Ann. Hum. Genet. (2000), **64**, 395–412
Printed in Great Britain

Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa

P. MALASPINA[1], F. CRUCIANI[2], P. SANTOLAMAZZA[2], A. TORRONI[2,3], A. PANGRAZIO[2], N. AKAR[4], V. BAKALLI[5], R. BRDICKA[6], J. JARUZELSKA[7], A. KOZLOV[8], B. MALYARCHUK[9], S. Q. MEHDI[10], E. MICHALODIMITRAKIS[11], L. VARESI[12], M. M. MEMMI[12], G. VONA[13], R. VILLEMS[14], J. PARIK[14], V. ROMANO[15], M. STEFAN[16], M. STENICO[17], L. TERRENATO[1], A. NOVELLETTO[1,18] AND R. SCOZZARI[2]

[1] *Department of Biology, University of Rome 'Tor Vergata', Italy*
[2] *Department of Genetics and Molecular Biology, University of Rome 'La Sapienza', Italy*
[3] *Institute of Biochemistry, University of Urbino, Italy*
[4] *Pediatrics Department, Ankara University, Turkey*
[5] *Haematology Unit, Girokaster Hospital, Albania*
[6] *Institute for Hematology and Blood Transfusion, Prague, Czech Republic*
[7] *Polish Academy of Sciences, Poznan, Poland*
[8] *Arct-An C Laboratory, Moscow, Russian Federation*
[9] *I.B.P.N., Magadan, Russian Federation*
[10] *A.Q. Khan Research Laboratory, Islamabad, Pakistan*
[11] *Department of Forensic Sciences, University of Crete, Heraklion, Greece*
[12] *Faculty of Sciences and Technics, University of Corte, France*
[13] *Department of Experimental Biology, Anthropology Section, University of Cagliari, Italy*
[14] *Department of Evolutionary Biology, Tartu University and Estonian Biocentre, Tartu, Estonia*
[15] *Department of Biopathology and Biomedical Methodology, University of Palermo, Italy*
[16] *Genetics Department, University of Bucharest, Rumania*
[17] *Department of Biology, University of Ferrara, Italy*
[18] *Department of Cell Biology, University of Calabria, Italy*

(Received 2.5.00. Accepted 19.7.00)

SUMMARY

We typed 1801 males from 55 locations for the Y-specific binary markers YAP, DYZ3, SRY_{10831} and the (CA)n microsatellites YCAII and DYS413. Phylogenetic relationships of chromosomes with the same binary haplotype were condensed in seven large one-step networks, which accounted for 95% of all chromosomes. Their coalescence ages were estimated based on microsatellite diversity. The three largest and oldest networks undergo sharp frequency changes in three areas. The more recent network 3.1A clearly discriminates between Western and Eastern European populations. Pairwise Fst showed an overall increment with increasing geographic distance but with a slope greatly reduced when compared to previous reports. By sectioning the entire data set according to geographic and linguistic criteria, we found higher Fst-on-distance slopes within Europe than in West Asia or across the two continents.

INTRODUCTION

Genetic diversity markers of the human Y chromosome are gaining an ever increasing value in understanding human microevolutionary processes. Its uniparental inheritance renders the non-recombinant portion of this chromosome (NRPY) a unique tool to describe its phylogeny and to make inferences on its diversity assuming the concept of a progressive accumulation of

Correspondence: Dr Patrizia Malaspina, Department of Biology, University of Rome 'Tor Vergata', Via della Ricerca Scientifica 00133 Rome, Italy. Tel: +39-06-72594321; Fax: +39-06-2023500.
E-mail: patrizia.malaspina@uniroma2.it

mutations over time. In this context, the NRPY can be regarded as the male counterpart of the mtDNA. However, a higher rate of divergence among populations for this chromosome with respect to mtDNA seems to be attributable to the difference in the female vs. male migration rate (Seielstad *et al*. 1998). This results in a higher level of population structuring which may be more pronounced in some continents and is often significantly detectable in neighbouring populations (Scozzari *et al*. 1997; Karafet *et al*. 1998).

Different classes of markers are able to reveal different subsets of the total variation, this ability being a function of the locus-specific mutation rate (Kayser *et al*. 2000). Binary polymorphisms, where mutational events are considered rare or even unique, have been used to unequivocally reconstruct phylogenies of Y chromosomes sampled in all continents (Altheide & Hammer, 1997; Hammer *et al*. 1997, 1998; Underhill *et al*. 1997). Such reconstruction of phylogenies can be further improved by a description of the geographical distribution of each lineage which, in turn, allows one to make inferences about the major male-specific population movements in the past, up to several tens of thousands of years ago (Hammer *et al*. 1998; Karafet *et al*. 1999; Santos *et al*. 1999; Hill *et al*. 2000). The joint use of binary markers with highly mutable mini- and micro-satellites becomes a powerful tool in analysing recent lineages and in defining population processes on smaller geographical and time scales (Zerjal *et al*. 1997; Hurles *et al*. 1998; Jobling *et al*. 1998; Kittles *et al*. 1998; Thomas *et al*. 1998; Scozzari *et al*. 1999).

Several papers have shown that the patterns of microsatellite length variation closely approach the expectations of the stepwise model which assumes mutational events that involve a single repeat unit (Di Rienzo *et al*. 1994, 1998; Cooper *et al*. 1996). Coalescence estimates derived from microsatellite variance are highly valuable since they allow one, in principle, to date the antiquity of lineages; these methods can in fact complement those which are sequence-based (Goldstein *et al*. 1995, 1996; Slatkin & Rannala, 1997).

The age of populations carrying a repertoire of lineages must keep within the time limits estimated in this way. However, very wide margins of uncertainty are present. As a matter of fact, an appropriate sample coverage over a vast geographical area is a necessary requirement, not often fulfilled, to attain an exhaustive description of the variation within each lineage.

Uncommon microsatellite multirepeat mutational events which could be regarded as a disturbance in the regular process of accumulation of variation are instead an important resource, since they may identify groups of chromosomes with a common ancestry (Forster *et al*. 1998) that cannot be otherwise recognized by binary markers.

We have previously defined six major one-step haplotype networks by combining binary and microsatellite markers and described their frequency in 33 populations from Europe, North Africa and West Asia (Malaspina *et al*. 1998). While a remarkable degree of geographic specificity emerged for all of the networks, it could be argued that the overall result may have been conditioned by an uneven coverage (for a discussion see Thomas *et al*. 2000), with an overrepresentation of Western European countries.

In the present paper, not only have we adopted the same analysis, but we have doubled the overall sample size by increasing the number of sampled locations to 55 (by adding populations mostly from Northern and Central Europe and West Asia) and added a third binary marker.

Here, we also investigated the patterns of population divergence across the surveyed area by introducing a spatial analysis of Fst. For the first time, we show that the pattern of such variation for the Y chromosome is anything but uniform and varies greatly depending upon the geographical areas where the populations are living and on the linguistic family they belong to.

MATERIALS AND METHODS

Subjects

1801 males were sampled in 55 locations (Table 1, Fig. 1A). Thirty-two of these samples, previously described by Malaspina *et al.* (1998), were further typed in this study for SRY_{10831} (see below). The previous mixed Turkish sample (Scozzari *et al.* 1997; Malaspina *et al.* 1998) was replaced by geographically well-characterized groups. The previous Sicilian sample (Malaspina *et al.* 1998) was here assigned to West Sicily, and increased by 44 subjects.

DNA was prepared by standard techniques from either fresh venous blood, dried blood absorbed on filter paper, or hair roots. Immortalized lymphoblastoid cell lines were used for the Pakistani samples.

Markers

Three binary polymorphisms and two complex dinucleotide polymorphic systems were studied. The presence/absence of the YAP element was assayed by PCR as described (Hammer & Horai, 1995). The presence/absence of the alphoid *Hin*dIII site (DYZ3) was tested by PCR followed by digestion (Santos *et al.* 1995). The A/G base substitution at position 10831 of the SRY sequence (Whitfield *et al.* 1995) (SRY_{10831} or SRY-1532) was assayed by either one of the two previously described independent methods (Santos *et al.* 1999; Scozzari *et al.* 1999). The YCAII and DYS413 polymorphic systems were analysed according to Mathias *et al.* (1994) and Malaspina *et al.* (1997). These systems consist of two Y-specific loci each, both containing a (CA)n microsatellite, which are co-amplified during the corresponding PCR reactions. The larger and smaller PCR fragments generated for each system were assigned to the allelic classes a and b, respectively. Whenever a single band was observed, two fragments of the same size were assumed (Mathias *et al.* 1994).

All DNA polymorphisms were assayed on all individuals except SRY_{10831}. The latter was tested on all subjects lacking the alphoid *Hin*dIII site and on 485 subjects with the alphoid *Hin*dIII site.

