1 STR-based genetic structure of the Berber population of Bejaia (Northern Algeria) and
its relationships to various ethnic groups

5Nadir AMIR a, Mohamed SAHNOUNEb, LounesCHIKHIc de and Djebbar ATMANIa∗

9Laboratoire de Biochimie Appliquée, Faculté des Sciences de la Nature et de la Vie, Université de Bejaia, 06000 Algeria.
11Laboratoire d’Ecologie et Environnement, Faculté des Sciences de la Nature et de la Vie, Université de Bejaia, 06000 Algeria
13CNRS, Université Paul Sabatier, ENFA; UMR5174 EDB (Laboratoire Évolution & Diversité Biologique) ; 118, Route de Narbonne, F-31062 Toulouse, France.
15Université Toulouse 3 Paul Sabatier, CNRS ; UMR5174 EDB, F-31062 Toulouse, France.
16Instituto Gulbenkian de Ciência, Rua da Quinta Grande, nº6, 2780-56, Oeiras, Portugal.
18*Corresponding author: DjebbarAtmani
19Laboratoire de Biochimie Appliquée, Faculté des Sciences de la Nature et de la Vie, Université de Bejaia, 06000 Algeria.
21E-mail: djatmani@yahoo.com
22Tel/Fax : +21334214762
Abstract

Patterns of genetic variation in human populations have been described for decades. However, North Africa has received little attention and Algeria, in particular, is poorly studied. Here we genotyped a Berber-speaking population from Algeria using 15 short tandem repeat (STR) loci D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818 and FGA from the commercially available AmpF/STR Identifiler kit. Altogether 150 unrelated North Algerian individuals were sampled across 10 administrative regions or towns from the Bejaia Wilaya (administrative district). We found that all of the STR loci met Hardy–Weinberg equilibrium expectations, after Bonferroni correction and that the Berber-speaking population of Bejaia presented a high level of observed heterozygosity for the 15 STR system (>0.7). Genetic parameters of forensic interest such as combined power of discrimination (PD) and combined probability of exclusion (PE) showed values higher than 0.999, suggesting that this set of STRs can be used for forensic studies. Our results were also compared to those published for 42 other human populations analyzed with the same set. We found that the Bejaia sample clustered with several North African populations but that some geographically close populations, including the Berber-speaking Mozabite from Algeria were closer to Near-Eastern populations. While we were able to detect some genetic structure among samples, we found that it was not correlated to language (Berber-speaking versus Arab-speaking) or to geography (east versus west). In other words, no significant genetic differences were found between the Berber-speaking and the Arab-speaking populations of North Africa. The genetic closeness of European, North African and Near-Eastern populations suggest that North Africa should be integrated in models aiming at reconstructing the demographic history of Europe. Similarly, the genetic proximity with sub-Saharan Africa is a reminder of the links that connect all African regions.

Keywords: STR diversity, Forensics, Berber/Arab-speaking populations, North Africa, Continuity.
631. Introduction

