MS205 Minisatellite Diversity in Basques: Evidence for a Pre-Neolithic Component

Santos Alonso[1] and John A.L. Armour

Division of Genetics, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK

A number of studies have suggested that Basques might be a relic of Mesolithic Europeans who escaped much of the homogenization brought about by the Neolithic expansion. In an attempt to add new insights into this hypothesis, MS205 minisatellite diversity has been investigated by Minisatellite Variant Repeat (MVR) analysis

lites have become an increasingly popular alternative in evolutionary analysis and have also been applied to the study of Basque diversity (Iriondo et al. 1997; Garcia et al. 1998). However, size constraint may limit the range of microsatellite alleles (Garza et al. 1995), with recurrent mutation causing an eventual loss of their phylogenetic information (Nauta and Weising 1996; Feldman et al. 1997). Additionally, different microsatellite loci are likely to show different size constraints and different mutation rates.

In this regard, the analysis of loci with lower effective population size, in which drift tends to dominate over mutation, such as those on the Y chromosome (effective size one-fourth of nuclear loci), showed the Basques as an extreme of the European frequency distribution because of their low haplotype-diversity value (Scozzari et al. 1997; but see Perez-Lezaun et al. 1997). Other Y-chromosome polymorphisms support the ''outlier'' behavior of Basques (Lucotte and Hazout 1996; Semino et al. 1996).

Here, we examine Basque diversity in the context of other world populations already studied, using another class of variable number of tandem repeat (VNTR) loci, minisatellites (Armour et al. 1998). The autosomal minisatellite MS205 not only shows huge allelic diversity with a heterozygosity value close to 100% (Armour et al. 1993, 1996) but other additional useful properties: Its size distribution, so far restricted to between 1 and 5 kb, allows Minisatellite Variant Repeat analysis by PCR (MVR–PCR) (Jeffreys et al. 1991). Thus, minisatellite alleles are not only differentiated by length but also by the internal arrangements of variant repeat units. Furthermore, a considerable body of data is available

group, the 155 for t11, or the 87 observed for t12, and most of its lineages share alleles that can be related by a single apparent mutational event. This may be a reflection of a lower mutation rate for this group; in this regard, at least for t10.7, a 10-fold reduction in mutation rate has been observed in analysis of sperm DNA (May et al. 1996). Another slowly mutating allele (May et al. 1996), belonging to the t12 group (5′-ttttttttttttattatatattatttttaa-3′) is also found at relatively high frequency in European samples (four times in 236 Basque alleles, three in the 106 alleles of the Northern European sample, three in the 46 UK alleles, and once in the Castilian sample of 48 alleles). A similar allele (5′-tttttttttttattatatattatttttttaaa-3′) is found five times in Basques only. Differential allele mutation rates may therefore have favored drift to higher frequency for those alleles with the lowest mutation rate. This is especially conspicuous in the Surui population, in which t10.7 accounts for 75% of alleles. The extent to which varying average mutation rates at this minisatellite may explain different evolutionary rates for populations with different allele compositions remains to be studied.

The parameter $\upsilon = 4N_e\mu$ has been estimated as $\upsilon_H = 135.9 \pm 23.6$. This allows us to estimate the long-term mean effective population size. Thus, for the total Basque population, using an average mutation rate of 0.4%, an approximate value of 8500 is obtained, in fair agreement with the conventional Figure of 10,000 (Harpending et al. 1998). An estimate (υ_k) of the same parameter υ from the number of observed alleles can be used instead to obtain the expected number of rare alleles (in this case, arbitrarily, those found only once) under an infinite allele model (IAM) and compare this to the observed value. For the Basque population, there is a nonsignificant excess of observed rare alleles (100 observed vs. 96.4 ± 9.8 expected, $P(n \geq 100) = 0.37$; $\upsilon_k = 162.3 \pm 23.1$), suggesting a fit to the model.

To represent the similarities between populations graphically, the allele frequency distributions were used as an initial approach by means of correspondence analysis (Fig. 1). The first two factors explained only ~17% of the total inertia. Axis 1 showed differentiation of the African populations, whereas axis 2, apart from displaying the outlier behavior of Melanesians, shows Basques forming a homogeneous group with the rest of the European populations.