A complete listing of the data is available at URL www.unical.it/dipartimenti/biologia/genetica.html.

Statistical analyses

One-step networks of adjacent microsatellite haplotypes on chromosomes with the same allelic states for binary polymorphisms were constructed as described (Cooper *et al.* 1996; Malaspina *et al.* 1998). Briefly, each network groups 'adjacent' haplotypes, i.e. haplotypes that differ for the insertion or deletion of a single CA unit at a single locus.

Estimates of coalescence times (t) for each network were obtained by two methods, based on the stepwise mutation model. The first method relies on average squared distance (ASD), a parameter linearly related to coalescence time (Goldstein *et al.* 1995; Slatkin, 1995). We calculated the squared difference (in CA units) for every microsatellite allelic series, between the value of each individual and the value found in the most frequent haplotype of the corresponding network (see Table 2, col. 3). The average values for chromosomes belonging to the same network were then averaged over the four microsatellite allelic series and then divided by the mutation rate (μ). The second method assumes a star-shaped genealogy characteristic of rapid population growth. In these conditions $V = \mu t$ where V is the average variance of microsatellite repeat counts (Thomas *et al.* 2000). It is to be observed that the results of the two methods converge when the ancestral haplotype carries the allele with the average CA size at each locus.

In both methods the value of 5.6×10^{-4} was used as the mutation rate (Weber & Wong, 1993), an intermediate value between that of Gyapay *et al.* (1994) and that of Kayser *et al.* (2000) for dinucleotide microsatellites.

Confidence intervals for both methods were obtained by bootstrapping, by performing two hundred random samplings from the entire data

Table 1. *The 55 population samples included in this study. Relative frequencies of the seven largest networks, eight minor networks (pooled) and haplotypes not joined to any network*

Region	Population	Linguistic family[a, b]	Sample size (n)	Larger networks							Minor networks	Unclassif. haplotypes
				1.1	1.2	1.3	2.1	3.1G	3.1A	1.4		
Northern Europe	Norwegian[c]	I.E.	8	0.25	0.00	0.00	0.00	0.13	0.50	0.00	0.00	0.13
	Lithuanian[c]	I.E.	15	0.40	0.00	0.00	0.07	0.00	0.40	0.00	0.07	0.07
	Estonian	Uralic	74	0.54	0.01	0.00	0.04	0.03	0.36	0.00	0.00	0.01
	Komi-Permiak	Uralic	31	0.65	0.06	0.00	0.03	0.06	0.19	0.00	0.00	0.00
	Danish[c]	I.E.	35	0.23	0.09	0.00	0.03	0.54	0.06	0.00	0.03	0.03
	Mordovian	Uralic	62	0.40	0.05	0.00	0.02	0.06	0.39	0.00	0.05	0.03
Great Britain	Londoners[c]	I.E.	20	0.30	0.00	0.00	0.00	0.65	0.05	0.00	0.00	0.00
Iberian peninsula	Northern Portuguese[c]	I.E.	26	0.19	0.00	0.00	0.08	0.58	0.00	0.00	0.00	0.15
	Southern Portuguese[c]	I.E.	26	0.31	0.04	0.00	0.08	0.46	0.00	0.00	0.00	0.12
	Central Spaniard[c]	I.E.	22	0.23	0.14	0.09	0.09	0.45	0.00	0.00	0.00	0.00
	Basque[c]	Basque	28	0.04	0.04	0.00	0.00	0.93	0.00	0.00	0.00	0.00
	Southern Spaniard[c]	I.E.	62	0.19	0.00	0.00	0.05	0.68	0.02	0.02	0.02	0.03
Italian peninsula	Ligurian[c]	I.E.	20	0.15	0.05	0.05	0.25	0.50	0.00	0.00	0.00	0.00
	Trentine	I.E.	30	0.20	0.03	0.00	0.00	0.67	0.07	0.00	0.00	0.03
	Venetian[c]	I.E.	20	0.15	0.20	0.00	0.10	0.45	0.10	0.00	0.00	0.00
	Latium[c]	I.E.	76	0.36	0.12	0.03	0.13	0.29	0.03	0.01	0.00	0.04
	Apulian[c]	I.E.	20	0.20	0.20	0.00	0.20	0.30	0.10	0.00	0.00	0.00
	Calabrian[c]	I.E.	28	0.21	0.25	0.00	0.11	0.32	0.07	0.00	0.00	0.04
	Lucanian[c]	I.E.	24	0.17	0.13	0.04	0.25	0.29	0.00	0.08	0.00	0.04
Sicily	Western Sicilian[c]	I.E.	65	0.12	0.15	0.00	0.18	0.43	0.03	0.03	0.02	0.03
	North-East Sicilian	I.E.	46	0.22	0.28	0.00	0.26	0.15	0.04	0.00	0.00	0.04
Sardinia	Northern Sardinian[c]	I.E.	189	0.24	0.05	0.37	0.11	0.20	0.01	0.01	0.00	0.02
	Southern Sardinian[c]	I.E.	29	0.17	0.14	0.41	0.03	0.24	0.00	0.00	0.00	0.00
Corsica	Corsican	I.E.	90	0.24	0.02	0.02	0.04	0.59	0.00	0.04	0.01	0.02
Central Europe	French	I.E.	26	0.19	0.00	0.04	0.04	0.62	0.08	0.00	0.00	0.04
	Slovakian[c]	I.E.	23	0.43	0.00	0.00	0.09	0.00	0.39	0.00	0.00	0.09
	Northern Rumanian[c]	I.E.	27	0.26	0.26	0.00	0.04	0.30	0.15	0.00	0.00	0.00
	Eastern Rumanian[c]	I.E.	18	0.39	0.06	0.00	0.11	0.11	0.28	0.00	0.00	0.06
	Western Rumanian	I.E.	22	0.36	0.09	0.00	0.05	0.23	0.18	0.00	0.00	0.09
	Ukrainian	I.E.	6	0.33	0.00	0.00	0.17	0.00	0.50	0.00	0.00	0.00
	Eastern Czech	I.E.	40	0.20	0.03	0.00	0.05	0.30	0.40	0.00	0.00	0.03
	Polish	I.E.	36	0.31	0.06	0.00	0.03	0.19	0.39	0.00	0.03	0.00
South-East Europe	Albanian[c]	I.E.	33	0.24	0.06	0.00	0.33	0.24	0.12	0.00	0.00	0.00
	Continental Greek[c]	I.E.	28	0.36	0.11	0.00	0.25	0.07	0.21	0.00	0.00	0.00

Crete island	Cretean[c]	I.E.	83	0.25	0.31	0.00	0.14	0.16	0.05	0.01	0.02	0.05
West Asia	North-East Turkish	Altaic	11	0.55	0.09	0.00	0.00	0.09	0.18	0.00	0.09	0.00
	Central Anatolian	Altaic	18	0.33	0.22	0.00	0.06	0.06	0.11	0.00	0.11	0.11
	South-West Turkish	Altaic	29	0.38	0.31	0.00	0.10	0.03	0.10	0.03	0.00	0.03
	South-East Turkish	Altaic	13	0.46	0.23	0.00	0.00	0.08	0.15	0.00	0.00	0.08
	Turkish Cypriot	Altaic	22	0.18	0.27	0.00	0.23	0.05	0.14	0.00	0.05	0.09
	Omani[c]	A.A.	13	0.38	0.08	0.00	0.08	0.08	0.08	0.00	0.00	0.31
	United Arab Emirate[c]	A.A.	35	0.51	0.06	0.00	0.14	0.06	0.17	0.00	0.03	0.03
	Iranian[c]	I.E.	7	0.29	0.14	0.00	0.14	0.14	0.14	0.00	0.14	0.00
	Pathan[c]	I.E.	22	0.50	0.00	0.00	0.09	0.18	0.23	0.00	0.00	0.00
	Sindhi[c]	I.E.	20	0.30	0.05	0.00	0.05	0.15	0.35	0.00	0.05	0.05
	Baluchi	I.E.	27	0.41	0.07	0.00	0.07	0.19	0.19	0.00	0.00	0.07
	Brahui	Dravidian	15	0.20	0.27	0.00	0.00	0.07	0.33	0.00	0.07	0.07
	Burusho	Burushaski[d]	19	0.21	0.05	0.00	0.00	0.16	0.53	0.00	0.05	0.00
	Hazara	I.E.	13	0.31	0.08	0.00	0.00	0.62	0.00	0.00	0.00	0.00
	Kalash	I.E.	20	0.80	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
	Punjabi	I.E.	13	0.23	0.00	0.00	0.00	0.15	0.46	0.00	0.08	0.08
	Makrani	I.E.	14	0.29	0.07	0.00	0.00	0.21	0.21	0.00	0.07	0.14
North Africa (NA)	Moroccan Arabs[e]	A.A.	56	0.27	0.00	0.00	0.70	0.02	0.00	0.00	0.00	0.02
	Northern Egyptian[c]	A.A.	24	0.13	0.13	0.00	0.58	0.13	0.00	0.04	0.00	0.00
	Southern Egyptian[c]	A.A.	22	0.32	0.05	0.00	0.23	0.09	0.05	0.05	0.00	0.23
Total			1801									

[a] From Grimes (1996).
[b] A.A., Afro-Asiatic.
I.E., Indo-European.
[c] Sample described in Malaspina *et al.* (1998).
[d] Linguistic isolate.
[e] Sample described in Scozzari *et al.* (1999).

