European Journal of Human Genetics (1998) 6, 365–375

http://www.stockton-press.co.uk/ejhg

ORIGINAL PAPER

mtDNA analysis of the Galician population: a genetic edge of European variation

Antonio Salas[1], David Comas[2], María Victoria Lareu[1], Jaume Bertranpetit[2] and Angel Carracedo[1]

[1]*Institute of Legal Medicine, University of Santiago de Compostela, Spain*
[2]*Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Spain*

Analysis of mitochondrial DNA (mtDNA) variation has become a useful tool for human population studies. We analysed the first hypervariable region of mitochondrial DNA control region (position 16024–16383) in 92 unrelated individuals from Galicia (Spain), a relatively isolated European population at the westernmost continental edge. Fifty different sequences defined by 56 variable positions were found. The frequency of the reference sequence reaches in Galicians its maximum value in Europe. Moreover, several genetic indexes confirm the low variability of our sample in comparison to data from 11 European and Middle Eastern populations. A parsimony tree of the sequences reveals a high simplicity of the tree, with few and small well defined clusters. These results place Galicians on the genetic edge of the European variation, bringing together all the traits of a cul-de-sac **population with a striking similarity to the Basque population. The present results are fully compatible with a population expansion model in Europe during the Upper Paleolithic age. The genetic evidence revealed by the analysis of mtDNA shows the Galician population at the edge of a demographic expansion towards Europe from the Middle East.**

Keywords: mtDNA; control region; Galicians; neighbour joining tree; pairwise difference distribution

Introduction

Much has been learnt about human population history and evolution through genetic analysis, Europe being the most comprehensively studied area in the world. Until recently, most of the information came from what are known as 'classical genetic marker' studies, where the analyses were based on geographic variation of allele frequencies for expressed genetic polymorphisms. When genetic information has been widely available for a given geographic region and data has been properly handled (be it through genetic distances, principal components or other numerical tools) it has been possible to make interesting and innovative proposals about the knowledge of our past. The exhaustive compilation of Cavalli-Sforza *et al*[1] stresses the long standing interpretation of the genetic variation in Europe as being primarily shaped by the demographic impact of the Neolithic expansion. Besides this principal general pattern, some populations have shown genetic peculiarities (in the sense that they show clear

Correspondence: J Bertranpetit, Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain. Tel: 34 93 4021461; Fax: 34 93 4110887; E-mail: jaumeb@porthos.bio.ub.es
Received 27 May 1997; revised 19 January 1998; accepted 26 January 1998

differentiation from neighbouring populations) which can be understood in terms of differences in allele frequencies due to specific events of genetic drift in their history. This is the case for the Basques, Sardinians, Finns, Saami, Icelanders and several populations of the Caucasus. As drift is the main factor for the genetic differences of these populations in relation to their geographic neighbours, allele frequency comparisons may not inform us about the origin of a given population. Rather, they refer to the last drift episode (bottleneck, founder effect) which has taken place in population history, ie the last episode of isolation with a small sample size. However, the genetic origin of the population may be much older. This is what seems to be revealed in the analysis of other genomic regions, mostly DNA sequences. The amount of variation in sequences may be maintained through a bottleneck due to the fact that most genetic lineages (or representative of most phylogenetic branches) may survive through dramatic changes in population size while other sequences may be lost. Although some sequence studies exist for autosomic regions (see Ref 2 for important methodological developments), mitochondrial DNA (mtDNA) has been largely the molecule of choice for most of the recent studies, especially in Europe. Three partial and somewhat contradictory syntheses have been published recently on mtDNA variation and the origin of Europeans.[3–5]

The present analysis tries to shed light on the general European genetic variation view by studying the westernmost European geographical edge (the Roman *Finisterrae*, the end of the Earth) through the analysis of the mtDNA hypervariable region I of the Galician population. Galicia is a region situated in the northwest extreme of the Iberian Peninsula (Figure 1) with specific characteristics. Its geographical isolation from the rest of the Iberian Peninsula, in addition to its migratory patterns (with high emigration rates throughout centuries and almost no immigration), has preserved its identity, language, economy and especially its own cultural identity.

Within the Iberian Peninsula not only Galicia but also the rest of the Cantabrian Coast showed a specific pattern in prehistoric times. During Paleolithic and Mesolithic times, all the Cantabrian region was a very homogenous area, differentiated from the rest of Iberia. Ecological and nutritional resources were similar along all the North Iberian coast, richer than in the rest of the Peninsula. The Basques, outliers in the genetic landscape of Europe, as shown by classical genetic markers,[1,6] are also located in this Cantabrian area. Moreover, Galicia seems to have been influenced by Atlantic cultures mainly during the first millennium BC[7,8] and Celtic influences are still found in some cultural traits.

The proximity of the present population to the genetically well-known Basques and its peripheral position in European geography make Galicia especially interesting for the understanding of population affinities, history and for refining the geography of genetic variation in Europe.

