



ARTICLE

# Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations

Michele Belledi<sup>1</sup>, Estella S Poloni<sup>2</sup>, Rosa Casalotti<sup>1</sup>, Franco Conterio<sup>1</sup>, Ilia Mikerezi<sup>3</sup>, James Tagliavini<sup>1</sup> and Laurent Excoffier<sup>2</sup>

<sup>1</sup>*Dipartimento di Biologia Evolutiva e Funzionale, Università di Parma, Italy;* <sup>2</sup>*Département d'Anthropologie et Ecologie, Université de Genève, Switzerland;* <sup>3</sup>*Fakulteti I Shkenkave Natyrore, Universiteti I Tiranës, Albania*

Mitochondrial DNA HV1 sequences and Y chromosome haplotypes (DYS19 STR and YAP) were characterised in an Albanian sample and compared with those of several other Indo-European populations from the European continent. No significant difference was observed between Albanians and most other Europeans, despite the fact that Albanians are clearly different from all other Indo-Europeans linguistically. We observe a general lack of genetic structure among Indo-European populations for both maternal and paternal polymorphisms, as well as low levels of correlation between linguistics and genetics, even though slightly more significant for the Y chromosome than for mtDNA. Altogether, our results show that the linguistic structure of continental Indo-European populations is not reflected in the variability of the mitochondrial and Y chromosome markers. This discrepancy could be due to very recent differentiation of Indo-European populations in Europe and/or substantial amounts of gene flow among these populations. *European Journal of Human Genetics* (2000) 8, 480–486.

**Keywords:** human genetic diversity; mitochondrial DNA; Y-chromosome; linguistics; AMOVA; Albania

## Introduction

Mitochondrial DNA (mtDNA) and Y chromosome polymorphisms have been studied extensively in the context of human population genetics. They are very convenient because of the lack of recombination and their haploid mode of transmission.<sup>1</sup> Their simultaneous analysis in a set of populations also raises the interesting possibility of contrasting evolutionary processes experienced by males and females.<sup>2,3</sup>

As a contribution to the evaluation of the biological history of the Albanian population, we have studied the sequence variability of the first hypervariable segment of mtDNA control region (HV1) in 42 individuals and that of Y-specific haplotypes based on microsatellite DYS19 and the Alu insertion (YAP) in 56 individuals. The Albanian population had never been examined for these polymorphisms. Its study is of particular interest in the context of the settlement of the

European continent, due to the fact that the Albanian language is a separate lineage of the Indo-European linguistic family. It is indeed very distinct from the Italic, Greek, Celtic, Germanic and Balto-Slavic sub-families that represent the vast majority of the languages spoken in Europe. A recent study on blood groups distributions (ABO, MN and Rh) suggested that Albanians may be indeed quite different from other Balkan populations.<sup>4</sup>

Albanian diversity for these haploid molecular markers was compared with that of other published samples from continental Europe, in order to evaluate the level of differentiation among Indo-Europeans and to check the correlation between genetic and linguistic affinities for these populations.

## Materials and methods

### Samples

Specimens of hairs were taken, and preserved in alcohol, from Albanian individuals, born and residing in 24 Albanian districts and in the adjacent regions of Macedonia, Kosovo and Montenegro. Individual DNA was extracted from 2–3 dried hair roots using standard protocols.<sup>5</sup>

Correspondence: Michele Belledi, Dip. Biologia Evolutiva e Funzionale, Università di Parma, viale delle Scienze, 43100 Parma, Italy. Tel: +39 521 905150; Fax: +39 521 905151; E-mail: belledi@irisbioc.bio.unipr.it

Received 23 March 1999; revised 25 October 1999; accepted 10 November 1999

### mtDNA amplification and sequencing

HV1 sequences were PCR amplified according to published forward<sup>6</sup> and reverse<sup>7</sup> primers in a standard PCR reaction mix. Direct sequencing was performed with L15996 and H16401 primers.<sup>8</sup>

### Microsatellite (STR) and YAP analyses

DYS19 STR and YAP insertions were amplified according to published protocols.<sup>9,10</sup> DNA samples typed by sequencing were used as ladders to assign allele sizes.<sup>11</sup>