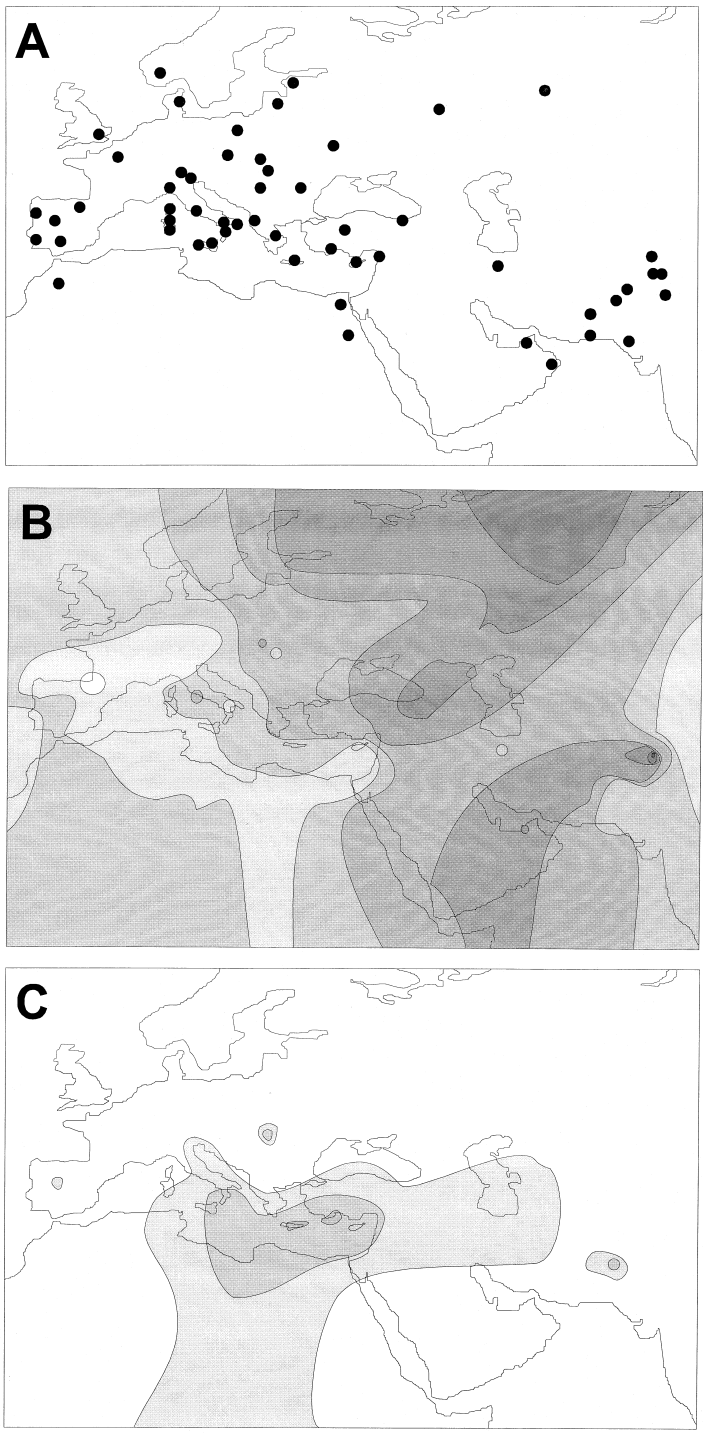


Fig. 1. For legend see opposite.

set, each consisting of 50% of the subjects. The 2.5 and 97.5 percentiles of the resulting distributions were obtained. This method gives a measure of the confidence intervals related to the representation of individuals within the genealogical group and not a measure of the confidence on the accuracy of the age estimate from the accumulated mutations. It should be appreciated that the uncertainty in the mutation rate, in the shape of the genealogy, as well as in the mutation process, significantly broadens these confidence intervals.

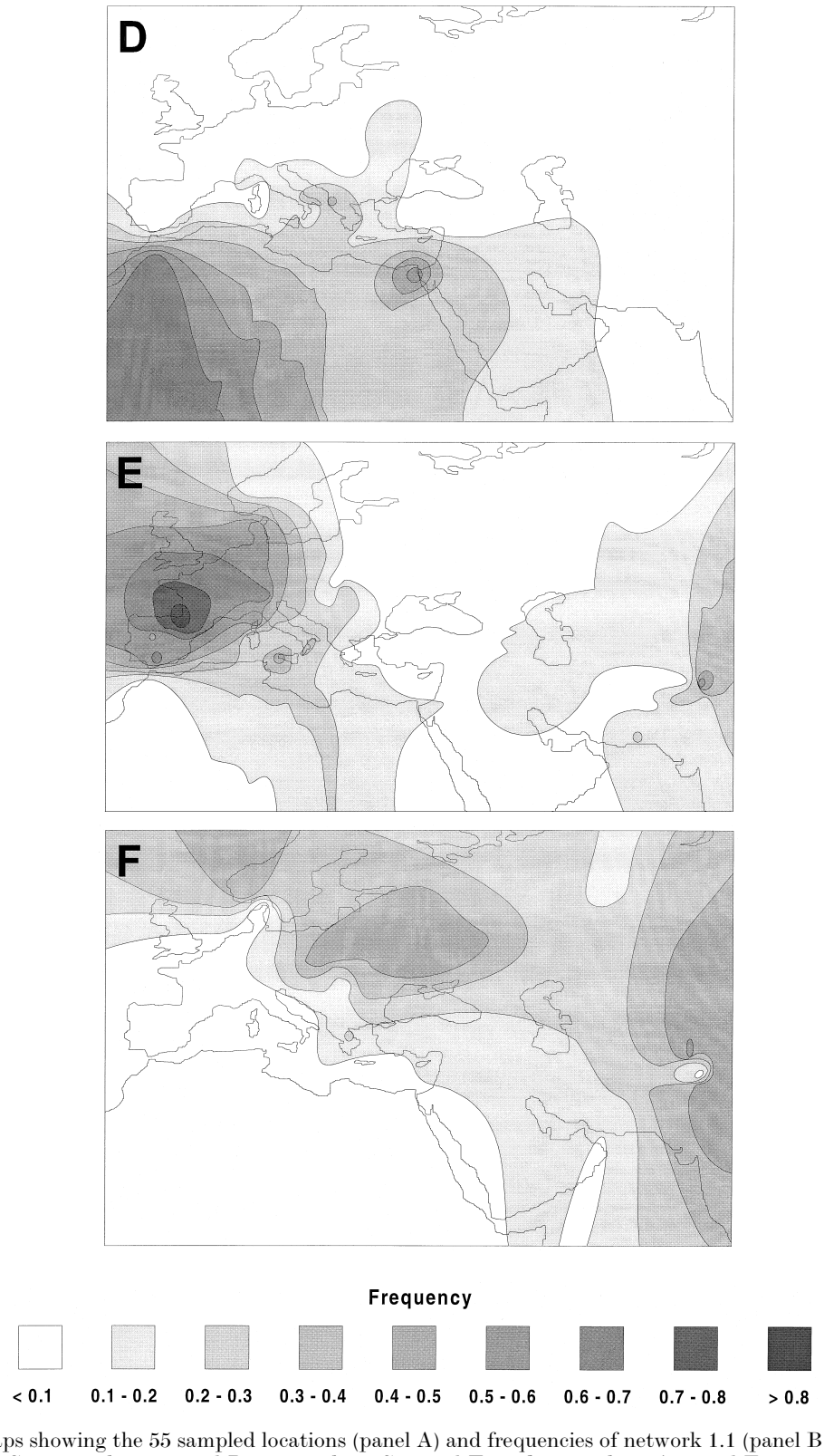


Fig. 1. Maps showing the 55 sampled locations (panel A) and frequencies of network 1.1 (panel B), network 1.2 (panel C), network 2.1 (panel D), network 3.1G (panel E) and network 3.1A (panel F).

