Study of a trimeric tandem repeat locus (SBMA) in the Basque population: Comparison with other populations

M.I. ARRIETA[a], B. MARTINEZ[a], J.M. MILLAN[b], A. GIL[a], E. MONROS[b], T. NUÑEZ[a], M. TELEZ[a], F. MARTINEZ[b]

[a] *Departamento de Biología Animal y Genética, Universidad del País Vasco, Spain*
[b] *Unidad de Genética, Hospital "La Fe", Valencia, Spain*

SUMMARY

Microsatellites can be highly unstable and show a high level of polymorphism between individuals. Here we present the analysis of the CAG trinucleotide repeat polymorphism at the SBMA locus in 57 phenotypically normal individuals rigorously assigned to the Spanish Basque population. Results are compared with 100 Spanish non-Basque individuals who were already analyzed by us (175 alleles). This is the first study undertaken in these populations for this marker. In addition, we compared our results with those published for other populations. Relative allele frequencies showed differences between the samples and no unimodal distribution. The expected heterozygosity in the Basque sample was slightly lower than in the non-Basque sample. Conformity with Hardy-Weinberg equilibrium was verified by three tests. When compared with published data, the predominant alleles appear to be the same in the various populations. There are more differences between Basques and other Caucasoid samples than between non-Basques and Caucasoid samples. Population relationships were also examined by dendrograms based on genetic distances. The results obtained showed some peculiarities in the Basque population. The high degree of similarity with other dendrograms based on different markers and the efficiency of this STR marker in differentiating closely related populations, support the potential usefulness of microsatellites as tools for human population studies.

ARRIETA MI, PASTOR G, GONZALEZ E, IBARRONDO A, LOSTAO CM (1987b) a-b palmar interdigital ridge-count in the Basque population. Anthrop Anz *45*: 57-61.

ARRIETA MI, SALZAR L, LOSTAO CM (1987c) Finger ridge count in Basque populations: Univariate and multivariate comparison with other Spanish populations. Ann Hum Biol *14*: 507-516.

ARRIETA MI, CRIADO B, MARTINEZ B, SALAZAR L, LOSTAO CM (1990a) Analysis of digital patterns in the Basque Valley of Deba: Multivariate comparison with other populations. Hum Biol *62*: 553-572.

ARRIETA MI, MARTINEZ B, CRIADO B, SIMON A, SALAZAR L, LOSTAO CM (1990b) Dermatoglyphic analysis of autistic Basque children. Am J Med Genet *31*: 1-9.

ARRIETA MI, CRIADO B, MARTINEZ B, SIMON A, SALAZAR L LOSTAO CM (1991) Characterization of digital dermatoglyphic features in Salazar Valley: Comparison with other Basque valleys. Ann Hum Biol *18*: 221-233.

ARRIETA MI, MARTINEZ B, CRIADO B, LOBADO N, LOSTAO CM (1992) Genetic and dermatoglyphic distances among Basque valleys. Hum Biol *64*: 705-716.

ARRIETA MI, MARTINEZ B, NUNEZ M, GIL A, CRIADO B, TELEZ M, LOSTAO CM (1995) a-b ridge count in a Basque population: Fluctuating asymmetry and comparison with other populations. Hum Biol *67*: 121-133.

ARRIETA MI, ECHARRI A, NUNEZ T, GIL A, CRIADO B, MARTINEZ B (1996) Cytogenetics of autosomal fragile sites: A Basque population study. Am J Hum Biol *8*: (4) (in press).

BARBUJANI G, PILASTRO A (1993) Genetic evidence on origin and dispersal of populations speaking languages of the Nostratic macrofamily. Proc Natl Acad Sci USA *90*: 4670-4673.

BERTORELLE G, BERTRANPETIT J, CALAFELL F, NASIDZE IS, BARBUJANI G (1995) Do Basque and Caucasian speaking populations share non-Indo-European ancestors? Eur J Hum Genet *3*: 256-263.

BERTRANPETIT J, CAVALLI-SFORZA LL (1991) A genetic reconstruction of the history of the population of the Iberian Peninsula. Ann Hum Genet *55*: 51-67.

