European Journal of Human Genetics (2000) 8, 964–974

www.nature.com/ejhg

ARTICLE

mtDNA hypervariable region II (HVII) sequences in human evolution studies

Antonio Salas[1], Victoria Lareu[1], Francesc Calafell[2], Jaume Bertranpetit[2] and Ángel Carracedo[1]

[1]*Unidad de Genética Forense, Departamento de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, Galicia;* [2]*Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Catalonia, Spain*

Variation in human mitochondrial DNA (mtDNA) has been used to infer the origin and migration patterns in human populations. mtDNA analysis has been focused mainly on the first hypervariable region (HVI). Nevertheless, although many studies of the second hypervariable region (HVII) have been carried out during recent years, the correlation between the first and the second hypervariable regions has not been well established. We have analysed 71 individuals from a relatively isolated region at the westernmost edge of continental Europe (Galicia, NW Iberian peninsula) and we have used available HVII sequence information from another 17 European and African populations. The results show high concordance between the two hypervariable regions, not only in variability levels but also in other phylogenetic aspects. The study of the population structure through an AMOVA analysis shows a low level of heterogeneity in the European populations. Nevertheless, we have found some inconsistency in the results, which are related to the mutation rate in these two hypervariable regions. These results are compatible with a high heterogeneity of mutation rates across the HVII region and stress the interest of HVII in population and forensic genetics. *European Journal of Human Genetics* (2000) **8**, 964–974.

Keywords: mtDNA; control region; mutation rate; mismatch distribution; pairwise differences; neighbour joining tree

Introduction

The analysis of genetic variation in the nucleotide sequences of mitochondrial DNA has allowed to unravel evolutionary aspects concerning the origin of modern human populations and the clarification of ancient human migration patterns (see [1–6] among others). The analysis of mtDNA sequence variation has been focused especially on the study of the first and the second hypervariable regions (HVI and HVII, respectively). Some specific features of mtDNA make it a suitable tool for human population issues:

(a) mtDNA is maternally inherited;

(b) it does not recombine;[7]

(c) mtDNA sequences evolve much faster than the average nuclear genes,[8,9]

(d) there are thousands (1000–10 000) of mtDNA molecule copies per cell.

Although most of the mtDNA sequencing studies have been focused on the HVI region, over the last years sequences of the HVII one have been determined in many populations representing most of the main continental groups (see[10–16] among others). HVI and HVII regions seem to have similar evolutionary processes; however, variation patterns in the two regions have not been explored in depth yet. Since the mtDNA molecule is transmitted as a non-recombination unit, the entire molecule has a single history. Nevertheless, HVI and HVII show different mutation rates,[17] and, if the difference in mutation rates were broad enough, variation patterns in the two regions could reflect different past events. Moreover, it is well known that there are several segments in the HVII region which are involved in the functional aspects

Correspondence: Professor Dr Ángel Carracedo, Institute of Legal Medicine, Genetics Service, Faculty of Medicine, C/San Francisco s/n, E-15705, Santiago de Compostela, Spain. Tel: +34 981 58 23 27; Fax: 34 981 58 03 36; E-mail: apimlang@usc.es
Received 7 October 1999; revised 14 July 2000; accepted 26 July 2000

related to the replication and translation in the mtDNA molecule.[18–20] The functional aspects of those sites could constrain the amount of neutral variation. Different studies have tried to estimate the genealogical mutation rate of HVI and HVII, but a consensus does not seem to have been reached.[21–24] Besides the differences in mutation rates between HVI and HVII, it also seems that mutation rates are more heterogeneous across nucleotide positions in HVII than in HVI.[25] It is not known to what extent the differences in mutation rate and heterogeneity between HVI and HVII could affect the evolutionary inferences drawn from the study of HVI. To that effect, we have typed HVII in a Galician sample (NW Spain), for which HVI sequences were known,[26] and we have compared HVI and HVII variation, within the same population and within the framework of the most extensively studied continent, namely Europe. We have used several analytical tools, such as pairwise difference means and distributions,[27] genetic distances among populations, Tajima's test,[28] AMOVA[29] and others, and we have compared the results obtained with each region.

Material and methods

Population samples, mtDNA amplification and sequencing

Blood fluid samples were obtained from 71 maternally unrelated individuals from Galicia (NW Spain), all of them were natives and Galician speakers. DNA was extracted using the method described in Salas *et al*.[26] The amplification of HVII was carried out in a Perkin Elmer 480-A Thermocycler. The temperature profile for 32 cycles of amplification was 95°C for 10 s, 60°C for 30 s, and 72°C for 30 s. Primers and PCR strategy described by Wilson *et al*[30] have been used. PCR product purification and sequencing were performed as in Salas *et al*,[26] except for the DNA sequencing kit (Sequencing Kit with dRhodamine Terminators; Perkin Elmer, Applied Biosystems, Foster City, CA, USA) and the automated sequencer (ABI 377, Applied Biosystems) used.

A computer file with the sequences is available by E-mail on request to apimlase@usc.es

Numerical analysis

CLUSTAL W (1.5) Multiple Sequence Alignment Program[31] was used to align the HVII sequences. The final information for each individual was a string of 360 bp belonging to the mtDNA hypervariable region II (HVII), from base positions 48 to 407.[32] However, a segment of 260 bp (from base position 63 to 322) was used to compare with other populations because that is the stretch for which most information is available.

Data from 10 different populations was used for comparison (see Results, Table 2): Galicians (present study), British,[10] Tuscans,[12] Austrians,[13] Bulgarians and Turks,[14] Koreans,[15] Biaka Pygmies, Mbuti Pygmies, and !Kung.[16] Additionally, sequences from Cornish, Hebrideans, Orcadians, Northumbrians, Northern Irish, Icelandic, French, and Danes[11] were used in some analyses but not on those that required a direct comparison of HVII and HVI sequences since data one exactly the same individuals was not available for those populations.

The mean number of nucleotide pairwise differences and the number of segregating sites were estimated and used to compute Tajima's D statistic.[28] The pairwise difference distribution was obtained, as well as the τ parameter from the two-parameter model by Harpending *et al*.[33] For the Galician population, standard errors were computed from 1000 bootstrap iterations.[34] Sequence and nucleotide diversity were estimated as described by Salas *et al*.[26] Phylogenetic comparisons of individuals in each population were carried out using reduced median networks.[35] To construct the network some positions that present a mutation rate six times higher than the average for the HVII region[17] have been removed from the original sample as discussed below. Genetic distance matrices between populations were obtained by using the intermatch–mismatch genetic distance. Standard errors were estimated by bootstrap.[34] Distances between populations were depicted as neighbour-joining trees,[36] by using the PHYLIP packaged.[37] To test the degree of congruence between genetic distance in the two different hypervariable regions, a Mantel matrix comparison test[38,39] was carried out using the TFPGA program.[40] The apportionment of genetic variation between and within populations was estimated by AMOVA,[29] by means of the Arlequin package.[41] In some cases, parameters obtained from HVII and from HVI were compared through their ratio adjusted by segment length (260 bp for HVII and 360 for HVI).