Table 2. *Main features of 15 one-step networks of Y chromosome haplotypes*

Network	N of subjects (% of the entire study)	Major haplotype in the network (CA units)[b]	*n* of carriers	Variance of CA units within each network YCAIIa	YCAIIb	DYS413a	DYS413b	Estimated coalescence age in generations[a] (95% C.I.) ASD method	CA variance method
1.1	523 (29.0)	22-22-22-21	44	0.90	1.49	1.26	0.86	5292 (5036–5599)	2013 (1884–2151)
1.2	166 (9.2)	22-19-17-17	91	0.42	0.32	0.19	0.16	500 (356–620)	482 (348–598)
1.3	91 (5.1)	21-11-21-21	72	0.14	0.00	0.10	0.13	172 (96–256)	165 (89–245)
2.1	216 (12.0)	21-19-24-23	38	0.32	0.12	0.86	1.09	1961 (1744–2200)	1066 (956–1179)
3.1G	480 (26.7)	23-19-23-23	253	0.76	0.17	0.37	0.53	932 (783–1076)	817 (697–938)
3.1A	222 (12.3)	23-19-22-22	136	0.52	0.01	0.08	0.34	475 (318–653)	424 (286–571)
1.4	15 (0.8)	24-23-20-20	5	0.86	0.31	0.00	0.00	506 (223–765)	522 (223–856)
8 minor networks	22 (1.2)								
Unclassified	66 (3.7)								
Total	1801								

[a] Assuming a mutation rate of 5.6×10^{-4} (Weber & Wong, 1993).
[b] Reported as YCAIIa-YCAIIb-DYS413a-DYS413b.

The overall network frequencies in the 55 populations (Table 1) were used to construct the geographical maps and to compute pairwise Fst values. Frequency maps were drawn with the Surfer System v. 4.15 (Golden Software Inc.) using the Kriging procedure (Delfiner, 1976). We used an 84×43 grid and estimates at each grid node were obtained considering a maximum of 10 nearest points in each quadrant. In order to represent metric distances among locations, a transformation of the actual longitude was adopted ($\text{Long}' = \text{Long} \times \cos(\text{Lat})$). With this and other methods isophlets are highly sensitive to the frequencies of the points closest to them. As a matter of fact, isophlets in North Africa may be more conspicuously affected by frequencies from Southern European locations than by the three African locations. The peculiarity of this method is that the estimated values of the variable coincide with the observed values at the sampled locations. This is therefore the best method to group populations with similar network frequencies into a definite number of belts. We chose to show isophlets over the bodies of water in order to better show the frequencies obtained in the Mediterranean islands. The change of frequencies over the seas thus merely reflects differences in frequencies among coastal populations.

The total Fst, and the off-diagonal 55×54 matrix of pairwise Fst values, were obtained by using the Arlequin package v. 1.1 (Schneider *et al.* 1997). The entire set of data was analysed by selecting values corresponding to pairwise comparisons that satisfied the geographical and linguistic criteria as described in the Results section (Table 3, col. 1). Fst-on-distance slopes, intercepts and their standard errors were obtained by ordinary linear regression analysis. In order to take into account the internal correlation in the matrices of pairwise Fst and distance measures, significance of Fst-on-distance regression was assayed by the Mantel test (Sokal & Rohlf, 1995) with 10000 iterations, when applicable. All calculations were performed with SPSS version 6.1.3.

RESULTS

By combining the results of YAP and alphoid *Hin*dIII markers, only two out of 1801 subjects turned out to carry the YAP element and to lack the alphoid *Hin*dIII site. This can be explained through a low rate of recurrence of the loss of alphoid units containing the *Hin*dIII site (Santos *et al.* 1996; Scozzari *et al.* 1999). The three combinations YAP−/*Hin*dIII+, YAP+/*Hin*dIII+ and YAP−/HindIII− are referred to as frames 1, 2 and 3, respectively (Persichetti *et al.* 1992); overall, 824, 233 and 742 chromosomes were in turn assigned to these frames (Table 1).

While Hammer *et al.* (1998) demonstrated that both an ancient A to G transition and a more recent G to A reversion occurred at nucleotide position SRY_{10831} in the Y chromosome phylogeny, other authors (Kwok *et al.* 1996; Santos *et al.* 1999) concluded that the G to A reversion occurred on chromosomes lacking the alphoid *Hin*dIII site. Therefore, we performed the typing of SRY_{10831} on all of the 742 frame 3 chromosomes, resulting in two subsets (505 and 237) of chromosomes carrying G and A alleles, respectively. We also tested 385 frame 1 chromosomes and 100 frame 2 chromosomes, all of which carried SRY_{10831}(G).

We analysed the variation of microsatellite markers in each of the three frames. The search of all adjacent relationships (Cooper *et al.* 1996) among microsatellite haplotypes resulted in 7 large and 8 very small (no more than 3 haplotypes and 4 subjects each) networks, which group 96% of all chromosomes (Table 2). Doubling the sample size with respect to our previous work resulted in a modest increase of novel haplotypes within each network. It is worth pointing out that we did not find haplotypes joining the previously identified major networks, indicating that the molecular definition of each network was not the result of incomplete sampling but rather reflects an actual discontinuity in haplotype composition in terms of number of CA repeats across the four microsatellite loci. Moreover, increasing the number of

sampled locations and expanding the surveyed area did not reveal populations harbouring previously hidden large quotas of the Y chromosome variation detectable with the markers here used.

Networks 1.2 and 1.3 retain their major haplotype, i.e. a haplotype whose frequency is much higher than any other haplotype of its own network. Network 1.1 does not show such a prevalence of a single haplotype since there are three haplotypes, each encompassing 30 or more subjects. Network 1.4, consisting of only 15 chromosomes, has the major haplotype 24–23–20–20 (CA units at YCAIIa-YCAIIb-DYS413a-DYS413b) found in 5 subjects. In network 2.1 the major haplotype is 21-19-24-23 (Table 2, col. 3) whereas in the smaller data set (Malaspina *et al*. 1998) it was 22-19-22-21 which now encompasses 30 subjects. This discrepancy exemplifies the effect of population structuring as in the case of haplotype 22-19-22-21 found at high frequencies among Moroccan Arabs, which is to be considered over-represented in the previous data set.

As far as frame 3 is concerned, network 3.1 was split into 3.1G and 3.1A according to the SRY_{10831} results. Network 3.1G includes a greater amount of microsatellite variation according to the number of haplotypes and CA variance (Table 2 cols. 5–8). Some microsatellite haplotypes, identical in state, are found in both networks, although the frequency of their major haplotype differentiated them (Table 2, col. 3). As a matter of fact, major haplotype 23-19-23-23 for network 3.1G accounts for only 2.3% of network 3.1A. Conversely, major haplotype 23-19-22-22 for network 3.1A accounts for only 1.9% in network 3.1G.

Geographic distribution of networks

The overall network frequencies (Table 1) were used to construct the five maps shown in Figure 1.

Network 1.1 (Fig. 1B) reaches its maximum frequency in the small sample of Kalash from Pakistan, represented mainly by the unusual haplotype 21-19-20-20. In Europe, a peak in the North-East (0.65) is shown, with a decreasing cline towards the South-West. The minimum incidence is observed among Basques (0.04).

Network 1.2 (Fig. 1C) is confirmed to be present mainly in Mediterranean populations. It reaches frequencies exceeding 0.30 only in Crete and South-Western Turkey. From these two areas, decreasing frequencies were found in all directions except for Northern Rumania (0.26 vs. < 0.10 in 4 neighbouring populations) and Central Spain (0.14 vs. < 0.10 in 4 neighbouring populations).

Network 1.3 (map not shown) identifies a group of chromosomes peculiar to the Sardinian population (Ciminelli *et al*. 1995; Quintana-Murci *et al*. 1999*b*). This network reaches frequencies of 0.41 and 0.37 in Southern and Northern Sardinian locations, respectively, dropping to 0.02 within the short geographical distance which separates Sardinia from Corsica.

Network 2.1 (Fig. 1D) shows the highest frequencies in Northern Africa, decreasing along the South–North direction. The change is particularly sharp at the Strait of Gibraltar, with frequencies dropping from 0.70 to < 0.10 in all of the Iberian peninsula and to 0.0 in Basques.