BIANCALANA V, SERVILLE F, POMMIER J, JULIEN J, HANAUER A, MANDEL JL (1992) Moderate instability of the trinucleotide repeat in spino bulbar muscular atrophy. Hum Mol Genet *1*: 255-258.

BOWCOCK AM, KIDD JR, MOUNTAIN JL (1991) Drift, amixture and selection in human evolution: A study with DNA polymorphisms. PNAS *88*: 839-843.

BOWCOCK AM, RUIZ-LINARES A, TOMFOHRDE J, MINCH E, KIDD JR, CAVALLI-SFORZA LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature *68*: 455-457.

CANN RL, STONEKING M, WILSON AC (1987) Mitochondrial DNA and Human Evolution. Nature *325*: 31-36.

CAVALLI-SFORZA LL (1988) The Basque population and ancient migrations in Europe. Munibe (Antropologia y Arqueologia) Suplemento *6*: 129-137.

CAVALLI-SFORZA LL, PIAZZA A (1993) Human genomic diversity in Europe: A summary of recent research and prospects for the future. Eur J Hum Genet *1*: 3-18.

CAVALLI-SFORZA LL, PIAZZA A, MENOZZI T, MOUNTAIN J (1988) Reconstruction of human evolution: Bringing together gentic, archaeological, and linguistic data. Proc Natl Acad Sci USA *85*: 6002-6006.

COLLIGNON R (1894) Les Basques. Nem Soc Anthrop *3*: 1-64.

CHAKRABOORTY R, FORNAGE M, GUEGUEN R, BOERWINKLE E (1991) Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism with a population. In: DNA fingerprinting approaches and applications. T. Burke, G. Dolf, J. Jeffrey, R.Wolff, Birkhauser verlag Basel, Switzerland.

DEKA R, CHAKRABORTY R, FERRELL RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. Genomics *11*: 83-92.

DEKA R, CHAKRABORTY R, DECROO S, ROTHHAMMER F, BARTON SA, FERRELL RE (1992) Characteristics of polymorphism at a VNTR locus 3' to the apolipoprotein B gene in five human populations. Am J Hum Genet *51*: 1325-1333.

DEKA R, JIN L, SHRIVER MD, YU LM, DECROO S, HUNRIESER J, BUNKER CH, FERRELL RE, CHAKRABORTY R (1995) Population genetics of dinucleotide (dC-dA)n (dG-dT)n polymorphism in world populations. Am J Hum Genet *56*:461-474.

EDWARDS A, HAMMOND HA, JIN L, CASKEY CT, CHAKRABORTY R (1992) Genetic variations at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics *12*: 241-253.

FERLINI A, ZANETTI A, GUIDETTI D (1991) Refinement of spinal and bulbar muscular atrophy (SBMA) localization by linkage analysis. Cytogenet Cell Genet *58*: 2064.

FISHBECK KH, IONASESCU V, RITTER AW, IONASESCU R, DAVIES K, BALL S, BOSCH P, BURNS T, HAUSMANOWA-PETRUSEWICZ I, BORKOWSKA J (1986) Localization of the gene for X-linked spinal muscular atrophy. Neurology *36*: 1595.

FU YH, KHUL DPA, PIZZUTI A, PIERETTI M, SUTCLIFFE JS, RICHARDS S, VERKERK AJMH, HOLDEN JJA, FENWICK RG, WARREN ST, COSTRA BA, NELSON DL, CASKEY CT (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. Cell *67*: 1047-1058.

GUGLIELMINO CR, SILVESTRI A (1995) Surname sampling for the study of the genetic structure of an Italian province. Hum Biol *67*: 613-628.

KAWAGUCHI Y, OKAMOTO T, TANIWAKI M, AIZAWA M, INOUE M, KATAYAMA S, KAWAKAMI H, NAKAMURA S, NISHIMURA M, AKIGUCHI I, KIMURA J, NARUMIYA S, KAKIZUKA A (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nature Genet *8*: 221-227.

KENNEDY WR, ALTER M, SUNG JH (1968) Progressive proximal spinal and bulbar muscular distrophy of late onset: A sex-linked recessive trait. Neurology *18*: 171.

KNIGHT SJL, FLANNERY AV, HIRST MC, CAMPBELL L, CHRISTODOULOU Z, PHELPS SR, POINTON J, MIDDLETON-PRICE HR, BARNICOAT A, PEMBREY ME, HOLLAND J, COSTRA BA, BOBROW M, DAVIES KE (1993) Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell *74*: 127-134.