### Data analysis

Statistical analyses of the samples were carried out using the Arlequin software package.<sup>12</sup> Gene diversity indexes were computed for both mtDNA sequences and Y specific haplotypes (on the basis of the number of repeats at locus DYS19 and the presence/absence of the Alu insertion). The level of genetic structure within the European populations was assessed with an AMOVA analysis<sup>13</sup> by comparing haplotype frequencies ( $F_{ST}$  statistics) or by taking molecular differences into account ( $\Phi_{ST}$  statistics). For Y haplotypes, we used the information on allelic similarity or dissimilarity, instead of the number of repeat differences between DYS19 alleles, because the variance of the resulting statistics is large when only a few loci are examined.<sup>14</sup> A hierarchical structure of populations was tested by AMOVA, in which populations were grouped into the major sub-families of the Indo-European linguistic family (see Tables 2 and 3).  $F$  and  $\Phi$  statistic significance were assessed by a permutation procedure (100000 permutations).<sup>13</sup> The mismatch distribution of HV1 sequences was computed to check for the sign of a potential population demographic expansion.<sup>15</sup> The parameters of a stepwise demographic expansion

$$\theta_0 = 2N_0\mu, \theta_1 = 2N_1\mu, \text{ and } \tau = 2t\mu,$$

(where  $N_0$  and  $N_1$  are the population sizes before and after the instantaneous expansion, respectively,  $t$  is the number of generations since that expansion occurred, and  $\mu$  is the mutation rate per generation for the whole sequence) were estimated by the method of least-squares,<sup>15</sup> as implemented in the Arlequin software.<sup>12</sup> The inferred expansion model was tested by a parametric bootstrap method, based on the sum of squared differences (SSD) between the observed and expected mismatch distributions, as described by Schneider and Excoffier.<sup>16</sup> The selective neutrality of HV1 sequences and the demographic equilibrium of the Albanian sample were examined using Tajima's  $D$ <sup>17</sup> and Fu's  $F_s$ <sup>18</sup> statistics, the significance of which were assessed by simulations based on the coalescent algorithm described in Hudson.<sup>19</sup> Pairwise genetic distances between populations were computed as described in Slatkin.<sup>14</sup> The Indo-European linguistic classification by Ruhlen<sup>20</sup> was used to compute linguistic distances between populations, as described in Poloni *et al.*<sup>2</sup> The significance of the correlation between genetic and linguistic distances was evaluated by a Mantel test.<sup>21</sup>

## Results

### MtDNA

Thirty-one different HV1 sequences are found among 42 individuals (Table 1). As in the rest of Europe, the Cambridge sequence<sup>22</sup> is quite frequent in Albania (16.7%). Of the 31 Albanian sequences, 21 are similar to those previously described in other populations for overlapping nucleotides.<sup>23</sup> The other 10 HV1 sequences are unique to Albanians. With the only exception of sequence 22, which matches the L1a African haplogroup HVI motifs, the other Albanian sequences clearly display the nucleotide substitution pattern described in Europe.<sup>24</sup> According to the classification outlined by Macaulay *et al.*<sup>24</sup> for HVI, three sequences belong to haplogroup J, three to haplogroup T, six to haplogroup U5, one to haplogroup K and one to haplogroup V. The nucleotide diversity of Albanians ( $h = 0.0147$ ) falls within the range of variation of other Indo-Europeans from continental Europe (Table 2). Albanian mtDNA sequences show a clear unimodal mismatch distribution, typical of populations having gone through a recent expansion.<sup>25,15</sup> A similar pattern is observed in all other European populations with the exception of the Saami.<sup>26</sup> The least-squares estimates of the parameters of a stepwise expansion are as follows:  $\theta_0 = 0.60$  CI<sub>95%</sub>(0–2.00),  $\theta_1 = 43.48$  CI<sub>95%</sub>(13.99–7202.23), and  $\tau = 3.64$  CI<sub>95%</sub>(1.95–6.02). The  $P$ -value of the SSD statistic is 0.693, validating the hypothesis of a recent expansion corresponding to the above estimated demographic parameters. Assuming a divergence rate of 33% per million years,<sup>27</sup> a  $\tau$  value of 3.64 corresponds to about 36788 years (CI<sub>95%</sub>[19647–60808]). Assuming the mutation rate is correct, the molecular diversity of Albanians is in agreement with a late Pleistocene expansion as is the majority of European populations.<sup>26</sup> These results are strengthened by significant negative values of Tajima's  $D$  (–2.0308,  $P = 0.0004$ ) and Fu's  $F_s$  (–25.54;  $P = 0$  with 10000 simulations), all indicative of a recent demographic expansion.<sup>17,18,28</sup> The hypotheses of selective neutrality and population equilibrium are also rejected for most tested samples using Tajima's  $D$ , and for all samples using Fu's  $F_s$  (Table 2).