Material and Methods

Population Samples

Blood samples were obtained from 92 healthy autochthonous unrelated individuals scattered through Galicia. DNA was extracted using a phenol–chloroform procedure[9] and quantified using a Hoefer DyNA Quant 200 Fluoremeter (Pharmacia, Uppsala, Sweden).

mtDNA Amplification and Sequencing

A semi-nested PCR was performed in order to amplify the first hypervariable segment of the mtDNA control region. The first amplification was carried out in a Perkin Elmer 480 A Thermocyler (Perkin Elmer, Foster City, CA, USA) using

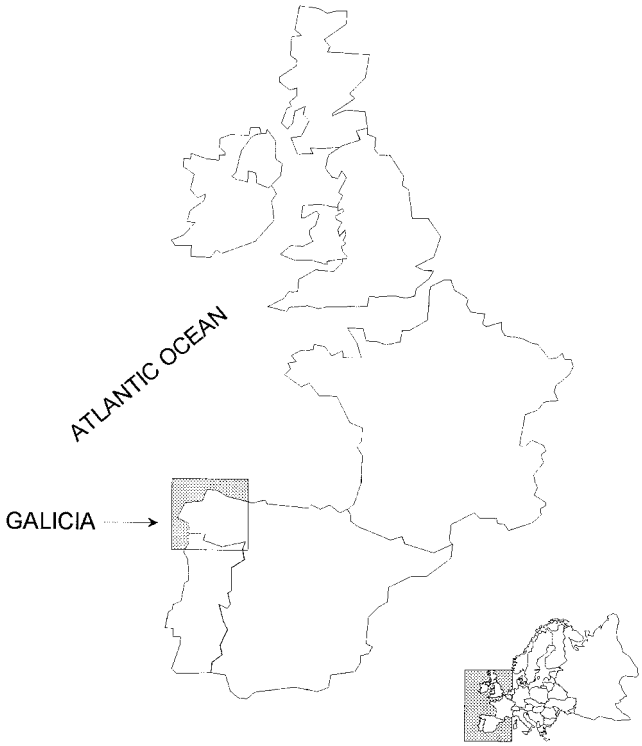

Figure 1 *Location of Galacian population in the western corner of Europe*

5 ng of DNA in 25 µl reaction volume, following a temperature profile for 32 cycles of amplification at 95°C for 1 min, 55°C for 1 min, and 72°C for 1 min. A segment of 1021 base pairs was amplified using the L15997 (5'-CACCATTAGCACCCAAAGCT-3')[10] and H408 (5'-CTGTTAAAAGTGCATACCGCCA-3')[11] primers. The nomenclature of the primers refers to the light and heavy chains of the mtDNA (L or H), and the numbers identify the position of the primer 3' ends in the reference Cambridge sequence.[12] The second PCR amplification was performed using primers L15997[10] and H16401 (5'-TGATTTCACGGAAGGATGGTG-3'),[11] which amplified a fragment of 443 bp, with a temperature profile for 32 cycles of amplification of 95°C for 1 min, 60°C for 1 min, and 72°C for 1 min. Positive and negative controls were checked in the PCR amplifications in order to detect possible contamination.

The PCR products were purified with MICROSPIN™ HR S-300 columns (Pharmacia Biotech, Uppsala, Sweden) before the cycle sequencing. The sequence reactions were carried out using the PCR Fentomol Sequencing Kit (Promega, Madison, USA) with 100 ng of template DNA and 0.5 µM of fluorescently labelled sequencing primers (L15997 and H16401). The sequencing profile for 10 cycles was, 95°C for 30 s, 55°C for 30 s and 70° for 90 s, followed by an extension cycle at 72°C for 5 min.

The sequence products were denatured with deionizide formamide and run in a 6% PAGE gel, and analysed in an ALF automatic sequencer (Pharmacia, Uppsala, Sweden).

Site 73 of the second hypervariable region has also been tested in 71 individuals under a more exhaustive analysis of mtDNA sequence (Salas *et al*, manuscript in preparation).

Computer Analysis

The alignment of the sequences obtained was performed using the CLUSTAL W (1.5) Multiple Sequence Alignment program.[13] The final information for each individual was a string of 360 characters belonging to the mtDNA hypervariable region I (HVI), from base position (16 024 to 16 383.[12] Sequences are available by e-mail on request to apimlase@uscmail.usc.es. For most calculations, the standard package PHYLIP 3.5c[14] was used, and some programs were specifically written.

To test the internal diversity of the sample, several parameters were computed. Nucleotide diversity[15] was estimated as $(n/n\text{-}1)\ (1/l)\ \Sigma_{i=1}^{l}\ (1\text{-}x_i^2)$, where n is sample size, l is sequence length (360, in the present study) and x_i is the frequency of each nucleotide at position i. Similarly, sequence diversity was estimated as $(n/n\text{-}1)\ \Sigma_{i=1}^{k}\ (1\text{-}p_i^2)$, where p_i is the frequency of each of the k different sequences in the sample. Finally, Shannon's measure of information H, defined as $H = -\Sigma\ p_i \log_2 p_i$ (where p_i is the sample frequency of the ith sequence), and H' (the ratio of H to its maximum value for a given sample size: $-\log_2 (1/n)$, where n is the sample size),[16] were calculated in order to measure the genetic diversity of the sample.

Pairwise difference distribution was computed, and the τ parameter from the two-parameter model of Harpending *et al*[17] was obtained. Standard errors were computed from 1000 bootstrap iterations: resampled sequences of the same length (360 characters) were obtained by sampling sites with replacement.