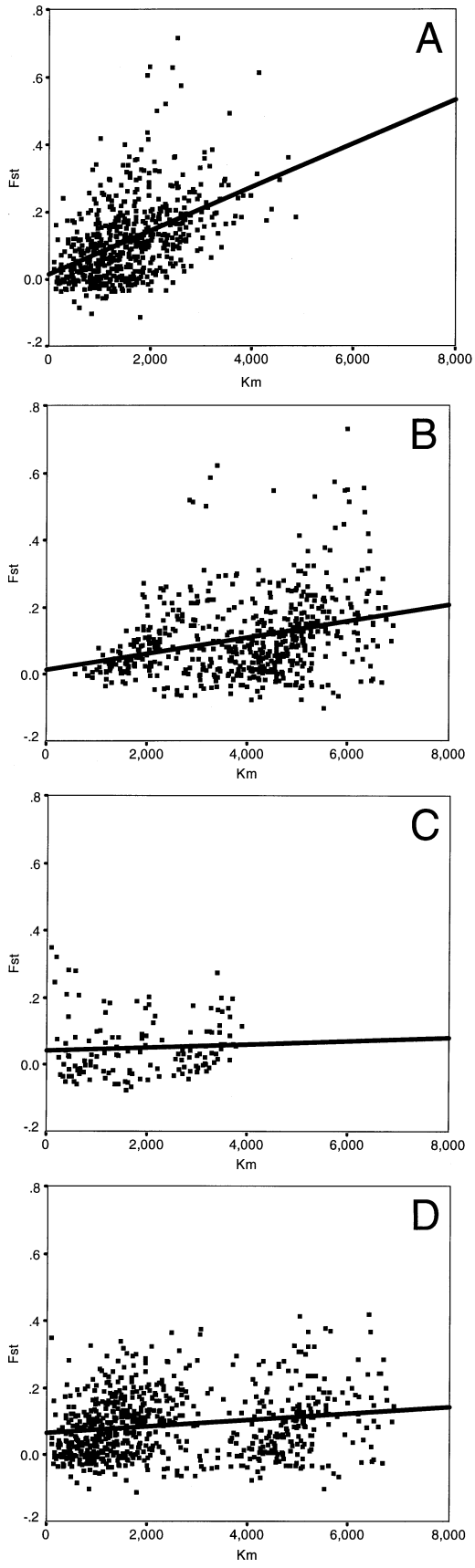
Network 3.1G (Fig. 1E) is detected in Europe, in a very particular portion of the total area previously described for frame 3 chromosomes (Malaspina *et al*. 1998). In Basques this network reaches a frequency of 0.93, 73% of which is represented by haplotype 23-19-23-23. Frequencies $\geqslant 0.15$ are found in 7 out of 9 populations from Pakistan.

Network 3.1A (Fig. 1F) also covers a specific subset of the whole area occupied by frame 3 chromosomes. In addition to Northern Europe, this network is also frequent in the Brahui, Burusho and Punjabi of Pakistan (> 0.30).

Network 1.4 (map not shown) is observed only at low frequencies (< 0.05) in populations of the Central-Eastern Mediterranean area.

Network coalescence

Two methods were used to estimate network coalescence age. In the first one the calculation of ASD is made by assuming the ancestral haplo-



type to be the most common in each network. For networks 1.2, 1.3, 3.1G, 3.1A and 1.4 the commonest haplotype is also constituted by the modal microsatellite alleles (Ruiz-Linares *et al.* 1999) within each network, bringing further support to the choice as the ancestral candidate.

When comparing the results, in two cases the estimates calculated with one method were outside the confidence limits with respect to the other (Table 2). These cases refer to the networks where the major haplotype did not outnumber all other haplotypes. It is also plausible that the assumption of a star-phylogeny is not appropriate for these ancient networks.

Overall, of the six largest one-step networks, three (1.1, 2.1 and 3.1G) coalesce in the Palaeolithic, two (1.2 and 3.1A) seem to coalesce in a window of time post-dating the Last Glacial Maximum and one (1.3) dates back to the last three to four millennia of Mediterranean history.

Fst analysis

When the frequencies of the 15 networks (Table 1) were used to compute the overall Fst, a value of 0.142 was obtained ($p < 0.0001$). This confirms the very high among-population divergence revealed by Y-chromosome polymorphisms. In the set of 1485 pairwise comparisons some populations systematically produced higher-than-average Fst values, namely Basques, Southern Spaniards, Moroccan Arabs and Kalash (41, 19, 39 and 20 values which were above the 90th percentile of the Fst distribution, respectively). Indeed, these populations showed up in one or more of the maps.

We then exploited the subdivision of our population sample into 55 distinct locations to analyse the sources of the overall Fst heterogeneity, by matching the pairwise values with

Fig. 2. Scatterplots of Fst (Y-axis) vs. geographic distance (X-axis) for pairwise inter-population comparisons in subsets of data selected according to the following criteria: European vs. European (panel A); European vs. West Asian (panel B); West Asian vs. West Asian (panel C); Indo-European-speaking vs. Indo-European-speaking (panel D). The regression line is also reported.

Table 3. *Slope and intercept of linear regression of Fst on distance for different criteria of pairwise population comparisons*

Comparison	n	Slope ± s.e.[a] (P)	Intercept ± s.e. (P)
All data	1485	1.83 ± 0.19 (= 0.0003)*	0.068 ± 0.006 (<0.0001)
Geography			
Europe–Europe	595	6.49 ± 0.50 (= 0.0001)*	0.014 ± 0.009 (n.s.)
W. Asia–W. Asia	136	0.49 ± 0.67 (n.s.)*	0.041 ± 0.015 (<0.01)
N. Africa–N. Africa	3	0.47 ± 4.29 (n.s.)*	0.092 ± 0.125 (n.s.)
Europe–W. Asia	595	2.43 ± 0.32 (<0.0001)	0.012 ± 0.013 (n.s.)
Europe–N. Africa	105	0.14 ± 1.54 (n.s.)	0.219 ± 0.042 (<0.0001)
W. Asia–N. Africa	51	5.22 ± 0.78 (<0.0001)	0.018 ± 0.031 (n.s.)
Island–mainland	376	1.79 ± 0.30 (<0.0001)	0.085 ± 0.009 (<0.0001)
Linguistics[b]			
I.E.–I.E.	741	0.97 ± 0.19 (= 0.0126)*	0.064 ± 0.006 (<0.0001)
I.E.-Basque	39	4.80 ± 1.27 (<0.001)	0.256 ± 0.038 (<0.0001)
I.E.–A.A.	195	0.84 ± 0.63 (n.s.)	0.138 ± 0.022 (<0.0001)
I.E.–Altaic	195	4.78 ± 0.65 (<0.0001)	−0.022 ± 0.016 (n.s.)
I.E.–Uralic	117	2.86 ± 1.04 (<0.005)	0.070 ± 0.030 (<0.05)
I.E.–Dravidian	39	1.90 ± 0.81 (<0.05)	0.019 ± 0.036 (n.s.)
I.E.–Bu.	39	2.05 ± 0.93 (<0.05)	0.042 ± 0.042 (n.s.)
Basque–A.A.	5	−2.09 ± 1.53 (n.s.)	0.645 ± 0.064 (<0.002)
Basque–Altaic	5	17.64 ± 8.31 (n.s.)	−0.002 ± 0.260 (n.s.)
Basque–Uralic	3	1.35 ± 7.80 (n.s.)	0.513 ± 0.271 (n.s.)
A.A.–A.A.	10	4.22 ± 1.65 (= 0.0187)*	0.004 ± 0.057 (n.s.)
A.A.–Altaic	25	3.36 ± 2.45 (n.s.)	0.033 ± 0.059 (n.s.)
A.A.–Uralic	15	1.23 ± 6.27 (n.s.)	0.136 ± 0.240 (n.s.)
A.A.–Dravidian	5	6.13 ± 1.62 (<0.05)	−0.034 ± 0.060 (n.s.)
A.A.–Bu.	5	6.44 ± 1.77 (<0.05)	−0.007 ± 0.071 (n.s.)
Altaic–Altaic	10	6.62 ± 5.51 (n.s.)*	−0.045 ± 0.031 (n.s.)
Altaic–Uralic	15	7.82 ± 3.72 (n.s.)	−0.122 ± 0.089 (n.s.)
Altaic–Dravidian	5	−4.31 ± 3.86 (n.s.)	0.143 ± 0.120 (n.s.)
Altaic–Bu.	5	4.54 ± 2.67 (n.s.)	−0.042 ± 0.090 (n.s.)
Uralic–Uralic	3	−4.12 ± 2.62 (n.s.)*	0.076 ± 0.034 (n.s.)
Uralic–Dravidian	3	2.51 ± 7.92 (n.s.)	0.005 ± 0.283 (n.s.)
Uralic–Bu.	3	−0.87 ± 10.5 (n.s.)	0.127 ± 0.344 (n.s.)
Combined (partial listing)			
I.E.–I.E.; Europe-Europe	465	5.40 ± 0.54 (<0.0001)	0.015 ± 0.008 (n.s.)
I.E.–Basque; Europe-Europe	31	20.7 ± 3.20 (<0.0001)	0.032 ± 0.051 (n.s.)

[a] $\times 10^{-5}$

[b] A.A., Afro-Asiatic.
I.E., Indo-European.
Bu., Burushaski.

* Significance determined by Mantel test (10000 iterations).

the corresponding geographic distances of the populations being compared. There was a clear trend towards larger Fst values with increasing distance, in addition to few high Fst values for distances < 2500 km. The average increase of Fst was estimated at 1.8×10^{-5}/km (Table 3).