### Y chromosome

In the Albanian sample, 14.3% of Y chromosomes bear the Alu insertion (YAP<sup>+</sup> chromosomes). At the DYS19 locus, the alleles observed are: A (186 bp), B (190 bp), C (194 bp), D (198 bp) and E (202 bp), with frequencies of 19.6%, 37.5%, 33.9%, 7.1% and 1.8%, respectively. We analysed Y chromosome variability by combining the information relative to both DYS19 and YAP polymorphisms. A total of 7 DYS19/YAP haplotypes were observed, the more frequent being B/YAP<sup>–</sup> (35.7%) and C/YAP<sup>–</sup> (33.9%). The other observed haplotypes are: A/YAP<sup>+</sup> (12.5%), A/YAP<sup>–</sup> (7.1%), D/YAP<sup>–</sup> (7.1%), E/YAP<sup>–</sup> (1.8%), and B/YAP<sup>+</sup> (1.8%). In Albania, as in the rest of Europe,<sup>29</sup> the DYS19 allele most frequently associated with the Alu insertion is allele A. This tight association was tested by an exact test of linkage disequilibrium,<sup>30</sup> and found to be



**Table 2** Continental Indo-European sample genetic properties for mtDNA HV1. Sequences of other populations' samples are included in the data set by Handt *et al*<sup>3</sup>

|                     | <i>N</i> | <i>k</i> | <i>m</i> | <i>S</i> | <i>h</i>      | <i>D</i>            | <i>F<sub>S</sub></i> | Indo-European sub-family |
|---------------------|----------|----------|----------|----------|---------------|---------------------|----------------------|--------------------------|
| Albania             | 42       | 31       | 300      | 45       | 0.015 (0.008) | -2.031 <sup>a</sup> | -25.54 <sup>a</sup>  | Albanian                 |
| Bulgaria            | 30       | 22       | 360      | 37       | 0.013 (0.007) | -1.878 <sup>c</sup> | -14.39 <sup>a</sup>  | Balto-Slavic             |
| Catalonia           | 15       | 10       | 255      | 13       | 0.012 (0.008) | -0.878 ns           | -3.99 <sup>c</sup>   | Italic                   |
| Spain               | 41       | 21       | 302      | 38       | 0.019 (0.010) | -1.283 ns           | -6.51 <sup>c</sup>   | Italic                   |
| Portugal            | 54       | 37       | 302      | 39       | 0.012 (0.007) | -1.978 <sup>b</sup> | -26.11 <sup>a</sup>  | Italic                   |
| Trento              | 20       | 20       | 360      | 39       | 0.017 (0.009) | -1.771 <sup>c</sup> | -17.20 <sup>a</sup>  | Italic                   |
| Tuscany             | 52       | 40       | 360      | 55       | 0.014 (0.008) | -2.025 <sup>a</sup> | -25.53 <sup>a</sup>  | Italic                   |
| Sardinia            | 69       | 46       | 385      | 53       | 0.011 (0.006) | -2.035 <sup>a</sup> | -25.80 <sup>a</sup>  | Italic                   |
| Denmark             | 33       | 26       | 287      | 29       | 0.019 (0.010) | -0.897 ns           | -18.79 <sup>a</sup>  | Germanic                 |
| Iceland             | 39       | 29       | 360      | 32       | 0.014 (0.008) | -1.167 ns           | -22.19 <sup>a</sup>  | Germanic                 |
| England             | 100      | 71       | 360      | 67       | 0.012 (0.007) | -2.141 <sup>a</sup> | -25.71 <sup>a</sup>  | Germanic                 |
| Cornish             | 69       | 43       | 276      | 50       | 0.013 (0.007) | -2.170 <sup>a</sup> | -26.11 <sup>a</sup>  | Celtic                   |
| Welsh               | 92       | 45       | 277      | 47       | 0.012 (0.007) | -2.105 <sup>a</sup> | -26.40 <sup>a</sup>  | Celtic                   |
| Bavaria             | 49       | 34       | 276      | 36       | 0.014 (0.008) | -1.783 <sup>c</sup> | -25.96 <sup>a</sup>  | Germanic                 |
| North Germany       | 100      | 73       | 282      | 60       | 0.017 (0.009) | -1.863 <sup>b</sup> | -25.54 <sup>a</sup>  | Germanic                 |
| German-Switzerland  | 44       | 14       | 225      | 14       | 0.009 (0.006) | -1.492 <sup>c</sup> | -7.83 <sup>a</sup>   | Germanic                 |
| Latin-Switzerland   | 16       | 10       | 225      | 10       | 0.010 (0.006) | -1.646 <sup>c</sup> | -6.69 <sup>a</sup>   | Italic                   |
| Romansh-Switzerland | 16       | 10       | 224      | 13       | 0.011 (0.007) | -1.453 ns           | -4.72 <sup>a</sup>   | Italic                   |

