Autosomal short tandem repeat genetic variation of the Basques in Spain

Aim To examine population genetic structure and hypotheses of the origin of the modern Basque population in Spain using autosomal short tandem repeat (STR) data from individuals living in 27 mountain villages in the provinces of Alava, Vizcaya, Guipuzcoa, and Navarre, by comparing Basque autosomal STR variation with that of neighboring populations in Europe, as well as proposed ancestral populations in North Africa and the Caucasus.

Methods Allele frequencies for 9 autosomal STR loci (D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA, and vWA) and several population genetic parameters were determined for the 4 provinces in the Basque region of Spain (n = 377). Heterozygosity within the Basque population was measured using a locus-by-locus analysis of molecular variance. Relationships between the Basques and other populations were examined using a multidimensional scaling (MDS) plot of Shriver’s $D_{ST}$ distance matrix.

Results Heterozygosity levels in the Basque provinces were on the low end of the European distribution (0.805-0.812). The MDS plot of genetic distances revealed that the Basques differed from both the Caucasian and North African populations with respect to autosomal STR variation.

Conclusions Autosomal STR analysis does not support the hypotheses of a recent common ancestor between the Basques and populations either from the Caucasus or North Africa.
The question of Basque origins has interested scholars since the 1800s, when Aranzadi suggested, based on cranial morphology, that the Basques were an ancient relic population (1). The Basque language, Euskara, is most widely accepted as an isolate, unrelated to any other extant language in Europe (2). Many hypotheses of a relationship between Basques and other populations have been put forward based on linguistic analyses. These proposed linguistic connections include the ancient languages of Iberian, Minoan, Etruscan, Pictish, Sumerian, and Aquitanian, as well as extinct Uralic (such as Finnish), Caucasian (such as Georgian), African (especially Berber), and Native American languages, and even Japanese (3-10). It has been suggested that populations in the Basque region and the Caucasus, which speak non-Indo-European agglutinative languages, could be remnants of a 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 (6). Alternatively, 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, but discoveries of Iberian inscriptions which were not translatable using Euskara weakened this hypothesis on linguistic grounds. 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 (11). There are also linguists who conclude that Basque is an autochthonous language, which developed in situ in the Iberian Peninsula, and once had a wider range, but has also been less affected than other languages in historical times (2).

Studies of human blood types in the mid-20th century bore out the distinctiveness of the Basques, distinguishing them from other European populations by a low frequency of ABO*B (1.1%) and a high frequency of RH*cde (between 30.5%-35.6%) (12-15). Since then, the Basque population either used small samples that were often collected in urban areas of a single province (when collection location was reported) to study relationships between the Basques and other populations (49,54,58-60). When larger samples were collected (48,50-52,57), the studies most often presented allele frequencies and population genetic parameters for only a few loci (Table 1). This caused forensic databases provide a wealth of comparative data.

The present study characterizes autosomal STR genotypes from the Basque region of Spain to examine population substructure and genetic relationships with other groups, including testing of the proposed genetic affinities between the Basques and populations in the Caucasus and North Africa. We predict that if the Basques share a common ancestor (or have experienced more recent migration and resultant gene flow) with either populations in the Caucasus or North Africa, allele frequencies of autosomal STR loci will be similar among the Basques and these proposed related populations and genetic distances between these populations will be low. Alternatively, if the Basques are an autochthonous European population (with no recent gene flow from groups in North Africa or the Caucasus), autosomal STR frequencies will be within the range of other populations on the European continent, and the Basques will be more genetically similar to other European groups. Previous studies presenting STR data from the Basque population either used small samples that were often collected in urban areas of a single province (when collection location was reported) to study relationships between the Basques and other populations (49,54,58-60). When larger samples were collected (48,50-52,57), the studies most often presented allele frequencies and population genetic parameters for only a few loci (Table 1). This study represents one of the most comprehensive samples of Basques yet analyzed for autosomal STR variation, with 377 individuals in 27 mountain villages from throughout the Basque region of Spain.

MATERIALS AND METHODS

Population

To test the hypotheses of genetic similarity between Basques and other populations, buccal DNA samples were collected from 652 autochthonous (those who claimed 4 Basque grandparents) participants of both sexes, in mountain villages throughout the Basque region of northern Spain. Six villages were sampled in Alava (n= 143), 17 villages in Vizcaya (n = 237), 10 villages in Guipuzcoa (n = 220), and 2 villages in Navarre (n = 52). A
A sub-sample of individuals from 27 villages (N = 478) was screened for autosomal STR genotypes (Table 1) (Figure 1). The samples were collected during summer field seasons between 2000 and 2002. Lab analysis was performed in 2003-2004, and statistical analysis was conducted between 2005-2007 (with some additional analyses done as part of KY’s dissertation in 2009 and revised analyses for publication in 2011). This study was approved by the University of Kansas Human Subjects Committee (HSCL #11955), and participants provided written informed consent.

