Recent Male-Mediated Gene Flow over a Linguistic Barrier in Iberia, Suggested by Analysis of a Y-Chromosomal DNA Polymorphism

Matthew E. Hurles,[1,*] Reiner Veitia,[2,*] Eduardo Arroyo,[3] Manuel Armenteros,[4]
Jaume Bertranpetit,[5,†] Anna Pérez-Lezaun,[5,†] Elena Bosch,[5,†] Maria Shlumukova,[1,‡]
Anne Cambon-Thomsen,[6] Ken McElreavey,[2] Adolfo López de Munain,[7] Arne Röhl,[8]
Ian J. Wilson,[9,§] Lalji Singh,[10] Arpita Pandya,[11] Fabrício R. Santos,[11,‖] Chris Tyler-Smith,[11]
and Mark A. Jobling[1]

[1]Department of Genetics, University of Leicester, University Road, Leicester, United Kingdom; [2]Unité d'Immunogénétique Humaine, Institut Pasteur, Paris; [3]Universidad Complutense de Madrid, Facultad de Medicina, and [4]Centro de Investigacion y Criminalistica, Laboratorio de ADN, Policia Judicial, Guardia Civil, Madrid; [5]Unitat d'Anthropologia, Departement de Biologia Animal, Universitat de Barcelona, Barcelona; [6]INSERM U518, Faculté de Médecine, Toulouse; [7]Servicio de Neurologia, Arantzazuko Amaren Ospitalea, San Sebastian, Spain; [8]Mathematisches Seminar, Universität Hamburg, Hamburg; [9] School of Biological Sciences, Queen Mary and Westfield College, London; [10]Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad; India; and [11]CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, Oxford

Summary

We have examined the worldwide distribution of a Y-chromosomal base-substitution polymorphism, the T/C transition at SRY-2627, where the T allele defines haplogroup 22; sequencing of primate homologues shows that the ancestral state cannot be determined unambiguously but is probably the C allele. Of 1,191 human Y chromosomes analyzed, 33 belong to haplogroup 22. Twenty-nine come from Iberia, and the highest frequencies are in Basques (11%; $n = 117$) and Catalans (22%; $n = 32$). Microsatellite and minisatellite (MSY1) diversity analysis shows that non-Iberian haplogroup-22 chromosomes are not significantly different from Iberian ones. The simplest interpretation of these data is that haplogroup 22 arose in Iberia and that non-Iberian cases reflect Iberian emigrants. Several different methods were used to date the origin of the polymorphism: microsatellite data gave ages of 1,650, 2,700, 3,100, or 3,450 years, and MSY1 gave ages of 1,000, 2,300, or 2,650 years, although 95% confidence intervals on all of these figures are wide. The age of the split between Basque and Catalan haplogroup-22 chromosomes was calculated as only 20% of the age of the lineage as a whole. This study thus provides evidence for direct or indirect gene flow over the substantial linguistic barrier between the Indo-European and non–Indo-European–speaking populations of the Catalans and the Basques, during the past few thousand years.

Introduction

Most regions of sharp genetic change within Europe correspond to linguistic boundaries (Barbujani and Sokal 1990). The Basques speak a non–Indo-European language with no close affinities to any other extant language (Ruhlen 1991), and this linguistic uniqueness has led to the idea that the Basques may represent a Mesolithic relict population, isolated, by linguistic and geographic barriers, from cultural and genetic exchange. Although specific archaeological evidence for such a picture is lacking (Collins 1986), genetic analysis certainly lends some support to the view of the Basques as an isolate, in light of unusual frequencies of alleles in blood groups such as rhesus and ABO (Mourant 1947, 1983) and of disease alleles in the calpain-3 gene that are responsible for limb-girdle muscular dystrophy (Urtasun et al. 1998). Albeit to a lesser extent, mtDNA sequences (Bertranpetit et al. 1995; Côrte-Real et al. 1996; Comas et al. 1997) and HLA types (Comas et al. 1998) also support this view. The use of multiple autosomal loci in principal-component analysis (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994*a*, 1994*b*), in the calculation of various genetic-distance measures (Calafell and Bertranpetit 1994*b*) and in a method de-

signed to detect genetic boundaries (Barbujani 1991), clearly distinguishes the Basque region from the rest of the Iberian peninsula.

The majority of the Y chromosome is nonrecombining, and so mutations on it represent a record of its evolutionary past that can be used in the reconstruction of population histories (Jobling and Tyler-Smith 1995; Mitchell and Hammer 1996). The presence of different polymorphic systems with different mutation rates and processes is a particular strength of the Y chromosome and allows us to use Y-chromosome markers "genealogically," defining lineages ("haplogroups") with slowly mutating biallelic polymorphisms such as base substitutions, which can be regarded as unique events in human evolution, and then examining diversity within these haplogroups, using polymorphisms that mutate more rapidly, such as microsatellites (Kayser et al. 1997) and the minisatellite, MSY1 (Jobling et al. 1998). This alleviates the problem of recurrent mutation at these loci and allows attempts to be made to date haplogroup origins.

A comparison of the correlation of languages with Y-chromosomal haplotypes (defined by the marker 49f [Ngo et al. 1986]) and with mtDNA haplotypes has suggested that the passing on of language from generation to generation is governed more by patrilineage than by matrilineage (Poloni et al. 1997). This certainly appears to be so for the Uralic-speaking Finns, who share most of their mtDNA lineages with Indo-European speakers (Lahermo et al. 1996) but, in contrast, approximately half of their Y-chromosome lineages with Central Asian Uralic speakers (Zerjal et al. 1997). In the case of the Basques, with their unique linguistic heritage, it is of particular interest to know whether their Y chromosomes are distinct from those of surrounding populations. Although a statistically significant difference has been shown between Basques and other populations, including Catalans, in studies using 49f (Lucotte and Hazout 1996; Poloni et al. 1997), a study using Y-chromosomal microsatellites (Pérez-Lezaun et al. 1997) finds no such difference. Here, we show that a specific Y-chromosomal lineage, which has a recent origin and is rare or absent in most parts of the world, is shared at high frequency between Basques and Catalans. This constitutes evidence for substantial recent male-mediated gene flow over a major linguistic barrier.

