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

5 ng of DNA in 25 l 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 M 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 $(n/n\text{-}1)\ (1/l)\ \sum_{i=1}^{l}(1\text{-}x_i^2)$, 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 $(n/n\text{-}1)\ \sum_{i=1}^{k}(1\text{-}p_i^2)$, 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 $H = -\ p_i \log_2 p_i$ (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 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

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1

Table 1 *Continued*

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

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.

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 = 1.33 mutations. As shown in Figure 3, Galician sequences do not fit a Poisson distribution with such a high value (^{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

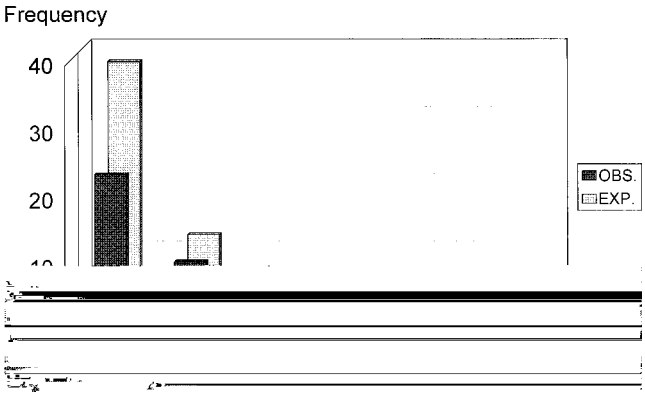

Figure 3 *Expected and observed distributions of the number of mutational events occurred from the reference sequence in the Galician population*

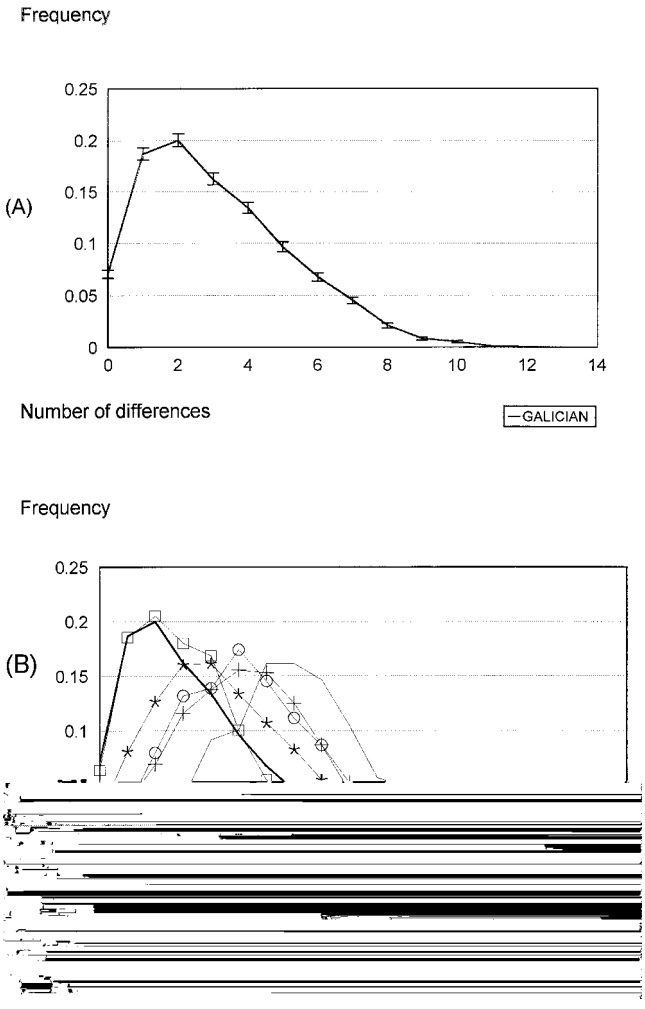

Figure 4 *(**A**) Nucleotide pairwise difference distribution in Galician population. Error bars were computed through 1000 bootstrap iterations. (**B**) Nucleotide pairwise differences distributions of some European and West Asian populations used for comparison*

iterations. From the observed distribution, the 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 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, 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%,

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

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

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.