KOIDE R, IKEUCHI T, ONODERA O, TANAKA H, IGARASHI S, ENDO K, TAKAHASHI H, KONDO R, ISHIKAWA A, HAYASHI T. SAITO M, TOMODA A, MIKE T, NAITO H, IKUTA F, TSUJI S (1994) Unstable expansion of CAG repeat in hereditary dentatorubralpalli-doluysian atrophy (DRPLA). Nature Genet *6*: 9-13.

KREMER EJ, PRITCHARD M, LYNCH M, YU S, HOLMAN K, BAKER E, WARREN ST, SCHLESSINGER D, SUTHERLAND GR, RICHARDS RI (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CGG)n. Science *252*: 1711-1714.

LA SPADA AR, WILSON AM, LUBAHN DB, HARDING AE, FISCHBECK KH (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature *352*: 77-79.

LERER I, MERIMS D, ABELIOVICH D, ZLOTOGORA J, GADOTH N (1996) Machado-Joseph disease: Correlation between the clinical features, the CAG repeat length and homozygosity for the mutation. Eur J Hum Genet *4*: 3-7.

MENOZZI P, PIAZZA A, CAVALLI-SFORZA LL (1978) Synthetic maps of human gene frequencies in Europeans. Science *201*: 786-792.

NAGAFUCHI *et al* (1994) Dentatorubral and pallidolyusian atrophy expansion of an unstable CAG trinucleotide on chromosome 12 p. Nature Genet *6*: 14.

ORR HT, GHUNG MY, BANFI S, KWIATKOWSKI TJ, SERVADIO A, BEAUDET AL, MCCALL AE, DUVICK LA, RANUM LPW, ZOGHBI HY (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellarataxia type 1. Nature Genet *4*: 221-226.

PIAZZA A, RENDINE G, ZEI G, MORONI A, CAVALLI-SFORZA LL (1987) Migration rates of human populations from surname distribution. Nature *329*: 714-716.

PIAZZA A, CAPELLO N, OLIVETTI E, RENDINE S (1998) A genetic history of Italy. Ann Hum Genet *52*: 203-213.

RENFREW C (1987) Archaeology and Language: The Puzzle of Indo-European Origins. London, Jonathan Cape.

RICHARDS RI, SUTHERLAND GR (1992) Dynamic mutations: A new class of mutations causing human disease. Cell *70*: 709-712.

SAITOU N, NEI M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol *4*: 406-425.

SANCHEZ-MAZAS A, LANGANEY A (1988) Common genetic pools between human populations. Hum Genet *78*: 161-166.

SAR M, STUMPF WE (1977) Androgen concentration in motor neurons of cranial nerves and spinal cord. Science *197*: 77.

SOKAL RR, ODEN NL, WILSON C (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature *351*:143-145.

SOKAL RR, ROHLF FJ (1969) Biometry. Freeman and company (Eds.), New York, N.Y., USA.

STAROSTIN SA (1990) A statistic evaluation of the time-depth and subgrouping of the nostratic macrofamily. In: Evolution. From molecules to culture (Ed. R .Dawkins), p 33. Pasadena, Cold Spring Harbor, USA.

TILLEY WD, MARCELLI M, WILSON JD, MCPHAUL MJ (1989) Characterization and expression of a cDNA encoding the human androgen receptor. Proc Natl Acad Sci USA *86*: 327-331.

VERKERK AJ, PERETTI M, SUTCLIFFE JS, FU YH, KUHL DPA, PIZZUTI A, REINER O, RICHARDS S, VICTORIA MF, ZHANG F, EUSSEN BE, VAN OMMEN GJB, BLONDEN LAJ, RIGGINS GJ, CHASTAIN JL, KUNST CB, GALJAARD H, CASKEY CT, NELSON DL, OOSTRA BA, WARREN ST (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X sindrome. Cell *65*: 905-914.

VIGILANT L, STONEKING M, HARPENDING H, HAWKES K, WILSON AC (1991) African populations and the evolution of human mitochondrial DNA. Science *253*: 1503-1507.