DNA analysis

DNA extraction was performed using a standard phenol:chloroform protocol. A portion of each sample was reserved for autosomal STR analysis using the Applied Biosystems Profiler Plus Kit (Foster City, CA; USA). The samples were characterized for 9 STR loci, including D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA, and vWA, plus the sex-determining amelogenin locus, using a multiplex PCR procedure according to manufacturer’s instructions (61). Amplified products were detected using an ABI 377 DNA sequencer. DNA fragments were sized and genotyped with GeneScan 3.1 and Genotyper 2.5 software (Applied Biosystems).

Comparative population data

To test hypotheses of population origins, autosomal STR data from our sample were compared to geographically and/or proposed related populations, including others

<table>
<thead>
<tr>
<th>Table 1. Previous autosomal short tandem repeat studies in the Basque population.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
</tr>
<tr>
<td>Rural autochthonous Basque females</td>
</tr>
<tr>
<td>Autochthonous Basques</td>
</tr>
<tr>
<td>Basque autochthonous residents</td>
</tr>
<tr>
<td>Unrelated autochthonous Basques</td>
</tr>
<tr>
<td>Basque Country autochthonous individuals</td>
</tr>
<tr>
<td>Unrelated “native Basques”</td>
</tr>
<tr>
<td>Unrelated Basque country residents</td>
</tr>
<tr>
<td>Unrelated autochthonous Basque students</td>
</tr>
<tr>
<td>Unrelated autochthonous Basques</td>
</tr>
<tr>
<td>Unrelated autochthonous Basques</td>
</tr>
<tr>
<td>Unrelated autochthonous Basques</td>
</tr>
</tbody>
</table>
from the Iberian Peninsula (Andalusia, Cantabria, Catalonia, Galicia, Murcia, Portugal, and Valencia) (54,62-67); Europe (Austria, Belgium, Bosnia, Germany, Greece, Hungary, Tuscany, Poland, Russia, Scotland, Serbia and Montenegro, Slovenia, and Switzerland) (68-81); North Africa (Egypt, Morocco) (82,83); the Middle East (Turkey) (84); and the Caucasus (Georgia) (85).

Statistical analysis

Allele frequencies were estimated using the gene counting method. Expected heterozygosity under Hardy-Weinberg equilibrium was estimated by the method of Guo and Thompson (86), and locus-by-locus analysis of molecular variance (AMOVA) was performed to examine genetic substructure among the Basques, using Arlequin 3.11 software (87). Genetic differentiation between the Basques and 27 comparative populations, including 1 Caucasian and 2 North African groups, was measured using heterozygosity and gene differentiation (GST). GST and gene diversity values for the Basques and comparative populations were calculated using DISPAN software (88). Genetic distances between populations were calculated using Shriver's distance (DSW), an adaptation of Nei's standard distance (D) weighted by the difference in number of repeats between alleles to account for the stepwise mutation pattern of tandem repeat loci. Genetic relationships between groups were examined using a multidimensional scaling (MDS) plot of the genetic distance matrix in NTSYS 2.1 software (89). MDS stress values were evaluated using the criteria of Sturrock and Rocha (90).

RESULTS

Examination of the autosomal STR data revealed that allelic dropout occurred in 21% of the sample (101 of 478), so that not all loci were amplified for every individual. Samples that did not amplify for all loci were removed from the analysis. Allele frequencies for the autosomal STR loci in each of the Basque provinces after correction for allelic dropout are available in the web extra material.

Observed heterozygosity values among the Basques ranged from 0.60526 (D5S818) to 0.92105 (vWA), with both extremes found in Navarre, likely as a result of small sample size (n = 38) (Table 2). In Alava and Vizcaya, only 2 loci had significantly lower heterozygosity values than expected. When the Bonferroni correction for multiple tests was applied, only the D8S1179 locus demonstrated an excess of homozygotes in all provinces, and bolded P values indicate those loci with lower than expected heterozygosity after correction of the data for allelic dropout (Table 2). This suggests that for the other STR loci examined, the expectations of Hardy-Weinberg equilibrium were met (Table 2).

