Report

Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe

1,* 2* 1 2 1
1,3 1,3 1,4 1,4 1,5
1,5 1,7 1,8 6
2 2 2 1 9
9 3 4 5
7 8 10 10
1,10 11 11 12 13
14 14 15 16 17
18,19 20 21 2
9 2 1 2

1

Previous studies revealed that Hg I reached a frequency of ~40%–50% in two distinct regions—in Nordic populations of Scandinavia and, in southern Europe, around the Dinaric Alps—each showing different background STR modal haplotypes (Semino et al. 2000; Passarino et al. 2002; Barać et al. 2003). In addition, subclade I-M26 (Underhill et al. 2000) reaches a very high frequency (~40%) (Semino et al. 2000; Passarino et al. 2001; Francalacci et al. 2003) in Sardinia, particularly in the "archaic area" (Cappello et al. 1996; Zei et al. 2003), and is associated with the peculiar YCAIIb-11 allele (Ciminelli et al. 1995; Caglia et al. 1997; Quintana-Murci et al. 1999; Malaspina et al. 2000; Scozzari et al. 2001). Overall, these observations suggest that Hg I could have played a central role in the process of human recolonization of Europe from isolated refuge areas after the LGM and suggest the likelihood that a comprehensive phylogeographic study should be able to localize the in situ origin and spread of principal male founders.

In the present study, the M170 A→C transversion, which defines Hg I, was assessed in a total of 7,574 subjects, including 6,095 Y chromosomes from 48 European populations and 1,479 individuals from 12 populations of surrounding regions (the Near East, Macaronesia, Central Asia, and the Caucasus). The results are reported in table 1, together with 407 additional members of Hg I out of 3,859 Y chromosomes extracted from the literature. Of the 1,104 Y chromosomes from the present study (1,060 from European subjects and 44 from the adjacent regions) that showed the derived M170 C-allele, 236, representative of the entire collection, were first examined for all the Hg I mutations known to date—namely, M21, M26, P37, M72, M223, M227, M253, M258, M284, and M307, whose phylogenetic relationships are illustrated in figure 1*A*. Genotyping was performed in a hierarchical way, and methods are provided in the legend to figure 1. The M258 and M307 mutations were observed in all of the Hg I and I1a Y chromosomes, respectively, whereas the M21, M72, and M284 mutations were not found. Thus, all of these were subsequently omitted in the remainder of the survey. To evaluate the differentiation of the I subclades (fig. 2), 533 Hg I Y chromosomes from 34 population samples were examined for the microsatellites DYS19, DYS388, DYS390, DYS391, DYS392, and

.1erformed3,856from
M170 date—na11.2 TD[herf8, i-340()-200(Tj/F3 1 Tf0.75.69 TD-0.0001 TB[(.)-370.1(Ge-65TJ-21.8049 11.2 TD[)40()-200(Rted)

Table 1

Frequencies of Haplogroup I and its Subhaplogroups

Region and Population	Sample Size	Hg I		Frequency of I Subhaplogroup[a] (%)						[b]
			%	I* M170	I1a* M253	I1a4 M227	I1b*P37	I1b2 M26	I1c M223	
Western Europe:										
Portuguese[c]	303	16	5.3	1.3	1.3		.7	.3	1.6	.808
Andalusian[c,d]	103	4	3.9	2.9				1.0		
Catalan[c,d]	32	1	3.1		3.1					
Basque (Spanish, French)[c,d]	100	6	6.0					6.0		.000
Bearnais[c]	26	2	7.7					7.7		
French (Southern France)[c]	38	6	15.8	5.3	5.3				5.3	.800
French (Low Normandy)[c]	42	10	23.8	4.8	11.9			2.4	4.8	.733
French (Lyon, Poitier)[c]	99	4	4.0		2.0			1.0	1.0	
Swiss[c]	144	11	7.6	.7	5.6				1.4	.473
Irish (Rush)[e]	76	8	10.5	NT	NT	NT	NT	2.6	NT	
Welsh[e]	196	16	8.1	NT	NT	NT	NT	.5	NT	
English[e]	945	174	18.4	NT	NT	NT	NT	.7	NT	
Scottish[e]	178	20	11.2	NT	NT	NT	NT		NT	
Scottish (Scottish Isles)[e]	272	45	16.5	NT	NT	NT	NT	.4	NT	
German[c,d]	16	6	37.5		25.0				12.5	.733
Dutch										

