The Basques in the Genetic Landscape of Europe

by

Kristin Leigh Young
M.A., University of Kansas, 2001

Submitted to the Department of Anthropology and the Faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Michael H. Crawford
Chairperson

James Mielke

Deborah Smith

Ivana Radovanovich

William Woods

Date Defended: June 16, 2009

The Dissertation Committee for Kristin L. Young certifies
that this is the approved version of the following dissertation:

The Basques in the Genetic Landscape of Europe

by

M.A., University of Kansas, 2001

Michael H. Crawford
Chairperson

James Mielke

Deborah Smith

Ivana Radovanovich

William Woods

Date Approved: June 16, 2009

Abstract

This study examines the position of the Basques in the genetic landscape of Europe using molecular genetic systems. Biparental markers (autosomal STRs and classical genetic markers) and uniparental markers (mtDNA haplogroups and HVS-I sequences, as well as Y-chromosome STR haplotypes) are used to address the origin and structure of the Basque population of Spain, as well as their role in the peopling of Europe. Three hypotheses of Basque origins are tested: 1) The Basques share a recent common ancestor with populations of the Caucasus; 2) The Basques are descendants of ancient Iberian populations who migrated from North Africa during the Neolithic; and 3) The Basques are a remnant population, the descendants of Paleolithic Europeans, who evolved *in situ*, with little gene flow from Neolithic farmers. The question of heterogeneity within the Basque population is addressed, and uniparental markers are examined for evidence of Neolithic ancestry in the Basque Provinces. Analysis of the molecular systems does not support a recent common ancestor between the Basques and populations either from the Caucasus or North Africa. While analysis of classical markers reveals the effects of genetic drift on the Basque population as a whole, AMOVA analysis of molecular markers demonstrates genetic homogeneity and little genetic structure between provinces (autosomal STRs: $V_a = -0.095$, $F_{CT} = -0.0036$, $p = 0.878$; Y-STRs: 1.71%, $p = 0.0369$; HVS-I sequences: 1.03%, $\phi_{ST} = 0.0103$, $p = 0.0308$). In addition, heterozygosity versus r_{ii} of autosomal STRs reveals that the impact of genetic drift is mediated by the influence of gene flow in three of the Basque Provinces. Gene and haplotype diversity levels in the Basque population are on the low end of the European distribution (autosomal STRs: 0.0805, Y-STRs: 0.9421, mtDNA: 0.0114), but other European populations have equivalent or lower levels. Distribution of uniparental haplogroups demonstrates varying levels of Neolithic admixture in the Basque population, with both Neolithic maternal lineages (J) and paternal lineages (E1b1b, G2, J2a) present. While these results do not suggest that the ancient Basque population had direct contact with Neolithic farmers, the presence of these markers cautions against using the Basques as a proxy Paleolithic population in genetic studies. However, the Basques do have high frequencies of other uniparental haplogroups considered to be of Paleolithic origin (Y-chromosome: R1b, mtDNA: H, U5), and analysis of demographic processes using HVS-I sequences places the date of population expansion among the Basques squarely in the Paleolithic, arguing against the complete replacement demic diffusion model of the Neolithic transition.

This dissertation is dedicated to my father, for saying I could be anything I wanted to be, and actually believing it. To my mother, for her unconditional love and support. To my children: Kaity, Brae, Hannah, and the little one on the way, for their inspiration and encouragement of my life outside of graduate school. To my husband, Brandon, my cheerleader, therapist, and best friend. To John, who was always proud of me. And to my grandmothers, Dennia Faye and Mary Ann, whose courage and independence were my guiding light.

Acknowledgements

I would especially like to thank the Basque participants whose participation made this study possible. I would also like to thank Dr. Arantza Apraiz-Gonzalez, a Basque post-doc in our laboratory who spent three field seasons in her home country collecting samples.

In addition, I would like to thank my committee members: Dr. Michael Crawford, Dr. James Mielke, Dr. Deborah Smith, Dr. Ivana Radovanovic, and Dr. William Woods; my collaborators: Dr. Ranjan Deka, Dr. Guangyn Sun, Dr. Eric Devor, and Dr. Mike Grose; my fellow graduate students in the Laboratory of Biological Anthropology: Geetha Chittoor, Kristie Beaty, Anne Justice, Chris Krawczak, Norberto Baldi-Salas, Jennifer Rack, Dr. Rohina Rubicz, Dr. Phillip Melton, Dr. Mark Zlojutro, and Orion Graf, for their continued friendship and encouragement; the departmental office staff, Judy Ross, Carol Archinal, and Kathleen Womack, who do their best to make this process as smooth as possible; and my family, for their unwavering support.

This research was funded by a National Geographic Society Grant to the University of Kansas Laboratory of Biological Anthropology (Project 6935-00) and a Carroll D. Clark award to the author.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE: INTRODUCTION

The debate concerning the Paleolithic versus Neolithic contributions to the modern European gene pool is currently a matter of degree. Did the Paleolithic inhabitants of Europe contribute to the modern gene pool? If so, how much? Or are modern Europeans the result of a replacement of the Paleolithic groups by technologically advanced Neolithic farmers? Given the linguistic and genetic distinctiveness of the Basques in Europe (Izagirre *et al.* 2001), and their relative isolation and distance from the epicenter of agricultural development during the Neolithic (beginning about 8,500 BC), they seem an ideal population in which to trace Paleolithic genomic signatures. The Basque language, *Euskara*, is most widely accepted as an isolate, unrelated to any other extant language in Europe. Genetically, the Basques are outliers in the European distribution for several classical markers, including blood groups ABO, Rhesus and MNS, erythrocytic enzyme adenylate kinase (AK), and immunoglobulin GM (Young 2007).

Hypotheses proposed to account for the uniqueness of the Basques in Europe include: a) The Basques share a recent common ancestor with populations of the Caucasus region (Bertorelle *et al.* 1995); b) The Basques are descendants of ancient Iberian populations who migrated to the peninsula from North Africa, possibly following drastic climate changes in the Sahara between 8000 and 4000 BC (Arnaiz-Villena *et al.* 2002; Arnaiz-Villena *et al.* 1997b; Sanchez-Velasco *et al.* 2003); and c) The Basques are a remnant population, the descendants of Paleolithic Europeans, who

evolved *in situ*, with little gene flow from various Neolithic populations to the present (Calafell and Bertranpetit 1994a; Calafell and Bertranpetit 1994c).

This last hypothesis also speaks to the peopling of Europe debate, which involves two competing models. The demic diffusion model (DDM) states that the majority of genetic variation present in modern Europeans is the result of the bands of Neolithic farmers spreading their technology (and genes) into Europe with the advent of agriculture (Ammerman and Cavalli-Sforza 1984). The cultural diffusion model (CDM) posits that while technology spread into Europe 10,000 years ago, people did not, leaving the Paleolithic gene pool largely intact (Novelletto 2007). Genetic evidence is used in support of both models. Synthetic gene map analyses of classical markers note a cline for several loci in the first principal component, accounting for approximately 27% of the total variation and spreading from southeast to northwest through Europe (Cavalli-Sforza *et al.* 1994). This cline is interpreted as a genetic signature of the DDM model, with a correlation of 0.89 between the first principal component of gene frequencies and temporal spread of agriculture into Europe (Ammerman and Cavalli-Sforza 1984). Similar clines are reported for autosomal markers HLA-DQA plus 6 short tandem repeat loci (STRs), indicating "directional population expansion" (Chikhi *et al.* 1998: 9055). Analysis of Y-chromosome STRs produce comparable results, with clinal variation in the frequencies of 12 of 27 alleles revealing maximum effective divergence times (τ_{max}) ranging from 281 – 10296 years, well within the Neolithic (Casalotti *et al.* 1999). This suggests that most of the variation present in the European Y-chromosome dates to this period, but research in

Southeastern Europe suggests that this Neolithic expansion includes endogenous European haplogroups (Battaglia *et al.* 2009).

In support of the CDM hypothesis, evidence comes from paleoanthropology as well as genetic analyses. Advocates note that the Paleolithic expansion into Europe occurred from the same area as the Neolithic expansion, implying that the gradient seen in some classical markers might instead be a Paleolithic signal (Barbujani *et al.* 1998). In the Basque region, there is strong archaeological evidence of Paleolithic human occupation (Bertranpetit *et al.* 1995). Analysis of prehistoric skeletal remains in Iberia show little transition in cranial morphology or evidence of dental caries between the Mesolithic and Neolithic, a transition which would be expected if the Mesolithic hunter/gatherers had been replaced by farmers with different dietary patterns (Jackes *et al.* 1997). Y-chromosome analyses reveal one haplotype (R1*M173) which appears to show evidence of expansion after the Last Glacial Maximum (Wells *et al.* 2001), and a "high degree of non-Neolithic ancestry" in populations of Iberia (Flores *et al.* 2004). Studies of Y-chromosome haplogroups among the Basques have demonstrated high frequencies of haplogroup R1b, the most common haplogroup in Western Europe (Alonso *et al.* 2005; Calderon *et al.* 2003; Flores *et al.* 2004; Lucotte and Hazout 1996; Lucotte and Loirat 1999; Quintana-Murci *et al.* 1999). Examination of Y-haplogroup diversity in Southeastern Europe demonstrated that the timing of the spread of agriculture in the region overlaps with the expansion of European Haplogroup I-M423, rather than Neolithic haplogroups G or J, suggesting a cultural diffusion of agricultural technologies by autochthonous

groups (Battaglia *et al.* 2009). Analysis of mitochondrial HVS-I sequences in Europe present little clinal variation, and divergence dates suggest that many of the mitochondrial haplotypes in Europe have a pre-Neolithic origin (Richards *et al.* 1996). Haplogroup V has been proposed as a signal of population expansion after the Last Glacial Maximum (LGM), but the absence of this haplogroup in an ancient sample called this hypothesis in to question (Izagirre and de la Rua 1999; Torroni *et al.* 1998). In addition, the presence of 15 individuals in the same sample belonging to haplogroup J, which has been linked to the expansion of Neolithic groups into Europe, suggests some level of admixture.

This investigation examines the origin and structure of the Basque population of Spain, and their role in the peopling of Europe. Previous molecular studies conducted among the Basque population used small sample sizes (N = 45-73), often collected in urban areas of a single province (when collection location was reported), to determine Basque origins (Aguirre *et al.* 1991a; Bertranpetit *et al.* 1995; Perez-Miranda *et al.* 2003). The data used in this dissertation represent one of the most comprehensive samples of Basques yet collected, with 652 individuals in 35 mountain villages from throughout the four Basque Provinces of Spain analyzed for mtDNA RFLP haplogroups, mtDNA control region sequences, Y-STR haplotypes, and autosomal STRs**.** These data allow the examination of maternal, paternal, and autosomal genetic variation, and analysis of concordance between these various types of data. Using both mitochondrial and Y-chromosome markers allows for the assessment of migration and gene flow by sex, as mtDNA is inherited maternally,

while Y-chromosome markers pass only from father to son. Haplotypes which have been identified as having either a Paleolithic or a Neolithic origin are also examined to calculate the relative contributions of each to the modern Basque gene pool. Heterogeneity within the Basques is explored, to determine if the provinces vary with respect to the presence or percentage of Paleolithic haplotypes or have undergone different rates of Neolithic admixture. The addition of autosomal STR analysis provides information on the overall genetic variation present in this population. Data from classical markers, such as blood groups, red cell enzymes, plasma proteins, and HLA haplotypes, are compiled from the literature for comparison. All of these systems are contrasted with data from other European and Mediterranean groups to determine which populations are most genetically similar to the Basques, and assist in realizing the place of the Basques in the history of the European continent.

This study contributes not only to the anthropological understanding of the Basque population, but speaks to phylogeographic knowledge of human migration, and peopling events in general. Collectively, the data present a singular opportunity in anthropological genetic studies. With data from mountain villages throughout the Spanish Basque country, this study performs a comprehensive analysis of the place of the Basques in Europe; homogeneity of the population (Aguirre *et al.* 1991a); possible origins and their influence on the peopling of Europe; and the effects of evolutionary forces recorded in three molecular systems, one maternal (mtDNA), one paternal (Y-chromosome), and one autosomal (STRs).

The following chapters present the population history of the Basques, including previous genetic studies and a discussion of hypotheses of population origins. Sample collection procedures, as well as laboratory and analytical methods used in the present study are described. Intra- and interpopulation results for each of the molecular systems examined (autosomal STRs, Y-chromosome STRs, and mtDNA haplogroups and HIVS-I sequences) are presented. Finally, implications of these results on the hypotheses of the origins of the Basques and the peopling of Europe are discussed.

CHAPTER TWO: LITERATURE REVIEW

Population History

The question of Basque origins is one that has interested scholars since at least 1889, when Aranzadi suggested, based on cranial morphology, that the Basques were an ancient relict population (Calafell *et al*. 1994a). In 1946, Boule went so far as to state that the Basques might be descendants of Cro-Magnon (Mourant 1974). It was also known that the Basque language, *Euskara*, was one of the few non-Indo-European languages on the European continent. It is widely accepted as a linguistic isolate, not related to any other extant languages (Trask 1996). The discovery of human blood types bore out the distinctiveness of the Basques. This early work distinguished the Basques from other European populations by a low frequency of the ABO histo-blood group B allele (1.1%) (Boyd and Boyd 1937). Subsequent studies also revealed a high level of the Rh negative (RH*cde) allele (between 30.5-35.6%) (Chalmers 1948; Chalmers *et al.* 1949; Etcheverry 1945), leading Mourant to state that "the Basques are a relict population which at least in Spain has suffered no significant admixture of elements akin to the general western European population" (1947: 505).

Figure 1. Map of the Basque Country, with Spanish provinces Alava, Guipuzkoa, Navarre and Vizcaya, and French provinces Labourd, Basse-Navarre, and Soule. Capitals of each province are also indicated. Adapated from Perez-Miranda et al. (2003).

The Basque country historically comprised three provinces in southern France (Labourd, Basse-Navarre, and Soule) and four provinces in northern Spain (Alava, Guipuzkoa, Vizcaya and Navarre) in the western Pyrenees along the Bay of Biscay (Figure 1). Today, only the Spanish provinces remain politically autonomous. Alava, Guipuzkoa and Vizcaya form the Basque Country Autonomous Community, while Navarre is its own autonomous unit – the Navarre Chartered Community (Calderon 2002). The three French provinces were incorporated into the department of Pyrénées-Atlantiques. Toponymy suggests that the Basque region was once much

larger than it is today, extending perhaps from as far north as the Garonne river in France (See Figure 2), to as far south as the Ebro River Valley in Spain, stretching from the Bay of Biscay to the Central Pyrenees. (Collins 1990; Levine 1967). Analysis of classical genetic markers supports this view, with French occidental populations clustering with Basques (Calafell and Bertranpetit 1993b).

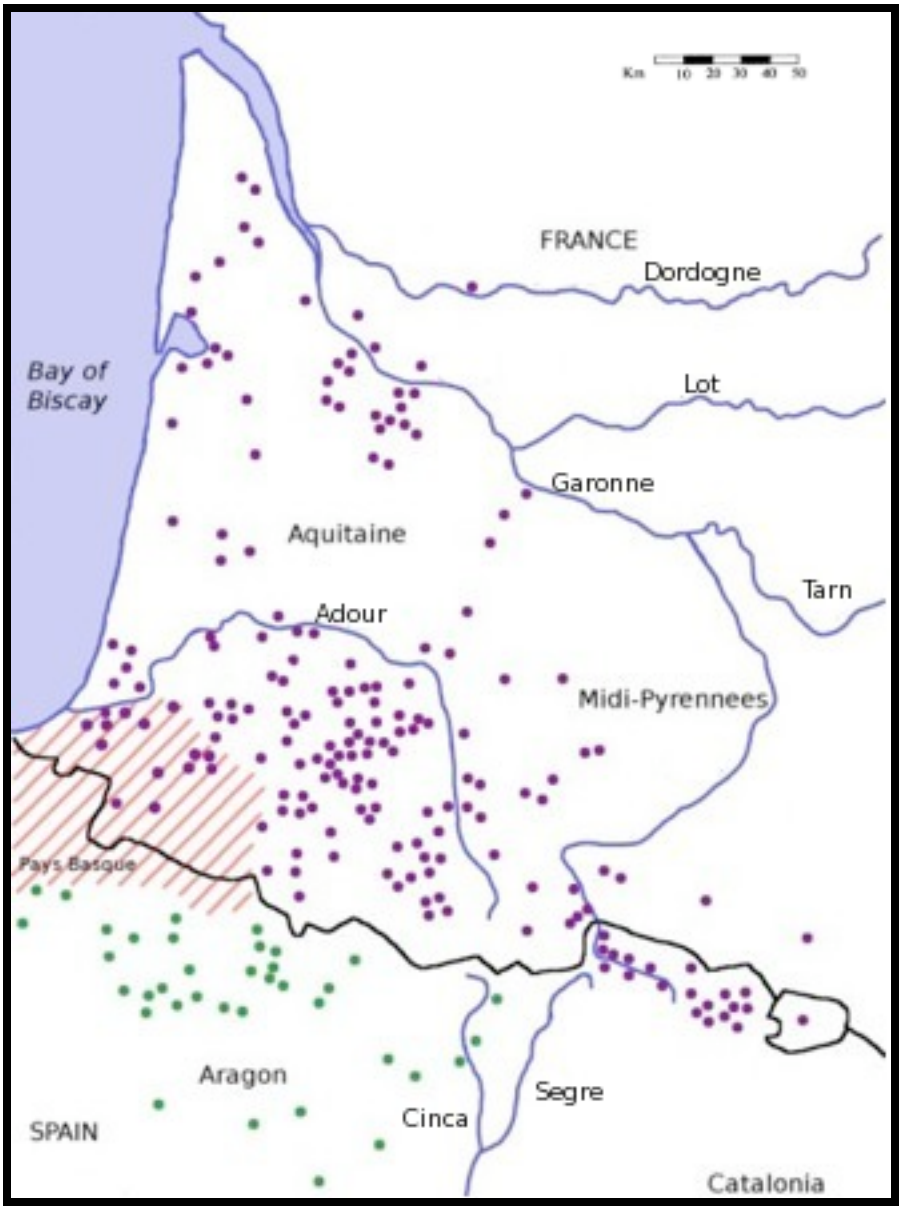

Figure 2. Map of place names with Basque endings. Purple dots: Names ending in –os, -ous, -osse, -ost, -oz. Green dots: Names ending in –ues, -ueste. After Bernard and Ruffie (1976). Pink lines indicate current extension of Basque territory.

Many hypotheses of a relationship between Basques and other populations have been proposed based on linguistic analyses. As R.L. Trask states:

> "As the only non-Indo-European language in western Europe, Basque has fascinated generations of serious linguists, dilettantes, and out-and-out cranks into devoting large chunks of their careers to dogged comparisons with whatever other languages have taken

> their fancy, comparisons which have proved to be, at best, unrewarding, and at their all-too-frequent worst, fanciful, incompetent, and downright crazy" (Trask 1996: 65-66).

These proposed linguistic connections include the ancient languages of Iberian, Minoan, Etruscan, Pictish, Sumerian, and Aquitanian, as well as extant Uralic (such as Finnish), Caucasian (such as Georgian), African (especially Berber), and sundry Native American languages, and even Japanese (Blaud 1974; Goedde *et al.* 1972a; Piazza *et al.* 1988b). Only the relationship of *Euskara* to Aquitanian, an extinct language from present-day France, is well-supported, and thus it is regarded by experts in the field as a "direct ancestor of [the] Basque [language]" (Trask 1996). Basque place names are found throughout the Aquitaine region of France (Figure 2), which could reflect this ancestral relationship between Euskara and Aquitanian, or alternatively, the previous extent of the Basque language. This has not prevented comparisons between Basque populations and other purported relatives, including common ancestry with populations from the Caucasus (Cavalli-Sforza 1988), the Berbers of Africa (Arnaiz-Villena *et al.* 1997a), or being a relict European population predating the Indo-European invasion (Mourant 1947).

A relationship between Basque and some, but not necessarily all, Caucasian languages was proposed as early as the 1920s (Bertorelle *et al.* 1995). Citing a French source dated 1925, Cavalli-Sforza and Piazza (1993) claimed a link between Basque and northern Caucasian languages, such as Abkhaz. A link to the southern Caucasian language, Georgian, has also been proposed (Calderon *et al.* 1993). However, Caucasian languages themselves are not a cohesive group, and while some

linguists see similarities between Basque and some aspects of the northern or southern Caucasian languages, they have been attributed either to sloppy technique, a shared Euro-African substratum, or similarities in the evolution of language itself (Jiménez 2001). As an example of the first case, Jiménez (2001) cites a comparative study of Basque and Caucasian plant names (Bouda *et al.* 1955) in which the words in the various languages have been parsed differently, and often do not have the same meaning in Basque as they do in the Caucasian languages. Linguists have not yet reached a consensus on the relationship, if any, between Caucasian languages and Basque. It has been suggested that both populations, which speak non-Indo-European agglutinative languages, could be remnants of the Mesolithic European population and have been less affected than the rest of the continent by the Neolithic Revolution for the same reason – both inhabit mountainous regions that were less hospitable to agricultural pursuits (Bertorelle *et al.* 1995).

The Vasco-Iberian hypothesis holds that languages related to Basque were spoken throughout the Iberian Peninsula prior to Roman conquest. A genetic relationship between *Euskara* and Iberian was favored in the late 1700s, with Basque considered the last remnant of this larger language family, preserved due to the remoteness of their territory in the foothills of the Pyrenees (Collins 1990: 10). Discoveries of Iberian inscriptions which were not translatable using *Euskara* weakened this hypothesis on linguistic grounds, but because Iberians were believed to have migrated from North Africa, and a connection between Iberian and Basque had been proposed,

genetic similarities between Basques and North Africans have also been sought (Jiménez 2001: 116).

Other research suggests that the Basques are remnants of an ancient European population (de Mouzon *et al.* 1979; Goedde *et al.* 1972a; Ohayon *et al.* 1980). While Calderon *et al.* (1998) assert that this concept is "a version of the multiregional hypothesis of the rise of modern humans from *Homo erectus*" (687), most studies suggest that the Basques are pre-Indo-European, not pre-anatomically modern *Homo sapiens.* Cavalli-Sforza and Piazza state that the Basques are "probably the most direct descendants of the earliest *post-Neanderthal* settlers of Europe (emphasis added)," whose distinct genetic makeup might result from the "remarkable isolation of western from eastern Europe at the time of the last glaciation" (1993: 11; Piazza 1993). Recent work on the genetics of cystic fibrosis (CF), a genetic disorder most often afflicting individuals of European descent, supports this idea. Analysis of the ΔF508 mutation (a 3bp deletion at codon 508 of the cystic fibrosis transmembrane conductance regulator protein) in Iberia determined that it was present at a frequency of 87% in Basque families, compared to a low of 33.3% in neighboring Navarre and a high of 66.7% in Madrid. (Casals *et al.* 1993; Casals *et al.* 1992). Combined with the greater haplotype diversity in Basque CF chromosomes compared to those found in other European populations, these data suggest that the ΔF508 mutation is an "old molecular defect" that originated in European populations during the Paleolithic, and consequently, that Basques are a relict population (Casals *et al.* 1992).

Some linguists conclude that Basque is an autochthonous language, which developed *in situ* in the Iberian Peninsula and once had a wider range, but has also had contact with other languages in historical times (Collins 1990; Jiménez 2001). Similarities between *Euskara* and Iberian are attributed to a long period of cultural contact; place names in the heart of the Basque country, such as Gernika, reflect a Celtic influence (the –ika suffix is Celtic, not Euskara), picked up during the Celtic invasion of the Iberian peninsula; and there are many Latin borrowings found in Basque, as well. Unlike the other non-Indo-European languages present at the time of Roman conquest, however, Basque survived, at least in the more inhospitable reaches of their domain. The lowland areas of Navarre and Alava were acculturated and Latinized, and the Basque language was lost, as it was north of Pyrenees in Aquitaine. However, the mountains were a safe haven for the Basque language, as they provided few resources of interest for conquerors and lacked large cultural centers where foreign influence could gain a foothold.

At the time of Roman contact, there was little evidence of tribal structure and no higher order political organization. Instead, Basque society appears to have been built on extended family structure, with the largest social unit being valley communities.

> The Basques thus present an extraordinary paradox with their remarkable capacity for survival as a racial and linguistic group, matched by an extreme cultural fragility outside a small nuclear homeland consisting of the western Pyrenees and a coastal region between the mountains and the Bay of Biscay. This raises fundamental questions about the nature of Basque identity (Collins 1992: 41).

Their identity as Basques seems to be based, even now, primarily on language (Urla 1993). As the Basques had no written records, the few early references come from

travelers in historical times. These descriptions must be considered critically, as they are written by outsiders and tend to be less than objective. *The Pilgrim's Guide to Santiago de Compostela* (ca. 1140) provides a typical, although heavily edited, example:

> …The Navarrese and Basques are held to be exactly alike in their food, their clothing and their language, but the Basques are held to be of whiter complexion than the Navarrese… If you heard them speak, you would be reminded of the barking of dogs. For their speech is utterly barbarous….
>
> This is a barbarous race unlike all other races in customs and in character, full of malice, swarthy in color, evil of face, depraved, perverse, perfidious, empty of faith and corrupt, libidinous, drunken, experienced in all violence…in everything inimical to our French people (Constable 1997:139-141).

Stripping this quote of its xenophobic overtones, one can conclude that the Basques and the Navarrese were essentially the same people, and they spoke the same language, which was unintelligible to this French traveler.

Evidence from the archaeological record is more concrete, though little exists in the way of material culture to identify specific sites as unequivocally Basque (Collins 1992). There is good evidence for Paleolithic occupation of the Basque country by modern humans (See Table 1), dating from the Aurignacian (34,000 BP) through the Magdalenian (10,000 BP) (Bertranpetit *et al.* 1995; Haarmann 1998). These are mostly cave sites and rock shelters, some with cave art and mobilary art pieces (Enamorado 1997). Some regard this long period of occupation as support for the descent of proto-Basque peoples from the "indigenous Neolithic/Bronze Age inhabitants of the mountainous zones" (Collins 1990:30), but other research, particularly in craniometrics, suggests a longer period of proto-Basque occupation.

Table 1. Sites of hominin occupation in the Basque Country. Adapted from (Enamorado 1997).

Site	*Province*	*Timeframe*	*Tool Tradition*	*Species*
Axlor	Vizcaya	42,000 BP	Mousterian	Neandertal
Amalda	Guipuzkoa	28,000-19,000 BP	Gravettian & Solutrean	*H. sapiens*
Lezetxiki	Guipuzkoa	150,000-80,000 BP		*H. hiedelbergensis*
			Mousterian	Neandertals
Arrillor	Alava	19,000-10,000 BC	Magdalenian	*H. sapiens*
Santimamine	Vizcaya	80,000-10,000BP	Mousterian-Magdalenian	Neandertal
				H. sapiens
Kurtzia	Vizcaya	34,000-23,000 PB	Aurignacian	*H. sapiens*
Lumentxa	Guipuzkoa	34,000-23,000 PB	Aurignacian	*H. sapiens*
Bolinkoba	Vizcaya	28,000-19,000 BP	Gravettian & Solutrean	*H. sapiens*
Labeko Koba	Guipuzkoa	34,000-30,000BP	Aurignacian	*H. sapiens*
Usategi	Guipuzkoa	22,000 – 19,000 BP	Solutrean	*H. sapiens*
Ekain	Guipuzkoa	16,500 – 12,050	Magdalenian	*H. sapiens*
Aitzbitarte	Guipuzkoa	34,000-10,000 BP	Aurignacian - Magdalenian	
Abauntz	Navarre	22,000 – 19,000 BP	Solutrean	*H. sapiens*
Ermittia	Guipuzkoa	22,000 – 19,000 BP	Solutrean	*H. sapiens*
Atxeta	Vizcaya	18,000 BP to 12,150 BP	Magdalenian	*H. sapiens*
Atxurra	Vizcaya	28,000 BP- 23,000 BP	Gravettian	*H. sapiens*

Craniometrics, the anthropometry of head shape, has a somewhat shady history in physical anthropology, as these measurements were first used to define various races (Closson 1897). Of particular interest was the cephalic index, a ratio of the width of the skull relative to its length (measured front to back). The two extremes of the head shape continuum were brachycephalic (broad, rather short skulls, with an index of 85 or greater) and dolichocephalic (skulls relatively longer than they were wide, with an index of 71-75). Measurements taken on fossil skulls led to the assertion that the Upper Paleolithic inhabitants of Europe, including Cro-Magnon, were of the dolichocephalic type (Broca 1878).

In 1864, 96% of a sample of 60 contemporary Basque skulls from Guipuzkoa 96% were reported as dolichocephalic, although the mean cephalic index for this sample was 77.7, outside Broca's defined dolichocephalic range. In the same letter to

the Anthropological Society of London, only 12% of 26 fossil skulls from a Bronze Age cave deposit in the Basque country were described as dolichocephalic (Broca 1864). These facts did not prevent the association of Basque with Cro-Magnon based on cephalic index, however, or the assertion that this "race" may have come from North Africa where similar cephalic indices had been reported. However, a reanalysis of Broca's sample supplemented by the addition of 19 skulls, noted that the "Basque skull is typically European in all respects: there is nothing to suggest that it is more closely related to non-European types than any other Western European forms are" (Morant 1929:71). A more recent multivariate analysis of 20 craniometric variables in 13 Iberian populations demonstrates the unique position of Basques in the Iberian Peninsula (Fox *et al.* 1996). Regardless of sex, the Basques were distinct in every analysis performed. The differences between the Basques and other Iberian populations could not be accounted for solely by geographic distance (Mantel T=0.1942, p= 0.73) and was instead attributed to greater age of the Basque population relative to the others.

Basque Genetic Studies: Classical Markers

Blood Groups

Early research into genetic differences between populations was not performed using genes, but rather their products, including blood groups, red-cell enzymes, and plasma proteins (Table 2). First discovered by Landsteiner and Wiener in 1900, the utility of blood groups in anthropological research was revealed when they were found to be present at varying frequencies in different populations. The 26

blood groups in humans are antigens, molecules which sit on the surfaces of red blood cells and other tissues, identifying them as self. The genetic distinctiveness of the Basques was first reported in blood groups such as ABO, Rhesus (Rh), Duffy (FY), and Kell (K) (Boyd and Boyd 1937; Chalmers 1948; Chalmers *et al.* 1949; Dausset 1972; Etcheverry 1945; Mourant 1947). In one French province, the frequency of ABO*B was found to be less than 1%, and it was due entirely to mixed Basques in the sample (Nijenhuis 1956). The frequency of Rh negative individuals among Basques has been reported at over 50% in Guipuzkoa, and over 42% in the French provinces, some of the highest frequencies of Rh*cde in the world (Levine *et al.* 1977; MacClancy 1993). The frequency of FY*A in Labourd was reported as lower than other European populations at 34.5% (Levine *et al.* 1974). Labourd was also noted for a low frequency of K*K (2.9%) (Dausset 1972). However, some researchers would argue that the Basques are not exceptional with respect to blood groups, as other populations in the region have similar frequencies for these loci (Arnaiz-Villena *et al.* 1999). Examination of the distribution of allele frequencies collected from the literature for various classical markers bears out this point (populations for which data were collected are given in Appendix 1), demonstrating that Basque populations are outliers at only three blood group loci (MNS, RH, and ABO). Basques in Labourd are outside the European range (0.21-0.28) for MNS*MS (0.3329) (Figure 3).

Table 2. Classical polymorphisms studied among the Basques.

Locus	***N***	***Locus***	***N***
Blood Groups		*Red Cell Enzymes*	
ABO (A subtypes)	23	Acid Phosphotase (ACP)	33
*ABO**	40	Adenosine Deaminase (ADA)	30
*Rhesus (RH)**	36	*Adenylate Kinase (AK)**	32
MN	31	Esterase D (ESD)	25
*MNS**	29	Glyoxylase I (GLO1)	32
Duffy (FY)	28	Phosphoglucomutase (PGM1)	31
P	27	6-Phosphogluconate Dehydrogenase (6-PGD)	24
Lutheran (Lu)	17	*Plasma Proteins*	
Kell (K)	25	Ceruloplasmin (CP)	31
ABH Secretion (Se)	19	Group Specific Component (GC)	25
Lewis (Le)	10	Haptoglobin (HP)	25
Kidd (Jk)	18	Transferrin (TF)	32
		α-1-antitrypsin (P)	9
		*Immunoglobulin (GM)**	28
		Immunoglobulins (KM)	25

**Indicates loci those for which some Basque populations are significantly different from other European populations. N indicates the number of populations for which data were gathered from the literature. Information on HLA loci is provided in Appendices 2 and 3.*

The Basques do have exceptionally high levels of RH*cde (Figure 4). Basques in Labourd (0.5021) and Guipuzkoa (0.5336) are outliers, but Basques in Alava (0.4572) are not. Basques are not the only population with elevated RH*cde levels, however, as the Scots also have a high frequency of this allele (0.4752) (Brown 1965).

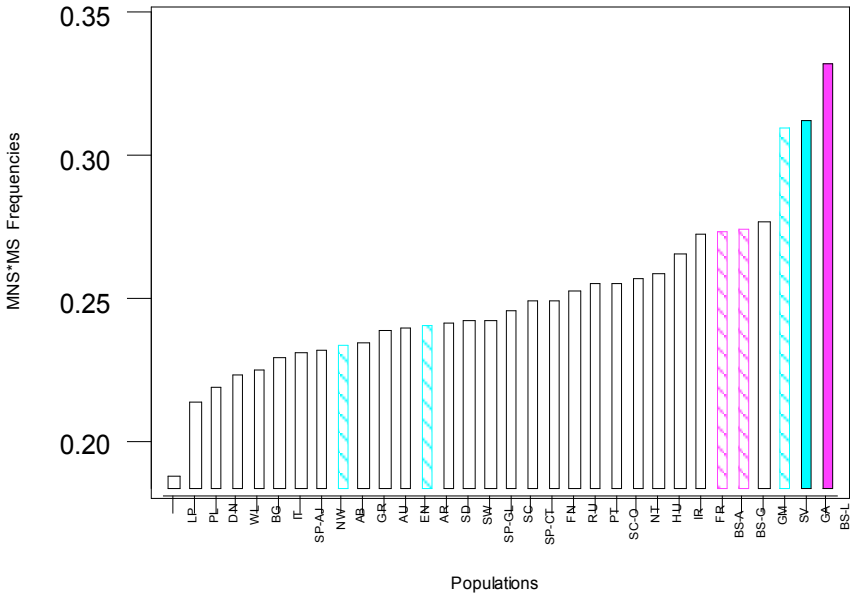

Figure 3. Frequency distribution of MNS*MS allele in European populations. Solid color bars indicate outliers: Basques from Labourd (pink) and Georgians (turquoise). Other Basque populations in Alava (BS-A) and Guipuzkoa (BS-G) are indicated as pink-striped bars. Other Caucasian populations, including the Svani (SV), Armenians (AR) and Abkhazians (AB) are shown as turquoise-striped bars. See Appendix 1 for other population abbreviations.

As for ABO*B, the Basque populations were found to be on the low end of the European frequency distribution (Figure 5), with the community of Basses-Pyrenees lacking this allele. Other populations on the Iberian Peninsula, including those in Alpujarra, Valencia, and Murcia, had frequencies of ABO*B below 5%, as did the other Basque populations. Frequencies of FY*A are low among the Basques (0.3107-0.3576), but other populations in the Iberian Peninsula, including the Aranese of the Aran Valley (0.3463) and the Spanish of Cataluna (0.3654), also have low FY*A levels (Figure 6). The only outlying population for the FY*A allele is the Lapps, which have a high frequency (0.5528). For all other blood groups, frequencies

among Basques were similar to those elsewhere in Europe (Manzano *et al.* 1996a; Nijenhuis 1956).

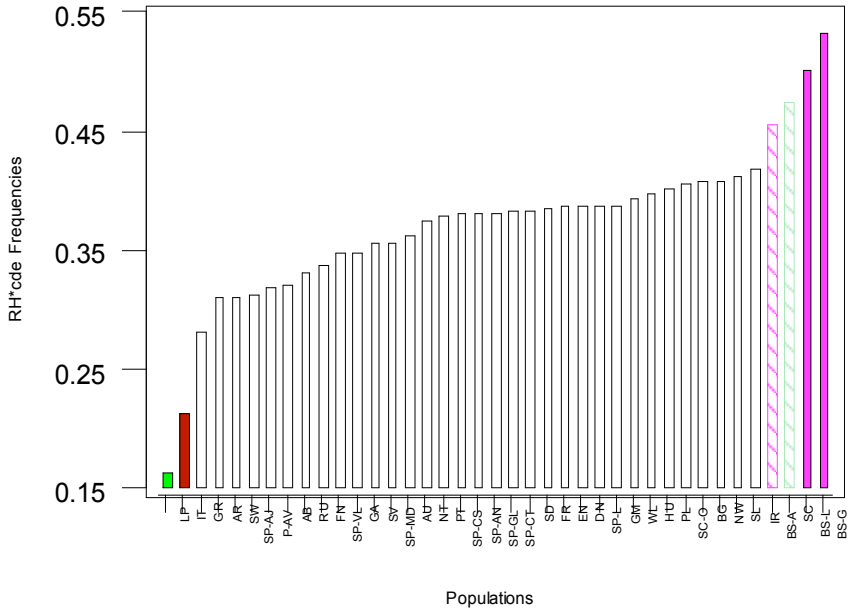

Figure 4. Frequency distribution of RH*cde allele. Outliers are shown as solid bars: Basques in Guipuzkoa (BS-G) and Labourd (BS-L) (pink), Lapps (lime), Italy (brown). The other Basque population from Alava (BS-A) is shown as a striped-pink bar. Scotland (SC), which has a frequency comparable to the Basques, is shown as a green-striped bar.

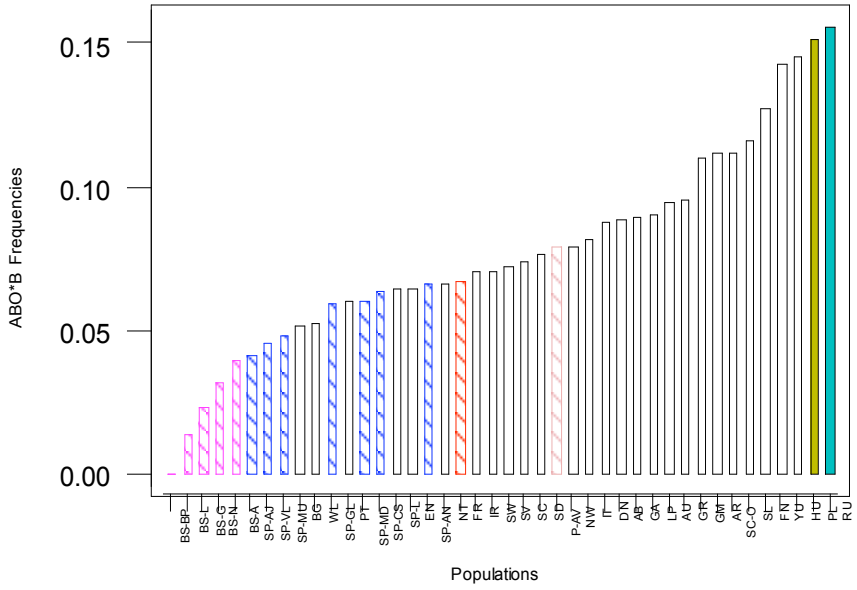

Figure 5. Frequency distribution of ABO*B among European populations. The Basque populations are shown as pink-striped bars. Spanish populations are in blue, France in red, and other Pyrenean populations in lilac. The three outlying populations are the Basques from Basses-Pyrenees (BS-BP), the Poles (olive) and the Russians (teal). The other Basque groups fall between the Basque outlier and other Iberian populations in Spain.

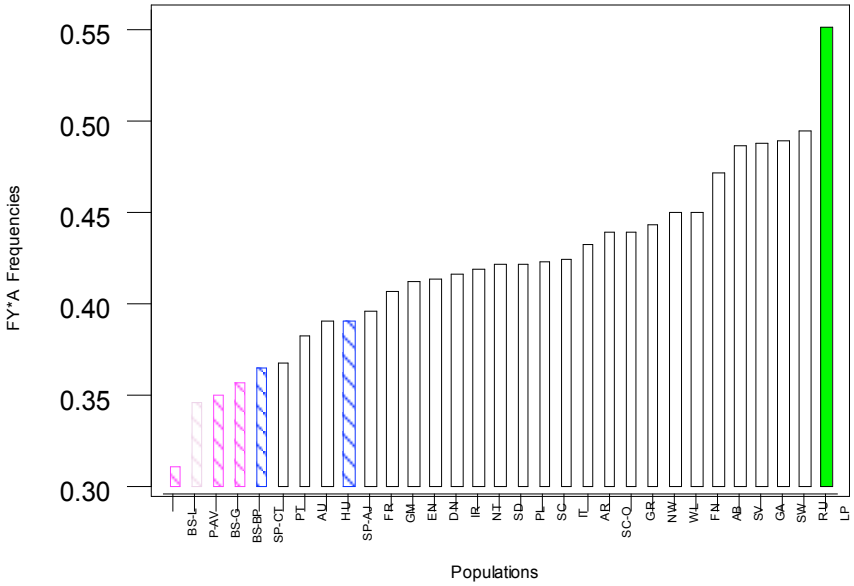

Figure 6. Frequency distribution of FY*A in European populations. The Lapps, the single outlying population, are shown in lime. The Basque populations (Labourd, Guipuzkoa, and Basses-Pyrenees) are pink-striped. Other Iberian populations are shown as blue-striped bars (Spanish in Cataluna and Alpujarra) and a lilac-striped bar (Aran Valley).

Erythrocyte Enzymes

Blood groups are not the only mucopolysaccharides and mucoproteins useful in population genetic studies. Red-cell enzymes, which catalyze various cellular reactions, have been shown to have electrophoretic variants. These variants differ according to size and electrical charge as a result of mutations in the genes coding for specific enzymes (Table 2), and are present at varying frequencies in human populations. Research on red-cell enzymes among the Basques has shown that certain loci appear to distinguish some Basque groups from other European populations, while others do not. In an analysis of subpopulations in Vizcaya province, a low

frequency of the acid phosphatase 1 A allele (ACP1*A - 0.2750), one of the highest frequencies in Europe of ACP1*B (0.7180), and a low frequency of the adenosine deaminase 2 allele (ADA*2 - 0.0210) were reported (Aguirre *et al.* 1991b). No significant difference between Basque samples in Gernika and Arratia (both in the province of Vizcaya) and other Spanish samples in Castile and Leon for frequencies at the ACP1 locus has been found (de Pancorbo *et al.* 1986). A comparison of four Spanish populations demonstrated that the Basques had a relatively low frequency of the ADA*2 allele (0.0284), and a high frequency of ACP*B (0.7170) (Goedde *et al.* 1972a). High levels of AK*2 have been reported in Arratia (0.0500) and Guipuzkoa (0.0650), which have some of the highest frequencies of AK*2 in Europe (Aguirre *et al.* 1989a; Manzano *et al.* 1996a).

Although the Basque populations studied have low levels of ACP*A, they are not significantly different from other European populations. Similar frequencies of ACP*A are also found among populations in Georgia (0.2820), Italy (0.2781), Spain (Galicia – 0.2984, Castile – 0.252, and Leon – 0.2330), and Greece (0.2200). And while Basques are at the low end of the European range for ADA*2 frequencies, they are joined by Scots in the Orkneys and Spanish groups in Castile, the Central Meseta, and Madrid. Examination of red-cell enzyme frequency distributions revealed only one locus for which Basques fall outside the European range, adenylate kinase (AK) (Figure 7). Basques in Guipuzkoa have a comparably low frequency of AK*1 (0.935), though the Georgians are lower (0.915). Two other populations, the Lapps of Norway (0.9872) and Spaniards from Galicia (0.9779) -- a province just north of

Portugal, were outliers at this locus with high frequencies of AK*1. Other Basque populations in Vizcaya, Navarre, Basses-Pyrenees and Labourd are not outliers for this allele. In fact, the latter three populations have frequencies in the high end of the range. Overall, frequencies for all other red-cell enzymes were similar to those reported elsewhere in Europe (de Pancorbo *et al.* 1989; Garcia-Orad *et al.* 1987).

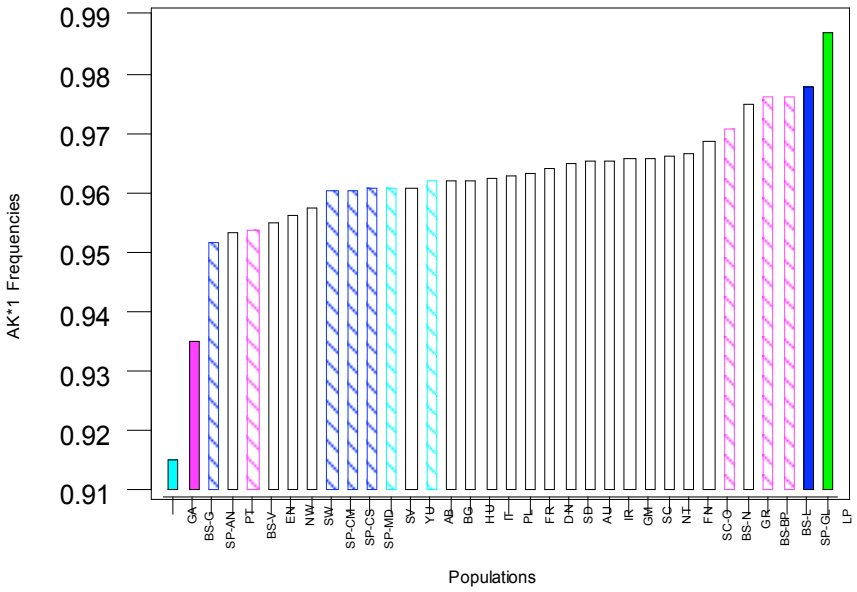

Figure 7. Distribution of AK*1 in Europe. Outliers are Lapps (lime), Spanish in Galicia (blue), Basques in Guipuzkoa (pink), and Georgians (turquoise). Other Basque, Spanish, and Caucasian populations are shown as hatched bars in the corresponding color.

Plasma Proteins

Proteins found in blood plasma serve a variety of biological functions, from vitamin-binding and ion transport, to enzyme inhibitors and antibodies. Like erythrocytic enzymes, plasma proteins also have electrophoretic variants which differ in frequency in human populations. No significant difference between Basques and

British, Canadian, American, or Iberian samples was found for the iron-transporting protein haptoglobin (HP) (Allison *et al.* 1958; Constans and Viau 1975; Planas 1966). For the transporter proteins transferrin (TF) and group-specific component (GC), the Basques are within the European range. While the Basques in Alava have been reported to "generally show the highest values in the Iberian Peninsula for the GC*2 allele" (Manzano *et al.* 1993b), the Spanish from the southern province of Andalusia have higher frequencies of GC*2 (0.3796), and the Spanish in Galicia (0.3589), as well as Hungarians (0.3261) and Poles (0.3337), are within the range of Basque populations for this allele (Figure 8).

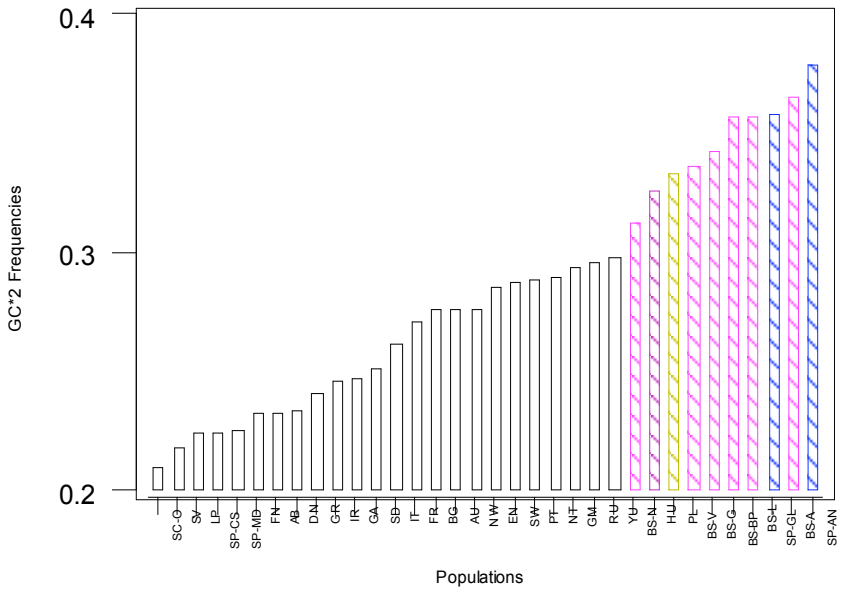

Figure 8. Frequency distribution of Group-Specific Compnent (GC*2) in European populations. Basque groups are in pink, Spanish populations in blue, Hungarians are purple, and Poles olive. Basque populations cluster near the high end of the range, as do Spanish populations and other from Eastern Europe.

Other plasma proteins that have been used in anthropological research are immunoglobulins. These are antibodies, or proteins which bind foreign cells, such as bacteria and viruses. There are five classes of immunoglobulins: IgG (γ), IgA (α), IgM (μ), IgD (δ), and IgE (ε) (Grubb 1970). In population studies, gammaglobulins (Gm) have been most informative. Three subclasses of gammaglobulins have been identified: (G1m, G2m, G3m), comprised of heavy and light polypeptide chains which form a flexible Y-shaped molecule (Figure 9). The light chains, also called kappa chains (KM), have only three alleles, whose frequencies vary little between human groups, so they are of limited utility in anthropological studies. The gamma immunoglobulins (GM), however, are one of the most polymorphic of the classical genetic systems used in human population analysis. Each subclass has multiple alleles (see Table 3), which, due to their presence on the same chromosome, tend to be inherited together as haplotypes. These haplotypes vary by population, so that some are found primarily in African groups, others in Asia, and others still in Europe, presenting the potential to address questions of admixture and population origins (Stevenson and Schanfield 1981).

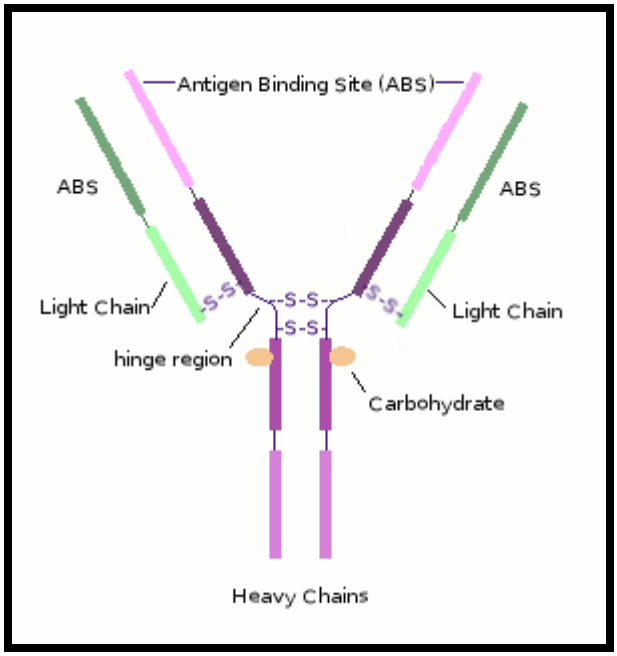

Figure 9. Schematic of an immunogobulin (IgG) molecule, showing the heavy (gamma) and light (kappa) chains, as well as the variable antigen binding site.

Several immunoglobulin allotype analyses have been performed on the Basques. Analysis of 11 allotypes in the villages of Macaye and Ahetze in the French Pyrenees presented data on three haplotypes, GM*Z,A;G, GM*Z,A,X;G and GM*F;B (Levine *et al.* 1974). These haplotypes are characteristic of European populations, demonstrating that these Basques were within the western European range. The issue of admixture or the possibility of non-European origin was not addressed. Analysis of 16 immunoglobulin allotypes in 11 communities in the French Pyrenees, including the Basque villages of Macaye, Mauleon, and St. Jean Pied de Port, revealed that the European haplotype GM*F;B was the most frequent (66-72%), while other haplotypes (including GM*Z,A,B;S,T; GM*Z,A;B; and

Table 3. Immunoglobulin allotypes used in population studies

Subclass	*Alphameric*	*Numeric*
Heavy Chain		
IgG1 (G1m)	(a)	(1)
	(x)	(2)
	(f)	(3)
IgG2 (G2m)	(z)	(17)
IgG3 (G3m)	(n)	(23)
	(b0)	(11)
	(b1)	(5)
	(b3)	(13)
	(b4)	(14)
	(b5)	(10)
	(c3)	(6)
	(c5)	(24)
	(g)	(21)
	(s)	(15)
	(t)	(16)
	(u)	(26)
	(v)	(27)
Light Chain		
Km		(1)
		(2)
		(3)

GM*Z,A;C,B) were found at frequencies of 1% or less (Dugoujon *et al.* 1989). A similarity between eastern and western Pyrenean populations for this system, due primarily to high frequencies of GM*F;B, suggests settlement of the region by a single population prior to Indo-European arrival, or similar patterns of migration on

both ends of the Pyrenees (Dugoujon *et al.* 1989: 48-49). Further analysis of these data using the 'Mobile Node Method,' which generates a distorted 'genetic similarity' map by minimizing differences between genetic and geographic distance matrices, illustrated general "agree[ment] with the hypothesis of an ancient population of the French Pyrenees issuing from the same origin and with little differentiation between later populations" (Hazout *et al.* 1991: 172-173).

GM analysis in Spain was conducted among both students at secondary and vocational schools in the three provinces of the Basque Autonomous Community (BAC) and in unrelated healthy rural adults in Alava and Guipuzkoa (Calderon *et al.* 1998; Esteban *et al.* 1998). Similar to French Basques, the most frequent haplotypes found among the Basques in Spain are GM*F;B, GM*Z,A;G, and GM*Z,A,X;G. Haplotype GM*Z,A;B was absent in the rural areas of Guipuzkoa, but present at polymorphic levels (>1%) in Alava, Vizcaya, and the education centers in Guipuzkoa. This haplotype is also found at appreciable frequency in populations scattered throughout the Pyrenees, as well as France, Spain, Yugoslavia, and Finland (Calderon *et al.* 1998) (Figure 10).

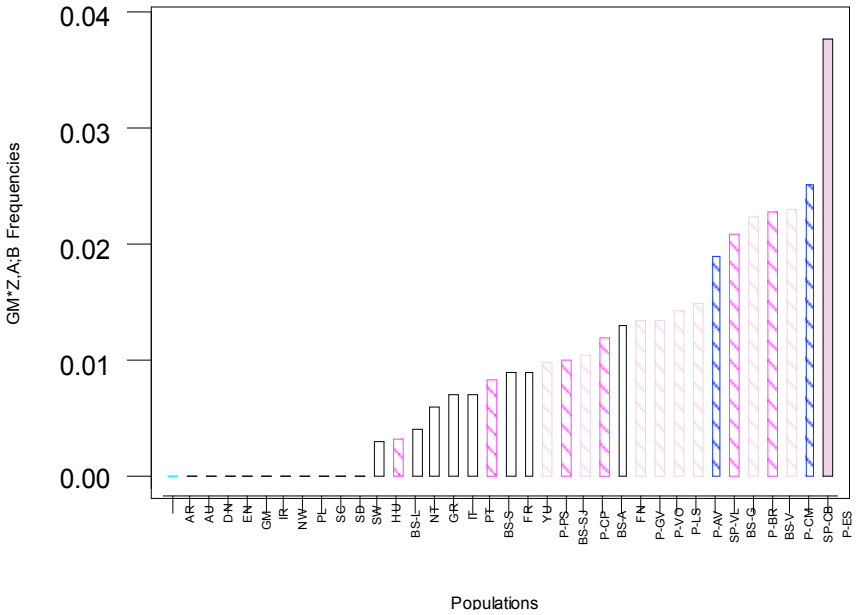

Figure 10. Distribution of GM*Z,A;B in European populations. The outlier is the Pyrenean population of Esparro (lilac), Basque populations are in pink, Spanish in blue and other Pyrenean populations in hatched lilac bars.

GM*Z,A;B is present at high frequencies in African populations, leading some researchers to attribute its presence in European populations to African admixture. However, this haplotype could result from recombination between two of the common European haplotypes, GM*Z,A;G and GM*F;B, and the African variant is often associated with another immunoglobulin allotype, IgA^2 (AM*2), while the GM*Z,A;B haplotype found in Valencia, Spain, is associated with IgA^1 (AM*1) (Schanfield *et al.* 1981). Unfortunately, data on IgA is absent for most other European populations, but it does call into question the African origin of this haplotype in Europe. Frequencies of GM*Z,A;B also increase as one travels south

and east across Europe, and the presence of this haplotype in Asian populations suggests a possible Asian origin for this haplotype in Europe (Stevenson and Schanfield 1981). Haplotype GM*Z,A;C,B (of known African origin) was found only in the village of Esparros in the Baronnies, and was attributable to descent from a West Indian immigrant, but was absent in all other autochthonous Pyrenean communities, including the Basques in France (Dugoujon *et al.* 1989; Giraldo *et al.* 1998).

