A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa

ARREDI, Barbara, et al.

Abstract

We have typed 275 men from five populations in Algeria, Tunisia, and Egypt with a set of 119 binary markers and 15 microsatellites from the Y chromosome, and we have analyzed the results together with published data from Moroccan populations. North African Y-chromosomal diversity is geographically structured and fits the pattern expected under an isolation-by-distance model. Autocorrelation analyses reveal an east-west cline of genetic variation that extends into the Middle East and is compatible with a hypothesis of demic expansion. This expansion must have involved relatively small numbers of Y chromosomes to account for the reduction in gene diversity towards the West that accompanied the frequency increase of Y haplogroup E3b2, but gene flow must have been maintained to explain the observed pattern of isolation-by-distance. Since the estimates of the times to the most recent common ancestor (TMRCAs) of the most common haplogroups are quite recent, we suggest that the North African pattern of Y-chromosomal variation is largely of Neolithic origin. Thus, we propose that the Neolithic transition in this part of the world [...]
A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa

Barbara Arredi,1,2,3 Estella S. Poloni,3 Silvia Paracchini,2,7 Tatiana Zerjal,2 Dahmani M. Fathallah,4 Mohamed Makrelouf,5 Vincenzo L. Pascali1, Andrea Novelletto,6 and Chris Tyler-Smith2,7

1Istituto di Medicina Legale, Università Cattolica del Sacro Cuore di Roma, Rome; 2Department of Biochemistry, University of Oxford, Oxford, United Kingdom; 3Department of Anthropology, University of Geneva, Geneva; 4Institut Pasteur de Tunis, Tunis, Tunisia; 5Laboratoire Central, Chu Bab El Oued, Alger, Algeria; 6Department of Cell Biology, University of Calabria, Rende, Italy; and 7The Wellcome Trust Sanger Institute, Hinxton, United Kingdom

We have typed 275 men from five populations in Algeria, Tunisia, and Egypt with a set of 119 binary markers and 15 microsatellites from the Y chromosome, and we have analyzed the results together with published data from Moroccan populations. North African Y-chromosomal diversity is geographically structured and fits the pattern expected under an isolation-by-distance model. Autocorrelation analyses reveal an east-west cline of genetic variation that extends into the Middle East and is compatible with a hypothesis of demic expansion. This expansion must have involved relatively small numbers of Y chromosomes to account for the reduction in gene diversity towards the West that accompanied the frequency increase of Y haplogroup E3b2, but gene flow must have been maintained to explain the observed pattern of isolation-by-distance. Since the estimates of the times to the most recent common ancestor (TMRCAs) of the most common haplogroups are quite recent, we suggest that the North African pattern of Y-chromosomal variation is largely of Neolithic origin. Thus, we propose that the Neolithic transition in this part of the world was accompanied by demic diffusion of Afro-Asiatic-speaking pastoralists from the Middle East.

North Africa showed (1) that about half of the lineages belonged to the E haplogroups otherwise observed mainly in sub-Saharan Africa and (2) that most of the rest fell into haplogroup U6 (Salas et al. 2002), which perhaps originated in the Near East and spread into North Africa ∼30 thousand years (KYA) ago (Ko et al. 2003). Y-chromosomal studies are potentially highly informative about the origin of male-specific lineages, because of the detailed haplotypes that can be obtained and their high geographical specificity (Jobling and Tyler-Smith 2003), but previous studies have been restricted to limited regions of North Africa (Bosch et al. 1999, 2001; Flores et al. 2001; Manni et al. 2002; Luis et al. 2004). Together, these genetic analyses highlighted the similarity between northeastern Africa and the Middle East and the clear genetic differentiation between northwestern Africa and both sub-Saharan Africa and Europe, including Iberia. The Sahara and Mediterranean, despite the narrow width of the Strait of Gibraltar, seem to have acted as effective long-term barriers to Y-chromosomal gene flow.

