doi: 10.1111/j.1469-1809.2007.00413.x

Close Genetic Relationship Between Semitic-speaking and Indo-European-speaking Groups in Iran

I. Nasidze[1,*], D. Quinque[1], M. Rahmani[2], S. A. Alemohamad[3] and M. Stoneking[1]

[1]*Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Deutscher Platz 6, D-04103, Leipzig, Germany*

[2]*Department of Molecular Genetics, Cardiovascular Research Center, Imam Hospital, Tehran University of Medical Sciences, Tehran, Iran*

[3]*Department of Human Genetics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran*

Summary

As part of a continuing investigation of the extent to which the genetic and linguistic relationships of populations are correlated, we analyzed mtDNA HV1 sequences, eleven Y chromosome bi-allelic markers, and 9 Y-STR loci in two neighboring groups from the southwest of Iran who speak languages belonging to different families: Indo-European-speaking Bakhtiari, and Semitic-speaking Arabs. Both mtDNA and the Y chromosome, showed a close relatedness of these groups with each other and with neighboring geographic groups, irrespective of the language spoken. Moreover, Semitic-speaking North African groups are more distant genetically from Semitic-speaking groups from the Near East and Iran. Thus, geographical proximity better explains genetic relatedness between populations than does linguistic relatedness in this part of the world.

Keywords: Iran, Bakhtiari, Arabs, Y chromosome, mtDNA

Introduction

Determining the extent to which the genetic and linguistic relationships of populations are correlated can provide insights into population history. However, any such correlation is confounded by geography, as populations that are geographically close to one another tend to be genetically related, and they tend to speak related languages. One informative approach toward disentangling the relative impact of geography vs. linguistic relationships is to analyze the genetic relationships of groups whose geographic neighbors are not their linguistic neighbors (Stoneking, 2005). Previously we have analyzed a number of such examples of neighboring groups, who speak different languages (Nasidze & Stoneking, 2001; Nasidze et al., 2004, 2005, 2006); in some cases we find that the geographical proximity of linguistically-different groups best explains their genetic relationships [e.g., Turkic-speaking Azerbaijanians in the Caucasus (Nasidze & Stoneking, 2001) or Turkic-speaking Gagauz in Moldavia (Nasidze et al., 2007), while in other cases the linguistic similarity of geographically-distant groups best explains the genetic relationships [e.g., Mongolian-speaking Kalmyks, who are surrounded by Slavic-speaking groups (Nasidze et al., 2005)]. Here we analyze the genetic relationships of two neighboring groups in southwest Iran who speak languages belonging to different linguistic families: Semitic-speaking Iranian Arabs, and Indo-European-speaking Bakhtiari.

The Bakhtiari tribal group lives in southwest Iran, in a mountainous region in the Khuzestan and Esfahan provinces (Figure 1). The Bakhtiari speak a dialect of the Luri language called Bakhtiari, which belongs to the Western Iranian branch of Indo-European (Ethnologue, 2000). Iranian Arabs mainly occupy the Khuzestan province in Iran (Figure 1). Historical evidence indicates that their ancestors, Arab tribes such as the *Bakr bin Wael* and *Bani Tamim,* entered Iran in the seventh century A.D., although there may have been an earlier Arabic presence in Iran (Morony, 2006). Iranian Arabs speak a Semitic language, which belongs to the Afro-Asiatic linguistic family (Ethnologue, 2000).

MtDNA HV1 (first hypervariable segment) sequence variability was previously studied in a Luri-speaking group, along with other groups from Iran, Pakistan, and Central Asia (Quintana-Murci et al., 2004). However, no specific inferences were made regarding this group; furthermore, it is not clear whether the samples analyzed came from a

* Corresponding author: Max Plank Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: nasidze@eva.mpg.de

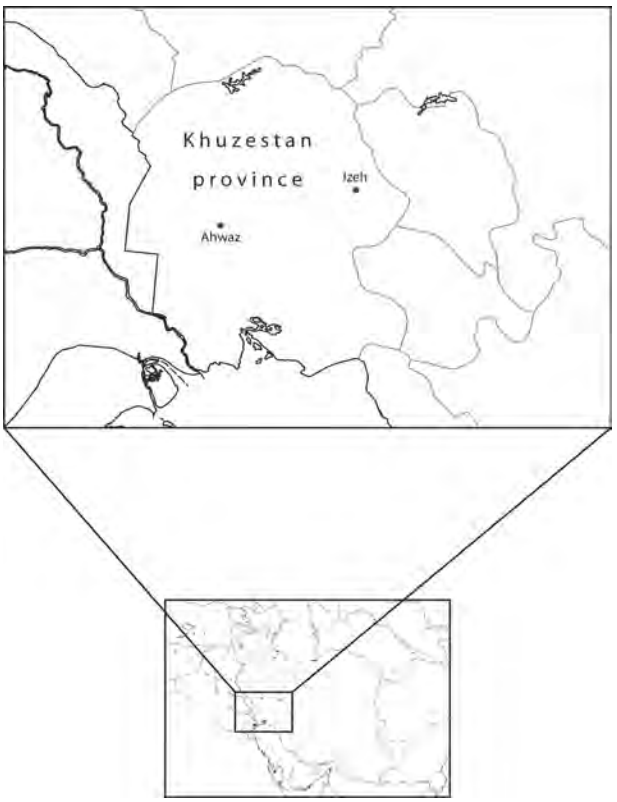

Figure 1 A map of the geographic location of the sampling sites for the Iranian Arabs and Bakhtiari.

Bakhtiari tribe, or from one of the other 10 Luri-speaking groups in Iran (Ethnologue, 2000). Y chromosome haplogroups have also been analyzed in several groups from the region (Quintana-Murci et al., 2001; Regueiro et al., 2006). Quintana-Murci et al. (2001) focused on the distribution of Y- haplogroups HG3 and HG9; HG9 was found in remarkably high frequency in the southeastern Caspian region and the Zagros Mountains (which includes the region where Luri-speaking groups live). However, the specific origin of the sample from the Zagros Mountains was not specified. The study by Regueiro et al. (2006) analyzed two groups of Persians, from the north and south of Iran. No further details concerning the origin(s) of these groups were provided.

Iranian Arabs have not been studied genetically to date. Although there are a number of studies describing mtDNA sequence variation, Y-SNP and Y-STR variation, and Alu insertion and HLA class I and II polymorphisms in different Arab groups from the Near East and North Africa (Almawi et al., 2004a; Almawi et al., 2004b; Luis et al., 2004; Semino et al., 2004; Chbel et al., 2003; Lucotte and Mercier, 2003; Richards et al., 2003; Al-Hussein et al., 2002; Cruciani et al., 2002; Manni et al., 2002; Bosch et al., 2001; Nebel et al., 2001; Semino et al., 2000; Scozzari et al., 1999), most of these focused on larger geographic scale questions and analyses, rather then concentrating on specific questions concerning the relationships of the diverse Arab groups. Moreover, for those studies that analyzed Y-haplogroups, the Y-SNP markers typed differ considerably across studies, making comparisons across populations difficult.