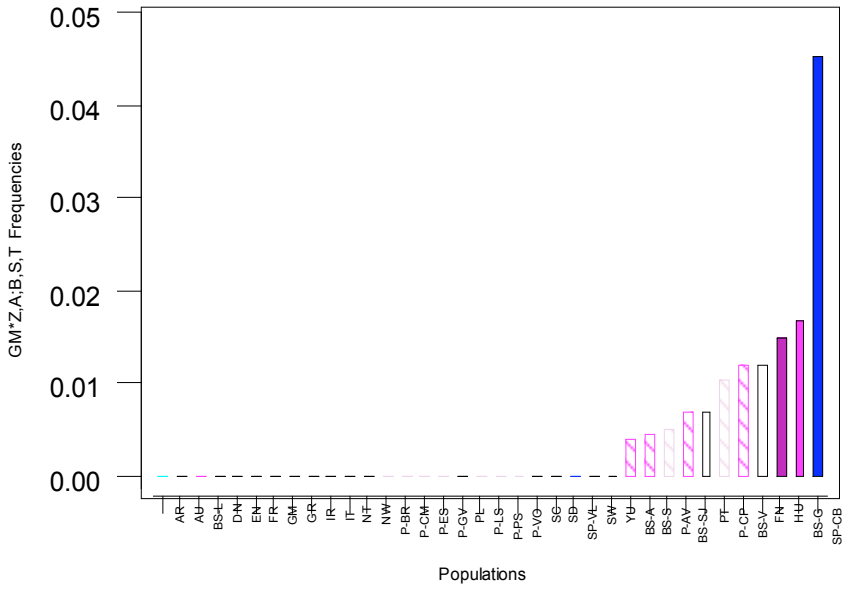

Figure 11. Distribution of GM*Z,A;B,S,T frequencies in European populations. Outliers are the Spanish in Cantabria (blue), Basques from Guipuzkoa (pink) and Hungarians (purple). Other Basque populations are shown and pink hatched bars, other Pyrenean populations as lilac hatched bars. This Asian marker is absent in most European populations, including Armenians from the Caucasus (turquoise), and is present at less than polymorphic frequencies (< 1%) in many Pyrenean populations, including Basques.

Haplotype GM*Z,A;B,S,T was found at polymorphic frequencies among Basques in Guipuzkoa and Vizcaya, but at lower frequencies in Alava (0.0040), Soule

(0.0050), and Basse-Navarre (0.0070). (Calderon *et al.* 1998; Dugoujon *et al.* 1989; Esteban *et al.* 1998). There were three outlier populations for this allele, including the Spanish community of Monte de Pas (Cantabria), the Basques of Guipuzkoa, and Hungarians (Figure 11). This haplotype has the highest frequency in Asia (China – 0.4400, Buryat – 0.3100), with a cline in all directions from there, leading Calderon *et al.* (1998) to hypothesize that the Basque population might be descendants of Neolithic immigrants from the Caucasus, east of the Caspian Sea. The authors do not elaborate on why the presence of this particular haplotype suggests a special origin for the Basques in particular, and unfortunately immunoglobulin data for Caucasian populations were unavailable in the literature, so their hypothesis could not be tested using immunoglobulin markers (Rychkov 2000).

Human Leukocyte Antigens

The Major Histocompatibility Complex (MHC) serves as the front line of the immune system. In humans, the MHC was first demonstrated in white blood cells (leukocytes), and so is known as the HLA (human leukocyte antigen) system. Three classes of HLA glycoproteins have been identified: HLA-I, HLA-II, and HLA-III. HLA-I molecules are found on all cells with a nucleus, and play a role in self-identification, protecting the body cells from being attacked by the immune system. When self-identification is compromised, autoimmune disorders, such as systemic lupus and rheumatoid arthritis, can result from the production of autoantibodies – antibodies which attack one's own cells (Mange and Mange 1990). HLA-I molecules also provide binding sites for viral antigens, tagging infected body cells for

destruction by cytotoxic T-cells (Alberts *et al.* 1989; Klitz *et al.* 1986). HLA-II molecules are found primarily on immune response cells, where they bind foreign cells (bacteria and other pathogens) and present them to lymphocytes for neutralization. HLA-III molecules are involved in the complement cascade, comprised of proteins (Bf, C2-C5) secreted into the plasma during the immune response, particularly in regards to bacterial invaders (Gruen and Weissman 2001; Lachmann and Hobart 1979; Martinez *et al.* 2001; Svejgaard 1979). Many of the HLA genes are located in close proximity on the short arm of chromosome 6 (Figure 12), and therefore tend to be inherited together, so the literature contains data on both allelic and haplotypic frequencies. Three class I loci have been frequently utilized in population studies (HLA-A, HLA-B, and HLA-C), along with three class II loci (DP, DQ, and DR). The genes coding for these antigens are some of the most polymorphic in humans, with 429 HLA-A, 751 HLA-B, 219 HLA-C, 514 HLA-DR, 101 HLA-DQ, and 144 HLA-DP alleles confirmed as of 2005, providing a wealth of data for anthropological genetic research (Marsh *et al.* 2005). The most common class I HLA alleles are described in Appendix 2, while the most common class II alleles are listed in Appendix 3.

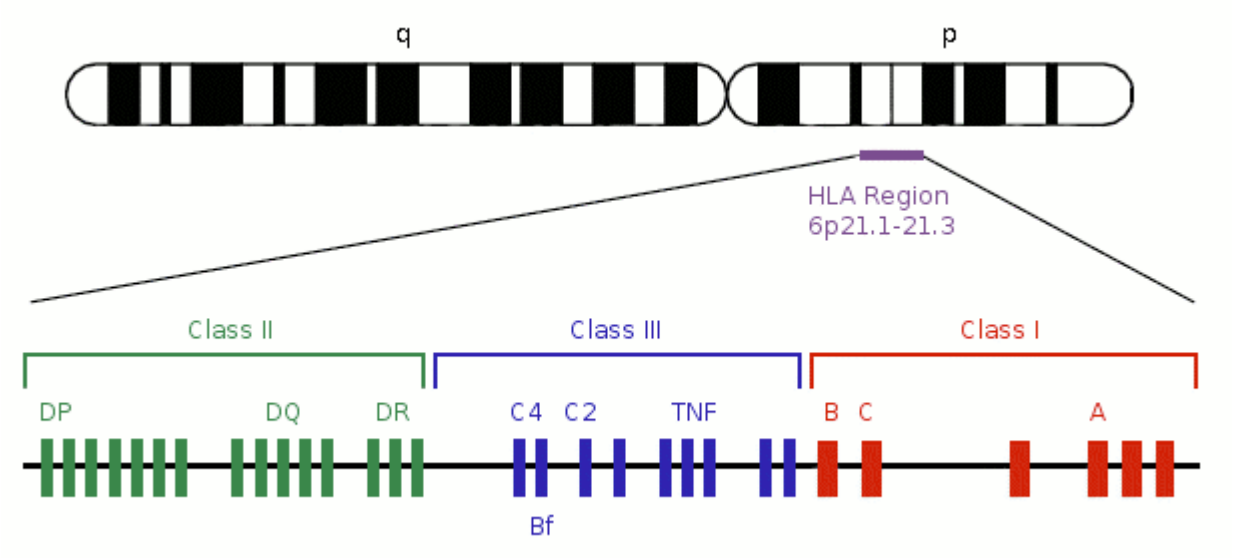

Figure 12. Physical map of the Human Leukocyte Antigen (HLA) Region on the short arm (p) of chromosome 6. Loci which have been examined in anthropological genetic studies are indicated. Adapted from Mehra and Kaur (2003).

HLA analysis has been the middle-ground between classic genetic analysis using antisera (a method employed since the discovery of blood groups), and molecular techniques using DNA sequencing. Unfortunately, the nomenclature has changed over time, as serologically-identified alleles were found to have multiple DNA sequence-identified alleles, and the need to distinguish alleles at the DNA level became apparent (Bodmer 1997). For example, HLA-A*01 contains 19 alleles identified to date, while HLA-B*58 contains 10. This makes compiling data from the literature difficult, as matching outmoded haplotypes with current nomenclature becomes confusing (Svejgaard 1979). Due to the abundance of alleles in this system, and contrary to Piazza *et al.'s* assertion that "a gene in common does not prove a common genetic origin" (1988b: 171), Arnaiz-Villena *et al.* recommend *not* combining HLA data with other classical markers for analysis, stating that it dilutes

the discriminatory power of the HLA system (1995; 1999; 1997a). Many specialists in the field have followed this approach, so HLA data are often presented and analyzed without the consideration of input from other systems.

Much research has been performed examining HLA allelic frequencies in Europe, and among the Basques in particular. Comparison of distributions of HLA-A and –B alleles throughout Europe shows four distinct clines. HLA B*7, B*8, and B*15 have high frequencies in northern Europe, fading toward the Mediterranean (Ryder *et al.* 1978). HLA B*5 shows the opposite trend, having the highest values in the Mediterranean, with low values in northern Europe. Other alleles identified at the time showed no clear pattern. Early work on HLA among the Basques was carried out in the French province of Labourd with a small sample of 88 "apparently unrelated" individuals in the total population of 500 (Dausset 1972). High levels of both A*29 and A*30 were reported. Haplotypes A*1-B*17 (0.0047); A*10-Da*31 (0.0033); A*11-B*27 (0.0040); A*30-B*18 (0.0056); and A*29-B*12 (0.0079) were the most frequent, and the Basques differed from both the neighboring French population and the Sardinians (Dausset 1972). A comparison of Basque populations in St. Jean-Pied-de-Port and Hasparren to non-Basques groups in the Pyrenean villages of l'Ouzom and Bareges found similar high frequencies of A*29, A*19.2 (A*30 and A*31), B*17, and B*18 in the Basques and in l'Ouzom, which were attributed to the likely distribution of Basque peoples in prehistoric times (de Mouzon *et al.* 1980). The most frequent biallelic class I haplotypes among the Basques in this study were A*19.2-B*35 (0.133), A*29-B*12 (0.1254), and A*1-B*17 (0.0807). Examination of HLA

frequencies in Basque students at the University of Bilbao (100 from Vizcaya, 50 from Guipuzkoa, 13 from Alava and eight from Navarre) found that the most frequent haplotypes were A*1-B*8 (0.1660), A*2-B*44 (0.0098), and A*3-B*7 (0.0069) (Calderon *et al.* 1993). All three haplotypes are characteristic of Western European populations. In comparison to neighboring populations in France and Spain, Basques have higher frequencies of A*19.2 (0.0440), A*29 (0.0640), B*18 (0.0800), B*12 (0.2530), B*7 (0.0130), and B*8 (0.0140). Differences between French and Spanish Basques for the HLA*A and HLA*B class I loci were also noted, with French Basques having elevated levels of A*29, A*19.2, B*17 and B*18, while Spanish Basques have higher levels of B*8.

Analysis of the class II HLA-DQA1 locus among 250 Basques from Guipuzkoa showed three alleles with frequencies above 20%: DQA1*0102 (0.2100), DQA1*0501 (0.2040), and DQA1*0201 (0.2100), which is the highest frequency for this allele yet reported (Iriondo *et al.* 1996). Guipuzkoans were not significantly different from populations in France (χ^2=1.86, p=0.8680), but they were significantly different from other populations in Spain, including Galicia, Andalusia, Valencia and Madrid, as well as populations in Algeria (Iriondo *et al.* 1996). Examination of the HLA-DQA1 locus in Alava, Guipuzkoa, and Vizcaya found no significant differences between the three Spanish provinces, but did find statistically significant differences between the Basques and other Spanish populations, with higher frequencies of HLA DQA1*0201 among Basques (0.1920) and lower frequencies of HLA DQA1*04* (which includes alleles 0401, 0501, and 0601) (0.2260) than in Galicians and

Catalonians (Esparza *et al.* 1995). Characterization of the Basques of Navarre for this locus found higher frequencies of HLA DQA1*0201 (0.2589), as well as high levels of DQA1*0501 – two alleles - (0.1920), DQA1*0102 – three alleles - (0.1741) and DQA1*0101 – two alleles - (0.1295) (Perez-Miranda *et al.* 2003). Analysis of additional class II loci, DPA1 and DPB1, in the Basques of Navarre revealed just two alleles with a frequency above 10%, DPA1*0201 (0.1853) and DPB1*0401 (0.3073) (Perez-Miranda *et al.* 2004). Both DPA1 and DPB1 are composite "alleles." DPA1*0201 has six sequence-identified alleles, while DPB1*0401 has two, but the authors did not test for these, meaning that these "alleles" are likely present at lower frequency among the Basques.

While allelic data are informative, haplotypic data are considered better measures of relationships between populations (Grimaldi *et al.* 2001), as "gene frequencies could have been changed by a variety of circumstances and various fusions [of populations] have been ignored…[so that] this method gives only a rough approximation of the origin of a population" (Degos and Dausset 1974: 195). Due to the high number of alleles at each HLA locus, haplotype frequencies are low, often less than 0.5% for the most common haplotypes worldwide, in comparison to Rh haplotypes involving three codominant loci, which can reach levels of up to 30% (Comas *et al.* 1998b; Svejgaard 1979). Grimaldi *et al.* (2001) use a benchmark of 2% to describe *frequent* HLA haplotypes used in population analyses. In a comparison of Mediterranean island populations with continental groups using extended haplotype analysis of Class I HLA loci, they reported seven common HLA Class I haplotypes

among French Basques: A*01-Cw*07-B*0801 (0.0320); A*02-Cw*0501-B*44(12) (0.0260); A*02-Cw*07-B*7 (0.0510); A*11.1-Cw*01-B*27 (0.0320), reported as A*01-B*27 in Dausset *et al.* (1972); A*23-Cw*04-B*44(12) (0.0380); A*29-Cw*1601-B*44(12) (0.0770), reported as A*29-B*12 by Dausset *et al.* (1972); A*30-Cw*0501-B*18 (0.0700), reported as A*30-B*18 (Dausset 1972; Grimaldi *et al.* 2001). Among Spanish Basques from Guipuzkoa, six class I haplotypes occur at a frequency higher than 2%: A*29-Cw*1601-B*44 (0.1002); A*30-Cw*0501-B*18 (0.0606); A*0201-Cw*0303-B*15 (0.0502); A*01-Cw*07-B*08 (0.0437); A*03-Cw*07-B*07 (0.0404); and A*0201-Cw*0501-B*44 (0.0303) (Comas *et al.* 1998b). The French and Spanish Basques have similar haplotype frequencies, with the exception of A*11.1-Cw*01-B*27, which is present in French Basques only, and A*0201-Cw*0303-B*15 and A*03-Cw*07-B*07, which are absent in French Basques. Table 4 presents the most frequent HLA haplotypes among Basques, as well as other populations in which those haplotypes are found.

Analysis of extended class I and class II haplotypes revealed four with frequencies higher than 2% in Spanish Basques: A*29-C*BL-B*44-DR*07-DQ*02 (0.0670); A*30-Cw*05-B*18-DR*03-DQ*02 (0.0360); A*01-Cw*07-B*08-DR*03-DQ*02 (0.0240); and A*03-Cw*07-B*07-DR*15-DQ*06 (0.0240) (Martinez-Laso *et al.* 1995b). The A*29 haplotype reported by Martinez-Laso *et al.* (1995) is presumably the same A*29-Cw*1601-B*44 haplotype reported in Comas *et al.* (1998b) and Grimaldi *et al.* (2001), as these studies were published more recently and additional C alleles have been identified. This haplotype is considered to be of

Western European origin (Arnaiz-Villena *et al.* 1997a; Tsuji *et al.* 1992), due to its presence in several Western European countries.

The A*30 haplotype (A*30-Cw*0501-B*18 in French and Spanish Basques) is found at appreciable frequency in several Mediterranean populations (Table 4). This haplotype is most frequent among Sardinians (0.1140) and has a frequency of less than 2% in Algeria (0.0150), Portugal (0.0170) and Morocco (0.0100) (Gomez-Casado *et al.* 2000; Grimaldi *et al.* 2001; Martinez-Laso *et al.* 1995b; Spinola *et al.* 2005). Arnaiz-Villena's team attribute an "Iberian-paleo-North-African" origin to this haplotype (Martinez-Laso *et al.* 1995b). Why it is not described as a Mediterranean paleo-European haplotype originating in Sardinia is not clear, if we follow other authors (Calderon *et al.* 1998) in assuming that the area of highest frequency indicates origin.

The A*01 haplotype (reported as the A*01-Cw*07-B*0801 haplotype in French Basques) is found across Europe (Imanishi *et al.* 1991; Spinola *et al.* 2005) and is described as a Celtic/Central European haplotype (Arnaiz-Villena *et al.* 1997a). It is also the most frequent "Caucasoid haplotype" (Rittner and Bertrams 1981). The Basques in Spain have haplotype frequencies (0.0240) similar to those found in Italy (0.0230), Portugal (0.0240), and France (0.0250), while the highest levels of the A*01 haplotype are found in Cornwall (0.0840) and the former Yugoslavia (0.0770).

Table 4. Common HLA haplotypes found in Basque populations. Frequency (F), sample size (N), and other populations in which the haplotype is found are indicated.

Frequent Haplotypes	*F*	*N*	*Population*	*Also found in*	*Reference*
A*01-Cw*07-B*08	0.0437	99	Guipuzkoa	Catalans, Spain, Italy, France, Greece	Comas *et al.* (1998b)
A*01-Cw*07-B*0801	0.0320	67	French Basques		Grimaldi *et al.* (2001)
A*01-Cw*07-B*08-DR*03-DQ*02	0.0240	82	Guipuzkoa	Spain, Britain, Denmark, Germany, Austria, Yugoslavia, Hungary, Italy, France, Romania, Portugal, Cornwall	Martinez-Laso *et al.* (1995a)
A*0201-Cw*0303-B*15	0.0502	99	Guipuzkoa	Catalans	Comas *et al.* (1998b)
A*02-Cw*0501-B*44(12)	0.0260	67	French Basques	Catalans, Spain, France, Italy, Balearic Islands	Grimaldi *et al.* (2001)
A*0201-Cw*0501-B*44	0.0303	99	Guipuzkoa		Comas *et al.* (1998b)
A*02-Cw*07-B*07	0.0510	67	French Basques	Spain, Corsica, Balearic Islands	Grimaldi *et al.* (2001)
A*02-Cw*07*B*07-DR*15-DQ*06	0.0180	82	Guipuzkoa	Spain, Cornwall, Austria, Algeria	Martinez-Laso *et al.* (1995a)
A*02-Cw*07-B*07-DR*15-DQ*01	0.0350	57	French Basques	Cornwall, Britain, Austria	Moreno *et al.* (1991)
A*03-Cw*07-B*07	0.0404	99	Guipuzkoa	Catalans	Comas *et al.* (1998b)
A*03-Cw*07-B*07-DR*15-DQ*06	0.0240	82	Guipuzkoa	Spain, Denmark, Austria, Czech, Algeria, Germany, Portugal, Yugoslavia	Martinez-Laso *et al.* (1995a)
A*11-Cw*01-B*27-DR*01-DQ*01	0.0440	57	French Basques	none	Moreno *et al.* (1991)

Frequent Haplotypes	*F*	*N*	*Population*	*Also found in*	*Reference*
A*11-Cw*01-B*27-DR*01-DQ*05	0.0120	82	Guipuzkoa	Spain, French	Martinez-Laso *et al.* (1995a)
A*11.1-Cw*01-B*27	0.0320	67	French Basques	none	Grimaldi *et al.* (2001)
A*2301-Cw*04-B*44(12)	0.0380	67	French Basques	none	Grimaldi *et al.* (2001)
A*29-Cw*1601-B*44	0.1002	99	Guipuzkoa	Catalans, Spain, France, Italy, Balearic Islands, Corsica	Comas *et al.* (1998b)
A*29-Cw*1601-B*44(12)	0.0770	67	French Basques		Grimaldi *et al.* (2001)
A*29-C*BL-B*44-DR*07-DQ*02	0.0670	82	Guipuzkoa	Cornwall, Spain, Balearic Islands, Corsica, France, Portugal, Denmark	Martinez-Laso *et al.* (1995a)
A*30-C*BL-B*18-DR*03-DQ*02	0.0470	57	French Basques	Spain	Moreno *et al.* (1991)
A*30-Cw*0501-B*18	0.0606	99	Guipuzkoa	Catalans, Spain, Corsica, Balearic Islands, Sardinia	Comas *et al.* (1998b)
A*30-Cw*0501-B*18	0.0700	67	French Basques		Grimaldi *et al.* (2001)
A*30-Cw*0501-B*18-DR*03-DQ*02	0.0360	82	Guipuzkoa	Spain, Sardinia, Algeria, Morocco, Portugal	Martinez-Laso *et al.* (1995a)

The A*03 haplotype was not reported in French Basques but is found at frequencies below 2% on the Iberian Peninsula and in North Africa. It is more frequent in other European populations, such as Danes (0.0360), Austrians (0.0320), Germans (0.0250), and Spanish Basques (0.0240) (Imanishi *et al.* 1991). Martinez-Laso *et al.* (1995b) decline to speculate on the origin of this haplotype, but Arnaiz-Villena *et al.* (1997a) describe it as a Northern European haplotype without

attributing any special relationship between Algerians and, for example, Danes, based on sharing a single HLA haplotype. However, a connection is made by the same authors between Basques and north-African populations based on one common haplotype - A*30-Cw*05-B*18*DR*03-DQ*02 - which is most frequent in Sardinia, and has a frequency of less than 2% in Algeria.

Class II HLA haplotype data are considered by some more discriminative than even class I HLA haplotype data, given that class II data are sequence-identified, while most of the class I data were serologically-identified and contain multiple alleles. However, it is now recognized that some class II alleles are also composite, and therefore contain multiple alleles themselves, although perhaps not to the same degree. DRB1*0701 has two alleles, and DQB1*0201 has five. The same argument for why HLA data should be used independently to examine population origins has now been used to show that the class I haplotype data are "generic" and "tend to homogenize comparisons" based solely on class II data (Gomez-Casado *et al.* 2000: 241). Comas *et al.* (1998b) describe 6 four-locus class II haplotypes common among Spanish Basques: DRB1*0701-DQA1*0201-DPA1*0103-DPB1*0401 (0.1393); DRB1*0701-DQA1*0201-DPA1*0201-DPB1*1101 (0.0668); DRB1*0101-DQA1*0101-DPA1*0103-DPB1*0401 (0.0658); DRB1*0301-DQA1*0501-DPA1*0103-DPB1*0202 (0.0647); DRB1*1501-DQA1*0102-DPA1*0103-DPB1*0401 (0.0611); and DRB1*0301-DQA1*0501-DPA1*0103-DPB1*0401 (0.0300). All of these haplotypes are found at appreciable frequency among Catalans in Spain, although the frequency of the DRB1*0701 haplotype is higher in Basques,

the highest reported in Europe (Comas *et al.* 1998b). Most populations have not been studied for the DRB1-DQA1-DPA1-DPB1 haplotype, so comparative data are not easily obtained. More often populations are typed for possibly two class II loci, in addition to the three standard class I loci. The most frequent 2-locus class II haplotype found among the Basques of Navarre was DPA1*0103-DPB1*0401 (0.3000) (Perez-Miranda *et al.* 2004), a component of 4 of the six common class II haplotypes described by Comas *et al.* (1998b). When those four haplotypes are summed to get the overall frequency of DPA1*0103-DPB1*0401 in Spanish Basques, the result is essentially the same (0.2962). The partial DPA1*01-DPB1*0401 haplotype is common in many populations, ranging from Swedes (0.4630) and French (0.4040), to Zuni (0.3300) and San Bushmen (0.0210) (Imanishi *et al.* 1991). It is also found at much lower frequency in Japan (0.0560) and China (0.0680) (Imanishi *et al.* 1991). However, these populations were not typed for the specific DPA1**0103* sequence, so it is impossible to ensure that they do in fact share the *same* haplotype. Given the sheer number of class II alleles (see Appendix 3), the more loci that are typed, the more discriminative the analysis becomes. Reliance on class II haplotype data to determine population relationships may not be advisable, especially in the absence of data from other markers, since many populations have not yet been typed with specificity and population data for all loci are lacking.

Basque Genetic Studies: Molecular Systems

Advances in technology and statistical methods have allowed for the analysis of increasing numbers of molecular markers, including autosomal and Y-chromosome

microsatellites (STRs), and mitochondrial DNA haplogroups and sequences. These data have been used to corroborate, or challenge, hypotheses of population origin and dispersal based on classical genetic markers.

Microsatellites

Microsatellites are sequences of 2-6 bases tandemly repeated 10-30 times, which are found scattered throughout the genome. These short tandem repeats (STRs) are considered selectively neutral, and therefore appropriate for population genetic studies. For microsatellites, genotype is determined not by protein product or DNA sequence, but by copy number. The system is also codominant, so that an individual who carried alleles measuring 13 and 17 repeats would express both. Mutation rates for these loci are relatively high compared to other types of markers in the nuclear genome (Table 5). Thirteen of these STR loci comprise the Combined DNA Index System (CODIS), used for forensic purposes. These loci have also been widely implemented in anthropological genetics, as identification databases provide a wealth of comparative data.

Table 5. Estimated mutation rates for selected molecular markers

Molecular Marker	***Estimated Mutation Rate (mutations/base/generation)***
Microsatellites	$10^{-3} – 10^{-4}$
Base substitutions (SNPs)	$10^{-7} – 10^{-8}$
mtDNA (whole genome)	3.4×10^{-7}
mtDNA (hypervariable region I)	3.6×10^{-6}

Several autosomal STR studies have been conducted among the Basques, but most were "announcements of population data" (Garcia *et al.* 1998; Garcia *et al.* 2001; Perez-Miranda *et al.* 2005b; Perez-Miranda *et al.* 2005c), analyses of only a

few loci (Alonso *et al.* 1995; Arrieta *et al.* 1997), or they reported scant results beyond tests for Hardy-Weinberg equilibrium and descriptive statistics (Iriondo *et al.* 1997; Perez-Lezaun *et al.* 2000). One phylogenetic analysis based on six STR loci from Navarre demonstrated that the Navarrese form a cluster with Basque populations from other Spanish provinces, off a branch that includes populations from northeastern Spain (Iriondo *et al.* 1999).

Y-Chromosome

The Y-chromosome provides a different perspective on questions of population history. Y markers are paternally inherited, do not undergo recombination, and reflect male lineages and migration. Two types of marker systems are in common use: (1) Y STRs – identical to their autosomal counterparts, these loci vary in the number of repeats and experience high mutation rates; and (2) Y SNPs – Single Nucleotide Polymorphisims, or biallelic markers. These are point mutations which evolve at a much slower rate, and typically are taken to occur only once per lineage.

Several studies, conducted using one or only a few Y-chromosome markers (Calderon *et al.* 2003; Lucotte and Hazout 1996; Quintana-Murci *et al.* 1999), demonstrate that Basques appear to harbor some ancient Y haplotypes at relatively high frequencies. Haplotype 15, based on p49/TaqI analysis of the DYS19 locus, has a frequency of 72% among French Basques. Analysis of 26 SNPs in various Iberian populations, including 45 Basques, revealed that they have a high frequency of the haplogroup R1 (62.3%), the most common in Western Europe (Flores *et al.* 2004). A

continental study of 11 Y SNPs incorporating 26 Basque samples showed "minor genetic barriers" separating them from neighboring populations, possibly due to the presence of a relatively young haplogroup (Hg 22) derived from the more common haplogroup 1 (Rosser *et al.* 2000).

Two studies on Y-STR frequencies in Iberia (Garcia *et al.* 2004; Gonzalez-Neira *et al.* 2000) noted significant differences between the Basques and other Iberian populations in pairwise comparisons. Y STR data have been found to have limited utility in deep-time evolutionary studies, because of the high mutation rate and consequent homoplasy, making determination of phylogenetic relationships difficult. However, STR haplotype variation appears mostly compartmentalized or is restricted to particular SNP backgrounds, in a comparative study of North African and Iberian Y lineages (Bosch *et al.* 1999). Out of 56 STR haplotypes, only one was found present in more than one SNP lineage. Similarly, copy number greater than 15 at the DYS388 locus can be used as an indicator of Y-chromosome haplogroup J (Flores *et al.* 2004).

In evolutionary studies, STR data are typically combined with Y SNP data, which have a slower mutation rate, and are used to construct the backbone of the phylogeny on which the STRs reside. Bosch *et al.* (2001) used this technique, and included a sample of 44 Basques, in their analysis of gene flow between northwest Africa and Iberia. The Basques showed a high frequency (57%) of Haplogroup 104, which is common in Europe, but rarely found in North Africa. A study of 168 Spanish Basques, using both Y SNPs and Y STRs, found that they have a high

proportion of the most frequent European haplotype, but also demonstrate a low level of STR diversity (Alonso *et al.* 2005). STR data have also been found useful for studies of intrapopulation heterogeneity, with STR haplotype grouping demonstrating geographic clines in Europe (Gusmao *et al.* 2003).

Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a circular, double-stranded extranuclear genome found in the mitochondria (Figure 13). It is inherited through the maternal line, from the mitochondria present in the ovum at fertilization. Genetic analysis of mtDNA involves restriction fragment length polymorphisms (RFLPs), as well as DNA sequences. RFLPs are used to define haplogroups, based on the presence/absence of enzyme-specific diagnostic cut sites. Such mtDNA haplogroups have been found to be generally continent specific. In Europe, for example, nine mtDNA haplogroups are common: H, I, J, K, T, U, V, W, X. Sequencing of hypervariable segments (HVS-I and II) of the control region (the d-loop in Figure 13) can further refine the haplogroup designations into haplotypes. For example, RFLP analysis can classify a sample as haplogroup H by the absence of an AluI cut site at position 7025 (Izagirre and de la Rua 1999), while characteristic mutations within the hypervariable region can assist in refining the classification to a particular haplotype (i.e., an A-G transition at position 16162 defines haplotype H1a) (Loogvali *et al.* 2004).

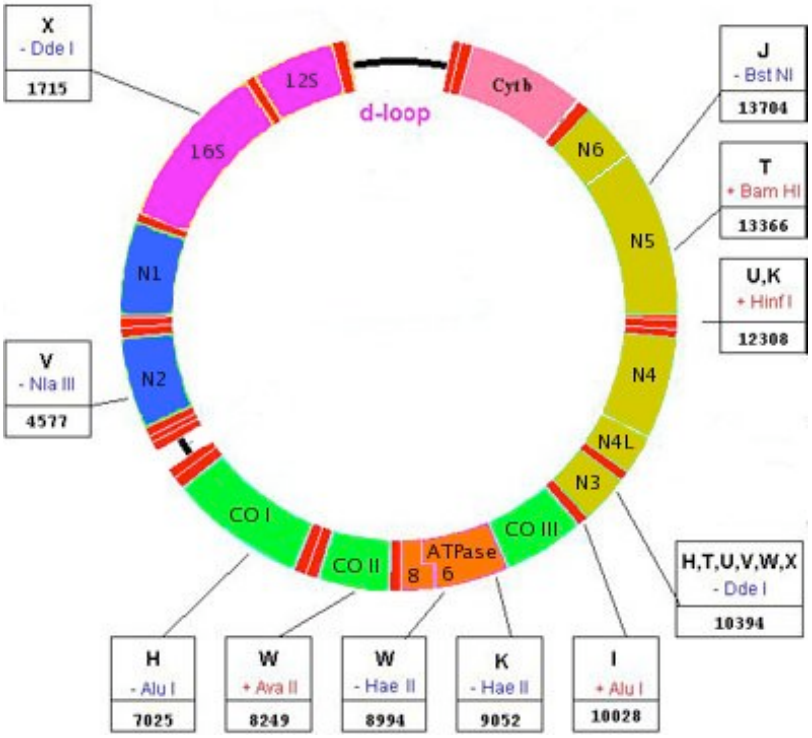

Figure 13. Diagram of mitochondrial genome, with restriction sites for European haplogroups highlighted.

Analysis of mitochondrial DNA, both haplogroup frequencies and control region sequences, has become a standard and widely used technique for inferring population history and evolution. The advantages of mtDNA, similar to those of Y markers, are that mtDNA exhibits uniparental inheritance and no recombination, allowing for the analysis of maternal lineages and gene flow. Among the Basques, both contemporary and ancient populations have been examined using this system.

Several haplogroup specific studies have been conducted, incorporating Basque samples in an attempt to flesh out the peopling of Europe and the origin of various haplogroups. A high frequency (20%) of Haplogroup V in Guipuzkoa was

presented as evidence that the Basque country was the homeland for this haplogroup, which may have originated during the last ice age, when the Franco-Cantabrian region was an ice free refugium (Torroni *et al.* 1998). The highest frequencies of subclades H1 (27.8%) and H3 (13.9%) were reported in the Spanish Basque country, H being the most frequent haplogroup in Europe (Achilli *et al.* 2004). In addition, the presence of the U8a haplotype among Basques has been attributed to Paleolithic occupation of the Basque region (Gonzalez *et al.* 2006).

Mitochondrial haplogroup analysis of 121 teeth from 4 prehistoric sites in the Basque region of Spain, whose occupation dates from the Neolithic to the Bronze Age, demonstrates the lack of haplogroups I and W, which correlates well with analyses of modern Basque populations (Izagirre and de la Rua 1999). Contrary to analysis of contemporary Basques, haplogroup V was also lacking. Analysis of mtDNA haplogroup frequencies of 37 individuals from a 6-7th century site in Alava show a temporal disparity in the Basque country, with the historical site resembling contemporary Atlantic fringe populations in their distribution of mtDNA haplogroups, while the prehistoric sites differ in higher frequencies of haplogroup K (Alzualde *et al.* 2005).

There have been seven studies examining control region sequence variation among the Basques (Alzualde *et al.* 2006; Alzualde *et al.* 2005; Bertranpetit *et al.* 1995; Corte-Real *et al.* 1996; Gonzalez *et al.* 2003; Gonzalez *et al.* 2006; Izagirre and de la Rua 1999). One of the earliest analyzed 45 HVS-I sequences from Guipuzkoa (Bertranpetit *et al.* 1995). At the time, relatively little comparative data were

available, and the authors were able to compare the Basques to only one other European population, the Sardinians. Of the 27 different sequences present in the Basque sample, 1/3 were identical to the Cambridge reference sequence (CRS), which is known to be Haplogroup H. Sixty-one HVS-I sequences from Alava and Vizcaya, included as part of a larger analysis of genetic diversity on the Iberian Peninsula, also showed a preponderance of sequences belonging to Haplogroup H (Corte-Real *et al.* 1996). Thirty-four unrelated control region sequences were also analyzed from the 6-7th century site in Alava discussed previously (Alzualde *et al.* 2006). Three different J haplotypes were reported, as well as one designated M1c. Haplogroup J has been described as a signature of the Neolithic expansion into Europe from the Middle East, while M1c is common in North Africa, suggesting that, historically, this particular population did not experience the genetic isolation often proposed to account for the distinctiveness of contemporary Basques. To date, mtDNA analyses among Basque populations have presented conflicting results. On the one hand, studies of contemporary populations suggest that the Basques are a relatively isolated population with deep European roots, while ancient DNA, conversely, demonstrates contact (but not necessarily gene flow) with possible non-European populations, at least in historical times.

Biological anthropologists use all of the systems discussed previously to examine the evolutionary relationships among different populations, and the evolutionary forces that may have affected them, such as gene flow, genetic drift, or selection. Piazza *et al.* caution that it is necessary to distinguish "a similarity by

common environment [biology, from] common culture" (1988b: 171). They suggest that differences be judged not based solely on a *single* gene (i.e., Basques have a high frequency of Rh negative (RH*cde), so obviously they are different from every other European population). Rather, it is the shared differences which appear in many systems that can provide the most insight into a population's evolutionary history.

Basque Origins: Hypotheses

Basque-Caucasian Hypothesis

The Basque-Caucasian hypothesis has been examined using classical genetic markers (Aguirre *et al.* 1991a; Bertorelle *et al.* 1995). Cluster analysis demonstrated that subpopulations sampled in Vizcaya were more genetically similar to each other than to other European populations or Caucasian groups outside Europe, such as those in Asia Minor and the Middle East (Aguirre *et al.* 1991a). Comparison of Basque and Caucasian populations using 10 blood group and serum protein loci revealed that both non-Indo-European groups were more genetically similar to their neighbors than to each other (Bertorelle *et al.* 1995). However, genetic distances between Basques and North Caucasian populations were lower than expected for three loci (ACP1, GLO1 and Kell), while genetic distances between Basques and South Caucasians were lower than expected at five loci, the three found in the North Caucasus plus ABO and HP. Regardless, Basques clustered with neighboring non-Basque groups in 99.71% of the 10,000 bootstrapped Neighbor-Joining trees, and in 67% of the 10,000 bootstrapped UPGMA trees.

The proposed relationship between Basques and Caucasian populations has also been explored using HLA data. The Svani (a Kartvelian-speaking population) and the Basques were found to have only one five-locus extended haplotype in common, A*01-B*8-DRB1*03-DQA1*0501-DQB1*0201 (Sanchez-Velasco and Leyva-Cobian 2001). This is the A*01-*Cw*07*-B*08-DR*03-DQ*02 haplotype described earlier, the most frequent HLA haplotype found in Europeans, and is also present in the Svani at a frequency of 0.0125 (less than 2%) and among Basques at 2%. The authors conclude that HLA analysis, in agreement with analyses based on other classical systems, does not support the hypothesis of a relationship between these groups.

Recent studies of molecular markers find little similarity between Basques and populations in the Caucasus. Analysis of Y-SNP haplogroups found that F_{ST} values between Basques and Caucasian groups were much greater than between Basques and surrounding Indo-European populations (Nasidze *et al.* 2003). While comparison of mtDNA sequences did reveal greater affinity between European groups and Caucasians than between West Asians and Caucasians (Nasidze and Stoneking 2001), the addition of populations from Iran resulted in a genetic picture in which the Caucasus fell between groups from Europe and Asia Minor with respect to mtDNA sequence variation (Nasidze *et al.* 2004). As with Y-SNPs, genetic distances based on mtDNA sequences were greater between Basques and Caucasians that between Basques and Indo-Europeans, lending credence to the hypothesis of no genetic relationship between Basques and Caucasian populations.

Vasco-Iberian Hypothesis

An extensive review of the literature provided few studies testing the Vasco-Iberian hypothesis based on traditional genetic markers. Bosch *et al.* (1997) did examine the population history of North Africa using classical systems and compared them to other Mediterranean groups. They found that Basques cluster with populations from Andalusia in southern Spain, while the Berbers were genetically similar to West African groups, and overall emphasized the genetic distinction of North Africa. This study did not, however, include other European populations for comparison, as the primary focus was elsewhere.

The majority of studies supporting a relationship between the Basques and North African populations have been published by Arnaiz-Villena and his colleagues. From their research, they suggest a link between the Basques and the Hamites specifically, a Paleolithic population which eventually diverged into the Berbers and the Egyptians (Arnaiz-Villena *et al.* 1995; Arnaiz-Villena *et al.* 2002; Arnaiz-Villena *et al.* 2001; Arnaiz-Villena *et al.* 1997a; Arnaiz-Villena *et al.* 1981; Martinez-Laso *et al.* 1995b). The assertions are based on tenuous and controversial linguistic evidence, at times stating that Basque is a remnant of ancient Iberian, other times declaring Basque to be part of a larger "Usko-Mediterranean" family which includes extant Dene-Caucasian languages and Berber, as well as extinct languages such as Etruscan and Minoan (see Trask quote above) (Arnaiz-Villena *et al.* 2002). In an examination of the "correlation between languages and genes" among the Usko-Mediterranean peoples, no statistical analysis (i.e., a Mantel test) of such a correlation is provided

(Arnaiz-Villena *et al.* 2001). Examination of class I HLA-A, -B, and –C and class II HLA-DR and –DQ allele frequencies in Algeria found significant differences between Algerians and Basques from Guipuzkoa, with lower frequencies of alleles A*29, B*44, and DRB1*0101 in Algerians, and higher frequencies of alleles A*24, B*53, B*49, B*35, B*38, and DRB1*0102. In their 2002 article, "Population genetic relationships between Mediterranean populations determined by HLA distribution and a historic perspective," Arnaiz-Villena *et al.* use high resolution (allele-level) data from just two class II loci (DR and DQ) to determine historic and prehistoric affinities between several Mediterranean groups. Two of their figures are reproduced here for clarification. Figure 14 shows a dendrogram with the Basques and Algerians highlighted in pink. Notice that they are on the same branch, implying a genetic relationship. But also note that this branch occurs in only 50% of the trees. In fact, 75% of the branches in the phenogram occur in only half the trees, demonstrating a lack of "statistical robustness" for this analysis (Bosch *et al.* 1997). Figure 15 presents a correspondence analysis from the same data, again with Basque and Algerian populations highlighted. Here, the Algerian group clusters with Spanish and Portuguese populations, not Basques. The Basques fall between the French and Moroccan groups and the Souss population, Berber speakers from southern Morocco. All of these groups are included in a single large cluster by the authors. However, in Figure 14, the Souss branch off near the Greeks and Moroccan Jews, nowhere near the other populations seen in the large cluster in Figure 15.

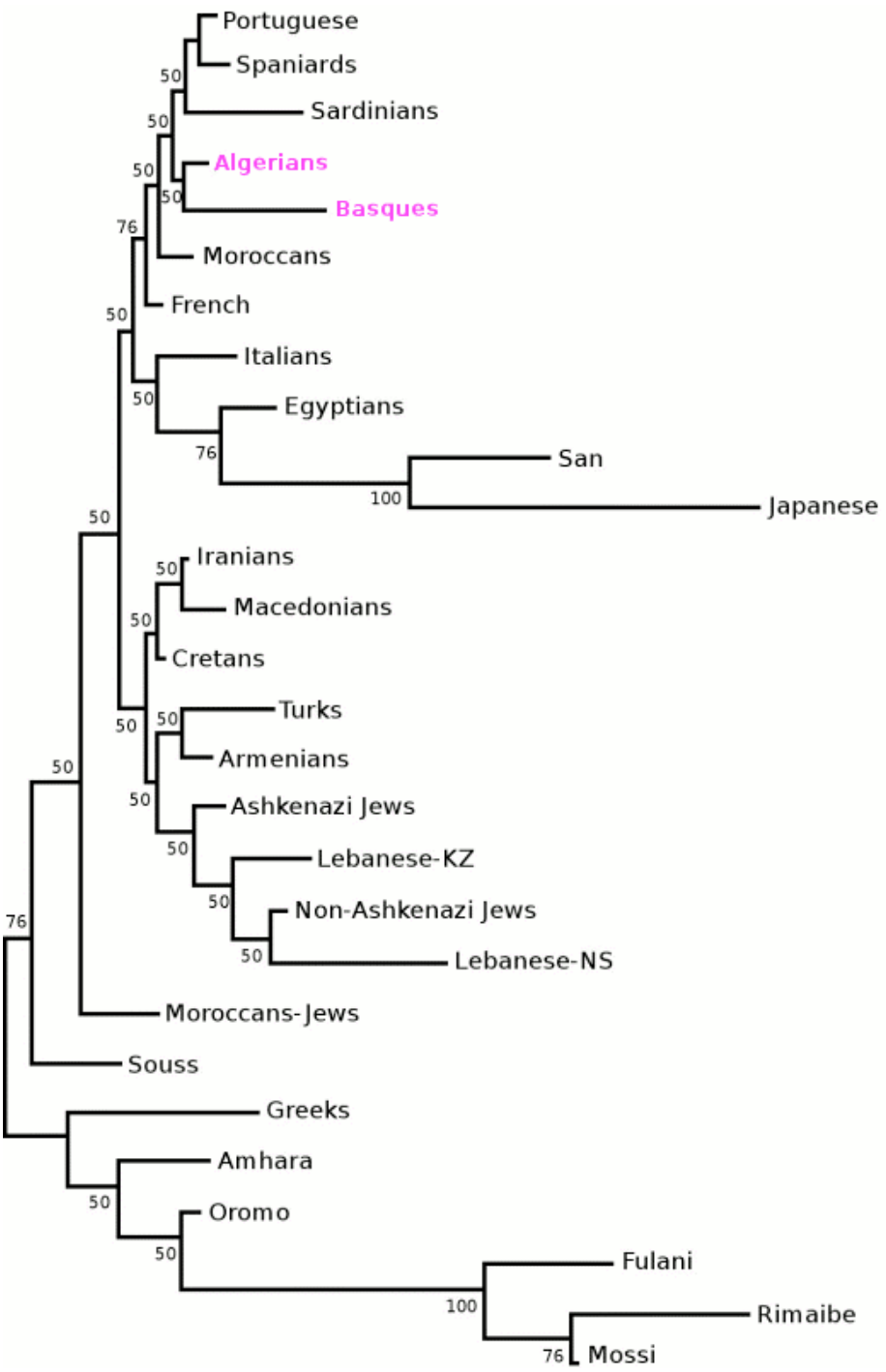

Figure 14. Phylogenetic tree redrawn from Arnaiz-Villena et al. (2002). In this analysis, the Basques cluster with Algerians in 50% of the bootstrapped trees. However, it is based on data from only two HLA loci (DR and DQ).

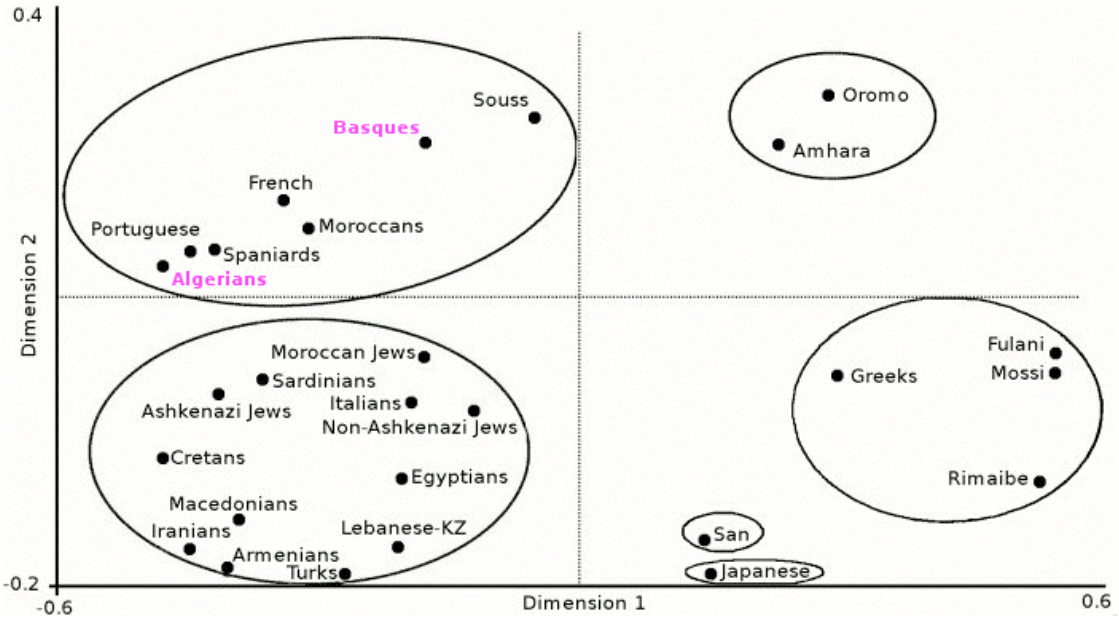

Figure 15. Correspondence analysis redrawn from Arnaiz-Villena (2002), based on the same HLA-DR and –DQ data. Algerians cluster closest to Spanish and Portuguese populations, while the Basques fall between the French and Souss populations.

Sanchez-Velasco *et al.* (2003) added several isolated populations from northern Spain (Cabuernigos, Cantabrians, and Pasiegos) to Arnaiz-Villena's earlier work, and while still insisting that Basques "are close to North Africans in most analyses," their figures do not support this assertion. The Neighbor-Joining dendrogram based on the same alleles used in the previous two figures shows that the Basques cluster from the Spanish branch with the other northern Iberian populations (Figure 16). Algerians form a second cluster off the Spanish branch with two other North African populations (El Jadida and Souss). Correspondence analysis including intermediate-resolution HLA-A and –B data in addition to the class II loci described previously again places the Basques with other isolated northern Iberian groups. Algerians cluster with other North African and European populations, including

Spanish, French, Italians, and Sardinians (Figure 17), possibly reflecting, in part, gene flow from Europe in colonial times (Bosch *et al.* 1997), rather than a common Paleolithic ancestor.

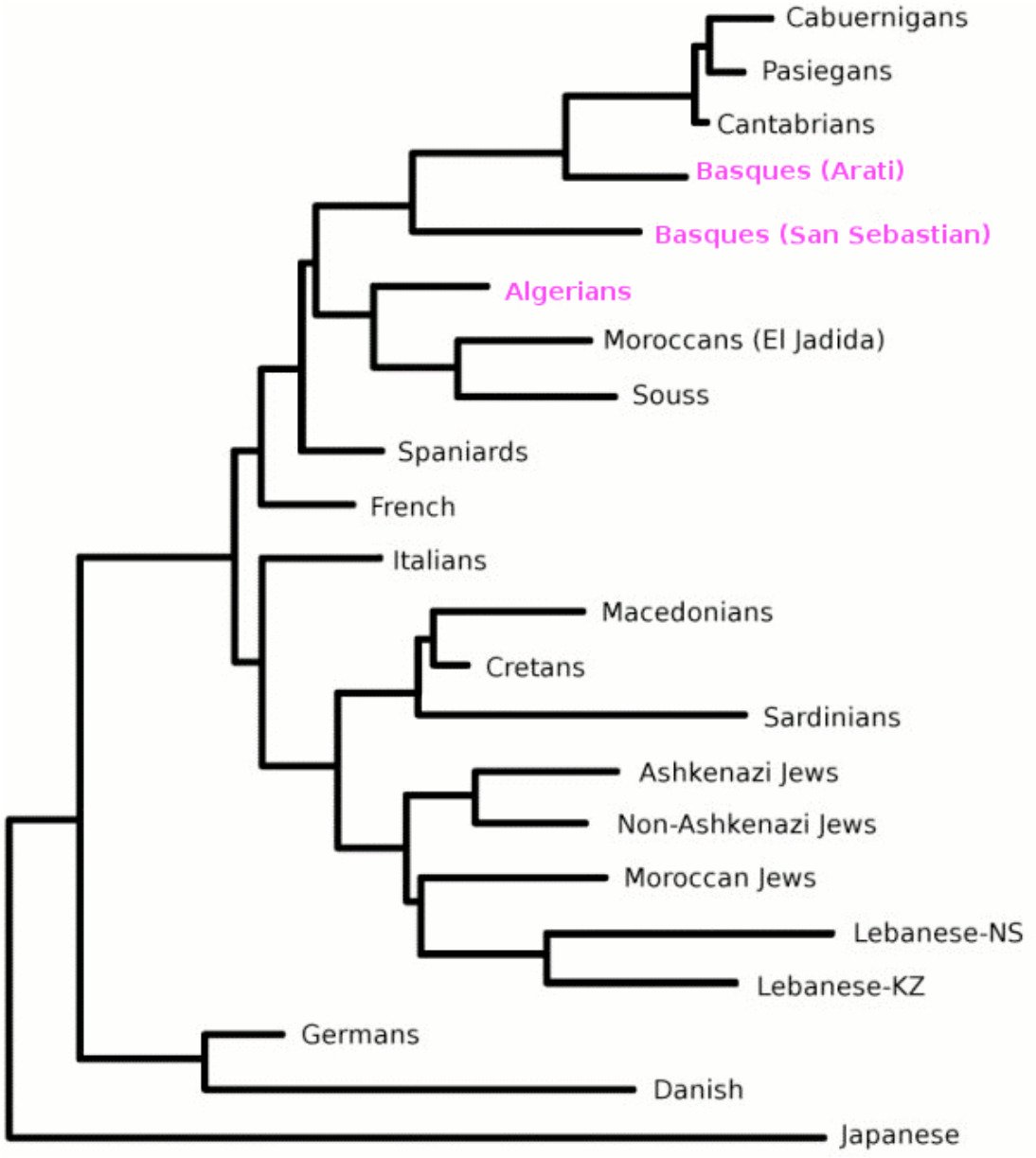

Figure 16. Neighbor-Joining dendrogram redrawn from Sanchez-Velasco et al. (2003). The Basques cluster with other northern Iberian populations.

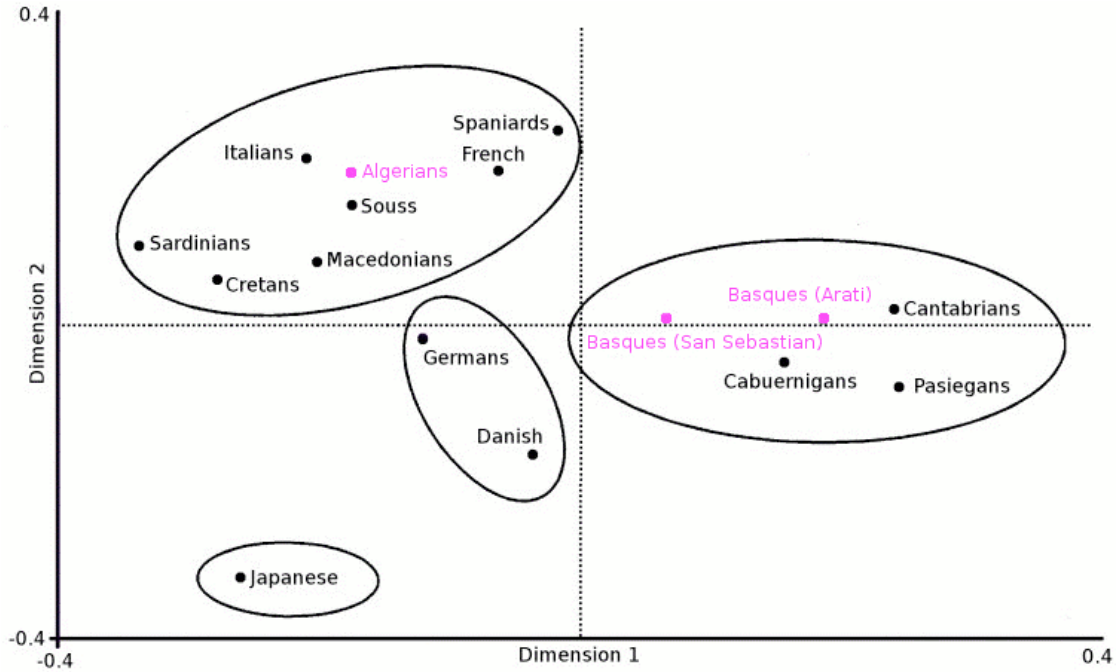

Figure 17. Correspondence analysis of DA genetic distances based on HLA-A, -B, -DRB1, and DQB1 data. The Basques cluster with other northern Iberian populations, while the Algerians cluster with other North African and European populations.

Contrary to Arnaiz-Villena's hypothesis, multidimensional scaling of the R-Matrix for 10 alleles at the DQA1 locus (Figure 18) demonstrated that the Basques from Vizcaya and Alava clustered with other European populations, while Basques in Guipuzkoa and Navarre cluster together (Perez-Miranda *et al.* 2003). None of the Basque groups fall near the Algerian population, suggesting that the DQA1 locus does not support a common origin for Basques and Algerians.

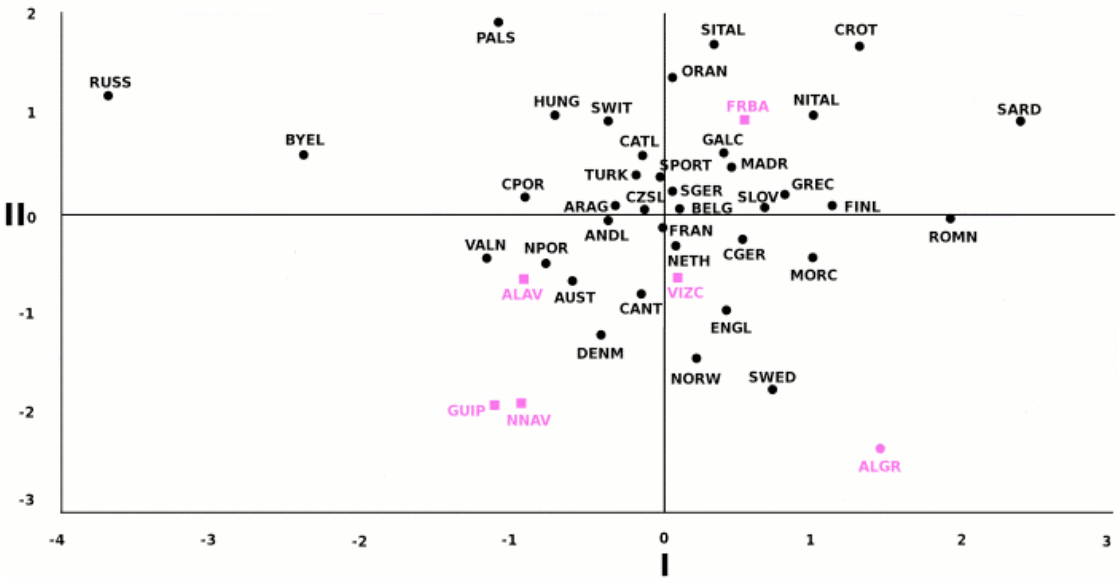

Figure 18. MDS analysis of HLA-DQA1 data from many European populations and Algeria (reproduced from Perez-Miranda et al. (2003). Basques from Guipuzkoa (GUIP), Northern Navarre (NNAV), Alava (ALAV), Vizcaya (VIZC), French Basques (FRBA) are shown as pink squares, while Algerians (ALGR) are represented as a pink circle. Basques from Guipuzkoa and Northern Navarre cluster together, while the other Basque groups cluster near other European populations. None of the Basque groups cluster near Algerians.

A more comprehensive comparison of Basques and Algerian populations using HLA-A, HLA-B, HLA-C, HLA-DRB, and HLA-DQA allelic data placed the Basques on the Spanish branch with Catalans (Figure 19), with the Algerians closer to Italians and Romanians (Comas *et al.* 1998b). Garcia Fernandez *et al's* (1997a). Analysis of the relationships between Basques and other populations based on HLA-A, -B, -DR, and –DQ allele frequency data show that, in fact, the Basques are distinct from other European populations, but still lie in that section of the tree, separated from the Algerians by Danish and Spanish groups (Figure 20). In an analysis of serological HLA-A and HLA-B data in addition to HLA-DRB1, HLA-DQA1, and HLA-DQB1 sequence data, Basques also cluster with other European groups (Portuguese and Spanish), not Algerians (Arnaiz-Villena *et al.* 1999). The results

repeatedly published by Arnaiz-Villena's group are not reproduced by other researchers.