To provide a more complete description of the North
African pattern of Y-chromosomal variation, we have analyzed five additional populations: Algerian Arabs, Algerian Berbers, Tunisians, and North and South Egyptians (table 1). Binary polymorphisms (Underhill et al. 2000), including 122F (Casanova et al. 1985), were typed in the hierarchical fashion described elsewhere (Rosser et al. 2000; Paracchini et al. 2002), allowing the allelic states at 119 markers defining 117 haplogroups to be measured or inferred from the Y phylogeny (fig. 1A). In the North African sample, 30 binary markers were found to be polymorphic, identifying 23 different haplogroups (Rosser et al. 2000; Paracchini et al. 2002), allowing the phylogenetically related haplogroups to be classified into clusters, the frequencies of which are shown schematically in fig. 1B. With the existing data from Morocco (Bosch et al. 2001), the combined set now spans the northern part of the continent. In addition, samples from southern Europe, the Middle East, and sub-Saharan Africa were included in some analyses (Semino et al. 2000; Underhill et al. 2000; Cruciani et al. 2002). Our results reveal four main conclusions about the male-lineage variation in North Africa.

First, as shown in fig. 1B, the lineages that are most prevalent in North Africa are distinct from those in the regions to the immediate north and south: Europe and sub-Saharan Africa. This is illustrated by even a cursory examination of the commonest haplogroups: E3b2 is the most common haplogroup in North Africa, forming 42% of the combined sample. In contrast, R1b made up 55% of a mixed European sample (Underhill et al. 2000) and was even higher (77%) in the Iberian sample examined by Bosch et al. (2001), whereas E3a predominate in many sub-Saharan areas, being present at 64% in a pooled sample (Underhill et al. 2000; Cruciani et al. 2002). Such a finding is not surprising, in the light of the earlier genetic studies, but has an important implication: despite haplogroups shared at low frequency, suggesting limited gene flow, North African populations have a genetic history largely distinct from both Europe and sub-Saharan Africa over the timescales needed for the Y-chromosomal differentiation to develop.

Second, just two haplogroups predominate within North Africa, together making up almost two-thirds of the male lineages: E3b2 and J* (42% and 20%, respectively). E3b2 is rare outside North Africa (Cruciani et al. 2004; Semino et al. 2004 and references therein), and is otherwise known only from Mali, Niger, and Sudan to the immediate south, and the Near East and Southern Europe at very low frequencies. Haplogroup J reaches its highest frequencies in the Middle East (Semino et al. 2004 and references therein), whereas the J-276 lineage (equivalent to J* here) is most frequent in Palestinian Arabs and Bedouins. Lineages can rise to high frequency because of biological selection, social selection, and/or neutral drift. There is a suggestion that weak negative selection due to partial deletion of genes needed for spermatogenesis could act on both E3b2 and J (Repping et al. 2003), but this would tend to decrease their frequency, and there is no evidence for positive selection. It therefore seems likely that their increase was due to drift despite any negative selection, implying that male effective population size has been small. Indeed, gene diversity values increase along a latitudinal axis from west to east (fig. 2), and much of this variation is accounted for by haplogroup E3b2, which decreases in frequency in a corresponding fashion from ∼76% in the Saharawis in Morocco to ∼10% in Egypt (fig. 2). The same haplogroup has increased in frequency in many different populations within North Africa, so there must have been gene flow between them.

Third, there is strong geographical structure to the Y-chromosomal variation within the region. There is a high and significant correlation observed between genetic and geographical distances (τ = 0.55, P < 0.0005). Multidimensional scaling (MDS) analysis of genetic distances (Slatkin 1995) based on pairwise ΦST estimates (calculated using the program Arlequin) between 17 of the samples in fig. 1B showed a close correspondence with their relative geographical locations (fig. 3). Indeed, the positions of the samples in the MDS plot describe a latitudinal axis, from North Africa and the Middle East in the upper part to Central and southern Africa in the lower part. Furthermore, the pattern of genetic affinities among the North African samples parallels the west-east

---

Table 1

<table>
<thead>
<tr>
<th>Populations Studied</th>
<th>Code</th>
<th>Origina</th>
<th>Language</th>
<th>Sample Size</th>
<th>k(SNP)b</th>
<th>k(STR)c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algerian Arabs</td>
<td>Al-Ara</td>
<td>Alger</td>
<td>Arabic</td>
<td>35</td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td>Algerian Berbers</td>
<td>Al-Ber</td>
<td>Tizi Ouzu</td>
<td>Berber-Tamazigh</td>
<td>19</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>North Egyptians</td>
<td>N-Eg</td>
<td>Mansoura</td>
<td>Arabic</td>
<td>44</td>
<td>14</td>
<td>37</td>
</tr>
<tr>
<td>South Egyptians</td>
<td>S-Eg</td>
<td>Luxor</td>
<td>Arabic</td>
<td>29</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>Tunisians</td>
<td>Tun</td>
<td>Tunis</td>
<td>Arabic</td>
<td>148</td>
<td>15</td>
<td>107</td>
</tr>
</tbody>
</table>