The results of the AMOVA (Table 3) suggested no obvious genetic structuring between provinces, as indicated by the among-groups covariance component (V_{ad} = -0.095). The lack of structure among provinces was confirmed by the global estimate of the fixation index among groups (F_{CT} = -0.0036, Table 3).
A small amount of subdivision was found between villages within provinces (1.309% total variation, $F_{SC} = 0.0131$, $P = 0.001$). Examination of the locus-by-locus results revealed that 3 loci made significant contributions to the differences between villages: D7S820 ($F_{SC} = 0.0332$, $P = 0.023$), vWA ($F_{SC} = 0.0185$, $P = 0.045$), and D18S51 ($F_{SC} = 0.0319$, $P = 0.021$). The majority of variation, however, was found between individuals within villages (99% total variation).

Average heterozygosity values by population ranged from 0.803 in Morocco to a high of 0.820 in Scotland (Table 4). Among the Basques, heterozygosity was lowest in Alava (0.805) and highest in Vizcaya (0.812). This was within the range of heterozygosity values seen in other modern Iberian populations (0.804-0.815). Total gene diversity between subpopulations ($H_I$) was high, ranging from 0.724 for D5S818 to 0.878 for D18S51. However, most of this diversity is explained by variation between individuals within subpopulations ($H_J$). The percentage of gene differentiation between subpopulations relative to the total gene differentiation ($G_{st}$) ranged from a high of 0.009 for D13S317 and D21S11 to a low of 0.006 for D3S1358, FGA, D7S820, vWA, and D18S51.

Visual representation of genetic distances between populations using multidimensional scaling (Figure 2) showed that the Basque groups clustered together on the right side of the plot, near their neighbors in Cantabria. The North African and Georgian populations were found near the bottom center of plot, differentiated from the other European groups. A stress value of 0.169, well below the threshold of 0.317 for 27 populations in two dimensions, demonstrates that the plot is an accurate representation of the genetic distance matrix. A Mantel test of matrix correlation between the original distance matrix and the MDS matrix also demonstrated that the MDS plot was an accurate represent of the genetic distances between populations (correlation coefficient: $r = 0.93498$, $t$ test: $t = 5.7717$, $P = 1.0$).

**DISCUSSION**

Our study of the autosomal STR variation did not support the hypotheses of a recent common ancestor between the Basques and populations either from the Caucasus or North Africa. Allelic dropout was noted for several samples in the present analysis, raising the possibility of a technical or genomic error in the typing of samples. At low sample DNA concentrations, the Profiler Kit is known to preferentially amplify short alleles and homozygotes (93). 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. The excess of homozygotes at D8S1179, even after correction for allelic dropout, was of particular concern. Concordance studies of autosomal STR typing across different multiplex kits have reported issues with the D8S1179 locus in certain populations, principally with alleles 15-18 using the Profiler Plus Kit (94). Reports of D8S1179 from previous studies among Basques give frequency ranges for alleles 15-17 between: 15 (0.110-0.210), 16 (0.010-0.029), and 17 (0.005-0.007) (54,56-58). D8S1179*18 has not been previously reported in Basques. Frequencies for D8S1179 alleles 15-17 in the present study fall within the ranges previously re-