Subjects and Methods

Subjects

Gifts (providers) of DNA samples from autochthonous males, defined in most cases on the basis of grandpaternal birthplace, were as follows: Castilians (Santos Alonso and John Armour), Galicians (Marisol Rodri-

guez-Calvo), León (Carlos Polanco), Belarusians (Yuri Dubrova), Germans (Manfred Kayser), and other DNA samples (our own collections). The Catalan samples (from Girona) and 51 of the Basque samples (from Guipúzcoa [and denoted by the suffix "v" in table 2]) have been described elsewhere (Pérez-Lezaun et al. 1997); Basque samples denoted as "m336"–"m365" are from Zumaya in Guipúzcoa, and the remaining Basques are from Pyrenées Atlantiques. All samples were taken with appropriate informed consent.

DNA Sequencing

Direct sequencing of 1.2-kb PCR products amplified from human and primate DNAs by use of the SRY-2627 primers R1 and F1 (Veitia et al. 1997) was performed by use of F1 as sequencing primer and with BigDye technology (Perkin-Elmer) on an ABI377 sequencer (Applied Biosystems).

Typing of SRY-2627 and 92R7

SRY-2627 (previously referred to as "SRY-2628" [Veitia et al. 1997]) was typed by PCR using the R1 and F1 primers (Veitia et al. 1997), followed by *Bsi*HKAI digestion (fig. 1*A*). Apparent T-allele chromosomes were verified by use of *Ban*I. A 709-bp amplicon containing the 92R7 (Mathias et al. 1994) polymorphism was amplified by use of the primers 5′-GAC CCG CTG TAG ACC TGA CT-3′ and 5′-GCC TAT CTA CTT CAG TGA TTT CT-3′, in an MJR PTC-200 thermal cycler (33 cycles of 94°C for 30 s, 62°C for 30 s, and 72°C for 60 s). Then, typing by *Hin*dIII digestion was done, to give 197- and 512-bp fragments from the C allele; the 709-bp product remains, since there is more than one copy of the locus on the Y chromosome, and since only a subset contains the polymorphic site. In this study, haplogroup 1 is defined by the 92R7 T allele in the presence of the SRY-1532 G allele (Hurles et al. 1998).

Microsatellite and MSY1 Haplotyping

Seven Y-specific microsatellites (*DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392,* and *DYS393*) were typed. Primer sequences are those given by Kayser et al. (1997).

PCR reactions (94°C for 10 min; 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and then 72°C for 10 min) were performed in the presence of a fluorescently labeled dUTP derivative (R6G; Applied Biosystems) diluted 1:600 with respect to dTTP. Products were electrophoresed on an ABI373A sequencer, and data were analyzed by GeneScan software (Applied Biosystems). Three-state microsatellite variant repeat (MVR)-PCR at the minisatellite MSY1 was performed as described elsewhere (Jobling et al. 1998).
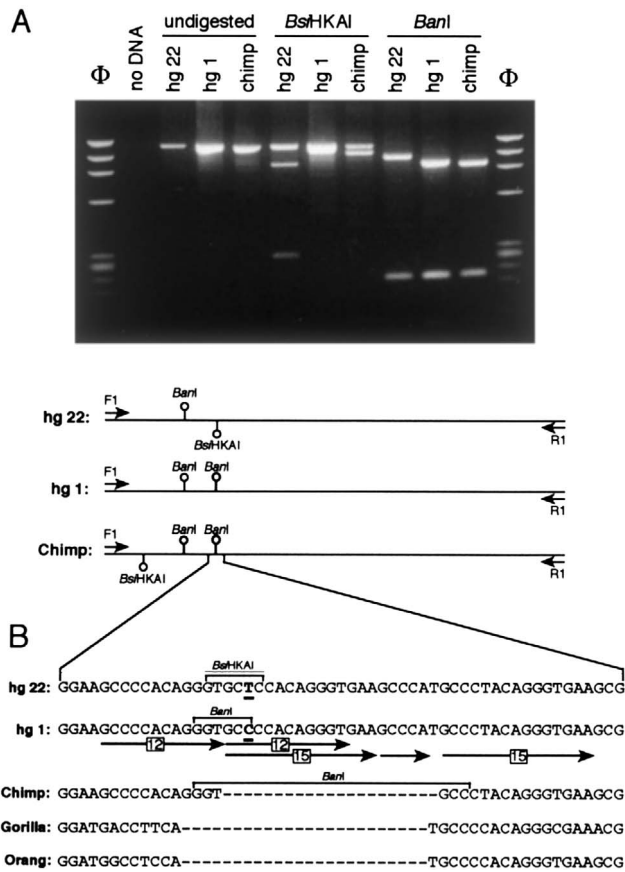
Figure 1 SRY-2627 polymorphism. *A,* Detection by PCR-RFLP analysis, and restriction map of PCR fragment. Digestion of the 1,242-bp product by *Bsi*HKAI in the T allele (haplogroup 22) chromosomes gives fragments of 945 and 297 bp (digests shown are partial). Products of 946, 210, and 86 bp are obtained by digestion of the C allele (here, haplogroup 1) chromosomes with *Ban*I. The fragment generated from chimp DNA contains a *Ban*I site at the polymorphic position and therefore resembles the human C allele. hg = Haplogroup; Φ = *Hae*III-digested ΦX174 phage DNA. Short arrows indicate PCR primers. *B,* Sequences of the SRY-2627 C and T alleles and of the homologous regions in DNAs of great apes. Arrows under the haplogroup 1 sequence indicate complete and partial copies of 12- and 15-bp direct repeats. Because of these, several alignments are possible, and the ancestral state cannot be unambiguously deduced; however, the polymorphism is at the third base of one of the 15-bp repeats in human DNA, and all other copies of the repeat, in both humans and apes, share a C at this position, suggesting that C is ancestral. In the alignment shown, the *Ban*I site in chimp DNA is not homologous to that in DNA of the human C allele.