Results

Genetic diversity in the Galician population; homogeneous levels of variability in European populations in HVII

A total of 71 individuals from the Galician population have shown 47 different sequences defined by 35 segregating nucleotide positions (between nucleotides 63 and 322). Thirty-one positions show substitutions (28 transitions and three transversions), and four positions represented length polymorphism (Table 1). Out of the 47 different mtDNA sequences, 39 were found once, three were found twice, two were found three times, and three different sequences were shared by four, five and 11 individuals, respectively (Table 1).

Cambridge Reference Sequence (CRS)[32] presents a track of seven cytosines from positions 303 through to 309. An additional C was found in 29 individuals (40.8%), and three individuals presented two additional Cs in this stretch. A total of 64 individuals (90.1%) were also found to have an additional cytosine with respect to the CRS in the cytosine tract 311–315. This is the usual pattern found in all the samples analysed from any continent. The frequency of one

Table 1 Variable positions for 71 mitochondrial HVII sequences (from base position 48 to 407). Dots indicate the presence of the same nucleotide as in the CRS.[32] A dash in SEQ 32 indicates a deletion in this position. Numbers at the top represent Anderson's numeration. Haplogroups were assigned according to HVI and HVII motifs. HVI sequences for the same individuals have been published by Salas *et al*.[26] Nucleotides in bold correspond to HVII haplogroup-defining motifs. HVI motifs are given after the haplogroup designation. The haplogroup designed as H? corresponds to sequences bearing the Cambridge Reference Sequence at HVI and a G at position 73 of HVII. Several sequences have been grouped under 'other' because no clear sequence motif could be identified in either HVI or in HVII. The sequence identification code on the left of the Table corresponds to that used in Salas *et al*.[26] In those cases where several individuals have the same HVI but different HVII sequences, the notation has been changed by adding an identification to the previous notation (i.e. SEQ1.1 in the present paper is one of the 24 SEQ1 sequences published in Table 1 of Salas *et al*.[26] SEQ1.2 is another of these 24 sequences but with a different HVII sequence from SEQ1.1)

	111111111111112222222222222333333		
	556677901145578888899000022233346900014 6		
	190423334960222568958047825858923599950 7		
CRS	TTTCTAAGCTTCTTCGCAATCATGTCGGAATCACC---CA	*HAPLOGROUP*	
	aba		
SEQ52	...T.....G..C....	H	
SEQ1.14	.....G..C.C..	H	
SEQ1.9	.....G..C.C..	H	
SEQ1.8	.....G..C.C..	H	
SEQ1.6	.....G..C.C..	H	
SEQ20	..C.....G..C.C..	H	
SEQ31.2	.....G.....	H	
SEQ32	.....G.-..C..	H	
SEQ31.3	.....G....C..	H	
SEQ28.2	.....G....C..	H	
SEQ28.1	.....G....C..	H	
SEQ5	.....G....C..	H	
SEQ1.5	.....G....C..	H	
SEQ25	.....G....C..	H	
SEQ2.2	.....G....C..	H	
SEQ21	.....G....C..	H	
SEQ1.10	.....G..C.C.G	H	
SEQ31.1	.C.....G....C..	H	
SEQ15	..........T.....G....C..	H	
SEQ24	.....G.....G....C..	H	
SEQ13.1	.....T.....C.....G..C....	H	
SEQ13.2	.....T.....C.....G....C..	H	
SEQ1.1	.....C.....G....C..	H	
SEQ1.2	.....C.....C.....G.....	H	
SEQ19	.....T.C.C.....C.....G..CCC..	H	
SEQ1.12	.....C.....G..CCC..	H	
SEQ41.2	.....C.....G..C.C..	H	
SEQ41.1	.....C.....G..C.C..	H	
SEQ12	.....C.....G..C.C..	H	
SEQ63	.....G...C.....C.C..	H	
SEQ1.2	.....G.....G..C.C..	H	
SEQ44	G.....G.....G..C.C..	H	
SEQ1.11	.....C.G..C.C..	H	
SEQ33	.....C.....C.G..C.C..	H	
SEQ1.13	.....C.....C.C..	H	
SEQ10	.....C.....G....C..	H	
SEQ8	.....C.....G....C..	H	
SEQ4	.....C.....G....C..	H	
SEQ2.3	.....C.....G....C..	H	
SEQ1.4	.....C.....G....T..G.....	H	
SEQ1.16	.....*G*.....T.....G....C..	H?	(CRS)
SEQ1.15	.....*G*.....T.....G..C....	H?	(CRS)
SEQ1.7	.....*G*.....C.....G....C..	H?	(CRS)
SEQ22	....**C**.....C.....G..C.C..	V	(16298C)
SEQ42	.....G....C..	K	(16224C/16311C)
SEQ43	.....G..T.....G..C.C..	K	(16224C/16311C)
SEQ29	.....G.....C.....G....C..	U4	(16356C)
SEQ39	.....G.....G....C..	U5	(16270T)

