Alu insertion polymorphisms in the Balkans and the origins of the Aromuns

D. Comas1,∗, H. Schmid1,2, S. Braeuer1,2, C. Flaiz1,2, A. Busquets1, F. Calafell1, J. Bertranpetit1, H.-G. Scheil3, W. Huckenbeck4, L. Efremovska5 and H. Schmidt2

1Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona
2Department of Anthropology, University of Ulm, Germany
3Institute for Human Genetics and Anthropology, Heinrich-Heine-University Düsseldorf, Germany
4Institute for Legal Medicine, Heinrich-Heine University, Düsseldorf, Germany
5Institute of Physiology and Anthropology, Medical Faculty of Skopje, Republic of Macedonia

Summary

We have analysed 11 human-specific Alu insertion polymorphisms in the Balkans to elucidate the origins of the Aromuns, a linguistic isolate inhabiting scattered areas in the Balkan Peninsula. Four Aromun samples (two from the Republic of Macedonia, one from Albania, and one from Romania) and five neighbouring populations (Macedonians, Albanians, Romanians, Greeks, and Turks) were analysed by means of genetic distances, principal components and analyses of the molecular variance (AMOVA). Three hypotheses were tested: Aromuns are Romanophonic Greeks; the result of a Romanian southward migration; or local descendants of the Thracians. The analyses show that the Aromuns do not constitute a homogeneous group separated from the rest of the Balkan populations. Grouping by language or geography does not explain the genetic differences observed in the region, suggesting a lack of genetic structure in the area. Aromuns do not seem to be particularly related to Greeks, Romanians, or to other Romance speakers. The Aromuns might have their origin to the south of the Danube river, with extensive gene flow with the neighbouring populations. The present results suggest a common ancestry of all Balkan populations, including Aromuns, with a lack of correlation between genetic differentiation and language or ethnicity, stressing that no major migration barriers have existed in the making of the complex Balkan human puzzle.

Introduction

Aromuns, also known as Aromani (with variants like Romani, Romeni, Rumeni) or Vlachs (the name given by their Slav neighbours), constitute a linguistic isolate scattered through the central Balkans and, although no firm statistical data about their population exists, their number is estimated at around 600,000 to 1,000,000. Most Aromuns live in northern Greece, while others inhabit Albania, Macedonia, and Bulgaria; after 1920, some Aromuns migrated north to Romania (Schmidt et al. 2000). The Romance language spoken by the Aromuns is closely related to classical Latin and is included within the Eastern group of continental Romance languages, jointly with Romanian, the only other eastern European Romance language spoken (Ruhlen, 1991).

The origin of this population is still unclear, although the most accepted theory is that the Aromuns are a pre-Slav population of Thracian origin who crossed with the Roman settlers (Rohr, 1987). The Thracians inhabited most of the Balkan Peninsula; no written testimony remains of the Thracian language. They founded the Macedonian Kingdom, which had its imperial heyday under Alexander the Great (336-323 BC). In 168 BC and 146 BC the Roman Empire conquered Macedonia and Greece respectively. Latin replaced Thracian, but not Greek, probably given the prestige that the latter enjoyed. In the course of this macedo-romanic
fusion a new population identity with a Neolatin dialect emerged, the Aromuns. In the 9th and 10th centuries the Slavs migrated to the Balkans and the Slavic languages replaced the Romance dialects in the Balkan Peninsula almost completely. Only the Romanians and Aromuns retained their Romance languages, and both groups of speakers constitute a linguistic isolate within the Balkans since they are surrounded by neighbours speaking languages belonging to different Indo-European branches (Slavic, Albanian and Greek). The extent to which this linguistic isolation implied a demographic isolation of the Aromuns can be detected using genetic methods.

On the basis of these facts several hypotheses about the ethnogenesis of the Aromuns were proposed (Schmidt et al. 2000), each of which would have yielded different genetic outputs:

1. The Aromuns are Latinised Greeks, either with a null demographic Roman contribution (that is, in an elite dominance model; Renfrew, 1994), or with Roman admixture. The genetic consequence of this situation would be that Aromuns are closely related to Greeks.

2. The Aromuns and the Dacoromanians (the present day Romanians) have a common origin and both are the descendants of admixture between Dacians, who lived north of the Danube, and Roman settlers. Later, Aromuns would have migrated south of the Danube. If this hypothesis is correct, the Aromuns should be genetically more similar to Romanians than to other Balkan populations.

