THE ROYAL SOCIETY

Y-chromosome polymorphisms and the origins of the European gene pool

Rosa Casalotti[1]†, Lucia Simoni[2], Michele Belledi[1] and Guido Barbujani[3]*

[1]*Dipartimento di Biologia Evolutiva, Università di Parma, I-43100 Parma, Italy*
[2]*Dipartimento di Biologia Evoluzionistica e Sperimentale, Università di Bologna, via E. Selmi 1, 40126, Italy*
[3]*Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, 44100 Ferrara, Italy*

Gradients of allele frequencies have long been considered the main genetic characteristic of the European population, but mitochondrial DNA diversity seems to be distributed differently. One Alu insertion (YAP), five tetranucleotide (DYS19, DYS389B, DYS390, DYS391 and DYS393) and one trinucleotide (DYS392) microsatellite loci of the Y chromosome were analysed for geographical patterns in 59 European populations. Spatial correlograms showed clines for most markers, which paralleled the gradients previously observed for two restriction fragment length polymorphisms. Effective separation times between populations were estimated from genetic distances at microsatellite loci. Even after correcting for the possible effects of continuous local gene flow, the most distant Indo-European-speaking populations seem to have separated no more than 7000 years ago. The clinal patterns and the estimated, recent separation times between populations jointly suggest that Y-chromosome diversity in Europe largely reflects a directional demic expansion, which is unlikely to have occurred before the Neolithic period.

Keywords: DNA diversity; Y chromosome; microsatellites; spatial autocorrelation; Europe

1. INTRODUCTION

The European gene pool seems to have been formed in two main phases (Cavalli-Sforza *et al.* 1994): an initial Palaeolithic colonization and a later Neolithic demic diffusion. Both phases entailed a population expansion from the Levant (Ammerman & Cavalli-Sforza 1984). Later gene flow has also affected the distribution of genetic diversity (Sokal *et al.* 1993). Schematically, however, there are two alternative views on the origins of the European gene pool (outlined in figure 1).

(a) *A Palaeolithic origin of the European gene pool?*

According to this view, the current European gene pool largely originated in the Upper Palaeolithic, when the first anatomically modern humans moved in from the Near East. Radiocarbon evidence shows Upper Palaeolithic industries in south-eastern Europe more than 44 000 years ago and a westward and northward spread of these artefacts. Around 30 000 years ago, much of Europe was populated (see Richards *et al.* (1997) and references therein), although at low population densities (Birdsell 1968). Further demographic changes in the late Palaeolithic may have involved local extinctions and repopulations in response to climatic changes (see e.g. Torroni *et al.* 1998). The Neolithic diffusion of farming technologies is regarded as a consequence of a cultural, not demographic, process, which would have had limited effects on the composition of the European gene pool.

*Author for correspondence (bjg@dns.unife.it).
†Present address: Dipartimento di Biologia, Università di Roma 2—Tor Vergata, Italy.

(b) *A Neolithic origin of the European gene pool?*

According to this view, there was a large-scale population replacement in the Neolithic, between 10 000 and 5000 years ago. When technologies for food production were developed in the Levant (Renfrew 1987), a combination of demographic growth, individual dispersal and limited admixture with pre-existing European hunter–gatherers led to a westward diffusion of populations which did not inhabit Europe in the Palaeolithic (Harlan 1971; Ammerman & Cavalli-Sforza 1984). During that expansion, the populations of eastern Europe received, on average, greater shares of immigrating genes than the populations of central and western Europe (Menozzi *et al.* 1978). The Indo-European languages may also have spread in Europe through that Neolithic expansion (Renfrew 1987).

(c) *Palaeolithic colonization versus Neolithic demic diffusion*

Until recently, the extensive gradients of allele frequencies (Menozzi *et al.* 1978) and DNA variants (Semino *et al.* 1996; Chikhi *et al.* 1998*a*,*b*), comparisons with archaeological and linguistic evidence (Sokal *et al.* 1991; Barbujani *et al.* 1994) and computer simulations (Rendine *et al.* 1986; Barbujani *et al.* 1995) appeared to jointly support a major population replacement in the Neolithic. However, studies of mitochondrial DNA (mtDNA) variation showed little geographical structure in Europe (Richards *et al.* 1996, 1998) and allele genealogies that coalesce tens of thousands of years ago. By equating the ages of the mitochondrial alleles with the ages of the populations, Richards *et al.* (1996) proposed that 85% of the European gene pool was descended from the first Palaeolithic