Global patterns of genetic diversity are becoming increasingly important to reconstruct the demographic history of human populations. While some regions have received significant attention, others, like North Africa, have been generally less sampled and less studied. This is the case for Algeria despite its geographical position linking the Mediterranean area and Sub-Saharan Africa. Today the Algerian population is composed of two main linguistic groups, the Berber- and the Arab-speaking populations, and it is usually considered that the majority of the Algerians descend from Berbers and Arabs (Taïeb, 2004). However, the history of Algeria and North Africa is rather complex. For instance, the Berber-speaking region of Bejaia has witnessed many successive invasions and conquests that caused important cultural, linguistic and religious reshuffles among which the most important is probably the Arab conquest that started in the seventh century. Chronologically, the region was submitted to the influence of the Romans (33 BC), the Vandals (429 AC), the Byzantines (533 AC), the Arabs (647 AC), the Spanish (1510 AC), the Ottomans (1555 AC) and the French (1832 AC) (Cote, 1991; Laporte, 2004). In addition to these migrations, there have been internal reshuffles, with the introduction of Jewish and sub-Saharan African populations. At the fall of Andalusia (1610 AC), many of its expelled citizens came to establish settlements in Bejaia (see Gaid, 2008). Thus, while Berbers are likely to be the most ancient inhabitants of the region, gene flow, immigration and language switching may have obscured the relationships between neighboring or distant populations. Genetic data could therefore be useful to identify connections between populations speaking different languages today within Algeria or at a wider geographical scale. For instance, Henn et al. (2012), using genomic data, estimated that the North African populations are likely of Berber origin with substantial shared ancestry with the Near East and, to a lesser extent, eastern and western sub-Saharan Africa and Europe.
Though the number of studies on North Africa is relatively limited, there have been important studies using various markers that have contributed to the anthropogenetic characterization of North African Berber populations. These studies have focused on the GM immunoglobulin allotypic system (Dugoujon et al., 2004; Coudray et al., 2004; Coudray et al., 2006), others on mitochondrial DNA (Fadhlaoui-Zid et al., 2004; Ennafaa et al., 2009, Coudray et al., 2009), the Y chromosome (Arredi et al., 2004), autosomal microsatellites (STR) (Bosch et al., 2000; Bosch et al., 2001; Coudray et al., 2006; Coudray et al., 2007a; khodjet-el-khil et al., 2008, El Ossmani, 2010, Khodjet-El-Khil et al., 2012, Gaibar et al., 2012), SNP (Henn et al., 2012), and Alu Sequences (Gonzalez-Pérez et al., 2003). Very few studies have been carried out on Algerian Berber populations (Bosch et al., 2001; Achilli et al., 2005; Lefevre-Witier et al., 2006; Coudray et al., 2009; Pereira et al., 2010, Bekada et al., 2013).

The present study is part of a wider project on the anthropogenetic characterization of Algerian populations. In this paper we used 15 independent autosomal STR loci to genotype a sample of 150 individuals from the Berber-speaking population of the Bejaia wilaya to provide data on allele frequencies distribution and forensic parameters. The allele frequencies were exploited, using multidimensional scaling (MDS) and tree analysis (UPGMA), to assess the relationships between the Bejaia population and 42 other populations from North Africa, Sub-Saharan Africa, the Middle-East, Europe, Asia and South America. Analysis of molecular variance (AMOVA) was performed to assess the genetic structure of 17 populations (including Bejaia). A STRUCTURE analysis was also conducted.

Materials and methods

Population

Buccal swab samples were collected from unrelated healthy Berber-speaking donors (n=150 individuals, 300 gametes) from the Bejaia area in North Algeria (Fig.1), after written informed consent was obtained. Donors provided genealogical information for at least three
previous generations. Samples were collected in accordance with the ethical guidelines specified by the institutions involved in this study.

Figure 1

2.2. DNA extraction and amplification

Genomic DNA extraction was performed on the saliva samples with the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer’s recommendations. Fifteen independent autosomal tetranucleotide STR loci (see Butler, 2006), namely D13S317, D19D16S539, D2S1338, vWA, TPOX, D18S51, D5S818, FGA, D8S1179, D21S11, D7S820, D20D19S433, CSF1PO, TH01 and D3S1358, were coamplified in a multiplex PCR amplification reaction. Amplification was performed in a GeneAmp PCR system 9700 (Applied Biosystems, Foster City, CA) using the AmpF/STR Identifiler PCR Amplification Kit (Applied Biosystems, Foster City, CA) according to the manufacturer’s specifications.

2.3. Electrophoresis and genotyping

DNA fragments were separated by multi-capillary electrophoresis on an ABI Prism 3130xl Genetic Analyzer using the ABI GeneScan 500 LIZ internal size standard as a basis for comparison. Fragment sizes were obtained using the software GeneMapper® v3.2 (Applied Biosystems, Foster City, CA) and alleles were identified by comparison to an allelic ladder supplied by the manufacturer (Applied Biosystems, Foster City, CA).