In order to investigate the genetic relationships of Bakhtiari and Iranian Arabs with their geographic and linguistic neighbors, we present here data on mtDNA and non-recombining Y chromosome (NRY) variation in 53 Bakhtiari and 46 Iranian Arabs from the Khuzestan province of Iran, to address the following question: which best explains the genetic relationships of these groups, geographical proximity or linguistic relatedness? That is, are the geographically-neighboring, but linguistically-dissimilar, Bakhtiari and Iranian Arabs genetically more similar to one another, or are they genetically more similar to their respective geographically-remote, linguistic neighbors?

Materials and Methods

Samples and DNA Extraction

A total of 99 whole blood samples from unrelated males, representing two populations from Iran – Bakhtiari (53 samples) from Izeh, and Arabs from Ahwaz, referred to here as Iranian Arabs (46 samples) – were collected (Figure 1). Genomic DNA from blood samples was extracted using the QIAamp® DNA Blood Mini Kit (Qiagen GmbH, Germany), following the instructions of the manufacturer. Informed consent and information about birthplace, parents and grandparents was obtained from all donors.

MtDNA Analysis

The first hypervariable segment (HV1) of the mtDNA control region was amplified using primers L15996 and H16410 (Vigilant et al., 1989), as described previously (Redd et al., 1995). The nested primers L16001 (Cordaux et al., 2003) and H16401 (Vigilant et al., 1989) were used to determine sequences for both strands of the PCR products with the DNA Sequencing Kit (Perkin-Elmer), following the protocol recommended by the supplier, and an ABI 3700 automated DNA sequencer. Individuals with the "C-stretch" between positions 16184–16193, which is caused by the 16189C substitution, were sequenced again in each direction, so that each base was determined twice.

Published mtDNA HV1 sequences were included from groups from Iran, other parts of West Asia, and North Africa (Brakez et al., 2001; Al-Zahery et al., 2003; Corte-Real et al., 1996; DiRienzo and Wilson, 1991; Kivisild et al., 1999; Krings et al., 1999; Nasidze et al., 2004; Nasidze et al., 2006; Quintana-Murci et al., 2004).

MtDNA haplogroups that are most informative in southwest Asia, i.e. haplogroups *H, J, N, T,* and *U* (Quintana-Murci et al., 2004), were determined by PCR-RFLP assays of the relevant

SNP, using previously-described methods (Finnila et al., 2000; Quintana-Murci et al., 2004).

Y Chromosome Bi-allelic Markers

All 99 samples were typed for the X- and Y-linked zinc finger protein genes in order to confirm the gender of the sample (Wilson & Erlandsson, 1998). Genotyping was carried out for ten Y chromosomal SNP markers: M9, M17, M45, M89, M124, RPS4Y (M130), M170, M173, M172, and M201 [(Underhill et al., 2000) and references therein]; the YAP *Alu* insertion polymorphism (Hammer & Horai, 1995) was also typed. The markers M9 and RPS4Y were typed by means of PCR-RFLP as described elsewhere (Kayser et al., 2000). The markers M17, M124, M172, M173, M45 and M201 were typed using PIRA-PCR (primer introduced restriction analysis) assays (Yoshimoto et al., 1993) as described previously (Nasidze et al., 2004; Cordaux et al., 2004). M89 was typed as described by Ke et al. (2001), while the YAP *Alu* insertion was typed as described by Hammer and Horai (1995). The M170 genotypes were detected by direct sequencing of the PCR product on an ABI3700 using BigDye technology (ABI Biosystems), following the protocol supplied by the manufacturer. The samples were genotyped according to the hierarchical order of the markers (Underhill et al., 2000). The Y-SNP haplogroup nomenclature used here is according to the recommendations of the Y chromosome consortium (Underhill et al., 2002).

Published Y-SNP data for Iranian, West Asian, and North African groups (Nasidze et al., 2004; Nasidze et al., 2006; Semino et al., 2000; Arredi et al., 2004) were also included in some analyses.

Y Chromosome STRs

Samples belonging to Y-SNP haplogroups G^*(M201) and $J2^*$(M172) were genotyped for nine Y chromosome short tandem repeat (Y-STR) markers: DYS19 (DYS394), DYS385a, DYS385b, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. These loci were amplified in pentaplex and quadraplex PCRs and detected on an ABI PRISM 377 DNA sequencer (Applied Biosystems) as described elsewhere (Kayser et al., 1997; Kayser et al., 2001). In order to distinguish genotypes at DYS385a and DYS385b, an additional PCR was carried out as described in Kittler et al. (2003) and detected on an ABI PRISM 377 DNA sequencer (Applied Biosystems).

Statistical Analysis

Basic parameters of molecular diversity and population genetic structure, including analyses of molecular variance (AMOVA), were calculated using the software package Arlequin 2.000 (Schneider et al., 2000). The statistical significance of F_{st} values was estimated by permutation analysis, using 10,000 permutations. The statistical significance of the correlation between genetic distance matrices, based on the mtDNA HV1 sequences and the Y chromosome SNP data, was evaluated by the Mantel test with 10,000 permutations. The STATISTICA package (StatSoft Inc.) was used for multi-dimensional scaling (MDS) analysis (Kruskal, 1964). Network analysis for Y-STR and mtDNA HV1 sequence data was carried out using the software package NETWORK version 3.1 (Bandelt et al., 1999).

Results

MtDNA HVI Sequence Variability

A total of 377 bp of the mtDNA HV1 region, comprising nucleotide positions 16024 to 16400 (Anderson et al., 1981), were determined for 53 Bakhtiari and 46 Arabs from southwestern Iran. As a check on the accuracy of the HV1 sequences, we used the network method to search for so-called "phantom" mutations (Bandelt et al., 2002). No such artifacts were found in the HVI sequences (analysis not shown). The sequences will be deposited in Genbank at the time of publication and will also be available from the corresponding author upon request.

MtDNA HV1 sequence data were previously published for 17 Luri from southwestern Iran (Quintana-Murci et al., 2004). However, it is not clear if this group belongs to a Bakhtiari tribe, or to some other Luri-speaking group. Pairwise F_{st} comparisons did not show significant differences between the published Luri and new Bakhtiari HV1 sequences ($F_{st} = 0.0151$, $p = 0.117$); however, the published Luri data are also not significantly different from other Iranian groups, such as Gilaki and Mazandarani from northern Iran ($F_{st} = 0.019$ and 0.0003; $p = 0.056$ and 0.439 respectively), Kurds from northwestern Iran ($F_{st} = 0.011$, $p = 0.173$), or Iranians from Isfahan ($F_{st} = 0.002$, $p = 0.378$). Therefore, we decided not to combine the published Luri data with our Bakhtiari data.