Table continues on next page

Table 1 – *continued from previous page*

	111111111111112222222222222333333		
	556677901145578888899000022233346900014 6		
	190423334960222568958047825858923599950 7		
CRS	TTTCTAAGCTTCTTCGCAATCATGTCGGAATCACC---CA	*HAPLOGROUP*	
	aba		
SEQ38	.....G.....T.....G....C..	U5	(16270T)
SEQ27	.....G.....C.....G..C.C..	U6	(16172C/16219G)
SEQ50	.....G.....TC.....G..C....	J	(16069T/16126C)
SEQ48.1	.....G....C.....G.....T.AG...G....C..	J	(16069T/16126C)
SEQ49	.....G....C....A.G.....T.AG...G....C..	J	(16069T/16126C)
SEQ17	.....G.....A.....A....G**T**.C.C..	J	(16069T)
SEQ48.2	.....G....C.....T.AG...G**T**...C..	J	(16069T/16126C)
SEQ47.3	.....G.....G**T**.C.C..	J	(16069T/16126C)
SEQ47.1	.....G.....G**T**...C..	J	(16069T/16126C)
SEQ47.2	.....G.....G.....G**T**.CCC..	J	(16069T/16126C)
SEQ51	.....TG**T**...C..	J2	(16069T/16126C/16261T)
SEQ53	.....C.....G..C.C..	T	(16126C/16294T/16296T)
SEQ34	.....G.....**G**...**CA**..A.....G..C.C..	W	(16223T/16292T)
SEQ45	.....G...C.....**G**C..**CA**.....G....C..	W	(16223T/16292T)
SEQ35	.....G.....C.....G..C.CTG	X	(16223T/16278T)
SEQ33	.....G....CTC.T...GCT.C.....G..C.C..	X	(16223T/16278T)
SEQ7	.....G.....G....C..	OTHER	
SEQ26	.....G.....T.....G..C.C..	OTHER	
SEQ36	.....G.....T.....G.....G....C..	OTHER	
SEQ2.1	G....G.....G.....G..C.C..	OTHER	
SEQ40	.....G.A.....G.....G....C..	OTHER	
SEQ11	.....G.....C.....G.....G....C..	OTHER	
SEQ16	.....G.....G.....G....C..	OTHER	

Table 2 Results of the computation of several diversity indexes for the HVII region analysed in several populations distributed in different continents. The values in italics for *'Galician'* were calculated using the total information of the sample analysed in this work, from base position 48 to 407. The rest of the values were calculated using the information from base position 63 to base position 322. For parameters S, M, and τ[27] a length-adjusted ratio with values for HVI (np 16024–16383) is given for those populations in which information for both segments in the same individuals is available; τ ratio is only given for populations with a pairwise difference distribution conforming to Rogers and Harpending[27] distribution. Tajima's D for HVI is also given for comparison. The significance of the S_{II}/S_I ratio was estimated by means of a chi-square test; the significance of D is according to Tajima[28]

Populations	*N*	*K*	S_{II}	S_{II}/S_I	*H*	π	*M*	M_{II}/M_I	D_{II}	D_I	τ_{II}/τ_I
Galician	*71*	*51*	*40*	–	*0.978*	*0.0094*	*3.99*	–	*–1.683*	–	–
Galician	71	47	35	0.850	0.968	0.0143	3.80	1.678	–1.530	–2.328**	1.845
British	100	51	46	0.951	0.962	0.0146	3.87	1.203	–1.783*	–2.121*	0.998
Cornish	13	12	21	–	0.987	0.0183	4.85	–	–	–	–
Hebridean	19	16	21	–	0.983	0.0176	4.67	–	–	–	–
N. Irish	22	18	16	–	0.983	0.0153	4.06	–	–	–	–
Northumbrian	17	14	21	–	0.971	0.0186	4.82	–	–	–	–
Orcadian	62	37	40	–	0.965	0.0146	3.88	–	–	–	–
Icelandic	13	6	6	–	0.818	0.0064	1.70	–	–	–	–
French	12	10	12	–	0.970	0.0138	3.65	–	–	–	–
Danish	16	14	16	–	0.975	0.0141	3.74	–	–	–	–
Austrian	101	56	37	0.693*	0.968	0.0137	3.63	1.096	–1.524	–2.204**	1216
Tuscan	49	36	32	0.805	0.981	0.0157	4.16	1.143	–1.413	–2.060*	0.943
Bulgarian	30	24	19	0.710	0.984	0.0127	3.36	1.021	–1.039	–1.878*	1.152
Turkish	29	27	34	0.887	0.995	0.0187	4.96	1.041	–1.568	–1.922*	1.029
Korean	303	146	52	0.545***	0.974	0.0121	3.20	0.752	–1.768*	–2.178**	–
Biaka Pygmy	17	14	19	1.247	0.971	0.0283	7.50	1.280	1.318	1.219	–
Mbuti Pygmy	20	13	17	1.073	0.942	0.0188	4.99	0.807	0.158	1.463	–
!Kung	26	13	12	0.975	0.874	0.0091	2.40	1.044	–0.794	–1.038	–

N: sample size; *K*: number of different sequences found; *S*: number of variable positions (S_I at the HVI region and S_{II} at the HVII one); *H*: sequence diversity; π: nucleotide diversity; *M*: average number of pairwise difference; M_{II}/M_I: adjusted HVII-to-HVI mean pairwise difference ratio; D_{II}: Tajima's statistic for HVII; D_I: Tajima's statistic for HVI; τ_{II}/τ_I: adjusted HVII-to-HVI τ ratio.[27]
*0.01<*P*<0.05; **0.001<*P*<0.01; ***P<0.001.

additional C in the 303–309 track is between 28% and 69% in the European populations, while the frequency of two additional Cs is between 3% and 14%. Sequences with three additional Cs in the 303–309 track are rare and have only been reported in the British and Biaka Pygmy populations. Track 311–315 almost always presents one additional C. Length heteroplasmy has been detected in five samples (7%) at this homopolymeric stretch. In these cases the predominant mtDNA molecules are reported.

Sequence diversity (0.968) and nucleotide diversity (0.0143) are, in general, slightly lower in the Galician population than in the rest of the European populations (Table 2). Diversity index values for the second hypervariable region show Turks have the highest internal diversity. The distribution observed in the other European samples seems to suggest a gradient of variability from Turkey to Western Europe along the Mediterranean. In fact sequence diversity is significantly correlated with geographic distance from Turkey under a non-parametric test (Spearman's $r = 0.821$, $P < 0.05$) but the meaning of such significance is questionable for the small and irregular differences among samples. This trend is also present, and even more patent, in HVI.[26,42]

Phylogenetic lineage analysis in the Galician population: HVII *vs* HVI

HVII sequences were assigned to haplogroups according to the sequence motifs[43,44] present in HVII and in the previously studied HVI of the same individual[26] (Table 1). Seven sequences could not be firmly assigned to any haplogroup. The most salient features of the haplogroup distribution in the 64 assigned sequences are the relatively high frequencies of H (60.6%) and of haplogroup J (12.7%); the latter is not expected for a haplogroup suggested as having a Neolithic diffusion in Europe,[45] since Galicia was one of the regions with most recent Neolithisation.[46] Haplogroup V is present at a relatively low frequency for an Iberian population (1.4%).[47] Haplogroup H tends to correspond to CRS or to a few unspecific mutations in HVI and to 73A in HVII; the latter is also found in haplogroup V, while all other haplogroups are characterised by 73G. Out of 16 individuals with CRS at HVI, three (19%) bear 73G. And, conversely, out of the 20 individuals bearing clear non-HV motifs at HVI, three had 73A.