Received 7 June 1999 *Accepted* 12 July 1999
© 1999 The Royal Society

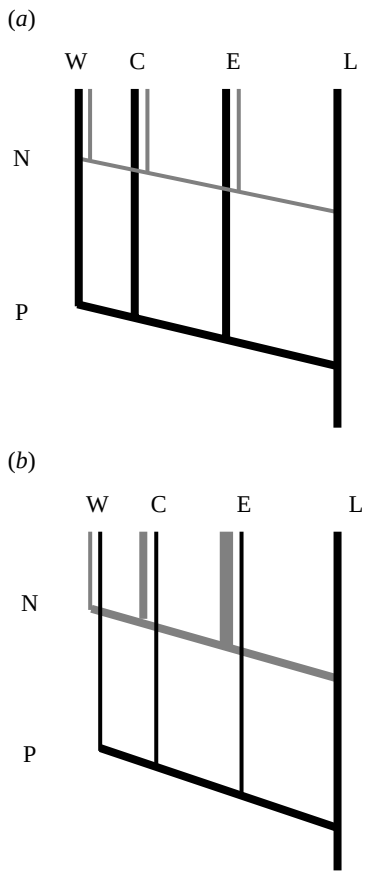

Figure 1. A schematic representation of the demographic processes involved in models of (*a*) Palaeolithic and (*b*) Neolithic origins of the European gene pool. The past is at the bottom of the figures, the present at the top; P is the Palaeolithic and N is the Neolithic. The thickness of the bars reflects the relative importance of genes that spread in the two periods in Western (W), Central (C) and Eastern (E) Europe from the Levant (L).

colonizers, with a limited contribution of Neolithic farming immigrants. Late Palaeolithic expansions from glacial refugia located in southern and central Europe could explain the absence of continentwide clines (Torroni *et al.* 1998; Sykes 1999).

Two crucial points are therefore the extent and origin of the gradients observed for many protein and DNA markers (Bertranpetit *et al.* 1996; Cavalli-Sforza & Minch 1997; Richards *et al.* 1997; Barbujani *et al.* 1998; Sykes 1999). Is the non-clinal pattern of mtDNA variation the exception or the rule at the DNA level? And does current DNA diversity suggest a recent separation of the European populations (i.e. at a moment compatible with a common ancestry in the Neolithic period) or an earlier subdivision (i.e. before the Neolithic farming technologies spread in Europe)? To address these questions, we studied the distributions of nine Y-chromosome markers in Europe, quantitatively describing patterns of spatial diversity and estimating probable dates of population split.

2. MATERIAL AND METHODS

(a) *The data*

The data set analysed was collected through an extensive search of published literature, integrated by unpublished information on Albanian, Italian and Turkish populations (figure 2). (The data set, including 16 562 individual records and bibliographical sources, is available from the corresponding author upon request.) Among the polymorphisms listed by Jobling & Tyler-Smith (1995), we considered five tetranucleotide (DYS19, DYS389B, DYS390, DYS391 and DYS393) and one trinucleotide (DYS392) microsatellite loci. By combining information on the DYS19 locus and the presence–absence of an Alu insertion (or YAP element, DYS287; Hammer 1994) two-locus haplotypes were also constructed, which will be hereafter referred to as DYS19/YAP. Not all markers were typed in all populations and so sample sizes varied across loci (see tables 1 and 2).

We also included in our analysis the data reported by Semino *et al.* (1996), referring to the 'European' alleles of the two restriction fragment length polymorphisms (RFLPs) systems p12f2/TaqI (DYS11) and 49a,f/TaqI (DYS1), namely the 8 kb fragment and the haplotype 15, respectively. In this way, nine Y-chromosome markers were analysed in parallel, two of which (DYS19 and DYS19/YAP) were not statistically independent. An additional factor reducing independence between the various sets of data, to an extent that proved impossible to quantify, is the fact that the same individuals may have been typed at more than one locus.

(b) *Spatial autocorrelation analysis*

We tested for spatial structure in the data by using two spatial autocorrelation methods designed for the treatment of frequency (SA: Sokal & Oden 1978) and DNA sequence (AIDA: Bertorelle & Barbujani 1995) data, respectively. Spatial autocorrelation measures the level of resemblance between samples (SA) or between individuals (AIDA) as a function of their distance in space. In this way, the spatial patterns produced by distinct classes of evolutionary phenomena can be objectively described and often recognized.