Data from different populations were used for comparison: 106 Basques,[18,19] 92 Welsh,[4] 54 Portuguese,[19] 69 Cornish,[4] 49 Bavarians,[4] 108 northern Germans,[4] 100 British,[20] 89 Spanish,[19,21] 49 Tuscans,[22] 96 Turks,[4,23,24] and 42 Middle Easteners.[25] A genetic distance matrix between populations was obtained by using the intermatch–mismatch distance: $D = d_{ij} - (d_{ii} + d_{jj})/2$, where d_{ij} is the mean number of intermatches between populations i and j, and d_{ii} and d_{jj} are the mean pairwise differences (mismatches) within populations i and j. This expression is known as the Jensen difference and was defined by Rao.[26] It is related to the pairwise difference distributions, which have been studied and modelled intensively.[17,27] Standard errors were estimated by bootstrap.[28] Neighbour-joining trees were built from the distance matrix using the options NEIGHBOUR and DRAWTREE in the PHYLIP package and the robustness of the clusters found was estimated by bootstrap[29] using the option SEQBOOT in the same package.

Results

Sequence diversity

A total of 53 different sequences defined by 56 variable positions were found (Table 1). All the polymorphisms observed in the sample were nucleotide substitutions, except for a deletion found in one individual of one A in the run of three As which goes from position 16 349 to 16 351. Of the 55 remaining variable positions, 50 were transitions, three were transversions and two presented both types of substitutions (positions 16 093 and 16 114). The variable sites present the following pattern: 41 T<–>C, 11 A<–>G, 2 T<–>G, 2 A<–>C, 1 A<–>T; 11.6:1 being the ratio between transitions and transversions. There is a clear bias in the proportion of the pyrimidine transitions (found 117 times across all sequences) with respect to purine transitions (found 24 times in all sequences). Nevertheless, this proportion is common in a large worldwide sequence set.[30] For all the positions, the most frequent nucleotide is that shown by the reference sequence, and only two positions present high levels of polymorphism: position 16 126 (T in the reference sequence) with a C in 11 individuals (12%), and position 16 311 (T in the reference sequence) with a C in 10 individuals (10.9%) in the sample.

It is interesting to note that 24 individuals presented the reference sequence,[12] the highest frequency (26.1%) observed in the whole set of populations used for comparison. Some of the remaining sequences were shared by a few individuals: three different sequences were found three times (9.8%), ten sequences twice (21.7%) and 39 individuals had unique sequences (42.4%); this represents a low level of variability within the sample and, phylogenetically, all sequences are closely related.

Table 1 mt DNA sequences of the hypervariable region HV1 found in 92 Galician individuals, compared to Ref. 12. Dots indicate identity with the reference sequence, and deletion is represented by a dash. N, indicates the number of individuals. Sequences are sorted according to the groups suggested by Richards *et al*.[4] The two last columns show the number of individuals (a total of 71) having an A or a G in position 73 of the hypervariables region HV2

	16025	16054	16069	16072	16086	16093	16094	16114	16124	16126	16129	16145	16162	16163	16164	16169	16172	16184	16189	16192	16193	16195	16207	16213	16216	16217	16218	16219	16221	16222	16223	16224	16234	16235	16239	16240	16241	16242	16243	16256	16261	16270	16278	16287	16288	16291	16292	16293	16294	16295	16296	16298	16304	16311	16319	16356	16362	N	73	
ANDERSON	T	A	C	C	T	T	T	C	T	T	G	G	A	A	A	C	T	C	T	C	C	T	A	G	A	T	C	A	C	C	C	T	C	A	C	A	A	C	T	C	C	C	C	C	T	C	C	A	C	C	C	T	T	T	A	T	T		A	G
SEQ1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	24	13	3
SEQ2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	2	1
SEQ3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	1	1	0
SEQ4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ5	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ6	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ7	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	–	–
SEQ10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	1	1	0
SEQ13	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	2	0
SEQ14	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	–	–
SEQ15	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ17	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	1	–	–
SEQ19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	1	1	0
SEQ20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ21	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	2	1	–
SEQ23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	1	–	–
SEQ24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	–	1
SEQ27	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	G	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ28	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	2	2	0
SEQ29	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	1	0	1
SEQ30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	1	–	–
SEQ31	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	3	0
SEQ32	.	.	.	.	.	G	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0
SEQ40	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	2	–	1
SEQ41	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	C	.	.	.	2	2	0
SEQ44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	C	.	.	.	1	1	0
SEQ46	.	.	.	.	.	.	.	.	.	.	A	.	.	G	T	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	1	–	–
SEQ47	.	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	0	3
SEQ48	.	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	2	0	2
SEQ49	G	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	1	0	1
SEQ50	.	.	T	.	.	.	.	.	.	C	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	1

Table 1 *Continued*

SEQ51	. . T . . . . . . C . A . . . . C . . . . T . . . . T	1	1	0
SEQ52	. . . . . . C . . C C	1	1	0
SEQ53	. . . . C . . . . T . T . C	2	1	–
SEQ33	. . . . A . . A . . . . A . . . . . . T . T . . . . T	1	0	1
SEQ34	. . . . T . . . . T . . T	1	0	1
SEQ35	. . . . . C . . C . . . . T . . . . T C	1	0	1
SEQ36	. . . . T . . . . T C . . . . T . . . . C . T	1	0	1
SEQ45	. . . . . . C . . . . T . . . . T . . . . . . C . . .	1	0	1
SEQ42	. . . . C . . . . C . . .	2	1	–
SEQ43	. . . . . C . . . . C . . . . C . . .	1	0	1
SEQ37	. . . . C T . . . . T	1	–	–
SEQ38	. . . . . C . . . . C . C T . . . . T	1	0	1
SEQ39	. . . . T . . . . T . T C	2	0	1

Sequence diversity (0.9295), nucleotide diversity (0.0087) and Shannon's index H' (0.865) shown in Table 2, present very low values in comparison to other European populations, reflecting again the phylogenetic simplicity in the mtDNA variation. The values obtained for these parameters show a clinical gradient from West Asia to Western Europe, with Basques and Galicians being the populations with the lowest diversity levels.