Microsatellite and Minisatellite Network Construction and Dating

Median joining networks were constructed by Network 1.1 (Bandelt et al. 1999); *DYS389I*-allele lengths (tables 2 and 3) were subtracted from *DYS389II*-allele lengths prior to analysis, since the former is contained within the latter (Cooper et al. 1996). Dating was done separately, by use of both microsatellite and MSY1 data,

by three different methods; for a description, see the Age of the SRY-2627 Mutation subsection. For two of the methods (Bertranpetit and Calafell 1996; Thomas et al. 1998), the root of a haplotype tree (microsatellite haplotype 21; MSY1 haplotype 6 [see table 2]) was chosen from pairwise differences, as that having the smallest number of mutational steps from all other chromosomes; this is identical to a haplotype constructed from the modal allele lengths. Note that uncertainty in root assignment has only a minimal effect on dating (data not shown). The average squared distance (ASD [Thomas et al. 1998]) was calculated, by Microsat 1.5d (Minch 1997), between a population of chromosomes and the root haplotype (this method was also used to date the divergence between Basque and Catalan haplogroup-22 chromosomes and to estimate diversity differences between haplogroups 1 and 22 and between European and Asian haplogroup-1 chromosomes). For all methods, we have assumed a generation time of 25 years and mutation rates of 2%–11% per generation (Jobling et al. [1998]; we also use the midpoint of this range, 6.5%), for MSY1, and 2.1×10^{-3} (95% confidence limits [95% CI] 0.6–4.9×10^{-3} [Heyer et al. 1997]), for microsatellites; for the third method (Goldstein et al. 1996), we consider N_e to be 4,900 (Hammer 1995). Throughout, mutation in each MSY1 repeat block is weighted for repeat number (Hurles et al. 1998), on the assumption that each repeat unit has an equal probability of mutating.

A fourth, coalescent-based method also was used, for microsatellite data only; it differs from a published one (Wilson and Balding 1998), by allowing for exponential growth in the population of Y chromosomes. This method uses a Markov-chain Monte Carlo simulation algorithm to generate simulated trees consistent with the observed haplotype data, sampling 10,000 of these trees at a rate proportional to their probability under a coalescent-with-exponential-growth model. No prior assumption is made about population size, but mutations are assumed to be stepwise, and the prior mutation rate is the same as that used in the other methods (Heyer et al. 1997). The output from this method includes probability distributions for T (tree height) and N (population size), from which a probability distribution for the time to the most recent common ancestor can be derived, again by use of a 25-year generation time. The standard coalescent model assumes that haplotypes are sampled at random from the whole population. However, when a population is growing rapidly, the coalescent-with-exponential-growth will be a good approximation to the genealogy of a haplogroup. Readers interested in this method are asked to contact I.J.W. Pairwise R_{ST} analysis was performed by use of the Arlequin package (Schneider et al. 1997).

Results

Ancestral State of the SRY-2627 Base-Substitution Polymorphism

A T/C transition polymorphism 2,627 bp 5′ of the start codon of the *SRY* gene has been described in previous reports (Bianchi et al. 1997; Veitia et al. 1997); the T allele defines a Y-chromosomal haplogroup that we term "haplogroup 22" and can be conveniently typed in a PCR-RFLP assay, since the T allele alone creates a *Bsi*HKAI site and the C allele alone creates a *Ban*I site (fig. 1*A*). In these previous reports, no direct evidence concerning ancestral state was available, and the rarer (T) allele was assumed to be the derived one. We compared the human sequence with its homologues in the great apes, in an attempt to determine ancestral state definitively (fig. 1*B*). This analysis shows that SRY-2627 lies within a region of direct repeats that has undergone a modest expansion in the human lineage since the human-chimp divergence. Therefore, ancestral state remains ambiguous, although a comparison of the sequences of different repeat copies suggests that C is far more likely than T.

Worldwide Survey for Haplogroup-22 Chromosomes

To examine the worldwide distribution of haplogroup 22, we performed an initial survey of 752 Y chromosomes. We found 10 haplogroup-22 chromosomes: singletons were found in England, in Germany, in a general French sample, and in a sample from a southwestern French population, the Béarnais, but the remaining six chromosomes were all found in Basques. In previous studies, SRY-2627/T-allele chromosomes had been found in France (Veitia et al. 1997) and also in South America (Bianchi et al. 1997), where they were at highest frequency in nonindigenous groups, who are likely to have an Iberian origin. We therefore intensified our survey within Iberia itself; practical difficulties in obtaining French DNA samples precluded a detailed survey of this region.

Haplogroup-22 Chromosomes in Iberia and France

We typed a further 439 Y chromosomes from Iberia, for SRY-2627, making a total of 469 Iberian chromosomes and 1,191 chromosomes worldwide. This survey yielded a further 23 chromosomes from haplogroup 22. The global distribution of this haplogroup is shown in table 1 and figure 2*A*, and the distribution within Iberia is shown in more detail in figure 2*B*. When the data are summed (and the very small Valencian sample is excluded), the populations in which we find haplogroup 22 at its highest frequency are the Basques (11%) and

Table 1

Populations Tested for SRY-2627, and Summary of Results

CONTINENT AND POPULATION	No. of Alleles	
	C	T
Europe:		
Iberia:		
Basque	104	13
Catalan	25	7
Galician	46	2
Andalucian	83	1
Madrid (urban)	60	2
Castilla la Mancha	23	0
Castilla y León	47	0
Extremaduran	18	0
Valencian	5	1
Other Spanish (not Basque or Catalan)	58	3
Other:		
Béarnais	13	1
French	33	1
British	63	1
German	48	1
Italian	39	0
Greek	19	0
Hungarian	34	0
Icelandic	27	0
Norwegian	46	0
Belarusian	23	0
Caucasus	16	0
Other	8	0
Asia:		
Chinese	40	0
Japanese	25	0
Indian	86	0
Gujarati	61	0
Sri Lankan	13	0
Tibetan	3	0
Mongolian	23	0
Other	5	0
Africa:		
Algerian	27	0
San	6	0
Biaka	4	0
Kenyan	7	0
Other	3	0
Oceania:		
Cook Islander	6	0
Australian	4	0
Melanesian	2	0
Americas:		
Various	5	0
Total	1,158	33

the Catalans (22%), whereas in other parts of Iberia that were sampled these chromosomes are rare or absent.