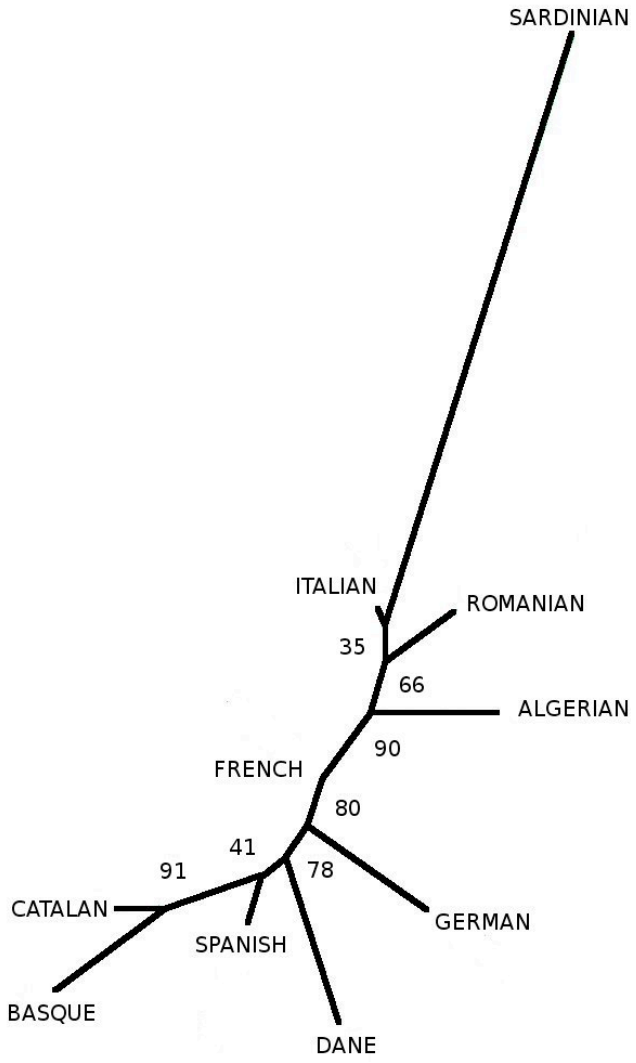

Figure 19. Neighbor-Joining tree based on allele frequencies for HLA-A, -B, -C, -DRB, and -DQA data. Basques segregate with other populations from the Iberian Peninsula, while Algerians do not (Comas et al. 1998b).

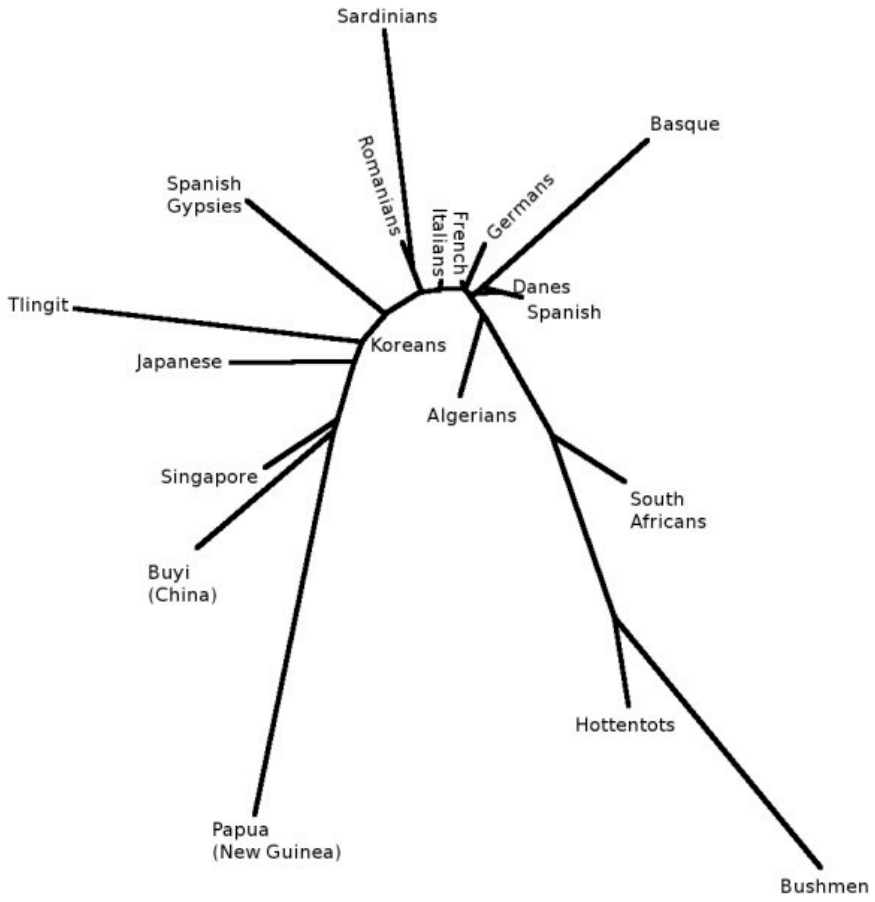

Figure 20. Neighbor-Joining tree redrawn from Garcia-Fernandez et al. (1997a) showing the relationships between world populations based on HLA-A, -B, -DR, and –DQ allele frequencies. Basques are in the European region of the tree, near Spanish, Danish, German and French populations, but also distinct from other groups, as shown by the extended branch length.

Debate has also arisen as to the ability of the HLA system to effectively reflect population history, insofar as it allows the discernment of population specific alleles or haplotypes. Comas *et al.* (1998b) emphasize the influence of genetic drift throughout much of human history, stating:

> it seems *incorrect* to allocate specific alleles or haplotypes to populations or to cultural events that may trigger a population expansion [but rather it is important] to take into consideration the basic properties of allele (or haplotype) frequency differences and changes as being highly dependent on demographic events in the past, when populations were much smaller and drift might have had its greatest effects [emphasis added]" (-8).

In fact, evidence has emerged that the HLA system may be a hot spot for selection. Mutation rates of HLA loci are comparable to those at other loci, with rates of 1.37 per site per billion years for HLA-A, 1.84 for HLA-B, 3.87 for HLA-C, and 1.18 for HLA-DRB1 (Raymond *et al.* 2005; Satta *et al.* 1993). However, analysis of HLA-A, -B, and –DR loci among the Portuguese found that all three had significantly higher heterozygosity values than expected (Spinola *et al.* 2005). Certain HLA alleles and haplotypes have also been linked to specific diseases. The Basques have been reported as having one of the highest frequencies of the properdin factor F1 (Bf*F1) in Europe (0.1393) (Ohayon *et al.* 1980), which has been associated (on a haplotype background including B*8, C4*Fs, and DR*03) with increased incidence of early onset diabetes (IDDM) (Alper *et al.* 1986; de Mouzon *et al.* 1979). The most frequent HLA haplotype among Europeans (A*1-Cw*7-B*8-DR*03-DQ*02) has been linked to multiple autoimmune disorders, including IDDM, celiac disease, systemic lupus, as well as increased mortality after HIV infection (Price *et al.* 1999). Among Basques, risk of early onset IDDM is increased in individuals who have HLA-DR*03, with a relative risk of 29.1 in individuals with that particular allele (Cambon-de Mouzon *et al.* 1982; Perez De Nanclares *et al.* 2000). HLA-DR*03 is found in two of the most frequent extended haplotypes among Basques (Table 4).

Irrespective of disease associations, the MHC appears to have been subject to balancing (overdominant) selection (Hughes and Nei 1989; Klein *et al.* 1993). Overdominant selection would be expected in a situation where heterozygous individuals have a selective advantage. In the HLA system, it seems obvious that heterozygotes would have such an advantage, as they have the potential to mount a more effective immune response against a wider variety of pathogens. Several class I and class II HLA loci have lower than expected homozygosity values (F), indicating the influence of balancing selection, while the class III loci (the complement system) appeared selectively neutral (Begovich *et al.* 1992; Klitz *et al.* 1986). Significant correlation between HLA diversity and "pathogen richness" has also been observed (Prugnolle *et al.* 2005). Noting that between 17-39% of the diversity seen in class I HLA genes is accounted for by geographic distance from East Africa, and thus the history of modern human world colonization, the researchers found an additional 8-11% of worldwide HLA diversity could be explained by the number of viruses present in a particular environment. Thus pathogen diversity is positively correlated with HLA allelic diversity.

Hughes and Nei (1988; 1989) conducted analyses of synonymous and nonsynonymous substitution rates in the antigen recognition sites (ARS) versus other regions of class I and class II loci. Synonymous substitutions (SS) typically occur in the third position of a codon, and so do not change the amino acid specified at that particular position (Strachan *et al.* 1984). For example, a point mutation occurring at the third position in the codon ACC, changing the C to an A (or a U or a G, in this

case), would not change the amino acid coded for, namely threonine. Nonsynonymous substitutions (NS) tend to occur at the first and second positions of the codon, resulting in a point mutation that changes the amino acid at that position (Hughes and Nei 1989). Using the previous example, the same substitution in the second position, changing codon ACC to AAC, changes the amino acid to asparagine. In the ARS of class I loci, the frequency of NS is significantly higher than the frequency of SS. For regions of class I loci not involved in antigen recognition, the pattern is reversed. They note:

> …in most eukaryotic genes…the number of nucleotide substitutions between polymorphic alleles is very small (0.0001-0.02 per nucleotide site) and…most of these substitutions are synonymous. The high values of [NS] and [SS] observed here are therefore unique to the polymorphic alleles at the MHC loci. Furthermore, our observation that [NS] is significantly greater than [SS] in a functionally important part of the gene seems to be without precedent (Hughes and Nei 1988: 169).

Further analysis indicates that the ratio of NS:SS in the ARS region of many HLA loci is likely double Hughes and Nei's estimates (Takahata *et al.* 1992). At the HLA*DPB1 locus, NS rates are nearly 200 times higher than SS rates at the ARS site (Gyllensten *et al.* 1996). This pattern is not seen in nonhuman primates, however, leading the authors to speculate that humans were subject to greater selection pressures as they colonized new areas after the human/chimp split. While the DQA*1 and DQA*3 alleles have higher overall NS/SS rates, allele DQA*4 does not (Erlich and Gyllensten 1991). Having a lower ratio of NS/SS than that expected under neutral conditions suggests that this allele has been under selective constraints, perhaps to maintain functionality as the α polypeptide chain produced by this gene

pairs with multiple β chains produced by the DQB locus. Perhaps these constraints have been operating for 5-20 million years, as this allele is also found in chimpanzees and gorillas, and thus predates, and has survived, multiple speciation events (Fan *et al.* 1989). Similar results, comparing human and chimpanzee sequence data, have been found for the class I HLA*A and *B loci, with the HLA-A1 allele 97.7% identical to a chimpanzee MHC allele (Ch25) (Lawlor *et al.* 1988). Other researchers suggest that HLA class II haplotypes have been evolving independently for 40 million years, as a result of the high levels of maximal pairwise divergence between dissimilar haplotypes, ranging from 2.1-9.3% for human-human sequence comparisons (Raymond *et al.* 2005). By contrast, the last common ancestor for human mtDNA dates to around 200,000 years ago. Analysis of SNP profiles suggest that balancing selection may occur not only at the HLA loci, but possibly across the whole subregion, as areas of high nucleotide diversity appear at other loci on chromosome 6q, including some retroviral sequences (Gaudieri *et al.* 2000).

HLA data may prove useful in elucidating hominoid evolution greater than 1 million years ago, but it appears that there is too high a rate of nonsynonymous substitution and that coalescence times are too great to make the system useful in studies of modern human populations which arose in the last 150-250,000 years (Takahata 1993). It seems unwise to infer recent population history from these data, especially from composite alleles or low-resolution haplotypes (neither of which can establish identity by descent as they do not distinguish sequence-variant alleles) and in the absence of other comparative genetic data. A study incorporating data using

five polymarker loci (Low Density Lipoprotein Receptor (LDLR), Glycophorin A (GYPA), Hemoglobin G gammaglobulin (HBGG), STR D7S8, and GC) and HLA*DQA1 to study relationships between the Basques and other populations worldwide illustrates this point (Brown *et al.* 2000). Considering only the polymarker loci, the Basques cluster most closely with other residents of the Basque country and then other residents of Spain. When allele frequency data from the HLA*DQA1 locus was added to the analysis, the Basques cluster more closely with populations from Tajikistan (a country of Indo-European speakers located just north of Afghanistan and Pakistan) than Basque country residents. The authors point out that the addition of the HLA locus produced a phylogeny that was "contrary to both the results of other phylogenetic studies and to scientific consensus" (Brown *et al.* 2000: 150).

Preliminary investigation of molecular systems also disagrees with the results of Arnaiz-Villena *et al.* Analysis of nine CODIS autosomal STR loci in a sample of 68 Basques from Vizcaya demonstrated similarity with Basques from Guipuzkoa and other Iberian populations, and distinction from North African groups in Morocco and the Maghreb, casting further doubt on the Vasco-Iberian hypothesis (Zlojutro *et al.* 2006).

Pre-Indo-European Hypothesis

Given that the Basques are not Indo-European, the question remains: who are they? The non-Indo-European nature of their language would mean, accepting Renfrew's assertion (1987) that Indo-European came into the continent with the

advent of agriculture around 10,000 BP, that the Basques were not part of the Neolithic migration from the Levant, at least from a linguistic perspective. However, Renfrew's hypothesis that the Neolithic farmers were Indo-European has been challenged by linguists, who point out that the earliest dates for a proto-Indo-European language range from 4500-2500 BC, and that the linguistic substrate in Anatolia is decidedly non-Indo-European (Mallory 1989). For Basques, this represents confirmation of their uniqueness, and plays a role, along with their language, in the formation of Basque identity and their quest, particularly in Spain, for political autonomy (Haarmann 1998; Urla 1993).

For European history, the question of Basque origins relates to the peopling of the continent. All inhabitants of Europe, whether Paleolithic or Neolithic, are immigrants. The issue is a microcosm of the Out-of-Africa/multiregional evolution debate: Did the Paleolithic inhabitants of Europe contribute to the modern gene pool? If so, to what degree? Alternatively, are modern Europeans the result of a replacement of the Paleolithic groups by more advanced Neolithic farmers? The debate has centered around two competing models. The Neolithic demic diffusion model (DDM), proposed by Ammerman and Cavalli-Sforza (1984) states that the majority of genetic variation present in modern Europeans is the result of the bands of Neolithic farmers spreading their technology (and genes) into Europe with the advent of agriculture. The cultural diffusion model (CDM) posits that while the technology spread into Europe 10,000 years ago, the people did not, so that there was a transfer

of technology, but not genes, leaving the Paleolithic gene pool largely intact (Novelletto 2007).

Exhaustive studies of the interplay between genetic, linguistic, and geographic variation in Europe have been conducted using classical markers (Barbujani *et al.* 1995; Derish and Sokal 1988; Sokal 1988; Sokal 1991a; Sokal 1991b; Sokal *et al.* 1989a; Sokal *et al.* 1993; Sokal and Menozzi 1982; Sokal *et al.* 1990; Sokal *et al.* 1989b; Sokal *et al.* 1988; Sokal *et al.* 1992a; Sokal *et al.* 1999a; Sokal *et al.* 1999b; Sokal *et al.* 1996; Sokal *et al.* 1991; Sokal *et al.* 1992b). Analysis of twenty-one loci (7 blood groups, 4 plasma proteins, 5 erythrocyte enzymes, immunoglobulins GM and KM, and HLA-A and B) for 3369 populations divided into five language phyla (Indo-European, Uralic, Altaic, Afro-Asiatic, and Basque) confirmed the genetic heterogeneity of the European continent, with significant correlations between genetic and linguistic distances reported, even when the effect of geographic distances was held constant (Sokal 1988). This means that the language families in Europe are significantly genetically different from each other. Also examined was the effect of single loci on heterogeneity, in which ABO, RH, TF, ACP1, and ADA were shown to be the primary contributors to the observed genetic differentiation. Recall that some Basque populations were outliers at the ABO and RH loci, and on the low end of the range for ACP1 and ADA. Perhaps the genetic distinctiveness of the language families of Europe can be partially explained by the inclusion of the Basques. Subsequent analysis of these data noted that the Basques and Finnic speakers were outliers from other European language families (Harding and Sokal 1988). The

distinction of Finnic peoples was attributed to the presence of Lapps, who speak a Finnic language but are genetically similar to other Siberian populations. The difference between Basque and other European groups is purported to be the result of their presumable descent from pre-Indo-Europeans. However, this study used a phenogram to describe the relationships between these populations, and Piazza *et al.* (1988) caution that known historical events in Europe (invasions, migrations) violate one of the primary assumptions of phylogenetic reconstruction – independent evolution between branches of the tree, so this type of representation might not be the best fit for European data.

A more thorough analysis of these data examined the effects of historical events on European gene frequencies using language family boundaries (Sokal *et al.* 1988). In essence, are gene frequencies between populations more different across language family boundaries than they are within a given language family? Language boundaries were found to correspond to areas of genetic change, and the most genetically distinct boundary was between Germanic and Romance speakers due primarily to differences at the HLA-B locus. The second distinct boundary fell between Basque and Romance speakers as a result of differences at the RH, TF, and GM loci. Evaluation of genetic differences between 12 language families in Europe (Albanian, Baltic, Celtic, Germanic, Greek, Romance, Slavic, Finnic, Ugric, Turkic, Semitic, and Basque), showed that peripheral populations (Basque, Celtic, Finnic and Greek) were outliers nearly 4 times more frequently than core populations, suggesting the possibility that clinal variation, in addition to linguistic and geographic distance

between populations, plays a role in the genetic differentiation of Europe.

The effects of specific historic events on genetic variation have also been examined using an ethnohistory database containing information on populations including their language, location, and significant events (such as migrations, settlements and invasions) between 2000 BC and 1970 AD based on a 5ºx5º quadrant map of Europe (Sokal 1991a). A stronger correlation was reported between historical movements of populations that occurred prior to 500 AD and modern gene frequencies (most often HLA loci) than with more recent events, lending credence to the idea that modern populations can be used to examine questions of ancient history. The areas with the most significant emigration of populations were the Balkans and central Europe, including Germany, reflecting perhaps the influx of Indo-Europeans into the continent in the first case, and the expansion of the Celtic and Germanic tribes in the second. Movements into northern Iberia appeared significant before 500 AD, but not after. This same database was used to develop "ethnohistorical distances" (based on Cavalli-Sforza's arc distance) between language families, to examine how they correlate with genetic and geographic distances (Sokal *et al.* 1993). These ethnohistorical distances were more highly correlated with geography than were the genetic distances, and significant partial correlations were reported between ethnohistorical distance and genetic distance for 10 systems, with geographic distance held constant. The significant systems include ABO (with A subtypes), MN, P, Kell, Duffy, TF, GC, PGM1, AK, and HLA-A. Plotting the correlation of ethno-historical distance and genetic distance over time, a substantial increase (-0.3300 to 0.3730)

between 1100 BC and 1000 BC was described, due mainly to Celtic expansion from central Europe. No significant correlations were reported for the Basque region, due perhaps to the lack of information early in this period, as only historically recorded events could be examined. Overall, a study of classical genetic markers (34 loci with 95 alleles) noted that European populations as a whole are more similar to each other than populations on other continents, with an F_{ST} of just 0.0142, compared to F_{ST} values ranging from 0.0393 in Australia to 0.0755 in the Americas (Cavalli-Sforza and Piazza 1993). Of the 26 populations studied, seven were consistently outliers in the construction of phylogenetic trees (Lapps, Sardinians, Greeks, Yugoslavs, Basques, Icelanders, and Finns). Four of the seven populations speak Indo-European languages (Sardinians, Greeks, Yugoslavs and Icelanders), while two speak Uralic languages (Finns and Lapps). Only the Basques are linguistic isolates.

Comparison of Basques and other populations using five blood groups (ABO, RH, MNS, Kell, and P1), five plasma proteins (GC, BF, HP, TF, PI), one erythrocyte enzyme (GLOI), the immunoglobulins GM and KM, as well as two HLA loci (A and B), found that Indo-European groups were more similar to Basques than non-Indo-European ones (Piazza *et al.* 1988b). Specifically, the greatest genetic similarity was found between the Basques and the population of Béarn in France, which the authors attribute to the distribution of Basques in ancient times, in agreement with analysis of the toponymy of the region (Piazza *et al.* 1988a). An analysis of eight Pyrenean populations, including Basques from France and Spain, using twenty-six classical markers (blood groups, isozymes, immunoglobulins and HLAs) determined that

Basques were distinct from other populations in the region, leading the authors to assert that the Basque differentiation occurred *in situ* over a long period of time, likely prior to the Neolithic Revolution (Calafell and Bertranpetit 1994b). Synthetic gene map analysis, popularized by Cavalli-Sforza, but now widely criticized for over-interpolation and generation of smooth clines from few data points (Sokal *et al.* 1999a; Sokal *et al.* 1999b), has also been applied to the study of Basque origins (Bertranpetit and Cavalli-Sforza 1991b). Using classical data for 20 systems from 11 locations in the Iberian Peninsula (compiled from 635 studies), differentiation between the Basque region and the rest of the peninsula was shown to account for the greatest amount of variation (27.1%) (Bertranpetit and Cavalli-Sforza 1991a). The authors maintain that this distinction reflects the occupation of that region since the Paleolithic. A computer simulation of fluctuations in gene frequencies on the peninsula from the Paleolithic through the post Neolithic found that greatest amount of variation (31.5%) was explained by the genetic peculiarity of the Basques. Analysis of actual gene frequency data matches the results found in the simulation, except that the actual data shows a larger, more gradual shift from the Basque region to the rest of the peninsula (Calafell and Bertranpetit 1993b).

Genetic evidence has also been used in support of both models of the peopling of Europe. Specifically, Cavalli-Sforza *et al.* (1994), in their synthetic gene map analysis of classical markers, noted a cline for several loci in the first principal component, accounting for approximately 27% of the total variation, spreading from southeast to northwest through Europe. They interpreted this cline as a genetic

signature of the DDM model noting a correlation coefficient of 0.89 between the first principal component of gene frequencies and temporal spread of agriculture into Europe (Ammerman and Cavalli-Sforza 1984). Similar clines were reported in a study of seven autosomal markers (HLA-DQA plus 6 STRs), indicating "directional population expansion" (Chikhi *et al.* 1998:9055). Some Y-chromosome studies produced comparable results, with clinal variation reported in the frequencies of 12 of 27 STR alleles examined in 16562 Y-chromosomes with maximum effective divergence times (τ_{max}) ranging from 281 – 10296 years, well within the Neolithic (Casalotti *et al.* 1999). This suggests that for the Y-chromosome, most of the variation present in the European gene pool dates to the Neolithic, though this analysis only accounts for male variation.

In support of the CDM hypothesis, advocates are quick to point out that the Paleolithic expansion into Europe occurred from the same area as the Neolithic expansion, implying that the gradient seen in some classical markers might be a Paleolithic signal (Barbujani *et al.* 1998). Additional evidence comes from paleoanthropology, as well as molecular data. In the Basque region, there is substantial archaeological evidence of Paleolithic human occupation, as described in Table 1 (Bertranpetit *et al.* 1995). Analysis of prehistoric skeletal remains in Iberia showed little transition in cranial morphology or evidence of dental caries between the Mesolithic and Neolithic, a transition which would be expected if the Mesolithic hunter/gatherers had been summarily replaced by farmers with different dietary patterns (Jackes *et al.* 1997). In anthropological genetics, analysis of European

mtDNA HVS I sequences demonstrated little clinal variation, and divergence dates suggested that many of the mitochondrial haplotypes in Europe had a pre-Neolithic origin (Richards *et al.* 1996). This work was criticized on several counts, from accusations that the authors were inferring population history from a single locus, to questions about the accuracy of the molecular clock, to the observation that the divergence time was accurate only if the assumption was made that the population had undergone a genetic bottleneck and all pre-existing variation had been lost before they migrated into Europe (Barbujani *et al.* 1998; Cavalli-Sforza and Minch 1997). The authors readily admitted that mtDNA could only reflect the maternal lineage, and concerns about the molecular clock proved unfounded (Macaulay *et al.* 1997; Richards *et al.* 1997). However, they did correct the divergence times, by removing variation present before the migration in the form of shared haplotypes (Sykes 1999). After correction, only one major haplogroup – J -- could be placed in the Neolithic. Haplogroups H, V, I, W, T, and K showed corrected divergence dates between 14,000-11,000 BP, while U5 had a divergence date of 50,000 BP. Phylogeographic analysis appears imperative, because Europe is fairly homogenous from a mitochondrial perspective, with an average of 40% of all European mtDNA belonging to haplogroup H (Roostalu *et al.* 2007). Mitochondrial haplogroup analysis of European populations would show little genetic diversity, and provide virtually no information on relationships between groups (Torroni *et al.* 1996; Torroni *et al.* 1994). Lineage analysis of the H haplogroup demonstrated that haplotypes H1 and H3, found at high frequencies among the Basques, appear to be a signature of

Paleolithic populations trapped by the Last Glacial Maximum in the Cantabrian refugium, which includes the present-day Basque country and was one of the few habitable regions in Europe 20,000 years ago (Achilli *et al.* 2004). A similar claim has been made for haplotype U8a and haplogroup V, which have high gene diversity rates among the Basques (Gonzalez *et al.* 2006; Torroni *et al.* 2001). U8a dates to around 28,000 BP and is absent in the Levant, while V was dated between 15,000 – 10,000 BP. A study of four Basque samples dating between the Neolithic and the Bronze Age, however, found a total absence of haplogroup V, which could be the result of random error due to limited sample size or could suggest that the divergence of this haplogroup is more recent (Izagirre and de la Rua 1999).

Further studies of Y-chromosome haplotypes revealed one (M173) which appears to show evidence of expansion after the Last Glacial Maximum (Wells *et al.* 2001), and a "high degree of non-Neolithic ancestry" in populations of Iberia (Flores *et al.* 2004). A subclade of this haplotype, M153, is reported to have the second highest frequency (7%) after M173, among a sample of Basque males, and a divergence date between 21,300-17,900 BP (Alonso *et al.* 2005). Other studies reported Y-chromosome haplotypes whose presence in Europe might be explained by Neolithic expansion (M89, J-M172, J-M267, E-M78, E-M123), or reflect more recent male gene flow in Iberia (Hg22) (Hurles *et al.* 1999; Semino *et al.* 2004).

For a time both sides of the debate were polarized by the notion that it had to be all or nothing, either all Europeans had mostly Neolithic genes, or mostly Paleolithic ones (Casalotti *et al.* 1999; Chikhi *et al.* 2002). This stems partially from

both sides drawing the history of an entire population from analysis of one or only a few loci. Single genes are not subject to the same evolutionary pressures as entire genomes (or populations), especially with regards to selection, as is the case with HLA loci. In addition, divergence times for specific loci or haplotypes do not necessarily indicate the age of the population in which it is found, or the time at which it arrived in its present location (Chikhi *et al.* 1998). It is, rather, the convergence of evidence from multiple lines of inquiry, including classical markers, mtDNA, Y-chromosome, autosomal STRs, archaeology, linguistics, paleoanthropology, and paleoecology, which provide the most comprehensive insight into the question of the role of the Basque population in the peopling of Europe.

CHAPTER THREE: MATERIALS AND METHODS

Sample Collection

During summer field sessions between 2000-2002, DNA samples were collected in mountain villages throughout the four Basque provinces of northern Spain by Dr. A. Apraiz, under the support of a National Geographic Society Grant to the University of Kansas Laboratory of Biological Anthropology (Project 6935-00). The Basque region of Spain is highlighted in Figure 1. Participant numbers by province, as well as the molecular systems analyzed, are shown in Table 6. This study was approved by the University of Kansas Human Subjects Committee (HSCL #11955). Participants signed informed consent statements (Appendix 4), and were provided with researcher contact details, should additional information be required. Buccal samples were collected from autochthonous participants (those who claimed four Basque grandparents) by rubbing a sterile wooden dowel along the cheeks to remove epithelial cells. The samples were stored in sterile TE until extraction in the Laboratory of Biological Anthropology at the University of Kansas. Data on classical genetic markers in these populations were compiled from the literature for comparative analysis.

Table 6. Basque Participants and Molecular Systems Characterized by Province

Province	*Villages (N)*	*Males*	*Females*	*Autosomal STRs*	*mtDNA RFLP*	*mtDNA sequences*	*Y-STRS*
Alava	6 (143)	45	64	96	110	36	41
Vizcaya	17 (237)	91	102	133	197	51	46
Guipuzkoa	10 (220)	64	100	128	173	42	58
Navarre	2 (52)	9	21	25	30	0	13
Total	35 (652)	209	287	382	510	129	158

Laboratory Methods

DNA extraction

DNA extraction was performed by the author at the Laboratory of Biological Anthropology (University of Kansas) using a standard phenol:chloroform protocol. Each sample was digested in a 55°C waterbath overnight in a mixture of 5X STE (100 µl), 10% SDS (25 µl), 20 mg/ml Proteinase K (25 µl), and 325 µl ddH_2O in order to lyse the cells. Then, 250 µl of 4°C 5M potassium acetate was added to the samples, which were incubated at -20°C for 15 minutes and then centrifuged at 10,000 RPM for 30 minutes to precipitate and pellet the proteins. The DNA was then cleansed of additional cellular debris using two phenol:chloroform:isoamyl alcohol (24:24:1) extractions. Next, the DNA was precipitated using two volumes of 4°C 95% ethanol and incubated at -20°C overnight. Finally, the DNA was washed with 4°C 70% ethanol (750µl), allowed to air dry, resuspended in 50µl sterile TE buffer (pH = 8.0), and stored at 4°C.

Autosomal STR Analysis

A portion of each sample was forwarded to Dr. Ranjan Deka (Department of Environmental Health, University of Cincinnati Medical Center) for STR analysis, which was performed by Dr. Guangyun Sun using the Applied Biosystems Profiler Plus Kit (Foster City, California). The samples were characterized for nine STR loci (detailed in Table 7), including D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA, and vWA, plus the sex-determining amelogenin locus.

Determination of STR profiles involved a multiplex PCR reaction (per sample) of: 9.5µl AMPF1STR PCR reaction mix; 5µl primer set solution; 0.5µl Ampli*Taq* Gold™ DNA polymerase; and 0.5-2.5ng genomic DNA. Amplifications were performed in a Perkin-Elmer 9600 PCR thermal cycler according to the following cycle conditions: preliminary denaturation for 11 minutes at 95°C; 26 cycles of denaturation for one minute at 94°C, annealing for one minute at 59°C, extension for one minute at 72°C; and a final extension for 45 minutes at 60°C followed by a hold at 4°C. Detection of the amplified products was accomplished using an ABI 377 DNA sequencer. PCR products and the AMPF1STR standard allelic ladder were individually combined with the same volume of formamide loading solution (formamide/blue dextran/internal standard GeneScan-500 ROX from Applied Biosystems in the ratio 5:1:1), denatured for five minutes at 95°C, and snap-cooled for five minutes on ice. Samples and allelic ladders were then electrophoresed on polyacrylamide denaturing sequencing gel for 2.5 hrs at 51°C and 3000V. GeneScan 3.1 and Genotyper 2.5 software (Applied Biosystems) were used for sizing and genotyping fragments.

Table 7. Autosomal STR loci used in the present analysis

Locus	*Repeat Sequence*	*Genomic Location*	*Reference*
D3S1358	AGAT	3p	(Li *et al.* 1993)
FGA	CTTT	4q28	(Mills *et al.* 1992)
D5S1818	AGAT	5q31-31	(Hudson *et al.* 1995)
D7S820	GATA	7q	(Green *et al.* 1991)
D8S1179	TATC	8	(Oldroyd *et al.* 1995)
vWA	TCTA	12p	(Kimpton *et al.* 1992)
D13S317	TATC	13q22-31	(Hudson *et al.* 1995)
D18S51	GAAA	18q21.3	(Urquhart *et al.* 1995)
D21S11	TCTA	21	(Sharma and Litt 1992)

Y-chromosome Analysis

Samples from male Basques were analysed for Y-STR haplotypes by Dr. Guangyun Sun in Dr. Ranjan Deka's laboratory (Department of Environmental Health, University of Cincinnati Medical Center). Y-STR profiles were determined using the Y-Plex™ 12 kit (Reliagene Technologies, Inc., New Orleans, Louisiana) for 12 Y loci, 11 STRs plus the sex-determining amelogenin locus, detailed in Table 8.

Table 8. Y-chromosome STRs used in the present analysis

STR Locus	*Repeat*	*Allele Range*
DYS392	TAT	6-18
DYS390	TCTA/TCTG	17-28
DYS385a/b	GAAA	7-25
DYS393	AGAT	8-17
DYS389I	TCTG/TCTA	10-17
DYS389II	TCTG/TCTA	24-34
DYS391	TCTA	6-14
DYS19	TAGA	10-19
DYS439	GATA	8-15
DYS438	TTTTC	6-14
Amelogenin	X, Y	--

Samples were amplified using the PCR reagents specified by the manufacturer: 10µl 2.5X Y-Plex™ primer mix; 0.5µl 5U/µl Ampli*Taq*™ Gold; 5µl (1-2ng) sample DNA; and ddH_2O to bring the final volume to 25µl per sample. PCR was conducted according to manufacturer's instructions using a Perkin Elmer 9600 thermal cycler under the following conditions: preliminary denaturation for 10 minutes at 95°C; 30 cycles denaturation for 60 seconds at 94°C, annealing for 60 seconds at 58°C, extension for 60 seconds at 70°C; final extension for 60 minutes at 60°C and a hold at 4°C. Detection of the amplified products was accomplished using

an ABI 377 DNA sequencer. PCR products were denatured with a formamide loading solution (formamide/blue dextran/internal standard GeneScan-500 ROX from Applied Biosystems in the ratio 5:1:1), heated for five minutes at 95°C, and snap-cooled for five minutes on ice. Samples and allelic ladders were then electrophoresed on 4% polyacrylamide denaturing sequencing gels for 2.5 hrs at 51°C and 3000V. GeneScan 3.1 and Genotyper 2.5 software (Applied Biosystems) were used for sizing and genotyping.

Mitochondrial DNA Analysis

Analysis of mitochondrial DNA included haplogroup assignment using restriction fragment length polymorphisms (RFLPs) and haplotype determination through sequencing of the first hypervariable segment of the control region (HVS-I). Mitochondrial haplogroups were established using a hierarchical approach (Santos *et al.* 2004), so that the most frequent European haplogroup -- (H) -- was tested for all samples first. Those samples that gave a negative result for H were then further tested for the other European haplogroups (U, K, V, I, J, T, W, and X). Halogrouping was performed by the author at the Laboratory of Biological Anthropology (University of Kansas). Sequencing the HVS-I region allows for the characterization of mtDNA haplotypes within each haplogroup, and was generally performed in the forward direction only, with the exception of those samples which have a T-C transition at position 16189. This transition creates a long run of cytosines, which can result in sequencing failure. Sequencing these samples in the reverse direction, from position 16400-16189, and then splicing the sequences together using sequence

alignment editing software, generates a complete sequence which can be used in further analyses.

Haplogroup diagnostic restriction sites were amplified using polymerase chain reaction (PCR) under the following conditions (per sample): 2.0µl 10X PCR Buffer; 1.2µl 25mM $MgCl_2$; 1.6µl 10mM deoxynucleotide mix (dNTP); 0.1µl 5U/µl *Taq* DNA polymerase; 0.6µl 10pmol/µl forward primer; 0.6µl 10pmol/µl reverse primer; 5.0-10µl DNA dilution (final concentration 10-100ng); and ddH_2O to bring the final reaction volume to 25µl. All PCR reagents were obtained from Promega (Madison, Wisconsin), with the exception of the primers, which were synthesized by Integrated DNA Technologies (IDT, Coralville, Iowa). A description of primers and annealing temperatures used in the haplogroup analysis is given in Table 9. PCR amplification was performed in either a Perkin-Elmer Applied Biosystems Gene Amp 2400 or 9700 according to the following cycle conditions: preliminary denaturation for one minute; followed by 35 cycles of denaturation for 40 seconds at 94°C; annealing for 30 seconds; and extension for 45 seconds at 75°C; followed by a hold at 4°C once amplification was complete. Success of PCR was determined by electrophoresis of a portion of each sample on a 1.5% SeaKem agarose gel, prepared with 1X TBE and stained with ethidium bromide, at 97V for approximately 1 hour. Gels were then photodocumented on a transilluminator under UV light.

The amplified DNA was then digested with the corresponding restriction enzymes (Table 9), which consistently identify and cleave specific DNA sequences

(typically 4-6 bases in length). Mitochondrial haplogroups are determined by the presence or absence of diagnostic restriction sites. These sites were identified using restriction digests under the following conditions (per sample): 2.0µl enzyme specific buffer (provided by the manufacturer); 1.0µl 100X bovine serum albumin (BSA); 0.5µl restriction enzyme (New England Biolabs, Beverly, Massachusetts); 7.5µl PCR DNA; and 9.0µl ddH_2O for a final volume of 20µl. Samples were digested for 10-18 hours at 37°C, and then reactions were terminated by the addition of 5µl of 3X loading dye (Promega, Madison, Wisconsin). The digests were then electrophoresed using 3% NuSieve agarose gels (ISC Bioexpress, Kaysville, UT), prepared with 1X TBE and stained with ethidium bromide, for 2 hours at 97V. The size of DNA fragments was measured against a standard ladder (25bp DNA step ladder, Promega, Madison, Wisconsin). Gels were photodocumented on a transilluminator under UV light.

The samples were sequenced in two locations: a portion of 100 samples from Alava and Vizcaya were sent to Dr. Ric Devor (IDT, Coralville, Iowa) for preliminary sequencing of the HVS-I mtDNA control region (16001-16400) using Big Dye Sequencing kits and an ABI 310 Sequencer (Applied Biosystems, Foster City, California). Seventy-five samples from Vizcaya and Guipuzkoa were sequenced by Dr. Mike Grose at the University of Kansas Sequencing Laboratory using Big Dye Sequencing kits and an ABI 3130 Sequencer (Applied Biosystems, Foster City, California). Sequencing reactions (per sample) consisted of: 4.0µl Big Dye Ready

reaction mix; 2.0µl Big Dye 5X sequencing buffer; 1.0µl forward or reverse primer; and 4.0µl sample DNA. The samples were amplified using the following PCR

Table 9. Primers used in mtDNA RFLP and sequence analyses

Haplogroup (Restriction Site)	Primers	Sequence (5' ⇒ 3')	Annealing Temperature
H (-7025 AluI)	FOR 6958 REV 7049	5' – CCTGACTGGCATTGTATT – 3' 5' – TGTAAAACGACGGCCAGTTGATAGGACATAGTGGAAGT – 3'	58°C
U/K (+12308 HinfI)	FOR 12216 REV 12338	5' – CACAAGAACTGCTAACTCATGC – 3' 5' – ATTACTTTTATTTGGAGTTGCACCAAGATT – 3'	55°C
K (-9052 HaeII)	FOR 9003 REV 9105	5' -- CCTAACCGCTAACATTAC – 3' 5' – TGTAAAACGACGGCCAGTGAAGATGATAAGTGTAGAGG – 3'	51°C
V (-4577 NlaIII)	FOR 4519 REV 4620	5' – CACTCATCACAGCGCTAAGC – 3' 5' – TGGCAGCTTCTGTGGAAC – 3'	55°C
J (-13704 BstNI)	FOR 13626 REV 13729	5' – CCTAACAGGTCAACCTCGCT – 3' 5' – TGTAAAACGACGGCCAGTCTGCGAATAGGCTTCCGGCT – 3'	64°C
T (BamHI)	FOR 13001 REV 13403	5' – GCAATTCAGCCCATTTAGGT – 3' 5' – ATATCTTGTTCATTGTTAAG – 3'	47°C
W (-8994 HaeIII	FOR 8908 REV 9033	5' – TTCTTACCACAAGGCACACC – 3' 5' – AGGTGGCCTGCAGTAATGT – 3'	65°C
X/I (-1715 DdeI)	FOR 1616 REV 1899	5' -- ACACAAAGCACCCAACTTACACTTAGGA – 3' 5' – CTTAGCTTTGGCTCTCCTTGC – 3'	59°C
X/I (± 10394 DdeI)	FOR 10235 REV 10569	5' – TATTACCTTCTTATTATTTTG – 3' 5' – CTAGGCATAGTAGGGAGGAT – 3'	48.2°C
I (AvaII)	FOR 8191 REV 8312	5' – ACCCACAGTTTCATGCCCAT – 3' 5' – TAAGTTAGCTTTACAGTGGGCT – 3'	59°C
HVS-I	15976 FOR 16401 REV	5' – CCACCATTAGCACCCAAAGCTAAG – 3' 5' – TGATTTCACGGAGGATGGTG – 3'	55°C

conditions: preliminary denaturation for 30 seconds at 96°C; 25 cycles of denaturation for 10 seconds at 96°C, annealing for five seconds at 50°C, extension for four minutes at 55°C; and a final hold at 4°C. The samples were cleaned in gel purification columns, and then dried by speed-vac. The samples were then treated with 20µl ABI template suppression buffer, heated for three minutes at 95°C, and snap-cooled on ice. Finally, the samples were placed in ABI tubes and loaded into the ABI sequencer, with the resulting chromatograms recorded by computer.

Mitochondrial sequences were edited using BioEdit software (Hall 2007) and aligned to the published human mtDNA reference sequence (Anderson *et al.* 1981). Nucleotides which differed from the reference sequence were recorded as mutations.

Analytical Methods

Genetic Diversity and Population Substructure

Hardy-Weinberg equilibrium (HWE) is a measure used to determine if a particular sample approximates being panmictic and does not demonstrate effects of evolutionary forces. For biallelic Mendelian systems, estimating HWE is a straightforward process. Hardy-Weinberg equilibrium is expressed as:

$$p^2 + 2pq + q^2 = 1 \qquad (1)$$

where p^2 is the frequency of the homozygous dominant genotype, $2pq$ is the genotype frequency of heterozygotes, and q^2 is the frequency of the homozygous recessive genotype. For systems with multiple alleles, such as STRs, expected heterozygosity

under Hardy-Weinberg equilibrium is estimated using a Markov chain constructed so that the probabilities of the genotype distribution expected under HWE match the allele counts found in the observed data (Guo and Thompson 1992). This estimate, analogous to Fisher's exact test, was computed using Arlequin 3.11 (Excoffier *et al.* 2005).

Wright's F-statistics, including F_{ST}, measure departures from HWE that result from population substructure, which causes a deficiency of heterozygotes in the total population. F_{ST} is defined as:

$$F_{ST} = (H_T - H_S) / H_T \quad (2)$$

where H_T is the expected heterozygosity in the total population, and H_S is the average expected heterozygosity among subpopulations (Wright 1951).

Like F_{ST}, the coefficient of gene differentiation (G_{ST}) measures genetic variation in substructured populations so that:

$$G_{ST} = D_{ST}/H_T \quad (3)$$

where D_{ST} is the gene differentiation among subpopulations (equivalent to $H_T - H_S$), and H_T is the gene diversity in the total population. Gene differentiation and gene diversity values for the Basques and comparative populations were calculated based on autosomal STR data using DISPAN (Ota 1993).

Gene diversity is a measure of the amount of genetic variation present in a population, and is another method used to test for equilibrium. It is based on haplotypic data (Y-STRs in the present study) and is equivalent to the more familiar heterozygosity measure for diploid data (Nei 1987). Gene diversity is defined as:

$$\hat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right) \tag{4}$$

where n is the number of genes in the sample, k equals the number of haplotypes, and p_i is the frequency of haplotype i in the sample.

An equivalent measure for mtDNA sequence data are nucleotide diversity, which is calculated as:

$$\pi = \sum_{ij}^{q} x_i x_j \pi_{ij} \tag{5}$$

where π is the average number of nucleotide differences between two sequences, x_i is the frequency of sequence i in the population, and π_{ij} is the number of nucleotide differences between sequences i and j at a given site (Nei and Li 1979). This measure was calculated for HVS-I sequence data using Arlequin 3.11 (Excoffier *et al.* 2005).

Tajima's D and Fu's F_S test for the effect of evolutionary forces in populations. These measures, based on the infinite allele model (IAM) with no recombination, use the parameter θ, or the average number of pairwise differences (mutations) between sequences in a sample. θ is calculated as:

$$\theta = 2N\mu \tag{6}$$

for haploid data, where N is the effective population size and μ is the mutation rate per sequence per generation.

Tajima's D, which assumes that the population is at equilibrium and that the sequences represent a random sample, compares the average number of nucleotide

differences $\left(\hat{\theta}_{\pi}\right)$ to the number of segregating sites $\left(\hat{\theta}_{S}\right)$ in the DNA sequence (Tajima 1989). *D* is defined as:

$$D = \frac{\hat{\theta}_{\pi} - \hat{\theta}_{S}}{\sqrt{Var\left(\hat{\theta}_{\pi} - \hat{\theta}_{S}\right)}} \tag{7}$$

A large number of segregating sites $\left(\hat{\theta}_{S}\right)$ relative to nucleotide differences $\left(\hat{\theta}_{\pi}\right)$ result from an increase in low frequency mutations, reflected in negative values of *D*. These indicate a population that has undergone an expansion event. Positive values of *D* result from a higher number of nucleotide differences relative to segregating sites and thus a greater number of high frequency mutations, suggesting a genetic bottleneck. Significant values of *D* could also result from selection or variations in mutation rate.

Fu's F_S tests specifically for population expansion and the effects of selection on surrounding neutral alleles through genetic hitchhiking (when selection is positive) or background selection (when deleterious alleles are being eliminated). Similar to Tajima's *D*, Fu's F_S measures the number of recent mutations in a population and is expressed as:

$$F_S = \ln\left(\frac{S'}{1 - S'}\right) \tag{8}$$

where $S' = p\left(k_0 \leq k \mid \theta = \hat{\theta}_{\pi}\right)$, or the probability of having k_0 alleles in a random sample when $\theta = \pi$ (Fu 1997). Large negative values of F_S indicate an excess of mutations from what would be expected under the neutral mutation theory, and

suggest that either selection or expansion has occurred in the population. Both Tajima's D and Fu's F_S were calculated for mitochondrial control region sequence data using Arlequin 3.11 (Excoffier *et al.* 2005).

Genetic Distance Measures

Genetic distances are employed to measure the relationship between populations and can be useful in reconstructing phylogenetic history. Various genetic distance measures have been developed to suit the evolutionary mechanisms of the molecular systems being studied. Classic genetic markers have few alleles, have a low mutation rate, and evolve according to an infinite allele model. According to this model, a gene 100 base pairs in length could mutate at any of those positions into one of the four possible nitrogenous bases, creating the potential for one billion (100^4) new alleles. STR loci, by contrast, have many alleles, have a comparatively high mutation rate, and evolve according to a stepwise mutation model (SMM). Unlike the infinite allele model, the SMM reduces the number of potential new alleles to those that are a few repeats larger or smaller than the current allele. Therefore, an allele of 14 repeats could mutate to one with 13 or 15 repeats, but not one with 35.

For classical genetic markers, Nei's D is the standard distance measure, and is defined as:

$$D = -\ln\left(\frac{\sum x_i y_i}{\sum x_i^2 \left(y_i^2\right)^{\frac{1}{2}}}\right) \tag{9}$$

where $\Sigma x_i y_i$ is the probability of two alleles being identical in two different populations, and Σx_i^2 and Σy_i^2 are the probabilities of two alleles being identical in population x and population y, respectively (Nei 1972).

Shriver's distance (D_{SW}) is an adaptation of Nei's standard distance (D) for STR loci, weighted by the difference in number of repeats between alleles, to account for the stepwise mutation pattern of tandem repeat loci, so that:

$$D_{SW} = d_{XYW} - \left(d_{XW} + d_{YW}\right)/2 \tag{10}$$

where d_{XYW} equals $\Sigma x_i y_i$ weighted by the absolute difference in repeat number between alleles averaged over all loci between populations, while d_{xw} and d_{YW} are the weighted probabilities within populations x and y (Shriver *et al.* 1995). This distance is appropriate for both autosomal (diploid) and Y-chromosome (haploid) STR data.

For sequence data, a variety of methods are available to measure genetic differences between nucleotide sequences. The simplest is merely a comparison of the number of pairwise differences, θ (Equation 6). However, this measure does not account for mutation rate heterogeneity within a particular sequence, and thus could lead to an underestimate of the true difference between sequences. In the mutational process, transitions, where a purine is substituted by a pyrimidine or vice versa (*e.g.,* C-A), are more common than transversions, where a base is substituted by one of a similar chemical structure (C-T). The Kimura-2-parameter distance addresses this issue, by correcting for differences in mutation rates between transversions and transitions, so that:

$$\hat{d} = \frac{1}{2}\log\left(1 - 2\hat{P} - \hat{Q}\right) - \frac{1}{4}\log\left(1 - 2\hat{Q}\right) \tag{11}$$

where $\hat{P}$ equals the frequency of transition sites in the sequence, and $\hat{Q}$ is the frequency of transversions (Kimura 1980).

Heterozygosity and distance from centroid

To study the effects of migration and genetic drift on local population structure, Harpending and Ward (1982) devised a method of regressing the mean per locus heterozygosity on the distance from the centroid (r_{ii}), the distance from the overall mean gene frequencies of the region being studied. Mean per locus heterozygosity, d_i, is calculated as 1 minus the sum of homozygous gene frequencies (p^2_k) (Equation 12).

$$d_i = 1 - \sum_{k=i}^{t} p_k^2 \tag{12}$$

According to this model, mean per locus heterozygosity should decrease with increasing genetic distance because regional gene frequency distributions are a function of local population structure, and are therefore much more affected by drift and migration than by selection and mutation. The latter two evolutionary forces are more likely to affect populations separated by large geographic distances, not those located in the same geographical region. Populations that fall below the theoretical regression line -- defined as mean heterozygosity ($1-r_{ii}$) -- experience significantly more drift (i.e., are more genetically isolated) than those that fall close to the line. Populations that have experienced significant gene flow are found well above the

theoretical regression line. These values were calculated for classical genetic marker data using ANTANA (Harpending and Rogers 1984).

Analysis of Molecular Variance (AMOVA)

Analysis of Molecular Variance (AMOVA) is comparable to a standard analysis of variance (ANOVA), except that the hierarchical analysis of variance in AMOVA is based on squared Euclidean distances between haplotypes (loci found on the same chromosome), such that:

$$SSD(T) = 1/4N \sum_{j=1}^{2N} \sum_{k=1}^{2N} \delta_{jk}^2 \tag{13}$$

where *SSD(T)* is the total sum of squared deviations between populations, *N* is the number of haplotypes, and δ_{jk}^2 is the squared difference between haplotypes *j* and *k*. In this procedure, individuals are assigned to groups (populations) based on certain criteria, such as shared language or geographic proximity. Because this method was developed for haplotypic data, or alleles that reside on the same chromosome, it is only appropriate for analysis of the uniparental markers. AMOVA is used to determine the sources of variation resulting from deviations from Hardy-Weinberg equilibrium within populations, within groups (i.e., language families or countries), and between populations within groups (Excoffier *et al.* 1992). This variation is described as Φ-statistics, analogous to Wright's F-statistics, which measure population subdivision. In AMOVA, three statistics are of interest: Φ_{ST}, Φ_{SC}, and Φ_{CT}. Φ_{ST} is defined as the amount of variation contained in a subpopulation relative to the total. Φ_{SC} measures the total genetic variation present in subpopulations

between groups, and Φ_{CT} describes the variation found within groups relative to the total variation. Significance values of covariance components and fixation indices, which are computed by a permutation test of 16,000 permutated distance matrices, measure the likelihood of having values greater those observed by chance (Excoffier *et al.* 1992).

To examine genetic substructure among the Basque Provinces using diploid autosomal data, genotypes for the nine autosomal STR loci of each individual were grouped by village within each of the four provinces. As autosomal STR loci are unlinked (i.e., located on different chromosomes), modifications to the standard AMOVA procedure were required: (1) the locus-by-locus AMOVA method was used, so that variance components and F-statistics were calculated for each locus individually, followed by estimations of variance components and fixation indices at each level of the hierarchy; (2) distances were calculated using the sum of squared differences between repeats at each locus, an approximation of R_{ST} (analogous to F_{ST}) which accounts for the stepwise mutation process of microsatellites. R_{ST} is calculated as:

$$R_{ST} = \frac{\bar{S} - S_W}{\bar{S}} \tag{14}$$

where $\bar{S}$ is the average squared size difference between all pairs of alleles at a given locus, and S_W is equal to the average sum of squared allele size differences within each population (Slatkin 1995). Population subdivision is measured within populations (F_{ST}), among populations within groups (F_{SC}), and among groups (F_{CT}).

AMOVA analysis was performed for all three molecular systems available in the present study using Arlequin 3.11 (Excoffier *et al.* 2005).

Spatial Analysis of Molecular Variance (SAMOVA)

Spatial analysis of molecular variance is another method of detecting variation within and between populations, but unlike AMOVA, group hierarchies are not determined *a priori*. Instead, the number of groups in the hierarchy (K) is chosen, and the SAMOVA algorithm identifies which grouping of populations maximizes F_{CT} by identifying genetic barriers between groups (Dupanloup *et al.* 2002). Using data from non-recombining genetic systems (mtDNA sequences and Y-STR haplotypes) and the geographical coordinates of the populations, the SAMOVA procedure first generates a series of Voronoi polygons from the geographical data, randomly partitions the populations into the specified K groups, and then computes the F_{CT} index. The procedure is repeated until the maximum value of F_{CT} for the designated number of groups is identified. Several runs are required to determine which value of K best fits the data, and therefore provides the highest F_{CT}, and then those genetic barriers can be overlaid on multidimensional scaling plots (see next section) to highlight areas of maximum genetic variation. SAMOVA analysis was performed for Y-STR haplotype data and mtDNA HVS-I sequence data using SAMOVA 1.0 (Dupanloup *et al.* 2002).

Network Analysis

Median-joining phylogenetic network analysis was employed to visualize relationships between haplotypes within the most common mitochondrial and Y-

chromosome haplogroups present in the Basque sample. This type of phylogenetic analysis generates a network of the most parsimonious trees, while at the same time identifying ambiguous states between haplotypes in the form of reticulations (Bandelt *et al*. 1999) These networks can then be examined for evidence of population expansion, in the form of an overall star-like structure. Networks can also be dated, by identifying ancestral and descendant nodes and then estimating ρ, which is the average distance from the derived haplotype to the ancestral node. For the mitochondrial data, a network for haplogroup H was constructed using sequences from Alava, Vizcaya and Guipuzkoa. A network of R1b haplotypes in the four Basque Provinces of Spain was generated from the Y-chromosome data. Given the large number of Y-STR haplotypes, these data were preprocessed using the star-contraction algorithm to create a skeleton network (Forster *et al.* 2001). After the median-joining network was generated, it was then post-processed using the maximum parsimony algorithm to eliminate unnecessary median vectors and simplify the network (Polzin and Daneschmand 2003). All networks were constructed using Network 4.510 (www.fluxus-engineering.com).

Mismatch/Intermatch Analysis

Mitochondrial HVS-I sequence data were used to generate mismatch/intermatch distributions to further examine population histories and estimate dates of expansion (Rogers and Harpending 1992). Mismatch analysis produces distributions of pairwise differences within populations which, according to coalescent theory, maintain a record of past population events including demographic

and spatial expansions, which leave unimodal waves in the distribution, or long periods of population stability, which result in multimodal distributions (Excoffier 2004; Hudson 1990; Rogers and Harpending 1992). Intermatch analysis is used to examine pairwise differences between populations, and to estimate divergence times between them (Sherry *et al.* 1994). Under the neutral mutation model with no recombination, the mean number of mutations is constant, and mutations accrue along lines of descent independent of population size or substructure. The number of pairwise differences between two sequences increases at a rate of $2u$, where u is equal to the mutation rate across the entire DNA sequence under study. In the present study, the HVS-I region being examined includes 276 nucleotides (m_T), so $u = m_T\mu$, or 276*16.5%/million years (Ward *et al.* 1991). Doubling this value gives the divergence rate ($2u$), and the date of population expansion can be estimated in units of mutational time using tau (τ), such that:

$$\tau = 2ut \tag{15}$$

where t is generation time. If τ is known, expansion time in years can be easily estimated by:

$$t = \tau/2u \tag{16}$$

For the present analysis, τ was estimated in two ways. First, τ was estimated using a method of moments (Rogers 1995), so that τ is expressed as:

$$\hat{\tau} = m - \sqrt{v - m} \tag{17}$$

where m is the mean of the observed mismatch distribution, while v is equal to the variance (mathematically, the first and second moments of the distribution). Second, τ was estimated using a generalized non-linear least-squares approach (Schneider and Excoffier 1999), so that the sum of squared deviations (SSD) between an observed mismatch distribution (F_{iobs}) and its expectation (F_{iexp}) are minimized, defined as:

$$SSD = \sum_{i=0}^{n} \left(F_{iobs} - F_{i\exp} \right)^2 \tag{18}$$

In the second case, 95% confidence intervals for τ are also generated using a parametric bootstrap approach. These confidence intervals can be converted into date ranges of the time estimate for population expansion using Equation (16).

The least-squares mismatch analysis was conducted to examine potential differences in population history between the Basques and other European groups using Arlequin 3.11 (Excoffier *et al.* 2005). Intermatch distributions between the Basques and other European groups, along with estimated divergence times, were generated using iWave (Sherry *et al.* 1994).

Ordination and Visualization Techniques

R-Matrix

R-Matrix analysis examines population structure and population history using principal components analysis of genetic distances based on normalized gene frequencies (Harpending and Jenkins 1973). A matrix of distances between

populations is calculated based on each allele, using the sample coefficient of kinship (R_{ij}) for each allele such that

$$R_{ij} = (p_i - \overline{p})(p_j - \overline{p}) / \overline{p}(1 - \overline{p}) \tag{19}$$

where p_i and p_j are frequencies of allele p in populations i and j, and $\overline{p}$ is the weighted mean gene frequency. Dividing the product of the difference of the gene frequencies from the mean by the mean times its complement normalizes the data and provides a statistic that is independent of the mean (Harpending and Ward 1982). These allele distance matrices are averaged into the overall R-matrix, a covariance matrix of genetic distances between populations which is weighted by effective size (not sample size) to examine population structure. The R-matrix is then used in a principal components analysis to provide a genetic map of relationships between groups. Principle component scores are scaled by dividing them by the square root of the eigenvalue to eliminate distortion of the true relationships between groups.

Additionally, this analysis also provides an S-matrix of the spread of alleles, which is generated by plotting the coefficients of the principal components. This is useful in identifying which alleles are contributing to population dispersal. R-Matrix analysis was performed using data on classical markers from the Basques and various comparative populations using ANTANA (Harpending and Rogers 1984).