The phylogenetic relationships between sequences were analysed using reduced median network.[35] Several known fast-mutating positions (discussed below) were weighted by 1/2 when constructing the network; length polymorphisms were weighted by 1/10, and transversions were given 10 times the weight of transitions. Although some positions were weighted in this way, the network obtained was highly reticulated and difficult to interpret (data not shown). When the haplogroup assignations were depicted on a median network (data not shown), some discrepancies could be observed with respect to the clustering of sequences. For instance, sequences belonging to the same haplogroup were situated in different clusters and even in opposite positions of the median network. It is evident that fast-mutating nucleotide positions introduced a high degree of reticulation (and thus, ambiguity) in the phylogeny blurring the real phylogenetic relationships among sequences in the sample that could be expected from their presumed haplogroup ascription. In this way, several nucleotide positions are found in two different states in many different haplogroup backgrounds. For instance, as can be observed in Table 1, 150T and 152C are both found in sequences presumably belonging to haplogroup H, U5, J, and X. This may be due to recurrent mutation, which adds noise to phylogenetic reconstruction.

To investigate further the effect of fast-mutating nucleotide positions on phylogeny reconstruction, a network has been carried out by dropping from the analysis recently described fast-mutating positions: 146, 150, 152, 182, 189, 195,[17,43] and the extremely high variable length polymorphisms 309a, 309b, and 315a (Figure 1). As shown in Figure 1, sequences belonging to each haplogroup tend now to cluster close. The network displays haplogroup J sequences at one edge of the tree and a central cluster containing mainly haplogroup H sequences, while two haplogroup W sequences also cluster together. As is to be expected, the level of reticulation of the network decreased clearly when the hypervariable positions were removed.

On the other hand, it can be also observed that the two most common sequences in the Galician sample were found to be in the centre of the starlike network (Figure 1). The most common sequence in the sample presents a G at position 263 and an additional C at 315 and it reaches the highest frequency in Galicia (15%), and decline in other East European populations: 12.2% in Tuscans, 11% in Austrian, 3% in Bulgarians and 3% in Turks. A similar gradient was described for CRS in HVI,[26] which is also the most frequent in Galicians. The second most common sequence contains an additional C at 309 and is also quite common across Europe.

Pairwise differences and demographic history

The mean pairwise difference among Galician HVII sequences was 3.99, which was reduced to 3.80 in the 260-bp segment between 63 and 322 used for comparison. The same individuals showed a mean pairwise difference of 3.13 for the 360-bp HVI. That is, corrected for segment length, the mean pairwise difference in HVII was 1.68 times greater than in HVI. Although not so evident, this trend is also present in most populations. Excess pairwise differences (corrected for the difference in sequence length) ranged from 2–4% in Bulgarians, Turks and !Kung (Table 2), to over 20% in British, Biaka Pygmies and Galicians, whilst Mbuti Pygmies and Koreans had larger corrected pairwise differences for HVI.

The number of segregating sites follows the opposite pattern: it is smaller in HVII. Out of the 360 bp sequenced in HVII for Galicians, 40 were found to be polymorphic, and 35

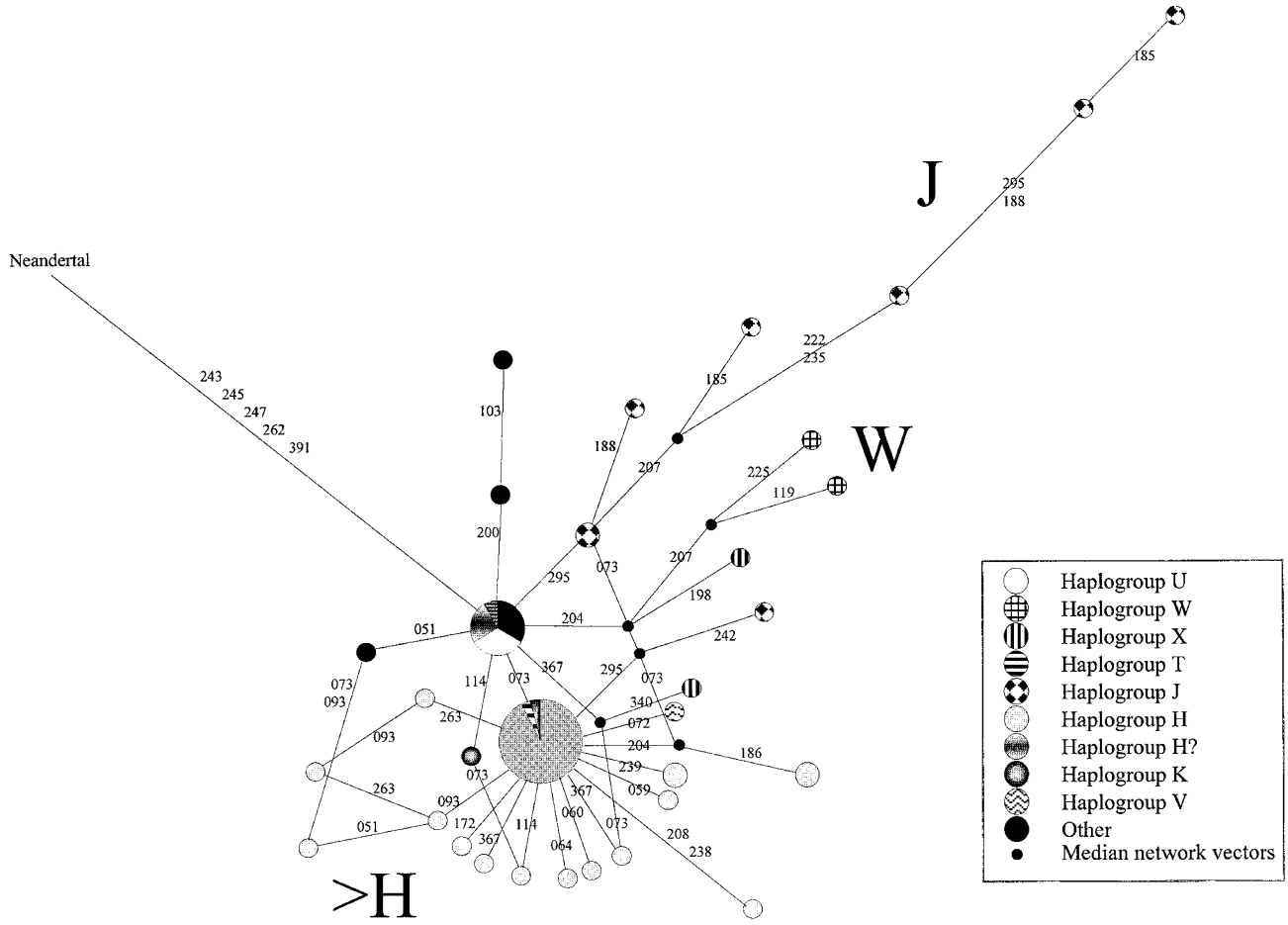

Figure 1 Median network constructed from Galician HVII of mtDNA sequences. Each circle represents a distinct haplotype. The area of the circles is proportional to the frequency of a given sequence and the branches are evolutionary events. The numbers on the branches denote mutations and correspond to base positions in the reference sequences.[32] Reticulations in the network indicate parallel events, which cannot be resolved using this method. The haplogroup designation based on HVI and HVII sequences is superimposed on the network. Empty nodes are introduced by the median algorithm. The Neandertal sequence[56] has been used as an outgroup.