Tree of Sequences: Maximum Parsimony

A maximum parsimony tree was built in order to observe the phylogenetic structure in the sampled sequences. Among all the possible maximum parsimonious trees, we represent in Figure 2 one which has a maximum number of sequences stemming from the most frequent substitutions (16 126 and 16 311).[31] Dark squares represent the intermediate steps not found in the present sample. In some cases more than one most parsimonious path leading to a sequence is shown. In this case some of the groups of sequences described by Richards *et al*[4] can be easily recognised. Nevertheless, we do not find strong support for any of the possible paths that may have generated these specific sequences.

Most of the sequences in the tree stem directly from the reference sequence, with a few further mutations leading to other lineages. This high frequency of the reference sequence and the low number of substitutions that link it to many of the other lineages in Galicia is consistent with the clinal variation observed in Europe with higher values in Western Europe.[31]

Despite the simplicity of the tree and the lack of a robust structure, some clusters should be mentioned:

(a) the C in position 16 126, which defines a small cluster of sequences in the tree, has been found to be polymorphic in European and West Asian populations and its frequency is clinal along Europe: from 50% in the Middle East to 7.5% in Basques. In the Galician population this polymorphism is present in 12% of the individuals sampled. This cluster of sequences corresponds to group 2 according to Richards *et al*,[4]

(b) The C at position 16 311 has a worldwide distribution with a high frequency in African populations. Its frequency in European populations is estimated to be around 20%[31] and its frequency in Galicia is 10.9%. No clear geographic patterns

Table 2 Sequence divergence. Population sources: Galician (present study), Basque,[18, 19] Welsh,[4] Portuguese,[19] Cornish,[4] Bavarian,[4] Northern German,[4] British,[20] Spanish,[19, 21] Tuscan,[22] Turk,[4, 23, 24] Middle Easters.[25] All values in all populations are based on the analysis of a fragment of 360 nucleotides, from 16024 to 16383[12]

	N	*K*	*A*	*B*	*C*	π	*J*	*H′*
Galician	92	53	56	3.13	0.869	0.0087	0.9295	0.865
Basque	106	52	52	2.95	0.819	0.0081	0.9362	0.863
Welsh	92	48	51	3.39	0.942	0.0094	0.9307	0.862
Portuguese	54	38	46	3.60	1.000	0.0100	0.9343	0.887
Cornish	69	45	52	3.89	1.080	0.0108	0.9650	0.925
Bavarian	49	37	37	3.94	1.094	0.0109	0.9838	0.967
Northern German	108	70	61	3.92	1.089	0.0109	0.9733	0.917
British	100	71	67	4.45	1.236	0.0123	0.9760	0.930
Spanish	89	70	69	5.02	1.394	0.0139	0.9834	0.952
Tuscan	49	40	55	5.03	1.397	0.0140	0.9685	0.938
Turk	96	79	82	5.45	1.514	0.0151	0.9879	0.961
Middle East	42	38	59	7.08	1.966	0.0197	0.9954	0.991

N: sample size. *K:* number of different sequences found. *A:* number of variable nucleotides positions. *B:* mean nucleotide pairwise differences. *C:* percentage average pairwise difference per nucleotide. π: nucleotide diversity. *J:* sequence diversity. *H′:* ratio of H (Shannon's index) to H_{max} (maximum value of H for a given sample size).

have been observed. It is not clear that it is a robust cluster and sequences with this substitution may have arisen through alternative ways.

Most of the 53 different sequences present a single change compared to the reference sequence, whereas only four differ in six positions (Figure 3). If we