Parameters summarizing some characteristics of the mtDNA HV1 sequence variability in these groups are presented in Table 1. The haplotype diversity in both newly studied groups was similar to other Indo-European-speaking and Semitic-speaking groups from West Asia and North Africa, while the mean number of pairwise differences (MPD) were slightly higher (Table 1). Tajima's D was negative and significantly different from zero in both groups, suggesting population expansion (Table 1).

Pairwise F_{st} comparisons (Table 2) showed that the Iranian Arab and Bakhtiari groups are not significantly different from each other ($F_{st} = 0.016$; $p = 0.019$). Iranian Arabs showed closer affinities with West Asian Indo-European-speaking groups than with either Bakhtiari or with Semitic-speaking groups from West Asia or North Africa; however, none of these comparisons are significantly different from

Table 1 MtDNA HV1 sequence variability among Arab and Bakhtiari populations and neighboring groups

Population	N	no. of haplotypes	Haplotype diversity and SE	MPD	Tajima's D	Source
Iranian Arabs	46	40	0.991+/−0.008	7.56	−1.86*	present study
Bakhtiari	53	44	0.991+/−0.006	7.14	−2.20**	present study
Other Indo-European speaking groups from Iran						
Iranians_Tehran	79	63	0.984+/−0.008	5.53	−2.05**	Nasidze et al., (2004)
Iranians_Isfahan	46	42	0.996+/−0.006	6.17	−2.13**	Nasidze et al., (2004)
Lur	17	15	0.978+/−0.031	5.52	−1.85*	Quintana-Murci et al. (2004)
Gilaki	87	73	0.995+/−0.003	6.40	−2.19**	Nasidze et al., (2006); Quintana-Murci et al., (2004)
Mazandarani	71	58	0.992+/−0.005	5.98	−2.02**	Nasidze et al., (2006); Quintana-Murci et al., (2004)
Kurds	20	19	0.995+/−0.018	6.13	−1.57*	Quintana-Murci et al. (2004)
Other Semitic speaking groups from West Asia						
Iraqi	113	92	0.993+/−0.003	5.46	−2.10**	Al-Zahery et al. (2003)
West Asians	42	41	0.999+/−0.006	6.98	−1.78*	Di Rienzo & Wilson (1991)
Yemen, Arabs	115	67	0.981+/−0.006	7.49	−1.62*	Kivisild et al. (2004)
Other Semitic speaking groups from North Africa						
Morocco, Arabs	154	97	0.983+/−0.005	5.67	−1.98*	Brakez et al. (2001)
Egypt, Arabs	102	86	0.995+/−0.003	7.98	−1.67*	Krings et al. (1999)
Algeria, Arabs	85	30	0.943+/−0.010	4.82	−1.10	Corte-Real et al. (1996)

* $P < 0.05$, ** $P < 0.01$

	Bakhtiari	Arabs	Indo-Eur.[1]	Semitic[2]	Semitic[3]
Bakhtiari		0.016	0.011	0.033	0.069
Iranian Arabs	0.001		0.005	0.014	0.026
Indo-European speakers[1]	0.048	0.036		0.020	0.051
W.A. Semitic speakers[2]	0.006	0.012	0.040		0.042
N.A. Semitic speakers[3]	0.160	0.160	0.191	0.126	

[1]Indo-European speaking groups from West Asia; [2]Semitic speakers from West Asia; [3]Semitic speakers from North Africa.

Table 2 Mean pairwise F_{st} values between Bakhtiari and Arabs, and Indo-European-speaking and Semitic- speaking groups. Below diagonal – paiwise F_{st} values based on Y-SNP haplogroups; above diagonal – pairwise F_{st} values based on mtDNA HVI sequences

one another. The Bakhtiari also showed the closest affinities with West Asian Indo-European-speaking groups.

An MDS plot (Figure 2a) based on the pairwise F_{st} values illustrates these patterns. Iranian Indo-European-speaking groups and West Asian Semitic-speaking groups are clustered together (with the exception of Arabs from Yemen), while North African Semitic-speaking groups are spread more widely, and separately from this cluster, in the MDS plot. The Bakhtiari are situated on one edge of the West Asian cluster, while the Iranian Arabs are at the opposite edge of this cluster (Figure 2a).

MtDNA haplogroups similarly indicate a closer relationship of Iranian Arabs with other Iranian groups than with other Semitic-speaking groups. In particular, the Iranian Arabs lack haplogroup L, which is found at frequencies of 9–14% in other Semitic-speaking groups from West Asia, and in higher frequencies (35–92%) in Semitic-speaking groups from North Africa (Figure 3). Network analyses of HV1 sequences, on the background of specific mtDNA haplogroups, also indicate considerable sharing of HV1 sequences between Iranian Arabs and Bakhtiari and other Iranian groups (Figure 4).