of those were found in the 260 bp stretch used for comparison. Compared to the 57 segregating sites of 360 bp in HVI, HVII shows 15% fewer segregating sites in Galicians. All other European samples also show a deficiency of segregating sites, from 5% in the British to 31% in the Austrian, the latter being statistically significant ($\chi^2 = 4.109$, 1 df, $P = 0.04$). As for the non-European samples, both Pygmy populations showed a slight excess of segregating sites in HVII, while the !Kung and the Koreans matched the European pattern.

The mean number of pairwise differences is an estimator of $\theta = 2n_e\mu$, and it is combined with the number of segregating sites in Tajima's[28] test for mutation-drift equilibrium. The results for the test in HVII and HVI are shown in Table 2; whilst Tajima's D is significantly smaller than zero for HVI in all European and Asian samples, it is significant for HVII in one of six European samples (British) and in the Koreans.

The distribution of nucleotide pairwise differences in the Galicians for HVII shows a unimodal, bell-shaped curve, that fits the distribution expected for a population having undergone an expansion in the past (Figure 2A;[27,33,48]). Similar distributions are obtained for other European populations (Figure 2B), although the modal value is 3 for Galicians, 4 for most other Europeans, and 5 for Tuscans and Turks. The fit to the Rogers and Harpending[27] distribution allows the estimation of τ. For Galicians, $\tau_{II} = 2.549$, while for HVI, the value obtained was $\tau_I = 1.913$. If an expansion is assumed, then the ratio τ_{II}/τ_I, together with the lengths of HVII and HVI can be used to derived μ_{II}/μ_I, which is 1.84 for Galicians. However, this extreme result seems to be a Galician particularity, since other European populations show μ_{II}/μ_I ratios from 0.943 to 1.216, with a median of 1.18. It has been shown that mutation rate heterogeneity may lead to an overestimate of τ,

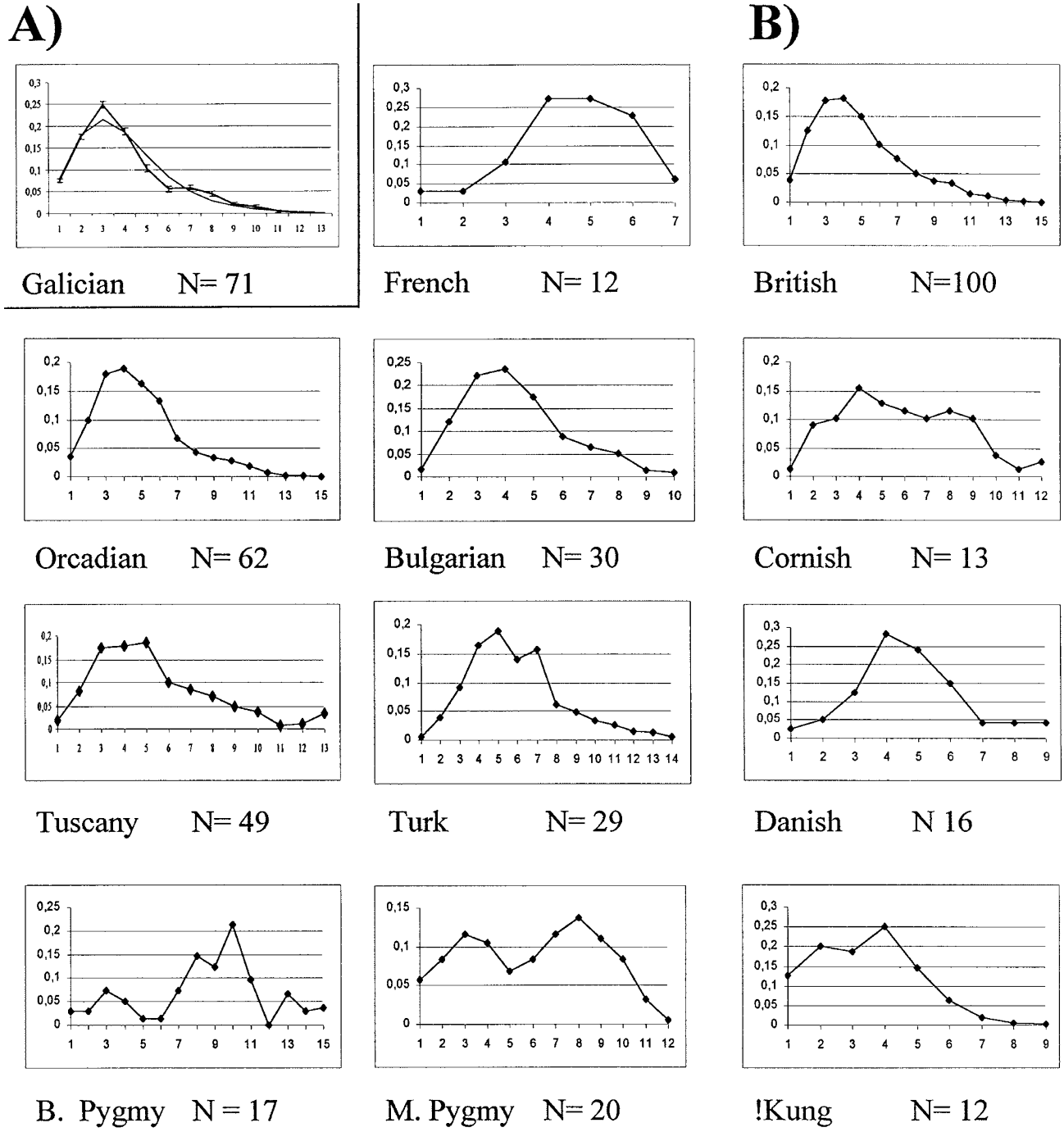

Figure 2 Mismatch distribution in some European and African populations. The abscissa gives the number of differences in a pair, and the ordinate gives the relative frequency number of pairs. As we can see for the Galician population at the top **A**, the distribution is very robust, as shown by the small errors in the different values (standard error was estimated by 1000 bootstrap iterations). The expected distribution for the Galician population is also shown in the same graph with values marked with squares **A**. Nucleotide pairwise differences distribution in the rest of European populations is clearly bell-shaped **B**. African populations show more irregular distributions **B**. N in the figure is the sample size.