2.4. RelPair analysis

To detect intra-population pairs of close relatives, we used the program RelPair Version 2.01 (Epstein et al., 2000). Each population was separately analyzed following the suggested settings of Pemberton et al. (2013), namely with a critical value set to 100 and a genotyping error rate of 0.008. When related individuals were identified, one of them was discarded from the analysis. In order to minimize the number of individuals removed, we preferentially omitted the individuals present in two or more related pairs while favoring those with less
missing data. We applied this analysis to all the populations for which we managed to obtain genotype data (See table 1 for populations’ codes). The number of individuals retained out of the initial number for each population is 116/150 (BJ), 40/44 (MB), 46/48 (SM), 94/105 (AN), 14190/104 (BH), 86/98 (SW), 87/99 (MA), 86/100 (CA), 73/80 (AM), 57/63 (BM), 57/59 (SH).

2.5. Statistical and phylogenetic analysis

Allele frequencies, expected (He) and observed (Ho) heterozygosity (Nei, 1987) and the exact test of Hardy-Weinberg equilibrium (Levene, 1949; Guo and Thompson, 1992) were computed using the Arlequin Software Version 3.5.1.2 (Excoffier and Lischer, 2010). The forensic parameters (matching probability (MP), power of discrimination (PD), polymorphism information content (PIC), probability of exclusion (PE) and typical paternity index (TPI)) were calculated using Powerstats Version 1.2 (http://www.promega.com/geneticidtools/powerstats/).

The expected number of genotypes was computed as \( Ng = \pi(k_j^2+k_j)/2 \) and the number of pairwise haplotype allele associations as \( Na = [(\sum k_j)^2 - \sum k_j^2]/2 \) (where \( k \) is the number of alleles at a considered locus and \( j \) the allele index). Bonferroni correction (Weir, 1996) was applied to adjust P values in Hardy-Weinberg assumptions (\( P = 0.05/15 = 0.0033 \) where 15 is the number of loci).

In order to determine the genetic relationship of our sample with other ethnic groups, we compared it to 42 populations from Europe, Asia, America and Africa using homologous microsatellite loci (Table 1). Pairwise uncorrected \( Fst \) distances between the 43 populations were used to perform a standard non-metric MDS using Statistica 8.0 (StatSoft, 2008) and infer a UPGMA tree using POPTREE2 (Takezaki et al., 2010) available at: http://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptree2/index.html. Tree robustness was evaluated using Bootstrap tests on 1000 permutations (Felsenstein, 1985). UPGMA rather than NJ method was used because it was more bootstrap-supported than the NJ one. Note that the trees were simply used as a graphical representation of the genetic distances computed.
They cannot be seen as a reliable representation of the relationships between populations due to the fact that such trees ignore the existence of gene flow, which is a crucial feature of human populations (Barbujani and Chikhi, 2007).

The MDS and Tree analyses were performed on all the 15 loci (including those with missing data) as well as after removing those with missing data (i.e. D16S539, D2S1338 and D19S433).

The significance of discriminance between groups in the MDS plot was determined using one-way ANOVA followed by unequal HSD (Honestly Significant Difference) test as implemented in Statistica 8.0 (StatSoft, 2008). The homogeneity of variances was checked using Levene’s and Cochran’s tests. When required, equality of variances was achieved by dividing data by the standard deviation values and comparing the standardized data.

Locus-by-locus allele frequency based AMOVA was performed using Arlequin v.3.5.1.2: Three plans of grouping were tested: (1) Grouping in relation to spoken language (Group 1 = Arab-speaking populations (RB, DM, MA, CA, AM, SH); Group 2 = Berber-speaking populations (BJ, MB, SM, AN, BH, AZ, KM, TN, LY, SW, BM)); (2) Grouping in accordance to geographical distribution (Group 1 = Western North African populations (SM, AN, BH, RB, AZ, KM, DM, AM, BM, SH); Group 2 = Central North African populations (BJ, MB, TN, LY); Group 3 = Eastern North African populations (SW, MA, CA); (3) Grouping in relation to UPGMA clustering (Group 1 = BJ, AZ, RB, LY, AN, BH, KM; Group 2 = MA, DM, SW, CA, Group 3 = MB, SM, SH, AM, BM; Group 4 = TN) (see Tab. 1 for population codes). An analysis of population structure was also carried out using the STRUCTURE software (Pritchard et al., 2000) but since our data were uninformative and did not lead to any clearly identifiable genetic clusters, the results are presented as supplementary material.
The correlations between genetic (uncorrected \( F_{st} \)) and geographical distances and between initial and final MDS distances were evaluated using Mantel test (Mantel, 1967; Smouse et al., 1992), and the fixation indices were tested using the permutation procedure (1000 iterations), as implemented in Arlequin 3.5.1.2.