In cross-sectioning this cloud of points, geographical and linguistic criteria for each pairwise comparison were taken into account (Table 3). Comparisons of European vs. other European populations showed a strong dependence of Fst on distance, contrasting with a significantly (> 6 s.e.) reduced slope when comparing European vs. West-Asian populations (Fig. 2A vs. 2B). Pairwise comparisons of West-Asian populations showed an even lower value for the slope (Fig. 2C). The high Fst values at very short distances corresponded to comparisons among some of the Pakistani populations. The reduced size of these samples, together with strong drift effects, may explain this result. Comparisons between European and North-African populations produced a non-significant value for the slope. However, this latter subset included some of the points with the highest Fst at short distance, yielding a very high value for the intercept.

Comparisons between populations living on islands vs. the mainland were also considered. The resulting cloud of points overlapped with those obtained for mainland vs. mainland comparisons and parameters of the interpolated line were very similar to those obtained for the entire group.

Complex patterns also emerged when language affiliation was considered. Pairs of Indo-European-speaking populations produced a very low Fst-on-distance slope (Fig. 2D). Comparisons involving one Indo-European-speaking population vs. a population speaking a language belonging to another family produced slopes significantly different from 0, with the exception of the Afro-Asiatic. In this case, the Moroccan population plays a major role in disrupting the overall regression producing Fst values > 0.35 for 9 populations at < 2000 km, in agreement with the drastic frequency change shown by networks 2.1 and 3.1G across the Strait of Gibraltar. The Basques produced by far the highest Fst-on-distance slope when compared to Indo-European-speaking populations.

The only other subset producing a significant regression on distance is that involving the Afro-Asiatic-speaking population comparisons.

Combining the plots of Figure 2 shows that the reduced slope of Fst on distance for Indo-European vs. Indo-European comparisons was indeed the result of two subgroups. The first included populations displaying a large divergence with increasing distance up to 3000 km, mostly consisting of European data. In the second subgroup, when doubling the distance across the Europe–Asia border, a corresponding Fst average increase was not found. When we took into account both the geographic and linguistic criteria and computed the regression, only in the 465 pairwise comparisons between Indo-European-speaking populations living in Europe, a slope of 5.40×10^{-5} was obtained.

DISCUSSION

The extension of the sample size and the expansion of the surveyed area has revealed the presence of the same networks obtained by the combined use of binary markers and microsatellites, as previously identified throughout Europe, North Africa and West Asia. Present findings strongly suggest that most of the Y chromosome diversity has been sampled and that the discontinuity in the haplotypic combinations is no longer the result of incomplete sampling. We also found that the mutation rate in the dinucleotide microsatellites is large enough to produce great variation during a period of a few hundred generations. Moreover, we identified at least two lineages generated by multirepeat microsatellite mutations which become a useful tool to address questions on the peopling of the Mediterranean area in the Neolithic period and in the periods thereafter.

In our survey we confirm the presence of minor groups of YAP+/alphoid *Hin*dIII− chromosomes (Santos *et al.* 1996; Malaspina *et al.* 1998; Scozzari *et al.* 1999) that can be explained by recurrent loss of alphoid units with *Hin*dIII site.

With the above exceptions, the group of alphoid *Hin*dIII− chromosomes is also identified by other mutations, i.e. the C-T mutation at 92R7, the *Xba*I mutation at M911 and the G-A mutation at DYS257 (Mathias *et al.* 1994; Bosch *et al.* 1999; Scozzari *et al.* 1999). In summary, the group of *Hin*dIII− chromosomes can be considered largely overlapping with haplotype 1C and its derivatives 1D and 1G (Hammer *et al.* 1998; Karafet *et al.* 1999), with haplotypes 13, 10, 1, 20, 6, 4 and their derivatives 31 and 32 (Santos *et al.* 1999) and haplogroup 1 (Hill *et al.* 2000).

The analysis of the G-A mutation at SRY_{10831} on all alphoid *Hin*dIII− chromosomes was also carried out. Allele A identifies haplotype 1D (Hammer *et al.* 1998) and haplotype 32 (Santos *et al.* 1999), as well as the ancestral haplotypes 1A and 19 (by the same authors, respectively). Given the geographic origin of our sample, a thorough search for these latter haplotypes was not carried out since they were expected to be very rare.

As far as our dating results are concerned, in three instances we could compare our estimates

with the available data. Network 2.1, grouping 92% of the YAP+ chromosomes here found, provided coalescence estimates that include the value of 20 KYA for the PN2-T mutation as reported by Hammer *et al.* (1998). Indeed, it has been shown that YAP+ chromosomes found in Europe and North Africa mostly carry the PN2-T mutation (Hammer *et al.* 1997; Scozzari *et al.* 1999 and unpublished data). The estimates for network 1.3 are compatible with an origin not dating earlier than the first human settlements in Sardinia (9 KYA, Cappello *et al.* 1996). Finally, the estimates for network 3.1A are in agreement with the results from Hammer *et al.* (1998) for the SRY_{10831}-G to A reversion. Our data indicate that this reversion occurred on a haplotype very similar to the major haplotype of network 3.1G. It is likely that this fact, and the short time allowed for molecular radiation within network 3.1A, contributed to the partial overlap of haplotypes belonging to the two networks.

Geographic distributions

The three largest and oldest networks were geographically widespread and revealed clines that cover the entire European continent. An area characterized by high frequencies was found in North-Eastern Europe (Mordovia and Komi-Permiak) for network 1.1. Network 2.1, reflecting an input of African chromosomes into Europe, displayed frequencies < 0.10 in all of Northern Europe and most of Turkey, whereas in South-Eastern Europe it contributes considerably to its gene pool. The splitting of network 3.1 (Malaspina *et al.* 1998) revealed a clearer focus in Western Europe of what is now called network 3.1G, fitting the distribution of haplotype 15 identified by the p49f/*Taq*I system (Semino *et al.* 1996; Lucotte & Loirat, 1999; Quintana-Murci *et al.* 1999*a*; Hill *et al.* 2000; Scozzari R., unpublished data). The above distributions are in line with one or more population movements from Asia into Europe and within Europe. Cavalli-Sforza *et al.* (1994, p. 292) interpreted their second PC map as the result of a similar process(es). An early expansion from Central Asia has been postulated (Santos *et al.* 1999) to explain the entry of haplotype 1 into Europe. Subsequently, a late Palaeolithic population expansion, as hypothesised by Torroni *et al.* (1998) on the basis of the distribution of mtDNA haplogroups V and H, may have increased the frequencies of network 3.1G in far Western Europe. A third, more recent, event (Zerjal *et al.* 1997) is marked by the distribution of the Tat-C mutation, which belongs to a small subgroup within our network 1.1 (Scozzari R., unpublished data). Other migratory movements may have occurred South-East of the Caspian Sea. The prevalence of YAP−, alphoid *Hin*dIII−, SRY_{10831}-G chromosomes in Pakistan may trace a connection between Asia and Central Africa, where high frequency spots were recently found (Scozzari *et al.* 1999).

In Europe the three largest networks undergo sharp frequency changes in three areas, i.e. across the Strait of Gibraltar, around the Basque region and across an imaginary line approximately connecting the Eastern Alps to the Baltic Sea. While the first two boundaries were also detected in analyses of population data for isoenzyme variation (Barbujani & Sokal, 1990; Simoni *et al.* 1999), the latter did not appear in the same analyses.

The smaller and more recent networks 1.2, 1.3 and 3.1A showed different patterns. Network 1.2 consisted of a group of chromosomes of the Eastern Mediterranean area and its geographic distribution may reflect an East-to-West population movement, possibly as part of a larger gradient revealed by p12f2 haplotypes (Semino *et al.* 1996; Scozzari R., unpublished data). Overall, this Y-chromosomal pattern fits the autosomal data condensed in the first PC by Cavalli-Sforza *et al.* (1994), and is interpreted as the result of the demic diffusion associated with the Neolithic spread of agriculture. Locally, this network establishes similarities between the Cretan and Southern Anatolian populations, in agreement with the hypothesis on the origins of the first occupants of Crete around 7000 B.C., reported by Renfrew (1998 and citations therein). The occurrence of this network at appreciable frequencies in the Southernmost part of Continental

Italy and Eastern Sicily is compatible with the further spread of these chromosomes during the Greek colonisation of the latter areas in the first millennium B.C. In this context, the different frequencies of networks 1.1 and 1.2 (0.22 vs. 0.12 and 0.28 vs. 0.15, respectively) found in Eastern and Western Sicily, are in agreement with a more pronounced Greek influence in the Eastern part of the island (Piazza *et al.* 1988).