WATKINS WS, BAMSHAD M, JORDE LB (1995) Population genetics of trinucleotide repeat polymorphisms. Hum Mol Gentile *4*: 1485-1491.

WIJSMAN E, ZEI G, MORONI A, CAVALLI-SFORZA LL (1984) Surnames in Sardinia: Computation of migration matrices from surname distributions in different periods. Ann Hum Genet *48*: 65-78.

ZAR JH (1984) Biostatistical analysis. (2nd edition). Prentice Hall, New Jersey (USA).

Dra. I. Arrieta
Dpto. Biologia Animal y Genetica
Universidad del Pais Vasco
Apdo. 644 - 48080 Bilbao
Spain

INTRODUCTION

One of the salient features of the mammal genome, in general, and of the human genome, in particular, is the vast excess of DNA sequences without a particular function. In this DNA group we can include the microsatellites or simple tandem repeated sequences, which include triplet and dinucleotide repeats. Microsatellites can be highly unstable and show high level of polymorphism. This is the reason why they are being used in the study of human genetic diversity and relationships [Edwards *et al* 1992; Bowcock *et al* 1994; Deka *et al* 1995; Watkins *et al* 1995].

The trinucleotide sequences are particularly interesting because unstable trinucleotide repeats, referred to as dynamic mutations [Richards and Sutherland 1992], have been found implicated in the pathology of some human diseases [Kremer *et al* 1991; Verkerk *et al* 1991; Biancalana *et al* 1992; Knight *et al* 1993; Orr *et al* 1993; Kawaguchi *et al* 1994; Koide *et al* 1994; Nagafuchi *et al* 1994; Lerer *et al* 1996]. Among these, spinal and bulbar muscular atrophy (SBMA), also known as Kennedy disease [Kennedy *et al* 1968], is an X-linked recessive disorder of the motor neurones.

The SBMA locus is located at the Xq12 chromosome [Fishbeck *et al* 1986; Ferlini *et al* 1991]. The results of some studies have proposed the AR gene (androgen receptor) as a candidate gene for SBMA [Sar and Stumpf 1977; La Spada *et al* 1991].

The androgen receptor gene contains a highly polymorphic CAG trinucleotide repeat in the 5' coding region. Expansion of the CAG repeat was found to be associated with the SBMA phenotype. The number of repeats in affected people ranges from 40 to 52 while in normal individuals the range is from 12 to 30. As the repeat number is also polymorphic in normal individuals, SBMA may be useful for interpopulation studies.

The Basque Country is located at the western end of the Pyrenees and comprises three provinces in France and four in Spain. The origin of the Basques is not exactly known. According to Collignon [1894], the Basques constitute the only present-day survivors of an ancient population of prehistoric Western Europe. According to Aranzadi [1905], the Basques represent only a local variant population among other populations in Southwestern Europe. More recent studies still maintain the controversy about their origin [Bertranpetit and Cavalli-Sforza 1991; Cavalli-Sforza and Piazza 1993; Bertorelle *et al* 1995].

The Basques speak an ancient language, "Euskara", with very distinct differences from what the surrounding populations speak. They are also culturally different from other European populations. Many studies on different traits showed Basque peculiarities when compared to other European populations [Piazza *et al* 1988; Bertranpetit and Cavalli-Sforza 1991;

Barbujani and Pilastro 1993; Cavalli-Sforza and Piazza 1993]. In our previous research we found some peculiarities and also some differences between the Basques and other Spanish populations [Arrieta *et al* 1987a,b,c; Arrieta *et al* 1990a,b; Arrieta *et al* 1991, 1992, 1995, 1996; Arrieta and Lostao 1988].

Here we present the analysis of the polymorphic CAG trinucleotide repeat at the SBMA locus in phenotypically normal and carefully selected individuals from the Spanish Basque population. The results are compared with those of a Spanish non-Basque sample previously analyzed by us. To our knowledge, this was the first study carried out in these populations. The objective is to find out if this polymorphism is also useful in differentiating individuals of Spanish Basque and non-Basque origin.

In addition, we compared our results with those published for other populations [Edwards *et al* 1992; Watkins *et al* 1995] obtaining a good demonstration of the usefulness of highly polymorphic microsatellites in unravelling the ancient history of mankind.