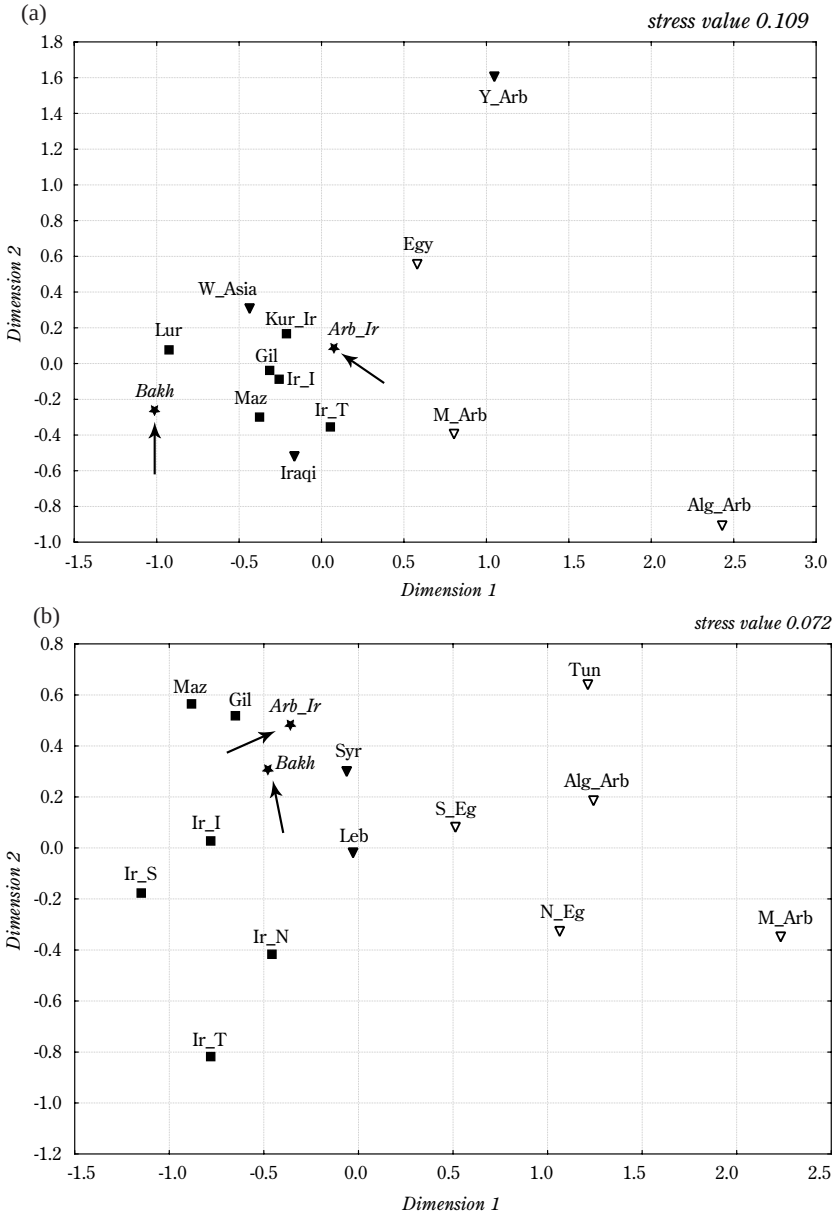

Figure 2 MDS plots based on pairwise F_{st} values, showing relationships among the Bakhtiari and Iranian Arab samples, Indo-European speaker groups from Iran, Semitic speaker groups from West Asia and North Africa. The Bakhtiari and Iranian Arab groups are represented by stars; Indo-European-speaking groups by squares; Semitic-speaking groups from North Africa by open triangles; and Semitic-speaking groups from West Asia by solid triangles. **A.** Based on mtDNA HVI sequence data. The stress value for the MDS plot is 0.109. **B.** Based on Y chromosome SNP data. The stress value for the MDS plot is 0.072. The names of the groups are abbreviated as follows: Arb_Ir –Arabs from Iran, Bakh – Bakhtiari, Lur – Luri, Maz – Mazandarani, Gil – Gilaki, Ir_I – Iranians from Isfahan, Ir˙T – Iranians from Tehran, Ir_N – Iranians from northern Iran, Ir_S – Iranians from Southern Iran, Kur_Ir – Kurds from Iran, Egy – Arabs from Egypt, Y_Arb – Arabs from Yemen, Alg_Arb – Arabs from Algeria, M_Arb – Arabs from Morocco, W_Asia – Arabs from West Asia, Iraqi – Arabs from Iraq, Syr – Syrians, Leb – Lebanese, N_Eg – Arabs from North Egypt, S_Eg – Arabs from South Egypt, Tun – Arabs from Tunisia.

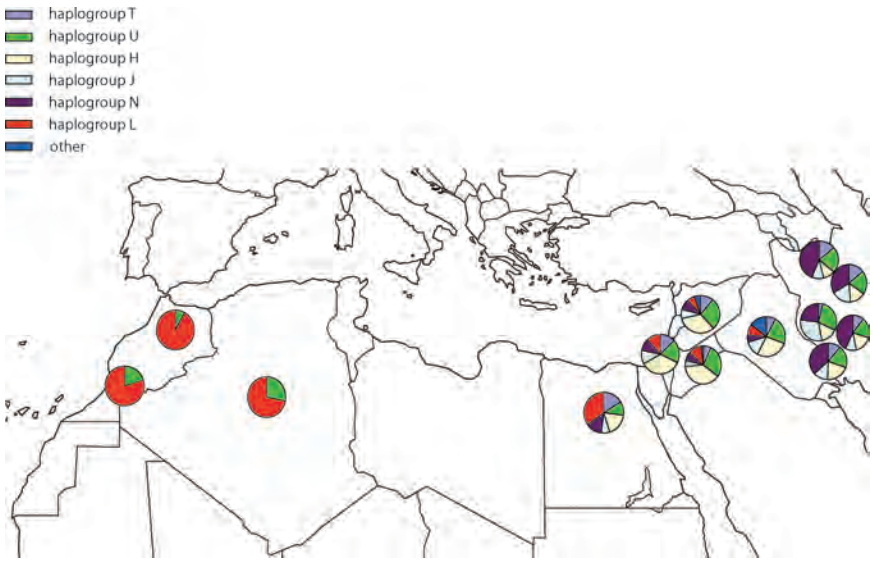

Figure 3 A map of the distribution of mtDNA haplogroups in Indo-European-speaking groups and in Semitic-speaking groups from West Asia, and North Africa.

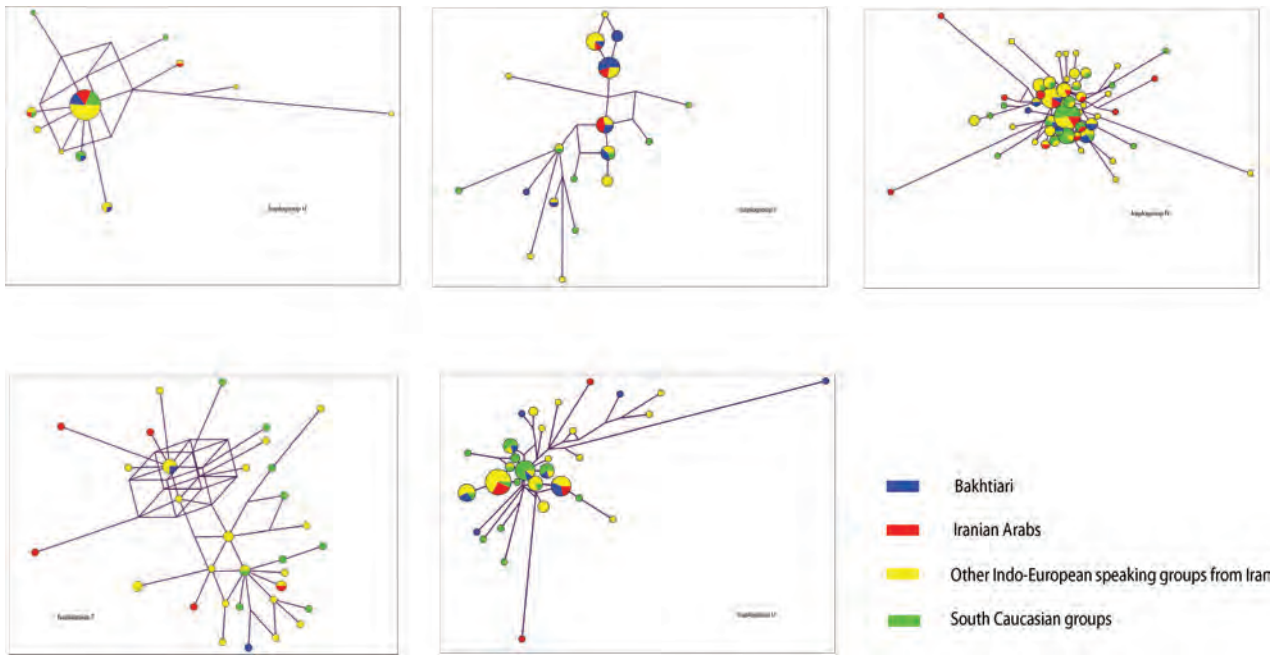

Figure 4 Median networks, based on mtDNA HVI sequences, for mtDNA haplogroups H, U, J, T and N. South Caucasian groups include Georgians, Armenians and Azeri. Other Indo-European groups from Iran are Gilaki and Mazandarani, and Iranians from Tehran and Isfahan.

