	.20	Niger-Congo	0°30′ N, 9°25′ E	Lucotte et al. (1994)
People from Cameroon	Cam	76	7	.63	Niger-Congo	3°51′ N, 11°31′ E	Lucotte et al. (1994)
People from Zaire	Zai	188	9	.34*	Niger-Congo	4°18′ S, 15°18′ E	Lucotte et al. (1994)
Mexican Indians	Mex	31	11	.75*	Amerind	17°05′ N, 96°41′ W	Torroni et al. (1994)
Polynesians	Pol	59	12	.78	Austronesian	21°09′ S, 175°14′ W	Spurdle et al. (1994*b*)
Ethiopians	Eth	57	10	.78	Afro-Asiatic	9°03′ N, 38°42′ E	G. Passarino (unpublished data)
Basques from Spain	Bas	52	10	.63*	Basque	43°15′ N, 2°56′ W	O. Semino (unpublished data)

(*continued*)

Table B1 (continued)

Sample	Code	No. of Chromosomes Analyzed	No. of Distinct Haplotypes Observed	Gene Diversity[a]	Linguistic Affiliation	Geographic Location	Reference
Basques from Spain	Bas	52	10	.63*	Basque	43°15′ N, 2°56′ W	O. Semino (unpublished data)
Catalans	Cat	28	11	.82	Indo-European	41°25′ N, 2°10′ E	O. Semino (unpublished data)
Chinese	Chi	193	25	.87	Sino-Tibetan	28°10′ N, 113°00′ E	Liu et al. (1994)
Mandenka (Senegal)	Man	64	12	.51**	Niger-Congo	12°35′ N, 12°09′ W	Excoffier et al. (1996)
Present study:							
Northern Sardinians 2	NoS2	117	21	.85	Indo-European	40°43′ N, 8°34′ E	A. S. Santachiara-Benerecetti (unpublished data)
Southern Sardinians 2	SoS2	98	17	.82	Indo-European	39°13′ N, 9°08′ E	A. S. Santachiara-Benerecetti (unpublished data)
Calabrians	Cal	91	20	.88	Indo-European	38°06′ N, 15°39′ E	A. S. Santachiara-Benerecetti (unpublished data)
Sicilians	Sic	84	21	.88	Indo-European	38°08′ N, 13°23′ E	A. S. Santachiara-Benerecetti (unpublished data)
Apulians	Apu	62	14	.88	Indo-European	41°07′ N, 16°52′ E	A. S. Santachiara-Benerecetti (unpublished data)
Turks from Istanbul	Tur1	77	19	.85	Altaic	41°02′ N, 28°57′ E	A. S. Santachiara-Benerecetti (unpublished data)
Turks from Konya	Tur2	131	24	.88	Altaic	37°51′ N, 32°30′ E	A. S. Santachiara-Benerecetti (unpublished data)
Greeks	Gre	90	20	.90	Indo-European	38°00′ N, 23°44′ E	A. S. Santachiara-Benerecetti (unpublished data)
Albanians	Alb1	56	15	.86	Indo-European	41°20′ N, 19°49′ E	A. S. Santachiara-Benerecetti (unpublished data)
Albanians in Calabria	Alb2	82	23	.91	Indo-European	38°06′ N, 15°39′ E	A. S. Santachiara-Benerecetti (unpublished data)
Algerians	Alg	58	13	.64**	Afro-Asiatic	36°50′ N, 3°00′ E	A. S. Santachiara-Benerecetti (unpublished data)
Tunisians	Tun	85	11	.73	Afro-Asiatic	36°50′ N, 10°13′ E	A. S. Santachiara-Benerecetti (unpublished data)
Lebanese	Leb	88	14	.84	Afro-Asiatic	33°52′ N, 35°30′ E	A. S. Santachiara-Benerecetti (unpublished data)

[a] Probability values (see below) indicate significant departure from selective neutrality and population equilibrium.

* $P < .05$.

** $P < .01$.

Appendix C

For the distribution of the most frequently occurring p49a,f/*Taq*I haplotypes in the linguistic groups discussed in the present study, see table C1.