### Table 1

<table>
<thead>
<tr>
<th>Locus</th>
<th>Observed Heterozygosity ( (H_o) )</th>
<th>Expected Heterozygosity ( (H_e) )</th>
<th>Hardy-Weinberg Equilibrium Test ( (P_h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2S1338</td>
<td>0.84</td>
<td>0.84</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>D5S818</td>
<td>0.93</td>
<td>0.93</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>D7S820</td>
<td>0.95</td>
<td>0.95</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>D13S317</td>
<td>0.96</td>
<td>0.96</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.97</td>
<td>0.97</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.98</td>
<td>0.98</td>
<td>0.77±0.05</td>
</tr>
<tr>
<td>FGA</td>
<td>0.99</td>
<td>0.99</td>
<td>0.77±0.05</td>
</tr>
</tbody>
</table>

The power of discrimination \( (PD) \), the probability of excluding paternity \( (PE) \) and the polymorphic information content \( (PIC) \) are displayed in Table 2. \( PD \) ranged from 0.911 to 0.975, \( PE \) from 0.388 to 0.624, and \( PIC \) from 0.740 to 0.900. All the 15 STR loci were highly polymorphic \( (PIC>0.7) \). The combined power of discrimination and combined probability of exclusion showed values higher than 0.999. With a \( PIC>0.8 \), seven of the fifteen loci \( (D21S11, D7S820, D3S1358, D2S1338, VWA, D18S51 and FGA) \) can be considered as very...
informative for genetic personal identification (with a combined PE=0.987). For the 15 loci (6 to 15 alleles per locus) the computed number of possible genotypes was $7.11 \times 10^{24}$.

Table 2

The standard non-metric multidimensional scaling (MDS) based on $F_{st}$ distances (15 loci) split the 17 North African populations including Bejaia (Table 1) into two main groups significantly discriminated (Fig. 2 A), one (North Africa 1, of 11 populations including Bejaia) close to the European populations, and the other (North Africa 2, of 5 populations), close to the Arabian Peninsula populations. As to Tunisia (TN), it behaved as outgroup to all other groups (Fig. 2 A). The population of Bejaia (BJ) went with the North Africa 1 populations. Its closest neighbors are Lybia (LY), Rabat (RB) and Azrou (AZ) of Morocco, all geographically close to Algeria (Fig. 2 A and B). The North African populations appeared to be the most heterogeneous in comparison with other regions groups included in this analysis (Fig. 2 A and B). Another MDS (Not shown) based on 12 loci (after removing those containing missing data, namely D16S539, D2S1338 and D19S433) gave roughly same results as above except that only MB (Mozabites) and CA (Copt Adaima) went with the Arabian peninsula populations.

Fig. 2

The UPGMA tree inferred using the $F_{st}$ distances between the 43 populations (Fig. 3) was congruent with the MDS results (Fig. 2) and exhibited higher bootstrap values than the neighbor-joining (NJ) tree (not shown,). However, most of the deepest nodes showed very low bootstrap values connecting reasonably well supported clusters. This tree emphasized the heterogeneity revealed by the MDS for the North African populations but the low bootstrap values suggest that caution is required in interpretation. The group containing BJ, AZ, RB, LY, AN, BH and KM appears as a sister cluster of that regrouping European and the three Middle East populations (LB, IR and IQ) but many topologies could explain the data and this mostly suggests that connections exist between all these populations. Four other populations, MA,
DM, SW and CA, exhibit longer branch but no clear clustering. As to MB, SM, SH, AM and BM, they formed a sister cluster to that constituted by the Arabian Peninsula populations (DB, OM, SA and YE), but again the low bootstrap values suggest caution. TN still behaved as an out-group to all other North African populations, hence confirming its isolated position in the MDS plot.