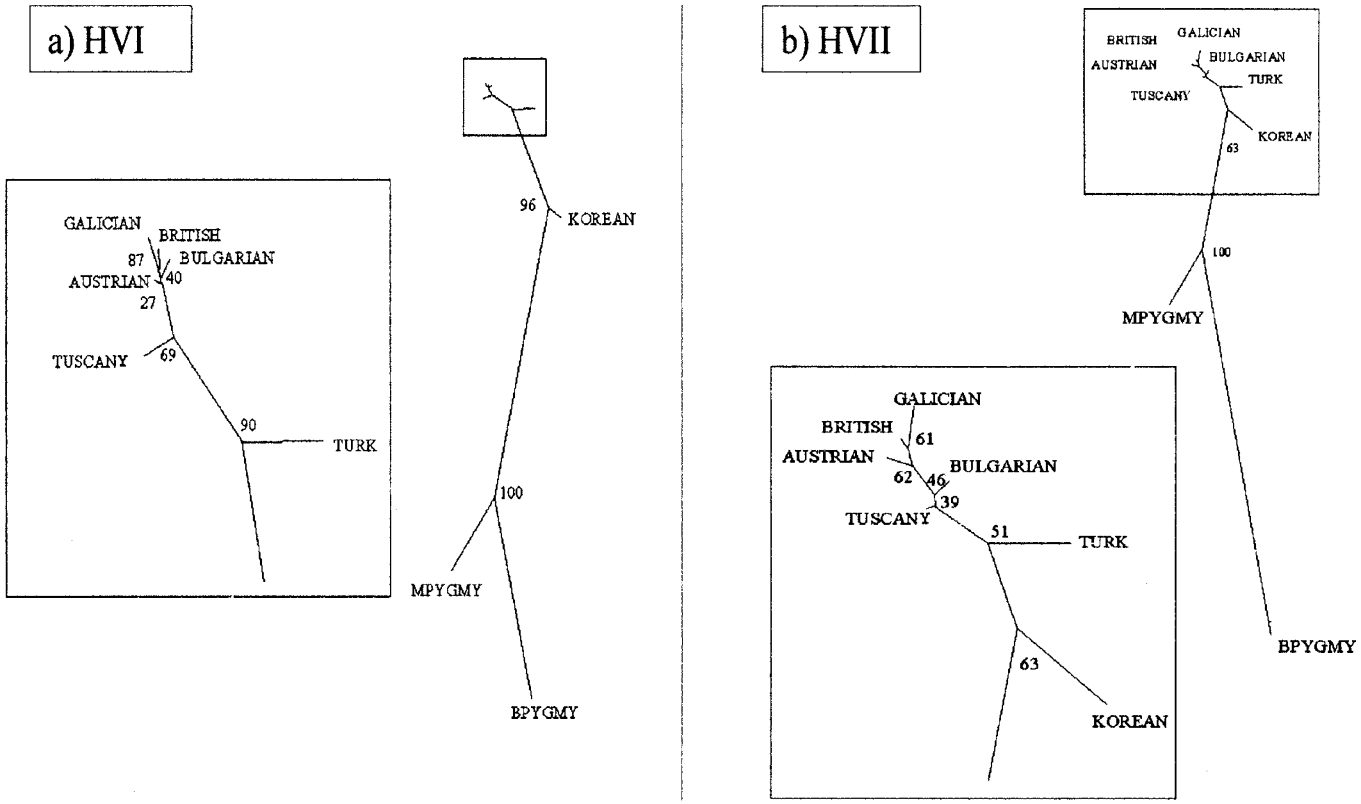

Figure 3 Unrooted neighbour-joining trees for the HV1 **a** and HV2 **b** regions in the same European, Asian and African populations. The numbers at the nodes are bootstrap percentages after 1000 resamplings. The insertions represent a neighbour-joining tree of non-African populations.

although to practically the same extent for the α rate heterogeneity parameters[17] inferred for HVI and HVII.[49] Then, mutation rate heterogeneity does not affect the τ_{II}/τ ratio and, therefore, it seems that HVII mutates, *on average*, slightly faster than HVI.

Genetic distances among populations

Intermatch–mismatch distances between populations were used to obtain unrooted trees (Figure 3) by neighbour-joining.[36]

In Figure 3B, a neighbour-joining tree constructed using the HVII data is shown. Two clusters can be clearly observed: the African populations at one edge, and the European populations at the opposite end. Asians (Koreans) are situated between them, nearest to Europeans. Among European populations, the Galician population is found at the extreme right of the tree. The European cluster in the neighbour-joining tree is very tight, and it is in agreement with the high genetic homogeneity found in these populations. The robustness of the tree is especially notable at the African node, where 100% of the bootstrapped trees are found. The Korean and African population node has a support of 63%, and the support is also high for the cluster linking the Galicians with the British (61%), and for the latter two populations plus the Austrians (62%).

HVI data from the same populations was used to construct a new neighbour-joining tree in order to find out whether the two hypervariable regions display the same phylogenetic pattern. The tree obtained (Figure 3A) shows a very high similarity with that obtained using the HVII data. The robustness of the tree is particularly notable in almost all the clusters. The high concordance in the topology of the HVII and HVI population trees is a result of the correlation between the two distance matrices, which was $r = 0.7989$ ($P = 0.001$; Mantel test with 999 permutations).

Apportionment of genetic variance

The apportionment of genetic variance within and between populations was estimated by means of AMOVA.[29] For HVII in 14 European populations, 0.86% of genetic variation was found to be between populations, whilst the remaining 99.14% was found within populations. A similar apportionment of genetic variation was found for HVI in Europe by Cavalli-Sforza and Minch[50] and by Melton *et al*[51] and, in the samples we have used for reference, the fraction of genetic variance found among populations was 0.31% in HVI. The analysis of both mtDNA regions agrees in that almost all genetic variation for mtDNA is found within populations, in contrast to nuclear loci, in which between-population variation reaches 15%.[52] Geography does not seem to have a clear-

cut impact on mtDNA genetic structure. When populations were grouped into the following regions: the United Kingdom, Iceland, Western Europe (Galicia), Central European populations (France, Denmark, Austria), Tuscany, and Southeastern European population (Bulgaria and Turkey), the genetic variation among these geographic groups was minimal (–0.11%, not significantly different from zero), while genetic variance between populations within these groups was 0.95% of the total.