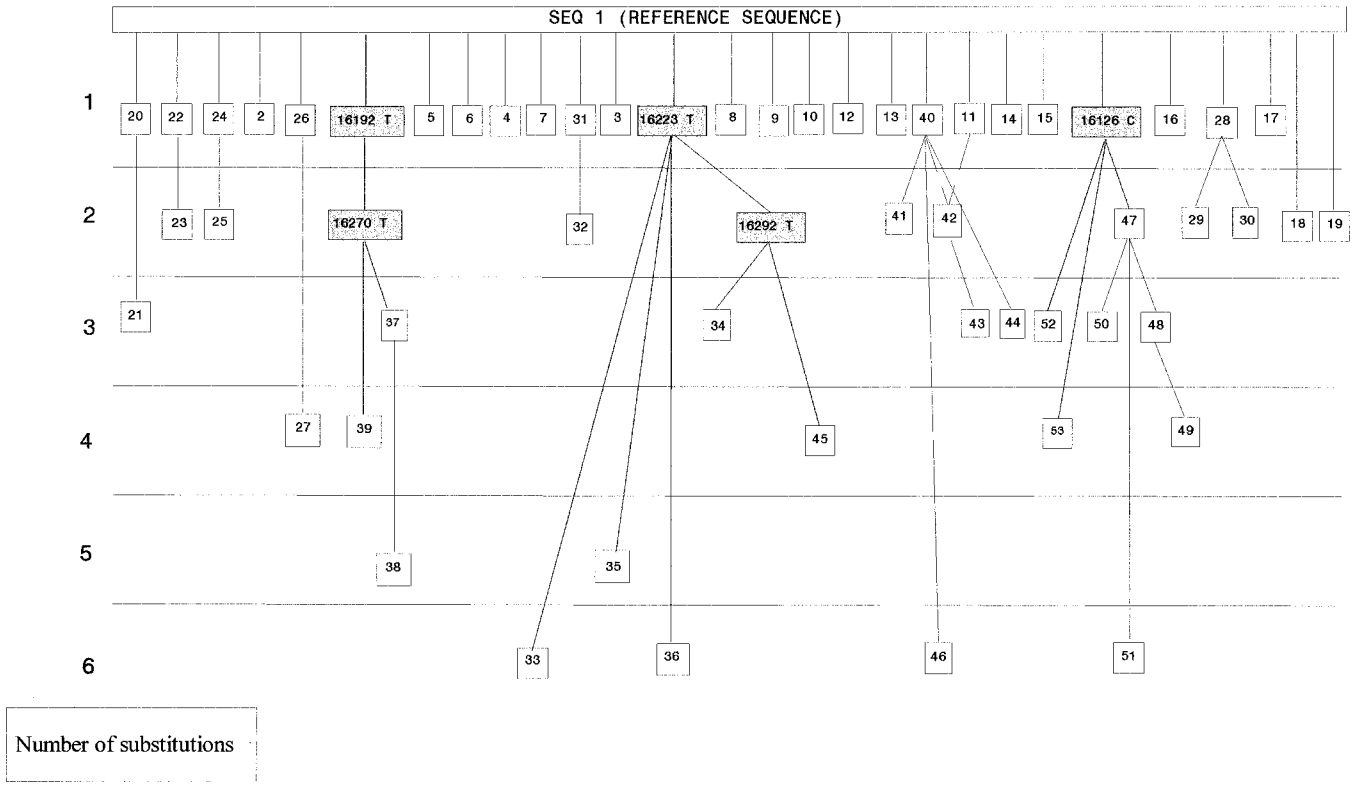

Figure 2 *Maximum-parsimony tree of the sequences, where the reference sequence represents the ancestral sequence. The rest of sequences could have evolved from this one. Numbers at the left indicate differences to the reference sequence*

consider the possibility that the reference sequence is the ancestor of all the sequences, and assume that mutations accumulate in a Poisson process,[32] the number of mutations relative to the reference would follow a Poisson distribution. From the sequences stemming from the reference by a single mutation, sequences have accumulated a further mean $\lambda = 1.33$ mutations. As shown in Figure 3, Galician sequences do not fit a Poisson distribution with such a high λ value ($\chi^2 = 17.83$, d.f. = 5, $P = 0.003$), having a clear over-representation of sequences with several substitutions. This could easily be understood if some variation had already existed in the founders of the Galician population. This fact is also supported by the sharing of Galician sequences (SEQ22, SEQ28, SEQ31, SEQ40, SEQ42, SEQ43 and SEQ47) by other European populations used for comparison. Nevertheless, the high frequency of the reference sequence in the present population is probably due to its high frequency in the founding population.

This lack of a branched structure and the star-like phylogeny found are compatible with a recent expansion of the Galician population[33] and therefore, except for the deeper clusters found, most of the new haplotypes should have been produced *in situ* in recent times.

Pairwise Differences

The mean pairwise difference in Galicians is 3.13, a value slightly higher than the value found in Basques and lower than the rest of European and West Asian populations (Table 2). This value is merely the result of the low diversity of the Galician population, its lack of a complex tree structure and the phylogenetic proximity of its sequences. As shown in Table 2, a clinal decrease of this value is patent from the Middle East to the Basques and Galicians. This decrease towards Western Europe is compatible with an ancient expansion from the Middle East to the Atlantic coast, reaching Galicia, one of the last edges of the European continent, in the last steps of the putative expansion. Other populations in the European far west, such as the Welsh, also fit in this pattern.

The Galician pairwise difference distribution, Figure 4(A), is clearly bell-shaped as expected in populations which have experienced a sudden expansion[27] with a peak at only two differences. This empirical distribution is very robust, as shown by the small errors of the different values estimated by 1000 bootstrap

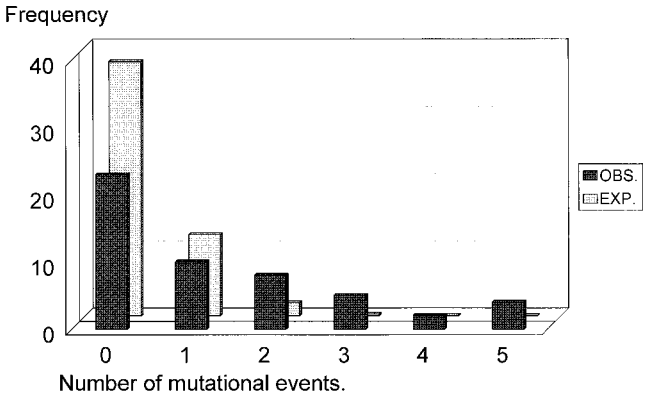

Figure 3 *Expected and observed distributions of the number of mutational events occurred from the reference sequence in the Galician population*