a Main city in the sampling area.
b Number of haplogroups defined by the SNPs.
c Number of haplotypes defined by 15 STRs.
Figure 1  A, Phylogeny of Y-chromosomal haplogroups. The name of each haplogroup is shown at the tip of the lineage (1) according to Underhill et al. (2000) and (2) according to the Y Chromosome Consortium (2002). The polymorphisms screened in this study are shown along the branches. Lineage colors correspond to the haplogroup cluster colors of fig. 1B. Lineages shown by dashed lines were not observed in any of the populations discussed. Haplogroups observed in our samples are shown in boldface. The § symbol denotes 12f2a.  B, Frequency distribution of Y haplogroup clusters in African, Middle Eastern, and European samples. Populations include the samples from table 1, as well as Moroccan Arabs (M-Ara), North Central Moroccan Berbers (NM-Ber), Saharawis (Sah), South Moroccan Berbers (SM-Ber) (Bosch et al. 2001); Sudanese (Sud), Ethiopians (Eth), Europeans (Europe), Malians (Mali), Central Africans (C-Afr), South Africans (S-Afr), Khoisan (Khoi), Middle Easterners (Mid-East) (Underhill et al. 2000); Ethiopian Jews (Eth-J), Mossi (Mossi), Rimaibe (Rim), Fali (Fali), Ouldeme (Ould), Bamileke (Bamil), and Ewondo (Ewo) (Cruciani et al. 2002).
Population gene diversity and haplogroup E3b2 frequency in North Africa as a function of longitude. The linear regressions of gene (i.e., haplogroup) diversity (grey line) and E3b2 frequency (black line) onto longitude are shown. The population codes are reported in table 1 and figure 1 (in italics for gene diversity and in roman type for E3b2 frequency).

Figure 3
MDS analyses based on binary marker genetic distances among the North African, sub-Saharan African, European, and Middle Eastern populations from fig. 1B, tested using 116 binary markers (stress value 0.15). The population codes are reported in table 1 and figure 1.
paths of the regions, although for the Sahara, episodes of more favorable climatic conditions could have relaxed this barrier at times, particularly during some intervals between ~10 KYA and ~5 KYA (Muzzolini 1993). There is no evident reason why it should have acted as a strong genetic barrier at such times, so, if there was substantial gene flow, the genetic differentiation between North and sub-Saharan Africa may postdate this period. A clinal pattern of haplogroup variation like the one we observe can be expected from an east-to-west population expansion, and the finding of lower E3b2 STR variation in the west than in central North Africa (table A2 [online only]), accompanied by a substantial increase in frequency of this haplogroup, is most readily explained by expansion into virtually uninhabited terrain by populations experiencing increasing drift (Barbujani et al. 1994).

The current distributions of the haplogroups can suggest geographical origins, and their TMRCAs provide some constraints on the times of their spread. The M35 lineage (see the phylogeny in fig. 1A for marker loca-

- **Figure 4** Spatial autocorrelation analyses. *A*, Correlogram of the AIDA II indices calculated for North Africa. *B*, Correlogram of the AIDA II indices for North Africa and the Middle East (Syrian, Turkish, and Lebanese samples of Semino et al. [2000]). *C*, Correlogram of the Moran’s I index for the haplogroup E3b2* (overall correlogram significance $P = .001$). ***, $P < .001$; ns, not significant.
Table 2