Discussion

We have analysed the HVII region of mtDNA in a sample of 71 individuals from Galicia, and we have compared the results with other European and non-European sequences, and with the well-established variation patterns of HVI in Europe. The phylogenetic analysis has confirmed that fewer HVII nucleotide positions are informative in the correlation that can be established with haplogroups defined from RFLP variation all around the mtDNA molecule. On the contrary, several nucleotide positions have shown variation in different presumed haplogroup backgrounds. The existence of fast mutation sites in a network introduces such an amount of noise that it blurs the phylogenetic structure given by the slow mutations. In addition, care has to be taken on the use of median network: although it condenses the whole amount of possible paths, it loses possibilities of phylogenetic interpretation, which is its goal. The detailed knowledge of the relative mutation rates for every nucleotide in HVI and HVII is of key importance for phylogenetic inferences since fast-mutating sites should be dropped in order to achieve a meaningful evolutionary picture.

HVII has shown higher mean pairwise differences, and when comparing τ values obtained from HVII and HVI, we have deduced that HVII mutates on average slightly faster than HVI. However, HVII also seems to show fewer segregating sites when compared with HVI. However, it was deduced from an analysis of 1222 different HVI and 385 different HVII sequences, that HVII is seen to mutate slower than HVI.[17] This discrepancy may be partially due to the different length of HVII analysed by us and by Meyer *et al*.[17] Whereas we considered HVII from positions 63 to 322, it was defined by Meyer *et al*[17] to be from 37 to 372. The region from 322 to 372 was found to be almost constant with only three variable nucleotides, only one of those with a higher than average mutation rate.[17] An additional factor in explaining the difference may be that Meyer *et al*[17] used worldwide sequences, while we focused on Europe. The particular haplogroup composition of European populations could contribute to the difference between the two studies. Finally, another important difference between the two studies is that Meyer *et al*[17] does not compare the two regions for each population, whilst our study does. This point is important since to ignore the population effect over the mtDNA background of a specific population could imply an important bias in the estimates of mutation rates for HVI and HVII. To solve this contradiction between the two studies would necessitate further study on this point.

On the whole, these observations are compatible with a higher heterogeneity of mutation rates across HVII, as observed by Aris-Brosou and Excoffier[25] on the basis of a single population sample, the Mandenkalu from Senegal. It seems that some positions, like 146, 150, 152, 182, 189 and 195 mutate fast and, in fact, Howell *et al*[21] detected mutations at some of these positions in mother–children transmissions. Mutation in these sites can contribute disproportionately to the overall mutation rate of HVII, which has been observed to be higher in genealogical studies.[21,23] It is expected that sites with low and intermediate mutation rates contribute to the correlation with mutations in other regions of mtDNA; a relative dearth of such sites may explain why there are fewer haplogroup-specific sequence motifs in HVII than in HVI.

Mutation rate heterogeneity, therefore, should be factored in when interpreting HVII sequence results. The overall high mutation rate has created sequence diversity in HVII, which is slightly higher than that in HVI. However, mutation rate is not so high as to obscure recent demographic processes that affect sequence diversity. For instance, a loss of genetic diversity due to a recent founder effect would be as visible in HVII sequences as in HVI. This shared pattern of sequence diversity may also explain the similar (and low) percentages of between-population genetic diversity in Europe detected by AMOVA, which may have been produced by a higher female migration rate.[53] HVII sequences also reflect some of the gradients described for HVI in Europe, such as the westward loss of sequence diversity and the higher frequency in Western Europe of sequences belonging to haplogroup H. For HVI, it has been suggested that the Palaeolithic and/or Neolithic expansions could be responsible for such patterns.[26,45,47,54] A slightly higher mutation rate and higher mutation rate heterogeneity do not seem to have concealed the effects of those demographic processes on HVII.

The high mutation rate of some positions generated high mean pairwise differences in HVII. However, since this is a property of HVII rather than of specific populations, the genetic distance based on the corrected interpopulation pairwise differences may be still a good measure of the differences between populations. The high correlation we observed between the HVII- and the HVI-based distance matrices implies that such distance is as reliable a tool in HVII as it is in HVI.

The deepest impact of mutation rate heterogeneity would be on Tajima's D. As Bertorelle and Slatkin[55] suggested, mutation rate heterogeneity will tend to increase Tajima's D statistic, even to suggest mutation-drift equilibrium when there is not. Usually, a significantly negative Tajima's D statistic (such as that found in European populations for HVI[54]) is interpreted as the result of population expansion or of positive selection and hitchhiking. It seems obvious that,

if any of these two factors affected HVI, the effects should also reach HVII, and we have observed non-significant (although negative) Tajima's D statistics in European HVII sequences. Such as Aris-Brosou and Excoffier[25] suggest, heterogeneity of mutation rates in combination with population expansion give slightly negative but non-significant D values. Therefore, mutation rate heterogeneity has contributed noise to the signals left on HVII by population expansion (or by selection if relevant), and any single value of Tajima's D for HVII should be interpreted with caution.

In summary, although some additional caution should be taken in some cases, we have shown that HVII sequences can be almost as valuable a tool as HVI sequences when used for human evolutionary inferences.

Acknowledgements

The authors would like to thank Dr Mark Miller for providing the TFPGA software. We would like to express our special thanks to two anonymous and referees one explicit for helpful comments on the manuscript. This work was supported in part by the grants from the Xunta de Galicia (XUGA 20816B96 and XUGA 20806B97) and by DGICT grant (PB95-0267-C02-01). AS received a postdoctoral fellowship from the Diputación de A Coruña. FC was supported by a postdoctoral return contract from CICYT (Spanish Government). The technical assistance of M Rodríguez is also acknowledged with appreciation.