Diversity Assessed by Use of Y-Specific Microsatellites and a Minisatellite

To illuminate the geographic and temporal origin of these haplogroup-22 chromosomes, we first examined their diversity, using seven highly polymorphic Y-specific
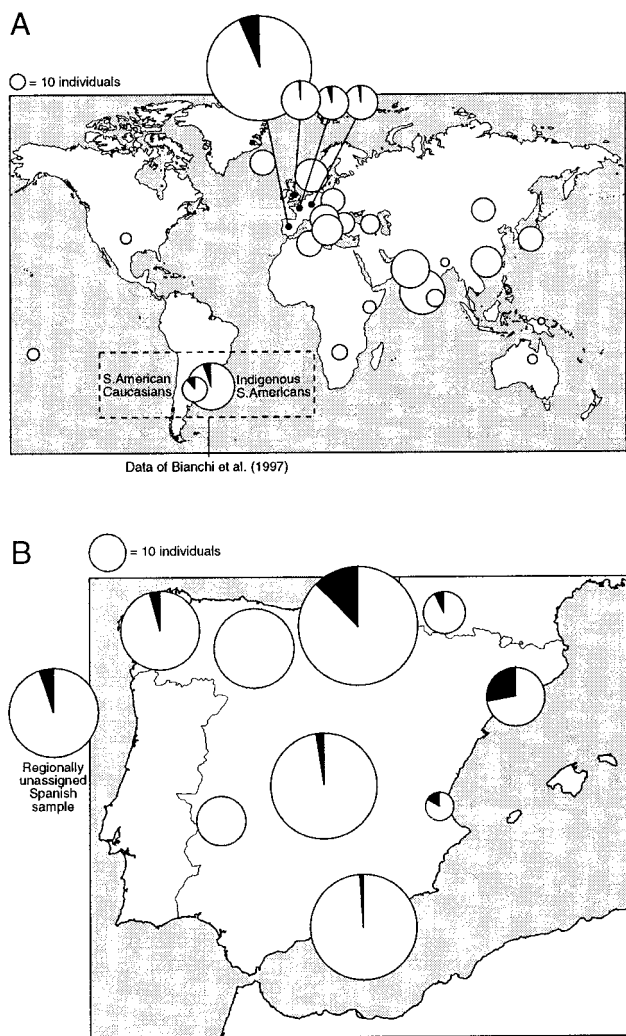
Figure 2   Geographical distribution of haplogroup-22 chromosomes. *A,* Worldwide distribution (see table 1). Data for South America are from Bianchi et al. (1997), as follows: pooled indigenous South Americans, 5/93 SRY-2627/T allele chromosomes; La Plata nonindigenous groups, 3/26 SRY-2627/T allele chromosomes. *B,* Distribution within Iberia and the Béarnais of southern France. Data on Madrid and Castilla la Mancha have been pooled.

microsatellites. Haplotypes were determined (table 2), and a median-joining network was constructed (fig. 3*A*). In this analysis, we included 2 SRY-2627/T-allele chromosomes of French origin that previously had been identified within the pedigrees catalogued by the Centre d'Étude du Polymorphisme Humain (CEPH) (Bianchi et al. 1997) and 1 SRY-2627/T-allele chromosome from the study by Veitia et al. (1997), together with the 33 identified in the present study, making a total of 36.

In addition, we determined the microsatellite diversity of 50 Asian and European Y chromosomes belonging to haplogroup 1 (table 3 and fig. 3*B*). This haplogroup is distinguished from haplogroup 22 only by the SRY-

2627 mutation and—if it is assumed that the ancestral state of SRY-2627 is indeed the C allele (Bianchi et al. 1997; Veitia et al. 1997)—is the ancestral haplogroup. When we compare the two microsatellite networks (fig. 3*A* and *B*), it is clear that haplogroup 1 is substantially more diverse; calculated ASD values are 0.290 for haplogroup 22, compared with 1.063 for haplogroup 1. Together with the much wider geographic distribution of haplogroup 1 (Santos and Tyler-Smith 1996), this is additional evidence that the ancestral state of SRY-2627 is the C allele.

We also determined "MSY1 codes," by MVR-PCR for the haplogroup-22 chromosomes (table 2); in this technique, the positions of three different classes of variant 25-bp repeat units along the MSY1 minisatellite array are mapped by use of discriminator primers specific to individual repeat types. Compared with other haplogroups analyzed (Jobling et al. 1998), this haplogroup has low diversity: four pairs, one set of three, and one set of seven males have identical MSY1 codes. Of the 35 chromosomes analyzed, 32 have MSY1 codes with the same modular structure (the order of blocks of different repeat types along the array), "1,3,4." If a single-step mutation model is assumed, these codes can also be assembled into a compact network (fig. 3*C*). Three chromosomes (all Iberian) have codes with structures that are more complex, and they are omitted from the network and from the dating calculations described below.

We can envisage two scenarios to explain the current geographic distribution of haplogroup 22: either the SRY-2627 mutation occurred outside Iberia, and individuals carrying it migrated into Iberia, where subsequent drift led to the high frequency in this region; or, alternatively, the origin was in Iberia, and the non-Iberian cases are explained by emigration. If the first explanation is correct, then we might expect that non-Iberian haplogroup-22 chromosomes would have higher haplotype diversity than is present in Iberian haplogroup-22 chromosomes and that these non-Iberian cases would include haplotypes at the peripheries of the networks, with the Iberian chromosomes forming a tighter cluster. This kind of partitioning is vividly displayed in the haplogroup-1 microsatellite network (fig. 3*B*), where Asian chromosomes lie at the network's periphery, consistent with an origin of this haplogroup outside Europe (see Karafet et al. [1999], who refer to the equivalent class of chromosomes as "haplotype 1C"), and show much more diversity than European chromosomes: diversity, measured in terms of ASD, is 1.762 for Asians, whereas it is only 0.359 for Europeans, a difference that is also supported by principal-components analysis (data not shown). In this scenario, we would also expect to see sharing of Asian and European haplotypes in the core of the network, and the absence of such sharing