3. The Aromuns are descendants of the Thracians, who lived south of the Danube, and afterwards may have mixed with Roman settlers. This hypothesis is more difficult to test since no extant population can be unequivocally identified with the descendants of the Thracians. Nevertheless, in this scenario Aromuns would be genetically close to other populations that settled south of the Danube and would be close to their neighbours speaking other languages.

There is scant genetic data about Aromuns: the Aromuns and other Balkan populations have been analysed for classical genetic markers (Schmidt et al. 2000; Scheil et al. 2001) and for four short tandem repeats (STRs; Huckenbeck et al. 2001). These studies have shown that the Aromuns do not constitute a homogeneous group of populations separated from their geographical neighbours.

Alu insertions abound in the human genome: they number over one million, or 11% of the total human genome sequence (Li et al. 2001). Some of these sequences have retroposed so recently within the human lineage that their insertion in a particular locus is polymorphic. Moreover, these human-specific polymorphisms are rapid and easy to type, apparently selectively neutral, have known ancestral states, and are identical by descent since they are the result of unique events (for a review of Alu insertions see Batzer & Deininger, 2002). Alu insertion polymorphisms in human evolution have been applied on a global scope (Batzer et al. 1994, 1996; Stoneking et al. 1997; Watkins et al. 2001), as well as a regional scale (Novick et al. 1998; Comas et al. 2000a; Nasidze et al. 2001).

In the present study, we have genotyped 11 human-specific Alu insertion polymorphisms in four different Aromun populations scattered over the Balkans, and in five neighbouring populations, speaking most widespread languages, in order to contrast several hypotheses regarding the origin of the Aromuns, and to assess the extent of genetic isolation that could be brought about by linguistic isolation.

**Material and Methods**

Eleven human specific Alu insertion polymorphisms (A25, B65, ACE, D1, APO, FXIIB, PV92, TPA25, HS2.43, HS3.23 and HS4.65) were typed in a total of 519 autochthonous individuals from the Balkans and Turkey (Figure 1) as described previously (Comas et al. 2000a). The sample comprised unrelated healthy blood donors; informed consent was obtained from all individuals participating in the study. The sample set comprised Albanians, Romanians, Macedonians, Northeastern Greeks (Thracia Province), Turks, and four different Aromun groups (Albanian Aromuns, Romanian Aromuns, Macedonian Aromuns from Stip, and Macedonian Aromuns from Krusevo). Additional European and Middle Eastern samples were used for comparison (Comas et al. 2000a; Romualdi et al. 2002).

Allele frequencies were calculated by direct counting; Hardy-Weinberg equilibrium was assessed by an exact
test (Guo & Thompson, 1992) provided by the Arlequin program (Schneider et al. 1996). Gene diversity for each population was calculated according to the formula 

\[ D = \frac{n}{n-1} \left(1 - \frac{\sum x^2}{n}\right), \]

where \( n \) is the number of gene copies in the sample and \( x \) the frequency of each allele.

\( F_{ST} \)-related genetic distances were computed between pairs of populations (Reynolds et al. 1983). A randomisation test was performed in order to test whether the distance between two populations differed (longer or shorter) from the distance between another pair of populations. The empirical null distribution of the genetic distances was obtained by bootstrapping 1,000 times over the frequencies of the eleven Alu loci typed, and repeating the distance calculations. From this distribution, a 95% confidence interval was obtained for each distance comparison.

Genetic distances were also represented in neighbour-joining trees (Saitou & Nei, 1987) by means of the PHYLIP 3.5c package (Felsenstein, 1989). The tree topology was assessed through 1,000 bootstrap iterations.

Principal component analysis (PCA) was performed on the correlation matrix of the Alu insertion frequencies analysed using the SPSS package.

In order to ascertain the proportion of the genetic variance due to differences within or between populations, genetic variance was hierarchically apportioned through the analysis of molecular variance (AMOVA) (Excoffier et al. 1992) performed with the Arlequin program (Schneider et al. 1996).

The correlation between genetic, geographic and linguistic matrices and their significance was assessed by Mantel test correlations with 1,000 permutations performed with the Arlequin program.