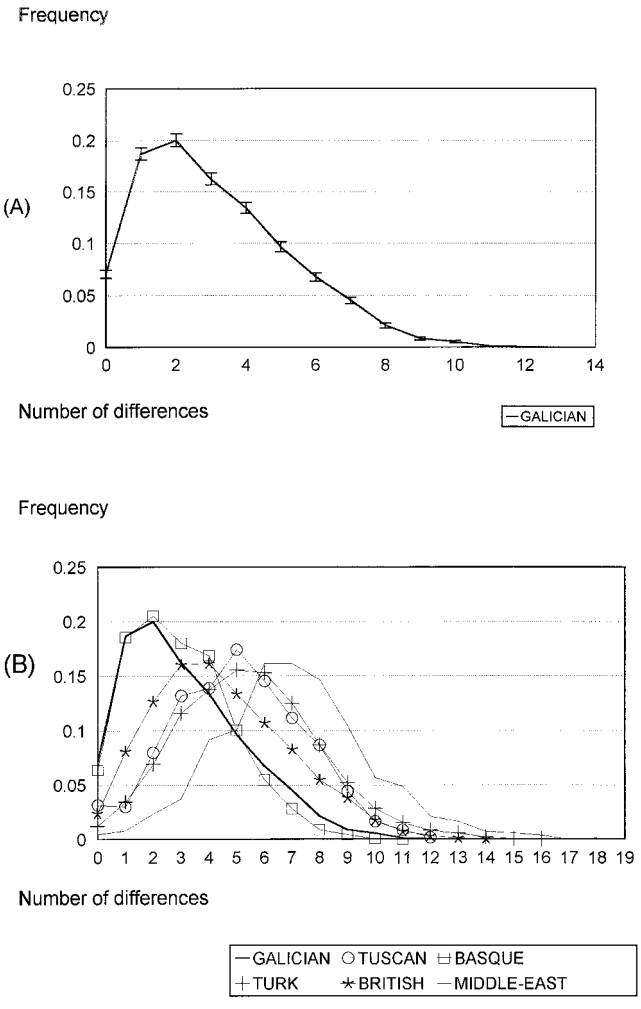

Figure 4 *(**A**) Nucleotide pairwise difference distribution in Galician population. Error bars were computed through 1000 bootstrap iterations. (**B**) Nucleotide pairwise differences distributions of some European and West Asian populations used for comparison*

iterations. From the observed distribution, the τ parameter, related to time since the putative expansion, can be estimated from the theoretical model proposed by Harpending *et al*.[17] In the present population this τ value is estimated to be 1.913 ± 0.045 (standard error computed from 1000 bootstrap iterations), the lowest value found in the European and West Asian populations (Table 3). As the theoretical model proposes, τ would increase with time after the expansion of the populations. In Figure 4(B), several European and West Asian pairwise difference distributions are shown. It can be seen that the peaks of the Western European populations remain at the left-hand side of the graph, whereas the West Asian populations tend to the right. The order of the peaks in the figure from left to right is Galician and Basque (with a very similar distribution), British, Tuscan, Turk and Middle Eastern. This pattern is highly correlated with the geographical position of the populations analysed. Again, the extreme position of Galicia within the European framework is shown.

Population Tree

Genetic distances between European and West Asian populations were calculated and a neighbour-joining tree was constructed. Its robustness was assessed by 1000 bootstrap iterations (Figure 5). The tree displays West Asian populations (Middle East and Turks) at one edge, Galician–Basques–Welsh at the opposite end, and the rest of the populations in between. The robustness of the tree is especially strong at these two edges where over 64% of the bootstrapped trees are found: Middle East and Turks have a bootstrap support of 64.3%, Galician and Basque, 66.8%, and Galician, Basque and Welsh, 68.9%. As seen before, in the present tree a geographic gradient from the Middle East to the West Atlantic populations is also shown. However, it is worth stressing the position of the Spanish population, closer to Central European groups than to the other two north Iberian groups. This fact might be due to an important bias as a result of grouping five different small samples from Spain in a single group of 89 individuals: 30 from northern Spain, 15 from Catalonia, 15 from Andalusia, 11 scattered throughout Spain,[19] and 18 Spanish mainlanders from different regions.[21] When a neighbour-joining tree is built with these groups considered separately (data not shown), only one of the sub-samples (15 Andalusians) is close to the West Asian groups. This fact produces the displacement of the total Spanish group to the central part of the tree. It will be interesting to clarify this point in the future.

Table 3 Population expansion times and τ parameter from Rogers and Harpending[27] model for several populations (Central Europe included German, Bavarian and Tuscan). Two different mutation rates were used to estimate expansion time dates: 4.14×10^{-6},[10] and 2.06×10^{-6},[38] The mean generation time was set to 20 years

	τ	$\mu = 4.14\times10^{-6}$	$\mu = 2.06\times10^{-6}$
Galician	1.913	12.835	25.796
Basque	2.258	15.150	30.448
Central Europe	3.495	23.450	47.128
Turk	4.502	30.206	60.707
Middle East	7.019	47.101	94.647

Site 73 in Hypervariable Region II

It has been suggested[34] that the nucleotide at position 73 of the hypervariable region II is a keystone in interpreting the European variation, as lineages carrying an A at this position would be strongly related to a cluster stemming from the reference sequence (called group 1[4] and group H,[35]) whereas the rest of the lineages would carry a G at this site with few exceptions.[35] The present results (see right-hand columns of Table 1) do not show the clear split of sequence clusters defined by position 73. Of the 16 individuals with the reference sequence, thus supposed to belong to group 1[4] or group H,[35] three (19%) present a G at position 73. Furthermore, of the 19 individuals typed for position 73 supposed to belong to the rest of the groups,[34,35] four (21%) presented an A. Therefore, it seems reasonable not to rely completely on the