Table 2

Microsatellite Haplotypes and MSY1 Codes of Haplogroup-22 Chromosomes

INDIVIDUAL (POPULATION)	NO. OF REPEATS UNITS IN[a]							MICROSATELLITE HAPLOTYPE[b]	MSY1	
	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393		Code[c]	Haplotype[d]
m337 (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{40}(4)_{18}$[e]	10
4301 (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{38}(4)_{19}$	7
46205 (Basque)	14	10	27	24	10	13	13	14	$(1)_{17}(3)_{38}(4)_{14}$	18
35v (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{39}(4)_{19}$	9
67c (Catalan)	14	10	27	24	10	13	13	14	$(1)_{15}(3)_{38}(4)_2(3)_1(4)_{15}$	22
m354 (Basque)	14	11	28	24	11	13	13	32	$(1)_{16}(3)_{38}(4)_{19}$[e]	7
m362 (Basque)	14	11	28	24	11	13	13	32	$(1)_{16}(3)_{38}(4)_{19}$[e]	7
m363 (Basque)	14	11	28	24	11	13	13	32	$(1)_{17}(3)_{37}(4)_{19}$[e]	17
ma19 (Madrid)	14	11	28	24	11	11	13	30	$(1)_{18}(3)_{39}(4)_{19}$	20
m62 (British)	14	10	28	24	11	13	13	33	$(1)_{15}(3)_{38}(4)_{20}$	3
m147 (French)[f]	14	10	29	24	10	13	13	34	$(1)_{16}(3)_{40}(4)_{19}$	11
m348 (Basque)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{41}(4)_{17}$[e]	14
m95 (French)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{39}(4)_{18}$[e]	8
ga29 (Galician)	14	10	27	24	11	13	13	21	$(1)_{14}(3)_{38}(4)_{18}$	1
7c (Catalan)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{38}(4)_{18}$	6
m341 (Basque)	14	10	27	24	12	13	13	35	$(1)_{14}(3)_{39}(4)_{19}$[e]	2
CEPH201 (French)[g]	14	10	27	24	12	13	13	35	$(1)_{16}(3)_{38}(4)_{19}$	7
6201 (Béarnais)	14	10	26	24	10	13	13	17	$(1)_{16}(3)_{38}(4)_{17}$	5
8v (Basque)	14	11	30	24	10	13	13	36	$(1)_{15}(3)_{39}(4)_{18}$	4
32v (Basque)	14	10	26	24	11	13	13	15	$(1)_{16}(3)_{41}(4)_{16}$	13
98v (Basque)	14	11	28	24	10	13	13	37	$(1)_{16}(3)_{38}(4)_{19}$	7
101v (Basque)	14	11	28	24	11	11	13	30	$(1)_{16}(3)_{38}(4)_{19}$	7
41c (Catalan)	14	10	27	23	10	13	13	38	$(1)_{15}(3)_{37}(4)_1(3)_2(4)_{16}$	21
45c (Catalan)	14	10	27	24	10	11	13	39	$(1)_{16}(3)_{41}(4)_{15}$	12
56c (Catalan)	14	10	27	24	11	11	13	25	$(1)_{16}(3)_{38}(4)_{18}$	6
81c (Catalan)	15	10	28	24	11	13	13	40	$(1)_{16}(3)_{38}(4)_{18}$	6
85c (Catalan)	14	10	27	24	10	11	14	41	$(1)_{16}(3)_{38}(4)_{17}$	5
ma20 (Madrid)	14	10	27	24	11	13	14	42	$(1)_{16}(3)_{44}(4)_{14}$	15
sp21 (Spanish)	15	10	26	24	11	13	13	43	$(1)_{16}(3)_{41}(4)_{16}$	13
ga22 (Galician)	14	9	26	24	11	13	13	23	$(1)_{18}(3)_{38}(4)_{17}$	19
sp77 (Spanish)	14	9	26	24	11	13	13	23	$(1)_{16}(3)_2(1)_1(3)_{40}(4)_1(3)_2(4)_{15}$	23
sp79 (Spanish)	15	10	27	24	10	13	13	29	$(1)_{17}(3)_{37}(4)_{19}$	17
sp123 (Valencian)	14	11	28	23	11	13	13	44	$(1)_{16}(3)_{39}(4)_{18}$	8
CEPH3501 (French)[g]	14	10	28	23	11	13	13	45	$(1)_{16}(3)_{38}(4)_{19}$	7
ge3202 (German)	14	10	29	24	11	13	13	46	$(1)_{16}(3)_{44}(4)_{16}$	16
alm1 (Andalucian)	14	11	28	24	11	12	13	47	Not done	…

[a] As defined by Kayser et al. (1997).

[b] See figure 3*A*.

[c] For example, "$(1)_{16}(3)_{40}(4)_{18}$" denotes that, 5′→3′, there are 16 type 1 repeats, 40 type 3 repeats, and 18 type 4 repeats (Jobling et al. 1998).

[d] See figure 3*C*.

[e] As determined by P. G. Taylor (Jobling et al. 1998).

[f] Identified by Veitia et al. (1997).

[g] Identified by Bianchi et al. (1997).

here is a function of the relatively small sample size—and limited geographic distribution—of the Asian chromosomes sampled. In contrast to this, all haplogroup-22 chromosomes cluster tightly in both microsatellite and minisatellite networks (fig. 3*A* and *C*), with, for the microsatellites, no pair of connected haplotypes differing by more than one mutation, indicating a probable Iberian origin.