Autocorrelation statistics are estimated by comparing all pairs of samples separated by arbitrary spatial lags (in this study 1–500 km, 500–1000 km and so on). Positive or negative values of the coefficients summarizing frequency (SA) or sequence (AIDA) differences in a given class indicate genetic similarity or dissimilarity, respectively. Therefore, a decreasing set of coefficients at increasing distances describes a genetic gradient, whereas isolation by distance is reflected in an asymptotic decline of autocorrelation, from positive significance at short distances to non-significant. AIDA differs from SA in that alleles of very different length contribute to negative autocorrelation more than alleles differing by one or two repeats only. In this way, the AIDA statistics reflect both allele frequency differences between samples and length differences between alleles.

(c) *Locating population splits in time*

Under a stepwise mutation model and mutation–drift equilibrium, the squared difference in average allele lengths between two populations $(\delta\mu)^2$ is linearly related with the time τ (generations) since population splits:

$$\tau = (\delta\mu)^2/2\beta,$$

where β is the mutation rate (Goldstein *et al.* 1995). By estimating τ one can put demographic processes into an approximate timeframe. One crucial assumption of the model is the absence of significant gene flow after two groups separated, which is unlikely to hold for geographically near populations. For that reason, we

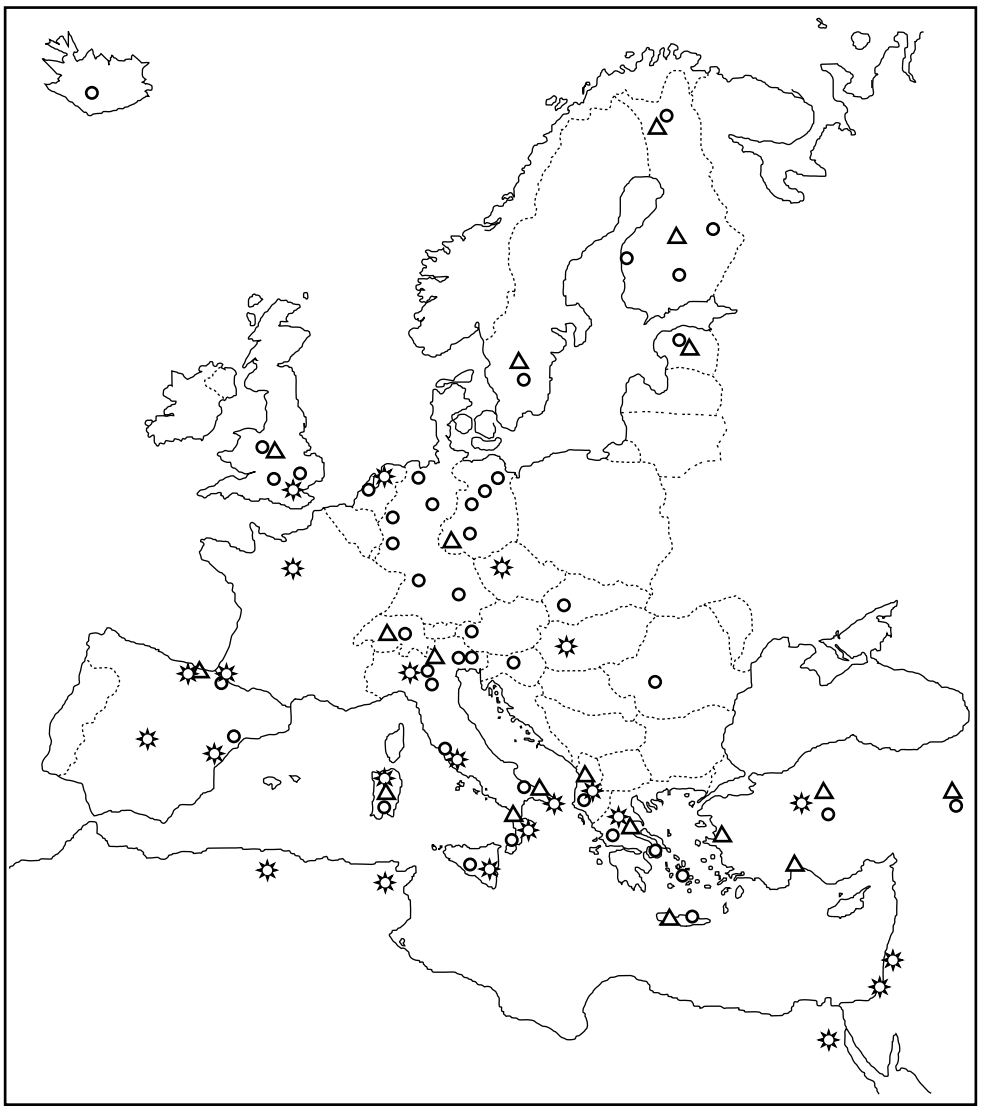

Figure 2. Distribution of population samples. The data on the p12f2 and 49a,f/TaqI RFLP polymorphisms are from Semino *et al.* (1996). Triangles, samples typed for DYS19/YAP; circles, samples typed for DYS19 and other STR loci; asterisks, samples typed for 49a,f/TaqI and 12f2/TaqI RFLPs.

decided to estimate t only between samples separated by large geographical distances, for which the assumption of negligible local gene flow is more robust.