Table 1 (continued)

Region and Population	Sample Size	Hg I		Frequency of I Subhaplogroup[a] (%)						[b]
			%	I* M170	I1a* M253	I1a4 M227	I1b*P37	I1b2 M26	I1c M223	
Central-eastern Europe:										
Polish[c,d]	191	34	17.8		5.8	1.0	9.9		1.0	.593
Czech and Slovak[c,d]	198	27	13.6	.5	4.5	.5	7.1		1.0	.635
Hungarian[c,d]	162	37	22.8	.6	9.9		11.1		1.2	.588
Byelorussian[c]	147	28	19.0	.7	2.7		15.0		.7	.373
Ukrainian[c,d]	585	128	21.9	.2	4.8	.3	16.1		.5	.415
Russian (northern, Pinega)[c]	127	6	4.7		.8		3.9			.333
Russian (Kostroma region)[c]	53	10	18.9		6.0		9.4		3.7	.688
Russian (Smolensk region)[c]	120	13	10.8		1.7		9.1			.283
Russian (Belgorod region)[c]	144	24	16.7		3.5		12.5		.7	.634
Russian (Cossacks)[c]	97	22	22.7	1.0	4.1		15.5		2.0	.515
Russian (Adygea)[c]	78	19	24.4	1.3	5.1		16.7		1.3	.508
Russian (Bashkortostan)[c]	50	3	6.0		4.0		2.0			
Udmurt[c,d]	132	3	2.3	NT	NT	NT	NT	NT	NT	
Mordvin[c]	83	16	19.3		12.0		2.4		4.8	.566
Komi[c]	110	5	4.5		3.6		.9			
Chuvashes[c]	80	9	11.3		7.5		1.3		2.5	.555
Tatar[c]	123	6	4.9	1.6	.8		2.4			.733
Near East:										
Turkish[c,d,k]	741	38	5.1	1.1	.9		2.3		.7	.723
Lebanese[c,d]	66	3	4.5			1.5	1.5		1.5	
Jewish[c,d]	150	2	1.3		.7		.7			
Iraqi[c,l]	176	1	.6	.6						
Iranian[c]	83	0								
Caucasus, Central Asia:										
Nogays[c]	61	3	4.9				4.9			
Adygeis[c]	138	6	4.3	1.4			2.9			.533
Karachais[c]	70	5	7.1				7.1			
Northern Caucasian[m]	114	7	6.1	NT	NT	NT	NT	NT	NT	
Southern Caucasian[m]	249	10	4.0	NT	NT	NT	NT	NT	NT	
Georgian[c,d]	63	0								
Central Asian[n]	984	15	1.5	NT	NT	NT	NT	NT	NT	

[a] NT = not tested.

[b] Haplogroup diversities ( ) were calculated as described by Nei (1987) if more than five Y chromosomes belonged to haplogroup I.

[c]