It can be argued that this approach disregards any information on phylogenetic relationships between alleles. We have therefore applied a procedure using values intended to reflect the average phylogenetic dissimilarity of alleles within and between populations. The permutation test of genetic differentiation between populations (Table 1) shows in principle the same population relationships. Although we need to bear in mind that the small sample size for some of the populations analyzed may reduce the power of the test employed (Hudson et al. 1992) and that more extensive bootstrapping would help reducing the variance associated with

Table 1. Test for Genetic Differentiation between Populations

Basques being a divergent European population that has retained a greater proportion of more ancestral alleles.

In this sense, estimates of allele age may provide a clue to the minimum age of a population. The advantage of MS205 in this regard is that its high mutation rate and its fit to IAM increase the chances for the generation of a number of population-specific alleles (or groups of alleles). If it is possible to estimate their age, we can infer a minimum estimate of the age of the population to which they belong. However, reconstructing the phylogeny of all groups of alleles is not straightforward, as some mutational events may produce, by chance, similar-by-state alleles (the chances of producing an allele identical to previously existing ones by recurring mutation are more remote, especially if alleles with an unusual internal structure are considered). Thus, estimation of allele ages has been restricted to independent single alleles rather than lineages until additional information is gathered.

For the Basques, allele t14.1 (58-ttttttttttttttatat-tattttatta-38) is the best example of a population-specific allele; it has only been found in Basques (four copies have been observed in 236 Basque alleles), and no other allele belonging to the t14 group has been detected in any world population.

We have based our analysis in a present day population size for the (autochthonous) Basques, of 0.5 million individuals. Calderon et al. (1998) argue that the present-day population in the Basque Country would be ~3 million people and that not more than one-fourth334

Table 2. Estimated Ages for Some Basque-Specific Alleles

Alleles	No. of copies each	Age (years)	Confidence interval
t8.39, t8.56, t10.23, t11.43, t11.61, t11.76, t11.115, t11.129, t12.33, t12.44, t12.65 , t12.77, t12.92, t17.6	2	15,079	(3,720–28,240)
t8.51, t11.37, t11.143, t12.39, t12.72	3	17,805	(8,860–30,360)
t12.57, t14.01	4	19,441	(11,260–31,700)
t8.63, t12.21	5	20,554	(12,900–32,760)
t8.45	6	21,442	(14,100–33,580)

confidence limits of 11,260–31,700 years. Other examples shown in Table 2 yield similarly pre-Neolithic estimates for the allele ages.

It can be argued that sample sizes analyzed so far are responsible for the nondetection of at least some of the alleged Basque-specific alleles in other world populations. Although this can hold true for some of them, many alleles observed twice have only been detected in a specific Basque subpopulation. On the other hand, one should expect alleles with a high frequency to be observed at least once in some other world population sample if they represent common ancestral alleles.

Calculated allele ages are likely to represent minimum ages, as we cannot ascertain if other closely related alleles, population specific or not, are derived from the allele under consideration or from other alleles observed in other populations. If the latter holds true, our inferences are not affected; if the former, these alleles would be part of the same lineage, and in discarding them we are not taking into account their frequency, thus producing younger estimates of the population age.

DISCUSSION

Although the Upper Paleolithic period offers a number of archaeological sites in the Basque Country, all of them in caves, the anthropological reconstruction of the local ethnogenic processes suffers from a scarcity and fragmentation of human paleontological remains (de la Rua 1995 and references therein). In this scenario, the analysis of genetic polymorphisms in the extant Basque (and world) population stands as a useful and complementary approach to unveil our links to the past.

As regards our more distant past, the comparison of the classical heterozygosity values to the intrapopulational averaged diversity values (Table 1) shows clearly that although most of the non-African populations have acquired high levels of diversity, this is mainly associated with mutations occurring in the hypervariable 3′ end that have generated a huge number of closely related alleles. This in principle would agree with the accumulation of diversity after an expansion event, in which genetic traits associated with low mutation rates show less post-expansion diversity compared to traits with higher mutation rates (Relethford 1997). Africans, on the other hand, despite showing similar heterozygosity values to Europeans, have a higher intrapopulational averaged allele dissimilarity (Table 1), reflecting a higher degree of slower-evolving, older diversity. This is consistent with African populations evolving for longer times and/or with greater long-term effective population sizes.