We used a mutation frequency of 2×10^{-3} per locus per generation, estimated from genealogies for tetranucleotide short tandem repeats (STRs) of the Y chromosome (Heyer *et al.* 1997). Because this figure was obtained after excluding the two most variable pedigrees from the analysis, it is probably an underestimate of the actual mutation rate. In addition, trinucleotides mutate at a higher frequency than tetranucleotides (Chakraborty *et al.* 1997), but here the mutation rate 2×10^{-3} was applied to DYS392 as well. Finally, the human generation time is often rounded at 20 years (see e.g. Perez-Lezaun *et al.* 1997*b*), but here we chose to consider each generation to cover 25 years. None of these factors seem likely to affect the estimated τ-values drastically, but the parameters that we chose tend, if anything, to inflate them.

3. RESULTS

Besides the two RFLP markers, the samples considered included 43 STR alleles, 20 of them polymorphic in the sense that their frequency over Europe exceeds 0.05 and 13 DYS19/YAP haplotypes, five of them polymorphic. Rare alleles are poorly informative on spatial processes, since their patterns are deeply affected by the random effects of sampling (Sokal *et al.* 1989) and so only the 27 polymorphic alleles were analysed (tables 1 and 2).

Twenty-two autocorrelation patterns were significant at the 0.05 level. For the biallelic polymorphisms identified by the 49a,f and p12f2 probes, the patterns observed are in agreement with previous findings based on visual inspection of data (Semino *et al.* 1996); the variation is approximately clinal for p12f2 and significantly so for 49a,f. Eleven microsatellite alleles showed a statistically significant clinal distribution, whereas for DYS19/YAP*190 (−) there is evidence of long-distance differentiation, but not a proper gradient of frequencies. At all loci except DYS390 and DYS391, one or more alleles were clinally distributed over Europe. After Bonferroni correction for multiple tests (i.e. after multiplying the highest observed significance for the number of correlograms, 27), the overall autocorrelation was still significant ($p < 0.025$). The binomial probability of observing 22 patterns significant at the 5% level in 27 tests by chance is virtually zero (10^{-24}).

Gradients of molecular diversity were detected by AIDA for all loci except DYS390, where autocorrelation

Table 1. *SA analysis of the frequencies of two biallelic RFLP markers, 20 polymorphic STR alleles and five polymorphic DYS19/YAP haplotypes*

(Patterns: n.s., non-significant; IBD, isolation by distance; C, cline; L, long-distance differentiation; NC, significant, non-clinal. *n*, number of samples considered.)