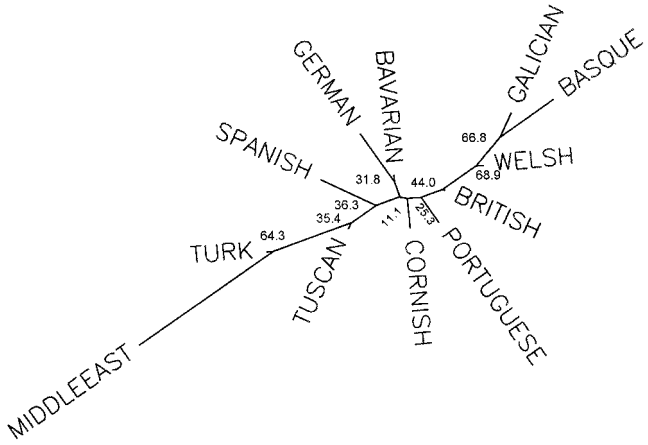

Figure 5 *Neighbour-joining tree built from mismatch-inter-match distances. Numbers at the nodes indicate bootstrap supports after 1000 iterations*

significance of the specific nucleotide present at this single site.

Discussion

The present results are fully compatible with an expansion population model in Europe during the Upper Paleolithic[5] which probably implies the replacement of the Neanderthals by anatomically modern humans. They show the extreme similarity between the two populations situated at the edges of the Cantabrian region (Galicians and Basques), which has been repeatedly shown by archaeologists to be a very homogenous area in prehistoric and mainly Paleolithic times.[36] Allele frequency data, nonetheless, tended to show important differences between these two populations.

The present results show that Galicians present a very low genetic diversity compared with the rest of the European populations. This is shown by the low nucleotide diversity, low sequence diversity and low Shannon index value. Moreover, the mean of the pairwise difference distribution is one of the lowest in Europe, and the frequency of the reference sequence reaches its maximum value in the present population.

The low genetic diversity indexes, the low mean pairwise differences, the simple maximum parsimony tree, the distance matrix between populations and the neighbour-joining tree of populations are compatible with an expansion into Europe from the Middle East, with Galicia at the edge of it. Figure 4B shows that the pairwise difference distributions in Europe and West Asia are compatible with a relatively recent population expansion.

All results are in accordance, as postulated by Comas *et al*,[5] with an expansion wave from the Eastern European populations to the West. The τ values obtained from the Harpending model[17] allows us to estimate this expansion for the Galician region between 13 000 and 26 000 years ago, depending on the mutation rates used. There is some debate on the mutation rate values for the control region of the mtDNA,[37] with recent family data giving higher values than those obtained phylogenetically.[10,38]

Archaeological data shows that the expansion of the Aurignacian culture from Central Europe to the Cantabrian area was very fast[39,40] and therefore it is unlikely that the cultural expansion would have been related to a demographic expansion and substitution. The demographic replacement of the area should have been posterior to the cultural replacement of the initial Upper Paleolithic, and it was homogenous in all of the Cantabrian coast. This replacement might be dated around the Last Glacial Maximum (LGM) – the Solutrean period 18 000 years ago[41] – when environmental conditions contributed to the isolation of groups. This isolation was postulated to have been particularly strong amongst the Basque population but now it also appears clearly in all the narrow Cantabrian littoral. The availability of marine resources in the Cantabrian coast and the relatively hospitable physical environment during the LGM allowed a relatively high population density in the area.[42]

The striking genetic similarity between Galicians and Basques at the mtDNA level may seem difficult to match with the great differences found in allele frequency studies with classic genetic markers.[6,43] Nonetheless, different sources of genetic data may explain different population events, and be compatible beyond apparent discrepancies. Allele frequencies are deeply affected by drift when population size is small and the stochastic nature of drift is likely to accumulate differences through time. Sequence composition, on the other hand, is much less prone to be influenced by population size and thus both populations could have had a similar origin in the late Upper Paleolithic. The passage of time until the late Neolithic growth would have allowed the two populations to diverge in allele frequency terms, preserving nonetheless the same basic mtDNA composition which does not seem to have much altered later in time.

In the specific case of Galicia, some archaeological events may be worth considering, keeping in mind that the whole Cantabrian area seems to have had a common heritage in prehistory. The effect of the Neolithic may now be envisaged as having had a poor population substitution in Atlantic Europe according to the availability model,[44] especially in the Cantabrian region.[45] Another important archaelogical event, often referred to as 'Celtic influences' in Galicia coming from the North Atlantic coast during the first millenium BC seems to have been produced by an elite contact with a small demographic effect.[7,46] Nonetheless, only if the migration had been quantitatively important could it have been detected by the present analysis due to the high homology of the mtDNA in the Atlantic European populations as a result of the Paleolithic expansion from the Middle East.

The genetic analysis of single populations may help not only to reconstruct the patterns of human history

but also, as in the present case, to understand discrepancies between different genetic analyses.

Acknowledgements

This work was supported in part by grants from the Dirección General de Investigación Científico Técnica (PB95-0267-C02-01) and from the Direcció General de Recerca, Generalitat de Catalunya (1996SGR00041) awarded to JB and DC and the grant from the Xunta de Galicia (XUGA 20801B95), given to MVL, AS, and AC. The help of F Calafell and R Ward is also acknowledged with appreciation. Hans Bandelt made an invaluable revision of the manuscript.