MATERIALS AND METHODS

We have typed the $(CAG)_n$ STR located near the SBMA locus in a sample of 57 unrelated apparently healthy autochthonous Spanish Basque females (114 alleles) and in a sample of 100 unrelated apparently healthy individuals (75 females and 25 males = 175 alleles) with Spanish non-Basque ancestry. Each individual ethnic allocation was based both on their surnames and on the paternal and maternal grandparents' birthplace. (In Spain, both the father's and the mother's surnames are used in sequential order; it is therefore easy to ascertain the four grandparents' or the eight great-grandparents' surnames).

Apart from the proven utility of surnames in sample selection [Wijsman *et al* 1984; Piazza *et al* 1987; Guglielmino and Silvestri 1995], in Basques they are a good criterion because of their marked difference from those of other Spanish regions.

The Basque sample was obtained by selecting individuals exclusively from rural areas of the Basque country and who had all eight surnames (the two surnames of their four grandparents) of Basque origin. The non-Basque sample was made up by individuals from the Valencia area (eastern Spain) and other parts of Spain with Spanish non-Basque surnames.

Genotype determinations - Analysis of Human Androgen Receptor (HUMARA) $(CAG)_n$ polymorphism was performed basically as described by Fu *et al* [1991] with minor modifications: amplified fragments were analyzed in native polyacrilamide gels that were then silver stained. A ladder

was previously obtained by mixing those PCR products which encompassed the whole variation in size observed in a first round of 32 samples (17 to 31 repeats). This ladder was then sized with different DNA size markers under variable conditions, including digestion of PCR products that yield much smaller sizes, and the number of repeats was deduced from the gene sequence [Tilley *et al* 1989].

Statistical analysis - Allele frequencies were estimated by gene counting. Expected heterozygosities (H) and standard errors (SE) for allele frequencies and H were estimated as described elsewhere [Edwards *et al* 1992]. Inter-sample comparisons of allele frequencies and of heterozygosity were carried out by means of Mann-Withney and chi-square tests [Zar 1984].

Following Chakraborty and Weiss [1991], Deka *et al* [1991, 1992, 1995] and Edwards *et al* [1992], three tests to detect deviation from Hardy-Weinberg equilibrium were used: a) a test based on the observed and expected number of heterozygotes, b) a test based on the total number of observed distinct homozygous and heterozygous genotypes and c) a likelihood-ratio test (G statistics) of observed and expected frequencies of every genotype.

Genetic distances between pairs of populations were estimated by using Prevosti's distance and the complement of common gene frequencies $D = 1\text{-}{}_2P/100$ where ${}_2P$ is the sum of minimum shared frequencies (in percentage) for every allele when comparing the two populations [Sánchez-Mazas and Langaney 1988]. We used these methods because of their simplicity and because the former is based on the analysis of differences, meanwhile the latter is based on the analysis of similarities. Furthermore, the sum of minimum shared frequencies among several populations (nP values) can easily be obtained and considered as the percentage of what the populations clustered at a given level still have in common.

The dendrogram was constructed on the basis of the genetic distances calculated as described above. Unrooted neighbour-joining trees were used [Saitou and Nei 1987].

RESULTS

Descriptive analysis and comparison with the Spanish non-Basque sample - The relative allele frequency distribution of polymorphism HUMARA $(CAG)_n$ in Spanish Basques and non-Basques appears not to be unimodal (Fig. 1 and Table I).

The spectrum of allelic variations is broader in non-Basques. The number of alleles observed in the Basque and in the non-Basque populations is 12 and 16, respectively. The most common allele coincides in the two populations and this allele is over-represented in Basques (28.07% of the total allele frequency).