Results

Table 1 shows the frequencies by population of the Alu insertion in the 11 loci typed. All loci were polymorphic in all populations, except for: APO in Albanians, Albanian Aromuns and Romanian Aromuns, where the insertion is fixed; for HS2.43 in Macedonian Aromuns from Stip and Greeks; and for HS4.65 in Greeks where the insertion is absent. Eight out of 99 tests for Hardy–Weinberg equilibrium showed significant departures from equilibrium. After applying the Bonferroni correction only two comparisons gave significant departures (APO in Macedonians and HS3.23 in Greeks). As none of the departures cluster by locus or by population, they probably represent random statistical fluctuations.

Table 2 shows the average gene diversity by population. The populations analysed show no significant
Table 2  Average gene diversity in the Balkans

<table>
<thead>
<tr>
<th></th>
<th>Average gene diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albanians</td>
<td>0.263 ± 0.154</td>
</tr>
<tr>
<td>Romanians</td>
<td>0.311 ± 0.177</td>
</tr>
<tr>
<td>Macedonians</td>
<td>0.313 ± 0.179</td>
</tr>
<tr>
<td>Greeks</td>
<td>0.236 ± 0.141</td>
</tr>
<tr>
<td>Turks</td>
<td>0.271 ± 0.161</td>
</tr>
<tr>
<td>Aromuns (Albania)</td>
<td>0.292 ± 0.168</td>
</tr>
<tr>
<td>Aromuns (Romania)</td>
<td>0.314 ± 0.179</td>
</tr>
<tr>
<td>Aromuns (Macedonia-Stip)</td>
<td>0.301 ± 0.172</td>
</tr>
<tr>
<td>Aromuns (Macedonia-Krusevo)</td>
<td>0.298 ± 0.171</td>
</tr>
</tbody>
</table>

In order to assess the relationship between the populations analysed, and compare them to other European and Middle Eastern samples, FST genetic distances were calculated (data not shown) and depicted in a neighbour-joining (NJ) tree (Figure 2A). Although a tree representation has several drawbacks when dealing with populations, it may be useful to recognize clusters of populations with statistical support as given by bootstrap. The tree clusters all Balkan populations, including Aromuns, with other European samples, whereas Hungarians, Armenians and Syrians clearly separate from this cluster. The node separating both groups shows strong bootstrap support after 1,000 iterations (78.6%); aside from this node, no other nodes show bootstrap supports over 55%, indicating a very unstructured pattern. When the analysis is focused on the Balkans (Figure 2B), Aromuns and non-Aromuns cluster without a clear
structure, and the only bootstrap value over 50% is in the branch joining Albanian Aromuns and Romanian Aromuns (60.8%).

In order to determine whether genetic distances between Aromun samples were significantly different from those between Aromuns and their neighbouring populations (Macedonians, Romanians, and Albanians), an empirical null distribution of genetic distances was calculated as described in the Material and Methods section. After correction for multiple testing, no significant differences between genetic distances were found, showing that each Aromun population is not genetically more similar to its non-Aromun neighbours than to other Aromuns groups.

Since a NJ tree imposes a bifurcating model onto a distance matrix, which may be inadequate for closely related populations, we also assessed the genetic relationship among the populations through principal component analysis. The first two principal components (PC) for all the populations account for 48.7% of the genetic variance observed, and their plot (Figure 3A) separates Armenians and Syrians from the rest of the populations. When the PC analysis is restricted to the Balkans, the plot of the two first PCs (54.9% of the genetic variance) shows a very homogenous pattern, without a clear structure.

Analyses of the Molecular Variance (AMOVA) were performed to establish the apportionment of the genetic variance found in the present sample set (Table 3). When Aromuns were considered as a single group, a significant percentage (1.02%) of the variation was assigned to differences among Aromun samples, suggesting that Aromuns do not constitute a genetically homogeneous group of populations. On the other hand, their geographical neighbours (Macedonians, Romanians, and Albanians) constituted a homogeneous group without significant differences among populations. When Aromuns and their geographical neighbours were grouped according to their Aromun affiliation or their country, no significant differences were found between groups, whereas significant differences were found within groups, suggesting that these classifications have no genetic consistence.