<sup>a</sup> $P < 0.005$ ; <sup>b</sup> $P < 0.01$ ; <sup>c</sup> $P < 0.05$ . *N*: sample size; *k*: number of different sequences; *m*: sequence length; *S*: number of polymorphic sites; *h*: gene diversity; *D*: Tajima's *D*; *F<sub>S</sub>*: Fu's *F<sub>S</sub>*.

very significant ( $P = 0$ ; 1000000 steps in the Markov Chain). In particular, we found a very strong association between allele A and YAP<sup>+</sup> ( $P = 0$ ), and between allele C and YAP<sup>-</sup> ( $P = 0.02$ ). The observed gene diversity for Albanians ( $h = 0.744$ , Table 3) is similar to that of other European populations, although slightly lower values are observed for some Northern European samples.

As reported in Table 4, a low but significant level of genetic differentiation among populations is observed for mtDNA HV1 sequences, both when molecular information is used ( $\Phi_{ST} = 0.011$ ,  $P < 0.00001$ ) and when it is not ( $F_{ST} = 0.021$ ,  $P < 0.00001$ ). Populations grouped within the Indo-European linguistic sub-families are also significantly differentiated ( $F_{SC}$  and  $\Phi_{SC}$  indexes in Table 4). However, our results fail to reveal any significant level of genetic differentiation between the linguistic sub-families ( $F_{CT}$  and  $\Phi_{CT}$  indexes in Table 4). Note that the AMOVA analyses for mtDNA HV1

sequences lead to significant *F* statistics that are higher than the  $\Phi$  statistics. This suggests that a substantial amount of evolutionary 'noise' is introduced in the analysis of genetic structure when molecular information is used, possibly because of frequent homoplastic events occurring in the D-loop. We also performed an AMOVA using only the frequencies of some nucleotide positions (16069, 16129, 16224, 16270, 16278, 16292, 16294 and 16298), which define mtDNA haplogroups previously described,<sup>24</sup> and found a similarly low and significant level of genetic structure for Indo-Europeans of Europe ( $F_{ST} = 0.015$ ,  $P < 0.00001$ ;  $F_{CT} = 0.004$ , NS).

For Y chromosome haplotypes, a low but significant level of differentiation is observed among populations when the analysis is based only on haplotype frequencies ( $F_{ST} = 0.021$ ,  $P = 0.043$ ), but it becomes not significant when the number of mutations between haplotypes is used ( $\Phi_{ST} = 0.017$ ,