### Table 3. Locus-by-locus analysis of molecular variance of 9 autosomal short tandem repeat loci.

<table>
<thead>
<tr>
<th>Locus</th>
<th>percent variance</th>
<th>$F_{CT}$</th>
<th>$P$</th>
<th>percent variance</th>
<th>$F_{SC}$</th>
<th>$P$</th>
<th>percent variance</th>
<th>$F_{ST}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>-0.026</td>
<td>-0.0003</td>
<td>0.353</td>
<td>1.257</td>
<td>0.0126</td>
<td>0.142</td>
<td>98.768</td>
<td>0.0123</td>
<td>0.153</td>
</tr>
<tr>
<td>FGA</td>
<td>-0.161</td>
<td>-0.0016</td>
<td>0.677</td>
<td>-0.591</td>
<td>-0.0059</td>
<td>0.814</td>
<td>100.752</td>
<td>-0.0075</td>
<td>0.849</td>
</tr>
<tr>
<td>D5S818</td>
<td>0.384</td>
<td>0.0038</td>
<td>0.182</td>
<td>-0.292</td>
<td>-0.0029</td>
<td>0.646</td>
<td>99.908</td>
<td>0.0009</td>
<td>0.544</td>
</tr>
<tr>
<td>D7S820</td>
<td>-1.036</td>
<td>-0.0104</td>
<td>0.939</td>
<td>3.350</td>
<td>0.0332</td>
<td>0.023</td>
<td>97.686</td>
<td>0.0231</td>
<td>0.032</td>
</tr>
<tr>
<td>D8S1179</td>
<td>-0.564</td>
<td>-0.0056</td>
<td>0.904</td>
<td>0.573</td>
<td>0.0057</td>
<td>0.203</td>
<td>99.990</td>
<td>0.0001</td>
<td>0.322</td>
</tr>
<tr>
<td>vWA</td>
<td>0.011</td>
<td>0.0001</td>
<td>0.371</td>
<td>1.851</td>
<td>0.0185</td>
<td>0.045</td>
<td>98.138</td>
<td>0.0186</td>
<td>0.045</td>
</tr>
<tr>
<td>D13S317</td>
<td>-0.318</td>
<td>-0.0032</td>
<td>0.641</td>
<td>0.110</td>
<td>0.0011</td>
<td>0.497</td>
<td>100.208</td>
<td>-0.0021</td>
<td>0.614</td>
</tr>
<tr>
<td>D18S51</td>
<td>-0.546</td>
<td>-0.0055</td>
<td>0.504</td>
<td>3.204</td>
<td>0.0319</td>
<td>0.021</td>
<td>97.342</td>
<td>0.0266</td>
<td>0.010</td>
</tr>
<tr>
<td>D21S11</td>
<td>-0.308</td>
<td>-0.0031</td>
<td>0.602</td>
<td>0.767</td>
<td>0.0076</td>
<td>0.316</td>
<td>99.541</td>
<td>0.0046</td>
<td>0.321</td>
</tr>
<tr>
<td>Global estimates</td>
<td>-0.355</td>
<td>-0.0036</td>
<td>0.878</td>
<td>1.309</td>
<td>0.0131</td>
<td>0.011</td>
<td>99.045</td>
<td>0.0096</td>
<td>0.019</td>
</tr>
</tbody>
</table>

| Covariance estimates | $V_c = -0.095$ | $V_c = 0.351$ | $V_c = 26.517$ |

*Abbreviations: $F_{CT}$ – fixation index among groups; $F_{SC}$ – fixation index among populations within groups; $F_{ST}$ – fixation index within populations.
ported for this population (54,56-58), and we also found no individuals with allele 18. In addition, the locus-by-locus AMOVA demonstrated that the Basque provinces were homogeneous with respect to autosomal STR variation, and the single locus found to not be in HWE (D8S1179) did not significantly contribute to differences between provinces. Therefore, we do no believe that the failure to meet HWE in this case represents a technical error, and we included the D8S1179 allele in the interpopulation analyses.

The present analysis of autosomal STR variation does not support either the Caucasian or Vasco-Iberian hypothesis of Basque origins. 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, these similarities have been attributed either to poor interpretation, a shared Euro-African substratum, or similarities in the evolution of language itself (11). Examination of the literature on the Basque-Caucasian hypothesis demonstrates little support from the genetic evidence (6,23,95-98). Cluster analysis of classical genetic markers showed 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 (99). Comparison of Basque and populations from the Caucasus 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 (6). Analysis of HLA data showed that the Svani

<table>
<thead>
<tr>
<th>TABLE 4. Gene diversity between populations based on autosomal short tandem repeat data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Alava*</td>
</tr>
<tr>
<td>Vizcaya*</td>
</tr>
<tr>
<td>Guipuzcoa*</td>
</tr>
<tr>
<td>Navarre*</td>
</tr>
<tr>
<td>Andalusia</td>
</tr>
<tr>
<td>Cantabria</td>
</tr>
<tr>
<td>Catalonia</td>
</tr>
<tr>
<td>Galicia</td>
</tr>
<tr>
<td>Murcia</td>
</tr>
<tr>
<td>Valencia</td>
</tr>
<tr>
<td>Austria</td>
</tr>
<tr>
<td>Belgium</td>
</tr>
<tr>
<td>Bosnia</td>
</tr>
<tr>
<td>Germany</td>
</tr>
<tr>
<td>Greece</td>
</tr>
<tr>
<td>Hungary</td>
</tr>
<tr>
<td>Tuscany</td>
</tr>
<tr>
<td>Poland</td>
</tr>
<tr>
<td>Portugal</td>
</tr>
<tr>
<td>Russia</td>
</tr>
<tr>
<td>Scotland</td>
</tr>
<tr>
<td>Slovenia</td>
</tr>
<tr>
<td>Switzerland</td>
</tr>
<tr>
<td>Egypt</td>
</tr>
<tr>
<td>Morocco</td>
</tr>
<tr>
<td>Turkey</td>
</tr>
<tr>
<td>Georgia</td>
</tr>
<tr>
<td>H^1</td>
</tr>
<tr>
<td>H^2</td>
</tr>
<tr>
<td>G_0.5</td>
</tr>
</tbody>
</table>