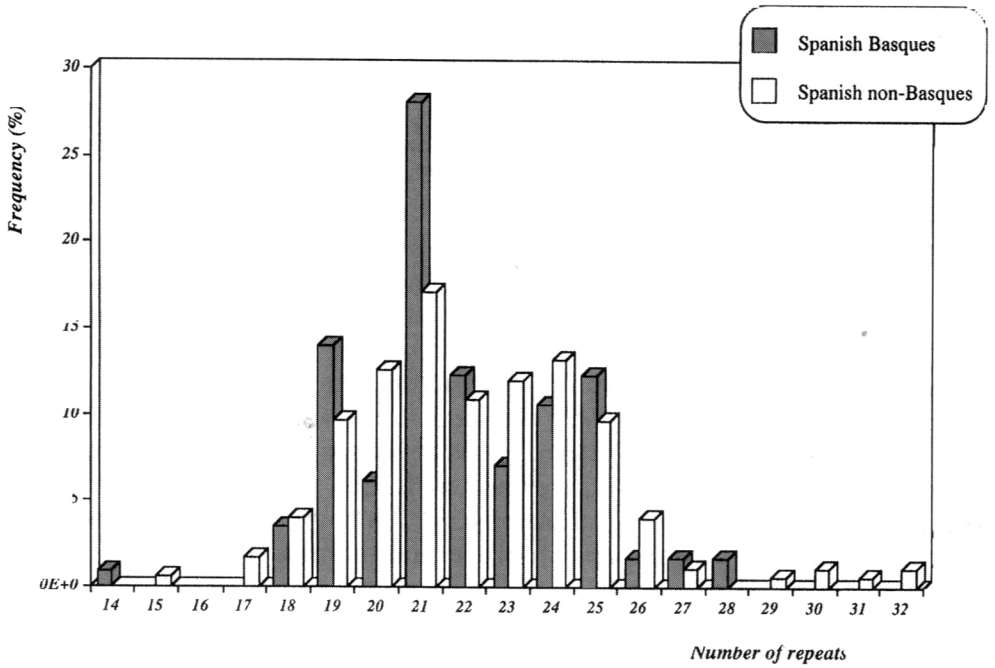

Fig. 1 - Allele frequency distribution for CAG trinucleotide repeat at the SBMA locus in Spanish Basque and non-Basque population.

Table I - Allele frequencies (%) at SBMA locus in Spanish Basque and non-Basque populations.

Allele	Basque	non-Basque
14	0.88	
15	0.57	
17	1.71	
18	3.51	4.00
19	14.04	9.71
20	6.14	12.57
21	28.07	17.14
22	12.28	10.86
23	7.02	12.00
24	10.53	13.14
25	12.28	9.71
26	1.75	4.00
27	1.75	1.14
28	1.75	
29	0.57	
30	1.14	
31	0.57	
32	1.14	

About 4% of the alleles in the non-Basque sample consisted of more than 28 CAG repeats. However, no allele above 28 CAG repeats was found in the Basque sample, although individuals were in the normal range of repeats.

The mean number and standard error of CAG repeats for Spanish Basque and non-Basque samples were 21.88 ± 0.23 and 21.70 ± 0.17, respectively. This difference was not significant ($Z = 0.96$, $p > 0.05$).

The expected heterozygosity in the Spanish Basque sample (0.847 ± 0.047) was very similar to that of the non-Basque sample (0.884 ± 0.036).

The results of the three tests, designed to detect a possible deviation from Hardy-Weinberg equilibrium, are the following: the first test, based on the observed and expected number of heterozygotes, did not detect heterozygote deficiencies ($p > 0.05$); the second test based on the total number of distinct homozygous and heterozygous genotypes observed, did not show a significant deviation from the expected values ($p > 0.05$); the third test [Edwards *et al* 1992] compared the observed genotype frequencies with the expected number by the expansion of the multinomial ($\sum p_i^2 = 1$). The likelihood-ratio test statistic result (-2 Ln L_0/L_1) is 84.38 ($p > 0.05$) where L_0 is the likelihood under Hardy-Weinberg equilibrium and L1 is the likelihood under general multinomial genotypic population conditions.

Comparison with other populations - We used published data from different ethnic groups: Afro-Americans (AFR), Caucasoids (CAU1, CAU2), Mexican-Americans (MEX), and Asians (ASI). AFR, CAU1, MEX and ASI were taken from Edwards *et al* [1992], and CAU2 from Watkins *et al* [1995]. CAU2 was analyzed separately from CAU1 as a test of sampling effect, since it is smaller in size (± 100 chromosomes) than the Spanish Basque (BAS) or Spanish non-Basque (SPA) samples.

Compared with the published data mentioned above, it appears that only in the Basque sample no allele above 28 CAG repeats was found. The frequencies of some specific alleles are clearly variable across the population groups (for example, alleles 17 and 18). Predominant alleles do not coincide between populations.