This study points to a European affiliation for the Basques. Nonetheless, the observation that a significant proportion of apparently Basque-specific alleles can be dated back to post-Aurignacian indicates a continuity of this population from prehistoric times several thousand years after the arrival of modern *Homo sapiens* in Europe, to the present day.

Archeological data suggest that the expansion of typically Aurignacian technology [attributed specifically to modern *H. sapiens* and most likely originating in the Middle East ~100 thousand years ago (kya)] into Europe, and the dispersal of the associated populations, could be linked to an East–West cline. However, strong evidence indicates that Aurignacian was already present in Northwestern and Northeastern Spain at least by 40 kya, earlier than in Southwestern France (~35 kya); here, it coexisted for several thousand years with the typical Neanderthal-associated Chatelperronian culture, which penetrated for a short distance into the Pyrenees and adjacent Northern Spain (Mellars 1992). Similarly, late Mousterian (Neanderthal) also coexists in some

sites in Southern Spain, like the Zafarraya site, where it seems to have persisted until well after the Last Pleniglacial. Solutrean and Magdalenian industries appear to arrive at the Cantabrian fringe (North

tion from the expected heterozygosity under the infinite allele model, following the procedure of Chakraborty and Daiger (1991). The expected number of rare alleles, and the significance of the difference from the observed value was also calculated as in Chakraborty and Daiger (1991) using u_k, the iteration estimator of u based on the expected number of alleles.

Multivariate Factor Correspondence analyses of allele frequencies in populations were performed using the ADE-4 package (Thioulouse et al. 1997).

To gain a deeper insight into the phylogenetic relationships among populations, customized computer programs were developed. One of them compares all pairs of alleles in the global sample and establishes a distance value for each comparison. In this procedure, data on the rate and pattern of mutational events at MS205, revealed by analysis of sperm mutants by SP–PCR (Jeffreys et al. 1994; May et al. 1996) were considered. The polarity of the mutation (with a greater penalty going to those mutations inferred to have taken place at the 58 end compared to those at the 38 end), the length of the region involved, and the kind of mutation involved are factors in the program. Thus, relationships among alleles differing by small insertion/deletion events, terminal and subterminal duplications of small blocks of repeats, and terminal and subterminal short gene-conversion events are included, as well as comparisons of the numbers of t-repeat types in the first run of ts from the 58 end.

Based on Nei's minimum distance (1987) and the works of Hudson et al. (1992) and Shriver et al. (1995), the averaged intrapopulational diversity, a value representing the average dissimilarity between pairs of alleles for each population, was estimated as

$$K_{\text{intra}.x} = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} / n^2$$

where n is the total number of alleles in the sample and d_{ij} the calculated (as indicated above) distance value for each corresponding allele pair. This figure represents the comparison of a population to itself and matches the classical heterozygosity

$$\left(1 - \left({}_i p_i^2\right)\right.$$

when a binary code (0 for identical, 1 for different alleles) is used as the distance value between alleles, disregarding any phylogenetic information. Confidence intervals were estimated by bootstrapping 100 times over alleles.

Subsequently, an averaged interpopulation diversity value between populations x and y,

$$K_{\text{inter}.xy} = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} d_{ij} / \sim n_x n_y !$$

was calculated by comparing all alleles in population x to all alleles in population y, where n_x and n_y are the corresponding sample sizes (numbers of alleles). This dissimilarity value was used as a measure of genetic distance as

$$d_{xy} = K_{\text{inter}.xy} \; 1 \; [(K_{\text{intra}.x} + K_{\text{intra}.y})/^2]$$

An estimate of the sampling distribution of this parameter d_{xy} was obtained by lumping both populations under comparison together and generating 100 pairs of random partitions of the total, each time of the same size as the two original samples, by shuffliD[(s 1 Tf0.6=d)Tj t2(w-500(20,000)- 1 Tf0.6=times.)-500(The)-500(number)- 1 Tf0.6=of times (in the(i(100)-391(simulations))-3imethe parameter was equal to greater55(5(than)-uao(the)-hanobserved)-4(to)netoas significance value of a testof fortesdifferentiationbetween populations.