Y-SNP Haplogroups

Overall, nine Y-SNP haplogroups were found in the Bakhtiari and Iranian Arab groups (Table 3). Haplogroup *J2* (M172)* was found at the highest frequency in the Bakhtiari, followed by haplogroups *F* (M89), G* (M201)* and *K* (M9)*; together, these four haplogroups account for more than 70% of Bakhtiari Y-chromosomes. Iranian Arabs showed a somewhat different pattern: the highest frequency was observed for haplogroup *F* (M89),* followed by haplogroups *J2* (M172), R1a1* (M17)* and *DE* (YAP)*; together, these four haplogroups account for more than 80% of Iranian Arab Y-chromosomes. The frequency of haplogroup *DE* (YAP)* in the Iranian Arabs is higher than

Table 3 Y chromosome haplogroup frequencies in Bakhtiari and Iranian Arabs, and in other relevant groups. HD, haplogroup diversity.

		Haplogroups												
Population	*N*	*DE** *YAP*	*C** *RPS4Y*	*K** *M9*	*P1* *M124*	*P** *M45*	*R1** *M173*	*R1a1** *M17*	*F** *M89*	*G** *M201*	*J2** *M172*	*I** *M170*	other	HD
Bakhiari	53	0.08	0	0.13	0.02	0.02	0.08	0.09	0.19	0.15	0.25	0	n/a	0.86
Iranian Arabs	47	0.11	0	0.04	0.02	0.02	0.04	0.11	0.32	0.06	0.28	0	n/a	0.81
Other Indo-European speaking groups from Iran														
Mazandarani[1]	50	0.04	0.02	0.06	0.04	0.04	0.14	0.06	0.06	0.14	0.40	0	n/a	0.80
Gilaki[1]	50	0.08	0.04	0	0	0	0.22	0.12	0.14	0.10	0.30	0	n/a	0.83
Iranians_Tehran[2]	80	0.06	0	0.10	0.01	0.04	0.08	0.20	0.03	0.05	0.10	0.34	n/a	0.82
Iranians_Isfahan[2]	50	0.02	0	0.14	0.02	0.06	0	0.18	0.22	0.06	0.20	0.10	n/a	0.86
North Iran[3]	33	0	0.03	0.12	0.03	0.09	0.21	0.03	0.09	0.15	0.24	n/a	n/a	0.87
South Iran[3]	117	0.07	0	0.10	0.01	0.03	0.09	0.16	0.16	0.13	0.23	n/a	0.03	0.86
Other Semitic speaking groups from West Asia														
Lebanese[4]	31	0.26	0.03	0.03	0	0	0.06	0.10	0.16	0.03	0.29	0.03	n/a	0.81
Syrian[4]	20	0.20	0	0	0	0	0.15	0.10	0.35	0	0.15	0.05	n/a	0.84
Other Semitic speaking groups from North Africa														
North Egypt[5]	44	0.55	n/a	0.05	0	0.05	0.09	0.02	0.16	n/a	0.09	0	n/a	0.67
South Egypt[5]	29	0.31	n/a	0.10	0	0	0.14	0	0.38	n/a	0.03	0.03	n/a	0.75
Tunisian Arabs[5]	148	0.51	n/a	0.01	0	0.01		0.07	0.37	n/a	0.03	0	n/a	0.60
Algerian Arabs[5]	35	0.57	n/a	0.03	0	0	0	0	0.34	n/a	0.06	0	n/a	0.57
Morrocan Arabs[6]	44	0.73	n/a	0.02	0	0	0.07	0	0.14	n/a	0.02	0.02	n/a	0.46

Data from: [1]Nasidze et al. (2006); [2]Nasidze et al. (2004); [3]Regueiro et al. (2006); [4]Semino et al. (2000); [5]Arredi et al. (2004); [6]Bosch et al. (2001).

that of any Indo-European-speaking group from Iran, but lower than that of any other Semitic-speaking group from either West Asia or North Africa (Table 3). Haplogroup diversity in the Bakhtiari and Arabs falls within the upper limit of the range of values observed in West Asia [Nasidze et al. (2004, 2006), Semino et al. (2000)] and is higher than those in North African Semitic-speaking groups [Arredi et al. (2004) and Bosch et al. (2001)].

Haplogroups *G* (M201)* and *C* (M130)* were not typed in the samples from North Africa by Arredi et al. (2004) and Bosch et al. (2001). We therefore used the same procedure as described elsewhere (Nasidze et al., 2004) in order to include these populations in the comparative analysis; namely, we assigned individuals belonging to these haplogroups to the haplogroup they would have been assigned to if these markers had not been analyzed.

The pairwise F_{st} value (Table 2) between Bakhtiari and Iranian Arabs was not significantly different from zero ($F_{st} = 0.001$, $p = 0.370$). In general, the F_{st} values are below 0.05 for all comparisons among West Asian groups, regardless of whether they are Indo-European speakers or Semitic speakers, and above 0.10 for all comparisons between West Asian groups and North African Semitic-speaking groups. The West Asian Semitic-speaking groups are the most similar of the West Asian groups to the North African Semitic-speaking groups. The MDS analysis (Figure 2b) illustrates these patterns. The Bakhtiari and Iranian Arab groups fall with other West Asian groups; North African Semitic-speaking groups are separate (in the first dimension), and the West Asian Semitic-speaking groups are between other West Asian groups and the North African Semitic-speaking groups.

Y Chromosome STRs

Haplogroup *J2* (M172)* was found in relatively high frequencies in the Iranian Arab and Bakhtiari groups, as well as in other groups from Iran. Haplogroup *G** (M201) was found with similar frequency in Iranian Arabs as in the Iranian groups from Tehran and Isfahan, but in higher frequency in the Bakhtiari, as with the Mazandarani and Gilaki groups from Iran (Nasidze et al., 2004, 2006). To further investigate the relationships of these groups based on these two Y-SNP haplogroups, we typed nine Y-STR loci in individuals with these two Y-SNP haplogroups. Median networks of the Y-STR haplotypes are shown in Figure 5. For both Y-SNP haplogroups, the Bakhtiari are more similar to other Iranian groups than to the Iranian Arabs. Moreover, there is very little sharing of Y-STR haplotypes between Iranian Arabs and other groups from Iran, in contrast to the situation with mtDNA HV1 sequences.

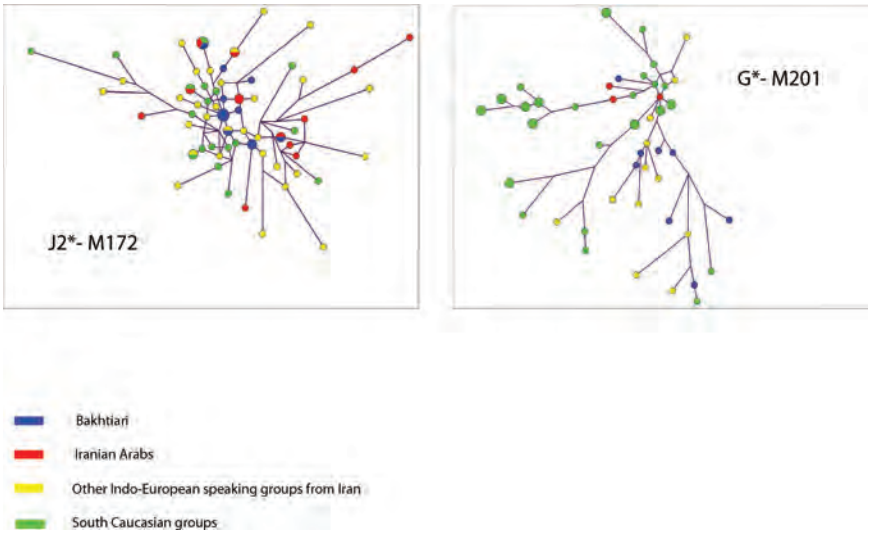

Figure 5 Median networks, based on Y-STR haplotypes, for Y-SNP haplogroups *J2* (M172)* and *G* (M201)*. Groups from South Caucasus and Iran are as listed in Figure 4.

Table 4 AMOVA results according to different classifications.

	mtDNA			Y-SNP		
Classification	Among groups	Among populations within groups	Within populations	Among groups	Among populations within groups	Within populations
Geography	1.19	2.9	95.9	14	4.25	81.75
Linguistic	1.39	2.64	95.96	10.1	5.65	84.25

Geography: West Asia vs North Africa.
Linguistic: Indo-European vs Semitic language families.

Comparison of Mitochondrial and Y-chromosome Data

The geographic and linguistic structure of Bakhtiari and Iranian Arabs, other Indo-European-speaking groups from Iran, and Semitic-speaking groups from West Asia and North Africa, as assessed by mtDNA and Y chromosome variation, was investigated by the AMOVA procedure (Table 4). As is typically seen in human populations, the within-populations proportion of the variance was much higher for mtDNA (about 96%) than for the Y chromosome (about 83%). For both genetic systems, a geographic classification of populations gave a slightly better fit (in terms of higher among-group variance and lower among-populations-within-groups variance) than did a linguistic classification (Table 4). The MDS plot for mtDNA (Figure 2A) shows that the Arabic group from Yemen is an outlier; removing this group from the AMOVA analysis increases the fit of the data to the geographic classification (among-groups component increases to 1.95% and the among-populations-within-groups component decreases to 1.80%) but does not improve the fit to the linguistic classification.

We also investigated the correlation between pairwise F_{st} values, based on mtDNA and Y-SNP data, in the Bakhtiari and Iranian Arabs and their geographic and linguistic neighbors, groups from Iran, the Near East and North Africa. The correlation was significant (Mantel test, $Z = 0.778$; $p = 0.017$), indicating concordant patterns of mtDNA and NRY variation in this part of the world.

Discussion

History has left considerable records that indicate intensive contacts between Arabs and different groups in Iran. Much of the current region known as Iran was conquered by Arab armies of the early Islamic state in the 7^{th}–8^{th} centuries A.D. (Morony, 2006). However, there may also have been contact with Arabic groups before these conquests (Morony, 2006). Since the time of the conquests, there have been numerous migrations of Arab settlers to Iran,

resulting eventually in the spread of Islam and in Arabic becoming the language of religion and literature in Iran. Therefore, the Arabic presence, beginning at least 1400 years ago, is clearly a major turning point in the history of Iran.

The genetic results are in good agreement with the above historical information. The Iranian Arab group shows close affinities with the Bakhtiari and other Iranian Indo-European-speaking groups for both mtDNA and the Y chromosome (Table 2, Figure 2). In fact, for both mtDNA and the Y chromosome, all of the Indo-European-speaking and Semitic-speaking groups from West Asia exhibit generally low levels of differentiation (i.e. F_{st} values are less than 0.05). The significant correlation between mtDNA and NRY F_{st} values, as shown by the Mantel test, further indicates that there are no substantial differences between patterns of mtDNA and NRY variation in this region of the world.

The lack of significant differentiation between west Asian Semitic-speaking and Indo-European-speaking groups indicates that language has not been a substantial barrier to gene flow in this part of the world. This conclusion receives further support when North African Semitic-speaking groups are also considered. West Asian Semitic-speaking groups are more similar genetically to West Asian Indo-European-speaking groups than they are to North African Semitic-speaking groups. This holds true for both mtDNA and NRY variation, as is quite evident in the F_{st} analyses (Table 2) and MDS plots (Figure 2). The North African Semitic-speaking groups differ from West Asian Semitic-speaking groups by having high frequencies of haplogroups that are characteristic of Africa, in particular mtDNA haplogroup L and Y-haplogroup *DE* (YAP)* (Figure 3, Table 3). These haplogroups are either absent from West Asia or found at much lower frequencies. Thus, appreciable mixing between Semitic-speaking and non-Semitic-speaking groups also appears to have occurred in North Africa, as in the case of west Asia. And, this mixing was not particularly sex-biased, since both mtDNA and NRY variation give the same general picture. It is also possible that such genetic similarity could be explained by much older phenomena, such as a long-standing occupation of agro-pastoralists in the region that predates the Persian and Arabic languages (Morony, 2006).

This case adds to our previous studies that have attempted to disentangle the relative influence of geography and language on the genetic relationships of groups whose geographic neighbors are different from their linguistic neighbors. Some general patterns are beginning to emerge from these studies of linguistic enclaves. One pattern is that observed in the present study, namely extensive mixing of groups speaking different languages. A similar pattern has been observed with Indo-European-speaking Armenians and Turkic-speaking Azerbaijanians in the Caucasus, who are genetically similar (for both mtDNA and NRY variation) to each other, and to their Caucasian-speaking geographic neighbors (Nasidze & Stoneking, 2001; Nasidze et al., 2004). In these cases, it is likely that the mixing has been so extensive as to also lead to language replacement; in any event, language does not appear to have been an appreciable barrier to gene flow in these instances.

In contrast to this pattern of extreme mixing is the absence of detectable mixing of a linguistic enclave with the geographic neighbors. An example is the Kalmyks, who speak a Mongolian language, have been in contact with Russian-speaking groups for 300 years, and yet have no detectable genetic admixture with their Russian neighbors (Nasidze et al., 2005). The Yakuts, a Turkic-speaking group in Siberia, also exhibit little or no signs of admixture with their neighboring Tungusic-speaking groups (Pakendorf et al., 2007).

In between these two extremes (complete mixture vs. no detectable mixture) is a pattern where some genetic similarities are detected between a linguistic enclave and both the geographic neighbors and the linguistic neighbors of the enclave. The Turkic-speaking Gagauz of Moldavia, who are surrounded by Indo-European-speaking groups, are an example of this pattern; moreover, the Gagauz show more similarity to other Turkic-speaking groups in patterns of NRY variation than in mtDNA variation, consistent with an effect of patrilocality (Nasidze et al., 2007). Another example would be the Gilaki and Mazandarani of the south Caspian region of Iran; although they speak Indo-Iranian languages closely-related to those of their neighboring Iranian groups, they are thought to have originated from the south Caucasus and do show affiliations with south Caucasian groups in their NRY variation, but affiliations with neighboring Iranian groups in their mtDNA variation (Nasidze et al., 2006). Evenks and Evens, two Tungusic-speaking groups who are widespread across Siberia and in contact with various other groups speaking diverse languages, may also fall into this category as there is some indication of admixture of some Evenk and Even groups with their geographic neighbors with respect to mtDNA (Pakendorf et al., 2007). The above framework for describing the relative influence of geography and language on the genetic relationships of linguistic enclaves is still tentative and highly speculative, based as it is on so few examples. Many more genetic studies of linguistic enclaves (i.e. groups who speak a different language from their geographic neighbors) are needed, and studies of autosomal markers are needed (in addition to mtDNA and NRY variation) in order to obtain accurate estimates of admixture between groups. Moreover, additional questions are

raised by the case studies to date. What social/historical factors influence the amount and type of admixture experienced by neighboring groups who speak different languages? Why, for example have the Kalmyks not admixed with their neighbors, while the Gagauz have admixed extensively with their neighbors? And when genetic admixture does occur between neighboring groups who speak different languages, are there detectable and/or predictable influences on the languages? Although the present study of Semitic-speaking groups adds to our understanding of what happens when groups speaking different languages come into contact, there is still much to be learned.

Acknowledgements

This study focuses on the relationships of populations as reflected by mtDNA and Y-chromosomal variation. It does not aim at ascribing ethnicity to individual groups, nor does it intend to evaluate the self-identification of such groups. We are grateful to the original donors for providing DNA samples. We thank Donald Stilo for useful discussion and suggestions. This research was supported by funding from the Max Planck Society, Germany.

References

Al-Hussein, K. A., Rama, N. R., Butt, A. I., Meyer, B., Rozemuller, E. & Tilanus, M. G. (2002) HLA class II sequence-based typing in normal Saudi individuals. *Tissue Antigens* **60**, 259–261.

Al-Zahery, N., Semino, O., Benuzzi, G., Magri, C., Passarino, G., Torroni, A. & Santachiara-Benerecetti, A. S. (2003) Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol Phylogenet Evol* **28**, 458–472.

Almawi, W. Y., Abou-Jaoude, M. M., Tamim, H., Al-Harbi, E. M., Finan, R. R., Wakim-Ghorayeb, S. F. & Motala, A. A. (2004a) Distribution of HLA class II (DRB1/DQB1) alleles and haplotypes among Bahraini and Lebanese Arabs. *Transplant Proc* **36**, 1844–1846.

Almawi, W. Y., Busson, M., Tamim, H., Al-Harbi, E. M., Finan, R. R., Wakim-Ghorayeb, S. F. & Motala, A. A. (2004b) HLA class II profile and distribution of HLA-DRB1 and HLA-DQB1 alleles and haplotypes among Lebanese and Bahraini Arabs. *Clin Diagn Lab Immunol* **11**, 770–774.

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., Pascali, V. L., Novelletto, A. & Tyler-Smith, C. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* **75**, 338–345.

Bandelt, H. J., Forster, P. & Rohl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37–48.

Bandelt, H. J., Quintana-Murci, L., Salas, A. & Macaulay, V. (2002) The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet* **71**, 1150–1160.

Bosch, E., Clarimon, J., Perez-Lezaun, A. & Calafell, F. (2001) STR data for 21 loci in northwestern Africa. *Forensic Sci Int* **116**, 41–51.

Brakez, Z., Bosch, E., Izaabel, H., Akhayat, O., Comas, D., Bertranpetit, J. & Calafell, F. (2001) Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Ann Hum Biol* **28**, 295–307.

Chbel, F., Nadifi, S., Martinez-Bouzas, C., Louahlia, S., Azeddoug, H. Martinez D. E. & Pancorbo, M. (2003) Population genetic data of eight tetrameric short tandem repeats (STRs) in Casablanca resident population to use in forensic casework. *Forensic Sci Int* **132**, 82–83.

Cordaux, R., Aunger, R., Bentley, G., Nasidze, I., Sirajuddin, S. M. & Stoneking, M. (2004) Independent origins of Indian caste and tribal paternal lineages. *Curr Biol* **14**, 231–235.

Cordaux, R., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. M. & Stoneking, M. (2003) Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur J Hum Genet* **11**, 253–264.

Corte-Real, H. B., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A., Papiha, S., Bertranpetit, J. & Sykes, B. C. (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* **60**, 331–350.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., Wallace, D. C., Oefner, P. J., Torroni, A., Cavalli-Sforza, L. L., Scozzari, R. & Underhill, P. A. (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* **70**, 1197–1214.

Dirienzo, A. & Wilson, A. C. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Science USA* **88**, 1597–1601.

Ethnologue (2000), (www.ethnologue.com).

Finnila, S., Hassinen, I. E., Ala-Kokko, L. & Majamaa, K. (2000) Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. *Am J Hum Genet* **66**, 1017–1026.

Hammer, M. F. & Horai, S. (1995) Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet* **56**, 951–962.

Kayser, M., Brauer, S., Weiss, G., Underhill, P. A., Roewer, L., Schiefenhovel, W. & Stoneking, M. (2000) Melanesian origin of Polynesian Y chromosomes. *Curr Biol* **10**, 1237–1246.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C. & Roewer, L. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine* **110**, 125–133, 141–149.