Table C1

Mean Frequency Distribution of Most Frequent p49a,f/*Taq*I Haplotypes in Linguistic Groups

	Frequency (%)			
Haplotype (ACDFI)	Niger-Congo[a]	Khoisans[b]	Afro-Asiatics[c]	Indo-Europeans
Ubiquitous:				
5 (20011)	11.0	…	22.6	9.6
11 (30011)	3.0	9.3	8.1	7.2
Frequently observed in at least two continental regions:				
4 (10011)	65.1	16.1	2.7	1.5
7 (20110)	.6	…	8.3	10.9
8 (20111)	.6	…	22.9	10.3
12 (30110)	.4	…	1.8	12.8
15 (31211)	1.1	…	.9	18.4
Observed at frequency ⩾10% in at least one sample:				
3 (10010)	5.6	3.1	.2	.6
10 (30010)	4.1	…	.9	2.4
13 (30111)	…	…	2.2	1.0
14 (31111)	…	…	.3	.6
18 (40111)	.1	…	2.3	1.1
24 (21111)	…	…	.6	1.8
26 (20001)	.3	20.6	2.2	…
31 (4/31211)	…	…	.1	.5
35 (31011)	…	…	2.0	5.1
37 (31201)	…	…	.1	.8
47 (3/21011)	…	…	.3	.8
50 (30000)	…	15.8	.1	…
51 (3/20011)	.2	21.6	.1	.2
60 (50111)	…	…	.5	.1
81 (10111)	…	…	1.7	.1
Other[d]	8.0	13.8	19.1	14.3

[a] Includes the Dama sample and all Niger-Congo–speaking population samples but the Lemba sample.
[b] Includes the Sekele San and Tsumkwe San samples.
[c] Includes all Afro-Asiatic–speaking population samples plus the Lemba sample.
[d] Cumulative frequencies of other haplotypes.

Acknowledgments

We are grateful to Peter Smouse and an anonymous reviewer for their helpful comments on an earlier draft of the manuscript. A.L. and L.E. were supported by Swiss FNRS grants 31-39847.93 and 32-047053.96, and A.S.S.-B. was supported by Italian MURST funds (40% and 60%).

References

Ammerman AJ, Cavalli-Sforza LL (1984) Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, et al (1996) Minisatellite diversity supports a recent African origin for modern humans. Nat Genet 13: 154–160

Astrinidis A, Kouvatsi A (1994) Mitochondrial DNA polymorphism in northern Greece. Hum Biol 66:601–611

Ballinger SW, Schurr TG, Torroni A, Gan YY, Hodge JA, Hassan K, Chen K-H, et al (1992) Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. Genetics 130:139–152

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

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

Bishop CE, Guellaën G, Geldwerth D, Voss R, Fellous M, Weissenbach J (1983) Single copy DNA sequences specific for the human Y chromosome. Nature 303:831–832

Bomhard AR, Kerns JC (1994) The Nostratic macrofamily, a study in distant linguistic relationship. Mouton de Gruyter, Berlin, New York

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evol-

utionary trees with polymorphic microsatellites. Nature 368: 455–457

Brega A, Scozzari R, Maccioni M, Iodice C, Wallace DC, Bianco I, Cao A, et al (1986) Mitochondrial DNA polymorphisms in Italy. I. Population data from Sardinia and Rome. Ann Hum Genet 50:327–338

Breuil S, Hallé L, Ruffié J, Lucotte G (1987) Polymorphisme d'une sonde Y-spécifique nommée p49 chez les Papous Baruya de Nouvelle-Guinée. Ann Génét 30:209–212

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

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

Ciminelli BM, Pompei F, Malaspina P, Hammer M, Persichetti F, Pignatti PF, Palena A, et al (1995) Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations. J Mol Evol 41:966–973

De Benedictis G, Passarino G, Leone O, Falcone E, Santachiara-Benerecetti AS, Boletini E, Semino O (1994) Mitochondrial DNA polymorphisms in a sample of Albanian population (Tirana). Int J Anthropol 9:129–135

De Benedictis G, Rose G, Caccio S, Picardi P, Quagliariello C (1989*a*) Mitochondrial DNA polymorphism in Calabria (southern Italy). Gene Geogr 3:33–40