Network 1.3 is confirmed to be basically confined to Sardinia, showing a particularly high frequency. Therefore, this network reinforces the previously observed genetic boundaries around the island (Barbujani & Sokal, 1990; Simoni *et al.* 1999).

The distribution of network 3.1A shows an almost complete complementarity to the distribution of its precursor, network 3.1G (see above), contributing to the sharp frequency change across Central Europe. In our maps this network shows the highest frequencies in Eastern Europe and in Pakistan. High frequencies of SRY_{10831}-A chromosomes were found in Northern Europe, Central Asia and India (Karafet *et al.* 1999; Santos *et al.* 1999; Zerjal *et al.* 1999), pointing to their recent entry in to Europe from the East. Our data indicated no difference in either CA variance or in the sets of microsatellite haplotypes contributing to this network in North-Eastern Europe vs. Pakistan, suggesting that its dispersal was fast and postdated the accumulation of the overall diversity of the network. Therefore our data favour a relatively early origin and a much later dispersal. Although error margins are considerably large, coalescence estimates raise the possibility that SRY_{10831}-A was already present in Eurasia at the end of the Palaeolithic, and possibly in one of the glacial refuges of Eastern Europe (Soffer, 1990; Dolukhanov, 1993). The striking similarity between the distribution of network 3.1A and that of allele ABO*B over the area also surveyed by Cavalli-Sforza *et al.* (1994, maps 107 and 108) deserves particular attention. The ABO*B frequencies contribute mainly to the fourth and third PC for Asia and Europe, respectively (Cavalli-Sforza *et al.* 1994, pp. 249, 291; Cavalli-Sforza, 1997). These authors have related this pattern to the expansion of the Kurgan culture, a three-wave process occurring between 5000 and 2900 B.C., which could also be responsible for the distribution of network 3.1A (Zerjal *et al.* 1999). Alternatively, other migratory movements from Central Asia westward may have brought both markers to the places where they are found at present.

Fst analysis

The analysis of our Fst-on-distance regressions showed that high values for the intercept most likely result from local discontinuities in allele frequencies that clearly appear in the corresponding maps. The same analyses also allowed inferences about the pattern of male-specific gene flow over the area under scrutiny.

First, among the conventional geographic borders between the continents here considered, the European–Asian border was not associated with high Fst-on-distance slopes: the highest slope was in fact observed within Europe, to which the Basque population has contributed to a large extent. In general, all pairwise inter-population comparisons, except those involving Basques, produced a slope of Fst-on-distance which was half or less than that reported by Seielstad *et al.* (1998). Their regression was based on 10 points, and was most probably inflated by the inclusion of the Basques. Indeed, we showed a marked heterogeneity of the dependence of Fst on distance, with many types of comparisons producing slopes as low as those reported for mitochondrial or autosomal markers.

Second, the geographic separation of the islands considered in this study does not seem to have acted as a relevant obstacle to gene flow, in agreement with the view that from the Mesolithic onwards the Mediterranean Sea represented less of a barrier and more of a bridge (Binder, 1989). In addition founder or drift effects in Sardinia and Crete could have contributed to the rise in frequency of some haplotypes.

Third, different linguistic affiliations were often associated with reduced levels of genetic similarity. This was not, however, an absolute

rule. Fst values as high as 0.30 and 0.40 were found for comparisons between pairs of Arabic- or Indo-European-speaking populations, respectively. In the former case, Arabic-speaking populations geographically far apart (Moroccan Arabs vs. Omani and UAE) produced high Fst. This is most likely because of the acquisition of the Arabic language in recent times by populations with a very different genetic background. As for the Indo-European-speaking populations, the Kalash largely diverged (average Fst = 0.25) from all other populations of this linguistic family. Also, large Fst values were observed between Indo-European-speaking populations residing in Europe. In this context, a discontinuity within this continent seems to be more relevant than the conventional boundary between Asia and Europe.

Our analysis merged two approaches to investigate different demographic scenarios that shaped the present-day distribution of genetic markers. The area of low Fst-on-distance slope (East Europe and West Asia) parallels the area with the highest frequencies of network 3.1A. Both suggest a fast and recent population dispersal. In this work, for the first time, the frequency and Fst-on-distance patterns are shown to complement each other in revealing the effects of peopling processes that led to the observed frequency distributions.

conclusions

This investigation has revealed sharp changes in Y-chromosomal frequencies in Central Europe and patterns of distance-dependence in the divergence between populations greatly varying throughout Europe, North Africa and West Asia. This is because in none of the extant populations the repertoire of Y-chromosomal lineages is the result of internal evolution, but rather, it is the result of admixture of peoples carrying lineages which originated in extremely distant locations and over long time spans. Further complexity may have resulted from mutual exclusion of lineages and/or peoples, two processes that were most likely reiterated during the Y-chromosome phylogeny. The consequences are twofold. First, the description of genetic diversity over Europe and any inference on the causes of its distribution must include data from Africa and Asia. In searching for the relationships between European populations, data from the neighbouring areas could be used, in principle, as outgroups are used in sequence analyses. Second, the vivid debate on the relative contribution of Palaeolithic and Neolithic populations to the extant European gene pool, so far mostly based on autosomal and mtDNA data, should take into account the marked difference between Western and Eastern Europe for Y-chromosomal markers. Averaging over the entire continent may not be meaningful and the different estimates need to be reviewed in light of the precise geographic origins of the population samples they have been based upon.

We gratefully acknowledge Jean-Paul Moisan and Damian Labuda for contributing French DNA samples. We also thank Dr. M. Lo Ponte for revising the paper. Work supported by CNR grants 98.00485.CT04 (AN), 97.00712.PF36 (LT), 97.00702.PF36 (RS) and PRIN MURST 1999.

references

Altheide, T. K. & Hammer, M. F. (1997). Evidence for a possible Asian origin of YAP+ Y chromosomes. *Am. J. Hum. Genet.* **61**, 462–466.

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

Binder, D. (1989). Aspects de la néolithisation dans les aires padane, provençale et ligure. In *Néolithisation*. British Archaeological Reports International Series 516. (eds. O. Aurenche & J. Chauvin), pp. 199–226. Oxford.

Bosch, E., Calafell, F., Santos, F. R., Pérez-Lezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C. & Bertranpetit, J. (1999). Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am. J. Hum. Genet.* **65**, 1623–1638.

Cappello, N., Rendine, S., Griffo, R., Mameli, G. E., Succa, V., Vona, G. & Piazza, A. (1996). Genetic analysis of Sardinia. I. Data on 12 polymorphisms in 21 linguistic domains. *Ann. Hum. Genet.* **60**, 125–141.

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

Cavalli-Sforza, L. L. (1997). Genes, peoples, and languages. *Proc. Natl. Acad. Sci. USA* 94, 7719–7724.

Ciminelli, B. M., Pompei, F., Malaspina, P., Hammer, M., Persichetti, F., Pignatti, P. F., Palena, A., Anagnou, N., Guanti, G., Jodice, C., Terrenato, L. &

Novelletto, A. (1995). Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations. *J. Mol. Evol.* **41**, 966–973.

Cooper, G., Amos, W., Hoffman, D. & Rubinsztein, D. C. (1996). Network analysis of human Y microsatellite haplotypes. *Hum. Mol. Genet.* **5**, 1759–1766.

Delfiner, P. (1976). Linear estimation of non-stationary spatial phenomena. In *Advanced geostatistics in the mining industry* (eds. M. Guarasio, M. David & C. Haijbegts), pp. 49–68. Dordrecht: Reidel.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. & Freimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* **91**, 3166–3170.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petz-Erler, M. L., Haines, G. K. & Barch, D. H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics* **148**, 1269–1284.

Dolukhanov, P. (1993). Foraging and farming groups in North-Eastern and North-Western Europe: identity and interaction. In *Cultural transformations and interactions in Eastern Europe* (eds. J. Chapman & P. Dolukhanov), pp. 122–145. Avenbury: Aldeshat.

Forster, P., Kayser, M., Meyer, E., Roewer, L., Pfeiffer, H., Benkmann, H. & Brinkmann, B. (1998). Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol. Biol. Evol.* **15**, 1108–1114.

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.

Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Ruiz-Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. (1996). Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol. Biol. Evol.* **13**, 1213–1218 (erratum: *Mol. Biol. Evol.* **14**, 354 [1997]).

Grimes, B. F. (1996). *Ethnologue*. Languages of the World, 13th ed. Dallas: SIL International.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M. & Weissenbach, J. (1994). The 1993–94 Généthon human genetic linkage map. *Nature Genet.* **7**, 246–249.

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

Hammer, M. F., Spurdle, A. B., Karafet, T., Bonner, M. R., Wood, E. T., Novelletto, A., Malaspina, P., Mitchell, R. J., Horai, S., Jenkins, T. & Zegura, S. L. (1997). The geographic distribution of human Y chromosome variation. *Genetics* **145**, 787–805.