![Figure 3](image_url)  
**Figure 3** Plot of the two principal components of the allele frequencies of the 11 Alu insertion loci typed. A) Middle Eastern and European populations, B) Populations from the Balkans.

**Table 3** Analyses of Molecular Variance (AMOVA) in the Balkans

<table>
<thead>
<tr>
<th>Groups</th>
<th>Among groups</th>
<th>Among populations</th>
<th>Within populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromuns</td>
<td>1.02***</td>
<td>98.98**</td>
<td></td>
</tr>
<tr>
<td>Aromuns’ neighbours</td>
<td>0.53 ns</td>
<td>99.47 ns</td>
<td></td>
</tr>
<tr>
<td>Aromuns vs neighbours</td>
<td>0.17 ns</td>
<td>0.83**</td>
<td>99.00**</td>
</tr>
<tr>
<td>Country</td>
<td>0.08 ns</td>
<td>0.87**</td>
<td>99.05**</td>
</tr>
<tr>
<td>All Balkan populations</td>
<td>0.84**</td>
<td>99.16**</td>
<td></td>
</tr>
<tr>
<td>Aromuns</td>
<td>0.09 ns</td>
<td>0.79**</td>
<td>99.12**</td>
</tr>
<tr>
<td>Country</td>
<td>-0.18 ns</td>
<td>0.99**</td>
<td>99.19**</td>
</tr>
<tr>
<td>Language</td>
<td>-0.96 ns</td>
<td>1.49**</td>
<td>99.47**</td>
</tr>
</tbody>
</table>

***(p < 0.01); ns: non-significant.**
When all studied Balkan populations were considered as a single group, the fraction of the genetic variance due to differences between populations was 0.84% (significantly different from 0, p < 0.0001). When the samples were divided into two groups, Aromuns versus non-Aromuns, the fraction of the genetic variance due to differences among groups was non-significant, whereas differences among populations within groups were significant (0.79%, p < 0.0001), indicating again that the Aromuns do not constitute a homogeneous group of samples separated from the rest of Balkan populations. If we group the samples according to current political country borders or linguistic group (Altaic, Romance, Slavic, Greek, Albanian), the genetic variance attributable to differences among populations was not significantly different from 0, suggesting that groups of populations defined either by language or by country are not genetically different from each other.

Genetic, linguistic, and geographic matrices were obtained in order to assess the correlation between these three variables. The partial correlation between the genetic and the geographic distance, corrected by the linguistic distance, was not significant (0.195, p = 0.071), neither was the correlation between the genetic and linguistic distances corrected by geography (−0.126, p = 0.748).

Discussion

This analysis of 11 Alu human-specific polymorphisms in the Balkans has shown that the allele frequencies found in these populations are very similar to other European populations. The Aromuns, an isolated linguistic group, are not an exception to the high homogeneity of European populations (Cavalli-Sforza et al. 1994). AMOVA has shown that the Aromuns do not represent a differentiated group compared to the rest of Balkan populations, and even when grouped by language no clear stratification exists. This result is in agreement with other studies that have failed to correlate genetic diversity with the linguistic affiliation at a regional level (for example, Comas et al. 2000b, using mtDNA), possibly because language replacement may be a much faster process than gene flow.

Although the Aromuns are a small linguistic and cultural isolate, their Alu insertion gene diversity is not reduced compared to the rest of European samples, as would be expected if they were also a genetic isolate. Although the gene diversity found may seem low (from 0.292 in Albanian Aromuns to 0.314 in Romanian Aromuns), it is similar to what is described for other Balkan populations, and it is worth noting that biallelic markers such as Alu insertion polymorphisms show diversity with a maximum of 0.5. However, a slight reduction may still be present and detectable with more sensitive systems, such as uniparentally transmitted mtDNA and the Y chromosome, which evolve with roughly one quarter of the effective population size of autosomal loci.

Three different scenarios have been postulated in order to explain the origin of the Aromuns. Each scenario would have led to a different genetic structure of the extant Aromun populations, which allows us to determine which is the most plausible scenario supported by the present Alu data.