Age of the SRY-2627 Mutation

From the microsatellite and minisatellite haplotype diversity, we can attempt to date the origin of the SRY-2627 mutation (table 4). We have used three different methods, one (Bertranpetit and Calafell 1996) based on the mean number of mutational steps from the network root, a second (Thomas et al. 1998) based on the ASD
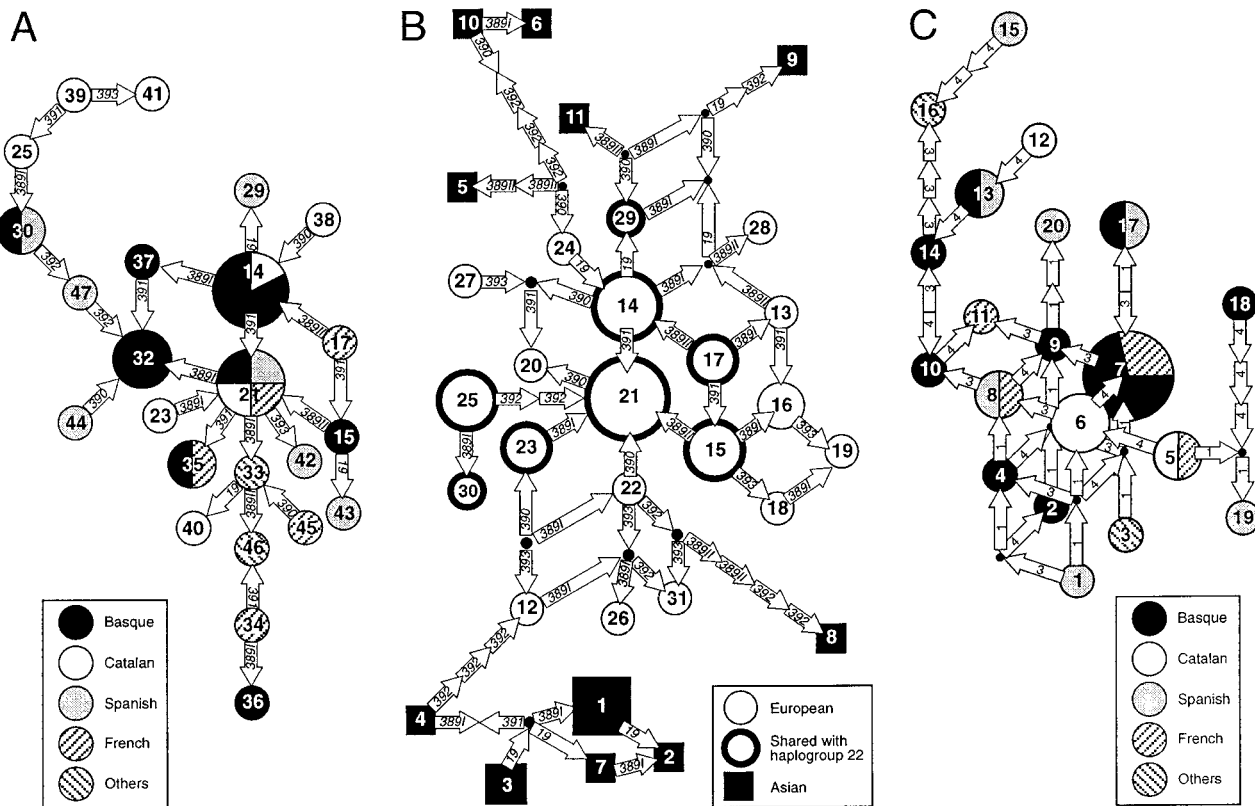
Figure 3   Median joining microsatellite haplotype networks for haplogroup-22 chromosomes (*A*) and haplogroup-1 chromosomes (*B*) and (*C*) MSY1 network for haplogroup 22 chromosomes. A microsatellite or MSY1 haplotype is represented by a circle, with its area proportional to the number of individuals having that haplotype; haplotype number (see tables 2 and 3) is given within each circle. A single arrow between circles represents a single mutational step in a haplotype, and its direction indicates an increase in allele length; the mutating microsatellite (e.g., "*390*," representing *DYS390*) or MSY1 repeat block (e.g., "4," representing the block of type 4 repeats) is indicated on the arrow. Arrows do not imply evolutionary pathways. A small blackened circle on a line indicates an intermediate, unobserved haplotype used as a node (i.e., median vector). Thick edges to circles highlight haplotypes shared by both haplogroups (haplogroup 1 only).

from the root, and a third (Goldstein et al. 1996) based on the extent of variance accumulated since the base substitution occurred on a single haplotype. We also have used a fourth, coalescent-based approach, for microsatellite data alone, that is an extension of a published method (Wilson and Balding 1998) and that has been described in the Subjects and Methods section.

Although the 95% CIs are wide, agreement between the different methods and systems is good (with five of the seven ages within the 2,300–3,500-year range) and suggests that the origin of the SRY-2627 polymorphism occurred a few thousand years before the present. Ages calculated from MSY1 data are consistently younger than those calculated from microsatellite data; this may be due to the omission of the more complex MSY1 alleles from the dating and also to the probable inadequacy of the single-step model for MSY1 mutation. The structure of the MSY1 network in itself provides information on possible deviations from this simple model. There are six connections between haplotypes that involve more

than one repeat unit: of these, four are confined to a single repeat block, and two involve a single-step increase in one repeat block, accompanied by a single-step decrease in an adjacent block. One interpretation of these observations is that multistep mutations can occur within blocks and, possibly, that mutations can occur that simultaneously expand one block and contract a neighboring block, perhaps by switching a boundary repeat from one type to another. Forthcoming direct data on mutation at MSY1 should show whether such events really occur and should allow us to use this locus in a more sophisticated way in the future.

The microsatellite networks for haplogroups 1 and 22 overlap substantially, with eight shared haplotypes; the two most common haplotypes in haplogroup 22 (frequency 9/36 when considered together) are also common (frequency 10/50) in haplogroup 1. This is consistent with an origin of haplogroup 22 on a haplogroup-1 background, followed by much parallel mutation, and insufficient time for substantial divergence of haplo-