Multidimensional Scaling

Multidimensional scaling (MDS) is a method used to create a map of population relationships from genetic distances, or more generally, to map the dissimilarity between objects (Manly, 1986). Unlike R-Matrix analysis,

multidimensional scaling allows the inclusion of all alleles at each locus, including those which may be present at low frequencies in only a few populations. It does not, however, provide information about which alleles affect population dispersal.

MDS does allow measurement of how well the "map" fits the actual data, through a variable known as stress, which is the normalized residual variance of the sample. The map which best fits the actual data will have the least stress, and a stress value less than 0.20 is considered an excellent fit (Kruskal, 1964). Recognizing that the stress value can be impacted by how many objects are being fit into a particular number of dimensions, Sturrock and Rocha (2000) developed a probability distribution for evaluating the significance of the stress variable given the number of objects projected in one, two or three dimensions. Using this table, a stress value of 0.349 would be an excellent fit ($p = 0.05$) for genetic distances of 39 populations plotted in two dimensions.

Neighbor-Joining

Neighbor-Joining (NJ) is a method developed to display genetic distances in a phylogenetic framework. The NJ tree is constructed using a least squares approach to minimize the sum of all branch lengths under the minimum evolution (ME) criterion, so that:

$$\min \sum_{1 \le i \le j \le n} \left(t_{ij} - d_{ij} \right)^2 \tag{20}$$

where t_{ij} is the sum of branch lengths connecting taxa i and j, and d_{ij} is the genetic distance between taxa i and j, is achieved (Felsenstein 2004). The algorithm begins

with a star phylogeny, assuming no relationship between groups, and then build nodes of neighbors with the smallest branch lengths (*i.e.*, those with the lowest genetic distance). As neighbors are found, new nodes are created and the algorithm runs iteratively until all neighbors have been determined so that the total number of nodes equals N - 2. Neighbor-Joining is appropriate for use will all types of distance matrices, so NJ trees were constructed using autosomal STR data (DISPAN), Y-STR haplotypes (MEGA 4.1), and mitochondrial control region sequence data (MEGA 4.1) (Excoffier *et al.* 2005; Ota 1993; Tamura *et al.* 2007).

To test how well an NJ tree represents the genetic variation present in the distance matrix on which it is based, an R^2 value can be calculated such that:

$$R^2 = 1 - \frac{\sum\left(\hat{D}_{ij} - \hat{d}_{ij}\right)^2}{\sum\left(\hat{D}_{ij} - \bar{\hat{D}}\right)^2} \tag{21}$$

where $\hat{D}_{ij}$ is the estimated genetic distance between populations i and j in the genetic distance matrix and $\hat{d}_{ij}$ is the observed genetic distance between populations i and j in the Neighbor-Joining tree (Kalinowski 2009). If the observed distances closely approximate the estimated distances, R^2 will have a value near 1.0, and the tree is a reasonable depiction of the relationships between groups. If $R^2 < 0.9$, the tree is not a good fit, and thus is not the best representation of population structure.

Interpolated Genetic Landscapes

Mitochondrial HVS-I sequence data and Y-STR haplotype data from the European populations were used in the construction of interpolated genetic distance landscapes (Miller 2006). In this procedure, a Delaunay triangulation connectivity network is constructed between sample sites based on geographic coordinates and pairwise nucleotide differences between individuals. Genetic distances between sample sites are then inferred and plotted on a uniform grid of the sample space using an inverse-weighted distance interpolation method. The x and y axes of the grid represent geographic coordinates, and the z axis plots the genetic differences between populations. Genetic similarities are shown as dips below the x/y plane, while genetic differences appear as peaks above the x/y plane. The pairwise genetic distances, z, are calculated as:

$$z = \frac{\sum_{i=1}^{n} w_i \times z_i}{\sum_{i=1}^{n} w_i} \tag{22}$$

where w_i is a weighting function given to each z_i which is inversely proportional to the difference between the sample location in the network and the actual geographic location of the sample. Interpolated genetic landscapes were construction using Alleles in Space (Miller 2005).

CHAPTER FOUR: RESULTS

The genetic origins of the Basques have been examined using classical markers, autosomal STRs, Y-STR haplotypes, and mitochondrial DNA. Both intra- and interpopulation analyses were conducted, the first to examine the question of Basque homogeneity (i.e., whether they can be considered a single population), and the second to examine the various hypotheses (Basque-Caucasian, Vasco-Iberian, and pre-Indo-European) put forward concerning Basque origins. In this chapter, data on gene and haplotype frequencies, along with measures of diversity and neutrality, are presented, followed by analysis of molecular variance within and between Basque provinces, and finally the results of interpopulation analysis including AMOVA (by geographic region and language family) and various phylogenetic methods (R-Matrix, MDS, Neighbor-Joining) plus SAMOVA are described.

Intrapopulation Analyses

Classical Markers

Allele frequencies for several classical genetic markers in the four Spanish Basque Provinces and the French Province of Labourd are presented in Table 10. These allele frequencies, along with comparable data from other European, North African and Caucasian populations, were collected from the literature for analysis using R-Matrix, MDS, and heterozygosity vs. r_{ii}. Alleles for which Basque populations are outliers in the European genetic landscape are highlighted in italics,

including ABO*B (0.0133-0.0394), RH*cde (0.4572-0.5336), MNS*MS (0.2740-0.3329), AK*1 (0.9350-0.9703), GM*Z,A;B,S,T (0.0080-0.0310).

Table 10. Frequencies of classical genetic markers among the Basques.

Allele	***Alava***	***Guipuzkoa***	***Navarre***	***Vizcaya***	***Labourd***
N	122-480	97-586	47-287	47-1116	63-384
Blood Groups					
ABO*A	0.2589	0.2297	0.2247		0.2090
*ABO*B*	0.0394	0.0233	0.0315		0.0133
ABO*O	0.7017	0.7470	0.7438		0.7777
RH*CDE (Rz)	0.0002	0.0395			
RH*CDe (R1)	0.3697	0.3580			0.3697
RH*cDE (R2)	0.1238	0.0313			0.1082
RH*cDe (Ro)	0.0341	0.0117			0.0200
RH*Cde (r')	0.0110	0.0211			0.0000
RH*cdE (r”)	0.0041	0.0048			0.0000
*RH*cde (r)*	0.4572	0.5336			0.5021
*MNS*MS*	0.2740	0.2750			0.3329
MNS*Ms	0.2708	0.2847			0.3104
MNS*NS	0.0896	0.0843			0.1105
MNS*Ns	0.3656	0.3560			0.2562
Fy*a		0.3505			0.3107
Fy*b		0.6495			0.6893
Jk*a					0.5713
Jk*b					0.4287
P1 (P-)	0.5148	0.5147			0.5430
P2 (P+)	0.4852	0.4853			0.4570
Le*(a+b-)	0.7077	0.6495			
Le*(a-b+)	0.2923	0.3505			
K*Kk		0.0299			0.0264
K*kk		0.9701			0.9736
Red Cell Enzymes					
GLO*1	0.4459			0.4570	
GLO*2	0.5541			0.5430	
PGD*A	0.9822		0.9878	0.9910	0.9966
PGD*C	0.0178		0.0122	0.0090	0.0034
ADA*1	0.9852	0.9760	0.9716	0.9800	
ADA*2	0.0148	0.0240	0.0284	0.0200	
ACP*A	0.2904	0.2450	0.2569	0.2750	0.2218
ACP*B	0.6832	0.7320	0.7170	0.7170	0.7676
ACP*C	0.0264	0.0230	0.0261	0.0070	0.0105

Allele	*Alava*	*Guipuzkoa*	*Navarre*	*Vizcaya*	*Labourd*
*AK*1*		0.9350	0.9703	0.9540	0.9652
AK*2		0.0650	0.0030	0.0460	0.0348
PGM1*1		0.6930	0.7617	0.7617	0.6980
PGM1*2		0.3070	0.2383	0.2383	0.3020
ESD*1	0.8935	0.9320	0.8850	0.9327	
ESD*2	0.0880	0.0560	0.1150	0.0584	
ESD*5	0.0185	0.0120		0.0089	
Plasma Proteins					
TF*C1	0.7932	0.8100			
TF*C2	0.1713	0.1230			
TF*C3	0.0324	0.0590			
TF*C6		0.0030			
TF*B	0.0031	0.0030	0.0174		
CP*A			0.0046		
CP*B			0.9954		
HP*1		0.3870	0.4416	0.4420	0.3810
HP*2		0.6130	0.5584	0.5770	0.6191
GC*1	0.6343	0.6570	0.6871	0.6630	0.6429
GC*2	0.3657	0.3430	0.3129	0.3370	0.3571
PI*M1	0.6115	0.6240	0.8767	0.7060	
PI*M2	0.1641	0.1890		0.1620	
PI*M3	0.1006	0.0790		0.0240	
PI*M4	0.0186				
PI*S	0.1006	0.1020	0.1164	0.0990	
PI*T	0.0031				
PI*Z	0.0015	0.0050	0.0069	0.0080	
GM *z,a,g1 (1,17,21)	0.0492	0.1443		0.1176	0.2339
GM*z,a,x,g1 (1,2,17,21)	0.0660	0.0000		0.0350	0.0994
GM*b1 (3,5)	0.4098	0.3402		0.3678	0.6667
GM*z,a;b,g	0.3525	0.4021		0.3176	
GM*z,a,x;b,g	0.0902	0.0410		0.0941	
GM*z,a;b	0.0246	0.0309		0.0471	
*GM*z,a;b,s,t*	0.0080	0.0310		0.0120	
KM*1+	0.1950	0.0800		0.1260	
KM*1-	0.8050	0.9200		0.8740	
Reference	1-3	2,4	4-6	2,4-10	11-12

[1]Manzano (1993b), [2]Calderon (1998), [3]Manzano (1996b),[4]Manzano (1996a), [5-6]Goedde (1972b; 1973), [7,10]Aguirre (1989b; 1991b), [8]Garcia-Orad (1990), [9]de Pancorbo (1989), [11]Vergnes (1980), [12]Levine (1974). Italics indicate outlier alleles.

Autosomal STRs

Examination of the autosomal STR data revealed that allelic dropout occurred in several samples, so that not all loci were amplified for every individual. At low sample DNA concentrations, the Profiler Kit preferentially amplifies short alleles and homozygotes (Pawlowski and Maciejewska 2000). Because the samples collected in this study were from buccal swabs, and only a portion of each sample was used for STR analysis, DNA concentrations were much lower than if the samples had been from whole blood. Therefore, samples which did not amplify for all loci were removed from the analysis. Table 11 shows the allele frequencies for the autosomal STR loci in each of the Basque Provinces after correction for allelic dropout.

Table 11. Basque Autosomal STR Frequencies and Exact Test of Hardy-Weinberg Equilibrium by Province.

Locus	Alava	Vizcaya	Guipuzkoa	Navarre
D3S1358	N = 96	N = 89	N = 154	N = 38
11	0.0052	0.0000	0.0000	0.0000
12	0.0052	0.0056	0.0097	0.0132
13	0.0208	0.0056	0.0065	0.0000
14	0.0833	0.1067	0.1429	0.0921
15	0.2969	0.2921	0.3442	0.2632
15.2	0.0052	0.0000	0.0000	0.0000
16	0.1875	0.3146	0.1786	0.2632
17	0.1771	0.1348	0.1071	0.1447
18	0.1927	0.1348	0.2045	0.2105
19	0.0260	0.0056	0.0065	0.0132
FGA				
17	0.0000	0.0112	0.0000	0.0000
18	0.0208	0.0449	0.0617	0.0395
19	0.0729	0.1292	0.1201	0.0921
20	0.1510	0.1461	0.1299	0.0658
20.2	0.0000	0.0000	0.0000	0.0132
21	0.2188	0.1292	0.1558	0.1842
21.2	0.0052	0.0000	0.0000	0.0000
22	0.1667	0.1348	0.0844	0.1447
22.2	0.000	0.0056	0.0032	0.0000
23	0.1094	0.1517	0.1916	0.1711
23.2	0.0052	0.0000	0.0000	0.0000

Locus	Alava	Vizcaya	Guipuzkoa	Navarre
24	0.1458	0.1180	0.1234	0.1447
25	0.0781	0.0674	0.0942	0.1053
26	0.0208	0.0506	0.0195	0.0395
27	0.0000	0.0000	0.0065	0.0000
28	0.0052	0.0112	0.0065	0.0000
29	0.0000	0.0000	0.0032	0.0000
D5S818				
8	0.0052	0.0056	0.0032	0.0000
9	0.0208	0.0393	0.0422	0.0132
10	0.0885	0.0955	0.1104	0.0658
11	0.3854	0.3876	0.3442	0.2895
12	0.3490	0.2978	0.3052	0.4079
13	0.1406	0.1629	0.1883	0.2237
14	0.0052	0.0112	0.0065	0.0000
15	0.0052	0.0000	0.0000	0.0000
D7S820				
7	0.0521	0.0169	0.0130	0.0526
8	0.0938	0.1966	0.1623	0.1974
9	0.1198	0.1236	0.1039	0.0789
10	0.3385	0.2191	0.2857	0.2500
11	0.1823	0.2640	0.2500	0.1711
12	0.1354	0.1180	0.1364	0.2105
13	0.0521	0.0562	0.0292	0.0263
14	0.0260	0.0056	0.0195	0.0132
D8S1179				
8	0.0208	0.0169	0.0162	0.0000
9	0.0260	0.0169	0.0195	0.0000
10	0.0833	0.0899	0.0747	0.1053
11	0.0469	0.0281	0.0422	0.0658
12	0.0885	0.0787	0.1234	0.1184
13	0.2656	0.2809	0.3052	0.3026
14	0.3333	0.3258	0.2500	0.2105
15	0.1198	0.1517	0.1494	0.1711
16	0.0104	0.0056	0.0195	0.0263
17	0.0052	0.0056	0.0000	0.0000
vWA				
12	0.0000	0.0000	0.0000	0.0132
13	0.0000	0.0000	0.0000	0.0132
14	0.1458	0.1517	0.1136	0.0921
15	0.1563	0.1180	0.1331	0.1974
16	0.2344	0.2472	0.1883	0.1974
17	0.2708	0.2416	0.3344	0.2632
18	0.1458	0.1629	0.1331	0.1579
19	0.0469	0.0730	0.0909	0.0658
20	0.0000	0.0000	0.0065	0.0000
21	0.0000	0.0056	0.0000	0.0000

Locus	Alava	Vizcaya	Guipuzkoa	Navarre
D13S317				
8	0.1875	0.1910	0.2175	0.1842
9	0.0365	0.0449	0.0455	0.0000
10	0.0365	0.0787	0.0227	0.0395
11	0.2604	0.2753	0.3247	0.3816
12	0.3646	0.2921	0.2500	0.2500
13	0.0781	0.0674	0.1039	0.1053
14	0.0365	0.0449	0.0325	0.0263
15	0.0000	0.0056	0.0032	0.0132
D18S51				
10	0.0208	0.0225	0.0130	0.0132
11	0.0156	0.0112	0.0227	0.0132
12	0.2188	0.1742	0.1786	0.1053
13	0.0990	0.0899	0.1558	0.1053
14	0.1198	0.1742	0.1396	0.1579
15	0.1406	0.1742	0.1623	0.1053
15.2	0.0000	0.0000	0.0032	0.0000
16	0.1146	0.1292	0.1136	0.1711
17	0.1615	0.0899	0.0877	0.2105
18	0.0469	0.0225	0.0422	0.0132
19	0.0313	0.0506	0.0130	0.0921
20	0.0156	0.0337	0.0390	0.0132
21	0.0052	0.0112	0.0195	0.0000
22	0.0052	0.0056	0.0097	0.0000
24	0.0052	0.0112	0.0000	0.0000
D21SS11				
26	0.0000	0.0056	0.0032	0.0000
27	0.0260	0.0393	0.0162	0.0395
28	0.1042	0.0787	0.0812	0.0921
29	0.1719	0.1685	0.2045	0.1579
29.2	0.0000	0.0056	0.0000	0.0000
30	0.2760	0.3146	0.2727	0.3289
30.2	0.0208	0.0843	0.0357	0.0263
31	0.0469	0.0506	0.0779	0.0132
31.2	0.0885	0.0843	0.0779	0.1316
32.0	0.0000	0.0000	0.0097	0.0000
32.2	0.1719	0.0843	0.1429	0.1711
33.2	0.0833	0.0730	0.0747	0.0395
34.2	0.0104	0.0056	0.0000	0.0000
35.2	0.0000	0.0056	0.0032	0.0000

Y-STR Haplotypes

Frequencies of the 89 Y-STR haplotypes found in the four Basque Provinces

are presented in Table 12. While several haplotypes are shared between provinces (H1 is found in Alava, Guipuzkoa, and Navarre), 92% of the haplotypes are unique to a single province, meaning that they represent distinct lineages. Y haplogroups were determined using a likelihood algorithm based on Y-STR haplotypes (Athey 2005; Athey 2006). Haplotype frequencies, haplogroup assignments, as well as fitness and probability of those assignments are shown in Table 13. All but three haplotypes were assigned to a single haplogroup with greater than 95% probability. Seventy-four of the eighty-nine haplotypes (83%) are R1b, the most frequent Western European haplogroup, especially along the Atlantic Fringe (Rootsi *et al.* 2004). Five other haplogroups were identified in the Basque sample: E1b1b (6%), J2a (3%), I2 (3%), G2a (2%), and L (1%). One haplotype, H57 (15-12-15-24-10-14-14-13-15-9-12), could not be unambiguously associated with any particular haplogroup.

Table 12. Frequencies of Y-STR Haplotypes in Four Spanish Basque Provinces

Haplotype	*Haplotype Definition*[a]	*Alava (32)*	*Vizcaya (25)*	*Guipuzkoa (58)*	*Navarre (11)*
H1	14-14-16-24-11-13-13-11-14-12-11	3		3	1
H2	14-14-16-23-9-11-13-12-17-10-12	3			
H3	13-12-17-24-10-11-12-17-18-10-11	3			
H4	14-13-17-24-11-13-13-12-14-12-11	2			
H5	14-13-16-24-11-13-13-11-14-12-11	2	1	1	
H6	14-12-16-24-11-13-13-11-14-12-12	2			
H7	15-13-17-23-11-13-13-11-16-12-12	1			
H8	15-13-16-24-12-13-13-11-15-12-11	1			
H9	14-14-17-23-10-13-13-11-11-12-13	1			
H10	14-14-16-24-11-13-13-12-14-12-11	1			
H11	14-14-16-24-11-13-13-11-15-12-12	1			
H12	14-14-16-24-11-13-13-11-14-12-12	1			1
H13	14-14-16-24-11-13-13-11-13-12-12	1			
H14	14-14-16-24-10-13-13-11-14-12-12	1			
H15	14-13-16-25-12-13-13-11-14-12-12	1			
H16	14-13-16-25-11-13-13-11-14-12-11	1			
H17	14-13-16-24-11-13-13-12-14-12-11	1		1	

Haplotype	*Haplotype Definition*[a]	*Alava (32)*	*Vizcaya (25)*	*Guipuzkoa (58)*	*Navarre (11)*
H18	14-13-16-24-11-13-13-11-14-12-13	1	2	1	
H19	14-13-16-24-11-13-13-11-14-12-12	1	1	8	
H20	14-13-16-24-10-13-13-11-15-12-12	1			
H21	14-13-15-24-11-13-13-11-14-12-14	1			
H22	14-12-17-24-11-13-13-11-14-12-11	1			
H23	14-12-16-24-10-13-13-11-14-12-12	1			
H24	14-14-16-23-10-13-13-11-11-12-13		2	1	
H25	16-13-17-23-9-12-12-13-13-9-12		1		
H26	15-13-16-24-9-11-12-13-15-10-10		1		
H27	15-13-16-24-11-13-13-12-14-12-11		1		
H28	15-13-16-24-11-13-13-11-15-12-12		1		
H29	14-14-16-24-11-13-13-11-15-12-11		1		
H30	14-14-16-24-10-13-13-11-14-12-11		1	1	
H31	14-14-16-23-10-13-13-11-11-12-14		1		
H32	14-14-15-25-11-13-13-11-14-12-12		1		
H33	14-13-17-24-11-13-13-11-15-12-12		1		
H34	14-13-17-23-11-13-13-11-15-12-13		1		
H35	14-13-16-25-10-13-12-11-13-12-12		1		
H36	14-13-16-24-11-14-13-11-14-12-12		1		
H37	14-13-16-24-11-13-13-12-14-12-12		1		
H38	14-13-16-24-11-13-13-11-11-12-12		1		
H39	14-13-16-24-11-13-12-11-14-13-12		1		
H40	14-13-16-24-10-13-13-11-13-12-12		1		
H41	14-13-16-23-11-13-13-11-16-12-13		1		
H42	14-13-15-24-11-13-13-11-15-14-12		1		
H43	13-12-17-25-10-11-12-17-18-10-11		1		
H44	14-13-16-24-11-13-14-11-14-12-12			3	
H45	17-13-15-24-9-11-13-12-12-10-13			2	
H46	15-14-17-24-11-15-13-11-14-12-12			2	
H47	14-13-16-24-10-13-13-11-14-12-12			2	
H48	17-13-16-23-11-12-13-14-16-10-11			1	
H49	15-14-17-23-10-14-11-13-17-10-13			1	
H50	15-14-16-24-11-13-13-12-14-12-12			1	
H51	15-13-17-24-11-13-13-12-15-12-11			1	
H52	15-13-16-24-11-13-13-11-15-12-13			1	
H53	15-13-16-24-10-13-13-11-14-12-12			1	
H54	15-13-16-23-11-13-13-11-15-12-13			1	
H55	15-12-19-22-10-11-14-14-14-10-11			1	
H56	15-12-18-22-10-10-14-15-16-10-12			1	
H57	15-12-15-24-10-14-14-13-15-9-12			1	
H58	15-12-15-23-10-11-13-12-12-10-12			1	
H59	14-15-16-24-11-13-13-11-15-12-11			1	
H60	14-14-18-24-10-13-13-11-14-12-12			1	
H61	14-14-16-24-11-13-13-11-15-13-11			1	
H62	14-14-16-24-10-13-13-11-11-12-12			1	

Haplotype	*Haplotype Definition*[a]	*Alava (32)*	*Vizcaya (25)*	*Guipuzkoa (58)*	*Navarre (11)*
H63	14-13-17-25-10-13-13-11-14-12-12			1	
H64	14-13-17-24-11-13-13-10-14-12-12			1	
H65	14-13-16-25-11-13-13-11-14-12-12			1	
H66	14-13-16-25-10-13-12-11-13-12-13			1	
H67	14-13-16-24-11-13-14-12-14-12-12			1	
H68	14-13-16-24-11-13-13-11-15-12-12			1	
H69	14-13-16-24-11-13-13-11-14-11-11			1	
H70	14-13-16-24-11-12-13-11-14-12-12			1	
H71	14-13-16-24-10-13-13-12-14-12-11			1	
H72	14-13-16-23-11-13-13-11-14-12-11			1	
H73	14-13-16-23-10-13-13-11-11-12-13			1	
H74	14-13-15-24-11-13-13-12-14-12-12			1	
H75	14-13-15-23-11-13-13-11-14-12-12			1	
H76	14-12-16-24-11-13-13-11-14-12-11			1	
H77	14-12-16-23-11-13-14-11-14-12-13			1	
H78	14-11-16-25-11-13-13-11-13-12-11			1	
H79	13-12-18-25-11-11-13-17-18-10-11			1	
H80	13-12-18-25-10-11-13-17-18-10-11			1	
H81	15-14-17-24-10-14-13-11-14-12-12				1
H82	14-14-17-24-10-13-13-11-14-12-11				1
H83	14-14-17-24-10-11-13-17-19-10-14				1
H84	14-14-16-24-11-13-14-11-14-12-12				1
H85	14-14-16-24-11-13-13-12-14-12-13				1
H86	14-14-16-24-10-13-13-11-13-12-11				1
H87	14-14-15-24-11-13-13-12-14-12-12				1
H88	14-13-16-24-11-13-12-12-14-12-11				1
H89	13-14-16-24-9-11-13-13-14-10-10				1

[a]Haplotype definition locus order: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS438, DYS439.

Table 13. Basque haplogroup identification from haplotype definition, with goodness of fit and probability.

Haplotype	*Definition*[a]	*Frequency*	*Haplogroup*	*Fitness Value*	*Probability*
H1	14-14-16-24-11-13-13-11-14-12-11	7	R1b	77	100.0%
H2	14-14-16-23-9-11-13-12-17-10-12	3	J2a	97	95.0%
H3	13-12-17-24-10-11-12-17-18-10-11	3	E1b1b	44	100.0%
H4	14-13-17-24-11-13-13-12-14-12-11	2	R1b	62	100.0%
H5	14-13-16-24-11-13-13-11-14-12-11	4	R1b	91	100.0%
H6	14-12-16-24-11-13-13-11-14-12-12	2	R1b	78	100.0%
H7	15-13-17-23-11-13-13-11-16-12-12	1	R1b	49	99.9%
H8	15-13-16-24-12-13-13-11-15-12-11	1	R1b	51	100.0%
H9	14-14-17-23-10-13-13-11-11-12-13	1	R1b	38	100.0%
H10	14-14-16-24-11-13-13-12-14-12-11	1	R1b	61	100.0%
H11	14-14-16-24-11-13-13-11-15-12-12	1	R1b	75	100.0%
H12	14-14-16-24-11-13-13-11-14-12-12	2	R1b	85	100.0%
H13	14-14-16-24-11-13-13-11-13-12-12	1	R1b	73	100.0%
H14	14-14-16-24-10-13-13-11-14-12-12	1	R1b	78	100.0%
H15	14-13-16-25-12-13-13-11-14-12-12	1	R1b	69	100.0%
H16	14-13-16-25-11-13-13-11-14-12-11	1	R1b	81	100.0%
H17	14-13-16-24-11-13-13-12-14-12-11	2	R1b	72	100.0%
H18	14-13-16-24-11-13-13-11-14-12-13	4	R1b	88	100.0%
H19	14-13-16-24-11-13-13-11-14-12-12	10	R1b	100	100.0%
H20	14-13-16-24-10-13-13-11-15-12-12	1	R1b	81	100.0%
H21	14-13-15-24-11-13-13-11-14-12-14	1	R1b	59	100.0%
H22	14-12-17-24-11-13-13-11-14-12-11	1	R1b	61	100.0%
H23	14-12-16-24-10-13-13-11-14-12-12	1	R1b	71	100.0%
H24	14-14-16-23-10-13-13-11-11-12-13	3	R1b	44	100.0%
H25	16-13-17-23-9-12-12-13-13-9-12	1	J2a	46	99.5%
H26	15-13-16-24-9-11-12-13-15-10-10	1	J2a	100	96.1%
H27	15-13-16-24-11-13-13-12-14-12-11	1	R1b	59	99.9%
H28	15-13-16-24-11-13-13-11-15-12-12	1	R1b	72	100.0%
H29	14-14-16-24-11-13-13-11-15-12-11	1	R1b	68	100.0%
H30	14-14-16-24-10-13-13-11-14-12-11	2	R1b	71	100.0%
H31	14-14-16-23-10-13-13-11-11-12-14	1	R1b	37	100.0%
H32	14-14-15-25-11-13-13-11-14-12-12	1	R1b	61	100.0%
H33	14-13-17-24-11-13-13-11-15-12-12	1	R1b	76	100.0%
H34	14-13-17-23-11-13-13-11-15-12-13	1	R1b	61	100.0%
H35	14-13-16-25-10-13-12-11-13-12-12	1	R1b	58	100.0%
H36	14-13-16-24-11-14-13-11-14-12-12	1	R1b	84	100.0%
H37	14-13-16-24-11-13-13-12-14-12-12	1	R1b	80	100.0%
H38	14-13-16-24-11-13-13-11-11-12-12	1	R1b	71	100.0%
H39	14-13-16-24-11-13-12-11-14-13-12	1	R1b	69	100.0%
H40	14-13-16-24-10-13-13-11-13-12-12	1	R1b	79	100.0%
H41	14-13-16-23-11-13-13-11-16-12-13	1	R1b	61	100.0%
H42	14-13-15-24-11-13-13-11-15-14-12	1	R1b	49	100.0%

Haplotype	*Definition*[a]	*Frequency*	*Haplogroup*	*Fitness Value*	*Probability*
H43	13-12-17-25-10-11-12-17-18-10-11	1	E1b1b	39	99.9%
H44	14-13-16-24-11-13-14-11-14-12-12	3	R1b	76	100.0%
H45	17-13-15-24-9-11-13-12-12-10-13	2	I2a2	36	95.6%
H46	15-14-17-24-11-15-13-11-14-12-12	2	R1b	41	99.4%
H47	14-13-16-24-10-13-13-11-14-12-12	2	R1b	92	100.0%
H48	17-13-16-23-11-12-13-14-16-10-11	1	I2	63	93.6%
H49	15-14-17-23-10-14-11-13-17-10-13	1	L	60	100.0%
H50	15-14-16-24-11-13-13-12-14-12-12	1	R1b	56	100.0%
H51	15-13-17-24-11-13-13-12-15-12-11	1	R1b	45	91.1%
H52	15-13-16-24-11-13-13-11-15-12-13	1	R1b	63	100.0%
H53	15-13-16-24-10-13-13-11-14-12-12	1	R1b	75	100.0%
H54	15-13-16-23-11-13-13-11-15-12-13	1	R1b	58	100.0%
H55	15-12-19-22-10-11-14-14-14-10-11	1	G2a	55	100.0%
H56	15-12-18-22-10-10-14-15-16-10-12	1	G2a	33	99.9%
H57	15-12-15-24-10-14-14-13-15-9-12	1	T/Q	22/19	62.8%/32.5%
H58	15-12-15-23-10-11-13-12-12-10-12	1	I2a2	42	97.0%
H59	14-15-16-24-11-13-13-11-15-12-11	1	R1b	49	100.0%
H60	14-14-18-24-10-13-13-11-14-12-12	1	R1b	55	100.0%
H61	14-14-16-24-11-13-13-11-15-13-11	1	R1b	58	100.0%
H62	14-14-16-24-10-13-13-11-11-12-12	1	R1b	55	100.0%
H63	14-13-17-25-10-13-13-11-14-12-12	1	R1b	71	100.0%
H64	14-13-17-24-11-13-13-10-14-12-12	1	R1b	65	100.0%
H65	14-13-16-25-11-13-13-11-14-12-12	1	R1b	90	100.0%
H66	14-13-16-25-10-13-12-11-13-12-13	1	R1b	51	100.0%
H67	14-13-16-24-11-13-14-12-14-12-12	1	R1b	61	100.0%
H68	14-13-16-24-11-13-13-11-15-12-12	1	R1b	88	100.0%
H69	14-13-16-24-11-13-13-11-14-11-11	1	R1b	77	100.0%
H70	14-13-16-24-11-12-13-11-14-12-12	1	R1b	71	100.0%
H71	14-13-16-24-10-13-13-12-14-12-11	1	R1b	66	100.0%
H72	14-13-16-23-11-13-13-11-14-12-11	1	R1b	83	100.0%
H73	14-13-16-23-10-13-13-11-11-12-13	1	R1b	52	100.0%
H74	14-13-15-24-11-13-13-12-14-12-12	1	R1b	64	100.0%
H75	14-13-15-23-11-13-13-11-14-12-12	1	R1b	73	100.0%
H76	14-12-16-24-11-13-13-11-14-12-11	1	R1b	70	100.0%
H77	14-12-16-23-11-13-14-11-14-12-13	1	R1b	47	100.0%
H78	14-11-16-25-11-13-13-11-13-12-11	1	R1b	43	100.0%
H79	13-12-18-25-11-11-13-17-18-10-11	1	E1b1b	38	100.0%
H80	13-12-18-25-10-11-13-17-18-10-11	1	E1b1b	47	100.0%
H81	15-14-17-24-10-14-13-11-14-12-12	1	R1b	46	99.5%
H82	14-14-17-24-10-13-13-11-14-12-11	1	R1b	61	100.0%
H83	14-14-17-24-10-11-13-17-19-10-14	1	E1b1b	37	100.0%
H84	14-14-16-24-11-13-14-11-14-12-12	1	R1b	64	100.0%
H85	14-14-16-24-11-13-13-12-14-12-13	1	R1b	59	100.0%
H86	14-14-16-24-10-13-13-11-13-12-11	1	R1b	61	100.0%
H87	14-14-15-24-11-13-13-12-14-12-12	1	R1b	54	100.0%

Haplotype	*Definition*[a]	*Frequency*	*Haplogroup*	*Fitness Value*	*Probability*
H88	14-13-16-24-11-13-12-12-14-12-11	1	R1b	59	100.0%
H89	13-14-16-24-9-11-13-13-14-10-10	1	E1b1b	35	99.9%

[a]Haplotype definition locus order: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS438, DYS439.

Percentages of Y haplogroups by province are shown in Figure 21. In all four provinces, over 75% of the Y-chromosomes have been assigned to Haplogroup R1b. One hundred six of the 126 chromosomes sampled belong to this haplogroup. All provinces also have haplogroup E1b1b, while Alava and Vizcaya also have J2a. Guipuzkoa shows the greatest variation, with L, I2, G2a and one unassigned haplotype also found there.

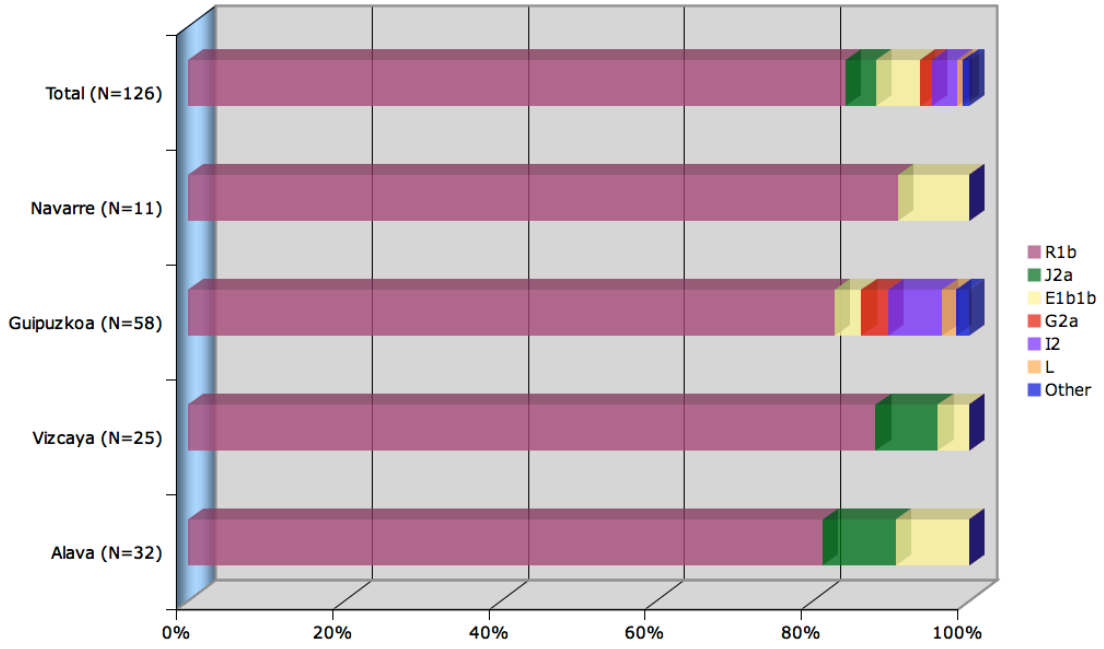

Figure 21. Y haplogroups in the Spanish Basque Provinces. The majority of Y-chromosomes are R1b.

A median-joining network of the 74 R1b haplotypes found among the Basques is displayed in Figure 22. R1b Y-chromosome haplotypes from Alava are shown in blue, Vizcaya in yellow, Guipuzkoa in green, and Navarre in red. The largest central node represents 24 Y-chromosome haplotypes in all four provinces. Each of the larger nodes represents a founder node in a star-like cluster, suggesting several expansion events in the history of this population.

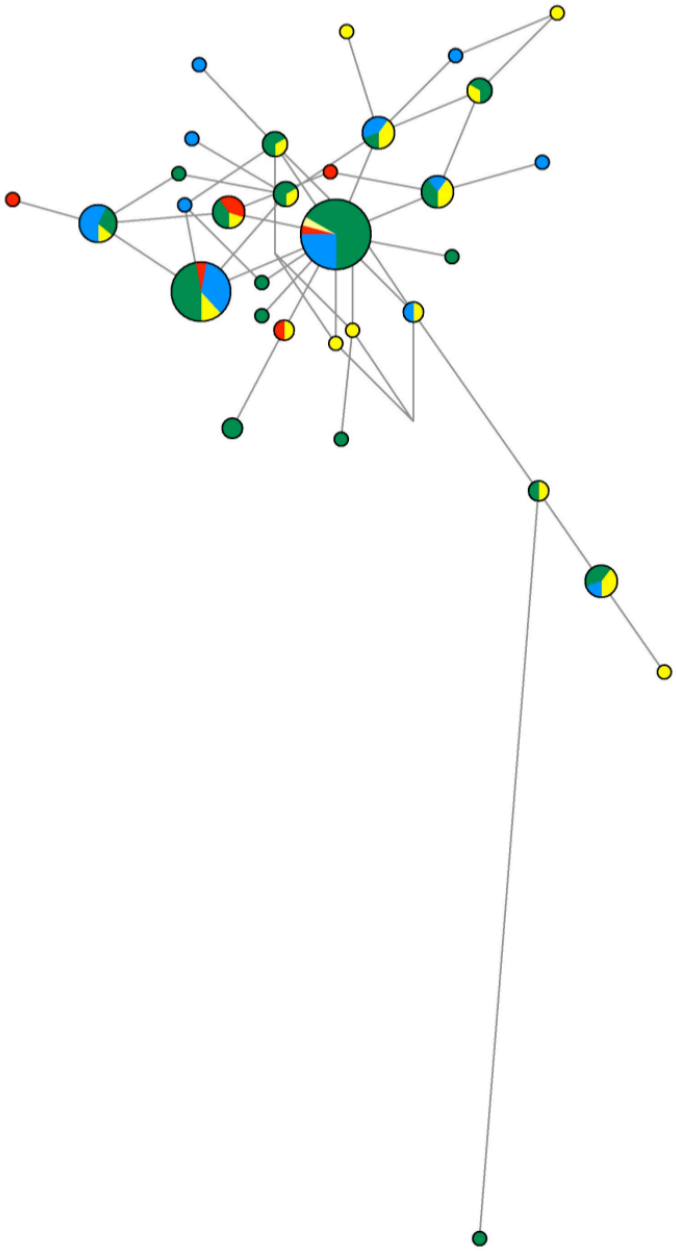

Figure 22. Skeleton median-joining network of all R1b Y-STR haplotypes in the four Basque Provinces. Alava is shown in blue, Vizcaya in yellow, Guipuzkoa in green, and Navarre in red.

Mitochondrial DNA

Restriction Fragment Length Polymorphisms

Mitochondrial DNA haplogroups found among the Basques, determined from analysis of restriction fragment length polymorphisms, are presented in Table 14. The sample size for Navarre included only 30 individuals, and so was not used in province level analyses (and thus is not listed separately in Table 14), but was included in the total haplogroup calculations. One hundred four individuals did not amplify for one or more haplogroups, and are thus listed as not determined. Eighty-five percent of the total sample belongs either to haplogroup H or U/K, the most frequent European haplogroups. Fifteen samples tested positive for 12308 *HinfI,* the determining RFLP site for haplogroups U and K, but then did not amplify for 9052 *HaeII*, the restriction site which distinguishes between those two haplogroups. As a result, they are coded as U/K. Haplogroup V comprises 6.6% of the total sample, while 5.6% of the sample is haplogroup J. Haplogroups I/X, T, W and N1b make up the remaining 2.8%.

Table 14. mtDNA haplogroups present among the Basques of Spain.

Province	*H*	*I/X*	*J*	*N1b*	*T*	*V*	*W*	*U/K*	*U*	*K*	*Not Determined*[a]	*Total*
Alava	71	0	3	0	2	5	3	1	21	4	8	118
Vizcaya	111	2	20	0	3	15	1	4	31	10	36	233
Guipuzkoa	102	1	6	1	1	14	1	10	27	10	52	225
Total	314	3	29	1	6	34	5	15	80	24	104	615

[a]*Individuals which did not amplify for one or more haplogroups.*

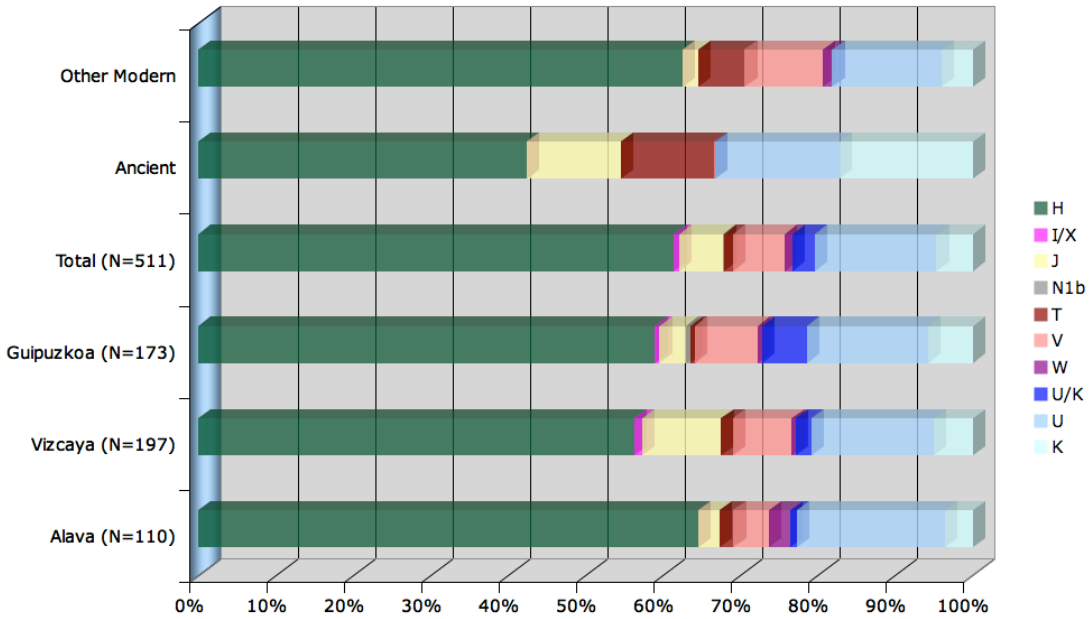

Figure 23. mtDNA haplogroups among Basques. Ancient and Other Modern Samples adapted from Alzualde et. al. (2005), other samples present study. In the ancient and other modern samples, haplogroups T and X are combined.

Percentages of each haplogroup by province, along with totals for the present study and comparative data from additional modern and ancient samples adapted from Alzualde *et al.* (2005), are displayed graphically in Figure 23. In the ancient sample, the most frequent haplogroups are H (42%), U/K (33%), T/X (12%), and J (12%). Haplogroups I and V are present at a frequency of 0.01% each, while haplogroup W is absent. In the additional modern sample, 62% are haplogroup H, 18% are U/K, and 10% are haplogroup V. Haplogroup J is present at a frequency of 2%, W is found at a frequency of 1.2%, and I is absent. In the present study, frequencies of haplogroups H and V are comparable to those found in the other modern sample, while frequencies of haplogroups U/K and J are comparable to those reported in the ancient sample.

Control Region Sequences

The results of sequencing the control region (HVS-I) of 129 Basque samples are presented in Table 15, with HVS-I positions given minus 16,000. The majority of sequences (55.8%) belong to haplogroup H, with 32% matching the Cambridge Reference Sequence (CRS). Other frequent haplogroups include U (14.7%), defined in part by a transition at 16270, and V (10.9%), defined by a transition at position 16298. Eight haplotypes display transversions: B164 has a T-G transverion at position 16235, B203 A-T at 16235, B602 C-A at 16328, B148 A-C at 16220, B633 C-G at 16176, B133 G-T at 16390, B141 C-A at 16114, and B441 G-C at 16129. All other mutations listed in Table 15 are transitions. The distribution of haplotypes in the provinces of Alava, Vizcaya, and Guipuzkoa are shown in Table 16. While many of the haplotypes are shared across provinces (e.g., B120 and B116), others occur in only a single province (B122 is found only in Vizcaya, while B199 is found in Alava). Of the ten common European mtDNA haplogroups, neither X nor I was detected in the sequence analysis, and only one sample each of haplogroups W and N1b was found.

Table 15. Basque HVS-I sequences (HVS-I positions shown -16,000).

	69	93	114	126	129	145	150	153	162	167	168	172	174	176	185	188	189	192	193	218	220	221	223	224	234	235	239	240	241	242	249	256	259	266	270	278	286	290	291	292	293	294	295	296	298	300	302	304	311	319	320	323	325	328	343	356	362	366	390	Hap	N
CRS	c	t	c	t	g	g	c	g	a	c	c	t	c	c	c	c	t	c	c	c	a	c	c	t	c	a	c	a	a	a	c	t	c	c	c	c	c	c	c	c	a	c	c	c	t	a	a	t	t	g	c	t	t	c	a	t	t	c	g	H	-
B120	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	23
B125	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	10
B123	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	H	3
B136	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	3
B478	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	H	3
B151	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	2
B198	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	2
B385	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	2
B104	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	2
B164	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	.	.	H	2
B186	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	2
B122	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B132	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	H	1
B161	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B175	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B203	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B304	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B463	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B543	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B119	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B155	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B190	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	A	.	.	.	.	.	.	.	.	.	H	1
B297	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	H	1
B299	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B318	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	H	1
B420	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	1
B134	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	H	1
B315	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	C	.	.	.	.	.	H	1
B602	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	H	1

Table 15 continued.

	69	93	114	126	129	145	150	153	162	167	168	172	174	176	185	188	189	192	193	218	220	221	223	224	234	235	239	240	241	242	249	256	259	266	270	278	286	290	291	292	293	294	295	296	298	300	302	304	311	319	320	323	325	328	343	356	362	366	390	Hap	N
CRS	c	t	c	t	g	g	c	g	a	c	c	t	c	c	c	c	t	c	c	c	a	c	c	t	c	a	c	a	a	c	t	c	c	c	c	c	c	c	c	c	a	c	c	c	t	a	a	t	t	g	c	t	t	c	a	t	t	c	g	H	--
B464	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	J	2
B145	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	J	2
B130	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	J	1
B140	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	J	1
B153	-	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	J	1
B169	-	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	J	1
B202	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	J	1
B182	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	J	1
B137	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	K	2
B148	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	K	1
B149	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	A	.	.	.	.	.	.	.	.	K	1
B519	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	K	1
B643	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	K	1
B195	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	T	.	.	.	.	.	.	.	K	1
B633	.	.	.	.	.	A	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	N1b	1
B172	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	T	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	T	2
B199	.	.	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	1
B144	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	1
B133	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	T	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	T	1
B116	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	U	7
B191	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	U	2
B185	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	U	1
B204	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	U	1
B205	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	U	1
B207	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	U	1
B141	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	U	1
B441	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	U	1
B296	T	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	U	1
B154	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	U	1
B526	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	C	.	.	U	1

Table 15 continued.

	69	93	114	126	129	145	150	153	162	167	168	172	174	176	185	188	189	192	193	218	220	221	223	224	234	235	239	240	241	242	249	256	259	266	270	278	286	290	291	292	293	294	295	296	298	300	302	304	311	319	320	323	325	328	343	356	362	366	390	Hap	N
CRS	c	t	c	t	g	g	c	g	a	c	c	t	c	c	c	c	t	c	c	c	a	c	c	t	c	a	c	a	a	c	t	c	c	c	c	c	c	c	c	c	a	c	c	c	t	a	a	t	t	g	c	t	t	c	a	t	t	c	g	H	--
B614	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	C	.	.	.	U	1
B152	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	8
B167	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	2
B101	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	1
B163	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	1
B179	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	1
B381	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	V	1
B382	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	W	1

Table 16. Frequencies of HVS-I sequences among Basques by province.

Haplotype	*Haplogroup*	*Alava*	*Vizcaya*	*Guipuzkoa*	*Total*
B120	H	6	6	11	23
B125	H	2	3	5	10
B123	H	0	3	0	3
B136	H	0	2	1	3
B478	H	0	0	3	3
B151	H	0	1	1	2
B198	H	2	0	0	2
B385	H	1	0	1	2
B104	H	1	1	0	2
B164	H	0	2	0	2
B186	H	2	0	0	2
B122	H	0	1	0	1
B132	H	0	1	0	1
B161	H	0	1	0	1
B175	H	0	1	0	1
B203	H	1	0	0	1
B304	H	1	0	0	1
B463	H	0	0	1	1
B543	H	0	0	1	1
B119	H	0	1	0	1
B155	H	0	1	0	1
B190	H	1	0	0	1
B297	H	1	0	0	1
B299	H	1	0	0	1
B318	H	1	0	0	1
B420	H	0	0	1	1
B134	H	0	1	0	1
B315	H	1	0	0	1
B602	H	0	0	1	1
B464	J	0	0	2	2
B145	J	1	1	0	2
B130	J	0	1	0	1
B140	J	0	1	0	1
B153	J	0	1	0	1
B169	J	0	1	0	1
B202	J	1	0	0	1
B182	J	1	0	0	1
B137	K	0	2	0	2
B148	K	0	1	0	1
B149	K	0	1	0	1
B519	K	0	0	1	1
B643	K	0	0	1	1
B195	K	1	0	0	1

Haplotype	*Haplogroup*	*Alava*	*Vizcaya*	*Guipuzkoa*	*Total*
B633	N1b	0	0	1	1
B172	T	0	1	1	2
B199	T	1	0	0	1
B144	T	0	1	0	1
B133	T	0	1	0	1
B116	U	1	1	5	7
B191	U	2	0	0	2
B185	U	1	0	0	1
B204	U	1	0	0	1
B205	U	1	0	0	1
B207	U	0	1	0	1
B141	U	0	1	0	1
B441	U	0	0	1	1
B296	U	1	0	0	1
B154	U	0	1	0	1
B526	U	0	0	1	1
B614	U	0	0	1	1
B152	V	1	5	2	8
B167	V	0	2	0	2
B101	V	0	1	0	1
B163	V	0	1	0	1
B179	V	0	1	0	1
B381	V	0	1	0	1
B382	W	1	0	0	1

Figure 24 displays a median-joining network of the Basque sequences belonging to Haplogroup H. The large central node contains sequences from Alava (blue), Vizcaya (yellow), and Guipuzkoa (green) and is representative of the Cambridge Reference Sequence (CRS). The second large node also contains sequences from all three provinces, and is distinguished from the CRS by a transition at position 16129. Nine sequences are found exclusively in Alava, thirteen sequences are found only in Vizcaya, and six in Guipuzkoa. The overall phylogeny is star-like, suggesting an expansion event.

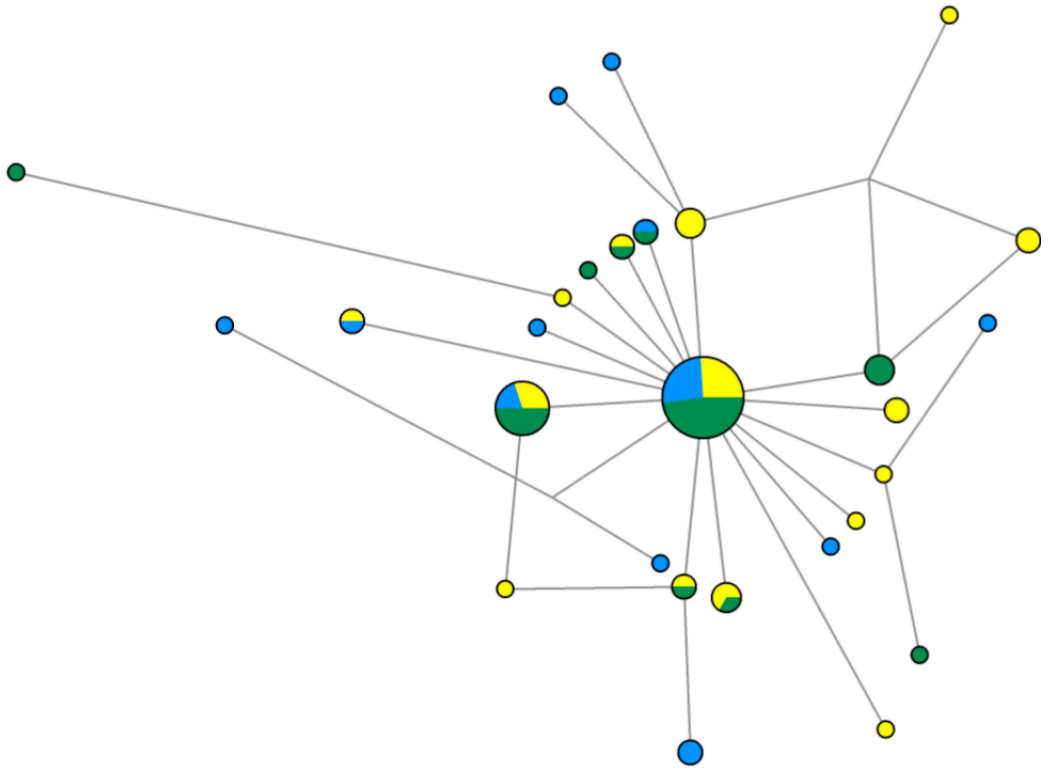

Figure 24. Network of Basque Haplogroup H sequences. Alava (blue), Vizcaya (yellow), Guipuzkoa (green).

Diversity and Neutrality Tests

Autosomal STRs

Preliminary examination of the autosomal STR data revealed three provinces that demonstrated an excess of homozygosity at multiple loci. In Alava and Guipuzkoa, five loci have lower than expected heterozygosity values, while in Vizcaya, seven of the nine loci are significantly less heterozygous than Hardy-Weinberg predictions due to allelic dropout. Figure 25 shows the results of this correction for allelic dropout on the exact test of HWE. The *p* values for the uncorrected data (short blue bars) are compared to the *p* values of the corrected data

(purple bars) from Guipuzkoa (Guo and Thompson 1992). For the uncorrected data, five of nine loci have significant *p* values ($p < 0.05$). In the corrected sample, only one locus (D8S) has a significant *p* value. All other loci now meet Hardy-Weinberg expectations.

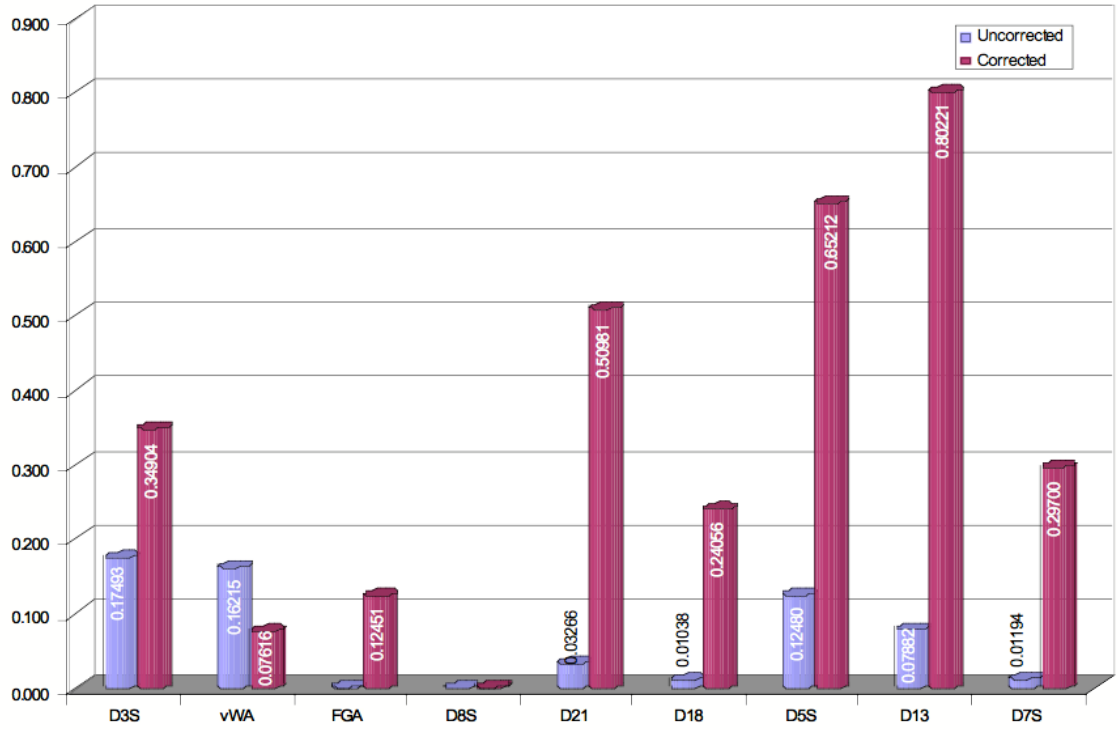

Figure 25. Comparison of p values from the exact test of Hardy-Weinberg equilibrium (Guo and Thompson 1992) for corrected and uncorrected STR data from Guipuzkoa. After removal of all individuals with missing data resulting from allelic dropout, only one locus (D8S) remains significant.

Table 17 presents observed and expected heterozygosity values for each of the nine autosomal STR loci in the four Basque Provinces. P values demonstrate the results of Guo and Thompson's exact test of HWE (1992) after correction of the data for allelic dropout. Observed heterozygosity values among the Basques range from 0.60526 (D5S818) to 0.92105 (vWA), with both extremes found in Navarre, likely as a result of small sample size (N=38). In Alava and Vizcaya, only two loci have

Table 17. Exact test of HWE for nine autosomal loci in four Basque Provinces.

