ORIGINAL ARTICLE

Mitochondrial DNA and Y-chromosomal stratification in Iran: relationship between Iran and the Arabian Peninsula

Maria C Terreros¹, Diane J Rowold¹, Sheyla Mirabal^{1,2} and Rene J Herrera¹

Modern day Iran is strategically located in the tri-continental corridor uniting Africa, Europe and Asia. Several ethnic groups belonging to distinct religions, speaking different languages and claiming divergent ancestries inhabit the region, generating a potentially diverse genetic reservoir. In addition, past pre-historical and historical events such as the out-of-Africa migrations, the Neolithic expansion from the Fertile Crescent, the Indo-Aryan treks from the Central Asian steppes, the westward Mongol expansions and the Muslim invasions may have chiseled their genetic fingerprints within the genealogical substrata of the Persians. On the other hand, the Iranian perimeter is bounded by the Zagros and Albrez mountain ranges, and the Dasht-e Kavir and Dash-e Lut deserts, which may have restricted gene flow from neighboring regions. By utilizing high-resolution mitochondrial DNA (mtDNA) markers and reanalyzing our previously published Y-chromosomal data, we have found a previously unexplored, genetic connection between Iranian populations and the Arabian Peninsula, likely the result of both ancient and recent gene flow. Furthermore, the regional distribution of mtDNA haplogroups J, I, U2 and U7 also provides evidence of barriers to gene flow posed by the two major Iranian deserts and the Zagros mountain range.

Journal of Human Genetics (2011) 56, 235–246; doi:10.1038/jhg.2010.174; published online 17 February 2011

Keywords: Arabian Peninsula; Iranian Plateau; mtDNA; Persia; Y-chromosome

INTRODUCTION

Present day Iran is land-bound by multiple nations, including Iraq to the west, Turkey, Armenia and Azerbaijan to the northwest, Turkmenistan to the northeast, and Afghanistan and Pakistan to the east. Its southern perimeter consists of the Persian Gulf, a maritime gate between Iran and the eastern coastal region of Arabian Peninsula, which encompasses Kuwait, Saudi Arabia, Bahrain, Qatar, UAE and Oman. The Strait of Hormuz, the narrowest section of this waterway, separates southwestern Iran from the northeast tip of Oman by a mere 46.7 km of shallow water littered with numerous tiny islands.^{1,2} Climatologists and historians have resolved that episodic sea level oscillations have exposed the ocean floor between these two regions at various points during the ancient past.¹ This land bridge may have been a critical segment of the southern coastal route facilitating the out-of-Africa migration of modern humans ~60,000 years before present.³

Iran's pivotal geographical location and its proximity to trans-continental migratory routes have led researchers to believe that the Iranian territory may have also played a key role in subsequent migrations, both prehistoric and historic, between Africa, Asia and Europe. This may be especially true for the Neolithic agricultural

diffusion^{4,5} given Iran's proximity to the Fertile Crescent where pastoral agriculturalists are postulated to have commenced their trek into both western and eastern Eurasia. Furthermore, Iran's location near the Central Asian steppes may have facilitated the westward-bound Mongol invasion in the thirteenth century AD,⁶ the ancient Silk Road passages,⁷ as well as the Central Asian settlement of pre-Islamic Iranian people, namely the Sogdians, Chorasmians, Scythians and Alans.⁸ Studies utilizing autosomal markers have thus far pointed to little or no differentiation between Iranian ethnic groups,^{9–11} and altogether they have been found to resemble most populations from the southern Balkan region and Anatolia.^{9–13}

Iran's presumed role as a cultural and ethnic nexus, however, may have been undermined by physical features capable of posing substantial barriers to demic movements. The Iranian Plateau is framed by two major mountain chains, the Kuhha ye Zagros mountains, bordering southern and western Iran, and the Kuhha ye Alborz range, which extends 998 km along Iran's northern boundary with the Caspian Sea. Both cordilleras border the elevated central Iranian Plateau, which is covered, primarily, by two major deserts, the Dasht-e Kavir in north central Iran and the Dash-e Lut in the southeast. Together, these topogeographical obstacles may have

¹Department of Molecular and Human Genetics, College of Medicine, Florida International University, Miami, FL, USA and ²Department of Medicine, Miller School of Medicine, University of Miami, Miami, FL, USA

Correspondence: Dr RJ Herrera, Department of Molecular and Human Genetics, College of Medicine, Florida International University, 11200 SW 8 Street, OE 304, Miami, FL 33199, USA.

E-mail: herrera@fiu.edu

Received 28 October 2010; revised 8 December 2010; accepted 14 December 2010; published online 17 February 2011

restricted gene flow across as well as within the Iranian perimeter. These geographical characteristics may have forced migrants to settle either in the northern or southern areas of Iran, steering clear of its harsh interior desert terrain.