Table 3

Microsatellite Haplotypes of Haplogroup-1 Chromosomes

| Individual (Population) | No. of Repeat Units in[a] | | | | | | | Microsatellite Haplotype |
	DYS19	DYS389I	DYS 389II	DYS390	DYS391	DYS392	DYS393	
m419 (Indian)	14	11	27	23	10	10	14	1
m432 (Indian)	14	11	27	23	10	10	14	1
m456 (Indian)	15	11	27	23	10	10	14	2
m464 (Indian)	13	10	27	23	10	10	14	3
m467 (Indian)	14	9	27	23	11	10	14	4
m503 (Indian)	13	10	29	23	10	13	13	5
m507 (Indian)	13	11	27	22	10	16	13	6
m620 (Indian)	15	10	27	23	10	10	14	7
m245 (Mongolian)	14	10	29	23	11	16	13	8
m283 (Mongolian)	16	11	27	23	10	14	13	9
m418 (Indian)	13	10	27	22	10	16	13	10
m434 (Indian)	14	11	27	23	10	10	14	1
m437 (Indian)	15	10	28	23	10	13	13	11
m469 (Indian)	14	11	27	23	10	10	14	1
m470 (Indian)	13	10	27	23	10	10	14	3
2001 (Basque)	14	9	27	23	11	13	14	12
2401 (Basque)	14	11	26	24	10	13	13	13
3101 (Basque)	14	10	27	24	10	13	13	14
3301 (Basque)	14	10	26	24	11	13	13	15
5501 (Basque)	14	11	26	24	11	13	13	16
6103 (Basque)	14	10	26	24	11	13	13	15
2801 (Béarnais)	14	11	26	24	11	13	13	16
3001 (Béarnais)	14	10	26	24	10	13	13	17
3701 (Béarnais)	14	10	26	24	11	13	14	18
3901 (Béarnais)	14	10	26	24	10	13	13	17
4604 (Béarnais)	14	10	26	24	11	13	13	15
4901 (Béarnais)	14	11	26	24	11	13	14	19
m256 (Irish)	14	10	27	25	11	13	13	20
m285 (Norwegian)	14	10	27	24	10	13	13	14
m288 (Norwegian)	14	10	27	24	11	13	13	21
m293 (Italian)	14	10	27	24	11	13	13	21
m366 (Italian)	14	10	27	23	11	13	13	22
7v (Basque)	14	9	27	24	11	13	13	23
16v (Basque)	14	9	27	24	11	13	13	23
21v (Basque)	13	10	27	24	10	13	13	24
23v (Basque)	14	10	27	24	11	11	13	25
24v (Basque)	14	10	27	24	11	13	13	21
25v (Basque)	14	11	27	23	11	13	14	26
30v (Basque)	14	10	27	25	10	13	12	27
41v (Basque)	14	11	28	24	10	13	13	28
44v (Basque)	14	10	27	24	10	13	13	14
8c (Catalan)	14	10	27	24	10	13	13	14
15c (Catalan)	14	10	27	24	11	13	13	21
17c (Catalan)	14	10	27	24	11	13	13	21
20c (Catalan)	14	10	27	24	11	11	13	25
26c (Catalan)	14	10	27	24	11	11	13	25
30c (Catalan)	14	10	27	24	11	13	13	21
32c (Catalan)	15	10	27	24	10	13	13	29
36c (Catalan)	14	11	27	24	11	11	13	30
43c (Catalan)	14	10	27	23	11	14	14	31

[a] As defined by Kayser et al. (1997).

group-22 microsatellite haplotypes from those in haplogroup 1. The same picture is also evident in MSY1 code diversity (Jobling et al. [1998], and data not shown).

Discussion

The Y chromosome has several properties that make it useful for evolutionary studies and that should make

Table 4

Estimates of Age of SRY-2627 Mutation

Method	Microsatellite Age (95% CI) (years)	MSY1 Age (Range) (years)
Mean mutations from root[a]	2,693 (1,154–9,425)	1,107 (604–3,320)
ASD[b]	3,452 (1,480–12,083)	2,632 (1,555–8,554)
Accumulated variance[c]	3,116 (1,166–16,001)	2,328 (1,217–10,143)
Coalescent[d]	1,650 (1,044–8,248)	Not done

[a] From Bertranpetit and Calafell (1996).
[b] From Goldstein et al. (1995).
[c] From Thomas et al. (1998).
[d] From Wilson and Balding (1998).

it simpler to analyze than the "grande dame" of molecular evolutionary biology, mtDNA. One of these properties is the Y chromosome's comparatively low base-substitution mutation rate: in the case of mtDNA, the rate is so high that many polymorphic bases have been multiply substituted since the human-chimp divergence, and trees cannot easily be rooted; on the Y chromosome, in contrast, unambiguous ancestral-state information should be obtained by analysis of the DNAs of other primates. Here, however, we have shown that this is not always straightforward.

Haplogroup-22 chromosomes are rare or absent in most of the world's populations and are most common in Iberia or in populations with substantial Iberian ancestry (Bianchi et al. 1997); within Iberia, the highest frequencies are found in Basques and Catalans, who speak languages belonging to different language families. Either the SRY-2627 mutation was present in a population that was ancestral to both populations and that spoke a single language, or it has occurred since linguistic divergence, implying gene flow over a linguistic barrier.

Contemporaneous evidence on linguistic prehistory does not exist, and, indeed, written records of Basque date back only 900 years (Collins 1986). However, its lack of linguistic relatives strongly suggests that the Basque language is ancient. Theories of Basque origins are many and varied. One theory, "Vasco-Iberism" (Lafon 1972), sees Basque as the last remnant of a language, Ibero, spoken in much of Iberia, including the northern part of modern Catalonia, before the Roman conquest; if this were true, then the linguistic divergence between Basques and Catalans might date back only 2 millennia, and our findings might then be taken to support the hypothesis. However, alleged similarities between Basque and Ibero rest on the scanty evidence of a few inscriptions and place names and are not supported by modern linguists (de Hoz 1995); Vasco-Iberism also seems inconsistent with information from sources such as Greek and Roman geographers (Collins 1986). Alternatively, it might be thought that contraction of

Basque from a previously greater territory could have resulted from the arrival of Indo-European speakers (Barbujani et al. 1994) during the Neolithic period, 4,000–6,000 years ago (Menozzi et al. 1978; Renfrew 1989)—dates that are included in our wide confidence intervals. However, the influence of Indo-European languages here was probably minor, with the non-Basque territory remaining non–Indo-European speaking until the arrival of the Romans.

To explore this issue further, we used microsatellite diversity to calculate ASD between the Basque and Catalan samples within haplogroup 22 and so to estimate the time of divergence of these two populations of haplogroup-22 chromosomes. ASD between all haplogroup-22 chromosomes and the root haplotype is 0.290 (equivalent to μt, where μ is the mutation rate and t the time in generations), and ASD between Basque and Catalan chromosomes, with correction for intrapopulation variance (equivalent to $2\mu t$, since we are no longer considering distance to a root) is 0.115. The age of divergence, as a percentage of the age of haplogroup 22, can then be calculated as the ASD between Basque and Catalan chromosomes, divided by twice the ASD between all haplogroup-22 chromosomes and the root, and is ~20%. Thus, the divergence between these populations of chromosomes is not ancient, and this supports the interpretation that there has been male-mediated gene flow directly between Basques and Catalans since the establishment of the languages now spoken. It also remains possible that haplogroup-22 chromosomes have been contributed to both populations by a third, unsampled population. In either case, genes have flowed over the substantial linguistic barrier that lies between Basque and an Indo-European language.