**Table 3** Continental Indo-European samples genetic properties for Y chromosome YAP/DYS19 haplotypes

|                | <i>N</i> | <i>k</i> | <i>h</i>      | Indo-European sub-family | Reference                            |
|----------------|----------|----------|---------------|--------------------------|--------------------------------------|
| Albania        | 56       | 7        | 0.744 (0.03)  | Albanian                 | Present study                        |
| Apulia         | 20       | 4        | 0.739 (0.055) | Italic                   | Ciminelli <i>et al</i> <sup>29</sup> |
| Calabria       | 26       | 6        | 0.702 (0.064) | Italic                   | Ciminelli <i>et al</i> <sup>29</sup> |
| Crete (Greeks) | 24       | 5        | 0.743 (0.052) | Greek                    | Ciminelli <i>et al</i> <sup>29</sup> |
| Sweden         | 40       | 4        | 0.541 (0.067) | Germanic                 | Sajantila <i>et al</i> <sup>42</sup> |
| Switzerland    | 51       | 5        | 0.657 (0.054) | Germanic                 | Sajantila <i>et al</i> <sup>42</sup> |
| England        | 19       | 4        | 0.708 (0.074) | Germanic                 | Hammer <i>et al</i> <sup>35</sup>    |
| Venetia        | 21       | 4        | 0.729 (0.058) | Italic                   | Hammer <i>et al</i> <sup>35</sup>    |
| North Sardinia | 28       | 5        | 0.794 (0.041) | Italic                   | Hammer <i>et al</i> <sup>35</sup>    |
| South Sardinia | 27       | 5        | 0.766 (0.039) | Italic                   | Hammer <i>et al</i> <sup>35</sup>    |
| Germany        | 30       | 5        | 0.667 (0.063) | Germanic                 | Hammer <i>et al</i> <sup>35</sup>    |

*N*: sample size; *k*: number of haplotypes; *h*: gene diversity

**Table 4** AMOVA analyses

|   | mtDNA              | Y chromosome       |
|---|--------------------|--------------------|
| Number of populations <sup>a</sup>            | 18                 | 11                 |
| Number of groups <sup>b</sup>                 | 5                  | 4                  |
| $F_{ST}$ (among populations)                  | 0.021 <sup>c</sup> | 0.021 <sup>d</sup> |
| $F_{SC}$ (among populations within groups)    | 0.024 <sup>c</sup> | 0.002 ns           |
| $F_{CT}$ (among groups)                       | -0.004 ns          | 0.019 ns           |
| $\Phi_{ST}$ (among populations)               | 0.011 <sup>c</sup> | 0.017 ns           |
| $\Phi_{SC}$ (among populations within groups) | 0.013 <sup>c</sup> | 0.003 ns           |
| $\Phi_{CT}$ (among groups)                    | -0.002 ns          | 0.014 ns           |

<sup>a</sup>Indo-European populations compared with Albanians for mtDNA HV1 sequences are listed in Table 2; Indo-European populations compared with Albanians for Y chromosomes are listed in Table 3; <sup>b</sup>Populations are grouped according to their linguistic classification into Indo-European sub-families; <sup>c</sup> $P < 0.0001$ ; <sup>d</sup> $P < 0.05$ ; ns: not significant.

$P = 0.061$ ). No significant level of genetic structure is detected either among populations within sub-families nor among sub-families (Table 4).

Population comparisons using both pairwise  $F_{ST}$  and  $\Phi_{ST}$  measures on Y chromosome diversity reveal that the Albanian sample is not significantly different from the other tested populations, with the exception of the Swedish sample. The same analyses performed on mtDNA haplotype frequencies shows that the Albanian sample is significantly different from the samples from Spain, Germany, Iceland, and German Switzerland. When molecular information is considered, Albania is only found significantly different from Denmark.

Comparisons of genetic, linguistic and geographic distance matrices are reported in Table 5. Values from 0 to 3 were assigned to the linguistic distances among pairs of languages within Indo-European sub-families, depending on their mutual relatedness. Linguistic distances among sub-families were varied from 4 (close relationship) to 16 (very distant relationship) to study the effect of different time depth of language evolution on the corresponding correlation coefficients. Our results show that the variability of mtDNA sequences among populations is not significantly correlated to the linguistic and geographic diversity. In contrast, linguistic information accounts for about 5% ( $r = 0.22$ ) of the

genetic variability between populations for the Y chromosome.

However, this contribution becomes not significant when the weight given to linguistic distances between Indo-European sub-families is increased (Table 5). Y chromosome diversity is also significantly correlated with geography. However, partial correlations of genetics with geography and linguistics are not significant, suggesting the impossibility of distinguishing independent geographic and linguistic factors which have contributed to the genetic differentiation of the populations (Table 5).