References

1 Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC: African populations and the evolution of mitochondrial DNA. *Science* 1991; **253**: 1503–1507.
2 Cann RL, Stoneking M, Wilson AC: Mitochondrial DNA and human evolution. *Nature* 1987; **325**: 31–36.
3 Jorde LB, Bamshad MJ, Watkins WS *et al*: Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* 1985; **57**: 523–538.
4 Ward RH, Frazier BI, Dew-Jager K, Pääbo S: Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci USA* 1991; **88**: 8720–8724.
5 Bertranpetit J, Sala J, Calafell F: Human mtDNA variation and the origin of Basques. *Ann Hum Genet* 1995; **59**: 63–81.
6 Mountain JL, Herbert JM, Bhattacharyya S *et al*: Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet* 1995; **56**: 979–992.
7 Giles RE, Blanc H, Cann HM, Wallace DC: Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci* 1980; **77**: 6715–6719.
8 Miyata TH, Hayashida R, Kikuno M *et al*: Molecular clock of silent substitution: at least six fold preponderance of silent changes in mitochondrial genes over those in nuclear genes. *J Mol Evol* 1982; **9**: 28–35.
9 Brown WM, George MJ, Wilson AC: Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 1979; **76**: 1967–1971.
10 Piercy R, Sullivan KM, Benson N, Gill P: The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int J Leg Med* 1996; **106**: 85–90.
11 Miller KWP, Dawson JL, Hagelberg E: A concordance of nucleotide substitutions in the first and second hypervariable segment of the human mtDNA control region. *Int J Legal Med* 1996; **109**: 107–113.
12 Francalacci P, Bertranpetit J, Calafell F, Underhill PA: Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phys Anthrop* 1996; **100**: 443–460.
13 Parson W, Parsons TJ, Scheithauer R, Holland MM: Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: Application of mtDNA sequence analysis to a forensic case. *Int J Legal Med* 1998; **111**: 124–132.
14 Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L: From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 1996; **60**: 35–49.
15 Lee SD, Shin CH, Kim KB, Lee YS, Lee JB: Sequence variation of mitochondrial DNA control region in Koreans. *Forensic Sci Int* 1997; **87**: 99–116.
16 Vigilant L: Control region sequences from African populations and the evolution of human mitochondrial DNA. PhD Thesis, University of California at Berkeley, 1990.
17 Meyer S, Weiss G, von Haeseler A: Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 1999; **152**: 1103–1110.
18 Crews S, Ojala D, Posakonoy J, Nishiguchi J, Attardi G: Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication. *Nature* 1979; **277**: 192–198.
19 Fisher RP, Topper JN, Clayton DA: Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. *Cell* 1987; **50**: 247–258.
20 Chang DD, Clayton DA: Priming of human mitochondrial DNA replication occurs at the light-strand promoter. *Proc Natl Acad Sci USA* 1985; **82**: 351–355.
21 Howell N, Kubacka I, Mackey DA: How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet* 1996; **59**: 501–509.
22 Macaulay VA, Richards MB, Forster P *et al*: mtDNA mutation rates – no need to panic. *Am J Hum Genet* 1997; **61**: 983–990.
23 Parsons TJ, Muniec DS, Sullivan K *et al*: A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet* 1997; **15**: 363–368.
24 Jazin E, Soodyall H, Jalonen P, Lindholm E, Stoneking M, Gyllensten U: Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nature* 1998; **18**: 109–110.
25 Aris-Brosou S, Excoffier L: The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol Biol Evol* 1996; **13**: 494–504.
26 Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A: mtDNA Analysis of the Galician population: A genetic edge of European variation. *Eur J Hum Genet* 1998; **6**: 365–375.
27 Rogers AR, Harpending H: Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 1992; **9**: 552–569.
28 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
29 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.
30 Wilson MR, Stoneking M, Holland MM, Di Rienzo JA, Budowle B: Guidelines for the use of mitochondrial DNA sequencing in forensic science. *Crime Lab Digest* 1993; **20**: 68–77.
31 Thompson JD, Higgins DG, Gibson TJ, Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; **22**: 4673–4680.
32 Anderson S, Bankier AT, Barrell BG *et al*: Sequence and organisation of the human mitochondrial genome. *Nature* 1981; **290**: 457–465.

33 Harpending HC, Sheny ST, Rogers AR, Stoneking M: The genetic structure of ancient human populations. *Curr Anthrop* 1993; **34**: 483–496.
34 Efron B: *The Jacknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics: Philadelphia, 1982.
35 Bandelt HJ, Forster P, Sykes BC, Richards MD: Mitochondrial portraits of human populations using median networks. *Genetics* 1995; **141**: 743–753.
36 Saitou N, Nei M: The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406.
37 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 1989; **5**: 164–166.
38 Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27**: 209–220.
39 Smouse PE, Lang JC, Sokal RR: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Sys Zool* 1986; **35**: 627–632.
40 Miller MP: TFPGA: Tools for Population Genetics Analysis. Version 4.00.950, 1995.
41 Schneider S, Kueffer JM, Roessli D, Excoffier L: ARLEQUIN software V 1.0, 1997.
42 Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Bertranpetit J: Mitochondrial DNA variation and the origin of the Europeans. *Hum Genet* 1997; **99**: 443–449.
43 Torroni A, Huoponen K, Francalacci P *et al*: Classification of European mtDNAs from an analysis of three European populations. *Genetics* 1996; **144**: 1835–1850.
44 Macaulay V, Richards M, Hickey E *et al*: The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* 1999; **64**: 232–249.
45 Richards MB, Macaulay VA, Bandelt HJ, Sykes BC: Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* 1998; **62**: 241–260.
46 Arias P: *De cazadores a campesinos. La transición del neolítico en la región cantábrica*. Universidad de Cantabria: Santander (Spain), 1991.
47 Torroni A, Bandelt HJ, D'Urbano L *et al*: mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 1998; **62**: 1137–1152.
48 Slatkin M, Hudson RR: Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 1991; **129**: 555–562.
49 Schneider S, Excoffier L: Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 1999; **152**: 1079–1089.
50 Cavalli-Sforza LL, Minch E: Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 1997; **61**: 247–254.
51 Melton T, Wilson M, Batzer M, Stoneking M: Extent of heterogeneity in mitochondrial DNA of European populations. *J Forensic Sci* 1997; **42**: 437–446.
52 Barbujani G, Stenico M, Excoffier L, Nigro L: Mitochondrial DNA sequence variation across linguistic and geographic boundaries in Italy. *Hum Biol* 1996; **68**: 201–215.
53 Seielstad MT, Minch E, Cavalli-Sforza LL: Genetic evidence for a higher female migration rate in humans. *Nat Genet* 1998; **20**: 278–280.
54 Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bertranpetit J: Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol Biol Evol* 1996; **13**: 1067–1077.
55 Bertorelle G, Slatkin M: The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Mol Biol Evol* 1995; **12**: 887–892.
56 Krings M, Geisert H, Schmitz RW, Krainitzki, Pääbo S: DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc Natl Acad Sci USA* 1999; **96**: 5581–5585.