Locus	Alava	Vizcaya	Guipuzkoa	Navarre
D3S1358	N = 96	N = 89	N = 154	N = 38
H_O	0.77320	0.76404	0.76623	0.68421
H_E	0.80615	0.77433	0.77829	0.80000
p	0.02230*	0.28137	0.34904	0.57454
FGA				
H_O	0.87500	0.85393	0.81818	0.81579
H_E	0.86044	0.88574	0.87578	0.87930
p	0.90451	0.01880*	0.12451	0.14558
D5S818				
H_O	0.65625	0.79775	0.72727	0.60526
H_E	0.70610	0.72780	0.74159	0.73439
p	0.14642	0.82662	0.65212	0.42251
D7S820				
H_O	0.71875	0.77528	0.78571	0.73684
H_E	0.80928	0.81553	0.80132	0.82596
p	0.07874	0.32608	0.29700	0.70250
D8S1179				
H_O	0.77083	0.76404	0.75325	0.86842
H_E	0.80988	0.80956	0.81006	0.83474
p	0.00000***	0.00000***	0.00082***	0.04749*
vWA				
H_O	0.85417	0.82022	0.79870	0.92105
H_E	0.81086	0.81946	0.79864	0.82561
p	0.05728	0.88346	0.07616	0.43085
D13S317				
H_O	0.76042	0.76404	0.75325	0.71053
H_E	0.76167	0.79794	0.78282	0.76596
p	0.27417	0.11321	0.80221	0.56594
D18S51				
H_O	0.80208	0.79775	0.81818	0.81579
H_E	0.86938	0.88282	0.87571	0.87053
p	0.12041	0.09120	0.24056	0.86468
D21SS11				
H_O	0.89583	0.83146	0.80519	0.86842
H_E	0.84004	0.84067	0.84073	0.82246
p	0.97004	0.05701	0.50981	0.53438

significantly lower heterozygosity values than expected. When the Bonferroni correction for multiple tests is applied, only the D8S locus demonstrates an excess of

homozygotes in all provinces. This suggests that for the other STR loci examined, the expectations of Hardy-Weinberg have been met (i.e., random mating, little population substructure, and absence of evolutionary forces).

The results of an Analysis of Molecular Variance (AMOVA) using autosomal STR genotypes of 377 individuals in 27 villages within the four Basque Provinces are shown in Table 18. There is no genetic structure between provinces, as indicated by the among groups covariance component (V_a = -0.095). The negative value results from the manner in which the covariance components are estimated, from the mean squares and lower level variances rather than as sums of squares (Excoffier *et al.* 1992). The lack of structure among provinces is confirmed by the global estimate of the fixation index among groups (F_{CT} = -0.0036, p = 0.892). A small amount of subdivision is evident between villages within provinces (1.309% total variation, F_{SC} = 0.0131, p = 0.001). Examination of the locus-by-locus results reveals that three loci make significant contributions to the differences between villages: D7S820 (F_{SC} = 0.0332, p = 0.009), vWA (F_{SC} = 0.0185, p = 0.050), D18S51 (F_{SC} = 0.0319, p = 0.012). The majority of variation, however, is found between individuals within villages (99.045% total variation).

Table 18. Locus-by-locus AMOVA of 9 Autosomal STR loci.

	Among Groups:			Among Populations:			Within Populations:		
Locus	% var	F_{CT}	*p*	% var	F_{SC}	*p*	% var	F_{ST}	*p*
D3S1358	-0.026	-0.0003	0.353	1.257	0.0126	0.142	98.768	0.0123	0.353
FGA	-0.161	-0.0016	0.677	-0.591	-0.0059	0.814	100.752	-0.0075	0.677
D5S818	0.384	0.0038	0.182	-0.292	-0.0029	0.646	99.908	0.0009	0.182
D7S820	-1.036	-0.0104	0.939	3.350	0.0332	0.117	97.686	0.0231	0.939
D8S1179	-0.564	-0.0056	0.904	0.573	0.0057	0.203	99.990	0.0001	0.904
vWA	0.011	0.0001	0.371	1.851	0.0185	0.052	98.138	0.0186	0.371
D13S317	-0.318	-0.0032	0.641	0.110	0.0011	0.497	100.208	-0.0021	0.641
D18S51	-0.546	-0.0055	0.504	3.204	0.0319	0.015	97.342	0.0266	0.504
D21S11	-0.308	-0.0031	0.602	0.767	0.0076	0.316	99.541	0.0046	0.602
Global Estimates	-0.355	-0.0036	0.878	1.309	0.0131	0.011	99.045	0.0096	0.019
Covariance Estimates		$V_a = -0.095$			$V_b = 0.351$			$V_c = 26.517$	

Y-STRs

Gene diversity levels by locus and overall haplotype diversities of 11 Y-STR loci are presented in Table 19. In Alava, the lowest gene diversity is found at DYS393 (0.175), while the highest is 0.667 (DYS389I). Gene diversity values in Vizcaya range from 0.297 (DYS392) to 0.770 (DYS385b), while in Guipuzkoa the low is 0.301 (DYS393) and the high is 0.617 (DYS439). As with the autosomal STR data, Navarre demonstrates the most extreme gene diversity values in the sample, 0.000 at DYS390 and 0.782 at DYS439, due to the small sample of males from this province (N=11). Haplotype diversity in Navarre is 1.000, indicating that every Y in this province is unique and belongs to its own lineage. Haplotype diversity levels in the other provinces are slightly lower, ranging from 0.976 in Alava to 0.993 in Vizcaya.

Table 19. Gene and haplotype diversity of 11 Y-STR loci in four Basque Provinces.

Locus	Alava N=32	Vizcaya N=25	Guipuzkoa N=58	Navarre N=11	Mean
DYS19	0.284	0.357	0.462	0.345	0.362
DYS389I	0.667	0.440	0.562	0.182	0.463
DYS389II	0.433	0.407	0.482	0.564	0.471
DYS390	0.373	0.540	0.472	0.000	0.346
DYS391	0.567	0.527	0.469	0.618	0.545
DYS392	0.315	0.297	0.390	0.473	0.369
DYS393	0.175	0.333	0.301	0.345	0.289
DYS385a	0.486	0.360	0.480	0.673	0.500
DYS385b	0.599	0.770	0.557	0.345	0.568
DYS438	0.315	0.363	0.352	0.327	0.339
DYS439	0.603	0.697	0.617	0.782	0.675
Haplotype Diversity	0.976 ± 0.014	0.993 ± 0.013	0.978 ± 0.012	1.000 ± 0.039	

Analysis of Molecular Variance (AMOVA) of Y-STR haplotypes among the four Basque provinces (Table 20) shows that while a small but significant amount of variation (1.71%, p=0.0369) is accounted for between provinces, 98.29% of the total variation is found between males within populations.

Table 20. AMOVA based on Y-STRs in four Basque Provinces.

Source of variation	Sum of squares	Variance components	Percentage of variation	p
Among populations	11.229	0.044	1.71	0.0369
Within populations	305.612	2.505	98.29	--
Total	316.841	2.549		

mtDNA sequences

Gene and nucleotide diversity, as well as two measures of selective neutrality (Tajima's D and Fu's F_s), in the three provinces for which HVS-I sequence data were collected are presented in Table 21. Gene diversity and nucleotide diversity show very little difference between provinces, unlike the diversity measures calculated

based on autosomal and Y-STR data. Tajima's D and Fu's F_s are both significantly negative in all three provinces, indicating populations that have undergone expansion.

Table 21. Measures of diversity and neutrality of HVS-I sequences in three Basque Provinces.

	Alava	Vizcaya	Guipuzkoa
Gene diversity	1.000 ± 0.006	1.000 ± 0.004	1.000 ± 0.005
Nucleotide diversity	0.011 ± 0.007	0.011 ± 0.007	0.010 ± 0.006
Tajima's D	-1.99 (0.009)	-2.09 (0.001)	-2.05 (0.009)
Fu's F_s	-26.15 (0.000)	-26.24 (0.000)	-26.33 (0.000)

The mismatch distribution of pairwise differences displayed in Figure 26 shows that both Alava (blue) and Vizcaya (yellow) have unimodal distributions, with a peak at three pairwise differences. A unimodal distribution is expected in populations that have undergone an expansion event, in agreement with the neutrality statistics. The distribution for Guipuzkoa (green), however, appears bimodal, with peaks at one and four pairwise differences. The raggedness index for this distribution is non-significant however (0.038, $p = 0.473$), and the sum of squared deviations goodness-of-it test does not reject the null hypothesis of a population expansion (SSD = 0.01, $p = 0.423$).

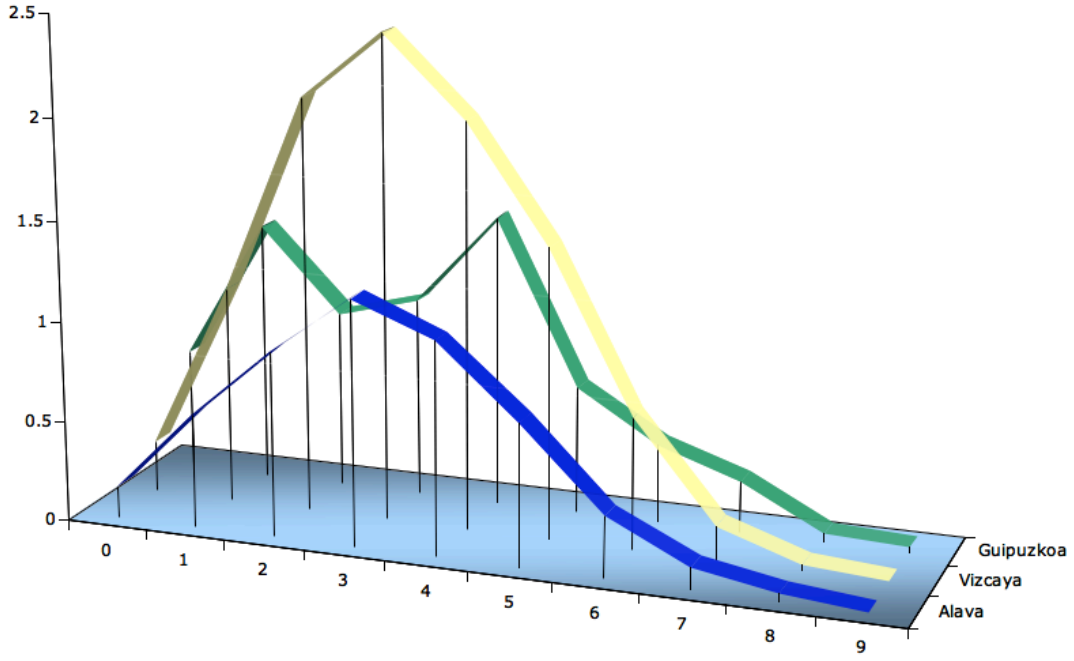

Figure 26. Mismatch distribution of HVS-I sequence pairwise differences in three Basque Provinces. Alava (blue) and Vizcaya (yellow) have a unimodal distribution, while Guipuzkoa has a bimodal distribution.

Examination of mismatch distributions for individual haplogroups demonstrates that J, K, and T each have unimodal peaks at three, while V has a unimodal peak at one difference, and haplogroup U has a unimodal peak at four differences (Figure 27). This suggests that haplogroup V has experienced a more recent expansion event, while the expansion of haplogroup U is more ancient. It is possible that the mismatch distribution for Guipuzkoa reflects these expansions, rather than representing a population that has had a relatively constant size.

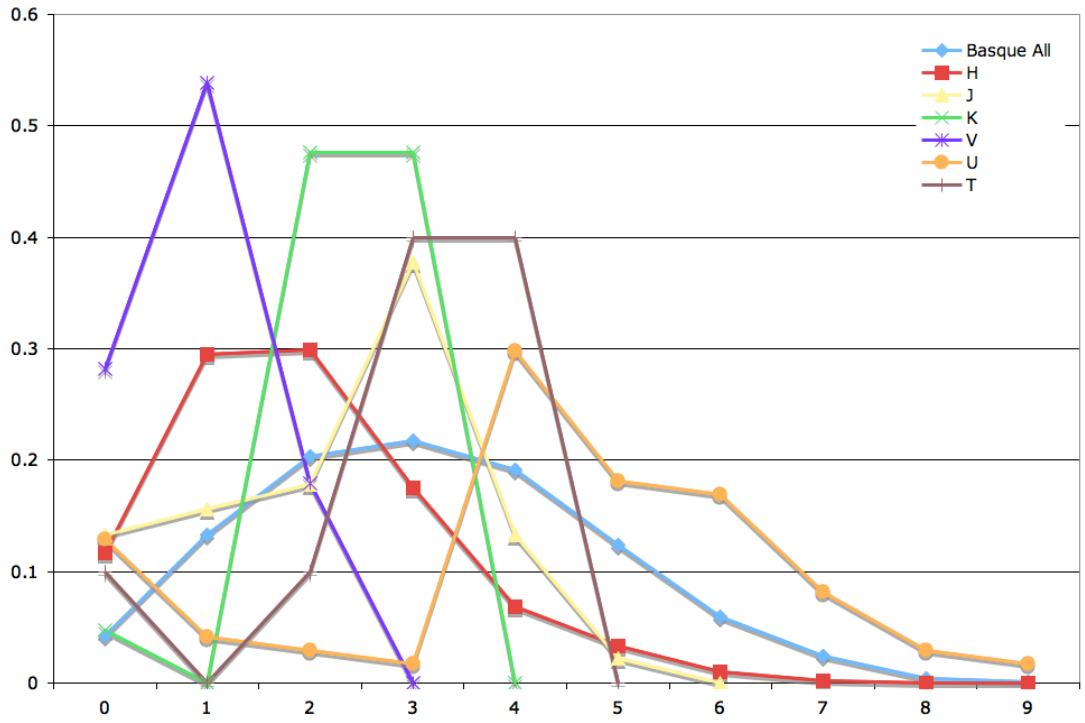

Figure 27. Mismatch distribution of mitochondrial haplogroups among the Basques. The overall distribution (light blue) shows a peak at 3 pairwise differences, while V has a unimodal peak at 1 difference, and U has a unimodal peak at 4 differences.

In agreement with the AMOVA results based on Y-STRs, AMOVA analysis of the HVS-I sequence data shows a small but significant difference between provinces (1.03%, p = 0.03079), but the majority of variation (98.97%) is found between individuals within populations (Table 22). These results suggest that the Basques can be considered a homogenous population.

Table 22. AMOVA based on mtDNA control region sequences in three Basque Provinces.

Source of variation	*Sum of squares*	*Variance components*	*Percentage of variation*	*ϕ statistics*	*p*
Among groups	4.640	0.01665	1.03	ϕ_{ST} = 0.0103	0.03079
Within populations	204.841	1.60032	98.97		
Total	209.481	1.61697			

Interpopulation Analyses

In addition to intrapopulation analyses, relationships between groups were examined to explore the various hypotheses proposed for the origin of the Basque population. For the molecular markers, diversity and neutrality measures were calculated, and AMOVAs were performed for the uniparental systems (Y-chromosomes and mtDNA) by both geographic region and language family to test for population structure. In addition, each of the three hypotheses of Basque origins (Basque-Caucasian, Vasco-Iberian, and Pre-Indo-European) was investigated using phylogenetic and phylogeographic techniques.

Diversity and Neutrality Measures

Autosomal STRs

Genetic differentiation between the Basques and twenty-seven comparative populations, including one Caucasian and two North African groups, was measured based on autosomal STR data using heterozygosity and G_{ST}. For each locus, Table 23 presents heterozygosity values by population, gene differentiation within subpopulations (H_S), gene differentiation between subpopulations (H_T), and the coefficient of gene differentiation (G_{ST}). Average heterozygosity values by population range from a low of 0.802 in Yemen to a high of 0.820 in Scotland.

Among the Basques, heterozygosity is lowest in Alava (0.805) and highest in Vizcaya (0.812). This is within the range of heterozygosity values seen in other Iberian populations (0.804-0.815). Total gene diversity between subpopulations (H_T) is high, ranging from 0.726 for D5S818 to 0.878 for D18S51. However, most of this diversity is explained by variation between individuals within subpopulations (H_S). The percentage of gene differentiation between subpopulations relative to the total gene differentiation (G_{ST}) ranges from a high of 0.010 for D13S317 to a low of 0.006 for D3S1358, FGA, and vWA.

Table 23. Gene diversity between populations based on autosomal STR data.

Population	D3S1358	FGA	D5S818	D7S820	D8S1179	vWA	D13S317	D18S51	D21S11	Average
1) Alava[1]	0.804	0.860	0.705	0.808	0.790	0.807	0.758	0.869	0.840	0.805
2) Vizcaya[1]	0.772	0.886	0.728	0.816	0.781	0.817	0.792	0.876	0.840	0.812
3) Guipuzkoa[1]	0.778	0.875	0.741	0.801	0.801	0.799	0.773	0.876	0.840	0.809
4) Navarre[1]	0.798	0.879	0.705	0.826	0.815	0.826	0.754	0.870	0.819	0.810
5) Andalusia	0.803	0.868	0.708	0.797	0.824	0.805	0.795	0.879	0.856	0.815
6) Cantabria	0.796	0.871	0.715	0.796	0.826	0.803	0.779	0.882	0.846	0.813
7) Catalonia	0.785	0.860	0.714	0.815	0.781	0.825	0.769	0.879	0.809	0.804
8) Galicia	0.786	0.855	0.712	0.796	0.817	0.822	0.795	0.880	0.830	0.810
9) Murcia	0.815	0.860	0.718	0.787	0.807	0.820	0.758	0.866	0.826	0.806
10) Valencia	0.800	0.872	0.702	0.803	0.826	0.807	0.781	0.875	0.839	0.812
11) Austria	0.806	0.864	0.709	0.808	0.814	0.806	0.801	0.872	0.854	0.815
12) Belgium	0.801	0.854	0.707	0.811	0.806	0.808	0.793	0.880	0.831	0.810
13) Bosnia	0.794	0.850	0.713	0.803	0.815	0.807	0.745	0.879	0.867	0.808
14) Germany	0.781	0.870	0.709	0.814	0.790	0.818	0.774	0.885	0.841	0.809
15) Greece	0.788	0.855	0.733	0.794	0.814	0.822	0.775	0.881	0.846	0.812
16) Hungary	0.794	0.864	0.728	0.798	0.809	0.805	0.787	0.886	0.855	0.814
17) Tuscany	0.788	0.865	0.722	0.796	0.832	0.793	0.745	0.868	0.852	0.807
18) Kosovo	0.778	0.868	0.719	0.798	0.798	0.819	0.794	0.852	0.830	0.806
19) Poland	0.802	0.863	0.717	0.812	0.797	0.804	0.759	0.873	0.866	0.810
20) Portugal	0.786	0.862	0.710	0.811	0.816	0.810	0.785	0.876	0.848	0.811
21) Russia	0.783	0.860	0.734	0.811	0.799	0.803	0.781	0.878	0.845	0.810
22) Scotland	0.799	0.856	0.727	0.804	0.834	0.811	0.828	0.864	0.855	0.820

Population	D3S1358	FGA	D5S818	D7S820	D8S1179	vWA	D13S317	D18S51	D21S11	Average
23) Vojvodina	0.773	0.850	0.690	0.796	0.784	0.810	0.786	0.883	0.854	0.803
24) Slovenia	0.797	0.876	0.718	0.810	0.780	0.809	0.785	0.879	0.855	0.812
25) Switzerland	0.791	0.869	0.725	0.821	0.830	0.809	0.773	0.877	0.842	0.815
26) Egypt	0.772	0.873	0.763	0.784	0.819	0.806	0.792	0.858	0.825	0.810
27) Morocco	0.779	0.851	0.727	0.772	0.824	0.822	0.748	0.878	0.831	0.803
28) Oman	0.784	0.865	0.754	0.777	0.842	0.787	0.757	0.879	0.856	0.811
29) Turkey	0.780	0.864	0.751	0.813	0.822	0.802	0.779	0.872	0.842	0.814
30) Yemen	0.780	0.850	0.756	0.787	0.833	0.776	0.739	0.869	0.827	0.802
31) Georgia	0.775	0.871	0.749	0.810	0.806	0.765	0.746	0.874	0.855	0.806
H_S[a]	0.787	0.861	0.720	0.799	0.808	0.805	0.772	0.872	0.839	0.807
H_T[b]	0.792	0.866	0.726	0.805	0.816	0.810	0.780	0.878	0.847	0.813
G_{ST}[c]	0.006	0.006	0.007	0.007	0.009	0.006	0.010	0.007	0.009	0.008

[1]Present Study, [a]Gene diversity within subpopulations., [b] Gene diversity between subpopulations, [c] Coefficient of gene differentiation (Nei 1973).

Y-STRs

For Y-STR haplotype data, gene and haplotype diversity, as well as the Garza-Williamson index, were calculated, and are presented in Table 24. The lowest gene diversity is seen among the South Moroccan Berbers (0.2725), while the highest value is found in Romania (0.6735). Gene diversity values for Basque populations range from 0.3966-0.5137. Haplotype diversity is lowest among the Mozabites of Algeria (0.8283), and highest in the Abkhazians (1.0000). A value of 1.0000 indicates that each haplotype is from a distinct lineage, not surprising given the sample size (just 12 individuals). Among the Basques, haplotype diversity ranges from 0.9148-0.9631, with the study population having a haplotype diversity of 0.9421. Of the populations in Europe, only Albanians from Kosovo (0.9406) and the Swedish Saami (0.9346) have lower haplotype diversity values than the study population. The Garza-Williamson index is lowest in the sample from Western Russia

(0.7770), while several populations (*e.g.*, Andalusia, Galicia) have G-W index values of 1.0000. None of the populations have values below 0.68, the threshold for populations which have experienced a recent bottleneck using data on 7 or more STR loci (Garza and Williamson 2001). For most of the populations examined, including the Basques (G-W range 0.8627-1.0000, study population 0.9886), effective male population size (N_e) not been drastically reduced.

Table 24. Gene and haplotype diversity, as well as Garza-Williamson index values for 55 populations based on Y-STR data.

Population	***N***	***Gene Diversity***	***Haplotype Diversity***	***G-W Index***
Basque (present study)	128	0.4268	0.9421 ± 0.0127	0.9886
Albania	101	0.5701	0.9455 ± 0.0134	1.0000
Andalusia	56	0.5725	0.9890 ± 0.0069	1.0000
Barcelona	224	0.5306	0.9778 ± 0.0053	0.9762
Basque (Brion)	29	0.5137	0.9631 ± 0.0204	0.9357
Basque (Garcia)	167	0.3966	0.9148 ± 0.0149	0.8627
Basque Residents	60	0.4822	0.9791 ± 0.0079	1.0000
Belgium	113	0.5830	0.9771 ± 0.0073	0.9022
Bosnia	181	0.4621	0.9734 ± 0.0054	0.9343
Bulgaria	126	0.5604	0.9874 ± 0.0047	1.0000
Cantabria	107	0.5387	0.9741 ± 0.0074	0.9796
Central Portugal	206	0.5808	0.9903 ± 0.0026	0.9841
Croatia	166	0.5317	0.9801 ± 0.0048	0.9748
Estonia	133	0.5968	0.9869 ± 0.0038	1.0000
France	100	0.5412	0.9834 ± 0.0052	0.8722
Galicia	53	0.5634	0.9746 ± 0.0124	1.0000
Germany	439	0.5907	0.9918 ± 0.0012	0.9667
Greece	69	0.6662	0.9983 ± 0.0028	0.9765
Hungary	115	0.6441	0.9963 ± 0.0017	0.9889
Ireland	151	0.4636	0.9679 ± 0.0075	0.9251
Kosovo Albanians	117	0.5500	0.9406 ± 0.0098	0.9287
Latvia	145	0.5841	0.9907 ± 0.0027	1.0000
Lithuania	152	0.5814	0.9884 ± 0.0031	1.0000
Moscow	85	0.5687	0.9776 ± 0.0079	0.9877
North Italy	104	0.5644	0.9733 ± 0.0095	0.8482
North Portugal	55	0.5234	0.9791 ± 0.0088	1.0000
Piedmont Italy	233	0.5914	0.9771 ± 0.0037	0.9512

Population	*N*	*Gene Diversity*	*Haplotype Diversity*	*G-W Index*
Poland	919	0.5428	0.9879 ± 0.0012	1.0000
Roma Portugal	125	0.5747	0.9080 ± 0.0122	0.9621
Romania	91	0.6735	0.9941 ± 0.0027	0.8389
Rome	125	0.6595	0.9961 ± 0.0017	0.9564
Serbia	185	0.6121	0.9797 ± 0.0041	1.0000
Sweden	344	0.6187	0.9837 ± 0.0028	0.9852
Swedish Saami	38	0.5519	0.9346 ± 0.0240	0.9167
Switzerland	125	0.5925	0.9902 ± 0.0031	0.8291
Valencia	140	0.5617	0.9852 ± 0.0043	0.8740
Western Russia	543	0.5974	0.9866 ± 0.0017	0.7770
Abazinians	14	0.6201	0.9780 ± 0.0345	0.9429
Abkhazia	12	0.6450	1.0000 ± 0.0340	0.9556
Armenia	100	0.6128	0.9945 ± 0.0026	0.9306
Azerbaijan	65	0.6245	0.9966 ± 0.0035	0.9753
Chechenya	19	0.4645	0.9708 ± 0.0273	0.9127
Darginians	22	0.4119	0.9827 ± 0.0183	0.9583
Georgia	77	0.6237	0.9877 ± 0.0046	1.0000
Ingushians	24	0.4084	0.8623 ± 0.0666	0.9841
Kabardians	55	0.6345	0.9845 ± 0.0062	1.0000
Lezginians	17	0.4622	0.9779 ± 0.0267	0.9306
North Ossetia - Ardon	28	0.6104	0.8889 ± 0.0500	0.9339
North Ossetia - Digora	25	0.5214	0.9600 ± 0.0207	0.9815
Rutulians	22	0.4712	0.9870 ± 0.0175	0.9234
Maghreb	20	0.5564	0.9351 ± 0.0042	0.9405
Mozabites Algeria	67	0.2970	0.8283 ± 0.0100	0.8878
Saharawis	30	0.4591	0.8558 ± 0.0077	1.0000
South Moroccan Berber	44	0.2725	0.9028 ± 0.0089	0.9357

mtDNA sequences

Table 25 presents nucleotide diversity values for HVS-I sequences, as well as two measures of selective neutrality (Tajima's *D* and Fu's F_S) for 52 populations. Nucleotide diversity ranges from a low of 0.0102 among Basque sequences collected from the literature, to a high of 0.0265 in the Moroccan Arabs. The Basque samples in the present study have the second lowest nucleotide diversity (0.0114), along with another Atlantic Fringe population, the Welsh. Values of Tajima's D are uniformly

negative and significant for European and Caucasian groups, suggesting that, from a mitochondrial perspective, these populations have experienced an expansion. This is confirmed by the highly significant negative values of Fu's F_S, which also indicate an expansion process acting on these populations, including the Basques. Among the North African populations, four had nonsignificant *D* values, Tata Tunisian Berbers, Algerian Mozabites, Moroccan Arabs, and Souss Moroccan Berbers, indicating that these populations have been through a bottleneck.

Table 25. Diversity and neutrality measures in 52 populations.

Population	*N*	*Nucleotide Diversity*	*Tajima's D*	*p*	*Fu's F_S*	*p*
Basque[1]	131	0.0114	-2.2347	0.0000	-26.3302	0.0000
Austria	99	0.0150	-2.1892	0.0000	-25.8355	0.0000
Basque	156	0.0102	-2.1623	0.0020	-26.4817	0.0000
Belgium	33	0.0133	-2.1837	0.0010	-26.2654	0.0000
Bosnia	144	0.0134	-2.1644	0.0000	-25.9169	0.0000
Brittany	62	0.0139	-1.9638	0.0070	-25.9929	0.0000
Bulgaria	141	0.0149	-2.1201	0.0000	-25.7191	0.0000
Central Portugal	162	0.0157	-2.2912	0.0000	-25.5705	0.0000
Cornwall	92	0.0128	-2.1523	0.0000	-26.1625	0.0000
Czech Republic	83	0.0150	-1.8405	0.0120	-25.8537	0.0000
Denmark	38	0.0127	-1.8313	0.0130	-14.0898	0.0000
England	242	0.0144	-2.2435	0.0000	-25.5208	0.0000
Estonia	149	0.0151	-1.8498	0.0080	-25.6779	0.0000
Finland	153	0.0134	-2.1087	0.0020	-25.9171	0.0000
France	379	0.0140	-2.2524	0.0000	-25.3638	0.0000
Germany	582	0.0136	-2.2316	0.0000	-25.1813	0.0000
Greece	179	0.0140	-2.1798	0.0010	-25.7464	0.0000
Hungary	78	0.0155	-2.1801	0.0000	-25.8040	0.0000
Ireland	300	0.0131	-2.2031	0.0000	-25.6119	0.0000
Italy	248	0.0165	-2.1600	0.0010	-25.2345	0.0000
Karelia	83	0.0134	-1.6779	0.0180	-26.0622	0.0000
North Portugal	183	0.0155	-2.1155	0.0000	-25.5225	0.0000
Norway	629	0.0138	-2.2561	0.0000	-25.1448	0.0000
Poland	473	0.0150	-2.1753	0.0000	-25.1370	0.0000
Romania	92	0.0152	-1.9705	0.0060	-25.8165	0.0000
Russia	379	0.0151	-2.0862	0.0000	-25.2211	0.0000
Sardinia	115	0.0142	-2.0650	0.0030	-25.8827	0.0000
Scotland	895	0.0145	-2.0769	0.0000	-24.8724	0.0000

Population	*N*	*Nucleotide Diversity*	*Tajima's D*	*p*	*Fu's F_S*	*p*
Sicily	196	0.0126	-2.1140	0.0000	-25.8978	0.0000
Slovenia	104	0.0134	-2.0924	0.0000	-26.0431	0.0000
South Portugal	195	0.0155	-2.0802	0.0010	-25.4956	0.0000
Spain Andalusia	114	0.0166	-2.1079	0.0030	-25.6010	0.0000
Spain Castile	38	0.0136	-1.9700	0.0080	-25.8585	0.0000
Spain Catalonia	61	0.0135	-1.8250	0.0120	-26.0352	0.0000
Spain Galicia	135	0.0128	-2.2315	0.0000	-26.0502	0.0000
Spain Leon	61	0.0127	-2.1492	0.0010	-26.1400	0.0000
Sweden	32	0.0159	-1.8895	0.0090	-24.8519	0.0000
Switzerland	224	0.0140	-2.2030	0.0010	-25.6315	0.0000
Wales	92	0.0114	-2.1053	0.0020	-26.4037	0.0000
Armenia	127	0.0195	-2.1112	0.0000	-25.0889	0.0000
Azerbaijan	43	0.0202	-1.8592	0.0100	-25.3829	0.0000
North Ossetia	61	0.0195	-1.6118	0.0270	-25.3349	0.0000
Georgia	28	0.0159	-1.4372	0.0500	-21.0856	0.0000
Adyegi	31	0.0177	-1.9957	0.0090	-21.0856	0.0000
Tunisia	42	0.0215	-1.9200	0.0090	-25.2890	0.0000
Tunisian Tata Berber	37	0.0219	-0.8730	0.2170	-5.8050	0.0450
Tunisian Matmata Berber	51	0.0164	-2.0930	0.0050	-19.0717	0.0000
Sened Tunisian Berber	37	0.0258	-1.4950	0.0500	-23.6183	0.0000
Algeria	27	0.0186	-1.6900	0.0260	-14.3780	0.0000
Mozabites Algeria	41	0.0170	-1.0148	0.1680	-25.5956	0.0000
Moroccan Arabs	34	0.0265	-1.0500	0.1680	-9.7060	0.0000
Souss Moroccan Berber	34	0.0156	-1.4570	0.051	-25.7780	0.0000

AMOVA

Two AMOVAs were performed to examine similarities between 55 populations using Y-STR haplotype data (Appendix 1). In the first, populations were grouped by geographical region: Europe, Caucasus, or North Africa (Table 26). In the second, populations were organized by language family: Indo-European (33), Caucasian (9), Afro-Asiatic (5), Uralic (2), Altaic (1), Kartevelian (1), and Basque (4) (Table 27). In both analyses, the majority of variation is accounted for within populations. However, a significant portion of the variation is found both among

populations within groups, and among the groups themselves. A greater amount of variation among groups is seen when the populations are organized by geographical region (21.87%) rather than language family (10.56%), possibly due in part to the geographic spread of the Indo-European language family in particular.

Table 26. AMOVA of Y-STR data in 55 populations, grouped by geographic region[a].

Source of variation	*Sum of squares*	*Variance components*	*Percentage of variation*	*ϕ statistics*	*p*
Among groups	6065.653	4.43644	21.87	$\phi_{CT} = 0.21869$	0.0000
Among populations within groups	14229.955	2.12822	10.49	$\phi_{SC} = 0.13427$	0.0000
Within populations	92430.287	13.72184	67.64	$\phi_{ST} = 0.32360$	0.0000
Total	112725.896	20.28649			

[a]Geographic regions (N): Europe (37), Caucasus (13), North Africa (5).

Table 27. AMOVA of Y-STR data in 55 populations, grouped by language family.[a]

Source of variation	*Sum of squares*	*Variance components*	*Percentage of variation*	*ϕ statistics*	*p*
Among groups	4897.784	1.91043	10.56	$\phi_{CT} = 0.10563$	0.0000
Among populations within groups	15397.824	2.45349	13.57	$\phi_{SC} = 0.15168$	0.0000
Within populations	92430.287	13.72184	75.87	$\phi_{ST} = 0.24129$	0.00228
Total	112725.896	18.08575			

[a]Language family (N): Indo-European (33), Caucasian (9), Afro-Asiatic (5), Uralic (2), Altaic (1), Kartevelian (1), and Basque (4).

AMOVAs were also conducted to examine relationships between populations using mtDNA sequence data (Appendix 1). As with the Y-STR data, populations were first grouped by geographical region: Europe, Caucasus, or North Africa (Table 28). Populations were also organized by language family: Indo-European (36), Caucasian (1), Afro-Asiatic (7), Uralic (4), Altaic (1), Kartevelian (1), and Basque (2)

(Table 29). In both analyses, significant variation is found at all hierarchical levels, with the majority of variation (>95%) being accounted for within populations. Slightly more variation was accounted for when the populations were grouped by geography (1.78%), rather than language (1.39%), though both values were significant.

Table 28. AMOVA of mtDNA sequence data in 52 populations, grouped by geography.[a]

Source of variation	*Sum of squares*	*Variance components*	*Percentage of variation*	*ϕ statistics*	*p*
Among groups	61.649	0.03916	1.78	$\phi_{CT} = 0.00504$	0.000
Among populations within groups	190.172	0.01087	0.50	$\phi_{SC} = 0.02279$	0.000
Within populations	17806.982	2.14568	97.72	$\phi_{ST} = 0.01783$	0.000
Total	18058.803	2.19571			

[a]Geographical regions (N): Europe (40), Caucasus (5), North Africa (7).

Table 29. AMOVA of mtDNA sequence data in 52 populations, grouped by language family.[a]

Source of variation	*Sum of squares*	*Variance components*	*Percentage of variation*	*ϕ statistics*	*p*
Among groups	84.318	0.03037	1.39	$\phi_{CT} = 0.01390$	0.000
Among populations within groups	167.503	0.00944	0.43	$\phi_{SC} = 0.00438$	0.000
Within populations	17806.982	2.14568	98.18	$\phi_{ST} = 0.01822$	0.000
Total	18058.803	2.18549			

[a] Language family (N): Indo-European (36), Caucasian (1), Afro-Asiatic (7), Uralic (4), Altaic (1), Kartevelian (1), and Basque (2).

Basque-Caucasian Hypothesis

Classical Markers

To examine the relationship between Basque groups and Caucasian populations, various phylogenetic analyses were performed using traditional markers, Y-STRs, and HVS-I sequence data. Figure 28 plots the results of R-Matrix analysis of data on 20 European populations collected from the literature (including Armenians from the Caucasus) using 14 alleles and demonstrates the distinctiveness of the Basques, with populations from Labourd and Guipuzkoa in the upper left quadrant.

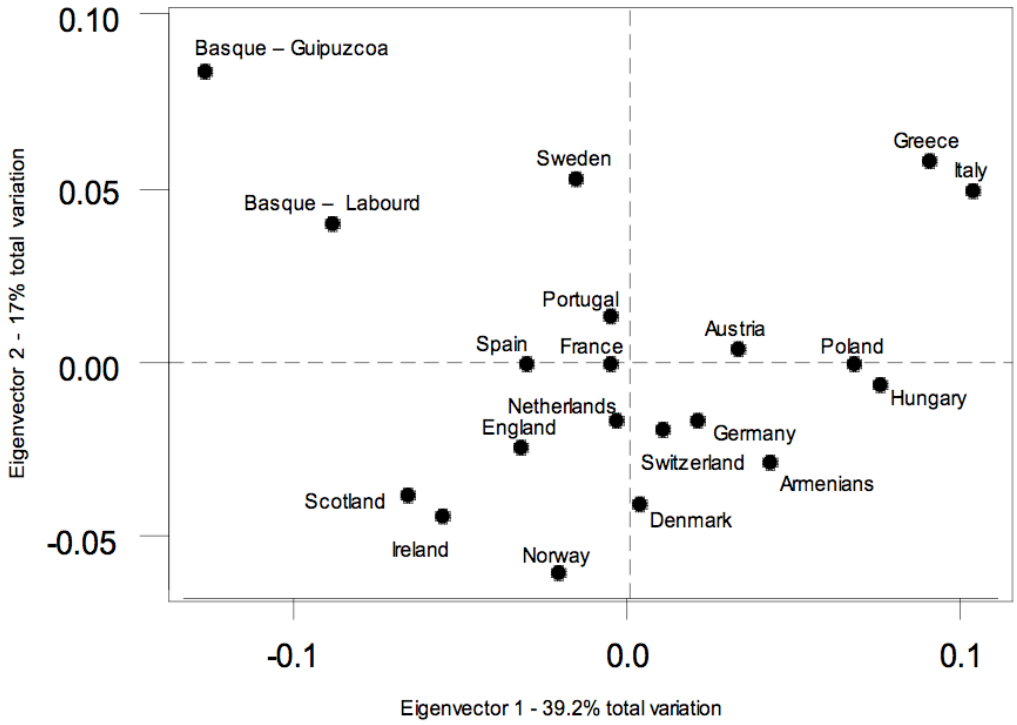

Figure 28. ***R-Matrix analysis of 20 European populations using 14 alleles. The first two principle components account for 56.2% of the variation. The Basque populations from Labourd and Guipuzkoa are distinct from other European populations, as are populations from Greece and Italy. Armenians (a Caucasian population) cluster with groups from Germany and Switzerland.***

The Basques are distinguished from other populations due to high frequencies of RH*cde (>50%), ABO*O (>74%), and GM*Z,A;G (20% in Labourd and 35% in Guipuzkoa) (Figure 29). The Armenians, a population that resides in the Caucasus but speaks an Indo-European language, clusters with populations from Germany, Switzerland, and Denmark. It should be noted that while this analysis *does* include GM haplotype frequencies, information on haplotypes GM*Z,A;B and GM*Z,A;B,S,T which are absent in many European populations, could not be included, as R-Matrix analysis does not allow for the inclusion of null data.

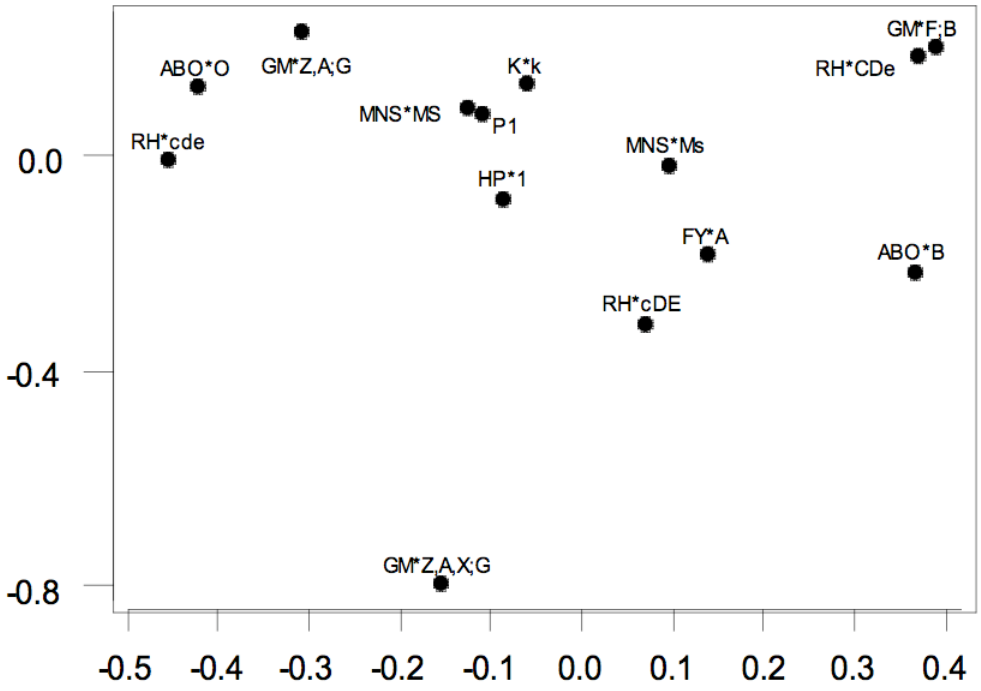

Figure 29. ***S-Matrix plot of alleles used in R-Matrix analysis. The Basques are distinct from other European populations due to high frequencies of RH*cde (>50%), ABO*O (>74%), and GM*Z,A;G (20% in Labourd and 35% in Guipuzkoa).***

Multidimensional scaling analysis of data on 18 European populations (Figure 30) using 37 alleles places the Basques from Guipuzkoa away from other European populations in the lower right corner of the plot, and far from the three Caucasian-speaking populations: Abkhazians, Georgians, and Svani. A final stress of 0.08074 indicates a good fit between the two dimensional scaling and the original genetic distance matrix. GM data were not available for all of these populations, the Caucasians in particular, and thus was not included in this analysis.

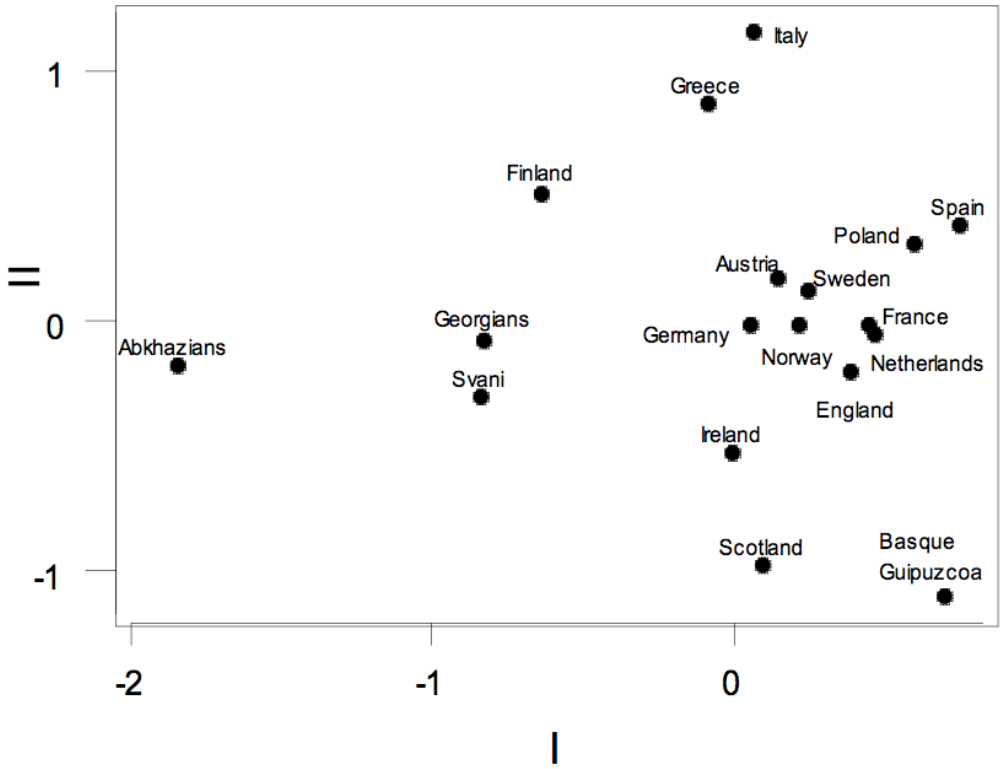

Figure 30. Multidimensional scaling analysis of 20 European populations using 6 blood group (ABO, MNS, P, RH, Kell, Duffy) and 2 serum protein (HP, GM) loci. The Armenians cluster with other European groups. The Basques from Guipuzkoa and Labourd are distinct from other European populations, including the Armenians. STRESS = 0.11482.

Multidimensional scaling analysis (final stress = 0.11482) of 20 European populations using 8 loci, including data on the Asian GM haplotypes (GM*Z,A;B and GM*Z,A;B;S,T), distinguishes the Basques of Labourd and Guipuzkoa from other groups, while Armenians cluster with other Indo-European populations from France and the Netherlands (Figure 31).

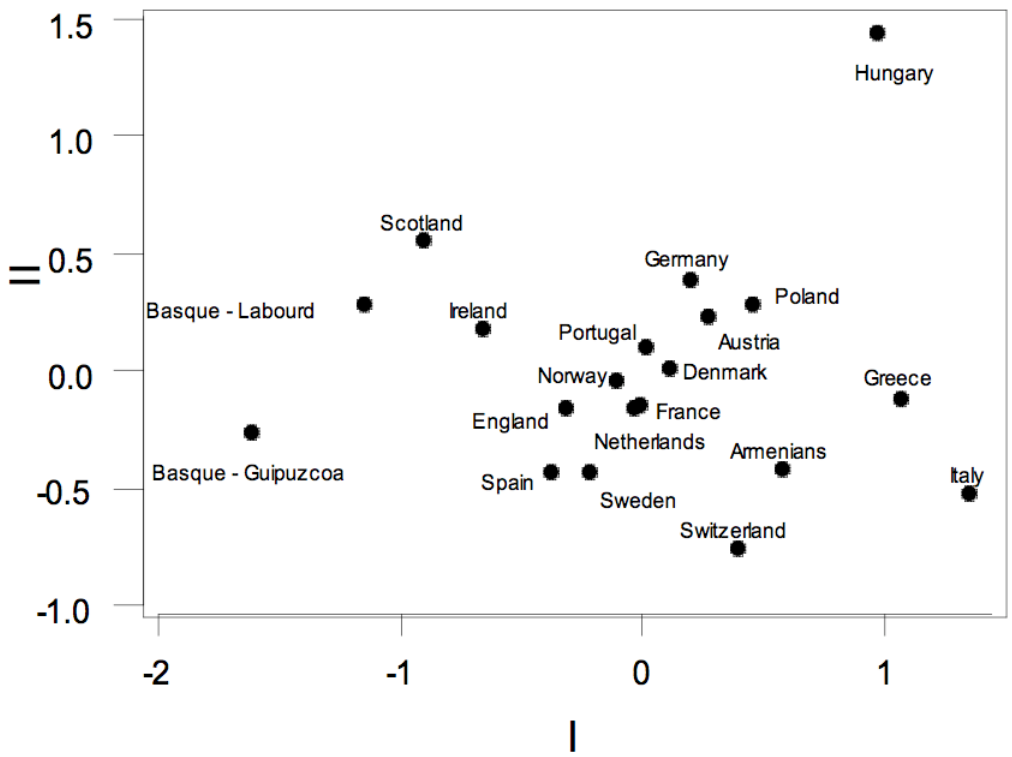

Figure 31. Multidimensional scaling analysis of 20 European populations using 6 blood group (ABO, MNS, RH, P, Kell, Duffy), 3 serum protein (GC, HP, TF), and 3 red cell enzyme loci (ACP, AK, PGM), with 37 alleles. The Basques are in the lower right corner, far from the Caucasian populations of Svani, Georgians, and Abkhazians. STRESS = 0.08074.

Y-STRs

The results of phylogenetic analysis of Y-STR haplotypes in 24 Iberian and Caucasian populations are shown in Figure 32. The Basque populations branch off the

Neighbor-Joining tree with four other Iberian populations (Galicia, Northern Portugal, Andalusia, and Valencia). The Caucasian populations form a separate branch, along with Iberian populations from Barcelona, Cantabria, and Central Portugal. An R^2 value of 0.963 shows a good fit between the NJ tree and the genetic distance matrix.

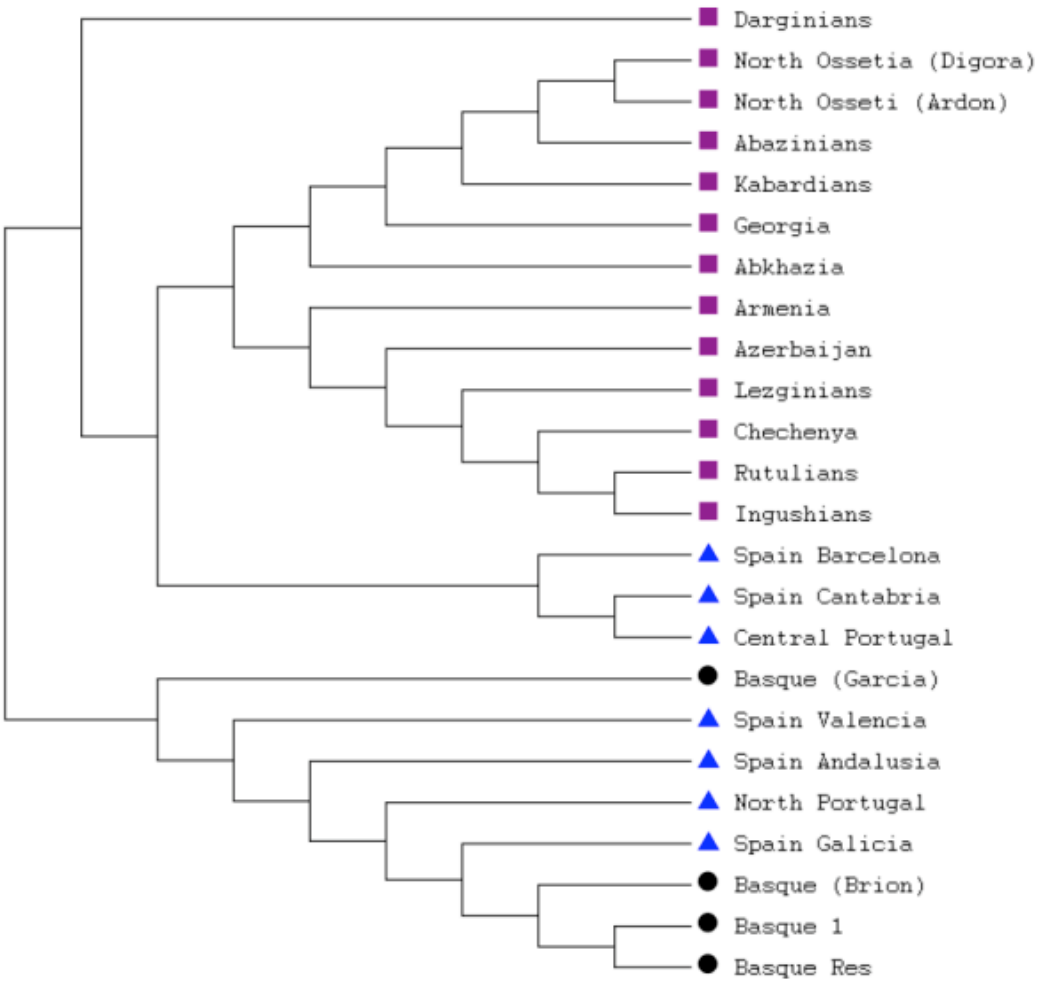

Figure 32. Neighbor-Joining tree based on Y-STR haplotypes data from 24 Basque (black circles), Iberian (blue triangles), and Caucasian (purple squares) populations. The study population is Basque 1. R^2=0.963.

SAMOVA analysis of Iberia and the Caucasus using Y-STR data (Table 30) revealed the greatest genetic variation is seen with three groups of populations: (1) North Ossetia (Digora) (2) the two Basque populations, and (3) the remaining Caucasian and Iberian populations (F_{CT} = 0.504165).

Table 30. Results of SAMOVA of Y-STR haplotype in Iberian and Caucasian populations. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample

Iberia + Caucasus	
K	F_{CT}
2	0.501625
3	0.504165[a]
4	0.502723
5	0.500688
6	0.473764
7	0.496032

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

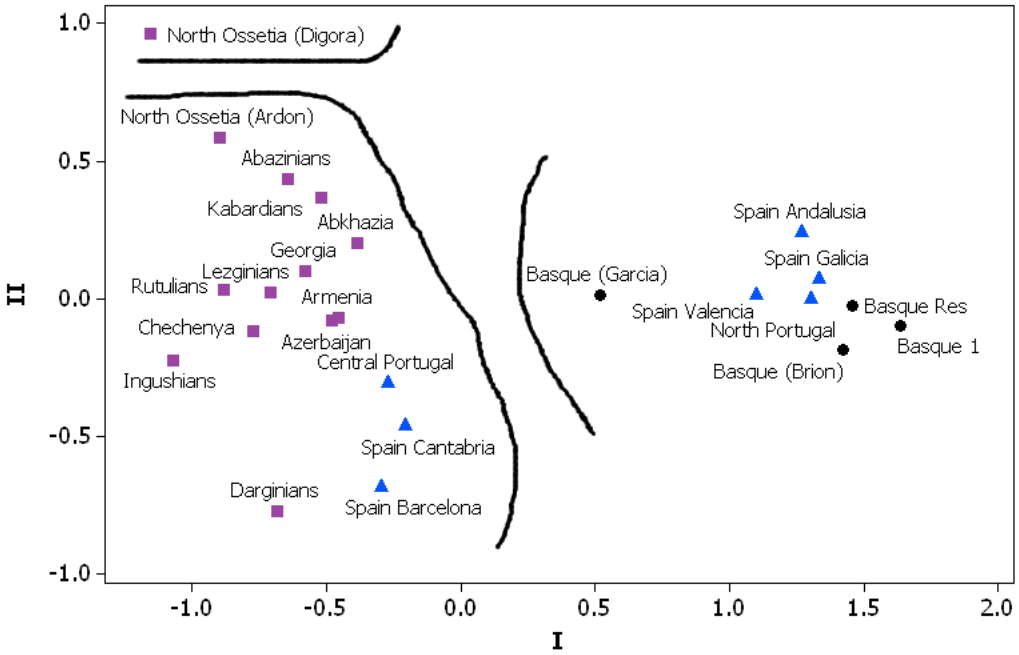

Figure 33. MDS plot of 24 Basque (black circles), Iberian (blue triangles), and Caucasian (purple squares) populations based on pairwise genetic distances between Y-STR haplotypes. Total variation on the first two axes = 79.43%, STRESS = 0.10322, r = 0.97142. Genetic barriers detected by SAMOVA (black lines), F_{CT} = 0.504165.

The multidimensional scaling plot of the same Y-STR haplotype pairwise genetic distance matrix is displayed in Figure 33. The total variation accounted for in the first two axes equals 79.43%, and the stress value (0.10322, p = 0.01) and correlation coefficient (r = 0.97142) indicate a good fit with the original data. As in the Neighbor-Joining tree seen above, the Basque populations cluster with other Iberian groups, while the Caucasian populations form a separate cluster. Genetic barriers detected by SAMOVA are indicated by black lines.

mtDNA sequences

Figure 34 displays a Neighbor-Joining tree of 15 Iberian and Caucasian populations constructed from a Kimura 2P genetic distance matrix based on mitochondrial control region sequence data. The Basque populations form their own branch at the top of the tree, off a branch with the northern Spanish province of Galicia. The five Caucasian populations form a distinct cluster at the bottom of the tree, separate from the Iberian populations. However, the R^2 value (0.838) indicates that this tree is not the best representation of the relationships between groups.

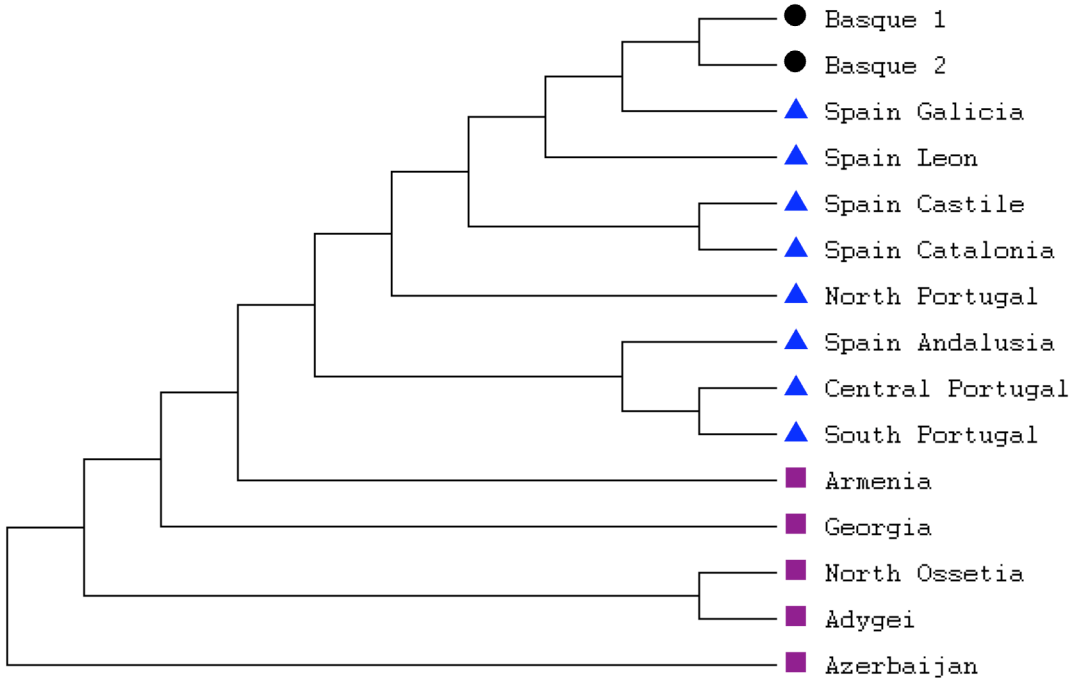

Figure 34. Neighbor-Joining tree of 15 Basque (black circles), Iberian (blue triangles), and Caucasian (purple squares) populations based on mtDNA HVS-I sequences. R^2=0.838.

In the SAMOVA analysis of Iberia and the Caucasus, the greatest genetic variation is seen with four groups of populations: (1) North Ossetia, (2) the two Basque populations, (3) Adygei, and (4) the remaining Caucasian and Iberian populations (FCT = 0.0126338).