## Discussion

Despite belonging to a separate branch of the Indo-European language family, the Albanian population is found to be very similar to most other European populations for mtDNA HV1, as attested by the low genetic distances observed between populations and the 36 HV1 Albanian sequences out of 42 shared with other populations. This result is in keeping with the observation of generally low but significant levels of variability in Europe,<sup>31,32</sup> with the exception of Ladins<sup>33</sup> and the Saami.<sup>34</sup> A similar pattern is observed for the two Y chromosome polymorphisms. As already pointed out,<sup>35</sup> YAP<sup>+</sup> chromosomes are less frequent in Northern (0–7%) than Southern Europe (8–20%), and the Albanians, with 14.3% of YAP<sup>+</sup> chromosomes, are quite typical of Southern European populations. Also as among Europeans in general, the DYS19 alleles B and C are the most frequent among Albanians.<sup>29</sup> In fact, most of the genetic distances between Albanians and the other European samples, inferred from DYS19/YAP haplotypes, are not significant.

The linguistic peculiarity of the Albanian population is thus not reflected in our genetic data. Actually, the AMOVA analyses reveal a general absence of genetic structure for both maternal and paternal markers associated with the differentiation of Indo-European linguistic sub-families in Europe (Table 4). This general lack of structure suggests either a very recent radiation of the major Indo-European language sub-families from Europe, or the occurrence of large amounts of

**Table 5** Correlations between matrices of genetic, geographic and linguistic distances for mtDNA and Y chromosome. Partial correlations are computed only for Y chromosome variation

|   | linguistic distance<br>between sub-families | mtDNA  |            | linguistic distance<br>between sub-families | Y chromosome |             |
|---|---|--------|------------|---|--------------|-------------|
| Number of samples   |   | 18     |            |   | 11           |             |
| Correlation geography–linguistics                                   | 4   | 0.360  | $P < 0.01$ | 4   | 0.560        | $P < 0.001$ |
| Correlation genetics–geography                                      |   | -0.001 | ns         |   | 0.335        | $P < 0.05$  |
| Correlation genetics–linguistics                                    | 4   | 0.047  | ns         | 4   | 0.224        | $P < 0.05$  |
| Correlation genetics–linguistics                                    | 6   | 0.041  | ns         | 6   | 0.224        | ns          |
| Correlation genetics–linguistics                                    | 8   | 0.038  | ns         | 8   | 0.220        | ns          |
| Correlation genetics–linguistics                                    | 16  | 0.032  | ns         | 16  | 0.213        | ns          |
| Partial correlation genetics–geography (controlled for linguistics) |   |        |            | 4   | 0.259        | ns          |
| Partial correlation genetics–linguistics (controlled for geography) |   |        |            | 4   | 0.050        | ns          |

ns: not significant

gene flow among European populations. Note that for mtDNA, the inclusion in the analysis of an Indo-European sample from the Indian sub-continent (Havik, Indic sub-family of languages)<sup>36</sup> raises the levels of genetic structure observed, but these values are still not significant ( $F_{CT} = 0.024$ ,  $P = 0.305$ ;  $\Phi_{CT} = 0.018$ ,  $P = 0.126$ ).

Even though the AMOVA analyses on mtDNA HV1 and Y chromosome polymorphisms are not strictly comparable, since the data sets available for the two molecular markers are somewhat different, a slightly lower level of genetic structure is observed for Y chromosome DYS19/YAP polymorphisms than for mtDNA HV1 sequences. An ascertainment bias cannot be totally excluded for one or both markers, but unfortunately another overlapping data set for the Y chromosome does not yet exist for a sufficient size of population to control for that problem. Nevertheless, for mtDNA we observe a low but significant level of differentiation between populations within the Indo-European linguistic sub-families ( $F_{SC} = 0.024$ ,  $P < 0.00001$ ;  $F_{ST} = 0.021$ ,  $P < 0.00001$ ). This raises the possibility that a female-specific genetic structure of Indo-European populations actually exists, but that the pattern associated with this structure is defined by a factor other than the history of language differentiation. In contrast, no alternative genetic structure among Indo-Europeans is apparent from the study of male-specific markers ( $F_{SC} = 0.002$ ,  $P = 0.496$ ;  $F_{ST} = 0.021$ ,  $P = 0.043$ ).