The genetic distance among populations defined abovewas wasto gener

J.R. Kidd, K.K. Kidd, J. Bertranpetit, S. Pääbo, and A.J. Jeffreys. 1996. Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.* 13: 154–160.

Armour, J.A.L., S. Alonso, S. Miles, L.J. Williams, and R.M. Badge. 1998. Minisatellites and mutation processes in tandemly repetitive DNA. In *Microsatellites: Evolution and applications* (ed. D.B. Goldstein and C. Schlötterer), Oxford University Press, Oxford, UK.

Bertranpetit, J. and L.L. Cavalli-Sforza. 1991. A genetic
t tmlern T¶tanTfŽ31ºÑth10ecD m Ta

rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Hum. Mol. Genet.* 5: 1823–1833.

Mellars, P.A. 1992. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. *Phil. Trans. R. Soc. Lond. B* 337: 225–234.

———. 1998. The Upper Paleolithic revolution. In *Prehistoric Europe. An illustrated history*, pp. 42–78 (ed. B. Cunliffe), Oxford University Press, Oxford, UK.

Mourant, A.E. 1947. The blood groups of the Basques. *Nature* 160: 505–506.

Nauta, M.J. and F.J. Weissing. 1996. Constraints on allele size at microsatellite loci: Implications for genetic differentiation. *Genetics* 143: 1021–1032.

Nei, M. 1987. *Molecular evolutionary genetics.* Columbia University Press, New York, NY.

Page, R.D.M. 1996. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12: 357–358.

Perez-Lezaun, A., F. Calafell, M. Seielstad, E. Mateu, D. Comas, E. Bosch, and J. Bertranpetit. 1997. Population genetics of Y chromosome short tandem repeats in humans. *J. Mol. Evol.* 45: 265–270.

Relethford, J.F. 1997. Mutation rate and excess African heterozygosity. *Hum. Biol.* 60: 785–792.

Richards, M., H. Côrte-Real, P. Forster, V. Macaulay, H. Wilkinson-Herbort, D. Demaine, S. Papiha, R. Hedges, H.J. Bandelt, and B. Sykes. 1996. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* 59: 185–203.

Richards, M., V. Macaulay, B. Sykes, P. Pettitt, R. Hedges, P. Forster, and H.J. Bandelt. 1997. Reply to Cavalli-Sforza and Minch. *Am. J. Hum. Genet.* 61: 251–254.

Rogers, A.R. and L.B. Jorde. 1996. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet.* 58: 1033–1041.

Royle, N.J., J.A.L. Armour, M. Webb, A. Thomas, and A.J. Jeffreys. 1992. A hypervariable locus D16S309 located at the distal end of 16p. *Nucleic Acids Res.* 20: 1162.

Sajantila, A. and S. Pääbo. 1995. Language replacement in Scandinavia. *Nat. Genet.* 11: 359–360.

Sajantila, A., P. Lahermo, T. Anttinen, M. Lukka, P. Sistonen, M. Savontaus, P. Aula, L. Beckman, L. Tranebjaerg, T. Gedde-Dahl et al. 1995. Genes and languages in Europe: An analysis of mitochondrial lineages. *Genome Res.* 5: 42–52.

Scozzari, R., F. Cruciani, P. Malaspina, P. Santolamaza, B.M. Ciminelli, A. Torroni, D. Modiano et al. 1997. Differential structuring of human populations for homologous X and Y microsatellite loci. *Am. J. Hum. Genet.* 61: 719–733.

Semino, O., G. Passarino, A. Brega, M. Fellous, and A.A. Santachiara-Benerecetti. 1996. A view of the Neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am. J. Hum. Genet.* 59: 964–968.

Shriver, M.D., L. Jin, E. Boerwinkle, R. Deka, R.E. Ferrell, and R. Chakraborty. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci.