Internal Diversification of Mitochondrial Haplogroup R0a Reveals Post-Last Glacial Maximum Demographic Expansions in South Arabia

Viktor Černý, *,1 Connie J. Mulligan,2 Verónica Fernandes,3 Nuno M. Silva,3 Farida Alshamali,4 Amy Non,2 Nourdin Harich,5 Lotfi Cherni,6 Amel Ben Ammar El Gaaied,6 Ali Al-Meeri,7 and Luísa Pereira3,8

1Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, The Czech Republic
2Department of Anthropology, University of Florida, Gainesville
3Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal
4General Department of Forensic Sciences & Criminology, Dubai Police General Headquarters, Dubai, United Arab Emirates
5Laboratoire d’Anthropogénétique, Département de Biologie, Université Chouaïb Doukkali, El Jadida, Morocco
6Laboratory of Genetics Immunology and Human Pathology, Faculty of Sciences of Tunis, Tunis, Tunisia
7Department of Clinical Biochemistry, Faculty of Medicine and Health Sciences, University of Sana’a, Sana’a, Yemen
8Faculdade de Medicina da Universidade do Porto, Porto, Portugal

*Corresponding author: E-mail: cerny@arup.cas.cz.

Associate editor: Sarah Tishkoff

Abstract

Widespread interest in the first successful Out of Africa dispersal of modern humans ~60–80 thousand years ago via a southern migration route has overshadowed the study of later periods of South Arabian prehistory. In this work, we show that the post-Last Glacial Maximum period of the past 20,000 years, during which climatic conditions were becoming more hospitable, has been a significant time in the formation of the extant genetic composition and population structure of this region. This conclusion is supported by the internal diversification displayed in the highly resolved phylogenetic tree of 89 whole mitochondrial genomes (71 being newly presented here) for haplogroup R0a—the most frequent and widespread haplogroup in Arabia. Additionally, two geographically specific clades (R0a1a1a and R0a2f1) have been identified in non–Arabic speaking peoples such as the Soqotri and Mahri living in the southern part of the Arabian Peninsula where a past refugium was identified by independent archaeological studies. Estimates of time to the most recent common ancestor of these lineages match the earliest archaeological evidence for seafaring activity in the peninsula in the sixth millennium BC.

Key words: South Arabia, phylogeography, R0a haplogroup, migrations.

Introduction

During the course of the past 10 years, the study of the first migration of anatomically modern humans Out of Africa through the Bab el-Mandab Strait has dominated both archaeological (Petraglia and Alsharekh 2003) and genetic (Quintana-Murci et al. 1999; Kivisild et al. 2004; Macaulay et al. 2005) investigations of the Arabian Peninsula. Consequently, this issue has overshadowed other periods of prehistory that might have been even more significant in the formation of regional extant population.

In particular, recent archaeological and palaeoclimatological investigations have revealed the importance of the terminal Pleistocene and Holocene in the population history of Arabian Peninsula (Petraglia and Rose 2009) and have identified three possible core zones (the Red Sea, the South Arabian [Dhofar], and the Persian Gulf Basins) that could have served as refugia during environmental downturns (Rose and Petraglia 2009). It appears that the period from 35 to 20 thousand years ago (henceforth referred to as KYA), as lake deposits in northern and southern parts of the Peninsula testify, was a favorably wet one. Subsequently, however, the Last Glacial Maximum (LGM), a period lasting several thousand years from around 20 KYA, has been marked by eolian processes (erosion and deposition caused by wind) and dune formation in the Rub’ al Khali desert as well as other places of Arabia (Bailey 2009; Parker 2009). This change to a harsher and more arid climate was likely associated with the general movement or exodus of hunter–gatherer populations to more hospitable locations.

New evidence points to a brief additional wet phase between 15 and 13 KYA (Parker 2009). This could be related to the Bölling-Allerød (BA) interstadial (dated also between 15 and 13 KYA), which, in the northern latitudes, is associated with the mass wasting of its ice sheets. According to Parker (2009), “The BA brief interlude may have permitted brief human entry into parts of Arabia either from outside the peninsula or from refugia within Arabia (e.g., Dhofar or the now submerged Arabo-Persian Gulf Basin).” The
most suitable places for population concentration were regions with high altitudes, which still today bear the highest population densities (Wilkinson 2009). The onset of the Younger Dryas (13.5 KYA) ended the climatically favorable period until the Early Holocene when monsoon precipitation again increased (~9.0 KYA). This last pluvial phase lasted until ~5.0 KYA when present climatic conditions were established (Parker 2009).

Archaeological research has revealed that southwest Arabia, especially the highlands, must have been well populated over the past 5,000–6,000 years and may have acted as a long-term population centre during much of the Holocene (Wilkinson 2009); it remains to be determined if it also acted as a refuge for populations during the Palaeolithic. Furthermore, phylogenetic analyses of linguistic data demonstrate an expansion of extend Semitic languages throughout southern Arabia and East Africa some 1,000–3,000 years ago and their initial entry into the region may have occurred thousands of years earlier (Kitchen et al. 2009).

The contemporary distribution of ancestral and derived mitochondrial DNA (mtDNA) sequences shows that modern human colonization of Eurasia may have been accomplished via a southern coastal route (~60–80 KYA (Forster et al. 2001; Metspalu et al. 2004; Macaulay et al. 2005; Thangaraj et al. 2005). Subsequent mtDNA diversification took place in south-western Asia, where haplogroups M and N (and its derivative R) evolved. In fact, precisely where these lineages first emerged is still under discussion. According to Metspalu et al. (2004), this process took place in three distinct regions in southern Asia (the Middle East, the Indus Valley, and East Asia) with only minor migrations between these regions. One of the most important results of such past mtDNA diversification processes is the absence or very low frequency of M haplogroups in Western Eurasia today, demonstrating that the division of an original “Out of Africa” gene pool can be placed somewhere west of the Indus Valley, with a later westward migration to the Middle East and back to Africa including only subbranch M1 (Forster et al. 2001; Olivieri et al. 2006).

Given its recent elevation to pivotal location in terms of past human migrations, the Arabian Peninsula became the focus of several genetic studies, both for mtDNA and for Y-chromosome markers (Richards et al. 2003; Kivisild et al. 2004; Luis et al. 2004; Abu-Amero et al. 2007; Rowold et al. 2007; Abu-Amero et al. 2008; Cadenas et al. 2008; Černý et al. 2008, 2009). Amplification of HVS-1 of the C.M. data set was performed using primer pair P23 (Gonder et al. 2009). From the combined set of HVS-1 sequences, we classified R0a haplotypes as those displaying substitutions at positions 16126 and 16362 and clearly not ascribed to any other haplogroup, such as J (which has an additional substitution at position 16069) or T (which also has a substitution at position 16294).

Materials and Methods
Samples
Yemeni R0a haplotypes came from two different sampling collections undertaken by two authors (V.C. and C.M.). Those (n = 22) from the V.C. Yemeni data set (n = 250) were published recently and their locations are described therein (Černý et al. 2008, 2009). The Yemeni data set of C.M. (n = 550), from which we have used 28 R0a haplotypes, come from highland areas (Dhamar, Amran, Hajja, and Taizz), lowland areas near the Red Sea (Zabid), desert areas (Al Bayda, Abyan, Jawf, Maarib, and Shabwa), and the eastern part of Yemen (Hadramawt and Al Maha). We also performed complete mtDNA sequencing on one Ethiopian Jewish R0a haplotype from the collection of C.M. In addition, we surveyed our Central, East, and North African data sets for the presence of R0a haplotypes, resulting in 2 samples out of 448 Chad Basin samples (Černý et al. 2007), 4 samples out of 149 from Somalia (unpublished), 4 out of 77 from Ethiopia (unpublished), 4 out of 102 from Sudan (unpublished), 4 out of 304 from Tunisia (Černý et al. 2009), and 2 out of 81 from Morocco (Harich et al. 2010). In total, 71 haplotypes were selected for complete genome analysis (see supplementary material 1, Supplementary Material online).

Laboratory Work
Amplification of the hypervariable segment 1 (HVS-1) of samples from the V.C. data set was reported previously (Černý et al. 2008, 2009). Amplification of HVS-1 of the C.M. data set was performed using primer pair P23 (Gonder et al. 2007). From the combined set of HVS-1 sequences, we classified R0a haplotypes as those displaying substitutions at positions 16126 and 16362 and clearly not ascribed to any other haplogroup, such as J (which has an additional substitution at position 16069) or T (which also has a substitution at position 16294).
Whole-genome sequencing of all Yemeni, Chad, and Ethiopian Jew (total 53) samples was conducted using 24 primer pairs (Gonder et al. 2007). In contrast, 18 samples from Somalia, Ethiopia, Sudan, Tunisia, and Morocco were analyzed using 32 primer pairs (Maca-Meyer et al. 2001; Pereira et al. 2007). Polymerase chain reaction (PCR) products were purified and sequenced using the forward primers only. However, in cases of the poly-C stretch between nucleotides 568–573 and 16184–16193, the reverse primers were used as well. Sequencing was performed on an ABI 3100 DNA Analyzer (Applied Biosystems, Foster City, CA). Chromatograms were evaluated by two independent observers (V.C. and L.P.) with the help of SeqScape (Applied Biosystems) and BioEdit version 7.0.4.1 (Hall 1999). In cases of ambiguous results, new PCR amplification and sequencing reactions were performed. The complete mtDNA sequences are deposited in GenBank database with accession numbers: HM185203–HM185273.

Phylogenetic Analyses
A preliminary reduced median network analysis (Bandelt et al. 1995) led to a suggested branching order for the tree, which was then constructed by hand. All variable positions were used except 16182C, 16183C, and 16519, as they are too recurrent and inconsistently reported in the literature. We compared our phylogenetic reconstruction with the trees published in two papers (Abu-Amero et al. 2007; Alvarez-Iglesias et al. 2009) and found the only differences to be in the placement of the highly recurrent substitution at position 58 and the insertion 60.1T (for which we used the most parsimonious inference), and in the nomenclature of some branches due to updated information in our expanded phylogeny.

We used 71 new Yemeni, Moroccan, Tunisian, Sudanese, Ethiopian, Somali, and Chad whole R0a genome sequences and 18 previously published whole R0a genome sequences (Achilli et al. 2004; Palanchamy et al. 2004; Abu-Amero et al. 2007; Gasparre et al. 2007; Behar et al. 2008; Costa et al. 2009; Hartmann et al. 2009). An inconsistency was detected between the sequences deposited in GenBank and the tree published in Abu-Amero et al. (2007) for position 16355, which was reported as a transversion to A in GenBank and a transition in the published tree; we decided to consider it as the transition as it seems to be the most probable variant when compared with other genome and HVS-1 sequences. For calculation of the TMRCA for specific clades in the phylogeny, the ρ statistic (mean divergence from inferred ancestral haplotype) was used with mutation rate estimates for complete sequence of one substitution in every 3,624 years corrected for the effect of purifying selection and for the synonymous substitutions of one substitution in every 7,884 years (Soares et al. 2009) using the calculator provided in that paper. Standard errors were calculated as in Saillard et al. (2000).

Interpolation Maps and Evaluation of Spatial Autocorrelation
To determine and visualize the geographical distribution of R0a and R0a1a, interpolation maps were drawn using the "Spatial Analyst Extension" of ArcView version 3.2 (www.esri.com/software/arcview). The “inverse distance weighted” (IDW) option with a power of two was used for the interpolation of the surface. IDW assumes that each input point has a local influence that decreases with distance. The geographic location used is the centre of the distribution area from where individual samples of each population were collected. This allows one to obtain the Morans I index and Z score (a statistic that assesses the significance of the null hypothesis, H0: there is no spatial pattern among the values associated with the geographic points in the study area). Correlograms for Morans I indices versus distances were obtained for R0a and R0a1a haplogroups using PaSSAGE software v 1.0 (www.passagesoftware.net). An interval of ten distance classes was used for all data points studied here and compared with ten distance classes for fewer data points (grouping some of the closest Yemen samples). The existence of a cline is assumed when a continuous decline trend composed of statistically significant points is observed.

Results
Frequencies and Geographic Distribution of R0a Haplogroups
In our population data sets of 800 HVS-1 sequences from Yemen, 176 samples (22%) were classified as R0a comprising 46 different haplotypes. The ancestral type, with only two diagnostic variants present in HVS-1 (16126–16362), is the most common (n = 43). This haplotype is present in 16 of 22 different Yemeni populations (see supplementary material 2, Supplementary Material online). The second most common haplotype (n = 38) is a one step descendant of the ancestral type (16126-16355-16362), which has been named R0a1a (Abu-Amero et al. 2007; Černý et al. 2009). These two haplotypes constitute almost half of the R0a pool in Yemen.

Among the next most frequent R0a haplotypes, R0a1a1a (16126-16172-16355-16362), previously identified only in Soqotra (Černý et al. 2009), was also observed in Al Mahra, the region of eastern Yemen near Omani Dhofar close to Soqotra. A similar frequency, but much wider geographical distribution, was also noted for R0a1a1a1a1a (16126-16355-16362) (Cˇerny´ et al. 2009). These two haplotypes constitute almost half of the R0a pool in Yemen.

The geographic distribution of R0a frequencies was plotted across the Arabian Peninsula and neighboring regions. Figure 1a shows the highest frequency of R0a on Soqotra Island and the neighboring territory of Al Mahra, where it is observed in approximately one-third of the samples. Other locations in Yemen and Saudi Arabia show lower frequencies of R0a (except for some points near the Red Sea and in the central region of Saudi Arabia). Radiating from the main highest R0a frequency focus, frequencies decrease in all directions. The global value of Moran’s I index (Moran’s I
**Fig. 1** Interpolation maps for R0a (a) and R0a1a (b) haplogroups.
index = 0.06; Z score = 1.9 standard deviations) shows that there is a 5–10% likelihood that this clustered pattern is the result of random chance. The plot of R0a1a is shown in figure 1b, and the global value of Moran’s I index (Moran’s I index = 0.01; Z score = 0.87 standard deviations) suggests that this pattern is random. The distribution of Moran’s I index versus distance classes (fig. 2a and b) shows that the distribution of both haplogroups is clinal, although some values of the distances classes are not statistically significant.

R0a Phylogenetic Tree Based on Whole-Genome Information

The R0a tree now includes 89 individuals, 71 of which are complete new sequences. Whole-genome sequencing revealed several coding region mutations important for the R0a topology. Figure 3 shows a simplified version of the R0a tree but the tree in the supplementary material 3 (Supplementary Material online) lists all mutations and all samples.

As was shown before (Abu-Amero et al. 2007), there are two main R0a branches, designated R0a1 and R0a2. R0a1 is defined by one substitution at position 827 and the largest proportion of its haplotypes, as explained in the previous section, bear substitutions at positions 146-8292-11761-16355 and form the previously named haplotype R0a1a. Most of the branches in R0a1a derive from the root. All Soqotra samples and one Tunisian sample belonging to this haplogroup share a substitution at position 13708; eight out of nine of these samples, and one sample from Al Mahra, share an additional substitution at position 16172, which we previously identified as haplogroup R0a1a1a (Černy et al. 2009). Thus, it seems that R0a1a1 may be widespread in Yemen but rare elsewhere, whereas
Table 1. Age Estimates and Standard Deviations (in years) for the TMRCA of R0a Haplogroups.

<table>
<thead>
<tr>
<th>Haplogroups</th>
<th>All Positions</th>
<th>Synonymous Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0a</td>
<td>22,588 ± 851</td>
<td>15,665 ± 1,157</td>
</tr>
<tr>
<td>R0a1</td>
<td>21,573 ± 1,325</td>
<td>16,216 ± 1,852</td>
</tr>
<tr>
<td>R0a1a</td>
<td>11,489 ± 993</td>
<td>9,681 ± 1,543</td>
</tr>
<tr>
<td>R0a1a1a</td>
<td>3,160 ± 943</td>
<td></td>
</tr>
<tr>
<td>R0a2</td>
<td>15,608 ± 933</td>
<td>14,484 ± 1,466</td>
</tr>
<tr>
<td>R0a2f1</td>
<td>5,801 ± 1,294</td>
<td></td>
</tr>
<tr>
<td>R0a2f1 + 11365</td>
<td>3,030 ± 1,120</td>
<td></td>
</tr>
<tr>
<td>R0a3</td>
<td>12,455 ± 3,327</td>
<td>13,146 ± 5,786</td>
</tr>
</tbody>
</table>

* Insufficient information for TMRCA estimation.

R0a1a1a is dominant in Soqotra and only appears in geographically close areas.

Haplogroup R0a2 (bearing an insertion at position 60 and substitutions at 2355-15674) also shows several branches deriving directly from the root, but it contains a much larger number of subclades. Some of these subclades were previously named by Alvarez-Iglesias et al. (2009), but the improved resolution of our tree supports some revisions in nomenclature (see supplementary material 3, Supplementary Material online). An interesting branch is the widespread R0a2f. It was characterized by a substitution at position 8251, which was observed in one Italian and two nomadic Arabs from Chad. Further derived R0a2f haplotypes were observed in a cluster of eight Yemeni sharing substitutions at positions 131-7837-12542-13708-13827. This R0a2f1, without identifying mutations within the HVS-1 region, and for that reason not previously recognized (Černý et al. 2009), is also specific to Soqotra and adjacent regions of Yemen (four haplotypes were found on Soqotra, three in Al Mahra, and one in Hadramawt).

Furthermore, we detected a new branch diverging from the root of the tree, which we named R0a3. This new haplogroup shares the T insertion in position 60 and substitution at position 15674 with R0a2, but it does not have the substitution at 2355 and has an additional substitution at 15466.

TMRCA Estimates

The complete sequencing of this large set of R0a samples allowed us to apply both mutation rate estimates recently proposed by Soares et al. (2009), one rate for all positions and one for synonymous substitutions (table 1). The TMRCA obtained with only synonymous substitutions are ~0.69–1.06 of the dates obtained when analyzing all positions. The relationship between the two TMRCA estimates (mean of 0.854) is not so close as that reported by Soares et al. (2009) (0.9452), demonstrating the randomness inherent in the emergence of neutral mutations as proposed by those authors as the most plausible explanation for differences in TMRCA estimates. For simplicity, in the following description, we refer only to the estimates obtained using all positions.

Overall, TMRCA estimates based on all positions, placed R0a and R0a1 in the time frame of 20,000–25,000 YBP (years before present) (22,588 ± 851 for R0a and 21,573 ± 1,325 for R0a1), ages somewhat older than the ones obtained by Abu-Amero et al. (2007), even if those estimates had large confidence intervals, that is, 18,959 ± 8,478 YBP and 18,993 ± 6,999 YBP for R0a (coding region and HVS-1, respectively) and 9,248 ± 7,604 YBP for R0a1 (coding region). Considerable population expansion seems to have occurred ~12,000–16,000 YBP, with the emergence of the frequent and widespread haplogroups R0a1a (11,489 ± 993 YBP) and R0a2 (15,608 ± 933 YBP) as well as the less frequent and newly identified R0a3 (12,455 ± 3,327 YBP). Other R0a haplogroups seem to have emerged in the period between 5,000 and 11,000 YBP, with some being restricted to the Arabian Peninsula (such as R0a2c and an R0a1a haplogroup defined by 3438-5120-5333 both haplogroups have TMRCA estimates of 8,784 ± 2,744 YBP), whereas other haplogroups were observed in North and East Africa (R0a2a—6,549 ± 2,045 YBP; R0a2b—10,063 ± 2,298 YBP; and an R0a2 haplogroup defined by a substitution at position 9128—6,094 ± 2,268 YBP). The two branches restricted to Soqotra and neighboring regions in Yemen display TMRCA of 3,160 ± 943 YBP (R0a1a1a and 3,030 ± 1,120 YBP (R0a2f1 plus a substitution at position 11365), overlapping but narrowing the estimates previously reported when based only on HVS-1 variants (3,363 ± 2,378 YBP; Černý et al. 2009). Curiously, both R0a1a1a and R0a2f1 branches share the highly recurrent substitution at position 13708, which is a replacement substitution on the protein-coding gene NDS, converting an alanine (neutral apolar) to a threonine (neutral polar) (Pereira et al. 2009; Soares et al. 2009).

Discussion

In this work, we provide genetic evidence in support of human demographic expansions in South Arabia during the last 20 KYA when the climate was rapidly improving. We show that there is a statistically significant clustering pattern of mitochondrial haplogroups R0a throughout the Arabian Peninsula, reaching its highest frequency (~30%) in southeast Yemen and especially on Yemen’s Soqotra Island. The distribution of Moran’s I index versus distance classes suggests a clinal distribution of R0a and R0a1a (fig. 1), although there are some nonsignificant values likely reflecting the heterogeneity of Arabia’s natural landscape where inhospitable desert alternates with more habitable oases and mountains. Furthermore, heterogeneity of sampling may play a role; in fact, when assaying fewer samples inside Yemen, the cline was clearer with four out of five points showing statistical significance (data not shown). Unfortunately, no available data from Oman were available to us.

Current evidence favors the introduction of R0 to Arabia from the Middle East, where the oldest lineages for the R clade are observed (Richards et al. 2000). However, it is possible that Arabia was an additional centre for R0 emergence. Our results show that several founders of R0a are present in southern Arabia, with TMRCA estimates suggesting population continuity between the terminal Pleistocene and Holocene. Some of these mtDNA lineages...
also spread to North and East Africa. Two periods of demographic expansions closely match two wet climatic periods: the end of the 35–20 KYA wet period and the brief wet phase between 15 and 13 KYA. The post-LGM expansion around 16 KYA seems to have been especially important (including the new R0a3 haplogroup and the very frequent haplogroup R0a1a with its many branches deriving directly from the root), which is concordant with the hypothesis of human resettlement of Arabia (Parker 2009).

Recent archaeological findings agree with the genetic evidence suggesting population continuity, especially in the Yemeni highlands where two Early Holocene pre-Neolithic and Neolithic sites have been identified (Fedele 2009). However, there also exists the hypothesis that the peopling of eastern Arabia by Pre-Pottery Neolithic B-related settlers was the result of widespread population dispersal during the Early Holocene (Uerpmann et al. 2009). Fedele (2009) examined this issue in the Yemeni highlands, where archaeological surveys at two sites attested to an Early Holocene “Pre-Neolithic” settlement throughout the eastern Yemen Plateau (and a single continuum from Pre-Neolithic to Neolithic) with features of its Pre-Neolithic industry displaying hints of similarity with East Africa rather than the Fertile Crescent. Further study (McCorriston and Martin 2009) focused on the Early Holocene pastoralists along the desert margins of southern Arabia and found evidence for cattle introduction from the Levant or possibly from northeastern Africa by the sixth millennium BC, if not earlier. They suggest multiple waves of expansion into Arabia, at different times, and even suggest that the earliest herd animals in southern Arabia were probably introduced as a pioneering strategy among local hunters.

Our results show that a substantial part of the Yemeni population is biologically related to one or more demographic expansion events that have taken place over the last 20 KYA. The southeast parts of Yemen, such as Al Mahra Hadramawt and mainly Soqotra might have played an important role in later demographic expansion, as R0a1a1a and R0a2f1 now testify. The place where these haplogroups were identified closely match a South Arabian refugium suggested by paleoclimatic data (Rose and Petraglia 2009) and show the importance of the Holocene in this area. It would be very informative to have R0a sequences from Oman (especially Dhofar) in order to more precisely localize the origin of these genetically apparent demographic upheavals. It is very interesting that the genetic dates for the introduction and expansion of Soqotran R0a clades (between 6 and 3 KYA) match the earliest evidence for seafaring activity in the peninsula in the sixth millennium BC (Boivin et al. 2009). Our study is thus an example of how results in human genetics can closely overlap with other fields of anthropology.

Supplementary Material
Supplementary materials 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
This project was supported by the Fulbright-Masaryk Fellowship of V.C. at the Department of Anthropology, University of Florida, the Ministry of Education of the Czech Republic, Grant number: KONTAKT ME 917, the Council of American Overseas Research Centers, the American Institute for Yemeni Studies, and the Portuguese Fundação para a Ciência e a Tecnologia (PTDC/ANT/66275/2006) (L.P.), (Instituto de Patologia e Imunologia Molecular da Universidade do Porto is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by the Portuguese Foundation for Science and Technology, FCT). This research was also supported by a grant from the National Science Foundation to C.J.M. (BSR-0518530).

References