The contention that these aforementioned geographical barriers may have restricted genetic flow within Iran and between Iran and neighboring regions is supported by Y-chromosome data reported by Wells *et al.*¹⁴ as well as the Central Asian mitochondrial DNA (mtDNA) analysis performed by Quintana-Murci *et al.*¹⁵ In the former study, examination of the paternal gene pools of north and south Iranians reveals a clear demarcation in the distribution of R1a1 lineages (defined by the M17/M198 mutations) between the two sections of the Plateau.¹⁴ Similarly, Quintana-Murci *et al.*¹⁵ found greater proportions of mtDNA haplogroups N1b, R2, HV2, U7, J2 and T* in northern Iran, whereas M*, N*, R5, B, pre-HV1, U2*, U2e and U3 lineages were higher in the south. A recent study, based on both Y-chromosome and mtDNA analyses, found little to no differences in ethnic groups (Indo-European speakers versus Semitic speakers) residing in close geographical proximity within Iran.¹⁶ Furthermore, another mtDNA investigation led to the conclusion that two Indo-Iranian-speaking Talysh groups from Iran and Azerbaijan, that claim a common ancestry, were genetically similar.¹⁷ In the same study, however, Y-chromosomal marker composition was shown to differ considerably between the Iranian and Azerbaijani Talysh, with the Azerbaijan Talysh more closely resembling the Azerbaijan neighbors than its Iranian counterpart.¹⁷ Results reported by Regueiro *et al.*¹ also indicate differential gene flow between northern and southern Iranian groups (divided by the Dasht-e Kavir and Dash-e Lut deserts) not only with respect to the R-M198 mutation, as illustrated by Wells *et al.*,¹⁴ but also with R-M269 as well. The same study also reveals significant divergence in the overall Y-haplogroup distributions between northern and southern Iranians as well as between both groups and other spatially separated Iranian populations (the Esfahan of Central Iran reported by Nasidze *et al.*¹⁸ and Uzbekistan discussed in the study by Wells *et al.*¹⁴). In spite of these efforts, a consensus has not yet been reached as to the source populations, overall genetic relationships and degree of stratification between different Iranian regions.

In the current inquiry, we explore Iran's relative importance, in contrasting roles, as a genetic nexus and as a genetic barrier from the perspective of high-resolution mtDNA analyses. We examine the maternal lineages of the 148 Iranian males featured in Regueiro *et al.*¹ using the same north/south divide outlined in that Y-chromosome analysis. The present investigation compares and contrasts the overall mtDNA haplogroup frequency patterns and the phylogenetic relationships observed among geographically relevant Eurasian and North African groups representing areas to the north, south, east and west of Iran (see Table 1). This study also expands upon the mtDNA analysis of Iran presented in the study by Quintana-Murci *et al.*¹⁵ by utilizing high-resolution markers and through inclusion of samples from the southern coast and northeast Iranian territories as well as several other pertinent regions of the Plateau. We have also reanalyzed our previously published Y-chromosomal data to include the counterpart reference populations herein used as a point of comparison with our mtDNA data in order to directly assess the contrasting patterns of maternal and paternal inheritance throughout the region.

MATERIALS AND METHODS

Population information

Whole blood was collected from a total of 148 healthy Iranian males whose ancestry can be traced back to at least two generations. Samples were obtained in strict compliance with National Institutes of Health guidelines as well as with

those indicated by the Institutional Review Board of Florida International University. Donors from throughout modern-day Iran were divided geographically into northern ($n=31$) and southern ($n=117$) groups segregated by a virtual line at the southern fringe of the Dash-e Kavir desert extending horizontally to Iran's eastern and western borders (see Supplementary Table 1 for the location of the collection of each individual included in the study). The Iranian groups (Iran North (IN) and Iran South (IS)) are analyzed phylogenetically and statistically in a tri-continental context against 44 previously published, geographically targeted reference populations from various neighboring regions: northwest Africa, the Caucasus, central Asia and southwest Asia^{5,15,19–28} (Table 1). Due to overlaps and differences in the resolution of the markers employed, populations marked with an asterisk (*) are only included in the NETWORK analysis and the time to most recent common ancestor estimates.

DNA extraction and sequencing

DNA from the 148 Iranian samples was extracted by standard phenol–chloroform methods as previously described,^{1,29,30} ethanol precipitated and diluted in 10 mM Tris-EDTA, pH 8.0. All samples were stored at -80°C when not in use. The mtDNA control region (nucleotide positions 15997–409) encompassing hypervariable regions I and II was PCR amplified with the single primer pair described in the study by Stoneking *et al.*³¹ Following amplification, the 1250 base pair amplicon was quantified with an Agilent Technologies (Santa Clara, CA, USA) 2100 Bioanalyzer. In all, 7–10 ng of the PCR product was then utilized for sequencing reactions (in both directions), employing the same primer pair used for PCR amplification and the dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, CA, USA). Sequenced products were separated in an ABI 3100 Genetic Analyzer with POP 6TM polymer (Applied Biosystems).