De Benedictis G, Rose G, Passarino G, Quagliariello C (1989*b*) Restriction fragment length polymorphism of human mitochondrial DNA in a sample population from Apulia (southern Italy). Ann Hum Genet 53:311–318

Deka R, Jin L, Shriver MD, Yu LM, Saha N, Barrantes R, Chakraborty R, et al (1996) Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Genome Res 6:1177–1184

Dolgopolsky AB (1989) Problems of Nostratic comparative phonology. In: Shevoroshkin V (ed) Reconstructing languages and cultures. Brockmeyer, Bochum, Germany, pp 90–98

Dorit RL, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183–1185

Ehret C (1984) Historical/linguistic evidence for early African food production. In: Clark JD, Brandt SA (eds) From hunters to farmers: the cause and consequences of food production in Africa. University of California Press, Berkeley, pp 26–35

Ellis N, Taylor A, Bengtsson BO, Kidd J, Rogers J, Goodfellow P (1990) Population structure of the human pseudoautosomal boundary. Nature 344:663–665

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, Pellegrini B, Sanchez-Mazas A, Simon C, Langaney A (1987) Genetics and history of sub-Saharan Africa. Yearbook Phys Anthropol 30:151–194

Excoffier L, Poloni ES, Santachiara-Benerecetti S, Semino O, Langaney A (1996) The molecular diversity of the Niokholo Mandenkalu from eastern Senegal: an insight into West Africa genetic history. In: Boyce AJ, Mascie-Taylor CGN (eds) Molecular biology and human diversity. Cambridge University Press, Cambridge, pp 141–155

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

Goldstein DB, Zhivotovsky LA, Nayar K, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13:1213–1218

Gomolka M, Hundrieser J, Nürnberg P, Roewer L, Epplen JT, Epplen C (1994) Selected di- and tetranucleotide microsatellites from chromosomes 7, 12, 14 and Y in various Eurasian populations. Hum Genet 93:592–596

Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol Biol Evol 12:334–345

Greenberg JH (1963) The languages of Africa. Mouton, The Hague

Hammer MF (1994) A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol 11:749–761

——— (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, et al (1997) The geographic distribution of human Y chromosome variation. Genetics 145: 787–805

Jakubiczka S, Arnemann J, Cooke HJ, Krawczak M, Schmidtke J (1989) A search for restriction fragment length polymorphism on the human Y chromosome. Hum Genet 84:86–88

Jobling MA (1994) A survey of long-range DNA polymorphisms on the human Y chromosome. Hum Mol Genet 3: 107–114

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell JR, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5: 1767–1775

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11: 449–456

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL (1983) Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J Mol Evol 19:255–271

Kaiser M, Shevoroshkin V (1988) Nostratic. Annu Rev Anthropol 17:309–329

Karafet T, Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, Romero F, et al (1997) Y chromosome markers and trans-Bering Strait dispersals. Am J Phys Anthropol 102: 301–314

Kruskal JB (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:28–42

Liu A, Hu L, Semino O, Brega A, Santachiara-Benerecetti AS (1994) Y chromosome specific RFLPs in Chinese confirm the genetic affinities between Amerindians and East Asians. Abstract presented at the 26th annual meeting of the European Society of Human Genetics, Paris, June 1–5

Lucotte G, David F (1992) Y-chromosome-specific haplotypes of Jews detected by probes 49f and 49a. Hum Biol 64: 757–761

Lucotte G, Gérard N, Krishnamoorthy R, David F, Aouizerate A, Galzot P (1994) Reduced variability in Y-chromosome-specific haplotypes for some central African populations. Hum Biol 66:519–526

Lucotte G, Guérin P, Hallé L, Loirat F, Hazout S (1989) Y chromosome DNA polymorphisms in two African populations. Am J Hum Genet 45:16–20

Lucotte G, Hazout S, Summers KM (1991) The p49/TaqI Y-specific DNA haplotypes in Australian aborigines. Gene Geogr 5:131–136

Lucotte G, Smets P, Ruffié J (1993) Y-chromosome-specific haplotype diversity in Ashkenazic and Sephardic Jews. Hum Biol 65:835–840

Lucotte G, Sriniva KR, Loirat F, Hazout S, Ruffié J (1990) The p49/TaqI Y-specific polymorphisms in three groups of Indians. Gene Geogr 4:21–28