Can we see evidence of this inferred gene flow in patterns of allele sharing at non–Y-chromosome loci? Published data on mtDNA (Côrte-Real et al. 1996) and HLA (Comas et al. 1998) in Basques and Catalans show no evidence for the sharing of any population-specific alleles or haplotypes. It is, however, striking that, whereas Basque and Catalan samples cluster significantly to-

gether in a neighbor-joining tree based on seven HLA loci (Comas et al. 1998), genetic distances calculated from mtDNA diversity are greatest between Catalans and all other Iberian samples, including Basques (Côrte-Real et al. 1996). This contrast between biparentally and maternally inherited loci may imply that the sharing of Y-chromosomal lineages that we observe is really a result of male-mediated gene flow, with little female-mediated flow and with autosomal markers reflecting an average of the two. Higher-resolution studies of Iberian Y-chromosome diversity, analyzing all available lineages, should further delineate genetic boundaries within this region.

In principle, the direction of gene flow between Basques and Catalans could be addressed by examination of the population distribution of root haplotypes; however, this has not been done here, because a combination of small sample size and uncertainty about the identification of these roots is likely to make such an analysis inaccurate.

The SRY-2627 polymorphism represents another example of the geographic specificity of Y-chromosome lineages, a phenomenon resulting from patrilocality (Seielstad et al. 1998) and cultural influences on mating practices, as well as from the small effective population size of the chromosome, which make it particularly susceptible to drift. When we find non-Iberian examples of haplogroup-22 chromosomes, they are likely to represent emigrants from Iberia. The finding of "Iberian" lineages in South America is not unexpected; their dates and places of origin are amenable to historical analysis, and they may provide a useful way to estimate the extent of admixture between indigenous people and Iberian colonists. Their occurrence in France, Germany, and England is more difficult to interpret, however. The young age of haplogroup 22 means that they cannot be adduced as support for the hypothesized "out of Iberia" migration 10,000–15,000 years ago, proposed on the basis of the distribution of mtDNA haplogroup V (Torroni et al. 1998). Population pairwise R_{ST} (Slatkin 1995) analysis of microsatellite data for the Iberian versus non-Iberian samples shows a significant difference ($P < .05$) between the two, which may be a sample-size effect but may also tell us that the emigrants are not very recent. In support of this, when information is available on the surnames of these individuals, these are typical of the populations in which they were found (data not shown). Surnames in most European populations came into existence after the 13th century (Hassall 1967), and therefore (if we set aside the complicating factors of non-paternity and local-surname adoption) this suggests that these emigrants may predate this period. This is no proposal for an early origin for tourism: there are many possible causes of such long-distance gene flow—for instance, it is known that the Roman army recruited cohorts of Basque soldiers, who served as far afield as Hungary, the lower Rhine, and northern England (Collins 1986; Perex Agorreta 1986).

References

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Barbujani G (1991) What do languages tell us about human microevolution? Trends Ecol Evol 6:151–156

Barbujani G, Pilastro A, de Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. paleolithic colonisation. Am J Phys Anthropol 95:137–154

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87:1816–1819

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds) Variation in the human genome. John Wiley & Sons, New York, pp 97–118

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

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59:63–81

Bianchi NO, Bailliet G, Bravi CM, Carnese RF, Rothhammer F, Martínez-Marignac VL, Pena SDJ (1997) Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am J Phys Anthropol 102:79–89

Calafell F, Bertranpetit J (1994a) Mountains and genes: population history of the Pyrenees. Hum Biol 66:823–842

——— (1994b) Principal component analysis of gene frequencies and the origin of Basques. Am J Phys Anthropol 93:201–215

Collins R (1986) The Basques. Blackwell, Oxford

Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Ber-

tranpetit J (1997) Mitochondrial DNA and the origin of the Europeans. Hum Genet 99:443–449

Comas D, Mateu E, Calafell F, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Bertranpetit J (1998) HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51:30–40

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

Côrte-Real HBSM, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, et al (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331–350

de Hoz J (1995) El poblamiento antiguo de los Pirineos desde el punto de vista lingüístico. In: Bertranpetit J, Vives E (eds) Munanyes i població—el passat dels Pirineus des d'una perspectiva multidisciplinària. Centre de Trobada de les Cultures Pirinenques, Andorra, pp 271–299

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13:1213–1218

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hassall WO (1967) History through surnames. Pergamon Press, Oxford

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, et al (1998) European Y-chromosomal lineages in Polynesia: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63:1793–1806

Jobling MA, Bouzekri N, Taylor PG (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). Hum Mol Genet 7:643–653

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11:449–456

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, et al (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am J Hum Genet 64:817–831

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133

Lafon R (1972) Basque: pour la comparison du basque et des langues caucasiques. Bedi Kartlisa 27:7–23

Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L, Savontaus M-L (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am J Hum Genet 58:1309–1322

Lucotte G, Hazout S (1996) Y chromosome DNA haplotypes in Basques. J Mol Evol 42:472–475

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786–792

Minch E (1997) Microsat 1.5d. Department of Genetics, University of Stanford, Stanford

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. Curr Opin Genet Dev 6:737–742

Mourant AE (1947) The blood groups of the Basques. Nature 160:505

———— (1983) Blood relations. Oxford University Press, Oxford

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J (1986) A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet 38:407–418

Perex Agorreta MJ (1986) Los vascones: el poblamiento en época romana. Gobierno de Navarra Departamento de Educación y Cultura, Institución Príncipe de Viana, Navarra, pp 63–69

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

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035

Renfrew C (1989) The origins of Indo-European languages. Sci Am 261:106–114

Ruhlen M (1991) A guide to the world's languages. Edward Arnold, London

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19:665–670

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–140

Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, et al (1998) mtDNA analysis reveals a major late Paleolithic population expansion from Southwestern to Northeastern Europe. Am J Hum Genet 62:1137–1152

Urtasun M, Sáenz A, Roudaut C, Poza JJ, Urtizberea JA, Cobo AM, Richard I, et al (1998) Limb-girdle muscular dystrophy in Guipúzcoa (Basque Country, Spain). Brain 121:1735–1747

Veitia R, Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, et al (1997) Mutations and sequence variants in the testis-determining region of the Y chromosome in

individuals with a 46,XY female phenotype. Hum Genet 99: 648–652

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183