The correlation study confirms the general lack of structure observed in the AMOVA analyses (Table 5). The linguistic fission history of Indo-European populations is not associated with mtDNA variation. When the Indian Havik sample is included in the analysis, the contribution of linguistics on mtDNA diversity increases by about 3%, but is still not significant. The very weak association observed between the Indo-European linguistic structure and the genetic distances among populations based on Y chromosome markers ( $r = 0.22$ , Table 5) suggests a possible correlation of the male-specific genetic radiation process with the differentiation of Indo-European language families. However, this correlation becomes non-significant when the weight attributed to linguistic differences between major Indo-European sub-families is increased. This could suggest that the differentiation of the Indo-European sub-families was indeed very recent, implying that the differentiation of genes proceeds at a slower pace than that of language. However, it is very likely that substantial amounts of gene flow have occurred among linguistic sub-families on the continent, which would have erased any former association between the linguistic and genetic radiation processes.

The low correlation between linguistic and genetic distances observed in this study for both maternal and paternal markers falls within the range of values observed for a large set of classical polymorphisms (correlation coefficients from  $-0.042$ , ns to  $0.455$ ,  $P = 0.004$ ).<sup>37</sup> This range of variation underlines the fact that, because of their particular history, different genes or regions of the genome will present distinct

patterns of variability among populations. Indeed, several studies have pointed to a correlation between linguistics and genetics among Indo-Europeans, but have also demonstrated the substantial effect of gene flow between populations in Europe in reducing the extent of their differentiation.<sup>38-41</sup>

In conclusion, on the basis of the polymorphisms here analysed, the Albanian population does not reveal any specific pattern that distinguishes it from the Indo-European gene pool of Europe. As for classic polymorphisms, the study of additional autosomal molecular markers could give us better clues to the genetic and linguistic history of Indo-European populations.

#### Acknowledgements

We are grateful to Giovanni Destro-Bisol, Eduardo Tarazona-Santos and Guido Barbujani for helpful comments on earlier versions of the paper. MB, RC, FC and JT were financially supported by 60% funding from the Italian MURST; ESP and LE were supported by Swiss National Science Foundation grants No. 32-047053-96 and 31-054059.98.

#### References

- 1 Jorde LB, Bamshad M, Rogers AR: Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays* 1998; **20**: 126-136.
- 2 Poloni ES, Passarino G, Stanachiaara Benerecetti AS, Semino O, Langaney A, Excoffier L: Human genetic affinities for Y chromosome p49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 1997; **61**: 1015-1035.
- 3 Seielstad MT, Minch E, Cavalli-Sforza LL: Genetic evidence for a higher female migration rate in humans. *Nat Genet* 1998; **20**: 278-280.
- 4 Susanne C, Bajrami Z, Kume K, Mikerezi I: Gene differentiation at the ABO, MN and Rhesus loci among Albanians and their relation with other Balkanic populations. *Gene Geography* 1996; **10**: 31-36.
- 5 Higuchi R, von Beroldingen CH, Sensabaugh GF, Erlich HA: DNA typing from single hairs. *Nature* 1988; **332**: 543-546.
- 6 Kocher TD, Thomas WK, Meyer A *et al*: Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc Natl Acad Sci USA* 1989; **86**: 6196-6200.
- 7 Tagliavini J, Battisti C, Conterio F: Polymorphic DdeI restriction sites in mitochondrial D-Loop from Emilian blood donors. *Gene Geography* 1993; **7**: 221-226.
- 8 Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC: African populations and the evolution of mitochondrial DNA. *Science* 1991; **253**: 1503-1507.
- 9 Roewer L, Arnemann J, Spurr NK, Greschik K-H, Epplen JT: Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum Genet* 1992; **89**: 389-394.
- 10 Jobling M, Tyler-Smith C: Fathers and sons: the Y chromosome and human evolution. *Trends Genet* 1995; **11**: 449-456.
- 11 Perez-Lezaun A, Calafell F, Seielstad M *et al*: Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol* 1997; **45**: 265-270.
- 12 Schneider S, Roessli D, Excoffier L: Arlequin vers. 2.0: A software for Population Genetic Data Analysis. Genetics and Biometry Laboratory, University of Geneva: Switzerland, 1999.
- 13 Excoffier L, Smouse PE, Quattro LM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: applications to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479-491.