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.

Hill, E.W., Jobling, M. A. & Bradley, D. G. (2000). Y-chromosome variation and Irish origins. *Nature* **404**, 351–352.

Hurles, M. E., Irven, C., Nicholson, J., Taylor, P. G., Santos, F. R., Loughlin, J., Jobling, M. A. & Sykes, B. C. (1998). European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. *Am. J. Hum. Genet.* **63**, 1793–1806.

Jobling, M. A., Bouzekri, N. & Taylor, P. G. (1998). Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum. Mol. Genet.* **7**, 643–653.

Karafet, T. M., Zegura, S. L., Posukh, O., Osipova, L., Bergen, A., Long, J., Goldman, D., Klitz, W., Harihara, S., De Knijff, P., Wiebe, V., Griffiths, R. C., Templeton, A. R., & Hammer, M. F. 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* **64**, 817–831.

Karafet, T., De Knijff, P., Wood, E., Ragland, J., Clark, A. & Hammer, M. F. (1998). Different patterns of variation at the X- and Y-chromosome-linked microsatellite loci DXYS156X and DXYS156Y in human populations. *Hum. Biol.* **70**, 979–992.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Krüger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M. & Sajantila, A. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* **66**, 1580–1588.

Kittles, R. A., Perola, M., Peltonen, L., Bergen, A. W., Aragon, R. A., Virkkunen, M., Linnoila, M., Goldman, D. & Long, J. C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* **62**, 1171–1179.

Kwok, C., Tyler-Smith, C., Mendonca, B. B., Hughes, I., Berkovitz, G. D., Goodfellow, P. N. & Hawkins, J. R. (1996). Mutation analysis of the 2 kb 5′ to SRY in XY females and XY intersex subjects. *J. Med. Genet.* **33**, 465–468.

Lucotte, G. & Loirat, F. (1999). Y-chromosome DNA haplotype 15 in Europe. *Hum. Biol.* **71**, 431–437.

Malaspina, P., Ciminelli, B., Viggiano, L., Jodice, C., Cruciani, F., Santolamazza, P., Sellitto, D., Scozzari, R., Terrenato, L., Rocchi, M. & Novelletto, A. (1997). Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology. *J. Mol. Evol.* **44**, 652–659.

Malaspina, P., Cruciani, F., Ciminelli, B. M., Terrenato, L., Santolamazza, P., Alonso, A., Banyko, J., Brdicka, R., Garcia, O., Gaudiano, C., Guanti, G., Kidd, K. K., Lavinha, J., Avila, M., Mandich, P., Moral, P., Qamar, R., Mehdi, S. Q., Ragusa, A., Stefanescu, G., Caraghin, M., Tyler-Smith, C., Scozzari, R. & Novelletto, A. (1998). Network analyses of Y-chromosomal types in Europe, Northern Africa and West Asia reveal specific patterns of geographic distribution. *Am. J. Hum. Genet.* **63**, 847–860.

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

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.

Piazza, A., Cappello, N., Olivetti, E. & Rendine, S. (1988). A genetic history of Italy. *Ann. Hum. Genet.* **52**, 203–213.

Quintana-Murci, L., Semino, O., Minch, E., Passarimo, G., Brega, A. & Santachiara-Benerecetti, A. S. (1999*a*). Further characteristics of proto-European Y chromosomes. *Eur. J. Hum. Genet.* **7**, 603–608.

Quintana-Murci, L., Semino, O., Poloni, E. S., Liu, A., Van Gijn, M., Brega, A., Nasidze, I. S., Maccioni, L., Cossu, G., Al-Zahery, N., Kidd, J. R., Kidd, K. K. & Santachiara-Benerecetti, A. S. (1999*b*). Y-chromosome specific YCAII, DYS19 and YAP polymorphisms: a comparative study. *Ann. Hum. Genet.* **63**, 153–156.

Renfrew, C. (1998). Word of Minos: the Minoan contribution to Mycenean Greek and the linguistic geography of the Bronze age Aegean. *Cambridge Archaeological J.* **8**, 239–264

Ruiz Linares, A., Ortiz-Barrientos, D., Figueroa, M., Mesa, N., Munera, J. G., Bedoya, G., Velez, I. D., Garcia, L. F., Perez-Lezaun, A., Bertranpetit, J., Feldman, M. W. & Goldstein, D. B. (1999). Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. *Proc. Natl. Acad. Sci. USA* **96**, 6312–6317.

Santos, F. R., Bianchi, N. O. & Pena, S. D. J. (1996). Worldwide distribution of human Y-chromosome haplotypes. *Genome Res.* **6**, 601–611.

Santos, F. R., Pena, S. D. J. & Tyler-Smith, C. (1995). PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. *Gene* **165**, 191–198.

Santos, F. R., Pandya, A., Tyler-Smith, C., Pena, S. D. J., Schanfield, M., Leonard, W. R., Osipova, L., Crawford, M. H. & Mitchell, R. J. (1999). The central Siberian origin for Native American Y chromosomes. *Am. J. Hum. Genet.* **64**, 619–628.

Schneider, S., Kueffer, J-M., Roessli, D. & Excoffier, L. (1997). *Arlequin ver. 1·1: a software for population genetic data analysis.* Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Scozzari, R., Cruciani, F., Malaspina, P., Santolamazza, P., Ciminelli, B. M., Torroni, A., Modiano, D., Wallace, D. C., Kidd, K. K., Olckers, A., Moral, P., Terrenato, L., Akar, N., Qamar, R., Mansoor, A., Mehdi, S. Q., Meloni, G., Vona, G., Cole, D. E. C., Cai, W. & Novelletto, A. (1997). Differential structuring of human populations for homologous X and Y microsatellite loci. *Am. J. Hum. Genet.* **61**, 719–733.

Scozzari, R., Cruciani, F., Santolamazza, P., Malaspina, P., Torroni, A., Sellitto, D., Arredi, B., Destro-Bisol, G., De Stefano, G., Rickards, O., Martinez-Labarga, C., Modiano, D., Biondi, G., Moral, P., Olckers, A., Wallace, D. C. & Novelletto, A. (1999). Combined use of biallelic and microsatellite Y chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* **65**, 829–846, erratum *Am. J. Hum. Genet.* **66**, 346 [2000].

Seielstad, M. T., 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.

Simoni, L., Gueresi, P., Pettener, D. & Barbujani, G. (1999). Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Hum. Biol.* **71**, 399–415.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

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

Soffer, O. (1990). The Russian plain at the Last Glacial Maximum. In *The world at* 18000 *BP vol. I High Latitudes* (eds. O. Soffer & C. Gamble), pp. 228–252. London: Unwin Hyman.

Sokal, R., & Rohlf, F. J. (1995). *Biometry.* 3rd edition. New York: W. H. Freeman and Co.

Thomas, M. G., Parfitt, T., Weiss, D. A., Weiss, A., Skorecki, K., Wilson, J. F., Le Roux, M., Bradman, N. & Goldstein, D. B. (2000). Y chromosomes traveling South: the Cohen modal haplotype and the origins of the Lemba – the'black Jews of Southern Africa'. *Am. J. Hum. Genet.* **66**, 674–686.

Thomas, M. G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N. & Goldstein, D. B. (1998). Origins of Old Testament priests. *Nature* **394**, 138–140.

Torroni, A., Bandelt, H.-J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M.-L., Bonné-Tamir, B. & Scozzari, R. (1998). MtDNA analysis reveals a major late paleolithic population expansion from Southwestern to Northeastern Europe. *Am. J. Hum. Genet.* **62**, 1137–1152.

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L. & Oefner, P. J. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7**, 996–1005.

Weber, J. L. & Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128.

Whitfield, L. S., Sulston, J. E. & Goodfellow, P. N. (1995). Sequence variation of the human Y chromosome. *Nature* **378**, 379–380.

Zerjal, T., Pandya, A., Santos, F. R., Adhikari, R., Tarazona, E., Kayser, M., Evgrafov, O., Singh, L., Thangaraj, K., Destro-Bisol, G., Thomas, M. G., Qamar, R., Mehdi, S. Q., Rosser, Z. H., Hurles, M. E., Jobling, M. A. & Tyler-Smith, C. (1999). The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In *Genomic Diversity: Applications in Human Population Genetics* (eds. S. S. Papiha, R. Deka & R. Chakraborty), pp. 91–102. Kluwer Academic/Plenum Publishers.

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Roewer, L., Santos, F. R., Schiefenhövel, W., Fretwell, N., Jobling, M. A., Harihara, S., Shimizu, K., Semjidmaa, D., Sajantila, A., Salo, P., Crawford, M. H., Ginter, E. K., Evgrafov, O. V. & Tyler-Smith, C. (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am. J. Hum. Genet.* **60**,1174–1183.