		distance class limits (km)				overall correlogram	
locus	allele (bp)	1–500	500–1000	1000–2000	> 2000	probability	pattern
p12f2 ($n=20$)	8 kb	0.11	0.07	0.00	-0.46^{**}	< 0.001	L
49a,f ($n=20$)	5 kb	0.40^{**}	0.11	−0.06	-0.48^{**}	< 0.001	C
DYS19 ($n=59$)	186	0.21^{**}	0.18^{**}	−0.04	-0.28^{**}	< 0.001	C
	190	0.29^{**}	0.19^{**}	−0.04	-0.32^{**}	< 0.001	C
	194	0.17^{**}	0.15^{**}	−0.04	-0.22^{**}	< 0.001	C
	198	0.20^{**}	−0.02	−0.04	-0.10^{**}	< 0.001	C
DYS389b ($n=17$)	247	0.32^{*}	0.01	0.03	-0.31^{**}	0.014	C
	251	0.63^{**}	0.57^{**}	−0.15	-0.53^{**}	< 0.001	C
	255	0.49^{*}	0.09	−0.17	−0.21	0.041	IBD
DYS390 ($n=24$)	207	0.10	0.15^{*}	-0.19^{*}	−0.06	0.113	n.s.
	211	0.14	0.10	−0.07	-0.17^{*}	0.116	n.s.
	215	0.31^{*}	0.16^{*}	-0.17^{*}	-0.16^{*}	0.061	n.s.
	219	0.63^{**}	−0.11	-0.34^{**}	0.11^{*}	< 0.001	NC
DYS391 ($n=22$)	283	0.35^{**}	−0.08	-0.24^{*}	−0.01	0.031	NC
	287	0.39^{**}	−0.10	-0.23^{*}	0.00	0.017	NC
DYS392 ($n=18$)	248	0.34	0.19	−0.18	−0.15	0.207	n.s.
	251	0.52^{**}	−0.21	-0.24^{*}	−0.02	0.007	NC
	254	0.50^{**}	0.09	−0.17	−0.15	0.033	IBD
	257	0.68^{**}	0.23^{*}	−0.02	-0.31^{**}	< 0.001	C
DYS393 ($n=20$)	120	0.64^{**}	0.13	-0.22^{*}	-0.29^{**}	< 0.001	C
	124	0.41^{**}	−0.01	−0.14	-0.18^{*}	0.012	C
	128	0.66^{**}	0.24^{**}	−0.13	-0.42^{**}	< 0.001	C
DYS19/YAP ($n=19$)[a]	186 (+)	0.38	0.42^{**}	−0.18	-0.26^{**}	0.002	NC
	190 (−)	0.52	0.24^{*}	0.18^{*}	-0.46^{**}	< 0.001	L
	194 (−)	0.72^{**}	0.07	-0.26^{*}	−0.08	0.013	NC
	194 (+)	0.88^{**}	0.24^{*}	0.02	-0.39^{**}	< 0.001	C
	198 (−)	−0.19	−0.18	0.02	0.03	0.743	n.s.

[a]Presence or absence of Alu insertion in parentheses.

Table 2. *AIDA analysis at six loci and for the DYS19/YAP haplotypes*

(Patterns: C, cline; IBD, isolation by distance. *n*, number of individuals considered.)

	distance class limits (km)					
locus	0	1–500	500–1000	1000–2000	> 2000	pattern
DYS19 ($n=4665$)	0.035^{***}	0.005^{***}	0.000^{*}	-0.006^{***}	-0.004^{***}	C
DYS389b ($n=1338$)	0.072^{***}	0.039^{***}	0.009^{***}	-0.016^{***}	-0.030^{***}	C
DYS390 ($n=1885$)	0.044^{***}	0.010^{***}	-0.003^{***}	-0.010^{***}	-0.005^{***}	IBD
DYS391 ($n=1265$)	0.059^{***}	0.017^{***}	-0.008^{**}	-0.007^{***}	-0.012^{***}	C
DYS392 ($n=1046$)	0.261^{***}	0.114^{***}	0.047^{***}	-0.042^{***}	-0.084^{***}	C
DYS393 ($n=1351$)	0.160^{***}	0.083^{***}	0.025^{***}	-0.029^{***}	-0.094^{***}	C
DYS19/YAP ($n=567$)	0.155^{***}	0.138^{***}	0.022^{***}	0.006^{***}	-0.060^{***}	C

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

was still significantly different from zero at all distance classes, but the negative peak was between 1000 and 2000 km (table 2). At the other five loci and for the DYS19/YAP haplotypes, autocorrelation was positive and significant, not only within samples, but also at the shortest distance classes and the molecular similarity between populations decreased with distance, reaching its highest levels in the last distance class.