Scenario 1: Aromuns are Latinised Greeks

The origin of the Aromuns could be traced to a group of Greeks who changed their language and adopted Latin after the arrival of Romans into the Balkans. This assumption would imply that Aromuns should be genetically more similar to Greeks than to the rest of neighbouring populations. In the present analysis, Northeastern Greeks have been shown to be genetically differentiated from the rest of the Balkan populations (see also Huckenbeck et al. 2001). This agrees with the results shown by classical genetic markers in other Greek samples (Cavalli-Sforza et al. 1994; Schmidt & Scheil, 2001). The Aromun samples differ from Greeks in the PC analysis and also presented considerable genetic distances from them. Moreover, the genetic distances found between Greeks and each Aromun sample are not significantly shorter than the distances found, for instance, between Aromuns and Romanians.

Within this scenario, a possible explanation for this difference found between Greeks and Aromuns would be that the origins of Aromuns could be traced back to Greeks who extensively interbred with Romans. This hypothesis would imply that the Aromuns present intermediate genetic features between Greeks and Southern and Western Romance populations. Since Alu allele frequencies are not available for Italians, Western
Romance populations (represented in the present analysis by French, Catalans and Andalusians) can be used as proxies. The genetic position of Aromuns was not intermediate between Greeks and the Western Romance speakers, which, with the caveat that the latter are not necessarily descendants of Romans, would be incompatible with Aromuns being the result of an admixture between Romans and Greeks.

Thus, the present data seems to refute this first hypothesis of the Aromuns being Latinised Greeks.

Scenario 2: Aromuns and Romanians are descendants of Dacians

This second scenario postulates that the Aromuns and Romanians have the same origin. Both may be descendants of ancient inhabitants living north of the Danube and mixed with Romans. This scenario would imply a close genetic relationship between Aromuns and Romanians. The present analysis shows that there is no close relation between the samples north of the Danube river (Romanians and Hungarians) and the Aromun samples, which agrees again with the results shown by classical genetic markers (Schmidt et al. 1998). Even Romanian Aromuns are very differentiated from Romanians, as shown by genetic distances and PC analysis. This could be explained by the fact that after 1920 many Aromuns coming from the south of the Balkans emigrated to Romania, where they settled in the Dobruja region and near Bucharest. This second hypothesis for the same origin of the Aromuns and Romanians, as descendants of the inhabitants north of the Danube and mixed with Romans, can also be rejected.

Scenario 3: Aromuns are Descendant of Thracians

The last hypothesis tested is the origin of the Aromuns as descendants of Thracians, who inhabited the region south of the Danube, and afterwards mixed with Romans. This scenario would be difficult to test since no extant population can be identified with the ancient Thracians. Nevertheless, this hypothesis would imply a close genetic relationship between Aromuns and the rest of the populations living in the Balkans south of the Danube. The present analysis has shown a close relationship between Balkan populations and Aromuns, although Aromuns of any given geographical region or political country do not present closer genetic distance to their neighbours, as seen in the AMOVA analysis. The genetic similarity of Aromuns to other Balkan populations could be compatible with the present scenario, but it is worth noting that Balkan populations do not differ distinctly from the other European populations. This is in agreement with the high homogeneity of European populations as shown by classical markers (Cavalli-Sforza et al. 1994) and mitochondrial DNA lineages (Simoni et al. 2000).

Nonetheless, a single origin for the Aromuns cannot be confirmed by the present data, given that the genetic heterogeneity found among Aromun samples is similar to that found among a set of linguistically unrelated Balkan populations. This pattern could be explained if Aromun populations had, in fact, different origins scattered in the Balkans, and had converged linguistically by adopting and preserving a Romance language. Or, alternatively, genetic drift could have erased the traces of a putative common origin. Alu insertion polymorphisms represent multiple independent autosomal loci, and the general trends observed in them escape the random specificities of each single locus, a feature that makes them useful in population genetics. However, they lack the well-defined phylogeographic structure of variation at the mtDNA or the Y chromosome, which may provide more precise details to the intriguing question of the origin of the Aromuns.

Acknowledgements

This research was supported by Dirección General de Investigación, Ministerio de Ciencia y Tecnología in Spain grant BOS2001-0794, and by Departament d’Universitats, Recerca i Societat de la Informació, Generalitat de Catalunya grant 2001SGR00285, and by the Deutsche Forschungsgemeinschaft grant Schm 580/4-1 and Schm 580/4-2. We are grateful to Mònica Vallès for technical assistance.

References


Received: 18 July 2003

Accepted: 1 October 2003