Kayser, M., Krawczak, M., Excoffier, L., Dieltjes, P., Corach, D., Pascali, V., Gehrig, C., Bernini, L., Jespersen, J., Bakker, E., Roewer, L. & De Knijff, P. (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *American Journal of Human Genetics* **68**, 990–1018.

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R. S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z. & Jin, L. (2001) African origin

of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* **292**, 1151–1153.

K, R., Erler, A., Br, S., Stoneking, M. & Kayser, M.*(2003)* Apparent intra-chromosomal exchange on the human Y chromosome explained by population history. *European Journal of Human Genetics* **11**, 304–314.

Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W. S., Dixon, M. E., Papiha, S. S., Mastana, S. S., Mir, M. R., Ferak, V. & Villems, R. (1999) Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol* **9**, 1331–1334.

Krings, M., Salem, A. E., Bauer, K., Geisert, H., Malek, A. K., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Utermann, G., Sajantila, A., Paabo, S. & Stoneking, M. (1999) mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am J Hum Genet* **64**, 1166–1176.

Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 1–27.

Lucotte, G. & Mercier, G. (2003) Y-chromosome DNA haplotypes in Jews: comparisons with Lebanese and Palestinians. *Genet Test* **7**, 67–71.

Luis, J. R., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Underhill, P. A., Cavalli-Sforza, L. L. & Herrera, R. J. (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am J Hum Genet* **74**, 532–544.

Manni, F., Leonardi, P., Barakat, A., Rouba, H., Heyer, E., Klintschar, M., McElreavy, K. & Quintana-Murci, L. (2002) Y-chromosome analysis in Egypt suggests a genetic regional continuity in Northeastern Africa. *Hum Biol* **74**, 645–658.

Morony, M. (2006) Arab II. Arab conquest of Iran. IN YARSHATER, E. (Ed.) *The Encyclopaedia Iranica*. Center for Iranian Studies, Columbia University (http://iranica.com/articlenavigation/index.html).

Nasidze, I., Ling, E. Y., Quinque, D., Dupanloup, I., Cordaux, R., Rychkov, S., Naumova, O., Zhukova, O., Sarraf-Zadegan, N., Naderi, G. A., Asgary, S., Sardas, S., Farhud, D. D., Sarkisian, T., Asadov, C., Kerimov, A. & Stoneking, M. (2004) Mitochondrial DNA and Y-chromosome variation in the caucasus. *Ann Hum Genet* **68**, 205–221.

Nasidze, I., Quinque, D., Dupanloup, I., Cordaux, R., Kokshunova, L. & Stoneking, M. (2005) Genetic evidence for the Mongolian ancestry of Kalmyks. *Am J Phys Anthropol* **128**, 846–854.

Nasidze, I., Quinque, D., Rahmani, M., Alemohamad, S. A. & Stoneking, M. (2006) Concomitant replacement of language and mtDNA in South Caspian populations of Iran. *Curr Biol* **16**, 668–673.

Nasidze, I., Quinque, D., Udina, I., Kunizheva, S. & Stoneking, M. (2007) The Gagauz, a linguistic enclave, are not a genetic isolate. *Ann Hum Genet* **71**, 379–389.

Nasidze, I. & Stoneking, M. (2001) Mitochondrial DNA variation and language replacements in the Caucasus. *Proc Biol Sci* **268**, 1197–1206.

Nebel, A., Filon, D., Brinkmann, B., Majumder, P. P., Faerman, M. & Oppenheim, A. (2001) The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am J Hum Genet* **69**, 1095–1112.

Pakendorf, B., Novgorodov, I. N., Osakovskij, V. L. & Stoneking, M. (2007) Mating patterns amongst Siberian reindeer herders: Inferences from mtDNA and Y-chromosomal analyses. *Am J Phys Anthropol* **133**, 1013–1027.

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A. S., Coppa, A., Ayub, Q., Mohyuddin, A., Tyler-Smith, C., Qasim Mehdi, S., Torroni, A. & McElreavey, K. (2004) Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am J Hum Genet* **74**, 827–845.

Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S. H., Hammer, M. F., Mehdi, S. Q., Ayub, Q., Qamar, R., Mohyuddin, A., Radhakrishna, U., Jobling, M. A., Tyler-Smith, C. & McElreavey, K. (2001) Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am J Hum Genet* **68**, 537–542.

Redd, A. J., Takezaki, N., Sherry, S. T., McGarvey, S. T., Sofro, A. S. M. & Stoneking, M. (1995) Evolutionary history of the COII/tRNA-lys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol* **12**, 604–615.

Regueiro, M., Cadenas, A.M., Gayden, T., Underhill, P.A., Herrera, R.J. (2006) Iran: tricontinental nexus for Y-chromosome driven migration. *Hum Hered* **61**, 132–143.

Richards, M., Rengo, C., Cruciani, F., Gratrix, F., Wilson, J. F., Scozzari, R., Macauley, V. & Torroni, A. (2003) Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. *Am J Hum Genet* **72**, 1058–1064.

Schneider, S., Roessli, D. & Excoffier, L. (2000) *Arlequin ver 2.000: A software for population genetics data analysis*, University of Geneva, Switzerland, Genetics and Biometry Laboratory.

Scozzari, R., Cruciani, F., Santolamazza, P., Malaspina, P., Torroni, A., Sellitto, D., Arredi, B., Destro-Bisol, G., De Stefano, G., Rickards, O., Martinez-Labarga, C., Modiano, D., Biondi, G., Moral, P., Olckers, A., Wallace, D. C. & Novelletto, A. (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* **65**, 829–846.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P. J., Zhivotovsky, L. A., King, R., Torroni, A., Cavalli-Sforza, L. L., Underhill, P. A. & Santachiara-Benerecetti, A. S. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* **74**, 1023–1034.

Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. S., Cavalli-Sforza, L. L. & Underhill, P. A. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155–1159.

Stoneking, M. (2005) Gene, geographie und Sprache. In: Hauksa, G. (Ed.) *Gene, Sprachen und ihre Evolution*. Regensburg, Universitätsverlag Regensburg GmbH. Germany.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L. & Oefner, P. J. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358–361.

Vigilant, L., Pennington, R., Harpending, H., Kocher, T. D. & Wilson, A. C. (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA* **86**, 9350–9354.

Wilson, J. & Erlandsson, R. (1998) Sexing of human and other primate DNA. *Biol. Chem*. 1287–1288.

Y Chromosome Consortium (2002) A nomenclature system for the tree of Human Y- chromosomal binary haplogroups. *Genome Res* **2**, 339–348.

Yoshimoto, K., Iwahana, H., Fukuda, A., Sano, T. & Itakura, M. (1993) Rare mutations of the Gs alpha subunit gene in human endocrine tumors. Mutation detection by polymerase chain reaction-primer-introduced restriction analysis. *Cancer* **72**, 1386–1393.

Received: 20 September 2007
Accepted: 11 October 2007