The separation times between populations estimated from allele-length differences (Goldstein *et al.* 1995) varied from a maximum of 14 875 years (at the DYS392 locus, between Basques and Italians) to virtually zero (in many cases, the gene pools of near populations never really separated). A few other large values were observed, all of them in comparisons involving either Finns or Basques, i.e. populations speaking non-Indo-European

Table 3. *Maximum effective divergence times (τ_{max}) between Indo-European-speaking populations (years) estimated from ten loci*

autosomic Chikhi *et al.* (1998*a*)[a]	τ_{max}	Y-chromosome this study[b]	τ_{max}
FES/FPS	4511	DYS19	10 296
FXIIIA	6042	DYS389b	281
TH01	8640	DYS390	3996
VWA31A	6325	DYS391	781
DYS392	4600	DYS393	900
mean	6380	mean	3476
mean of ten loci	4637		

[a] $\mu = 2.8 \times 10^{-4}$.
[b] $\mu = 2 \times 10^{-3}$.

languages. The maximum estimated separation between samples speaking a Indo-European language was 10 296 years (at the DYS19 locus, between Slovaks and Catalans).

The differences between loci were, as expected, substantial. Table 3 shows the highest values observed in comparisons between the most differentiated Indo-European-speaking samples (τ_{max}) for each locus, as well as comparable figures for four autosomic tetranucleotide STR loci (Chikhi *et al.* 1998*a*). The average τ_{max} was less than 5000 years.

4. DISCUSSION

A highly significant degree of geographical structuring was evident for several polymorphisms of the Y chromosome. The patterns identified by spatial autocorrelation were mostly clinal and therefore consistent with the effects of a directional population expansion. The fractions of loci showing significant clines seemed higher at the DNA than at the protein level (see Sokal *et al.* 1989; Barbujani *et al.* 1994; Chikhi *et al.* 1998*a*,*b*).

Both main models of the origin of the European gene pool assumed a population expansion from the Near East. On the contrary, a model based on mtDNA data and interpreting genetic diversity as a consequence of late Palaeolithic expansions from glacial refugia (Torroni *et al.* 1998; Sykes 1999) seemed at odds with this and other analyses of nuclear polymorphisms. Indeed, several independent expansions from southern and central Europe are extremely unlikely to have resulted in multilocus gradients encompassing the entire continent. To discriminate between the effects of the initial Palaeolithic colonization and of the Neolithic demic diffusion, one must place the observed pattern into a time-frame. Malaspina *et al.* (1998) analysed a combination of microsatellite, insertion/deletion and restriction polymorphisms of the Y chromosome, observed both clinal and non-clinal patterns in Europe and tentatively attributed the clines to the effects of a Palaeolithic expansion. On the contrary, in this study, the times estimated from microsatellite diversity tended to be short, much shorter in fact than 10 000 years. The statistical error is probably large here; all the figures in table 3 are independent estimates of the deepest split between populations. Approximate though they must be, however, figures between 10 296 and a few hundred years seem unlikely to represent random variates about a real separation time preceding the Neolithic period.

As Weiss (1984) remarked, these are 'effective separation times', that is to say these dates indicate moments at which one could locate the split between populations, had no local gene flow occurred afterwards. Therefore, these figures probably underestimate the depth of the genealogical relationships between populations. However, could continuous gene flow lead to misplacing in Neolithic times a separation that really occurred in the Paleolithic period? On the basis of a birth-and-death process (Slatkin & Rannala 1997), Chikhi *et al.* (1998*a*) showed that, if as many as 15 genes travelled from the Balkans to Iberia in each generation for 40 000 years, the τ-values would decrease by 9% at worst. This means that, even if we correct for extraordinarily high levels of continuous gene flow and increase our estimates by 9%, the average τ_{max} would become 5054 and the highest value observed at a single locus would be 11 223.

Other factors may affect the estimated values of τ. Goldstein *et al.*'s (1995) model assumed mutation–drift equilibrium which means that different alleles are lost by drift in different populations (leading to genetic divergence), but mutation tends to reintroduce them (increasing genetic similarity between samples). In a worldwide study of microsatellite diversity, Perez-Lezaun *et al.* (1997*a*) concluded that, if separation times are short, the effects of drift dominate over those of mutation. If their point is correct, τ should overestimate the time that has elapsed since the European population splits, which happened recently on an evolutionary scale. In synthesis, we do not think these dates should be taken literally, but (i) the calculations based on Chikhi *et al.*'s (1998*a*) model show that, if two gene pools really separated in Palaeolithic times, reasonable levels of successive gene flow cannot dramatically reduce the estimates of τ, and (ii) departures from mutation–drift equilibrium are unlikely to result in an underestimation of τ.