Locus-by-locus AMOVA revealed no significant difference between the Berber- and Arab-speaking groups, with $F_{ct}$ values that varied from -0.00246 to 0.00104 and percentages of variation from 0.02 to 0.25%. Similarly, the three geographical groups did not exhibit strong signals of differentiation with $F_{ct}$ values that varied from -0.00097 to 0.00454 and percentages of variation from 0.01 to 0.45%. Significant differences between these geographical groups were revealed only for D8S1197 and TH01 and no significant correlation ($R = -0.089, P = 0.68$) between genetic distances ($F_{st}$) and geographical distances was detected by the Mantel test. Groupings defined according to the UPGMA tree (Fig. 3) showed similarly low $F_{ct}$ values and percentages of variation of -0.00062-0.00689 and 0.01-0.69% respectively. For this grouping, significant differences were observed for 8 out of 15 loci (Table 3): D8S1197, D7S820, CSF1PO, D3S1358, TH01, D16S539, TPOX and FGA. However, for all the three plans of grouping (spoken language, geographical location and cluster affiliation), AMOVA revealed highly significant differences between populations within each the groups and for all loci (Supplementary Table 2).

The $F_{st}$ distances between the 17 North African populations (supplementary table 3) suggest that the closest populations to Bejaia were AZ and AN (Berber-speaking populations from Morocco) with $F_{st} = 0.005$ and $F_{st} = 0.006$, respectively; whereas the most distant populations from Bejaia were MB (Berber-speaking from Central South Algeria) and TN.
(Berber-speaking from Tunisia) both with $F_{st} = 0.029$. The closest populations to each other were BJ (Bejaia), AZ (Azrou) and AN (Asni) with $F_{st} = 0.005-0.006$, and the most distant populations were CA (Copt Adaima) and TN (Tunisia) with $F_{st} = 0.052$. Altogether these values were relatively low.

4. Discussion

These results constitute the first data reported on genetic diversity of the Bejaia population. The 15 STR loci were highly polymorphic with a significant proportion (40%) of rare alleles (Tab. 2 and ST. 2). The power of discrimination (0.911-0.975), the probability of exclusion (0.388-0.624) and the polymorphic information content (0.74-0.90) (Tab. 2) were high with combined $PD$ and $PE$ values higher than 0.999. As expected, the most polymorphic loci ($PIC > 0.8$) were also the most discriminating. Altogether these results strongly support the use of this set of genetic markers for forensic personal identification and paternity testing in the Bejaia region.

Regarding population structure, the different analyses provided concordant results but exhibited also slightly different levels of discrimination. The STRUCTURE and AMOVA analyses identified little genetic structure across North Africa and between major linguistic or geographical groupings. The MDS and Tree analyses (Figs. 2 and 3) confirmed this but identified possible subgroups. For instance, some populations, such as MB (Mozabites), SM (Berbers from South Morocco), SH (Sahrawis), AM (Arabs from Morocco) and BM (Northern Morocco Berbers), appeared closer to the Arabian peninsula populations, while others, including Bejaia (BJ), Lybia (LY), Asni (AN), Bouhria (BH) (Figs. 2 and 3), appeared closer to European and other Middle East (Lebanon, Iran and Iraq) populations. However, the main result is that the different populations are slightly differentiated from each other but the differentiation exists both among geographically close or distant populations and among populations speaking the same or a different language.
The genetic heterogeneity of North African populations with more or less affinities with Middle East, Europe and Sub-Saharan Africa has been suggested by authors using mtDNA (Plaza et al., 2003; Coudray et al., 2009), Y-chromosome DNA (Arredi et al., 2004; Capelli et al., 2006), STR markers (Capelli et al., 2006; El Ossmani et al., 2010) and SNPs (Botigué et al., 2013; Henn et al., 2012). In some studies, a West-to-East gradient, ranging from West Sahara to the Middle East has been described, which we could not detect in the present study, probably due to the limited number of markers or populations used. As suggested by the results of Henn et al. (2012), if larger numbers of populations and markers were sampled from North Africa and the neighboring regions (Europe, Middle-East and Sub-Saharan), one can expect a continuous complex with multi-polar gradients between the various ancestries admitted to North African populations (Maghrebi = Berber, Sub-Saharan, Middle-Eastern and European), as suggested by Serre and Pääbo (2004).