- 14 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 1995; **139**: 457–462.
- 15 Rogers AR, Harpending HC: Population growth makes waves in the distribution of pairwise genetic distances. *Mol Biol Evol* 1992; **9**: 552–569.
- 16 Schneider S, Excoffier L: Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: Application to human mitochondrial DNA. *Genetics* 1999; **152**: 1079–1089.
- 17 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
- 18 Fu Y-X: Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 1997; **147**: 915–925.
- 19 Hudson RR: Gene genealogies and the coalescent process. In: Futuyma DJ and Antonovics JD (eds). *Oxford Surveys in Evolutionary Biology*. Oxford University Press: New York, 1990; pp 1–44.
- 20 Ruhlen M: *A Guide to the World's Languages*. Stanford University Press: Stanford, 1991.
- 21 Smouse PE, Long JC, Sokal RR: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 1986; **35**: 627–632.
- 22 Anderson S, Bankier AT, Barrel BG, de Bruijn MHL *et al*: Sequence and organization of the human mitochondrial genome. *Nature* 1981; **290**: 457–465.
- 23 Handt O, Meyer S, von Haeseler A: Compilation of human mtDNA control region sequences. *Nucleic Acids Res* 1998; **26**: 126–129.
- 24 Macaulay V, Richards M, Hickey E *et al*: The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* 1999; **64**: 232–249.
- 25 Slatkin M, Hudson RR: Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 1991; **129**: 555–562.
- 26 Excoffier L, Schneider S: Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proc Natl Acad Sci USA* 1999; **96**: 10597–10602.
- 27 Ward RH, Frazier BL, Dew-Jager K, Paabo S: Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci USA* 1991; **88**: 8720–8724.
- 28 Aris-Brosou S, Excoffier L: The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol Biol Evol* 1996; **13**: 494–504.
- 29 Ciminelli B, Pompei P, Malaspina P *et al*: Recurrent simple tandem repeat during human Y-chromosome radiation in Caucasian subpopulations. *J Mol Evol* 1995; **41**: 966–973.
- 30 Slatkin M: Linkage disequilibrium in growing and stable populations. *Genetics* 1994; **137**: 331–336.
- 31 Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D: Human mitochondrial DNA variation and the origin of Basques. *Ann Hum Genet* 1995; **59**: 63–81.
- 32 Francalacci P, Bertranpetit J, Calafell F, Underhill PA: Sequence diversity in the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phys Anthropol* 1996; **100**: 443–460.
- 33 Stenico M, Nigro L, Bertorelle G *et al*: High mitochondrial sequence diversity in linguistic isolates of the Alps. *Am J Hum Genet* 1996; **59**: 1363–1375.
- 34 Sajantila A, Lahermo P, Anttinen T *et al*: Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res* 1995; **5**: 42–52.
- 35 Hammer M, Spurdle AB, Karafet T *et al*: The geographic distribution of human Y chromosome variation. *Genetics* 1997; **145**: 787–805.
- 36 Mountain JL, Hebert JM, Bhattacharyya S *et al*: Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet* 1995; **56**: 979–992.
- 37 Sokal RR, Oden NL, Thomson BA: Origins of Indo-Europeans: genetic evidence. *Proc Natl Acad Sci USA* 1992; **89**: 7669–7673.
- 38 Sokal RR, Oden NL, Thomson BA: Origins of the Indo-Europeans: genetic evidence. *Proc Natl Acad Sci USA* 1992; **89**: 7669–7673.
- 39 Sokal RR, Oden NL, Walker J, Di Giovanni D, Thomson BA: Historical population movements in Europe influence genetic relationships in modern samples. *Hum Biol* 1996; **68**: 873–898.
- 40 Barbujani G, Pilastro A: Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proc Natl Acad Sci USA* 1993; **90**: 4670–4673.
- 41 Cavalli-Sforza LL, Menozzi P, Piazza A: *History and Geography of Human Genes*. Princeton University Press: Princeton, NJ, 1994.
- 42 Sajantila A, Salem A-H, Savpöainen P, Bauer K, Gierig C, Pääbo S: Paternal and maternal lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci USA* 1996; **93**: 12035–12039.