The differences between the dates estimated in various studies may partly depend on their different assumptions. The coalescence time of a genealogy of alleles has been used to approximate the age of the European gene pool in some analyses of mtDNA (Richards *et al.* 1996; Sykes 1999) and Y-chromosome data (Malaspina *et al.* 1998). However, coalescence times are the ages of molecules, which reflect the age of the population only if that population passed through a bottleneck that erased all previously existing diversity. That might have been the case for some geographic or linguistic isolates, such as Finns (Sajantila *et al.* 1996). On the contrary, if a population was founded by a group of genetically differentiated individuals, the coalescence times will consistently overestimate its age (Saitou 1996; Barbujani *et al.* 1998). Can one safely assume that the entire European population evolved from a very limited number of ancestors?

The microsatellite loci of this study are not the best markers for addressing this question, because of their high mutation rate. However, most single-nucleotide polymorphisms (SNPs) are considered to have arisen only once in human evolution (Hammer *et al.* 1998). A strong founder effect at the origin of the European gene pool would then be expected reduce both SNP diversity within

Europe and allele sharing with other continents. In a worldwide survey of Y-chromosome SNPs, Hammer *et al.* (1998) described ten different haplotypes. Four of them are present in the five European populations typed and none of them is restricted to Europe. This result may not be confirmed by further studies of other polymorphisms, but it certainly does not support a strong founder effect at the origin of the European gene pool. As a consequence, we think that one should treat the coalescence times of European allele genealogies only as an upper bound of the populations' age. Palaeolithic coalescence times are fully consistent with a Neolithic subdivision of the European gene pool.

The questions of whether mtDNA really shows geographical patterns that are not typical of most nuclear markers and why still call for an answer. Possible explanations include selection and different migrational behaviours of males and females (Seielstad *et al.* 1998). As for the Y chromosome, populations are being typed for SNPs and it will soon be possible to see whether evolutionarily stabler polymorphisms do or do not confirm the clinal patterns shown in this study by fast-evolving microsatellites. Like any large-scale analysis of population data, this study is necessarily based on less than optimal collections of samples, which may have somewhat affected its results. However, the amount of structuring detected, its statistical significance, the estimated times of population split and previous analyses of autosomic protein and DNA variation (Cavalli-Sforza *et al.* 1994; Chikhi *et al.* 1998*a*,*b*) are in evident agreement with a model of population replacement in the Neolithic, accompanied by separation of what we could call regional gene pools. Alternative models should be considered only if, besides improving our understanding of other aspects of the European population structure, they can also fully account for all these observations.

Many thanks to Jeffrey Long, Rick Kittles, Michele Stenico and Giulietta Di Benedetto for giving us access to their unpublished data. This study was supported by grants from the Italian Ministry of University (COFIN 97) and the Universities of Ferrara and Bologna.

REFERENCES

Ammerman, A. J. & Cavalli-Sforza, L. L. 1984 *The Neolithic transition and the genetics of populations in Europe*. Princeton University Press.

Barbujani, G., Pilastro, A., DeDomenico, S. & Renfrew, C. 1994 Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonisation. *Am. J. Phys. Anthropol.* **95**, 137–154.

Barbujani, G., Sokal, R. R. & Oden, N. L. 1995 Indo-European origins: a computer-simulation test of five hypotheses. *Am. J. Phys. Anthropol.* **96**, 109–132.

Barbujani, G., Bertorelle, G. & Chikhi, L. 1998 Evidence for Paleolithic and Neolithic gene flow in Europe. *Am. J. Hum. Genet.* **62**, 488–491.

Bertorelle, G. & Barbujani, G. 1995 Analysis of DNA diversity by spatial autocorrelation. *Genetics* **139**, 811–819.

Bertranpetit, J., Calafell, F., Comas, D., Perez-Lezaun, A. & Mateu, E. 1996 Mitochondrial DNA sequences in Europe: an insight into population structure. In *Molecular biology and human diversity* (ed. A. J. Boyce & C. G. N. Mscie-Taylor), pp. 112–129. Cambridge University Press.

Birdsell, J. B. 1968 Some predictions for the Pleistocene based on equilibrium systems among recent hunter–gatherers. In *Man the hunter* (ed. R. Lee & I. De Vore), pp. 229–249. Hawthorne, NY: Aldine.

Cavalli-Sforza, L. L. & Minch, E. 1997 Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **61**, 247–251.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994 *The history and geography of human genes*. Princeton University Press.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. & Deka, R. 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl Acad. Sci. USA* **94**, 1041–1046.