Table 31. Results of SAMOVA of mtDNA sequences for Iberian and Caucasian populations. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample

Iberia + Caucasus	
K	F_{CT}
2	0.0118475
3	0.0121858
4	0.0126338[a]
5	0.0119823
6	0.0115201
7	0.0115067

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

The same populations were used to construct an MDS plot of the genetic distance matrix, with the genetic barriers determined by SAMOVA also displayed, in Figure 35. The Basques are separated from all other populations, while the Iberian groups form a cluster with several Caucasian populations, including those from Georgia, Armenia, and Azerbaijan. The plot accounts for 95.63% of the total variation in the sample, and the STRESS value of 0.14791 ($p = 0.01$, $r = 0.91578$), demonstrates a good fit with the original data, and thus an accurate representation of the relationships between these populations based on control region sequences. Unlike the NJ tree, SAMOVA analysis indicates that the North Ossetian population of Digora and the Adygei population are distinct from each other, in addition to being distinct from the other populations in the tree.

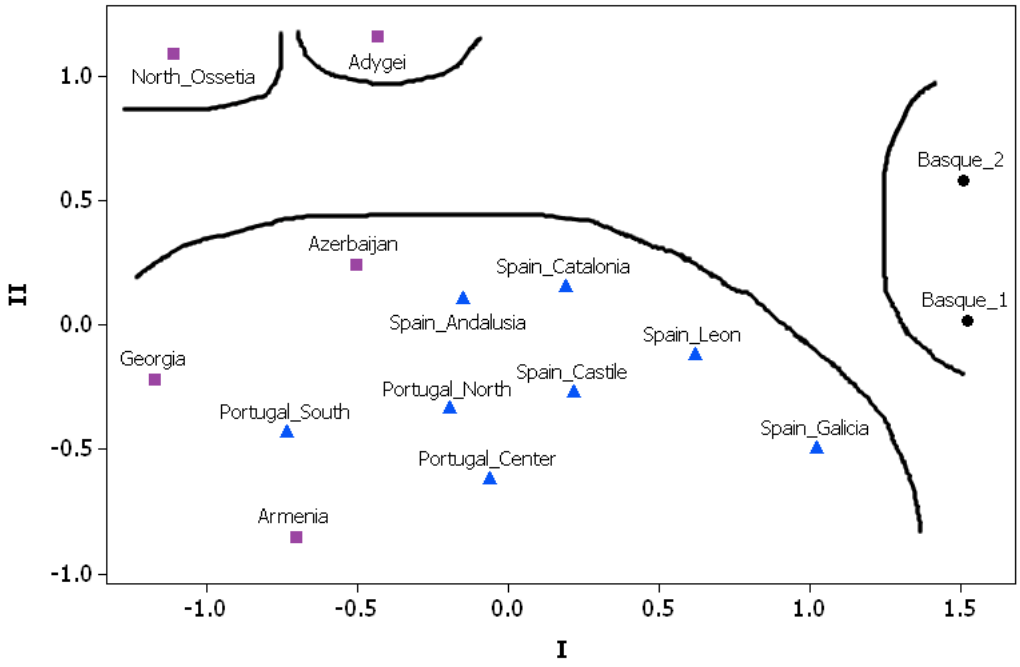

Figure 35. MDS plot of 15 Basque (black circles), Iberian (blue triangles), and Caucasian (purple squares) populations. Genetic barriers detected by SAMOVA shown as black lines. STRESS = 0.14791 (p=0.01), r=0.91578.

Vasco-Iberian Hypothesis

Classical Markers

To examine the Vasco-Iberian hypothesis using classical markers, data on 20 populations (including the Berbers from Algeria) with 31 alleles were gathered to perform an MDS analysis (Figure 36), which demonstrates that the Algerians are distinct from all other populations at the top right corner of the plot. The Basques from Labourd and Guipuzkoa cluster in the opposite corner from the Algerians, near other Atlantic fringe European populations from Scotland and Ireland.

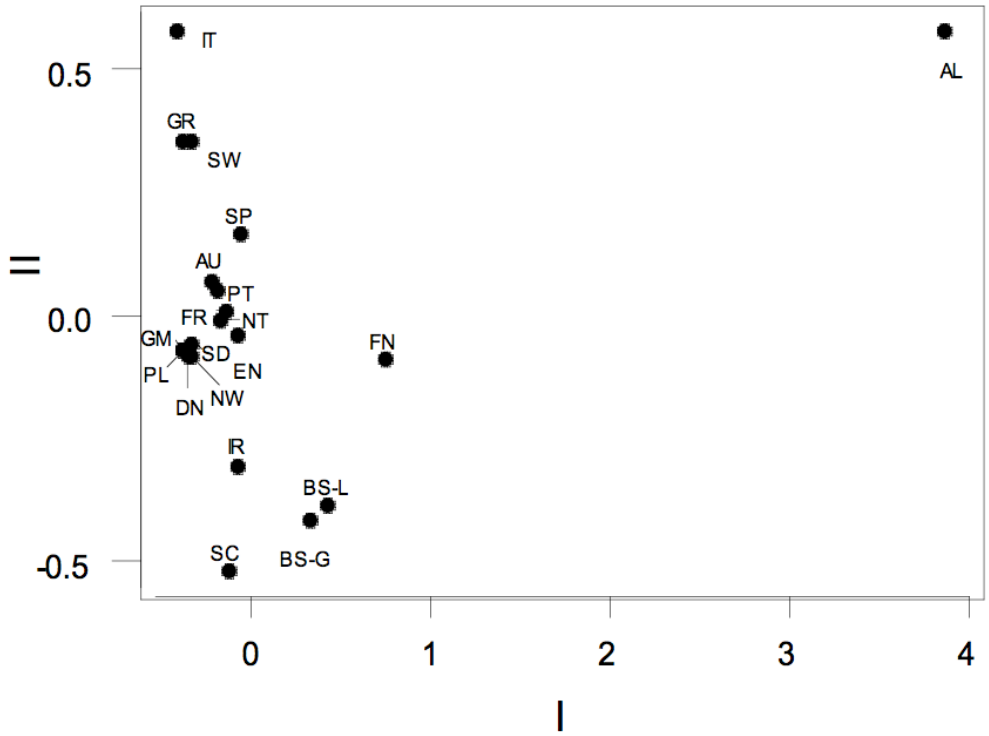

Figure 36. Multidimensional scaling analysis of 20 populations using 6 blood group (ABO, RH, MN, P, Kell, Duffy), 1 serum protein (HP), and 2 red cell enzyme loci (ACP, PGM), with 31 alleles. The North African population of Algeria (AL) is the most distinct at the top right. The two Basque populations, from Guipuzkoa (BS-G) and Labourd (BS-L), are in the lower left corner, near the populations of Scotland and Ireland.

Y-STRs

Figure 37 displays the results of a Neighbor-Joining tree constructed from the pairwise genetic distance matrix based on Y-STR haplotypes in Iberia and North Africa. The four Basque populations (black circles) form a cluster off a branch containing other Iberian populations (blue triangles), including Northern Portugal and the Spanish province of Galicia. The North African populations (green squares), with the exception of the Maghreb (which cluster with the Roma from Portugal), form a separate branch with three Iberian groups (Central Portugal, Barcelona and

Cantabria). The R^2 value for this tree is 0.983, indicating that it is a good representation of the relationships between populations.

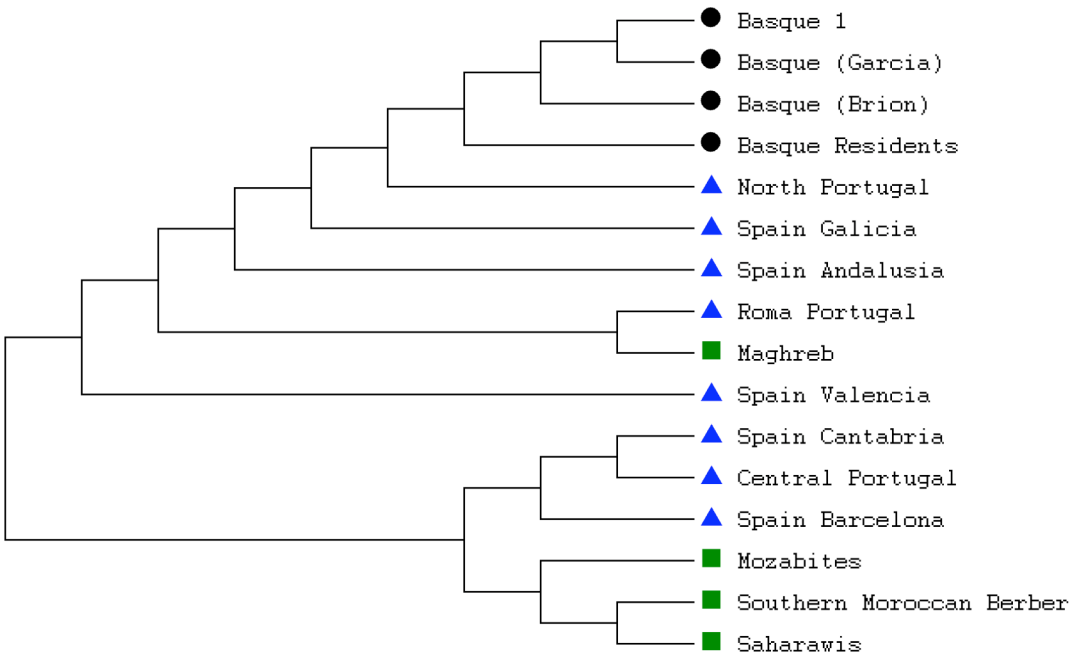

Figure 37. Neighbor-Joining tree based on pairwise genetic distances of Y-STR haplotype data of 16 Iberian and North African populations (R^2=0.983). Basque populations (black circles), Iberian groups (blue triangles), North Africa (green squares). The study population is Basque 1.

Spatial analysis of molecular variance (SAMOVA) was performed using Y-STR data to explore genetic barriers between North Africa and Iberia in the paternal lineage, and these results are presented in Table 32. The greatest genetic variance between groups of populations is found when K = 3 (F_{CT} = 0.480231), with groups comprised of (1) Mozabites, Saharawis, and Southern Moroccan Berbers, (2) Barcelona, Cantabria, and Central Portugal, and (3) the remaining ten populations, including all the Basque groups.

Table 32. Results of SAMOVA of Y-STR haplotypes. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample.

Iberia + North Africa	
K	F_{CT}
2	0.471021
3	0.480231[a]
4	0.478271
5	0.476047
6	0.465619
7	0.463665

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

Multidimensional scaling of the same distance matrix demonstrates a similar pattern (Figure 38). The MDS plot accounts for 82.92% of the total variation, and the stress value of 0.05476 (p = 0.01) indicates a good fit with the original distance matrix. The correlation between the multidimensional scaling plot and the original data (r =0.98218) demonstrates that the MDS plot is an accurate representation of the relationships between groups. The lines on the plot represent the three genetic barriers detected by SAMOVA.

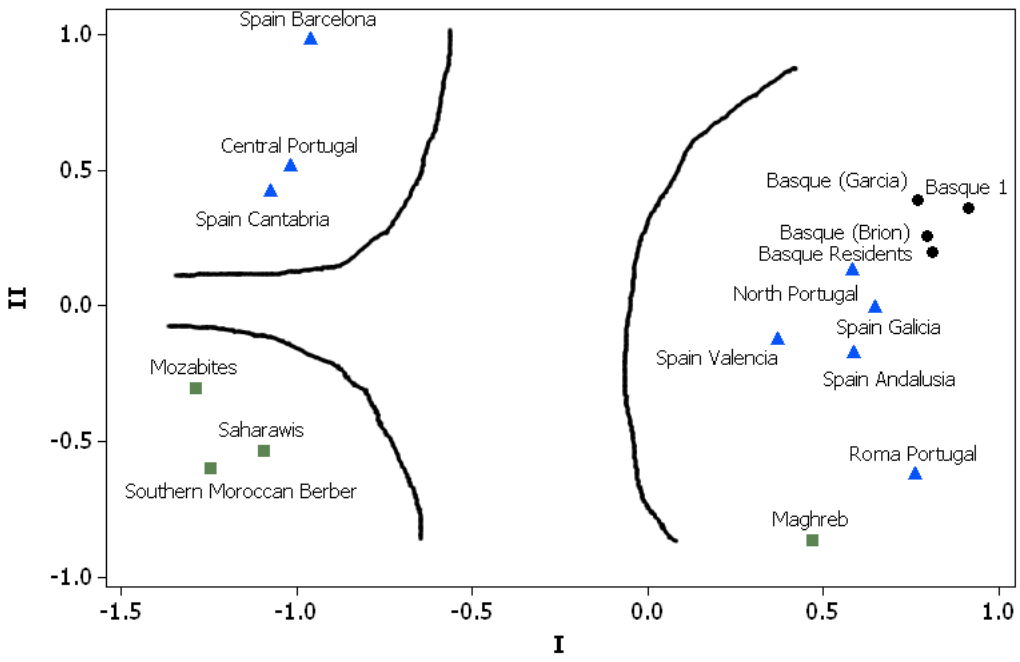

Figure 38. MDS plot of 16 Iberian and North African populations based on Y-STR haplotype data (STRESS=0.05476, r = 0.98218). Basque populations (black circles), Iberian groups (blue triangles), North Africa (green triangles). Genetic barriers detected by SAMOVA (black lines), F_{CT}= 0.480231).

mtDNA sequences

Figure 39 displays the Neighbor-Joining tree of Kimura 2P distances between 18 Iberian and North African populations based on mtDNA control region sequences, with an R^2 value of 0.919. The two Basque populations (black circles) form a branch at the top of the tree, clustered with other Iberian populations (blue triangles), while many of the North African populations have individual branches (green squares).

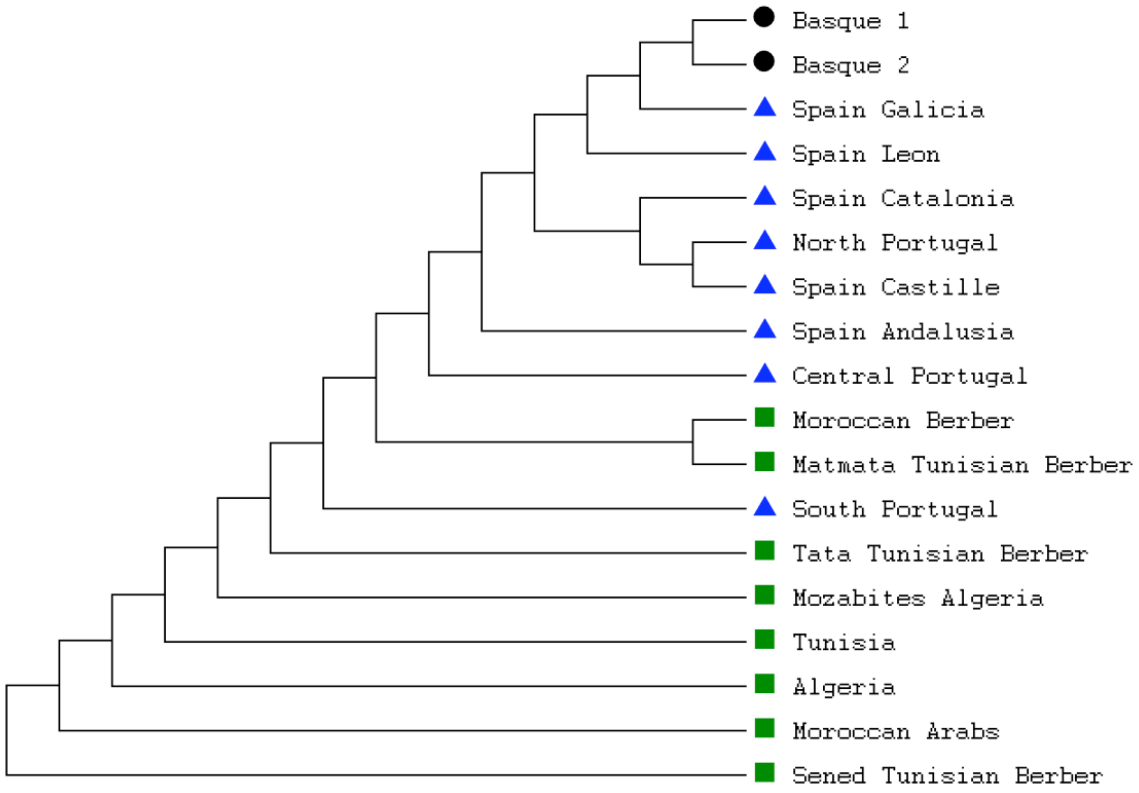

Figure 39. Neighbor-Joining tree of 18 Basque (black circles), Iberian (blue triangles), and North African (green squares) populations based on mtDNA HVS-I sequences. R^2=0.919.

In the SAMOVA analysis of Iberia and North Africa using mtDNA sequences, the greatest genetic variance between groups of populations is found when K = 4 (F_{CT} = 0.0770784), with groups comprised of (1) Mozabites, (2) Tata Tunisian Berbers, (3) Moroccan Arabs and Sened Tunisian Berbers, and (4) all other groups, including the Basques (Table 33).

Table 33. Results of SAMOVA of mtDNA sequence data. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample.[a]

Iberia + North Africa	
K	F_{CT}
2	0.0689103
3	0.0738478
4	0.0770784[a]
5	0.0767607
6	0.0716066
7	0.0672944

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

The same Kimura 2P genetic distance matrix was also used to construct an MDS plot, with the SAMOVA genetic barriers overlaid as black lines (Figure 40). The plot accounts for 80.22% of the total variation present in the sample, and the STRESS value of 0.06447 (r = 0.97222) indicates a very good fit with the original distance matrix. Genetically distinct populations include the Tata Tunisian Berbers, Mozabites from Algeria, and Moroccan Arabs and Sened Tunisian Berbers. The two Basque populations are found on the right side of the plot, surrounded by other populations from Iberia.

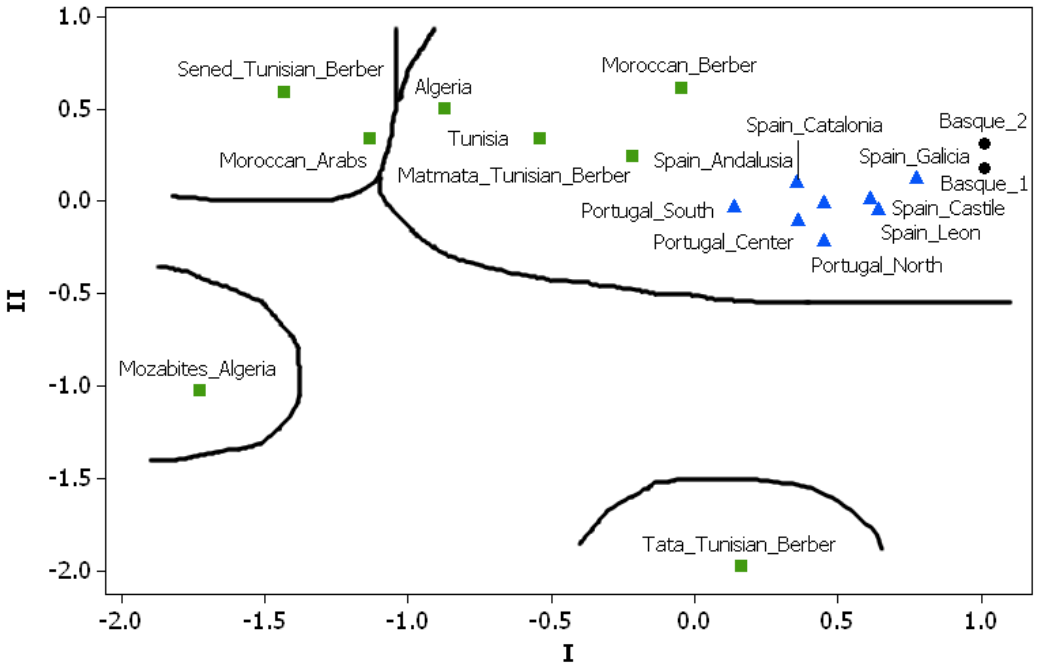

Figure 40. MDS plot of 18 Iberian and North African populations based on mtDNA control region sequences (STRESS=0.06447, r=0.97222). Basque populations (black circles), Iberian groups (blue triangles), North Africa (green triangles). Genetic barriers detected by SAMOVA (black lines), F_{CT} ***=0.0770784.***

Pre-Indo-European Hypothesis

Biparental Markers

Classical Markers

To examine the place of the Basque population in Europe, phylogenetic analyses using all four previously described marker systems were performed. In addition, mismatch and intermatch analyses, along with genetic landscape analysis, were conducted. R-Matrix analysis of data on 14 European populations based on 21 blood group, red cell enzyme, and plasma protein alleles collected from the literature, and accounting for 62.5% of the total variation present in the sample, reveals that the

Basques are distinct from other European groups (Figure 41) due primarily to high frequencies of RH*cde (>50%), GM*Z,A;G (35%), ACP*B (73%), and ADA*1 (>97%) (Figure 42). Populations closest to the Basques are neighboring groups in France and Spain, plus populations in Britain and Ireland.

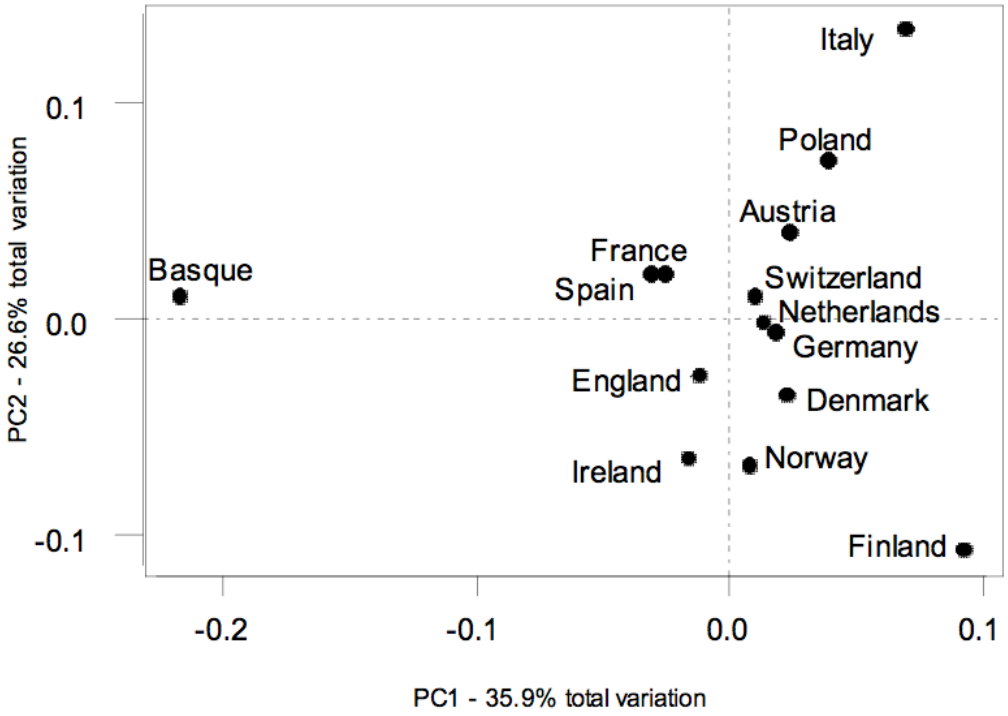

Figure 41. Principal component analysis of R-Matrix of 14 European populations using 13 loci (21 alleles). The first two principle components account for 62.5% of the total variation.

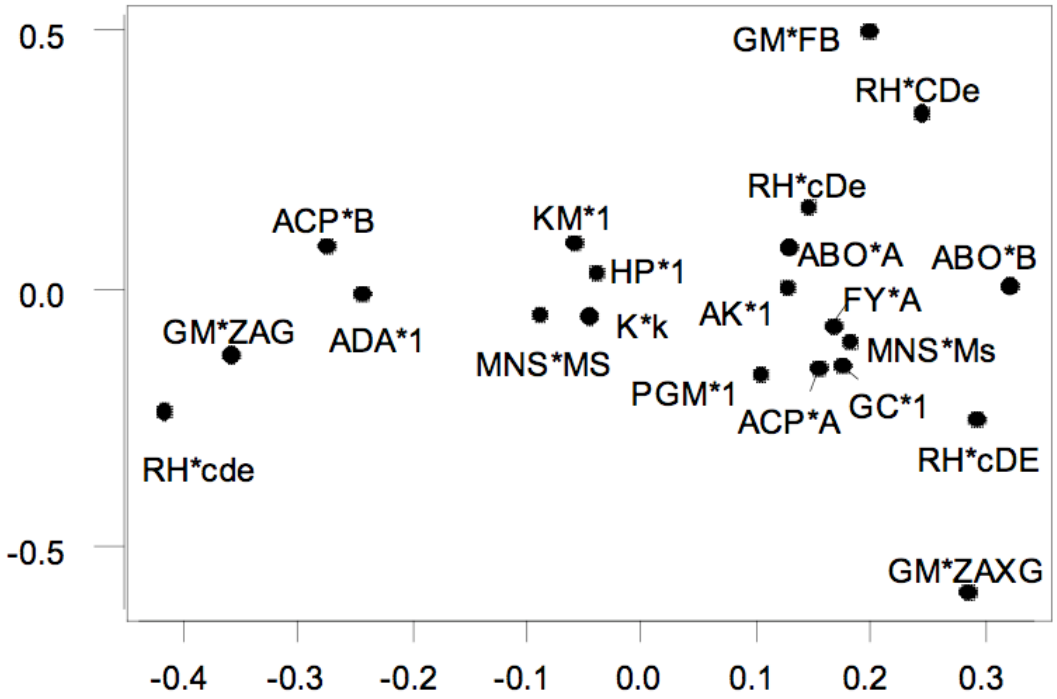

Figure 42. ***Plot of S-Matrix spread of alleles used in R-Matrix analysis. Basques are distinguished from other European populations by high frequencies of RH*cde (>50%), GM*Z,A;G (35%), ACP*B (73%), and ADA*1 (>97%).***

Heterozygosity vs. r_{ii} analysis using data from classical genetic markers in the same 14 European populations demonstrates that the Basques fall below the theoretical regression line, suggesting that they have experienced significant genetic drift but little gene flow (Figure 43), thus perhaps accounting for their genetic distinction.

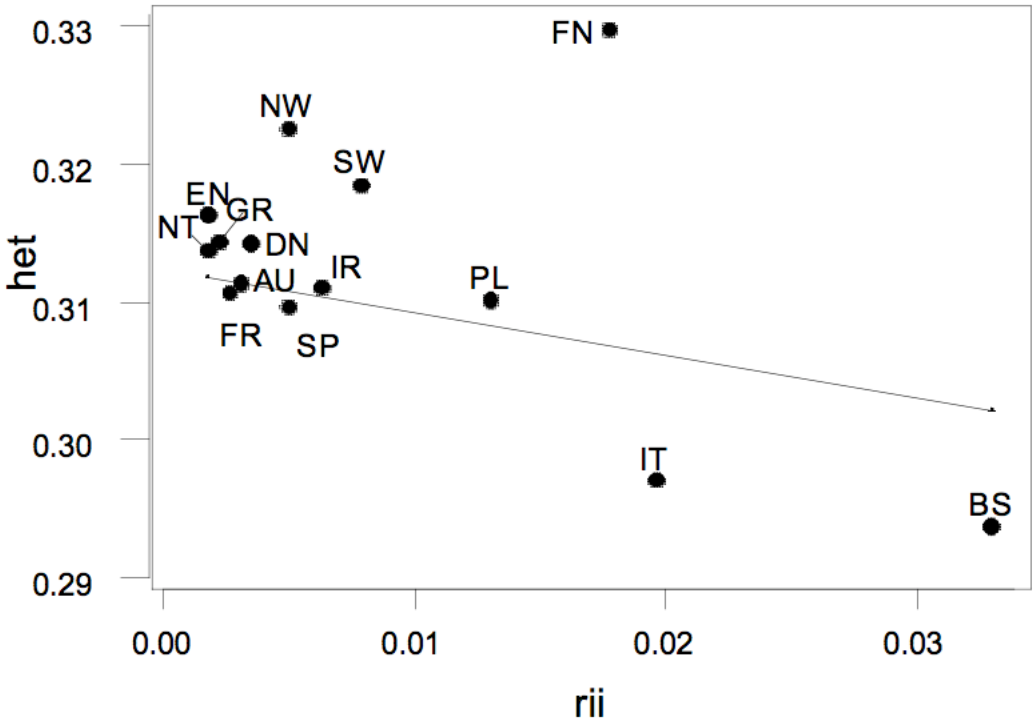

Figure 43. Heterozygosity vs. r_{ii} plot for 14 European populations. The Basques show the highest amount of genetic drift, with little gene flow.

Autosomal STRs

The results of the analysis of Shriver's D_{SW} genetic distances, calculated from autosomal STR data, into a bootstrapped Neighbor-Joining tree are shown in Figure 44. Included are 31 populations, the four Basque provinces from the present study, seven Iberian populations, 14 additional European groups, as well as two North African (Morocco and Egypt), three Middle Eastern (Oman, Yemen, and Turkish Caucasians), and the population from Georgia in the Caucasus. An R^2 value of 0.934 demonstrates that the tree is an accurate representation of the underlying population structure. The Basque Provinces cluster together (Guipuzkoa with Vizcaya, and Alava

with Navarre) in 71% of the 10,000 bootstrapped trees, along a branch encompassing other Iberian populations and Germany. The North African (Egypt and Morocco) and Middle Eastern (Yemen and Oman) populations form distinct branches at the opposite end of the tree, while Georgia clusters with Caucasians from neighboring Turkey. Other European populations form the branches between the Iberian Peninsula and the Middle East.

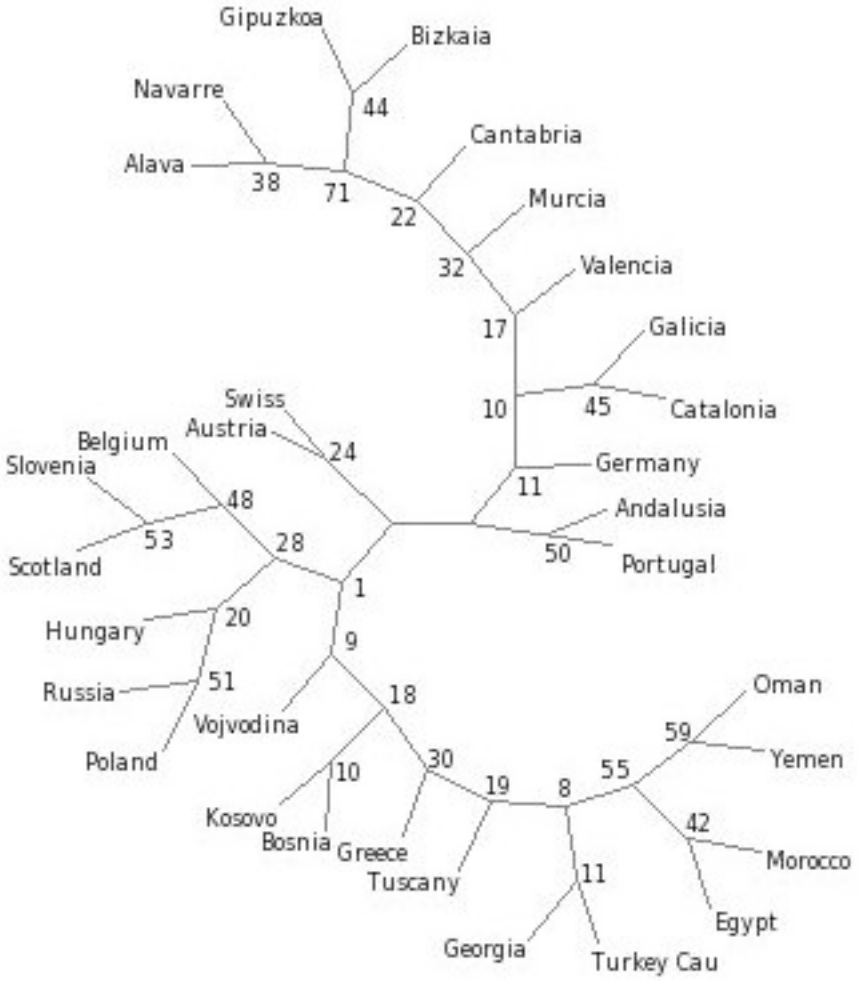

Figure 44. Neighbor-Joining tree of 31 populations based on D_{SW} genetic distances calculated from autosomal STR data. Numbers indicate branch percentages out of 10,000 bootstrapped trees. R^2=0.934.

Multidimensional scaling of the same distance matrix was also performed (Figure 45). The three dimensional plot accounts for 56.63% of the total variation in the sample, and an overall stress value of 0.13201 indicates a good fit between the original distance matrix and the projection of populations in three dimensions (p = 0.01). Comparison of the distance matrix and the MDS projection matrix using a Mantel test demonstrates a strong correlation (r=0.94984), denoting an accurate representation of the original data.

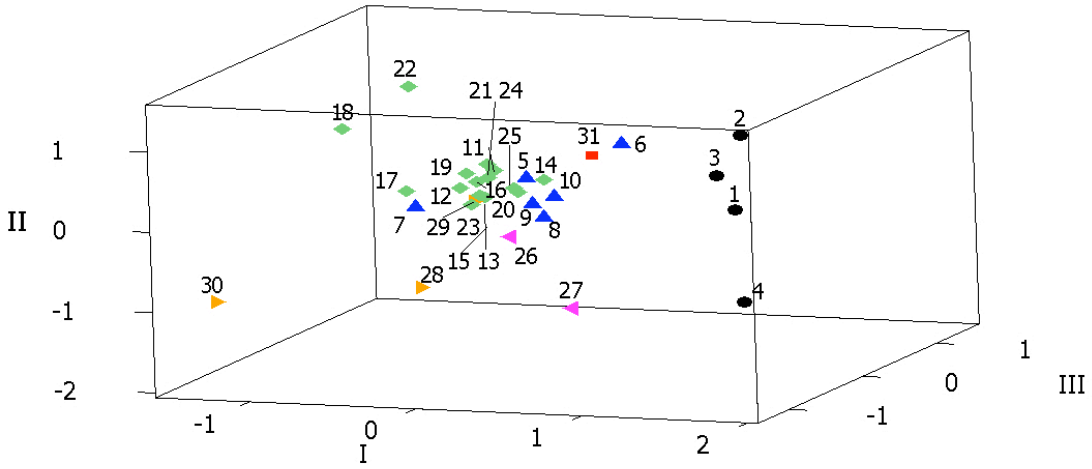

Figure 45. MDS plot of 31 populations using 9 autosomal STR loci. Basques (black circles), Iberian populations (blue triangles), other Europeans (green diamonds), Middle Eastern (orange triangles), North African (pink triangles), Caucasian (red square). STRESS = 0.13201, p = 0.01. Numbers refer to populations as listed in Table 23.

In Figure 45, the Basque Provinces are represented by black circles, other Iberian populations by blue triangles, and other European groups by green diamonds. Orange triangles indicate Middle Eastern populations, pink triangles North African groups, and the red square is the Caucasian population of Georgia. Numbers indicate

populations listed in Table 23. The Basque groups are separated along the first axis, distinct from the other populations. Iberians cluster near other European groups, with the exception of Cantabria (6), which is nearer to the Basque populations. Cantabrians are also closest to the Basques in the Neighbor-Joining tree. North African and Middle Eastern populations are separated from other groups along the second axis, with Yemen (30) and Morocco (21) being the most distinct. Georgians (31) cluster with other populations from Europe.

Examination of heterozygosity vs. distance from the centroid (r_{ii}) based on autosomal gene frequency data demonstrates a different pattern than the analysis of blood groups (Figure 46). Rather than showing evidence of significant drift, the molecular autosomal data reveal three provinces: Alava (BA-AV), Vizcaya (BA-VZ), and Guipuzkoa (BA-GP), which have relatively high r_{ii} values but also fall above the theoretical regression line, suggesting that the effects of genetic drift have been mediated somewhat by gene flow.

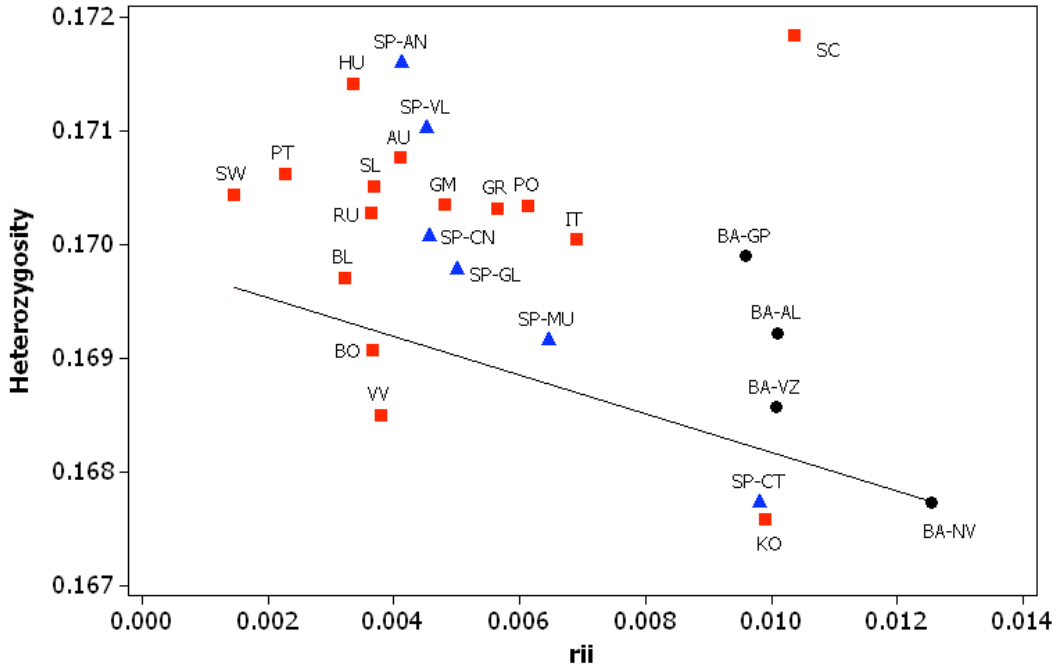

Figure 46. Plot of heterozygosity vs. distance from the centroid based on autosomal STRs. Basque groups shown as black circles, Iberian populations as blue triangles, and other Europeans as red squares. Three of the Basque provinces, Alava (BA-AL), Vizcaya (BA-VZ) and Guipuzkoa (BA-GP) fall above the theoretical regression line, while also having relatively high r_{ii} values, suggesting the effects of genetic drift have been mediated somewhat by gene flow.

Uniparental Markers

Y-STRs

A Neighbor-Joining tree of populations on the European continent based on Y-STR haplotype data are shown in Figure 47. The Basque groups are found at the top of the tree, along with the population from Ireland. The Iberian groups from Barcelona, Cantabria and Central Portugal form a distinct branch with Switzerland, while the other Iberian populations are scattered throughout the tree. The R^2 value for this tree is 0.922, which demonstrates a good fit between it and the genetic distance matrix.

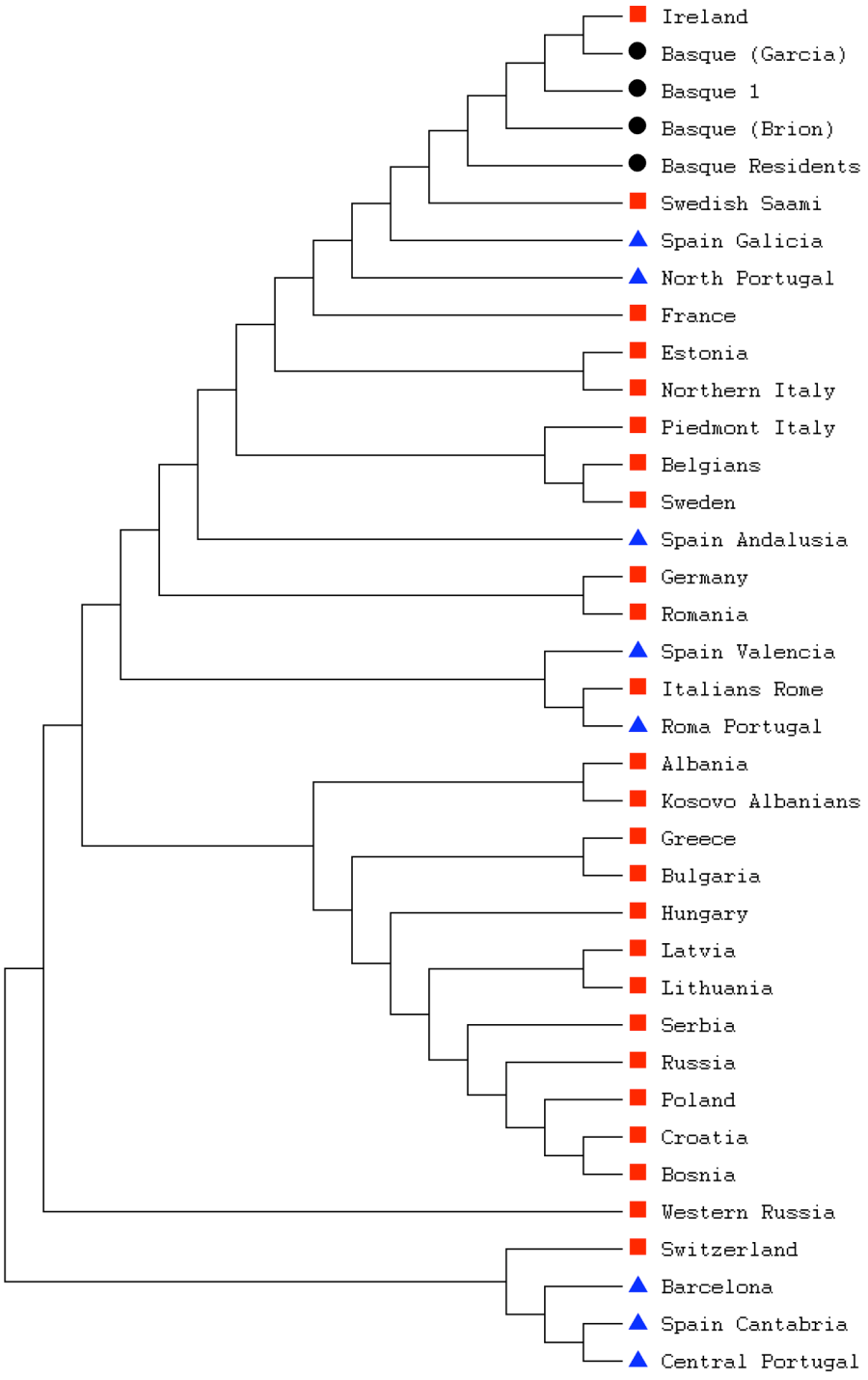

Figure 47. Neighbor-Joining tree of 37 European populations based on Y-STRs. Basque populations are indicated by black circles, other Iberian groups by blue triangles, and additional European populations by red squares. R^2=0.922.

Table 34. Results of SAMOVA of Y-STR haplotypes. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample

Europe	
K	F_{CT}
2	0.224858[a]
3	0.222243
4	0.217837
5	0.212403
6	0.175365
7	0.134018

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

Analysis of European populations using SAMOVA reveals the greatest genetic variance (F_{CT} = 0.224858) when the populations are divided into just two groups: (1) Barcelona, Cantabria, Central Portugal, and Switzerland (2) all other European populations, including the Basques (Table 34). The results of multidimensional scaling of the same Y-STR haplotype data from European populations are shown in Figure 48. The first two axes account for 79.93% of the total variation present in the sample, and the stress value (0.17431, p = 0.01) and Mantel correlation (r = 0.93612) demonstrate a good fit with the original distance matrix. The Basque groups cluster near the top right corner of the plot with the Irish population, with other European groups radiating from this point. The second axis distinguishes between Eastern and Western European groups, with many Slavic (*e.g.,* Russia, Poland, Bulgaria) and Baltic (Lithuania, Latvia) populations found in the bottom half of the plot. The Iberian populations, which were distinct in previous analyses (Central Portugal, Cantabria, and Barcelona), form a separate cluster in the top left corner. SAMOVA indicates that the population from Switzerland also belongs in this group.

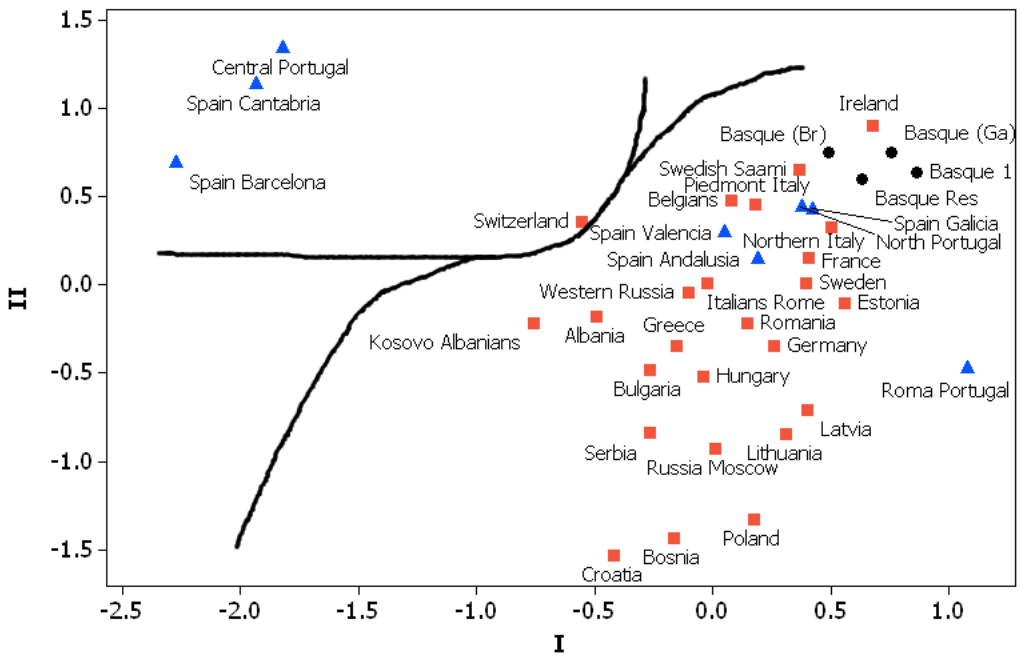

Figure 48. MDS plot of Y-STR haplotype pairwise genetic distances between 37 European populations. Basque groups are shown as black circles, Iberian populations as blue triangles, and other Europeans as red squares. Total variation in the first two axes = 79.93%, STRESS = 0.17431, r = 0.93612. Genetic barriers detected by SAMOVA (black lines), F_{CT} = 0.224858.

Figure 49 displays an interpolated genetic landscape of genetic distances between 28 European populations based on Y-STR data. In the plot, X and Y axes represent the geographic coordinates of the populations, while the Z axis shows differences between genetic distances. Areas of genetic similarity are indicated by red dips, while regions of genetic discontinuity are represented by blue peaks. There are two zones of genetic similarity in the figure. One separates the Northern European populations of Sweden and Estonia from all other groups. The second is surrounded by small peaks of genetic dissimilarity, and primarily separates populations in Iberia and the Mediterranean from the rest of Europe. The Basque populations and the Northwestern Iberian population of Galicia can been seen concentrated around that

area of genetic similarity, indicating a high degree of homogeneity among these groups.

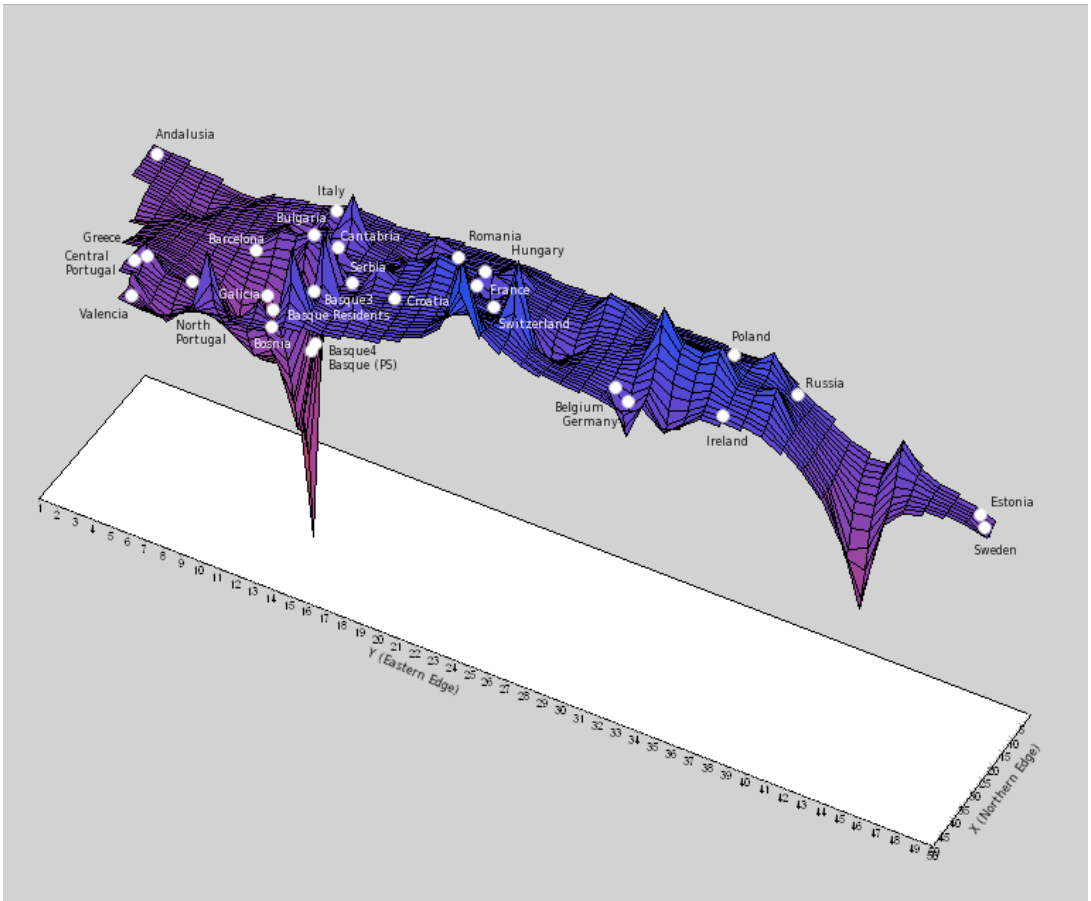

Figure 49. Interpolated genetic landscape based on Y-STR data in28 European populations. X and Y axes represent geographic coordinates, while the Z axis indicates variation in genetic distance. Red dips indicate areas of high genetic similarity, while blue points are indicative of genetic dissimilarity.

mtDNA sequences

Construction of a Neighbor-Joining tree from a Kimura 2P genetic distance matrix calculated from mtDNA sequence data are shown in Figure 50. The Basques form their own branch off a cluster with the northern Spanish province of Galicia as well as the Welsh population. Other Iberian populations are scattered throughout the tree. This particular tree had the lowest R^2 value for any of the NJ trees constructed

(0.699), indicating that it is not an accurate representation of the underlying population structure. A much better representation is seen in the MDS plot with SAMOVA genetic barriers (Figure 51), which accounts for 74% of the total variation. Unlike the NJ tree, the STRESS (0.15897, $p = 0.01$) and Mantel correlation ($r = 0.89094$) indicate a good fit with the original distance matrix. In this plot, the Basques are distinct in Europe, and this placement is confirmed by SAMOVA where the European populations show the highest F_{CT} (0.01016410) when divided into just two groups: (1) the Basques, and (2) all other European populations (Table 35). In the NJ tree, the Basques formed a branch off a cluster containing populations from Galicia, Ireland and Wales. These populations are near the Basques in the MDS plot, but SAMOVA analysis demonstrates that they actually form a cluster with all the other European groups.

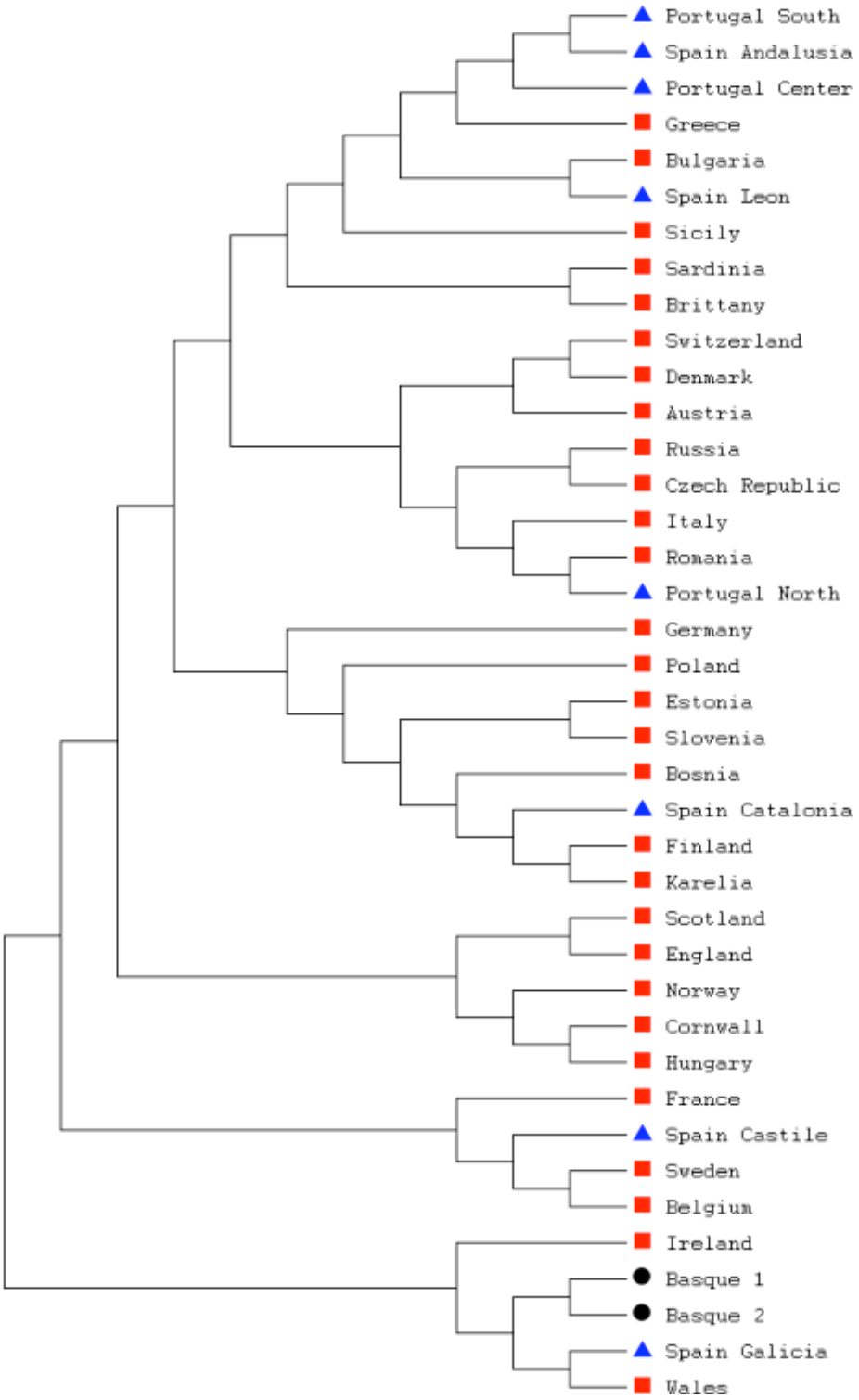

Figure 50. Neighbor-Joining tree of 39 European populations using mtDNA HVS-I data. Basques (black circles), Iberia (blue triangles), other European (red squares). R^2=0.699.

Table 35. Results of SAMOVA mtDNA HVS-I sequences in European populations. K = number of groups, F_{CT} = variance among groups relative to total variance in the sample.

Europe	
K	F_{CT}
2	0.01016410[a]
3	0.00754304
4	0.00935746
5	0.00875238
6	0.00802244
7	0.00747620

[a]Highest F_{CT} = greatest genetic variance between the number of groups (K).

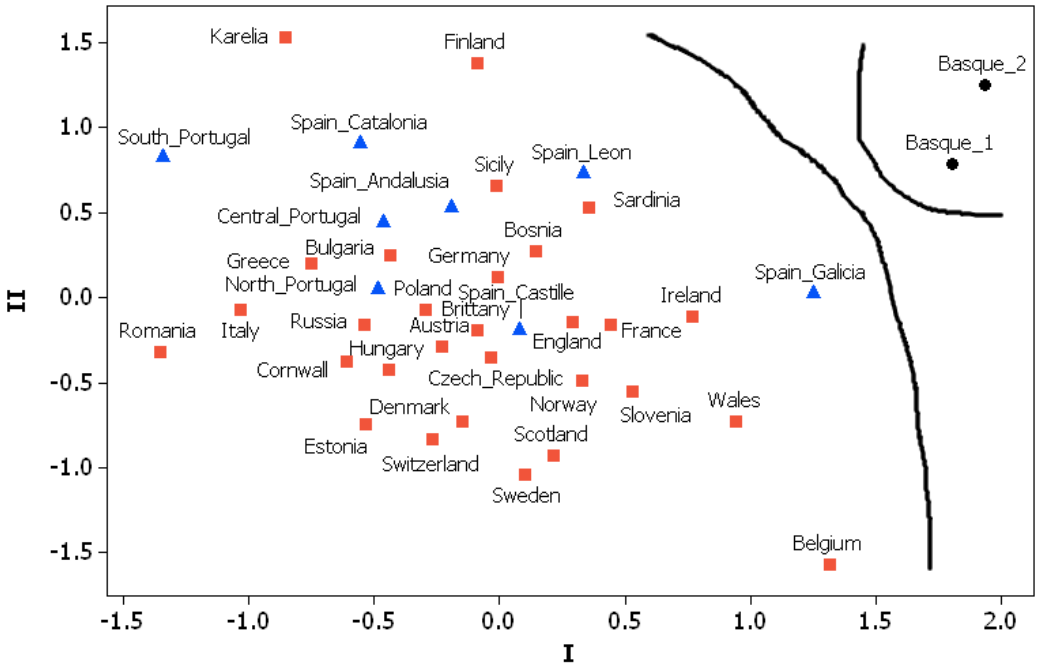

Figure 51. MDS plot of 39 European populations using mtDNA HVS-I data. Basques (black circles), Iberian populations (blue triangles), other Europeans (red squares). STRESS = 0.15879 (p=0.01), r = 0.89094).