<table>
<thead>
<tr>
<th>SAMPLED LINEAGES</th>
<th>MUTATION</th>
<th>NO. OF CHROMOSOMES</th>
<th>TMRCA [IN KY] (95% CI)</th>
<th>30 Years/Generation)</th>
<th>25 Years/Generation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td>390</td>
<td>13.95 (8.08–24.67)</td>
<td>25.87 (20.47–30.99)</td>
<td></td>
</tr>
<tr>
<td>E3a</td>
<td>M2</td>
<td>8</td>
<td>2.96 (1.83–4.25)</td>
<td>5.06 (3.01–7.61)</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>M33</td>
<td>3</td>
<td>3.38 (2.11–5.25)</td>
<td>5.6 (2.79–9.38)</td>
<td></td>
</tr>
<tr>
<td>E3b1</td>
<td>M78</td>
<td>38</td>
<td>4.48 (3.01–6.16)</td>
<td>8.10 (5.42–10.71)</td>
<td></td>
</tr>
<tr>
<td>E3b</td>
<td>M35</td>
<td>226</td>
<td>8.26 (5.18–12.37)</td>
<td>14.33 (9.32–19.19)</td>
<td></td>
</tr>
<tr>
<td>E3b2</td>
<td>M81</td>
<td>165</td>
<td>4.15 (2.84–5.97)</td>
<td>6.90 (5.91–8.19)</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>M89</td>
<td>152</td>
<td>9.77 (6.47–14.84)</td>
<td>19.05 (15.30–22.61)</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>12f2</td>
<td>93</td>
<td>6.78 (4.38–11.10)</td>
<td>7.86 (6.63–9.13)</td>
<td></td>
</tr>
<tr>
<td>J2</td>
<td>M172</td>
<td>14</td>
<td>3.67 (2.50–5.25)</td>
<td>4.98 (4.10–6.06)</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>M173</td>
<td>27</td>
<td>4.44 (3.04–6.62)</td>
<td>10.24 (8.54–12.94)</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>M170</td>
<td>2</td>
<td>2.83 (1.68–4.32)</td>
<td>5.29 (3.19–9.16)</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>M9</td>
<td>38</td>
<td>7.10 (4.38–11.11)</td>
<td>13.09 (11.56–18.16)</td>
<td></td>
</tr>
</tbody>
</table>

Note.—TMRCA estimates based on 8 STRs (DYS19, DYS388, DYS389b, DYS389I, DYS390, DYS391, DYS392, and DYS393) in nine North African populations (those from the present study and the Moroccan samples of Bosch et al. 2001). In these calculations, we incorporated the information from binary markers.

a The two parameters describing the population growth (alpha and beta) have been set as alpha prior uniform (0.03, 0.05) and beta prior uniform (0.10, 0.20), the microsatellite mutation rates used were from Weale et al. (2001) or gamma (2,1000) for the loci for which published estimates of mutation rates were not available.

b The two parameters describing the population growth (alpha and beta) have been set as in footnote a, the microsatellite mutation rate used was from Zhivotovsky et al. (2004).

In conclusion, we propose that the Y-chromosomal genetic structure observed in North Africa is mainly the result of an expansion of early food-producing societies. Moreover, following Arioti and Oxby (1997), we speculate that the economy of those societies relied initially more on herding than on agriculture, because pastoral economies probably supported lower numbers of individuals, thus favoring genetic drift, and showed more mobility than agriculturalists, thus allowing gene flow. Some authors believe that languages families are unlikely to be >10 KY old and that their diffusion was associated with the diffusion of agriculture (Diamond and Bellwood 2003). Since most of the languages spoken in North Africa and in nearby parts of Asia belong to the Afro-Asiatic family (Ruhlen 1991), this expansion could have involved people speaking a proto-Afro-Asiatic language. These people could have carried, among others, the E3b and J lineages, after which the M81 mutation arose within North Africa and expanded along with the Neolithic population into an environment containing few humans.

Acknowledgments

We thank the donors of the DNA samples, for making this work possible, and Guido Barbujani, Jaume Bertranpetit, Giovanni Destro-Bisol, Matthieu Honegger, Mark A. Jobling, Vincenzo Mandarino, Alicia Sanchez-Mazas, and Francesco Vecchi, for useful discussions and comments on the manuscript. B.A. was supported by a Ph.D. fellowship from the Universita Cattolica del Sacro Cuore di Roma, E.S.P. was supported by Swiss National Science Foundation grant 39-59375.99, V.L.P. was supported by grant PRIN-MIUR 2002 N.2002063871 from the Progetti Rilevanza Nazionale–Ministero dell’Istruzione, dell’Università e della Ricerca, A.N. was supported by grant PRIN-MIUR 2002, and C.T.-S. was supported by The Wellcome Trust.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Arlequin, http://lgb.unige.ch/arlequin/ (for population genetics software package)
Fluxus Engineering, http://www.fluxus-engineering.com/sharenet.htm (for NETWORK 3.1.1.1 phylogenetic network analysis software)
Ian Wilson’s download page, http://www.maths.abdn.ac.uk/~ijw/downloads/download.htm (for BATWING program)
References