*Present study.
†Gene diversity within subpopulations.
‡Gene diversity between subpopulations.
§Coefficient of gene differentiation (92).
(a Kartvelian-speaking population) and the Basques were found to share only a single five-locus extended haplotype, A*01-B*8-DRB1*03-DQA1*0501-DQB1*0201 (95). This is the most frequent HLA haplotype found in Europeans (100,101), and is present in the Svani at a frequency of 1.25% and among Basques at 2%, leading the authors to conclude that the HLA system does not support the hypothesis of a relationship between these groups.

Recent studies of molecular markers also found little similarity between Basques and populations living in the Caucasus region. Analysis of Y-SNP haplogroups found that $F_{ST}$ values between Basques and Caucasus-dwelling groups were much greater than between Basques and surrounding Indo-European populations (96). While comparison of mtDNA sequences did reveal greater affinity between European groups and Caucasians than between West Asians and Caucasians (97), the addition of populations from Iran resulted in a genetic picture in which the Caucasus groups fell between populations from Europe and Asia Minor with respect to mtDNA sequence variation (98). As with Y-SNPs, genetic distances based on mtDNA sequences were greater between Basques and Caucasians than between Basques and Indo-Europeans, lending credence to the hypothesis of no genetic relationship between Basques and Caucasian populations. The results of the present study agree with those previously published using other genetic markers, as genetic distances based on autosomal data place the Basque groups in a different quadrant of the MDS plot than the population from Georgia.

The Vasco-Iberian hypothesis is based partly on craniometrics, the anthropometry of head shape (102). Broca suggested, based on a sample of 60 skulls from Guipuzcoa, that the Basques were similar to populations in North Africa (103,104). However, a reanalysis of Broca’s sample supplemented by the addition of 19 skulls noted no greater similarity between Basques and North African groups than any other European populations with regards to head shape (105). A more recent multivariate analysis of 20 craniometrical...
nic variables in 13 Iberian populations demonstrates the unique position of Basques in the Iberian Peninsula (106). Regardless of sex, Basques were distinct in every analysis performed. The differences between Basques and other Iberian populations could not be accounted for solely by geographic distance and were instead attributed to greater age of the Basque population relative to the others.

The majority of genetic studies supporting a relationship between Basques and North African populations have been based on HLA data (8,107-111). Other genetic systems do not support a relationship between Basques and North African groups (112-114), and additional HLA analyses also found no evidence of a relationship between the two populations (29,115-119). Preliminary investigation of autosomal STRs in Vizcaya Province indicated similarity with the Basque province of Guipuzcoa, and distinction from North African groups in the Maghreb (58). The present study demonstrates the lack of relationship between Basques and populations of North Africa, as the Basque populations do not cluster near either North African population included in the MDS plot (Morocco and Egypt), but rather are found near neighboring Cantabria. Our results instead lend support to the hypothesis that the Basques are a distinct European population, with no detectable prehistoric connection to (or recent gene flow from) populations in the Caucasus or North Africa.

Dataset available from the first author upon request. The authors thank the Basque participants, Dr Arantza Apraiz for collection of the samples, and Dr Rohina Rubicz for her assistance with DNA extraction.

Funding This work is supported in part by National Geographic Society Grant (Project 6935-00) to the University of Kansas Laboratory of Biological Anthropology.

Ethical approval Received from the University of Kansas Human Subjects Committee (HSCL #11955).

Declaration of authorship KLY performed all of the statistical analyses and prepared and edited the manuscript. GS directed and performed all genotyping analyses included in this manuscript. RD directed genotyping and participated in drafting the manuscript. MHC directed and performed all genotyping analyses included in this manuscript.

Competing interests All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

References
18 Levine MH, von Hagen V, Quilici JC, Salmon D. Anthropology of a Basque village: A new hematopoyetic study. Cahiers
43 Gonzalez AM, Garcia Q, Larruga JM, Cabrera VM. The mitochondrial lineage U8a reveals a Paleolithic settlement in the Basque country.


Young et al: STR variation in Basques in Spain

www.cmj.hr


87 Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1:47-50. doi:10.1017/S1361953105050107


95 Sanchez-Velasco P, Leyva-Cobian F. The HLA class I and class II allele


www.cmj.hr