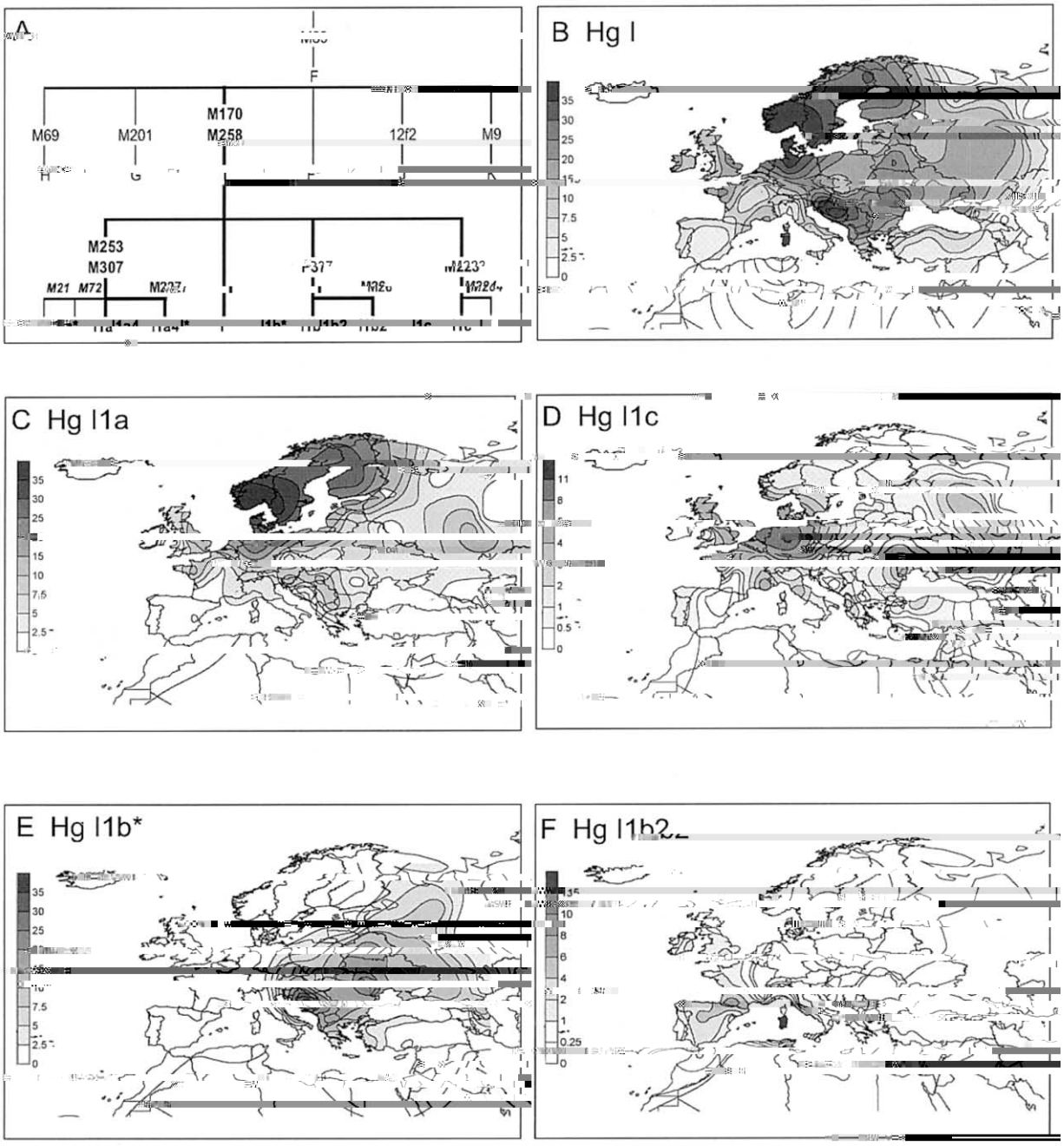

Figure 1 *A*, Phylogram of Hg I and its subclades within the context of the superhaplogroup F. Mutation labeling follows the Y Chromosome Consortium nomenclature (Y Chromosome Consortium 2002; Jobling and Tyler-Smith 2003). Markers M21 and M72 (Underhill et al. 2001) and the three new markers—M258 (a T→C transition at position 123), M284 (ACAAdel at position 105), and M307 (a G→A transversion at position 282)—were examined in a subset of 236 Y chromosomes, representative of the entire collection, by using the DHPLC method. The primers used for the new markers were as follows: F, 5′-tatatagcatatgttaaatgtttaggt-3′ and R, 5′-gacttttgaataatttgcatctttc-3′ for M258; F, 5′-ggcagttttcatttaagcaga-3′and R, 5′-agcgaaactttcagcacttc-3′ for M284; and F, 5′-ttattggcatttcaggaagtg-3′ and R, 5′-gggtgaggcaggaaaatagc-3′ for M307. When the DHPLC method was not used, M170 was detected as described by Ye et al. (2002); M253 was detected by using published primers (Cinnioğlu et al. 2004) and restriction analysis with cII; P37 was assayed by I digestion using the primers given by YCC (2002); and M223, M26, and the novel M227 (a C→G transversion at position 157) were studied by sequencing using published primers (Underhill et al. 2001) and the primers F, 5′-gagtgccaagctgaggatg-3′ and R, 5′-tccttgcagccgctgaggag-3′, respectively. A minority (= 67) of widely geographically distributed Hg I Y chromosomes (table 1) not tested for M258 and not harboring derived alleles at the sites M253, P37, and M223 were aggregated into paragroup I*. *B* , Frequency distribution of haplogroup I (*B*) and its subclades: I1a (*C*), I1c (*D*), I1b* (*E*), and I1b2 (*F*). Maps were obtained by applying the frequencies from table 1 in Surfer (version 7) software (Golden Software).