The interpolated genetic landscape based on genetic distances calculated from HVS-I sequence data in 35 European populations is presented in Figure 52. Broader regions of genetic similarity are seen in the mitochondrial data, as indicated by the purple areas on the plot. Two areas of strong genetic similarity are seen. The first

encompasses England, Ireland, Wales, and Germany. As with the genetic landscape based on Y-STR data, the second dip occurs in Northern Iberia, including the Basque populations along with groups from Galicia and Catalonia.

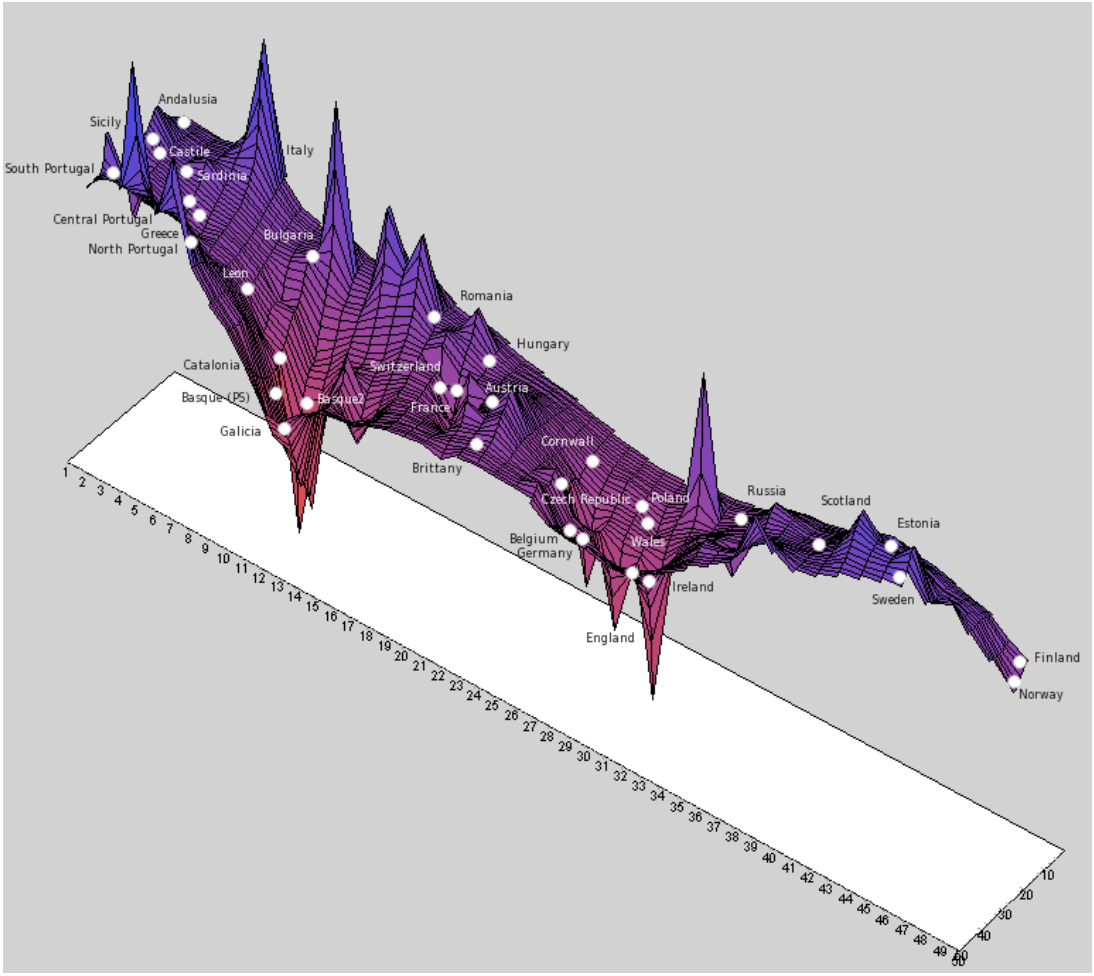

Figure 52. Interpolated genetic landscape based on HVS-I sequence data in 35 European populations. X and Y axes represent geographic coordinates, while the Z axis indicates differences in genetic distance. Red dips show areas of genetic similarity, while blue peaks display areas of genetic differentiation.

HVS-I sequence data were also used to perform mismatch analyses for all 39 European populations, and intermatch comparisons were made between the study population and other comparative groups, in order to estimate the time of population

expansion within groups, and coalescent times between groups. Tau (τ) values based on a least squares procedure, along with associated 95% confidence intervals, Harpending's raggedness index, and time estimates, are presented in Table 36. In all populations, low raggedness values (r) indicate expansion (0.0083-0.0367). Tau values range from 1.50 in Wales to 4.90 in Italy, while the two Basque populations have tau estimates of 2.90 and 3.20 (present study). These estimates translate into expansion times between 16,469-53,799 years for Europe, and 31,840-35,134 years among the Basques.

Table 36. Tau values and time estimates based on a least squares analysis of mismatch distributions in 39 European populations.

Population	*N*	τ	*95% CI*	*r*	*Time Estimate*	*Date Range*
Basque (Present Study)	131	3.20	2.166-3.750	0.0246	35,134	23,781-41,173
Austria	99	4.00	2.072-7.330	0.0084	43,917	22,749-80,479
Basque	156	2.90	1.551-3.777	0.0278	31,840	17,029-41,469
Belgium	33	3.00	1.994-4.008	0.0367	32,938	21,893-44,005
Bosnia	144	3.00	1.562-6.467	0.0173	32,938	17,150-71,003
Brittany	62	3.30	1.189-8.375	0.0126	36,232	13,055-91,952
Bulgaria	141	3.30	2.766-3.754	0.0137	36,232	30,369-41,217
Central Portugal	162	2.20	0.758-8.320	0.0101	24,155	8,322-91,348
Cornwall	92	2.20	1.697-2.777	0.0188	24,155	18,632-30,490
Czech Republic	83	3.80	3.076-5.311	0.0195	41,722	33,773-58,311
Denmark	38	3.20	1.209-7.609	0.0145	35,134	13,274-83,542
England	242	3.00	1.449-7.500	0.0118	32,938	15,909-82,345
Estonia	149	3.10	1.869-6.789	0.0125	34,036	20,520-74,539
Finland	153	3.90	1.723-5.744	0.0171	42,819	18,917-63,065
France	379	2.80	1.809-6.111	0.0146	30,742	19,862-67,095
Germany	582	3.30	2.428-4.951	0.0171	36,232	27,251-54,359
Greece	179	3.00	2.539-3.438	0.0169	32,938	27,877-37,747
Hungary	78	2.90	2.186-3.539	0.0186	31,840	24,001-38,856
Ireland	300	3.40	1.975-5.271	0.0155	37,330	21,684-57,872
Italy	248	4.90	2.746-6.059	0.0113	53,799	30,149-66,524
Karelia	83	2.70	1.365-5.773	0.0157	29,644	14,987-63,384
North Portugal	183	3.90	1.787-8.381	0.0083	42,819	19,620-92,018
Norway	629	3.20	1.863-6.439	0.0115	35,134	20,455-70,696
Poland	473	3.30	1.279-10.848	0.0101	36,232	14,043-119,104

Population	*N*	τ	*95% CI*	*r*	*Time Estimate*	*Date Range*
Romania	92	3.10	2.508-3.756	0.0174	34,036	27,536-41,238
Russia	379	3.50	2.215-6.023	0.0128	38,428	24,319-66,129
Sardinia	115	3.90	1.879-6.027	0.0125	42,819	20,630-66,173
Scotland	895	3.50	2.338-6.072	0.0123	38,428	25,669-66,667
Sicily	196	3.40	1.318-6.863	0.0093	37,330	14,471-75,351
Slovenia	104	2.40	1.873-2.920	0.0200	26,350	20,564-32,060
South Portugal	195	2.70	0.900-8.486	0.0077	29,644	9,881-93,170
Spain Andalusia	114	3.90	1.900-7.652	0.0091	42,819	20,861-84,014
Spain Castile	38	3.90	2.111-5.113	0.0202	42,819	23,177-56,137
Spain Catalonia	61	3.90	1.541-7.033	0.0163	42,819	16,919-77,218
Spain Galicia	135	1.70	1.309-2.195	0.0145	18,665	14,372-24,100
Spain Leon	61	2.40	0.967-7.842	0.0137	26,350	10,617-86,100
Sweden	32	3.80	2.740-5.865	0.0163	41,722	30,083-64,394
Switzerland	224	3.80	2.312-5.109	0.0164	41,722	25,384-56,094
Wales	92	1.50	0.541-6.420	0.0170	16,469	5,940-70,487

N = ***number of mtDNA sequences in the sample,*** τ ***= expansion time in mutational units,*** **r** ***= raggedness index, CI = confidence intervals for*** τ

Tau values estimated by a method of moments, along with time estimates and divergence times between the Basques in the present study and comparative European populations are presented in Table 37. Tau values range from 1.21 in Central Portugal to 5.36 in Scotland, with tau estimates among the Basque between 2.64 and 3.15 (present study). These estimates give expansion times between 13,326-58,847 years in Europe, and 28,958-34,599 years for the Basques. While the two methods of determining expansion times provide somewhat different values, all time estimates based on the method of moments fall within the date ranges based on the least squares approach. Divergence times between the Basques and other European populations suggest a common maternal ancestor between 16,967-47,377 years ago.

Table 37. Tau values and time estimates in 39 European populations, based on a method of moments analysis of mismatch and intermatch distributions.

Population	*Sample Size*	τ	*Time Estimate*	*Divergence Time*
Basque (Present Study)	131	3.15	34,559	--
Austria	99	3.04	33,376	33,114
Basque	156	2.64	28,958	32,635
Belgium	33	3.60	39,509	37,751
Bosnia	144	2.77	30,447	32,276
Brittany	62	3.16	34,656	35,040
Bulgaria	141	3.38	37,110	37,844
Central Portugal	162	1.21	13,263	16,967
Cornwall	92	2.54	27,921	31,270
Czech Republic	83	4.15	45,511	40,407
Denmark	38	2.34	25,672	28,629
England	142	2.90	31,835	32,210
Estonia	149	3.25	35,654	36,474
Finland	153	3.25	35,704	38,130
France	71	3.01	33,047	32,604
Germany	109	3.10	34,002	36,398
Greece	179	3.26	35,753	37,386
Hungary	78	3.19	34,983	33,055
Ireland	157	4.74	51,999	43,654
Italy	83	3.70	40,624	36,840
Karelia	83	2.88	31,599	34,111
North Portugal	183	2.98	32,743	31,374
Norway	74	3.47	38,060	35,891
Poland	37	2.62	28,813	34,365
Romania	92	3.55	38,959	36,653
Russia	50	3.79	41,656	37,939
Sardinia	115	2.67	29,294	29,914
Scotland	100	5.36	58,847	47,377
Sicily	196	2.36	25,896	28,740
Slovenia	104	2.59	28,428	31,210
South Portugal	195	2.58	28,305	27,592
Spain Andalusia	114	3.41	37,419	33,390
Spain Castile	38	3.77	41,410	38,258
Spain Catalonia	61	2.48	27,192	29,719
Spain Galicia	135	1.86	20,406	24,332
Spain Leon	61	2.19	24,099	26,923
Sweden	32	4.40	48,300	41,930
Switzerland	154	3.87	42,526	41,149
Wales	92	1.94	21,268	26,789

CHAPTER FIVE: DISCUSSION

Basque Heterogeneity

There has been debate in the literature concerning whether the Basques can be considered a single, homogenous population. While many studies treated them as such, and results of some research found no significant differences among Basque localities (Francalacci *et al.* 1996; Iriondo *et al.* 1996; Levine 1977; Pena *et al.* 1997), other analyses have observed the presence of heterogeneity both within and among provinces (Aguirre *et al.* 1989b; Manzano *et al.* 1996a; Manzano *et al.* 1993a; Vergnes *et al.* 1980). A study of 16 classical markers noted heterogeneity among provinces ($\chi^2 = 263.4$, $p = 0.01$) (Manzano *et al.* 1996a). Guipuzkoa and Alava differed in allele frequencies at four loci (Rh, K, ESD, and TF), Guipuzkoa and Vizcaya differed at six loci (Rh, P, Pi, ACP, AK, and HP), and Vizcaya and Alava showed significant heterogeneity values at six loci (ABO, Rh, P, ACP, ESD, TF). No explanation for this observed heterogeneity was provided in this case, though genetic drift between isolated valleys seems likely. Subsequent study of 2,414 Basques distributed by districts within provinces for 18 classical genetic markers demonstrated heterogeneity in only two loci, Rh ($\chi^2 = 145.18$, $p = 0.0000$) and MNS ($\chi^2 = 100.64$, $p = 0.0000$) (Manzano *et al.* 2002). No comparison of heterogeneity among provinces was reported. Analyses of autosomal STRs is Vizcaya found a sizeable D_A distance between that province and neighboring Guipuzkoa, although average gene diversity

values (0.796 in Guipuzkoa and 0.807 in Vizcaya) did not differ greatly (Zlojutro *et al.* 2006). Analysis of 6 autosomal STR loci (29 alleles) in 17 Basque districts in Spain found significant heterogeneity at two loci (HUMCSF1PO and HUMTPOX), attributed to pre-Neolithic genetic drift because the differences detected would have occurred when the populations were small and more sensitive to the effects of drift. However, after correcting for possible spatial autocorrelation and multiple tests, only one allele in 29 (CSF1P0*12, $p = 0.001$) was significantly different among groups of districts (Iriondo *et al.* 2003). Analysis of seven red cell enzymes demonstrated a lack of heterogeneity between the regions of Vizcaya Province. Only one locus, adenylate kinase (AK), showed significant heterogeneity between regions, with a range of 0.897-0.975 and an F_{ST} value of 0.012. This result was attributed to random genetic drift which did not have the same effect on all markers, with the caveat that the heterogeneity detected in the Basques might have been a result of the sampling strategy, with more internal subdivision defined for the Basques than for comparative populations. (Aguirre *et al.* 1991a). Locus-by-locus AMOVA of four HLA loci among Basque groups in Spain and France demonstrated that a majority of variation was accounted for by variation between individuals, while variation by province comprised only 1.62% of the total. Division of the Basque groups by country of origin showed no variation between the Basque populations of France and Spain (Comas *et al.* 1998a). Examination of the HLA*DQA1 locus in Navarre revealed no significant differences in allele frequencies between that province and Guipuzkoa ($\chi^2 = 10.19$, NS) or Vizcaya ($\chi^2 = 14.89$, $p = 0.061$), but did note significant

heterogeneity between Navarre and Alava ($\chi^2 = 19.39$, $p < 0.01$), which was attributed to the high frequency of a single allele (DQA1*02) (Perez-Miranda *et al.* 2003). AMOVA analysis of autosomal STRs showed significant variation among individuals within the four Basque Provinces in Spain ($F_{ST} = 0.0015$, $p = 0.0052$), but heterogeneity between provinces was not reported (Perez-Miranda *et al.* 2005a). Analysis of Y-chromosome STR haplotypes demonstrated significant heterogeneity between published data from Guipuzkoa and populations in Vizcaya ($F_{ST} = 0.02631$, $p = 0.04346$) and Guipuzkoa ($F_{ST} = 0.02991$, $p = 0.04495$) (Alonso *et al.* 2005; Bosch *et al.* 2001). Comparisons with additional Basque groups, 29 individuals from Brion *et. al.* (2003) of unknown geographic origin and 22 from Alava and Navarre, found no significant differences between either the study populations or the second Guipuzkoa group. After Bonferroni correction, no heterogeneity between any of the Basque samples is observed for the Y-STR haplotypes belonging to haplogroup R1b, and the heterogeneity present in the study sample is explained by a higher proportion of R1b in that province, rather than differentiation within the haplogroup itself. Y-STR haplotype analysis of Iberian population including 29 Basques found no increased homozygosity, but did report slightly lower haplotype diversity in this population (0.96, range of other Iberian groups: 0.97-0.99) (Gonzalez-Neira *et al.* 2000). For mtDNA sequence data, preliminary comparison between Alava and Vizcaya found no significant genetic differentiation between them (Corte-Real *et al.* 1996), but genetic diversity was lower in these provinces (2.76-3.24) than in other regions of Spain (3.53-5.77).

In the present study, no significant heterogeneity was detected between provinces in the locus-by-locus AMOVA of autosomal STRs (V_a = -0.095, F_{CT} = -0.0036, p = 0.878). An excess of homozygotes was confirmed for only one locus (D8S1179) in the provinces of Alava, Vizcaya, and Guipuzkoa, after correction for allelic dropout. Though a small but significant amount of microdifferentiation was detected among villages within provinces (1.309%, F_{SC} = 0.0131, p = 0.011), the majority of variation within the Basque sample (99.045%, F_{ST} = 0.0096, p = 0.019) is accounted for by differences between individuals. This suggests little genetic structure between Basque groups, and the possibility that recent gene flow between provinces has obscured the effects of drift. This hypothesis supported by the analysis of heterozygosity versus r_{ii} for autosomal STR data, where three of the four Basque Provinces fall above the theoretical regression line while still have relatively high r_{ii} values. AMOVA analysis of Y-STR haplotypes and mtDNA sequence data both confirm small but significant differences between provinces (Y-STRs: 1.71%, p = 0.0369; mtDNA: 1.03%, ϕ_{ST} = 0.0103, p = 0.0308), but the majority of variation is again comprised of differences between individuals (Y-STRs: 98.29%, mtDNA: 98.97%), suggesting little population subdivision between provinces. It appears that, in general, the literature and the present analysis agree that there is some minor detectable heterogeneity between Basque populations, what differs is how much significance is placed on variation found at one or a few loci. Given that estimates of F_{ST} are affected by the number of population subdivisions and effective population size, it seems wise to interpret these studies with caution (Cavalli-Sforza and Feldman

1990). In fact, meta-analysis of classical marker frequencies found no significant heterogeneity among Basques, once the sampling intensity of the region (7.28 times the rest of the Iberian Peninsula) had been accounted for (Calafell and Bertranpetit 1994c). After all, it is not unexpected that there would be some genetic substructure present among the Basques, given the impact of genetic drift on small, isolated villages, as shown by the analysis of heterozygosity vs. r_{ii} for classical markers in the present work. But these results are overall found to be a consequence of higher frequencies of certain alleles in a particular province, rather than the presence of different alleles.

Basque Origins

Basque-Caucasian Hypothesis

Examination of the literature on the Basque-Caucasian hypothesis demonstrates little support from the genetic evidence, not from analysis of classical markers (Aguirre *et al.* 1991a; Bertorelle *et al.* 1995), HLA alleles (Sanchez-Velasco and Leyva-Cobian 2001), Y-chromosome haplogroups (Nasidze *et al.* 2003), or mtDNA sequences (Nasidze *et al.* 2004; Nasidze and Stoneking 2001). The results of the present study agree with those previously published.

None of the molecular system analyzed corroborate a relationship between Basques and Caucasian populations. Examination of classical markers show that populations of the Caucasus do not cluster near Basque groups, neither in R-Matrix or MDS analyses. The bootstrap neighbor-joining tree of autosomal STR data cluster the Basque provinces together in 71% of trees, while Georgians cluster with Turks. Y-

STR analysis groups some Iberian populations (Cantabria, Barcelona, Central Portugal) with Caucasian groups (Azerbaijan, Darginians, etc.), but the Basques form a second cluster with other Iberian populations (Galicia, Valencia, North Portugal) and SAMOVA analysis confirms a genetic barrier between the Basques and Caucasian groups. Phylogeographic analysis of mitochondrial sequences clusters several Iberian groups (Portugal, Andalusia, Castile, etc.) with residents of Azerbaijan, Armenia, and Georgia). A genetic barrier exists between this cluster and the Basques, and the Caucasian populations of North Ossetia and the Adygei are singled out by additional genetic barriers. None of the results of the present study support the hypothesis of a common non-Indo-European ancestor for the Basques and Caucasians.

Vasco-Iberian Hypothesis

The Vasco-Iberian hypothesis of Basque origins, based primarily on the analysis of HLA alleles (Arnaiz-Villena *et al.* 1995; Arnaiz-Villena *et al.* 2002; Arnaiz-Villena *et al.* 2001; Arnaiz-Villena *et al.* 1997a; Arnaiz-Villena *et al.* 1981; Martinez-Laso *et al.* 1995a), has found little support in the literature (Bosch *et al.* 2001; Bosch *et al.* 1997; Brion *et al.* 2003). Studies which purport to demonstrate a relationship between Basques and North African groups have proven to be statistically weak (Sanchez-Velasco *et al.* 2003), while other analyses based on different HLA loci and Alu polymorphisms have found no relationship between North Africans and Basques (Arnaiz-Villena *et al.* 1999; Brown *et al.* 2000; Comas *et al.* 2000; Comas *et al.* 1998b; Garcia Fernandez *et al.* 1997b; Perez-Miranda *et al.*

2003).

In the present study, analysis of classical markers demonstrated the genetic difference between Algeria and all European populations, including the Basques. Preliminary investigation of autosomal STRs in Vizcaya indicated similarity with the Basque province of Guipuzkoa, and distinction from North African groups in Morocco and the Maghreb (Zlojutro *et al.* 2006). The neighbor-joining tree based on autosomal STRs groups North African populations in Egypt and Morocco with Middle Eastern groups in Yemen and Oman in 55% of the bootstrapped trees, while the Basques are found at the opposite end of 71% of bootstrapped trees. Analysis of uniparental markers exhibit the distinctiveness of North African groups, with a barrier separating Mozabites, Saharawis and Moroccan Berbers from Iberian populations in North Portugal, Galicia, Valencia, and four Basque populations based on Y-STRs. This is in agreement with previous studies, which showed a low frequency of haplogroup R1b in the Maghreb population, clearly distinguishing this North African group from Iberian populations (Brion *et al.* 2003). The addition of several more North African populations in the present analysis of mtDNA sequences resulted in genetic barriers separating Tata Berbers from Algerian Mozabites from Sened Berbers and Moroccan Arab. The final group comprised all Iberian populations and four North African groups (Algeria, Tunisia, Moroccan Berber, and Matmata Berber). In this final cluster, the two Basque groups are separated from the North African populations by all other Iberian groups. In addition, the North African subhaplogroup U6 was not detected in the current sample (Gonzalez *et al.* 2003;

Maca-Meyer *et al.* 2001; Maca-Meyer *et al.* 2003). Once again, the present study does not lend credence to this hypothesis of Basque origins.

Pre-Indo-European Hypothesis

Autosomal Markers

Genetic analyses of classical markers have confirmed the distinctiveness of the Basques in Europe, which is generally attributed both to their being non-Indo-European and undergoing a long period of *in situ* differentiation (Barbujani *et al.* 1995; Bertranpetit and Cavalli-Sforza 1991a; Calafell and Bertranpetit 1993a; Calafell and Bertranpetit 1994c; Cavalli-Sforza and Piazza 1993; Derish and Sokal 1988; Harding and Sokal 1988; Sokal 1988; Sokal 1991a; Sokal 1991b; Sokal *et al.* 1989a; Sokal *et al.* 1993; Sokal and Menozzi 1982; Sokal *et al.* 1990; Sokal *et al.* 1989b; Sokal *et al.* 1988; Sokal *et al.* 1992a; Sokal *et al.* 1999a; Sokal *et al.* 1999b; Sokal *et al.* 1996; Sokal *et al.* 1991; Sokal *et al.* 1992b). However, additional analyses have noted that Basques are genetically more similar to neighboring Indo-European groups (specifically Béarn) than to other non-Indo-European populations (Piazza *et al.* 1988a). In the present study, analysis of classical markers reveals that some Basque groups are outliers at 5 markers systems (ABO, MNS, Rh, AK, and GM), and are separated from other European groups in the R-Matrix analysis as a result of high frequencies of RH*cde and GM*Z,A;G. For molecular systems, a similar pattern emerges (Comas *et al.* 2000). The neighbor-joining tree based on autosomal STRs groups the Basque Provinces on a branch including all other Iberian populations. Additional populations in Northern and Eastern Europe form a separate

branch with Caucasian, North African, and Middle Eastern groups. Biparental markers suggest that the Basques are distinct in Europe, but still part of the European genetic landscape.

Y-Chromosome

Previous analysis of Y-haplogroups in the Basque country found 86% R1b (Alonso *et al.* 2005), which is consistent with the present study, where 83% of the Y-chromosomes were identified as this most common Western European haplogroup. Prior research has also reported the presence of haplogroups E, G2, and I in the Basque population and noted the absence of J in Guipuzkoa (Brion *et al.* 2003; Cruciani *et al.* 2004; Flores *et al.* 2004; Rootsi *et al.* 2004). The present analysis found haplogroup E1b1b in all four provinces and haplogroups G2a (1.6%) and I2 (3%) in Guipuzkoa. Haplogroup J was absent in Guipuzkoa, though J2a was detected in both Alava (2%) and Vizcaya (1.6%). In addition, Guipuzkoa also had a low frequency of haplogroup L (0.8%), which has not been reported previously among the Basques. This individual, from the town of Azpeitia, has mitochondrial haplogroup H. It is possible that this represents mixed Basque descent, or that haplogroup L is found at low frequencies throughout Southern Europe, including among the Basques.

Haplogroup R1b shows the highest frequencies in Western European populations, including France (52.2%), the Netherlands (70.4%), Germany (50%), Italy (62%), Denmark (41.7%), and Britain (68.8%) (Pericic *et al.* 2005; Semino *et al.* 2000a). In the British Isles, this haplogroup is found in 79-82% of males, while in Iberia it ranges from 56% in Portugal to 68% in Spain (Rosser *et al.* 2000). Also

found in haplogroup R1b is the Atlantic Modal Haplotype (AMH) -- defined as DYS19*14-DYS390*24-DYS391*11-DYS392*13-DYS393*13 (Wilson *et al.* 2001) --which is present at a frequency of 50% in the present Basque sample and found at frequencies ranging from 44% in Ireland to 70% in Wales. R1b is considered a Western European-specific haplogroup which diverged during the Upper Paleolithic (See Table 38), with the high frequencies found along the Atlantic Fringe being attributed to the results of genetic drift during the Last Glacial Maximum (LGM) (Alonso *et al.* 2005; Quintana-Murci *et al.* 1999; Semino *et al.* 2000b; Wells *et al.* 2001). Haplogroup I2a2 is reported among Basques (6%), the populations of Castile (19%), Bernais (7.7%), and Normandy (2.4%), as well as the Irish (2.6%), with a high of 40.9% in Sardinia (Flores *et al.* 2004; Rootsi *et al.* 2004). This European-specific haplogroup is believed to have originated in the Pyrenees prior to the LGM (Karafet *et al.* 2008; Lopez-Parra *et al.* 2009; Novelletto 2007; Rootsi *et al.* 2004). Haplogroup E1b1b has been reported in many other European populations, including Portugal (4.0%), Spain (3.2%), France (4.7%), Italy (10%), and Sardinia (3.5%). Among the French Basques, the frequency of E1b1b was 6.3% (Cruciani *et al.* 2004). Haplogroup E1b1b has a complex history, with evidence of several demographic expansion events (Cruciani *et al.* 2004; Cruciani *et al.* 2006; Cruciani *et al.* 2007). The most common variant of this haplogroup in Europe is defined by mutation E-M78, which is believed to have originated in the Horn of Africa, from which it spread to the Middle East and then into Southern Europe during the Neolithic (Semino *et al.* 2004). Haplogroup G2a has been reported in Italy (10%), Sardinia (14.1%), Catalonia

(8.3%), and Andalusia (2%), and reaches its highest frequency in Palestine (75%) (Francalacci *et al.* 2003; Semino *et al.* 2000a; Shen *et al.* 2004). Like E1b1b, G2a is considered a Neolithic Y-haplogroup (Cinnioglu *et al.* 2004), which spread from the Middle East into Europe with the advent of agriculture. Haplogroup J2a has been previously reported among the French Basques (13.6%), French (13%), Italians (12-16%), and Greeks (21%), with higher frequencies in Turkey (40%) and Lebanon (29%) (Semino *et al.* 2004; Semino *et al.* 2000a). It is also considered a marker of Neolithic expansion (Capelli *et al.* 2007; Rosser *et al.* 2000), but a more recent maritime route for the distribution of J2a from the Middle East through the Mediterranean has been proposed (Di Giacomo *et al.* 2004). Haplogroup L has recently been divided into three subclades, L1, L2, and L3. L1 is found at moderate frequency in India (6.3%), while L3 is more frequent in Pakistan (6.8%), and the majority of European and Anatolian L samples are believed to be L2 (Sengupta *et al.* 2006). Haplogroup L2 has been detected in several other European groups, including Andalusia (3.4%), Italy (5.4%), Greece (1.3%), and Hungary (2.2%). It is also found at low frequencies in Turkey (1.6 – 4.2%) and Syria (3.2%) (Cinnioglu *et al.* 2004; Semino *et al.* 2000a). The parental haplogroup L originated in East Africa around 30,000 years ago, while L2 dates to 14,600 years ago in Turkish populations assuming a model of continuous growth (Cinnioglu *et al.* 2004). Evidence from the Y-chromosome supports the Pre-Indo-European hypothesis, as all of the Y-haplogroups present in the Basque population are found in other European populations, though the Basques do have a higher frequency of R1b.

Table 38. Divergence time estimates for various European Y-chromosome haplogroups.

Haplogroup	*Divergence Time*	*Reference*
R1b (Basques)	17,900 – 21,300 BP	Alonso *et al.* (2005)
E1b1b	7,000-14,000 BP	Semino *et al.* (2004)
G2a	9,500-17,000 BP	Cinnioglu *et al.* (2004) Semino *et al.* (2000)
I2a2 (Pyrenees)	11,900 BP	Lopez-Parra *et al.* (2009)
J2a	4,700-10,000 BP	Di Giacomo (2004)
L2 (Turkey)	14,600 BP	Cinnioglu *et al.* (2004)

Mitochondrial DNA

Previous haplogroup analysis of Basque populations reported a preponderance of haplogroup H/HV (74%), with the addition of haplogroups U (11%), T (6%), K (5%), J (3%), and I/W/X (2%) (Achilli *et al.* 2004; Bertranpetit *et al.* 1995; Corte-Real *et al.* 1996). Historical populations (6th and 7th centuries AD) in the Basque region had lower frequencies of H (42%), but higher frequencies of U (16%), K (17%), T/X (12%), and J (12%) (Alzualde *et al.* 2005). Haplogroup V had a frequency of 0.68% in the historic population, much lower than the frequencies reported in modern Basque groups (3-20%) (Torroni *et al.* 2001). Prehistoric populations in the region, dating between the Neolithic and the Bronze Age, also have lower frequencies of H (33%), but higher frequencies of K (23%), U (19%), J (14%), and T/X (10%) (Izagirre and de la Rua 1999). Haplogroup V was absent in the prehistoric sample. In the present study, the most frequent haplogroup was H (63%), followed by U (16%). Haplogroups V (7%), J (6%), and K (5%) were present at moderate frequencies, while haplogroups T (1%), W (1%), I/X (0.6%) and N1b (0.2%) are found at low frequencies. The divergence time estimates of some of the

European haplogroups are shown in Figure 53. It should be noted that these separation times reflect the age of the particular genetic system, not the age of the population in which it occurs (Casalotti *et al.* 1999). Only one, haplogroup J, can be unequivocally dated to the Neolithic. It is found at frequencies between 7-14% in other European populations (Alzualde *et al.* 2005). Haplogroup U (in particular U5) is the oldest, dating to around 50,000 BP in the Early Upper Paleolithic, and is considered to have developed *in situ* in Europe and reflect the migration of anatomically modern humans into the continent (Richards *et al.* 2002; Salas *et al.* 1998; Sykes 1999). Of the 19 U mtDNA sequences identified in the current sample, 11 (58%) are U5. By contrast, haplogroup U8a, dating to around 28,000 BP and reported at a frequency of 1.1% in modern Basques, was not detected (Gonzalez *et al.* 2006). H – the most frequent haplogroup in many extant European populations -- is comparatively young, dating to 18,000 BP during the Middle/Late Upper Paleolithic (Pereira *et al.* 2005). The other haplogroups (T, I, V, W) date to between 14,000-11,000 BP (Francalacci *et al.* 1996). Haplogroup V was once believed, due to its relatively high frequency and heterogeneity, to have arisen in the Basque region during the Late Upper Paleolithic (Torroni *et al.* 2001). However, analysis in prehistoric and historic populations revealed low to no frequency of V in these groups, and recent work has suggested that the homeland for this haplogroup, while still European, be moved to Cantabria just west of the Basque country (Alzualde *et al.* 2005; Izagirre and de la Rua 1999; Lell and Wallace 2000; Richards *et al.* 2002).

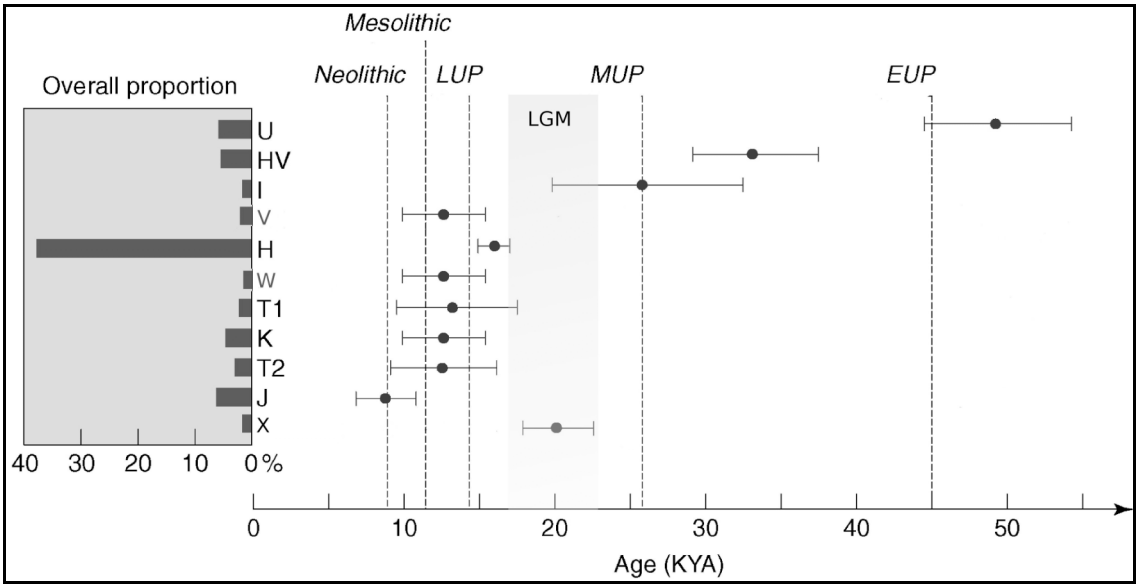

Figure 53. Estimates of ages of European mtDNA haplogroups. Adapted from Jobling et al. (2004) and Sykes (1999). Data based on Richards et al. (2000). LUP = start of Late Upper Paleolithic, MUP = start Middle Upper Paleolithic, EUP = start of Early Upper Paleolithic, LGM = Last Glacial Maximum.

Haplogroup N1b has not been reported previously in the Basque country, and was found in one male residing in the village of Ataun in Guipuzkoa, whose Y-haplogroup was determined as R1b. Haplogroup N1b is found at low frequencies elsewhere in Europe, and is present at high frequencies (10%) in Ashkenazi Jewish populations (Behar *et al.* 2004; Behar *et al.* 2006). However, the majority of Ashkenazi N1b sequences harbor a C-A transversion at position 16176, while the N1b identified in the Basque sample has the more common C-G transversion. This mutation has been reported in one Ashkenazim from Germany, but is more commonly found in non-Jewish groups, including populations in France (1.4%) and Italy (4.6%) and Eastern European groups in Bosnia (0.69%), Poland (0.23%), Croatia (2.8%) and the Czech Republic (1.1%) (Babalini *et al.* 2005; Dubut *et al.* 2004; Malyarchuk *et al.* 2003; Malyarchuk *et al.* 2002; Malyarchuk *et al.* 2006). It

has also been found at low frequencies in the Middle East (1-5%), the Caucasus (1-3%) as well as in Egypt (2.5-5%) (Gonzalez *et al.* 2008; Rowold *et al.* 2007). Haplogroup N1b is rare but not unusual in Europe, and is generally considered a Western Eurasian haplogroup, so the presence N1b among the Basques is not entirely unexpected. However, since autochthony was determined by Basque surnames and residence of four grandparents, it is possible that gene flow which occurred prior to the last three generations was unaccounted for. In agreement with the analysis of biparental markers and Y-chromosome, the mitochondrial DNA evidence demonstrates that the Basques are European, with high frequencies of the most common European haplogroup, H.

Peopling of Europe

Phylogeographic analysis of the Y-STR data revealed a high degree of genetic similarity in northern Iberia, including four Basque populations and Galicia. SAMOVA analysis separated Iberian populations in Central Portugal, Cantabria, and Barcelona, as well as Switzerland, from other European populations. The Basque populations cluster near populations in Ireland, Galicia, and the Swedish Saami, populations either on the Atlantic Fringe in the first two cases, or other non-Indo-European outliers in the last. Phylogeographic analysis of mtDNA sequence data also demonstrated an area of genetic similarity in northern Iberia, including both Basque groups and the populations from Galicia and Catalonia, while SAMOVA detected a genetic barrier between the Basques and all other European populations. Other studies have attributed the Basque distinction to little gene flow with surrounding groups and

a long period of genetic isolation (Aguirre *et al.* 1991a; Manzano *et al.* 1996a). Autosomal STR analysis confirms some gene flow in the Basque provinces, and the present study notes the presence of Neolithic haplogroups, both mtDNA (haplogroup J) and Y-chromosome (haplogroups J2a, G2a, and E1b1b). For mtDNA data, the Basques have approximately 6% Neolithic admixture, while for Y-chromosome data, they display 12% Neolithic ancestry. This means that while the Basques have not experienced the same degree of Neolithic influence as other European populations, Neolithic gene flow is still detectable, and therefore the Basques should not be considered a purely Paleolithic parental population. That being said, the population does harbor relatively high frequencies of haplogroups dated to the Paleolithic (mtDNA U5, Y-chromosome R1b), likely the result of genetic drift as interpreted from the heterozygosity vs. r_{ii} analysis of classical markers.

In regards to the Neolithic demic diffusion vs. cultural diffusion debate, neither seems to be strictly true, particularly in Iberia. The genetic evidence presents a more complex picture of European origins, with a SE-NW cline in frequencies for certain loci and haplotypes (mitochondrial haplogroups J, T1, and U3; Y-chromosome haplogroups J2a, G2a, and E1b1b), and a deeper time depth in Europe for others (mitochondrial haplogroups U8a, U5, H1, and H3; Y-chromosome haplogroup R1b) (Richards *et al.* 2002). A survey of human occupation of Europe during the Paleolithic/Neolithic transition explains how this result could occur (Pinhasi *et al.* 2000). Beginning in 14,000 BP, Europe was lightly populated, with a concentration of Late Paleolithic sites in present-day France and Germany. Over the

next 3,000 years, evidence of human occupation spread north and east, reaching Italy, the British Isles, and Scandinavia, with sites more sparsely distributed and little evidence of human activity in Greece or most of Eastern Europe. From 10,000 – 7,000 BP, evidence of Neolithic sites first appears in Anatolia, and spreads rapidly into those areas which were sparsely populated by Mesolithic hunter-gatherers. The foraging communities in Eastern Europe were not replaced by Neolithic farmers; they did not exist. Admixture models based on this archaeological data suggest regional variation across Europe (Lahr *et al.* 2000), with areas close to the epicenter of agricultural innovation and/or sparsely populated by Mesolithic groups showing little evidence of admixture and an almost exclusively Neolithic gene pool. Those areas which were densely populated during the Mesolithic and received few Neolithic migrants would show low admixture values and a high percentage of Mesolithic genes, while those which were more densely populated by Mesolithic groups followed by a moderate level of immigration by Neolithic farmers would have gene pools comprised approximately equally of Mesolithic and Neolithic genes. Given the evidence of the archaeological record, accounting for 27% of the genetic variation for the entire continent of Europe in one principal component is not that informative (Ammerman and Cavalli-Sforza 1984). It ignores the dynamic population movements during the Late Paleolithic and Mesolithic in Europe, as well as the impact of regional admixture. The authors now suggest that the first principal component (and the SE-NW cline observed for classical markers) be seen as a proxy for the amount of Neolithic admixture present in Europe, rather than as evidence of a

complete replacement (Cavalli-Sforza 2003).

Recent genetic studies concerning the peopling of Europe reflect a more nuanced approach. A study of admixture rates in contemporary European populations, using 22 Y-SNPs and putative Paleolithic (Basque) and Neolithic (Turkey, Iraq, Syria, and Lebanon) parental populations, found that the percentage of Neolithic admixture varies with distance from the Levant (Chikhi *et al.* 2002). Greece, Albania, and Macedonia have 100% Neolithic genes, while Sardinian, the Netherlands, and the Andalusia province of Spain have 0-16% Neolithic admixture. The authors interpret these results as evidence for the demic diffusion model, but, as the archaeological evidence demonstrates, modern populations in Greece would have few Paleolithic genes due to a lack of Paleolithic occupation of that region (though some have suggested this may be the result of sampling bias of the archaeological record in Greece). (Pinhasi 2000). Frequency and variance of Y-chromosome haplogroups in Central Europe suggest the diffusion of agriculture into this area was accomplished by populations bearing the indigenous I-M423 haplogroup, rather than haplogroups linked to the Neolithic transition (G and J) (Battaglia *et al.* 2009). The average Paleolithic contribution in 17 populations was found to be 51% for the Y-chromosome, which argues strongly against complete replacement. A meta-analysis of eight datasets, including mtDNA, Y SNPs, and autosomal markers, estimated Paleolithic/Neolithic admixture rates in 34 European populations (Dupanloup *et al.* 2004). As in the previous study, Basques were chosen to represent the Paleolithic parental population, while Near East and Anatolian groups were used as a proxy for

the Neolithic parental population. The Neolithic contribution in Europe varied with distance from the Levant, ranging from 80% in the Balkans to only 21% in the British Isles.

During the Paleolithic and into the Mesolithic, the Iberian Peninsula was sparsely populated, particularly in the interior. Most of the sites were coastal, reflecting the comparative wealth of resources (Straus 1991b; Zilhao 2000). The northern coast of Iberia, known as the Franco-Cantabrian region, was one of two refuge areas during the Last Glacial Maximum, the second being in the Ukraine. Paleolithic populations would have retreated to these areas during the height of the glaciation, and the archaeological record demonstrates an increase in occupation sites between 21,000 – 16,000 BP in this region of Iberia (47 sites compared to only 26 for the entire Early Upper Paleolithic) (Housley *et al.* 1997; Straus 1991a). Several mtDNA haplogroups (all present in the Basque population) have divergence times dating to the end of this period, including H, V, K, T and W, reflecting an expansion of these haplogroups as the glaciers retreated (McEvoy *et al.* 2004; Sykes 2003). Divergence times for Y-chromosome haplogroups R1b and I2a2 also date to this period, suggesting that these haplogroups are also of pre-Neolithic origin. Rather than demonstrating evidence of a 'wave of advance,' the Neolithic archaeological record in Iberia instead shows a leapfrog settlement pattern, again along the coasts (Zilhao 2000). Mesolithic populations were well established in these areas, and it is likely that small groups of Neolithic 'maritime pioneers' would have assimilated, or been assimilated by, the local Mesolithic groups (Lahr *et al.* 2000; Zilhao 1998; Zilhao

2003). In reference to the Basques, based on the archaeological evidence, the entire Iberian Peninsula was little affected by the Neolithic Transition, lending credence to the idea that the Basques are a remnant population, although one which has experienced some admixture since the end of the Last Glacial Maximum. The results of the present study agree with this hypothesis, with all genetic systems demonstrating that the Basques do not share a common ancestor with populations from the Caucasus or North Africa, but instead are a unique part of the genetic landscape of Europe.

CHAPTER SIX: CONCLUSION

This dissertation described the genetic composition of the Basques using three molecular marker systems, as well as data from classical genetic markers collected from the literature. Autosomal STRs, Y-chromosome STR haplotypes, mitochondrial DNA haplogroups, and mitochondrial HVS-I sequences were examined to facilitate the understanding of the place of this population in the genetic landscape of Europe. Whether the Basques share a common recent ancestor with populations from the Caucasus or North Africa, issues of heterogeneity among the Basques, and the degree of Paleolithic versus Neolithic ancestry were addressed.

The four genetic systems employed in the present study generate a consensus regarding hypotheses of Basque origins. Although Basques and Caucasians do both speak non-Indo-European agglutinative languages, a genetic relationship between these groups is not supported by the data. Genetic barriers were detected between the Basques and Caucasian groups using both uniparental marker systems, and phylogenetic analyses of autosomal markers revealed no relationship between them, contradicting the hypothesis of a common non-Indo-European ancestor. And while Basques and North African populations do share certain HLA haplotypes, they are found at very low frequencies in both populations (3.6% in Basques, < 2% in Algeria). These HLA markers are from a genomic region known to be under strong selective pressure (Brown *et al.* 2000; Martinez-Laso *et al.* 1995a; Satta *et al.* 1993; Takahata *et al.* 1992). The Vasco-Iberian hypothesis is not substantiated by analyses

of additional genetic systems, whether classical markers; autosomal, maternal, or paternal molecular systems; or even additional HLA analyses (Arnaiz-Villena *et al.* 1999; Bosch *et al.* 2001; Bosch *et al.* 1997; Brion *et al.* 2003; Brown *et al.* 2000; Comas *et al.* 2000; Comas *et al.* 1998b; Garcia Fernandez *et al.* 1997b; Perez-Miranda *et al.* 2003). Comparisons between the Basque population and other European groups, however, demonstrate both similarity and distinction. For the more than 20 classical loci examined, the Basque populations are outside the European range at only five. Analysis of uniparental markers reveals an absence of any non-European haplogroups, but higher frequencies of some, such as R1b and H. It is these higher frequencies for certain alleles, including RH*cde and GM*Z,A;G, which distinguish the Basques from other European populations, likely as the result of genetic drift. The diversity values calculated for various European groups demonstrate that while the Basques are on the low end of the European range for the molecular markers examined, there are other groups with comparable or lower diversity values (i.e., Catalonia, the Swedish Saami, Vojvodina and Wales). Phylogeographic analysis revealed high degrees of genetic similarity in populations of northern Iberia, using both Y-chromosome and mitochondrial data, however, based on mtDNA a genetic barrier was detected between the Basques and all other European groups.

While the Basques do display low levels of intrapopulation diversity at a few loci, the majority of variation in the population is accounted for by differences between individuals in all three molecular systems examined. This suggests that while

genetic drift may have played a role in the increased frequency of certain alleles in the Basque population, it has been ameliorated by gene flow. This gene flow is evident both in the analysis of heterozygosity versus r_{ii} in autosomal STR markers, as well as the haplogroup variation present in the uniparental genetic systems, as both maternal and paternal lineages contain Neolithic haplogroups (mtDNA: J, Y-STR: J2a, G2a, E1b1b). Examination of uniparental markers demonstrates that overall, the Basques have 7.14% Neolithic haplotypes in the paternal lineage, and a comparable level (5.68%) in the maternal line. By province, Vizcaya has the highest level of mitochondrial Neolithic admixture (10.13%), while Guipuzkoa shows the highest level of Neolithic admixture (5.5%) for the Y-chromosome. The presence of 15 individuals in an ancient Basque sample (5,000-3,400 BP) belonging to mitochondrial haplogroup J suggests that this admixture has not been recent (Izagirre and de la Rua 1999; Jackes *et al.* 1997). This result does not imply that ancient Basque populations experienced gene flow from Neolithic farmers, as an analysis of skeletal remains from the Neolithic Transition in Iberia found little evidence of population replacement (i.e., major dietary changes reflected in skeletal morphology or trace mineral composition). In addition, expansion times calculated from mitochondrial HVS-I sequences for European populations date overwhelmingly to the Paleolithic, meaning population expansions occurred well before the advent of agriculture. The percentage of Paleolithic ancestry found in European populations increases with increasing geographic distance from the Fertile Crescent, arguing strongly against the demic diffusion model of complete replacement of Paleolithic foragers by a wave of

advancing Neolithic farmers.

This study provides a comprehensive genetic view of the place of the Basques in the European genetic landscape. Concordance among the four genetic systems examined lends strength to the hypothesis that the Basques are a European population which has experienced significant genetic drift, but also some gene flow, since the Last Glacial Maximum.

LITERATURE CITED

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V and others. 2004. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *American Journal of Human Genetics* 75(5):910-918.

Aguirre A, Vicario A, Mazon LI, Estomba A, Martinez de Pancorbo M, Arrieta Pico V, Perez Elortondo F, and Lostao CM. 1991a. Are the Basques a single and a unique population? *American Journal of Human Genetics* 49(2):450-458.

Aguirre AI, Vicario A, Mazon LI, De Pancorbo MM, Arizti P, Estomba A, and Lostao CM. 1989a. AK1, PGD, GC and HP frequencies in the Basque population: a review. *Gene Geography* 3(1):41-51.

Aguirre AI, Vicario A, Mazon LI, De Pancorbo MM, Arizti P, Estomba A, and Lostao CM. 1989b. AK1, PGD, GC and HP frequencies in the Basque population: a review. *Gene Geography* 3(1):41-51.

Aguirre AI, Vicario A, Mazon LI, de Pancorbo MM, Estomba A, and Lostao C. 1991b. Acid phosphatase, adenosine deaminase and esterase D polymorphisms in the Spanish Basque population. *Human Heredity* 41(2):93-102.

Alberts B, Bray D, Lewis J, Raff M, Roberts K, and Watson JD. 1989. *Molecular Biology of the Cell*. New York: Garland Publishing.

Allison AC, Blumberg BS, and Ap R. 1958. Haptoglobin types in British, Spanish, Basque and Nigerian African populations. *Nature* 181(4612):824-825.

Alonso S, Castro A, Fernandez I, Gomez de Cedron M, Garcia-Orad A, Meyer E, and Martinez de Pancorbo M. 1995. Population study of 3 STR loci in the Basque Country (northern Spain). *International Journal of Legal Medicine* 107(5):239-245.

Alonso S, Flores C, Cabrera V, Alonso A, Martin P, Albarran C, Izagirre N, de la Rua C, and Garcia O. 2005. The place of the Basques in the European Y-chromosome diversity landscape. *European Journal of Human Genetics* 13(12):1293-1302.

Alper CA, Awdeh ZL, and Yunis EJ. 1986. Complotypes, extended haplotypes, male segregation distortion, and disease markers. *Human Immunology* 15(4):366-373.

Alzualde A, Izagirre N, Alonso S, Alonso A, Albarran C, Azkarate A, and de la Rua C. 2006. Insights into the "isolation" of the Basques: mtDNA lineages from the historical site of Aldaieta (6th-7th centuries AD). *American Journal of Physical Anthropology* 130(3):394-404.

Alzualde A, Izagirre N, Alonso S, Alonso A, and de la Rua C. 2005. Temporal mitochondrial DNA variation in the Basque Country: influence of Post-Neolithic events. *Annals of Human Genetics* 69(Pt 6):665-679.

Ammerman AJ, and Cavalli-Sforza LL. 1984. *The Neolithic Transition and the Genetics of Populations in Europe*. Princeton, N.J.: Princeton University Press. xv, 176 p. p.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F and others. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290(5806):457-465.

Arnaiz-Villena A, Benmamar D, Alvarez M, Diaz-Campos N, Varela P, Gomez-Casado E, and Martinez-Laso J. 1995. HLA allele and haplotype frequencies in Algerians. Relatedness to Spaniards and Basques. *Human Immunology* 43(4):259-268.

Arnaiz-Villena A, Gomez-Casado E, and Martinez-Laso J. 2002. Population genetic relationships between Mediterranean populations determined by HLA allele distribution and a historic perspective. *Tissue Antigens* 60(2):111-121.

Arnaiz-Villena A, Martinez-Laso J, and Alonso-Garcia J. 1999. Iberia: population genetics, anthropology, and linguistics. *Human Biology* 71(5):725-743.

Arnaiz-Villena A, Martinez-Laso J, and Alonso-Garcia J. 2001. The correlation between languages and genes: the Usko-Mediterranean peoples. *Human Immunology* 62(9):1051-1061.

Arnaiz-Villena A, Martinez-Laso J, and Gomez-Casado E. 1997a. Relatedness among Basques, Portuguese, Spaniards, and Algerians studied by *HLA* allelic frequencies and haplotypes. *Immunogenetics* 47:37-43.

Arnaiz-Villena A, Martinez-Laso J, Gomez-Casado E, Diaz-Campos N, Santos P, Martinho A, and Breda-Coimbra H. 1997b. Relatedness among Basques, Portuguese, Spaniards, and Algerians studied by HLA allelic frequencies and haplotypes. *Immunogenetics* 47(1):37-43.

Arnaiz-Villena A, Rodriguez de Cordoba S, Vela F, Pascual JC, Cervero J, and Bootello A. 1981. HLA antigens in a sample of the Spanish population: common features among Spaniards, Basques, and Sardinians. *Human Genetics* 58(3):344-348.

Arrieta MI, Martinez B, Millan JM, Gil A, Monros E, Nunez T, Telez M, and Martinez F. 1997. Study of a trimeric tandem repeat locus (SBMA) in the Basque population: comparison with other populations. *Gene Geography* 11(1):61-72.

Athey TW. 2005. Haplogroup prediction from Y-STR values using an allele-frequency approach. *Journal of Genetic Genealogy* 1:1-7.

Athey TW. 2006. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *Journal of Genetic Genealogy* 2:34-39.

Babalini C, Martinez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, Tarsi T, Contini I, Barac L, Janicijevic B, Martinovic Klaric I and others. 2005. The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. *European Journal of Human Genetics* 13(8):902-912.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16:37-48.

Barbujani G, Bertorelle G, and Chikhi L. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. *American Journal of Human Genetics* 62(2):488-492.

Barbujani G, Sokal RR, and Oden NL. 1995. Indo-European origins: a computer-simulation test of five hypotheses. *American Journal of Physical Anthropology* 96(2):109-132.

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, King RJ, Rootsi S, Marjanovic D, Primorac D and others. 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *European Journal of Human Genetics* 17(6):820-830.

Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, Erlich HA, and Klitz W. 1992. Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *Journal of Immunology* 148(1):249-258.

Behar DM, Hammer MF, Garrigan D, Villems R, Bonne-Tamir B, Richards M, Gurwitz D, Rosengarten D, Kaplan M, Della Pergola S and others. 2004. MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. *European Journal of Human Genetics* 12(5):355-364.

Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, Tzur S, Pereira L, Amorim A, Quintana-Murci L, Majamaa K and others. 2006. The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. *American Journal of Human Genetics* 78(3):487-497.

Bernard J, and Ruffie J. 1976. Hematologie et culture: Le peuplement de l'Europe de l'ouest. *Annales, Economies, Societes, Civilizations* 31:661-676.

Bertorelle G, Bertranpetit J, Calafell F, Nasidze IS, and Barbujani G. 1995. Do Basque- and Caucasian-speaking populations share non-Indo-European ancestors? *European Journal of Human Genetics* 3(4):256-263.

Bertranpetit J, and Cavalli-Sforza LL. 1991a. A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics* 55(Pt 1):51-67.

Bertranpetit J, and Cavalli-Sforza LL. 1991b. A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics* 55 (Pt 1):51-67.

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, and Comas D. 1995. Human mitochondrial DNA variation and the origin of Basques. *Annals of Human Genetics* 59(Pt 1):63-81.

Blaud HC. 1974. *The Basques*. San Francisco: R and E Research Associates. vi, 95 p. p.

Bodmer WF. 1997. HLA: what's in a name? A commentary on HLA nomenclature development over the years. *Tissue Antigens* 49(3 Pt 2):293-296.

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, and Bertranpetit J. 2001. High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *American Journal of Human Genetics* 68(4):1019-1029.

Bosch E, Calafell F, Perez-Lezaun A, Comas D, Mateu E, and Bertranpetit J. 1997. Population history of north Africa: evidence from classical genetic markers. *Human Biology* 69(3):295-311.

Bosch E, Calafell F, Santos FR, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, and Bertranpetit J. 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y-chromosome. *American Journal of Human Genetics* 65(6):1623-1638.

Boyd WC, and Boyd LG. 1937. New data on blood groups and other inherited factors in Europe and Egypt. *American Journal of Physical Anthropology* 23:49-70.

Brion M, Salas A, Gonzalez-Neira A, Lareu MV, and Carracedo A. 2003. Insights into Iberian population origins through the construction of highly informative Y-chromosome haplotypes using biallelic markers, STRs, and the MSY1 minisatellite. *American Journal of Physical Anthropology* 122(2):147-161.

Broca P. 1864. Letter to the Society regarding Basque crania. *Journal of the Anthropological Society of London* 2:cclxvii-cclxxiii.

Broca P. 1878. Translation of the Greater Part of the Address Delivered by M. Broca, President, at the Opening Meeting of the French Association for the Advancement of the Sciences, at the Havre Congress, 1877. *The Journal of the Anthropological Institute of Great Britain and Ireland* 7:187-200.

Brown ES. 1965. Distribution of the Abo and Rhesus (D) Blood Groups in the North of Scotland. *Heredity* 20:289-303.

Brown RJ, Rowold D, Tahir M, Barna C, Duncan G, and Herrer RJ. 2000. Distribution of the HLA-DQA1 and polymarker alleles in the Basque population of Spain. *Forensic Science International* 108(2):145-151.

Calafell F, and Bertranpetit J. 1993a. The genetic history of the Iberian peninsula: A simulation. *Current Anthropology* 66:823-842.

Calafell F, and Bertranpetit J. 1993b. The genetic history of the Iberian peninsula: A simulation. *Current Anthropology* 34:735-745.