While we did not find a clear geographical pattern, we can note that some geographical structure appear when the two dimensions of Figure 2 are considered separately. Indeed, the North African populations (including Tunisia) are distributed as a gradient between Sub-Saharan and European groups according to dimension 1 and as another gradient between Sub-Saharan and Arabian Peninsula groups. In relation to dimension 1, Sub-Saharan, North Africa (including Mozabites) and Arabian Peninsula populations are confounded. This suggests that a finer geographical sampling and a larger number of markers would be necessary to identify the regions through which gene flow connected all these regions.

At a smaller geographical scale and based on the MDS and tree analyses (Figs. 2-3 and Tab. 12), the closest neighbors of the Bejaia population (BJ) were Azrou (AZ), Rabat (RB), Lybia (LY), Asni (AN), Bouhria (BH) with genetic distances (Fst) lower than 0.01. The Mozabite (MB) and the Tunisian populations (TN) were more distant from the Bejaia population (both with Fst=0.029) despite their geographical closeness and the shared Berber language. The out
layer position of the Tunisian Chenini population (Figs. 2 and 3) can be attributed to genetic
drift due to the small size of the sampled population as proposed by Khodjet-El-Khil et al.
(2008) and Bentayebi et al. (2014).

These observations illustrate the absence of correlation of genetic distances with both
geographical distances and the spoken language demonstrated by the AMOVA and Mantel
tests (Tab. 2). The absence of correlation between genetic and geographical distances at this
geographical scale may be due to population relocations, isolations and genetic drift. Indeed,
most studies that have shown that genetic distances are correlated with geographical distances
(Ramachandran et al., 2005; Lao et al., 2008) were performed at large geographic scale. Studies
carried out at smaller scale (within regions such as North Africa) are likely more
influenced by population relocations and isolations (Ramachandran et al., 2005).

No significant genetic differences were found in this study between the Berber- and the Arab-
speaking populations (Tab. 3). This lack of differentiations between these two groups of
populations have also been found by several studies using classical markers (Bosh et al.,
1997), Alu insertion polymorphism (Coma et al., 2000), Y chromosome (Bosh et al., 2001),
mtDNA (Fadhlouai-Zid, 2004) and autosomal STRs (Bosh et al., 2000; Khodjet el khil et al.,
2008 and 2012). This suggests that either the presence of Arab-speaking groups in north
Africa was mostly a cultural process, with limited gene flow between Arabs and Berbers
(Bosh et al, 2000), or that these populations were genetically very similar when they met.

Our results show that language boundaries are not correlated with genetic distances for North
African populations, probably due to the fact that the Arabisation is recent in the region.
However, this is not necessarily a general rule since several authors found correlation between
language boundaries and genetic differentiation (Barbujani and Sokal, 1990; Chen et al.,
1995).
SNP-based STRUCTURE analysis (730,000 sites) of 7 North African populations in comparison with populations from Middle East (Qatar), Europe (Tuscan and Basque) and Sub-Saharan Africa (6 populations) has revealed a putative autochthonous North African ancestry (referred to as Maghrebi = Berber ancestry component) decreasing in frequency from Western Sahara eastward to Egypt (interrupted only by the isolated Tunisian Berber Chenini population) with a parallel and equal increase of the Middle Eastern and European ancestry components, with lesser and irregular Sub-Saharan influence (Henn et al., 2012). Our STR-based STRUCTURE analysis did not retrieve this east-to-west ancestry gradient in North African populations (SF. 1), and did not detect the genetic heterogeneity suggested by MDS, phylogenetic and AMOVA (Figs. 2-3 and Table 3) analyses. This may be due to the low number of markers used in our study (15 STR) and/or the sample size of the populations analyzed. As demonstrated by Pritchard et al., 2000, the accuracy of inferences improves with sample size, number of loci, and degree of divergence between populations. Our results are in agreement with previously reported observation (Bosch et al., 2000, Khodjet-El-Khil et al., 2008 and 2012).