References

1 Cavalli-Sforza LL, Menozzi P, Piazza A: *History and Geography of Human Genes*. Princeton University Press, Princeton, 1994.
2 Harding RM, Fullerton SM, Griffiths RC *et al*: Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 1997; **60**: 772–789.
3 Sajantila A, Lahermo P, Anttinen T *et al*: Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res* 1995; **5**: 42–52.
4 Richards M, Côrte Real H, Forster P *et al*: Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 1996; **59**: 185–203.
5 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.
6 Calafell F, Bertranpetit J: Principal component analysis of gene frequencies and the origin of Basques. *Am J Phys Antrop* 1994; **93**: 201–215.
7 Ruiz-Gálvez M: Canciones del muchacho viajero. *Veleia* 1990 **7**: 79–103.
8 Sherratt A: Plough and pastoralism: aspects of the secondary products revolution. In: Hodder I, Isaac G, Hammond N (eds): *Patterns of the Past*. Cambridge University Press, Cambridge, 1981.
9 Valverde E, Cabrero C, Cao R *et al*: Population genetics of three VNTR polymorphisms in two different Spanish populations. *Int J Legal Med* 1993; **151**: 251–256.
10 Ward RH, Frazier BL, Dew-Jager K, Pääbo S: Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci USA* 1991; **88**: 8720–8724.
11 Vigilant L, Pennington R, Harpending H, Kocher TD: Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci USA* 1989; **86**: 9350–9354.
12 Anderson S, Bankier AT, Barrell BG *et al*: Sequence and organisation of the human mitochondrial genome. *Nature* 1981; **290**: 457–465.
13 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. *Nucl Acid Res* 1994; **22**: 4673–4680.
14 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 1989; **5**: 164–166.
15 Nei M, Tajima F: DNA polymorphism detectable by restriction endonucleases. *Genetics* 1981; **97**: 145–163.
16 Magurran AE: *Ecological Diversity and its Measure*. Princeton University Press, Princeton, 1988.
17 Harpending HC, Sheny ST, Rogers AR, Stoneking M: The genetic structure of ancient human populations. *Curr Anthrop* 1993; **34**: 483–496.
18 Bertranpetit J, Sala I, Calafell F, Underhill P, Moral P, Comas D: Human mitochondrial DNA variation and the origin of the Basques. *Ann Hum Genet* 1995; **59**: 63–81.
19 Côrte-Real HBSM, Macaulay VA, Ricards MB *et al*: Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 1996; **60**: 31–350.
20 Piercy R, Sullivan KM, Benson N, and Gill P: The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int J Leg Med* 1996; **106**: 85–90.
21 Pinto F, González AM, Hernánez M, Larruga JM, Cabrera VM: Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann Hum Genet* 1996; **60**: 321–330.
22 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.
23 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.
24 Comas D, Calafell F, Mateu E, Pérez-Lezaun, Bertranpetit J: Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European population. *Mol Biol Evol* 1996; **13**: 1067–1077.
25 Di Rienzo A, Wilson AC: Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci USA* 1991; **88**: 1597–1601.
26 Rao CR: Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Bio* 1982; **21**: 24–43.
27 Rogers AR, Harpending H: Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 1992; **9**: 552–569.
28 Efron B: *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, 1982.
29 Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985; **39**: 785–791.
30 Wakeley J: Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J Mol Evol* 1993; **37**: 613–623.
31 Bertranpetit J, Calafell F, Comas D, Pérez-Lezaun A, Mateu E: Mitochondrial DNA sequences in Europe: an insight into population history. In: AJ Boyce and CGN Mascie-Taylor (eds): *Molecular Biology and Human Diversity*. Cambridge University Press, Cambridge, 1996, pp112–129.
32 Hudson RR: Gene genealogies and the coalescent process. *Oxford Surveys in evolutionary biology* 1990; **7**: 1–44.

33 Slatkin M, Hudson R: Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 1991; **129**: 555–562.
34 Wilkinson-Herbots HM, Richards MB, Forster P, Sykes C: Site 73 in hypervariable region II of the human mitochondrial genome and the origin of European populations. *Ann Hum Genet* 1996; **60**: 499–508.
35 Torroni A, Huoponen K, Francalacci P *et al*: Classification of European mtDNAs from an analysis of three European populations. *Genetics* 1996; **144**: 1835–1850.
36 Straus LG: Upper Paleolithic origins and radiocarbon calibration: more new evidence from Spain. *Evolutionary Anthropology* 1994; **2**: 195–198.
37 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.
38 Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N: Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci USA* 1995; **92**: 532–536.
39 Cabrera V, Bischoff J: Accelerator ^{14}C ages for basal Aurignacian at El Castillo, Spain. *J Archaeol Sci* 1989; **16**: 577–584.
40 Straus LG: Age of modern Europeans. *Nature* 1989; **342**: 476–477.
41 Soffer O, Gamble C (eds): *The World at 18000 BP*. Unwin Hyman, London, 1990.
42 Straus LG: Southwestern Europe at the Last Glacial Maximum. *Current Anthropology* 1991; **32**: 189–199.
43 Bertranpetit J, Cavalli-Sforza LL: A genetic reconstruction of the history of the population of the Iberian Peninsula. *Ann Hum Genet* 1991; **55**: 51–67.
44 Zvelebil M, Rowley-Conwy PR: Foragers and farmers in Atlantic Europe. In: M Zvelebil (ed). *Hunters in Transition*. Cambridge University Press, Cambridge, 1986, pp67–93.
45 Arias P: *De cazadores a campesinos. La Transición al Neolítico en la región Cantábrica*. Universidad de Cantabria, Santander, 1991.
46 Sherratt A: The emegence of élites: earlier Bronze Age Europe, 2500–1300 BC. In: Cunliffe B (ed): *The Oxford Illustrated Prehistory of Europe*. Oxford University Press, Oxford, 1994, pp244–276.