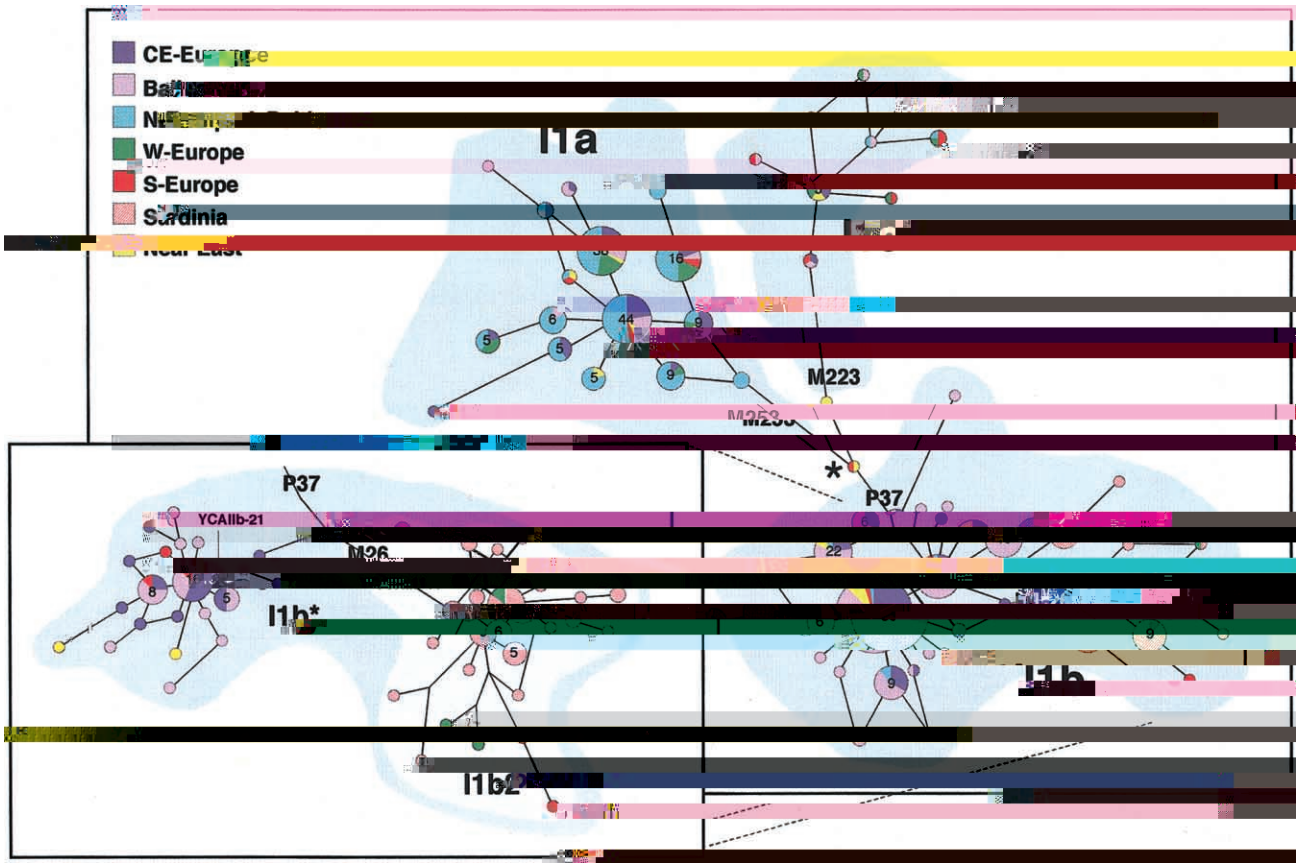

Figure 2

In central and eastern Europe, subhaplogroups I1a and I1b show overlapping frequency gradients, although with opposite post-LGM spreading. The divergent distributions of I1b2 and I1b* suggest that their separation occurred before the LGM and that the M26 mutation arose in a I1b Y chromosome from western Europe, most likely in a population in Iberia/southern France. The exceptionally high incidence of I1b2 in the archaic zone of Sardinia (Cappello et al. 1996; Zei et al. 2003) can be explained by the presence of I1b2 chromosomes among the first humans who colonized the island, ~9,000 years ago, followed by isolation and genetic drift. The extremely low frequency of I1b2 in the Scandinavian Peninsula, where the "western European" I1a Y chromosomes account for the large majority of Hg I, suggests, in addition, that the ancestral western European population(s), characterized by the M26 mutation, probably played a minor role in the colonization of that region. A geographic and genetic subdivision within the broad western refuge area, together with differences in initial sample size, genetic drift, and expansions, could also explain the quite different distribution of Hg I subhaplogroups with respect to the west-east decreasing gradient displayed by R1b, the most frequent subhaplogroup in western Europe.

The high STR diversity of the I1b* lineages in Bosnia supports the view that the P37 SNP might have been present in the Balkan area before the LGM, as previously proposed by Semino et al. (2000). Diversity values based on STR haplotypes for I1a are highest near Iberia but vary substantially in different populations (table 2). For I1b*, conversely, the highest values are in the Balkan populations—among Bosnians (0.93) and Croats (0.85)—coinciding with the area of its frequency peak, but equally high values were also observed for Czechs and Slovaks (0.90). The lowest values of I1b* were detected among Turks (0.76) and in our Moldavian sample (0.41).