Basques were significantly different from other Caucasoid samples for allele frequencies ($\chi^2 = 16.42$; $p = 0.05$, d.f. 9). However, Spanish non-Basque frequencies were not significantly different from the samples mentioned above ($\chi^2 = 3.70$; $p = 0.93$, d.f. 9). In relation to the other comparisons BAS is also more differentiated than SPA except for the AFR; in this case, both comparisons obtained the same probability ($p = 10^{-4}$).

As a measure of variation between populations we also compared heterozygosity values: their range is from 0.83 in CAU1 to 0.91 in AFR. The lowest values of heterozygosity appear in CAU1, BAS and SPA, following this order.

Table II - Matrix of genetic distances between the samples examined.

Population	BAS	SPA	CAU1	CAU2	AFR	MEX	ASI
BAS		0.213	0.212	0.223	0.381	0.305	0.297
SPA	0.212		0.133	0.155	0.403	0.235	0.223
CAU1	0.215	0.134		0.088	0.377	0.254	0.288
CAU2	0.222	0.155	0.093		0.342	0.256	0.299
AFR	0.376	0.399	0.377	0.347		0.374	0.391
MEX	0.305	0.235	0.255	0.262	0.375		0.178
ASI	0.295	0.222	0.288	0.303	0.390	0.177	

Values above the diagonal were calculated from the shared gene frequencies; values below the diagonal were calculated by Prevosti's distance.

Population relationships were also examined using genetic distances. The matrix of genetic distances between groups is shown in Table II. The values above the diagonal were calculated from the shared gene frequencies while the values below it were calculated by Prevosti's distance. It is noteworthy that in relation to all Caucasoid samples, CAU2 and CAU1 show the smallest genetic distance. BAS showed the greatest distances from the CAU1, CAU2 and SPA while the distances between SPA, CAU1 and CAU2 were similar.

Genetic relationships are shown more explicitly in the average linkage dendrogram (Fig. 2). The distance used is $(1_{-2}P/100)$ and the intergroup *nP*s are indicated at the nodes. Both angles and branch lengths are plotted proportionally to the mean *nP* common to all populations clustered at a node.

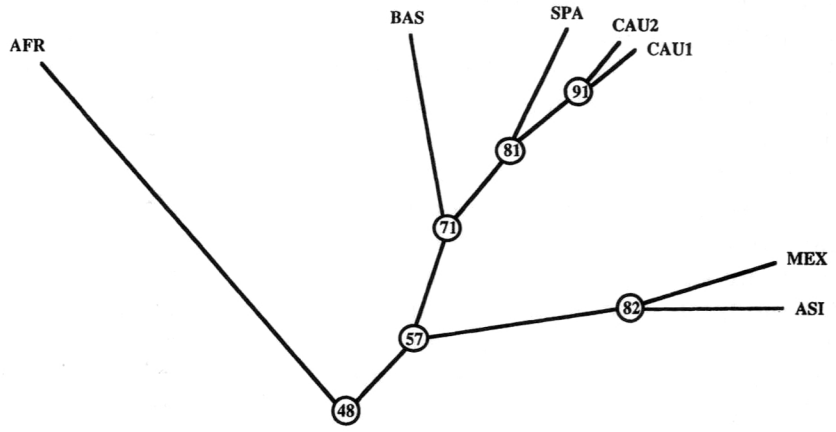

Fig. 2 - Average linkage dendrogram for samples. The distance is $(1\text{-}_2P/100)$ and intergroup nP_s are indicated at the nodes. Both angles and branch lengths are plotted proportional to the mean nP_s commmon to all populations clustered at a node.

Note the intermediate position of SPA between BAS and CAU1 and CAU2. The tree clearly identifies three groups. The population of Black African origin is distinctly separated from the rest, followed by the separation of Caucasoids (mainly of European origin) from Amerindians and Asiatics. Amerindians seem to be more closely related to Asiatics than to Caucasoids.

DISCUSSION

The analysis of the polymorphic CAG trinucleotide repeat at SBMA locus shows that, in relation to the allele size, the range of variation is more similar between the Spanish non-Basque sample and the Caucasoid samples previously analyzed [Edwards *et al* 1992; Watkins *et al* 1995] than between the Spanish non-Basque and Spanish-Basque samples.