Malaspina P, Persichetti F, Noveletto A, Jodice C, Terrenato L, Wolfe J, Ferraro M, et al (1990) The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet 54:297–305

Mantel NA (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Maruyama T, Fuerst PA (1985) Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics 111:675–689

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Maynard Smith J (1990) The Y of human relationships. Nature 344:591–592

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142: 1061–1064

Muller S, Gomolka M, Walter H (1994) The Y-specific SSLP of the locus DYS19 in four different European samples. Hum Hered 44:298–300

Nei M, Roychoudhury AK (1993) Evolutionary relationships of human populations on a global scale. Mol Biol Evol 10: 927–943

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J (1986) A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet 38:407–418

Oakey R, Tyler-Smith C (1990) Y chromosome DNA haplotyping suggest that most Europeans and Asian men are descended from one or two males. Genomics 7:325–330

Pääbo S (1995) The Y chromosome and the origin of all of us (men). Science 268:1141–1142

Pena SDJ, Santos FR, Bianchi NO, Bravi CM, Carnese FR, Rothhammer F, Gerelsaikhan T, et al (1995) A major founder Y-chromosome haplotype in Amerindians. Nat Genet 11:15–16

Persichetti F, Blasi P, Hammer M, Malaspina P, Jodice C, Terrenato L, Novelletto A (1992) Disequilibrium of multiple DNA markers on the human Y chromosome. Ann Hum Genet 56:303–310

Phillipson DW (1977) The later prehistory of eastern and southern Africa. Heinemann, London

Relethford, JH (1995) Genetics and modern human origins. Evol Anthropol 4:53–63

Renfrew C (1987) Archaeology and language : the puzzle of Indo-European Origins. Jonathan Cape, London

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

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics 105:767–779

Ritte U, Neufeld E, Broit M, Shavit D, Motro U (1993*a*) The differences among Jewish communities—maternal and paternal contributions. J Mol Evol 37:435–440

Ritte U, Neufeld E, Prager EM, Gross M, Hakim I, Khatib A, Bonné-Tamir B (1993*b*) Mitochondrial DNA affinity of several Jewish communities. Hum Biol 65:359–385

Roewer L, Arnemann J, Spurr NK, Grzeschik K-H, Epplen JT (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89:389–394

Rohlf FJ (1993) NTSYS-pc: numerical taxonomy and multivariate analysis system. Exeter Software, New York

Ruhlen M (1987) A guide to the world's languages. Vol 1: Classification. Edward Arnold, London

Ruiz Linares A, Nayar K, Goldstein DB, Hebert JM, Seielstad MT, Underhill PA, Lin AA, et al (1996) Geographic clustering of human Y-chromosomes haplotypes. Ann Hum Genet 60:401–408

Salem A-H, Badr FM, Gaballah MF, Pääbo S (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. Am J Hum Genet 59: 741–743

Sanchez-Mazas A, Bütler-Brunner E, Excoffier L, Ghanem N, Ben Salem M, Breguet G, Dard P, et al (1994) New data for AG haplotype frequencies in Caucasoid populations and selective neutrality of the AG polymorphism. Hum Biol 66: 27–48

Santachiara-Benerecetti AS, Semino O, Passarino G, Morpurgo GP, Fellous M, Modiano G (1992) Y-chromosome DNA polymorphisms in Ashkenazi and Sepharadi Jews. In: Bonné-Tamir B, Adam A (eds) Genetic diversity among Jews: diseases and markers at the DNA level. Oxford University Press, New York, pp 45–50

Santachiara-Benerecetti AS, Semino O, Passarino G, Torroni A, Brdicka R, Fellous M, Modiano G (1993) The common, Near-Eastern origin of Ashkenazi and Sephardi Jews supported by Y-chromosome similarity. Ann Hum Genet 57: 55–64

Santos FR, Gerelsaikhan T, Munkhtuja B, Oyunsuren T, Epplen JT, Pena SDJ (1996) Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism DYS19. Hum Genet 97:309–313

Santos F, Pena SDJ, Epplen JT (1993) Genetic and population

study of a Y-linked tetranucleotide repeat DNA polymorphism with a simple non-isotopic technique. Hum Genet 90: 655–656