Subhaplogroup I1c (fig. 1*D*) covers a wide range in Europe, with the highest frequencies (~5%–12%) in northwestern Europe and lower frequencies elsewhere. Its geographic and linguistic correlations across the continent were insignificant. However, the 49a,f system and the microsatellites YCAIIa-YCAIIb reveal that I1a and I1c harbor an identical compound haplotype (49a,f Ht12/YCAIIa-21/YCAIIb-19), which is different from those of I1b* (49a,f Ht10/YCAIIa-21/YCAIIb-21) and I1b2 (49a,f Ht12/YCAIIa-21/YCAIIb-11). These results indicate that subhaplogroups I1a and I1c may be part of a single monophyletic clade whose deep biallelic mutations are still undefined and that they probably share a common history of expansion. This scenario is also supported by the high positive correlation between the geographic distributions of I1a and I1c (= 0.75 when Fennoscandia is excluded; significance level 0.999).

Table 2

Haplotype Diversity of I1a and I1b*

	Diversity of[a]	
Population	I1a	I1b*
French	.972	…
Italian	.933	…
Swiss	.750	…
Norwegian	.809	…
Saami	.806	…
Swedish	.926	…
Estonian	.895	.867
Hungarian	.884	.746
Ukrainian	.782	.802
Czech and Slovak	.857	.904
Polish	…	.818
Croat	.830	.845
Bosnian	…	.929
Gagauz (Moldova)	…	.900
Moldavian	…	.410
Turkish	.800	.755

[a] Haplotype diversity values were calculated according to Nei (1987), using the STR (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) haplotype frequencies only when more than five Y chromosomes were found to belong to either I1a or I1b*.

Anatolia is at the easternmost fringe of the spread of haplogroup I, where it is found at higher frequencies in the regions that are geographically closer to Europe (Cinnioğlu et al. 2004). This observation, combined with a low haplotype diversity in Turkey plus exact haplotype matches with Europe, suggests that haplogroup I Y chromosomes in Turkey are due to migrations from Europe, as has been argued for a fraction of the Turkish mtDNAs (Richards et al. 2000).