Among all the populations used in the comparison, only in the Basque sample no allele above 28 CAG repeats was found, suggesting that its incidence in the Basque population could be lower than in other Caucaoid populations.

The BAS, SPA and CAU1 show the lowest heterozygosity values, possibly reflecting a lower level of genetic diversity in these populations.

The test to detect deviation from Hardy-Weinberg equilibrium at SBMA locus in AFR, CAU1, ASI and MEX [Edwards *et al* 1992] showed a deviation in the Black population. In this study, we followed the procedure of the authors mentioned above to see how this locus behaves in BAS. Our results are similar to those obtained for the CAU1, therefore, allele frequencies derived from the Spanish Basque population provide reasonable estimation of genotype frequencies as well.

It is difficult to interpret these differences. According to Bertorelle *et al* [1995], among other factors, allele frequency differences between groups depend on the history of the populations. Probably, the differences found reflect the evolutionary history of the alleles in the different groups. The CAG trinucleotide repeat is located within the coding region of exon 1 of the androgen receptor; therefore it is not difficult to imagine a biological basis of selection for or against specific allele sizes. Although this locus is very variable, there are size constraints that can be tolerated by the functional protein.

A large fraction of the allele frequency gradients currently observed in Europe are interpreted as a consequence of the spread of alleles of Near Eastern origin, probably as a consequence of Neolithic demic diffusion. According to Cavalli-Sforza [1988], Basques have only marginally been affected by the demic expansion that propagated in European farming culture [Menozzi *et al* 1978; Ammerman and Cavalli-Sforza 1984; Sokal *et al* 1991], and possibly languages [Renfrew 1987; Starostin 1990; Barbujani

and Pilastro 1993], of Near Eastern origin.According to Cavalli-Sforza and Piazza [1993], "Basques are probably the most direct descendants of the earliest post-Neanderthal settlers of Europe. Conservation of a distinct language must have been an important factor in maintaining social and genetic identity. It is very likely that the genetic uniqueness of Basques is a product of the remarkable isolation of western from eastern Europe at the time of last glaciation, which peaked around 18,000 years ago, and there may have been very limited genetic and cultural exchanges between the two halves of Europe during the early Paleolithic".

Our results support this point of view. In fact, the divergence of Basques from the Caucasoid branch seems to be an earlier event than the divergence of Amerindians from Asiatics (>14,000 years ago).

The dendrogram obtained results quite similar to that obtained from another study in highly informative immunological polymorphism [Sánchez-Mazas and Langaney 1988]. It is also interesting to point out that the three major groups (African, Caucasoid and Oriental) seem to diverge quite concomitantly, in good agreement with other trees based on nuclear genetic systems (classical markers, restriction fragment length polymorphism and microsatellite) [Cavalli-Sforza *et al* 1988; Bowcock *et al* 1991, 1994], in contrast with those based on mitochondrial DNA, which usually show long African branches [Cann *et al* 1987; Vigilant *et al* 1991].

On the other hand, previous studies that we carried out in the Basque population have shown differences between this population and other populations at the level of visible characteristics (dermatoglyphics) [Arrieta *et al* 1987a,b; Arrieta 1990a; Arrieta 1995; Arrieta and Lostao 1988] and at a cytogenetic level [Arrieta 1996]. This study points out that peculiarities in the Basque population are also present at a molecular level.

The high degree of similarity to other trees and its power to differentiate closely related populations show that microsatellites, and especially those which are more polymorphic, can be useful tools in the establishment of the history of the human species and in its spreading pattern throughout the world.

REFERENCES

AMMERMAN AJ, CAVALLI-SFORZA LL (1984) The neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press.

ARANZADI W (1905) Existe una raza euskara? Sus caracteres antropológicos. San Sebastian: Imp de la Provincia.

ARRIETA MI, LOSTAO CM (1988) Principal component analysis of digital dermatoglyphic patterns in a Basque population. Hum Biol *60*: 191-204.

ARRIETA MI, IBARRONDO MA, LOSTAO CM (1987a) Digital dermatoglyphics in the Basque population: Univariate and multivariate comparison with other Spanish populations. Am J Phys Anthrop *73*: 89-98.