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, et al (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nat Genet 14:292–299

Scozzari R, Torroni A, Semino O, Cruciani F, Spedini G, Santachiara-Benerecetti SA (1994) Genetic studies in Cameroon: mitochondrial DNA polymorphisms in Bamileke. Hum Biol 66:1–12

Scozzari R, Torroni A, Semino O, Sirugo G, Brega A, Santachiara-Benerecetti AS (1988) Genetic studies on the Senegal population. I. Mitochondrial DNA polymorphisms. Am J Hum Genet 43:534–544

Seielstad MT, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D, Cavalli-Sforza LL (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet 3:2159–2161

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

Semino O, Torroni A, Scozzari R, Brega A, De Benedictis G, Santachiara-Benerecetti AS (1989) Mitochondrial DNA polymorphisms in Italy. III. Population data from Sicily: a possible quantitation of maternal African ancestry. Ann Hum Genet 53:193–202

Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627–632

Soodyall H, Jenkins T (1992) Mitochondrial DNA polymorphisms in Khoisan populations from southern Africa. Ann Hum Genet 56:315–324

——— (1993) Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo. Ann Hum Biol 20:477–485

Spurdle AB, Hammer MF, Jenkins T (1994*a*) The Y *Alu* polymorphism in southern African populations and its relationship to other Y-specific polymorphisms. Am J Hum Genet 54:319–330

Spurdle A, Jenkins T (1992*a*) The search for Y chromosome polymorphism is extended to negroids. Hum Mol Genet 1: 169–170

——— (1992*b*) Y chromosome probe p49a detects complex *Pvu*II haplotypes and many new *Taq*I haplotypes in southern African populations. Am J Hum Genet 50:107–125

——— (1992*c*) The Y chromosome as a tool for studying human evolution. Curr Opin Genet Dev 2:487–491

——— (1996) The origins of the Lemba "Black Jews" of southern Africa: evidence from p12F2 and other Y-chromosome markers. Am J Hum Genet 59:1126–1133

Spurdle AB, Woodfield DG, Hammer MF, Jenkins T (1994*b*) The genetic affinity of Polynesians: evidence from Y chromosome polymorphisms. Ann Hum Genet 58:251–263

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518

Tiercy J-M, Sanchez-Mazas A, Excoffier L, Shi-Isaac X, Jeannet M, Mach B, Langaney A (1992) HLA-DR polymorphism in a Senegalese Mandenka population: DNA oligotyping and population genetics of DRB1 specificities. Am J Hum Genet 51:592–608

Torroni A, Chen Y-S, Semino O, Santachiara-Benerecetti AS, Scott CR, Lott MT, Winter M, et al (1994) mtDNA and Y-chromosome polymorphisms in four Native American populations from southern Mexico. Am J Hum Genet 54: 303–318

Torroni A, Semino O, Rose G, De Benedictis G, Brancati C, Santachiara-Benerecetti AS (1990*a*) Mitochondrial DNA polymorphisms in the Albanian population of Calabria (southern Italy). Int J Anthropol 5:97–104

Torroni A, Semino O, Scozzari R, Sirugo G, Spedini G, Abbas N, Fellous M, et al (1990*b*) Y chromosome DNA polymorphisms in human populations: differences between Caucasoids and Africans detected by 49a and 49f probes. Ann Hum Genet 54:287–296

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196–200

Vansina J (1984) Western Bantu expansions. J Afr Hist 25: 129–145

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405–417

——— (1986) The homozygosity test after a change in population size. Genetics 112:899–907

Whitfield LS, Hawkins TL, Goodfellow PN, Sulston J (1995*a*) 41 Kilobases of analyzed sequence from the pseudoautosomal and sex-determining regions of the short arm of the human Y chromosome. Genomics 27:306–311

Whitfield LS, Sulston JE, Goodfellow PN (1995*b*) Sequence variation of the human Y chromosome. Nature 378: 379–380

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

Zoosmann-Diskin A, Ticher A, Hakim I, Rubinstein A, Bonné-Tamir B (1991) Genetic affinities of Ethiopian Jews. Israel J Med Sci 27:245–251