Calafell F, and Bertranpetit J. 1994a. Mountains and genes: population history of the Pyrenees. *Human Biology* 66(5):823-842.

Calafell F, and Bertranpetit J. 1994b. Principal component analysis of gene frequencies and the origin of Basques. *American Journal of Physical Anthropology* 93(2):201-215.

Calafell F, and Bertranpetit J. 1994c. Principal component analysis of gene frequencies and the origin of Basques. *American Journal of Physical Anthropology* 93(2):201-215.

Calderon R. 2002. Genetic structure of the Basque herders of northern Spain. In: Leonard WR, and Crawford MH, editors. *Human Biology of Pastoral Populations.* Cambridge: Cambridge University Press. p 50-63.

Calderon R, Carrion M, Perez-Miranda A, Pena JA, Dugoujon JM, and Crouau-Roy B. 2003. Allele variation of DYS19 and Y-Alu insertion (YAP) polymorphisms in Basques: an insight into the peopling of Europe and the Mediterranean region. *Human Biology* 75(1):117-127.

Calderon R, Vidales C, Pena JA, Perez-Miranda A, and Dugoujon JM. 1998. Immunoglobulin allotypes (GM and KM) in Basques from Spain: approach to the origin of the Basque population. *Human Biology* 70(4):667-698.

Calderon R, Wentzel J, and Roberts DF. 1993. HLA frequencies in Basques in Spain and in neighbouring populations. *Annals of Human Biology* 20(2):109-120.

Cambon-de Mouzon A, Ohayon E, Hauptmann G, Sevin A, Abbal M, Sommer E, Vergnes H, and Ducos J. 1982. HLA-A, B, C, DR antigens, Bf, C4 and glyoxalase I (GLO) polymorphisms in French Basques with insulin-dependent diabetes mellitus (IDDM). *Tissue Antigens* 19:366-379.

Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia A, Vetrugno G, Tofanelli S, Onofri V, Tagliabracci A, Paoli G and others. 2007. Y-chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Molecular and Phylogenetic Evolution* 44(1):228-239.

Casalotti R, Simoni L, Belledi M, and Barbujani G. 1999. Y-chromosome polymorphisms and the origins of the European gene pool. *Proceedings: Biological Sciences* 266(1432):1959-1965.

Casals T, Nunes V, Palacio A, Gimenez J, Gaona A, Ibanez N, Morral N, and Estivill X. 1993. Cystic fibrosis in Spain: high frequency of mutation G542X in the Mediterranean coastal area. *Human Genetics* 91(1):66-70.

Casals T, Vazquez C, Lazaro C, Girbau E, Gimenez FJ, and Estivill X. 1992. Cystic fibrosis in the Basque country: high frequency of mutation delta F508 in patients of Basque origin. *American Journal of Human Genetics* 50(2):404-410.

Cavalli-Sforza LL. 2003. Returning to the Neolithic transition in Europe. In: Ammerman AJ, and Biagi P, editors. *The Widening Harvest: The Neolithic Transition in Europe.* Boston, MA: Archaeological Institute of America. p 297-313.

Cavalli-Sforza LL. 1988. The Basque population and ancient migrations in Europe. *Munibe (Antropolgia y Arqueologia)* Supplemento 6:129-137.

Cavalli-Sforza LL, and Feldman MW. 1990. Spatial subdivision of populations and estimates of genetic variation. *Theoretical Population Biology* 37(1):3-25.

Cavalli-Sforza LL, Menozzi P, and Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, N.J.: Princeton University Press. xi, 518 p. p.

Cavalli-Sforza LL, and Minch E. 1997. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *American Journal of Human Genetics* 61(1):247-254.

Cavalli-Sforza LL, and Piazza A. 1993. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *European Journal of Human Genetics* 1(1):3-18.

Chalmers JN, Elizabeth W. Ikin, A.E. Mourant. 1948. Basque blood groups. *Nature* 162(4105):27.

Chalmers JN, Ikin EW, and Mourant AE. 1949. The ABO, MN and Rh blood groups of the Basque people. *American Journal of Physical Anthropology* 7(4):529-544.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, and Barbujani G. 1998. Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proceedings of the National Academy of Sciences USA* 95(15):9053-9058.

Chikhi L, Nichols RA, Barbujani G, and Beaumont MA. 2002. Y genetic data support the Neolithic demic diffusion model. *Proceedings of the National Academy of Sciences USA* 99(17):11008-11013.

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K and others. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Human Genetics* 114(2):127-148.

Closson CC. 1897. The hierarchy of European races. *The American Journal of Sociology* 3(3):314-327.

Collins R. 1990. *The Basques*. Oxford, UK: Basil Blackwell Ltd.

Collins R. 1992. The ethnogenesis of the Basques. In: Collins R, editor. *Law, Culture and Regionalism in Early Medieval Spain*. Hampshire, Great Britain: Variorum. p 35-44.

Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, and Sajantila A. 2000. Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Human Genetics* 107(4):312-319.

Comas D, Calafell F, Mateu E, Perez-Lezaun A, and Bertranpetit J. 1998a. HLA evidence for the lack of genetic heterogeneity in Basques. *Annals of Human Genetics* 62(Pt 2):123-132.

Comas D, Mateu E, Calafell F, Perez-Lezaun A, Bosch E, Martinez-Arias R, and Bertranpetit J. 1998b. HLA class I and class II DNA typing and the origin of Basques. *Tissue Antigens* 51(1):30-40.

Constable OR. 1997. *Medieval Iberia : Readings from Christian, Muslim, and Jewish Sources*. Philadelphia: University of Pennsylvania Press. xxvii, 426 p. p.

Constans J, and Viau M. 1975. Distribution of haptoglobin subtypes in French Basques. *Human Heredity* 25(2):156-159.

Corte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, and Sykes BC. 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Annals of Human Genetics* 60(Pt 4):331-350.

Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B and others. 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y-chromosomes reveals multiple migratory events within and out of Africa. *American Journal of Human Genetics* 74(5):1014-1022.

Cruciani F, La Fratta R, Torroni A, Underhill PA, and Scozzari R. 2006. Molecular dissection of the Y-chromosome haplogroup E-M78 (E3b1a): a posteriori

evaluation of a microsatellite-network-based approach through six new biallelic markers. *Human Mutation* 27(8):831-832.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon JM, Crivellaro F, Benincasa T, Pascone R and others. 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Molecular Biology and Evolution* 24(6):1300-1311.

Dausset J, Legrand, L, Levine, MH, Qulici, JC, Colombani, M, Ruffie, J. 1972. Genetic structure and distribution of HL-A antigens in a Basque village. *Histocompatability Testing*:99-105.

de Mouzon A, Constans J, Sommer E, Sevin A, Quilici JC, Fernet P, Ducos J, and Ohayon E. 1980. HLA-A, B typing in Basque and other Pyrenean populations. *Tissue Antigens* 15(1):11-18.

de Mouzon A, Ohayon E, Ducos J, and Hamptmann G. 1979. Bf and C4 markers for insulin-dependent diabetes in Basques. *Lancet* 2(8156-8157):1364.

de Pancorbo MM, Mazon LI, de la Rica C, Vicario A, and Lostao CM. 1989. Some red cell enzymes and haptoglobin gene frequencies in two Basque regions and Leon. *Annals of Human Biology* 16(2):147-154.

de Pancorbo MM, Mazon LI, and Lostao CM. 1986. A cline in the acid phosphatase1 distribution in the Iberian Peninsula. *Annals of Human Biology* 13(3):297-300.

Degos L, and Dausset J. 1974. Human migrations and linkage disequilibrium of HL-A system. *Immunogenetics* 3:195-210.

Derish PA, and Sokal RR. 1988. A classification of European populations based on gene frequencies and cranial measurements: a map-quadrat approach. *Human Biology* 60(5):801-824.

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, Barbujani G, Papola F, Ciavarella G and others. 2004. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Human Genetics* 115(5):357-371.

Dubut V, Chollet L, Murail P, Cartault F, Beraud-Colomb E, Serre M, and Mogentale-Profizi N. 2004. mtDNA polymorphisms in five French groups: importance of regional sampling. *European Journal of Human Genetics* 12(4):293-300.

Dugoujon JM, Clayton J, Sevin A, Constans J, Loirat F, and Hazout S. 1989. Immunoglobulin (Gm and Km) allotypes in some Pyrenean populations of France. *Collegium Antropologicum* 13:43-50.

Dupanloup I, Bertorelle G, Chikhi L, and Barbujani G. 2004. Estimating the impact of prehistoric admixture on the genome of Europeans. *Molecular Biology and Evolution* 21(7):1361-1372.

Dupanloup I, Schneider S, and Excoffier L. 2002. A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* 11(12):2571-2581.

Enamorado J. 1997. Behavioural transformations during the Pleistocene. In: Díaz-Andreu M, and Keay SJ, editors. *The Archaeology of Iberia: The Dynamics of Change*. London: Routledge. p 34-64.

Erlich HA, and Gyllensten UB. 1991. The evolution of allelic diversity at the primate major histocompatibility complex class II loci. *Human Immunology* 30(2):110-118.

Esparza B, Pestoni C, Martin MD, Merino F, and Carracedo A. 1995. Distribution of the HLA-DQA1 polymorphism in the population of the Basque Country (Spain). *Gene Geography* 9(1):53-58.

Esteban E, Dugoujon JM, Guitard E, Senegas MT, Manzano C, de la Rua C, Valveny N, and Moral P. 1998. Genetic diversity in northern Spain (Basque Country and Cantabria): GM and KM variation related to demographic histories. *European Journal of Human Genetics* 6(4):315-324.

Etcheverry MA. 1945. El factor rhesus, su genetica e importancia clinica.

Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology* 13(4):853-864.

Excoffier L, Laval G, and Schneider JA. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1:47-50.

Excoffier L, Smouse PE, and Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131(2):479-491.

Fan WM, Kasahara M, Gutknecht J, Klein D, Mayer WE, Jonker M, and Klein J. 1989. Shared class II MHC polymorphisms between humans and chimpanzees. *Human Immunology* 26(2):107-121.

Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, Mass.: Sinauer Associates. xx, 664 p. p.

Flores C, Maca-Meyer N, Gonzalez AM, Oefner PJ, Shen P, Perez JA, Rojas A, Larruga JM, and Underhill PA. 2004. Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. *European Journal of Human Genetics* 12(10):855-863.

Forster P, Torroni A, Renfrew C, Röhl A. 2001. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Molecular Biology and Evolution* 18:1864-1881.

Fox CL, Gonzalez Martin A, and Vives Civit S. 1996. Cranial variation in the Iberian Peninsula and the Balearic Islands: inferences about the history of the population. *American Journal of Physical Anthropology* 99(3):413-428.

Francalacci P, Bertranpetit J, Calafell F, and Underhill PA. 1996. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *American Journal of Physical Anthropology* 100(4):443-460.

Francalacci P, Morelli L, Underhill PA, Lillie AS, Passarino G, Useli A, Madeddu R, Paoli G, Tofanelli S, Calo CM and others. 2003. Peopling of three

Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *American Journal of Physical Anthropology* 121(3):270-279.
Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147(2):915-925.
Garcia Fernandez E, Arrieta A, Rinon M, Maruri N, Arranz MC, Pena JA, and Garcia Masdevall MD. 1997a. Genetic polymorphisms of HLA class I and class II system in the Basque population. *Transplant Proceedings* 29(8):3707-3709.
Garcia Fernandez E, Arrieta A, Rinon M, Maruri N, Arranz MC, Pena JA, and Garcia Masdevall MD. 1997b. Genetic polymorphisms of HLA class I and class II system in the Basque population. *Transplant Proceedings* 29:3707-3709.
Garcia O, Martin P, Budowle B, Uriarte J, Albarran C, and Alonso A. 1998. Basque Country autochthonous population data on 7 short tandem repeat loci. *International Journal of Legal Medicine* 111(3):162-164.
Garcia O, Martin P, Gusmao L, Albarran C, Alonso S, de la Rua C, Flores C, Izagirre N, Penas R, Antonio Perez J and others. 2004. A Basque Country autochthonous population study of 11 Y-chromosome STR loci. *Forensic Science International* 145(1):65-68.
Garcia O, Uriarte I, Martin P, Albarran C, and Alonso A. 2001. STR data from Basque country autochthonous population. *Forensic Science International* 115(1-2):111-112.
Garcia-Orad A, Aguirre AI, Mazon LI, and de Pancorbo MM. 1987. Polymorphism of delta-aminolevulinic acid dehydratase in Basque populations. *Human Heredity* 37(5):321-322.
Garcia-Orad A, Arizti P, Esteban JL, Garcia-Orad C, Garcia-Arenzana C, Constans J, and de Pancorbo MM. 1990. Polymorphism of haptoglobin (HP), group specific component (GC) and alpha-1-antitrypsin (PI) in the resident population of the Basque Country (Spain). *Gene Geography* 4(1):43-51.
Garza J, and Williamson E. 2001. Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* 10(2):305-318.
Gaudieri S, Dawkins RL, Habara K, Kulski JK, and Gojobori T. 2000. SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Research* 10(10):1579-1586.
Giraldo MP, Vallet M, Guitard E, Senegas MT, Sevin A, Nogues RM, Aluja MP, and Dugoujon JM. 1998. GM and KM immunoglobulin allotypes in a Spanish Pyrenean population: Val d'Aran. *Annals of Human Biology* 25(5):453-465.
Goedde HW, Hirth L, Benkmann HG, Pellicer A, Pellicer T, Stahn M, and Singh S. 1972a. Population genetic studies of red cell enzyme polymorphisms in four Spanish populations. *Human Heredity* 22:552-560.
Goedde HW, Hirth L, Benkmann HG, Pellicer A, Pellicer T, Stahn M, and Singh S. 1972b. Population genetic studies of red cell enzyme polymorphisms in four Spanish populations. *Human Heredity* 22:552-560.

Goedde HW, Hirth L, Benkmann HG, Pellicer A, Pellicer T, Stahn M, and Singh S. 1973. Population genetic studies of serum protein polymorphisms in four Spanish populations. II. *Human Heredity* 23(2):135-146.

Gomez-Casado E, del Moral P, Martinez-Laso J, Garcia-Gomez A, Allende L, Silvera-Redondo C, Longas J, Gonzalez-Hevilla M, Kandil M, Zamora J and others. 2000. HLA genes in Arabic-speaking Moroccans: close relatedness to Berbers and Iberians. *Tissue Antigens* 55(3):239-249.

Gonzalez AM, Brehm A, Perez JA, Maca-Meyer N, Flores C, and Cabrera VM. 2003. Mitochondrial DNA affinities at the Atlantic fringe of Europe. *American Journal of Physical Anthropology* 120(4):391-404.

Gonzalez AM, Garcia O, Larruga JM, and Cabrera VM. 2006. The mitochondrial lineage U8a reveals a Paleolithic settlement in the Basque country. *BMC Genomics* 7(1):124.

Gonzalez AM, Karadsheh N, Maca-Meyer N, Flores C, Cabrera VM, and Larruga JM. 2008. Mitochondrial DNA variation in Jordanians and their genetic relationship to other Middle East populations. *Annals of Human Biology* 35(2):212 - 231.

Gonzalez-Neira A, Gusmao L, Brion M, Lareu MV, Amorim A, and Carracedo A. 2000. Distribution of Y-chromosome STR defined haplotypes in Iberia. *Forensic Science International* 110(2):117-126.

Green ED, Mohr RM, Idol JR, Jones M, Buckingham JM, Deaven LL, Moyzis RK, and Olson MV. 1991. Systematic generation of sequence-tagged sites for physical mapping of human chromosomes: application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* 11(3):548-564.

Grimaldi MC, Crouau-Roy B, Amoros JP, Cambon-Thomsen A, Carcassi C, Orru S, Viader C, and Contu L. 2001. West Mediterranean islands (Corsica, Balearic islands, Sardinia) and the Basque population: contribution of HLA class I molecular markers to their evolutionary history. *Tissue Antigens* 58(5):281-292.

Grubb R. 1970. *The Genetic Markers of Human Immunoglobulins.* New York: Springer-Verlag.

Gruen JR, and Weissman SM. 2001. Human MHC class III and IV genes and disease associations. *Frontiers in Bioscience* 6:D960-972.

Guo SW, and Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48(2):361-372.

Gusmao L, Sanchez-Diz P, Alves C, Beleza S, Lopes A, Carracedo A, and Amorim A. 2003. Grouping of Y-STR haplotypes discloses European geographic clines. *Forensic Science International* 134(2-3):172-179.

Gyllensten U, Bergstrom T, Josefsson A, Sundvall M, and Erlich HA. 1996. Rapid allelic diversification and intensified selection at antigen recognition sites of the Mhc class II DPB1 locus during hominoid evolution. *Tissue Antigens* 47(3):212-221.

Haarmann H. 1998. Basque ethnogenesis, acculturation, and the role of language contacts. *Fontes Linguae Vasconum*:25-42.
Hall T. 2007. BioEdit. Version 7.0.8. Carlsbad, CA: Ibis Biosciences.
Harding RM, and Sokal RR. 1988. Classification of the European language families by genetic distance. *Proceedings of the National Academy of Sciences USA* 85(23):9370-9372.
Harpending H, and Jenkins T. 1973. Genetic distances among southern African populations. In: Crawford MH, and Workman PL, editors. *Methods and Theories of Anthropological Genetics*. Albuquerque, NM: University of New Mexico Press.
Harpending H, and Rogers A. 1984. ANTANA: A package for multivariate data analysis: Distributed by the authors.
Harpending H, and Ward BE. 1982. Chemical systematics and human populations. In: Nitecki M, editor. *Biological Aspects of Evolutionary Biology*. Chicago: University of Chicago Press.
Hazout S, Dugoujon JM, Loirat F, and Constans J. 1991. Genetic similarity maps and immunoglobulin allotypes of eleven populations from the Pyrenees (France). *Annals of Human Genetics* 55 (Pt 2):161-174.
Housley RA, Gamble CS, Street M, and Pettitt P. 1997. Radiocarbon evidence for the lateglacial human recolonisation of Northern Europe. *Proceedings of the Prehistoric Society* 63:25-54.
Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*. Oxford: Oxford University Press. p 1-44.
Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, Baptista R, Kruglyak L, Xu SH and others. 1995. An STS-based map of the human genome. *Science* 270(5244):1945-1954.
Hughes AL, and Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335(6186):167-170.
Hughes AL, and Nei M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences USA* 86(3):958-962.
Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Perez-Lezaun A, Bosch E, Shlumukova M, Cambon-Thomsen A, McElreavey K and others. 1999. Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *American Journal of Human Genetics* 65(5):1437-1448.
Imanishi T, Akaza T, Kimura A, Tokunaga K, and Gojobori T. 1991. Allele and haplotype frequencies for HLA and complement loci in various ethnic groups. In: Tsuji K, Aizawa M, and Sasazuki T, editors. *Proceedings of the Eleventh International Histocompatibility Workshop and Conference*. Oxford: Oxford University Press. p 1065-1220.

Iriondo M, Barbero MC, Izagirre N, and Manzano C. 1997. Data on six short-tandem repeat polymorphisms in an autochthonous Basque population. *Human Heredity* 47(3):131-137.

Iriondo M, Barbero MC, and Manzano C. 2003. DNA polymorphisms detect ancient barriers to gene flow in Basques. *American Journal of Physical Anthropology* 122(1):73-84.

Iriondo M, de la Rua C, Barbero MC, Aguirre A, and Manzano C. 1999. Analysis of 6 short tandem repeat loci in Navarre (northern Spain). *Human Biology* 71(1):43-54.

Iriondo M, Manzano C, and de la Rua C. 1996. HLA-DQA1 in autochthonous Basques: description of a genocline for the DQA1*0201 allele in Europe. *International Journal of Legal Medicine* 109(4):181-185.

Izagirre N, and de la Rua C. 1999. An mtDNA analysis in ancient Basque populations: implications for haplogroup V as a marker for a major Paleolithic expansion from southwestern Europe. *American Journal of Human Genetics* 65(1):199-207.

Jackes M, Lubell D, and Meiklejohn C. 1997. On physical anthropological aspects of the Mesolithic-Neolithic transition in the Iberian Peninsula. *Current Anthropology* 38(5):839-846.

Jiménez EF. 2001. *Struggle and Survival of the Pre-Roman Languages of the Iberian Peninsula*. Lewiston, N.Y.: E. Mellen Press. iii, 155 p. p.

Kalinowski ST. 2009. How well do evolutionary trees describe genetic relationships among populations? *Heredity* 102(5):506-13.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, and Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research* 18(5):830-838.

Kimpton C, Walton A, and Gill P. 1992. A further tetranucleotide repeat polymorphism in the vWF gene. *Human Molecular Genetics* 1(4):287.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16(2):111-120.

Klein J, Satta Y, O'HUigin C, and Takahata N. 1993. The molecular descent of the major histocompatibility complex. *Annual Review of Immunology* 11:269-295.

Klitz W, Thomson G, and Baur MP. 1986. Contrasting evolutionary histories among tightly linked HLA loci. *American Journal of Human Genetics* 39(3):340-349.

Lachmann PJ, and Hobart MJ. 1979. The genetics of the complement system. *Ciba Foundation Symposia* (66):231-250.

Lahr MM, Foley RA, and Pinhasi R. 2000. Expected regional patterns of Mesolithic-Neolithic human population admixture in Europe based on archaeological evidence. In: Renfrew C, and Boyle K, editors. *Archaeogenetics: DNA and the Population Prehistory of Europe*. Cambridge: McDonald Institute for Archaeological Research. p 81-88.

Lawlor DA, Ward FE, Ennis PD, Jackson AP, and Parham P. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335(6187):268-271.

Lell JT, and Wallace DC. 2000. The peopling of Europe from the maternal and paternal perspectives. *American Journal of Human Genetics* 67(6):1376-1381.

Levine MH. 1967. The Basques. *Natural History* 76:44-51.

Levine MH. 1977. A hematological approach to Basque isolation in two French Basque Villages. *Annals of the New York Academy of Science* 293:185-193.

Levine MH, von Hagen V, Quilci J-C, and Salmon D. 1974. Anthropology of a Basque village: A new hemotypological study. *Cahiers d'anthropologie et d'ecologie humaine* II(3-4):159-171.

Levine MH, Von Hagen V, Ruffie J, and Darrasse H. 1977. A hematological approach to Basque isolation in two French Basque Villages. *Annals of the New York Academy of Science* 293:185-193.

Li H, Schmidt L, Wei MH, Hustad T, Lerman MI, Zbar B, and Tory K. 1993. Three tetranucleotide polymorphisms for loci: D3S1352; D3S1358; D3S1359. *Human Molecular Genetics* 2(8):1327.

Loogvali EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, Metspalu E, Tambets K, Reidla M, Tolk HV, Parik J and others. 2004. Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Molecular Biology and Evolution* 21(11):2012-2021.

Lopez-Parra AM, Gusmao L, Tavares L, Baeza C, Amorim A, Mesa MS, Prata MJ, and Arroyo-Pardo E. 2009. In search of the pre- and post-neolithic genetic substrates in Iberia: evidence from Y-chromosome in Pyrenean populations. *Annals of Human Genetics* 73(1):42-53.

Lucotte G, and Hazout S. 1996. Y-chromosome DNA haplotypes in Basques. *Journal of Molecular Evolution* 42(4):472-475.

Lucotte G, and Loirat F. 1999. Y-chromosome DNA haplotype 15 in Europe. *Human Biology* 71(3):431-437.

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, and Cabrera VM. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* 2:13.

Maca-Meyer N, Gonzalez AM, Pestano J, Flores C, Larruga JM, and Cabrera VM. 2003. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genetics* 4:15.

Macaulay VA, Richards MB, Forster P, Bendall KE, Watson E, Sykes B, and Bandelt HJ. 1997. mtDNA mutation rates--no need to panic. *American Journal of Human Genetics* 61(4):983-990.

MacClancy J. 1993. Biological Basques, sociologically speaking. In: Chapman M, editor. Social and Biological Aspects of Ethnicity. Oxford, UK: Oxford University Press.

Mallory JP. 1989. *In search of the Indo-Europeans : Language, Archaeology, and Myth*. New York, N.Y.: Thames and Hudson. 288 p. p.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnic K, and Miscicka-Sliwka D. 2003. Mitochondrial DNA variability in Bosnians and Slovenians. *Annals of Human Genetics* 67(Pt 5):412-425.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Wozniak M, and Miscicka-Sliwka D. 2002. Mitochondrial DNA variability in Poles and Russians. *Annals of Human Genetics* 66(Pt 4):261-283.

Malyarchuk BA, Vanecek T, Perkova MA, Derenko MV, and Sip M. 2006. Mitochondrial DNA variability in the Czech population, with application to the ethnic history of Slavs. *Human Biology* 78(6):681-696.

Mange AP, and Mange EJ. 1990. *Genetics: Human Aspects*. Sunderland, Massachusetts: Sinauer Associates.

Manzano C, Aguirre AI, Iriondo M, Martin M, Osaba L, and de la Rua C. 1996a. Genetic polymorphisms of the Basques from Gipuzkoa: genetic heterogeneity of the Basque population. *Annals of Human Biology* 23(4):285-296.

Manzano C, Aguirre AI, Madoz P, Ribo G, Osaba L, Moreno P, and De la Rua C. 1993a. New contribution to the genetics of the Basques: heterogeneity in the esterase D subtype distribution. *Human Heredity* 43(4):219-222.

Manzano C, de RC, Iriondo M, Mazon LI, Vicario A, and Aguirre A. 2002. Structuring the genetic heterogeneity of the Basque population: a view from classical polymorphisms. *Human Biology* 74(1):51-74.

Manzano C, Moral P, De la Rua C, and Moreno P. 1993b. Serum protein polymorphisms (GC, TF, and PI subtypes) in the Basque population of Alava. *Human Heredity* 43(2):121-125.

Manzano C, Orue JM, and de la Rua C. 1996b. The "Basqueness" of the Basques of Alava: a reappraisal from a multidisciplinary perspective. *American Journal of Physical Anthropology* 99(2):249-258.

Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Hurley CK, Mach B and others. 2005. Nomenclature for Factors of the HLA System, 2004. *Human Immunology* 66(5):571-636.

Martinez OP, Longman-Jacobsen N, Davies R, Chung EK, Yang Y, Gaudieri S, Dawkins RL, and Yu CY. 2001. Genetics of human complement component C4 and evolution the central MHC. *Frontiers in Bioscience* 6:D904-913.

Martinez-Laso J, De Juan D, Martinez-Quiles N, Gomez-Casado E, Cuadrado E, and Arnaiz-Villena A. 1995a. The contribution of the HLA-A, -B, -C and -DR, -DQ DNA typing to the study of the origins of Spaniards and Basques. *Tissue Antigens* 45(4):237-245.

Martinez-Laso J, de Juan D, Martinez-Quiles N, Gomez-Casado E, Cuadrado E, and Arnaiz-Villena A. 1995b. The contribution of the HLA-A, -B, -C, and -DR, -DQ DNA typing to the study of the origins of Spaniards and Basques. *Tissue Antigens* 45:237-245.

McEvoy B, Richards M, Forster P, and Bradley DG. 2004. The Longue Duree of genetic ancestry: multiple genetic marker systems and Celtic origins on the Atlantic facade of Europe. *American Journal of Human Genetics* 75(4):693-702.

Mehra NK, and Kaur G. 2003. MHC-based vaccination approaches: progress and perspectives. *Expert Reviews in Molecular Medicine* 2003:1-17.
Miller, MP. 2005. Alleles in space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity* 96(6):722-724.
Miller, MP, Bellinger, MR, Forsman, ED, Haig, SM. 2006. Effects of historical climate change, habitat connectivity, and vicariance on genetic structure and diversity across the range of the red tree vole (*Phenacomys longicaudus*) in the Pacific Northwestern United States. *Molecular Ecology* 15(1):145-159.
Mills KA, Even D, and Murray JC. 1992. Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). *Human Molecular Genetics* 1(9):779.
Morant GM. 1929. A contribution to Basque craniometry. *Biometrika* 21(1/4):67-84.
Moreno ME, de Pablo MR, Vilches C, Kreisler M, Ohayon E, and Cambon-Thomsen A. 1991. Analysis of class I and II antigens in Basque and Spanish Gypsy populations. In: Tsuji K, Aizawa M, and Sasazuki T, editors. HLA 1991: *Proceedings of the Eleventh International Histocompatibility Workshop and Conference*. Oxford: Oxford University Press. p 645-648.
Mourant AE. 1947. The blood groups of the Basques. *Nature* 160(4067):505-506.
Mourant AE. 1974. Les groupes sanguins des Basques. *Cahiers d'anthropologie et d'ecologie humaine* II((3-4)):149-151.
Nasidze I, Ling EY, Quinque D, Dupanloup I, Cordaux R, Rychkov S, Naumova O, Zhukova O, Sarraf-Zadegan N, Naderi GA and others. 2004. Mitochondrial DNA and Y-chromosome variation in the caucasus. *Annals of Human Genetics* 68(Pt 3):205-221.
Nasidze I, Sarkisian T, Kerimov A, and Stoneking M. 2003. Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. *Human Genetics* 112(3):255-261.
Nasidze I, and Stoneking M. 2001. Mitochondrial DNA variation and language replacements in the Caucasus. *Proceedings of Biological Sciences* 268(1472):1197-1206.
Nei M. 1972. Genetic Distance between Populations. *The American Naturalist* 106(949):283-292.
Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA* 70(12):3321-3323.
Nei M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press. x, 512 p. p.
Nei M, and Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences USA* 76(10):5269-5273.
Nijenhuis LE. 1956. Blood group frequencies in French Basques. *Acta Genetica, Basel* 6:531-535.
Novelletto A. 2007. Y-chromosome variation in Europe: continental and local processes in the formation of the extant gene pool. *Annals of Human Biology* 34(2):139-172.

Ohayon E, de Mouzon A, Hauptmann G, Klein J, Abbal M, Constans J, Mayer S, and Ducos J. 1980. High frequency of the properdin factor Bf F1 and its linkage to HLA in French Basques. *Journal of Immunogenetics* 7(6):441-445.

Oldroyd NJ, Urquhart AJ, Kimpton CP, Millican ES, Watson SK, Downes T, and Gill PD. 1995. A highly discriminating octoplex short tandem repeat polymerase chain reaction system suitable for human individual identification. *Electrophoresis* 16(3):334-337.

Ota T. 1993. DISPAN: Genetic Distance and Phylogenetic Analysis. University Park, PA: Institute of Molecular Evolutionary Genetics, Pennsylvania State University.

Pawlowski R, and Maciejewska A. 2000. Forensic validation of a multiplex containing nine STRs--population genetics in northern Poland. *International Journal of Legal Medicine* 114(1-2):45-49.

Pena JA, Morales B, and Calderon R. 1997. New method for comparing levels of microdifferentiation: application to migration matrices of two populations from the Basque Country (Spain). *Human Biology* 69(3):329-344.

Pereira L, Richards M, Goios A, Alonso A, Albarran C, Garcia O, Behar DM, Golge M, Hatina J, Al-Gazali L and others. 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Research* 15(1):19-24.

Perez De Nanclares G, Bilbao JR, Calvo B, and Castano L. 2000. Analysis of chromosome 6q in Basque families with type 1 diabetes. GEPV-N. Basque-Navarre Endocrinology and Paediatric Group. *Autoimmunity* 33(1):33-36.

Perez-Lezaun A, Calafell F, Clarimon J, Bosch E, Mateu E, Gusmao L, Amorim A, Benchemsi N, and Bertranpetit J. 2000. Allele frequencies of 13 short tandem repeats in population samples from the Iberian Peninsula and northern Africa. *International Journal of Legal Medicine* 113(4):208-214.

Perez-Miranda AM, Alfonso-Sanchez MA, Kalantar A, Garcia-Obregon S, de Pancorbo MM, Pena JA, and Herrera RJ. 2005a. Microsatellite data support subpopulation structuring among Basques. *Journal of Human Genetics* 50(8):403-414.

Perez-Miranda AM, Alfonso-Sanchez MA, Kalantar A, Pena JA, Pancorbo MM, and Herrera RJ. 2005b. Allelic frequencies of 13 STR loci in autochthonous Basques from the province of Vizcaya (Spain). *Forensic Science International* 152(2-3):259-262.

Perez-Miranda AM, Alfonso-Sanchez MA, Pena JA, and Calderon R. 2003. HLA-DQA1 polymorphism in autochthonous Basques from Navarre (Spain): genetic position within European and Mediterranean scopes. *Tissue Antigens* 61(6):465-474.

Perez-Miranda AM, Alfonso-Sanchez MA, Pena JA, de Pancorbo MM, and Herrera RJ. 2005c. Genetic polymorphisms at 13 STR loci in autochthonous Basques from the province of Alava (Spain). *Legal Medicine (Tokyo, Japan)* 7(1):58-61.

Perez-Miranda AM, Alfonso-Sanchez MA, Vidales MC, Calderon R, and Pena JA. 2004. Genetic polymorphism and linkage disequilibrium of the HLA-DP region in Basques from Navarre (Spain). *Tissue Antigens* 64(3):264-275.

Pericic M, Lauc LB, Klaric IM, Rootsi S, Janicijevic B, Rudan I, Terzic R, Colak I, Kvesic A, Popovic D and others. 2005. High-resolution phylogenetic analysis of southeastern Europe traces major episodes of paternal gene flow among Slavic populations. *Molecular Biology and Evolution* 22(10):1964-1975.

Piazza A. 1993. Who are the Europeans? *Science* 260:1767-1769.

Piazza A, Cappello N, Olivetti E, and Rendine S. 1988a. The Basques in Europe: A genetic analysis. *Munibe (Antropologia y Arqueologia)* Supplemento 6:169-177.

Piazza A, Cappello N, Olivetti E, and Rendine S. 1988b. The Basques in Europe: a genetic analysis. *Munibe (Antropolgia y Arqueologia)* Supplement 6:169-177.

Pinhasi R, Foley RA, and Lahr MM. 2000. Spatial and temporal patterns in the Mesolithic-Neolithic archaeological record of Europe. In: Renfrew C, and Boyle K, editors. *Archaeogenetics: DNA and the Population Prehistory of Europe*. Cambridge: McDonald Institute for Archaeological Research. p 45-56.

Planas J, M. Fuste, J. Vinas, J.L. Irizar. 1966. Haptoglobin types in the Iberian Peninsula. *Acta genetica, Basel* 16:371-376.

Polzin T, Daneschmand S.V. 2003. On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters* 31:12-20.

Price P, Witt C, Allcock R, Sayer D, Garlepp M, Kok CC, French M, Mallal S, and Christiansen F. 1999. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunolgy Review* 167:257-274.

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, and Balloux F. 2005. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Current Biology* 15(11):1022-1027.

Quintana-Murci L, Semino O, Minch E, Passarimo G, Brega A, and Santachiara-Benerecetti AS. 1999. Further characteristics of proto-European Y-chromosomes. *European Journal of Human Genetics* 7(5):603-608.

Raymond CK, Kas A, Paddock M, Qiu R, Zhou Y, Subramanian S, Chang J, Palmieri A, Haugen E, Kaul R and others. 2005. Ancient haplotypes of the HLA Class II region. *Genome Research* 15(9):1250-1257.

Renfrew, Colin (1987), *Archaeology and Language. The Puzzle of Indo-European Origins*, London, J. Cape.

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, and Sykes B. 1996. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *American Journal of Human Genetics* 59(1):185-203.

Richards M, Macaulay V, Sykes B, Pettitt P, Hedges R, Forster P, and Bandelt H-J. 1997. Reply to Cavalli-Sforza and Minch. *American Journal of Human Genetics* 61:251-254.

Richards M, Macaulay V, Torroni A, and Bandelt HJ. 2002. In search of geographical patterns in European mitochondrial DNA. *American Journal of Human Genetics* 71(5):1168-1174.

Rittner C, and Bertrams J. 1981. On the significance of C2, C4, and factor B polymorphisms in disease. *Human Genetics* 56:235-247.

Rogers AR. 1995. Genetic Evidence for a Pleistocene Population Explosion. *Evolution* 49(4):608-615.

Rogers AR, and Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* 9(3):552-569.

Roostalu U, Kutuev I, Loogvali EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, and Villems R. 2007. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. *Molecular Biology and Evolution* 24(2):436-448.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O and others. 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *American Journal of Human Genetics* 75(1):128-137.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G and others. 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics* 67(6):1526-1543.

Rowold DJ, Luis JR, Terreros MC, and Herrera RJ. 2007. Mitochondrial DNA geneflow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. *Journal of Human Genetics* 52(5):436-447.

Rychkov YG. 2000. *Gene Pool of Populations of Russia and Contiguous Countries*. Saint Petersburg: Nauka.

Ryder LP, Andersen E, and Svejgaard A. 1978. An HLA map of Europe. *Human Heredity* 28(3):171-200.

Salas A, Comas D, Lareu MV, Bertranpetit J, and Carracedo A. 1998. mtDNA analysis of the Galician population: a genetic edge of European variation. *European Journal of Human Genetics* 6(4):365-375.

Sanchez-Velasco P, Gomez-Casado E, Martinez-Laso J, Moscoso J, Zamora J, Lowy E, Silvera C, Cemborain A, Leyva-Cobian F, and Arnaiz-Villena A. 2003. HLA alleles in isolated populations from North Spain: origin of the Basques and the ancient Iberians. *Tissue Antigens* 61(5):384-392.

Sanchez-Velasco P, and Leyva-Cobian F. 2001. The HLA class I and class II allele frequencies studied at the DNA level in the Svanetian population (Upper Caucasus) and their relationships to Western European populations. *Tissue Antigens* 58(4):223-233.

Santos C, Montiel R, Angles N, Lima M, Francalacci P, Malgosa A, Abade A, and Aluja MP. 2004. Determination of human caucasian mitochondrial DNA

haplogroups by means of a hierarchical approach. *Human Biology* 76(3):431-453.

Satta Y, O'HUigin C, Takahata N, and Klein J. 1993. The synonymous substitution rate of the major histocompatibility complex loci in primates. *Proceedings of the National Academy of Sciences USA* 90(16):7480-7484.

Schanfield MS, Baylerian R, Maiquez J, and Carbonell F. 1981. Immunoglobulin allotypes in European populations. IV. Gm, Am and Km allotypic markers in Valencia, Spain. *Journal of Immunogenetics* 8(6):529-532.

Schneider S, and Excoffier L. 1999. Estimation of Past Demographic Parameters From the Distribution of Pairwise Differences When the Mutation Rates Vary Among Sites: Application to Human Mitochondrial DNA. *Genetics* 152(3):1079-1089.

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ and others. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *American Journal of Human Genetics* 74(5):1023-1034.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S and others. 2000a. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y-chromosome perspective. *Science* 290(5494):1155-1159.

Semino O, Passarino G, Quintana-Murci L, Liu A, Beres J, Czeizel A, and Santachiara-Benerecetti AS. 2000b. MtDNA and Y-chromosome polymorphisms in Hungary: inferences from the palaeolithic, neolithic and Uralic influences on the modern Hungarian gene pool. *European Journal of Human Genetics* 8(5):339-346.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A and others. 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *American Journal of Human Genetics* 78(2):202-221.

Sharma V, and Litt M. 1992. Tetranucleotide repeat polymorphism at the D21S11 locus. *Human Molecular Genetics* 1(1):67.

Shen P, Lavi T, Kivisild T, Chou V, Sengun D, Gefel D, Shpirer I, Woolf E, Hillel J, Feldman MW and others. 2004. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Human Mutation* 24(3):248-260.

Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, and Stoneking M. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. *Human Biology* 66(5):761-775.

Shriver MD, Jin L, Boerwinkle E, Deka R, Ferrell RE, and Chakraborty R. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology and Evolution* 12(5):914-920.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139(1):457-462.
Sokal RR. 1988. Genetic, geographic, and linguistic distances in Europe. *Proceedings of the National Academy of Sciences USA* 85(5):1722-1726.
Sokal RR. 1991a. Ancient movement patterns determine modern genetic variances in Europe. *Human Biology* 63(5):589-606.
Sokal RR. 1991b. The continental population structure of Europe. *Annual Review of Anthropology* 20:119-140.
Sokal RR, Harding RM, and Oden NL. 1989a. Spatial patterns of human gene frequencies in Europe. *American Journal of Physical Anthropology* 80(3):267-294.
Sokal RR, Jacquez GM, Oden NL, DiGiovanni D, Falsetti AB, McGee E, and Thomson BA. 1993. Genetic relationships of European populations reflect their ethnohistorical affinities. *American Journal of Physical Anthropology* 91(1):55-70.
Sokal RR, and Menozzi P. 1982. Spatial autocorrelation of HLA frequencies in Europe supports demic diffusion of early farmers. *American Naturalist* 119:1-17.
Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim JY, Thomson BA, Vaudor A, Harding RM, and Barbujani G. 1990. Genetics and language in European populations. *American Naturalist* 135:157-175.
Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim JY, and Vaudor A. 1989b. Genetic differences among language families in Europe. *American Journal of Physical Anthropology* 79(4):489-502.
Sokal RR, Oden NL, and Thomson BA. 1988. Genetic changes across language boundaries in Europe. *American Journal of Physical Anthropology* 76(3):337-361.
Sokal RR, Oden NL, and Thomson BA. 1992a. Origins of the Indo-Europeans: genetic evidence. *Proceedings of the National Academy of Sciences USA* 89(16):7669-7673.
Sokal RR, Oden NL, and Thomson BA. 1999a. A problem with synthetic maps. *Human Biology* 71(1):1-13; discussion 15-25.
Sokal RR, Oden NL, and Thomson BA. 1999b. Problems with synthetic maps remain: reply to Rendine *et al. Human Biology* 71(3):447-453.
Sokal RR, Oden NL, Walker J, Di Giovanni D, and Thomson BA. 1996. Historical population movements in Europe influence genetic relationships in modern samples. *Human Biology* 68(6):873-898.
Sokal RR, Oden NL, and Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* 351(6322):143-145.
Sokal RR, Wilson C, and Oden NL. 1992b. Patterns of population spread. *Nature* 355(6357):214.
Spinola H, Middleton D, and Brehm A. 2005. HLA genes in Portugal inferred from sequence-based typing: in the crossroad between Europe and Africa. *Tissue Antigens* 66(1):26-36.

Stevenson JC, and Schanfield MS. 1981. Immunoglobulin allotypes in European populations: III. Gm, Am and Km allotypes in people of European ancestry in the United states. *Human Biology* 53(4):521-542.

Strachan T, Sodoyer R, Damotte M, and Jordan BR. 1984. Complete nucleotide sequence of a functional class I HLA gene, HLA-A3: implications for the evolution of HLA genes. *EMBO Journal* 3(4):887-894.

Straus LG. 1991a. Human geography of the Late Upper Paleolithic in Western Europe: Present state of the question. *Journal of Anthropological Research* 47(2):259-278.

Straus LG. 1991b. Southwestern Europe at the Last Glacial Maximum. *Current Anthropology* 32(2):189-199.

Sturrock K, and Rocha J. 2000. A multidimensional scaling stress evaluation table. *Field Methods* 12(1):49-60.

Svejgaard A. 1979. *The HLA system: An introductory survey*. New York: S. Karger. vi, 111 p. p.

Sykes B. 1999. The molecular genetics of European ancestry. *Philosophical Transactions of the Royal Society B: Biological Sciences* 354:131-139.

Sykes B. 2003. European Ancestry: The mitochondrial landscape. In: Ammerman AJ, and Biagi P, editors. *The Widening Harvest: The Neolithic Transition in Europe*. Boston, MA: Archaeological Institute of America. p 315-326.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.

Takahata N. 1993. Allelic genealogy and human evolution. *Molecular Biology and Evolution* 10(1):2-22.

Takahata N, Satta Y, and Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130(4):925-938.

Tamura K, Dudley J, Nei M, and Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B and others. 1998. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *American Journal of Human Genetics* 62(5):1137-1152.

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E and others. 2001. A signal, from human mtDNA, of postglacial recolonization in Europe. *American Journal of Human Genetics* 69(4):844-852.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, and Wallace DC. 1996. Classification of European mtDNAs from an analysis of three European populations. *Genetics* 144(4):1835-1850.

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, and Wallace DC. 1994. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *American Journal of Human Genetics* 55(4):760-776.

Trask RL. 1996. Origin and relatives of the Basque language: Review of the evidence. In: Hualde JI, Lakarra JA, and Trask RL, editors. Towards a History of the Basque Language. Amsterdam: John Benjamins Publishing Company. p 65-99.

Tsuji K, Aizawa M, and Sasazuki T. 1992. *HLA 1991 : Proceedings of the Eleventh International Histocompatibility Workshop and Conference, held in Yokohama, Japan, 6-13 November, 1991*. Oxford ; New York: Oxford University Press.

Urla J. 1993. Cultural politics in an age of statistics: Numbers, nations, and the making of Basque identity. *American Ethnologist* 20(4):818-843.

Urquhart A, Oldroyd NJ, Kimpton CP, and Gill P. 1995. Highly discriminating heptaplex short tandem repeat PCR system for forensic identification. *BioTechniques* 18(1):116-118, 120-111.

Vergnes H, Constans J, Quilici JC, Lefevre-Witier P, Sevin J, and Stevens M. 1980. Study of red blood cell and serum enzymes in five Pyrenean communities and in a Basque population sample. *Human Heredity* 30(3):171-180.

Ward RH, Frazier BL, Dew-Jager K, and Paabo S. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences USA* 88(19):8720-8724.

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S and others. 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences USA* 98(18):10244-10249.

Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, and Goldstein DB. 2001. Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proceedings of the National Academy of Sciences USA* 98(9):5078-5083.

Wright S. 1951. Genetic structure of populations. *Annals of Eugenics* 15:323-354.

Young K. 2007. Classical genetic markers and Basque origins. Lawrence, KS: University of Kansas.

Zilhao J. 1998. On logical and empirical aspects of the Mesolithic-Neolithic Transition in the Iberian Peninsula. *Current Anthropology* 39(5):690-698.

Zilhao J. 2000. From the Mesolithic to the Neolithic in the Iberian Peninsula. In: Price TD, editor. *Europe's First Farmers*. Cambridge: Cambridge University Press. p 144-182.

Zilhao J. 2003. The Neolithic transition in Portugal and the role of demic diffusion in the spread of agriculture across West Mediterranean Europe. In: Ammerman AJ, and Biagi P, editors. *The Widening Harvest: The Neolithic transition in Europe*. Boston, MA: Archaeological Institute of America. p 207-223.

Zlojutro M, Roy R, Palikij J, and Crawford MH. 2006. Autosomal STR variation in a Basque population: Vizcaya Province. *Human Biology* 78(5):599-618.

Appendix 1. Populations used in the present study.

Populations	***Language Family***	***Label***	***Genetic Marker System***			
			Classical	*Autosomal STRs*	*YSTRs*	*mtDNA*
Abkhazia	North Caucasian (West Caucasian)	AB	●		●	
Albania	Indo-European (Albanian)	AN			●	●
Algeria	Afro-Asiatic (Semetic)	AL				●
Algeria - Berbers	Afro-Asiatic (Berber)	AL - B	●			
Austria	Indo-European (Germanic)	AU	●	●		●
Armenia	Indo-European (Armenian)	AR	●		●	●
Azerbaijan	Altaic (Turkic)	AZ			●	●
Baronnies - Esparros	Indo-European (Italic)	B-ES	●			
Balearic Islands	Indo-European (Italic)	BI	●			
Basque-Alava	Basque	BS-A	●	● PS	● PS	● PS
Basque – Basses-Pyrenees	Basque	BS-BP	●			
Basque - France	Basque	BS-FR	●			
Basque – Guipuzcoa	Basque	BS-G	●	● PS	● PS	● PS
Basque – Labourd	Basque	BS-L	●			
Basque - Navarre	Basque	BS-N	●	● PS	● PS	● PS
Basque - Soule	Basque	BS-S	●			
Basque - Spain	Basque	BS-SP	●	●	●	●
Basque – Basses-Navarre	Basque	BS-BN	●			
Basque - Vizcaya	Basque	BS-V	●	● PS	● PS	● PS
Bearn – Luz St. Sauveur	Indo-European (Italic)	BN-LS	●			
Bearn – Vallee de l'Ouzom	Indo-European (Italic)	BN-VO	●			
Belgium	Indo-European (Germanic)	BG	●	●	●	●
Bigorre - Gavarnie	Indo-Eurpean (Italic)	BI-G	●			
Bigorre - Barages	Indo-European (Italic)	BI-B	●			
Bosnia	Indo-European (Slavic)	BO		●	●	●
Brittany	Indo-European (Italic)	BR				●
Bulgaria	Indo-European (Slavic)	BU			●	●
Caucasus – Abazinian	North Caucasian (West Caucasian)	AN			●	

Populations	*Language Family*	*Label*	*Genetic Marker System*			
Caucasus – Adygei	North Caucasian (West Caucasian)	AD				●
Caucasus – Darginian	North Caucasian (East Caucasian)	DG			●	
Caucasus – Ingushian	North Caucasian (East Caucasian)	IN			●	
Caucasus – Kabardinian	North Caucasian (West Caucasian)	KB			●	
Caucasus – Lezginian	North Caucasian (East Caucasian)	LZ			●	
Caucasus – Rutulian	North Caucasian (East Caucasian)	RT			●	
Chechnya	North Caucasian (East Caucasian)	CC			●	
Cornwall	Indo-European (Germanic)	CW	●			●
Corsica	Indo-European (Italic)	CS	●			
Croatia	Indo-European (Slavic)	CR			●	
Czech Republic	Indo-European (Slavic)	CZ	●			
Denmark	Indo-European (Germanic)	DN	●			●
Egypt	Afro-Asiatic (Semitic)	EG	●	●		
England	Indo-European (Germanic)	EN	●			●
Estonia	Uralic (Finnic)	ES	●			●
Finland	Uralic (Finnic)	FN	●			●
France	Indo-European (Italic)	FR	●		●	●
Georgia	Kartvelian (Georgian)	GA		●	●	●
Germany	Indo-European (Germanic)	GM	●	●	●	●
Greece	Indo-European (Greek)	GR	●	●	●	●
Hungary	Uralic (Ugric)	HU	●	●	●	●
Ireland	Indo-European (Celtic)	IR	●	●	●	●
Italy	Indo-European (Italic)	IT	●		●	●
Karelia	Uralic (Finnic)	KR				●
Kosovo	Indo-European (Italic)	KO		●		
Kosovo – Albanian	Indo-European (Albanian)	KO-A			●	
Lapps	Uralic (Saami)	LP	●		●	
Latvia	Indo-European (Baltic)	LT			●	
Libya	Afro-Asiatic (Semitic)	LB	●			
Lithuania	Indo-European (Baltic)	LH			●	

Populations	*Language Family*	*Label*	*Genetic Marker System*			
Morocco	Afro-Asiatic (Arabic and Berber)	MR	●	●		
Morocco – Arab	Afro-Asiatic (Arabic)	MR-A				●
Morocco – Berber	Afro-Asiatic (Berber)	MR-B				●
Netherlands	Indo-European (Germanic)	NT	●			
North Ossetia	Indo-European (Indo-Iranian)	NO			●	●
Norway	Indo-European (Germanic)	NW	●			●
Oman	Afro-Asiatic (Semetic)	OM		●		
Poland	Indo-European (Slavic)	PL	●	●	●	●
Portugal	Indo-European (Italic)	PT	●	●	●	●
Portugal – Roma	Indo-European (Indo-Iranian)	PT-R			●	
Pyrenees – Aran Valley	Indo-European (Italic)	P-AV	●			
Pyrenees - Capcir	Indo-European (Italic)	P-CP	●			
Pyrenees – Pays de Sault	Indo-European (Italic)	P-PS	●			
Pyrenees - Camurac	Indo-European (Italic)	P-CM	●			
Romania	Indo-European (Italic)	RM	●		●	●
Russia	Indo-European (Slavic)	RU	●	●	●	●
Sardinia	Indo-European (Italic)	SR	●			●
Scotland	Indo-European (Celtic)	SC	●	●		●
Scotland – Orkney Islands	Indo-European (Celtic)	SC-O	●			
Serbia	Indo-European (Slavic)	SB		●	●	
Sicily	Indo-European (Italic)	SI				●
Slovenia	Indo-European (Slavic)	SL	●	●		●
Spain - Alpujarra	Indo-European (Italic)	SP-AJ	●			
Spain - Andalusia	Indo-European (Italic)	SP-AN	●	●	●	●
Spain - Barcelona	Indo-European (Italic)	SP-B			●	
Spain - Cantabria	Indo-European (Italic)	SP-CN	●	●	●	
Spain - Castile	Indo-European (Italic)	SP-CS	●			●
Spain - Cataluna	Indo-European (Italic)	SP-CT	●			
Spain – Catalonia	Indo-European (Italic)	SP-CA		●		●
Spain – Central Meseta	Indo-European (Italic)	SP-CM	●			
Spain – Central Plateau	Indo-European (Italic)	SP-CP	●			
Spain - Galicia	Indo-European (Italic)	SP-GL	●	●	●	●

Populations	*Language Family*	*Label*	*Genetic Marker System*			
Spain - Leon	Indo-European (Italic)	SP-LN	●			●
Spain – Madrid	Indo-European (Italic)	SP-MD	●			
Spain - Murcia	Indo-European (Italic)	SP-MU	●	●		
Spain - Valencia	Indo-European (Italic)	SP-VL	●	●	●	
Svani - Georgia	Kartvelian (Georgian)	SV	●			
Sweden	Indo-European (Germanic)	SD	●		●	●
Switzerland	Indo-European (Italic)	SW	●	●	●	●
Tunisia	Afro-Asiatic (Semitic)	TN	●			●
Tunisia – Berber	Afro-Asiatic (Berber)	TN-B				●
Turkey	Altaic (Turkic)	TU	●	●		
Wales	Indo-European (Celtic)	WL	●			●
Yemen	Afro-Asiatic (Semitic)	YM	●	●		
Yugoslavia	Indo-European (Slavic)	YU	●			

Appendix 2. HLA Class I Alleles.

Class I					
HLA-A – Common Alleles			***HLA-B – Common Alleles***		
Allele	Serological Specificity	Sequences	Allele	Serological Specificity	Sequences
A1		19	B7		30
A2		109	B8		16
A3		9	B13		10
A11		13	B14		6
A23	A9	9	B15		73
A24	A9	36	B18		18
A25	A10	4	B27		25
A26	A10	18	B35		44
A29	A19	6	B37		5
A30	A19	12	B38	B16	8
A31	A19	8	B39	B16	26
A32	A19	7	B40		44
A33	A19	6	B41		6
A34	A10	4	B42		4
A36		4	B44	B12	31
A43		1	B45	B12	6
A66	A10	4	B46		2
A68	A28	1	B47		4
A74	A19	8	B48		7
A80		1	B49	B21	3
HLA-C			B50	B21	4
Allele	Serological Specificity	Sequences	B51	B5	29
Cw1		7	B52	B5	4
Cw2		5	B53	B5	9
Cw3		15	B54	B22	2
Cw9	Cw3	4	B55	B22	12
Cw10	Cw3	5	B56	B22	8
Cw4		10	B57	B17	9
Cw5		5	B58	B17	10

HLA-C			***HLA-B – Common Alleles***		
Allele	Serological Specificity	Sequences	Allele	Serological Specificity	Sequences
Cw6		7	B58	B17	6
Cw7		16	B59		1
Cw8		9	B60	B40	6
Cw12		8	B61	B40	8
Cw14		5	B62	B15	17
Cw15		11	B63	B15	2
Cw16		3	B67		2
			B73		1
			B78		5
			B81		2
			B82		2
			B83		1

Adapted from Svejgaard et al. (1979), and Marsh et al. (2005).

Appendix 3 HLA Class II Alleles.

Class II					
DR (DRB1, DRB3/4/5)			***DQ (DQA1-DQB1)***		
Allele	Serological Specificity	Sequences	Allele	Serological Specificity	Sequences
DR1		8	DQ1		3
DR15	DR2	13	DQ5	DQ1	7
DR16	DR2	8	DQ6	DQ1	12
DR3		23	DQ2		4
DR17	DR3	7	DQ3		1
DR18	DR3	3	DQ7	DQ3	3
DR4		44	DQ8	DQ3	6
DR11	DR5	43	DQ9	DQ3	2
DR12	DR5	8	DQ4		2
DR13	DR6	52	***DP (DPA1-DPB1)***		
DR14	DR6	43	Allele	Serological Specificity	Sequences
DR7		6	DP1		12
DR8		24	DP2		17
DR9		2	DP3		4
DR10		1	DP4		5
DR19		1	DP5		2
			DP6		2

Adapted from Svejgaard et al. (1979), and Marsh et al. (2005).

Appendix 4. Informed Consent Statement

The Department of Anthropology at the University of Kansas supports the practice of protection for human subjects participating in research. The following information is provided for you to decide whether you wish to participate in the present study. You should be aware that even if you agree to participate, you are free to withdraw without penalty.

We are interested in studying the genealogies of Basque migrants to the United States as well as the inheritance of DNA markers, particularly mitochondrial DNA and Y-chromosome, in human populations. You will be participating in one session that will require completion of a genealogical questionnaire and collection of one buccal swab. This technique consists of a sterile equivalent of a toothbrush being gently stroked across the cheeks and gums.

Data from the questionnaires will be strictly private and number coded. The DNA extracted from the buccal swabs will be used solely to study those DNA markers. All DNA will be used up in the analyses. Only personnel working directly on this project will have access to the DNA.

Your participation is solicited although strictly voluntary. We assure you that your name will not be associated in any way with the research findings. The information will be identified only by a code number.

If you would like additional information concerning this study before or after it is complete, please feel free to contact me by phone or mail.

Sincerely,

Dr. Arantza Gonzalez Apraiz
Department of Anthropology
622 Fraser Hall
University of Kansas
Lawrence, Kansas 66045
Phone: 00 1 785-864-2606
Fax: 00 1 785-864-5224

Dr. Michael H. Crawford
Department of Anthropology
622 Fraser Hall
University of Kansas
Lawrence, Kansas 66045
Phone: 00 1 785-864-4170
Fax: 00 1 785-864-5224

Signature of subject agreeing to participate

With my signature I affirm that I am at least 18 years of age and have received a copy of the consent form to keep.

Signature of parent or guardian if individual under 18 years of age