Chikhi, L., Destro-Bisol, G., Bertorelle, G., Pascali, V. & Barbujani, G. 1998*a* 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., Destro-Bisol, G., Pascali, V., Baravelli, V., Dobosz, M. & Barbujani, G. 1998*b* Clinal variation in the nuclear DNA of Europeans. *Hum. Biol.* **70**, 643–657.

Goldstein, D. B., Ruiz-Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995 An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463–471.

Hammer, M. F. 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.

Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C., Templeton, A. R. & Zegura, S. L. 1998 Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* **15**, 427–441.

Harlan, J. R. 1971 Agricultural origins: centers and noncenters. *Science* **174**, 468–474.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. & DeKnijff, P. 1997 Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* **6**, 799–803.

Jobling, M. A. & Tyler-Smith, C. 1995 Fathers and sons: the Y chromosome and human evolution. *Trends Genet.* **11**, 449–456.

Malaspina, P. (and 23 others) 1998 Network analyses of Y-chromosomal types in Europe, Northern Africa, and Western Asia reveal specific patterns of geographic distribution. *Am. J. Hum. Genet.* **63**, 847–860.

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

Perez-Lezaun, A., Calafell, F., Mateu, E., Comas, D., Ruiz-Pacheco, R. & Bertranpetit, J. 1997*a* Microsatellite variation and the differentiation of modern humans. *Hum. Genet.* **99**, 1–7.

Perez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E. & Bertranpetit, J. 1997*b* Population genetics of Y-chromosome short tandem repeats in humans. *J. Mol. Evol.* **45**, 265–270.

Rendine, S., Piazza, A. & Cavalli-Sforza, L. L. 1986 Simulation and separation by principal components of multiple demic expansions in Europe. *Am. Nat.* **128**, 681–706.

Renfrew, C. 1987 *Archaeology and language. The puzzle of Indo-European origins*. London: Jonathan Cape.

Richards, M., Corte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H. J. & Sykes, B. 1996 Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **58**, 185–203.

Richards, M., Macaulay, V., Sykes, B., Pettit, P., Forster, P., Hedges, R. & Bandelt, H. J. 1997 Palaeolithic and Neolithic lineages in the European mitochondrial gene pool: a response to Cavalli-Sforza and Minch. *Am. J. Hum. Genet.* **61**, 251–254.

Richards, M., Macaulay, V., Bandelt, H. J. & Sykes, B. 1998 Phylogeography of mitochondrial DNA in Western Europe. *A. Hum. Genet.* **62**, 241–260.

Saitou, N. 1996 Contrasting gene trees and population trees of the evolution of modern humans. In *Molecular biology and human diversity* (ed. A. J. Boyce & C.G.N. Mascie-Taylor), pp. 265–282. Cambridge University Press.

Sajantila, A., Salem, A. H., Savolainen, P., Bauer, K., Gierig, C. & Pääbo, S. 1996 Paternal and maternal lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Natl Acad. Sci. USA* **93**, 12 035–12 039.

Seielstad, M., Minch, E. & Cavalli-Sforza, L. L. 1998 Genetic evidence for a higher female migration rate in humans. *Nature Genet.* **20**, 278–280.

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

Slatkin, M. & Rannala, B. 1997 Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.* **60**, 447–458.

Sokal, R. R. & Oden, N. L. 1978 Spatial autocorrelation in biology. *Biol. J. Linn. Soc.* **10**, 199–249.

Sokal, R. R., Harding. R. M. & Oden, N. L. 1989 Spatial patterns of human gene frequencies in Europe. *Am. J. Phys. Anthropol.* **80**, 267–294.

Sokal, R. R., Oden, N. L. & Wilson, C. 1991 New genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* **351**, 143–145.

Sokal, R. R., Jacquez, G. M., Oden, N. L., DiGiovanni, D., Falsetti, A. B., McGee, E. & Thomson, B. A. 1993 Genetic relationships of European populations reflect their ethnohistorical affinities. *Am. J. Phys. Anthropol.* **91**, 55–70.

Sykes, B. 1999. The molecular genetics of European ancestry. *Phil. Trans. R. Soc. Lond.* B **354**, 131–139.

Torroni, A. (and 10 others) 1998 mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am. J. Hum. Genet.* **62**, 1137–1152.

Weiss, K. M. 1984 On the number of members of the genus *Homo* who have ever lived, and some evolutionary implications. *Hum. Biol.* **56**, 637–649.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.