Altogether our results show that the language spoken today may not reflect the history of the populations, with several Arab-speaking populations being Berbers who shifted their language after the Arab conquest. Another possibility is that genetic drift in some of them has led to significant differences in allele frequencies which blurred the historical relationships. Also, admixture and gene flow between Arab-speaking and Berber-speaking population may have contributed to the present-day situation where linguistic and genetic distances are less correlated than they perhaps were in the past.

In this study, we do not wish to make strong statements and draw conclusions on these issues. Our aim was to identify useful markers for forensic studies and quantify genetic diversity in the Bejaia area compared to other previously analyzed populations. In order to reconstruct the
history of the Bejaia population we would need a better geographical sampling of Algeria and North Africa. We would also need to apply more complex and advanced statistical methods that those used here. In particular, it would be interesting to better understand the relationships between north Africa and the Andalusians of Moroccan origin who came to settle around Bejaia after the fall of Andalusia in 1610 (see Gaid, 2008). Similarly it would be interesting to quantify the impact of the various invaders of the Bejaia region during history. Historical texts and the genetic closeness of the Bejaia population to its neighbours found here suggests that these contributions were probably limited but it would still be interesting to quantify them using genomic approaches and inferential methods such as Approximate Bayesian Computation (ABC, Beaumont, 2010). One could for instance test whether it is true that Berbers were little impacted by external gene flow as a consequence of their taking refuge in difficultly accessible mountains. More populations and more different markers must be used before drawing decisive conclusions.

While it is not new to state that spoken languages do not constitute a reliable criterion of ethnic origin, our results show that it is also true in North Africa between Berber-speaking and Arab-speaking populations. This suggests that genomic studies using Mozabites as representative of Berber-speaking populations should perhaps be regarded as very approximate. Interestingly the genetic heterogeneity of the North African Berber populations together with their relative closeness to the European and Middle Eastern populations revealed here suggest that these populations should probably be more integrated in models aiming at understanding the recent demographic history of Europe, including both historical and prehistoric events such as the Neolithic transition.

Acknowledgements

We thank all the donors who made this study possible and Isa Pais for technical help in the laboratory (Instituto Gulbenkian de Ciência). We thank Elena Bosch, Maria Esther Esteban, Jean-Michel Dugoujon, and Ana Fernandez-Santander for providing STR data for North
African populations, Naoko Takezaki for help with tree software. LC is partly funded by the Laboratoire d’Excellence (LABEX) entitled TULIP (ANR-10-LABX-41).

Conflict of interest: the authors declare they have no conflict of interest.

References


El Ossmani, H., Bouchrif, B., Aboukhalid, R., Bouabdillah, M., Gazzaz, B., Zaoui, D.,

501Ennafaa, H., Cabrera, V.M., Abu-Amero, K.K., González, A.M., Amor, M.B., Bouhaha, R.,

504Epstein, M.P., Duren, M.L. and Boehnke M., 2000. Improved Inference of Relationship for

506Evanno, G., Regnaut, S., Goudet, J., (2005). Detecting the number of clusters of individuals
507using the software structure: a simulation study. Molecular Ecology 14 (8), 2611-2620.

509perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564-
510567.


514Evolution 39,783–791.

516Usefulness of autosomal STR polymorphisms beyond forensic purposes: data on Arabic- and

519Mimouni, Alger.

520Gonzalez-Pérez, E., Via, M., Esteban, E., Lopez-Alomar, A., Mazieres, S., Harich, N.,
522west Africa - genetic boundaries or melting pot?. Coll. Antropol. 27, 491–500.

523Gorostiza, A, Gonzalez-Martin, A., Ramirez, C. L., Sanchez, C., Barrot, C., Ortega, M.,
524Huguet, E., Corbella, J., Gené, M., 2007. Allele frequencies of the 15 AmpF/Str Identifiler
525loci in the population of Metztitlan (Estado de Hidalgo), México. Forensic Sci. Int. 166, 230–
526232.

527Guo, S. and Thompson, E. 1992 Performing the exact test of Hardy-Weinberg proportion for

530AmpF/STR Identifiler loci in Macedonians and Macedonian Romani (Gypsy). Forensic Sci.
531Int. 173, 220–224.