A temporal interpretation of the phylogeography based on the results of the STR length variation in the individual subhaplogroups of I (Zhivotovsky et al. 2004) is reported in table 3. The age of STR variation for I* was estimated as 24,000 ± 7,100 years, a value that is very close to the population divergence time (23,000 ± 7,700 years). This finding supports the earlier suggestion that haplogroup I originated from a pool of European pre-LGM, middle Upper Paleolithic Y chromosomes (Semino et al. 2000). Our time estimates hint that its initial spread in Europe may be linked to the diffusion of the largely pan-European Gravettian technology ~28,000–23,000 years ago (Djindjian 2000; Perles 2000). On the other hand, these values represent the lower limit of the age of M170 mutation. The precedent mutation (M89) (fig. 1*A*) defines the overarching superhaplogroup F, whose representatives span the entire non-African gene pool, likely predating the peopling of Europe (some 40,000–50,000 years ago). Potentially more informative are the estimates of subclade divergence times. Thus, it appears that I1a, I1b, and I1c all

Table 3

Age Estimates and Divergence Times of Haplogroup I Subclades

Age Estimate or Divergence Time	Haplogroup I Subclade				
	I*	I1a	I1b*	I1b2	I1c
Time since subclade divergence[a]	…	15.9 ± 5.2[b]	10.7 ± 4.8[b]	9.3 ± 7.6[c]	14.6 ± 3.8[b]
Age of STR variation[d]	24.0 ± 7.1	8.8 ± 3.2	7.6 ± 2.7	8.0 ± 4.0	13.2 ± 2.7
Time since population divergence[e]	23.0 ± 7.7[f]	6.8 ± 1.9[g]	7.1 ± 2.5	7.9 ± 3.6[h]	11.2 ± 2.3[i]

[a] The times, in thousands of years, when the subclades I1a, I1b*, and I1c diverged from I*, as well as when I1b2 diverged from I1b*, were estimated by using the $_D$ estimator: $_D = (\hat{D}_1 - 2\ _0)/2$ (Zhivotovsky et al. 2001, 2004). Here, $\hat{D}_1$ is the average squared difference between two alleles sampled from two populations; $_0$ is the within-population variance in the number of repeats in the ancestral population prior to its subdivision, estimated as a half square difference between the allele repeat scores at the founder haplotypes; and $$ is the effective mutation rate of 0.00069 per locus per 25 years (Zhivotovsky et al. 2004).

[b] Divergence from I*.

[c] Divergence from I1b*.

[d] The age of STR variation of a subclade was estimated as the average squared difference in the number of repeats between all sampled chromosomes and the founder haplotype, divided by $$. Ages of STR variation within clades I and I1b were estimated by using I* and I1b* Y chromosomes, respectively. This makes them statistically independent from the STR variation of their subclades, although they could be still biased because of uncertainties on founder haplotypes.

[e] The age of population expansion (divergence), estimated with $_D$, letting $_0 = 0$, gives its upper bound. Time since population divergence was analyzed only in populations with a sample size of at least five individuals; the estimates give an upper bound for the time of population expansion (divergence).

[f] Since all populations, except the Tofasbouupp2s indi8iduals;

Italian Ministry of the University: Progetti Ricerca Interesse Nazionale 2002 and 2003 (to A.T.); National Institutes of Health grant GM28428 (to the Stanford researchers); and Ministry of Science and Technology of the Republic of Croatia project number 0196005 (to P.R.).

Electronic-Database Information

Dupanloup I, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome P49a,f/ I haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035

Quintana-Murci L, Semino O, Poloni ES, Liu A, Van Gijn M, Passarino G, Brega A, Nasidze IS, Maccioni L, Cossu G, al-Zahery N, Kidd JR, Kidd KK, Santachiara-Benerecetti AS (1999) Y-chromosome specific YCAII, DYS19 and YAP polymorphisms in human populations: a comparative study. Ann Hum Genet 63:153–166

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, et al (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276

Roewer L, Arnemann AJ, Spurr NK, Grzeschik KH, Epplen JT (1992) Simple repeated sequences on the Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89:389–394

Roewer L, Kayser M, Dieltjes P, Nagy M, Bakker E, Krawczak M, de Knijff P (1996) Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. Hum Mol Genet 5: 1029–1033

Sanchez JJ, Borsting C, Hallenberg C, Buchard A, Hernandez A, Morling N (2003) Multiplex PCR and minisequencing of SNPs: a model with 35 Y chromosome SNPs. Forensic Sci Int 137:74–84

Scozzari R, Cruciani F, Pangrazio A, Santolamazza P, Vona G, Moral P, Latini V, Varesi L, Memmi MM, Romano V, De Leo G, Gennarelli M, Jaruzelska J, Villems R, Parik J, Macaulay V, Torroni A (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. Hum Immunol 62:871–884

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, San-