mtDNA haplotyping

The resulting sequences were aligned to the revised Cambridge Reference Sequence^{32,33} using the BioEdit sequence alignment software (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). Haplogroup assignment was based on the classification scheme outlined by Macaulay *et al.*³⁴ Following alignment, restriction fragment length polymorphism analysis³⁵ of coding sequences was performed to confirm designations of haplogroup assignments as previously described.³⁶ The tentative assignments of haplogroups H, HV, M, N, R, U, W, I, K, JT and T1 were substantiated by restriction endonuclease digestions at the following diagnostic sites: both $-14\,766$ MseI and -7025 AluI for haplogroup H, $-14\,766$ MseI for HV, $+10\,397$ AluI for M, $+10\,871$ MnlI for N, $+12\,703$ MboII for R, and both $+12\,703$ MboII and $+12\,308$ HinfI for haplogroup U, -8994 HaeIII for W, $+10\,032$ for I, -9052 HaeII for haplogroup K, $+4216$ NlaIII for JT, and $-12\,629$ AvaII for T1. The restriction endonuclease site at position 12308 was confirmed via a mismatched primer approach as previously reported.³⁶ Haplogroup H samples were typed for sub-haplogroups H1–H13 using restriction enzyme analyses and/or allele-specific PCR according to Achilli *et al.*,³⁷ Loogväli *et al.*,³⁸ Roostalu *et al.*³⁹ and Gayden *et al.*⁴⁰

Phylogenetic/statistical analyses

To assess the statistical significance of haplogroup frequencies, the Bayesian 0.95 credible region (0.95 CR) was calculated using the SAMPLING program provided by Vincent Macaulay. Bonferroni corrections and χ^2 analyses were also conducted to test for significant differences in the observed haplogroup frequencies between the IN and IS groups.

In order to ascertain the proportion of gene flow from the surrounding regions into the Iranian domain, admixture analyses based on frequency distributions were conducted using regional population groups as parentals and collections within the Iranian Plateau as hybrids with the aid of the SPSS version 16.0 software.^{41,42} Parental groups were assembled as specified in Table 1 for mtDNA and in Supplementary Table 2 for the Y-chromosome.^{1,14,18,43–60}

To gauge significant deviations from neutral theory,⁶¹ the Ewens-Watterson homozygosity test^{62,63} was performed on Iranian mtDNA control region sequences (nucleotide positions 16019–16569). In addition, mtDNA control region sequence diversity was evaluated by nucleotide diversity index.⁶⁴ Time

Table 1 Populations analyzed

Geographical area	N	Code	Country	Population	Language	References
Northeast Africa	47	TUN	Tunisia	Arabs	Afro-Asiatic	Plaza <i>et al.</i> ²⁰
	118	EGY	Egypt	Arabs/Berbers	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	53	ETH	Ethiopia	Arabs	Afro-Asiatic	Kivisild <i>et al.</i> ²²
Arabian Peninsula	105	OMA	Oman	Arabs	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	90	QAT	Qatar	Arabs	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	131	UAE	United Arab Emirates	Arabs	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	50	YEM	Yemen	Arabs	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	553	SAA	Saudi Arabia	Arabs	Afro-Asiatic	Abu-Amro <i>et al.</i> ²⁷
	29	BED	Arabia	Bedouin ^a	Afro-Asiatic	Richards <i>et al.</i> ⁵
Caucasus	40	AZE	Azerbaijan	Azeris	Altaic	Quintana-Murci <i>et al.</i> ¹⁵
	192	ARM	Armenia	Armenian ^a	Indo-European	Richards <i>et al.</i> ⁵
	139	GEO	Georgia	Georgian	Caucasian	Tambets <i>et al.</i> ¹⁹
Levant/Anatolia	69	SYR	Syria	Syrian	Afro-Asiatic	Richards <i>et al.</i> ⁵
	39	JOR	Jordan	Arabs	Afro-Asiatic	Rowold <i>et al.</i> ²⁵
	216	IRQ	Iraq	Iraqi	Afro-Asiatic	Richards <i>et al.</i> ⁵
	117	PAL	Palestine	Palestinian	Afro-Asiatic	Richards <i>et al.</i> ⁵
	388	TUR	Turkey	Anatolia	Altaic	Tambets <i>et al.</i> ¹⁹
Balkan Peninsula	130	HER	Herzegovina	Herzegovina	Indo-European	Cvjetan <i>et al.</i> ²¹
	247	BOS	Bosnia	Bosnia	Indo-European	Cvjetan <i>et al.</i> ²¹
	117	SER	Serbia	Serbia	Indo-European	Cvjetan <i>et al.</i> ²¹
	96	CRO	Croatia	Croatia	Indo-European	Cvjetan <i>et al.</i> ²¹
	37	MAC	Macedonia	Macedonia	Indo-European	Bosch <i>et al.</i> ²⁴
	42	ALB	Albania	Albania	Indo-European	Bosch <i>et al.</i> ²⁴
	25	GRE	Greece	Greece	Indo-European	Bosch <i>et al.</i> ²⁴
	42	PER	Iran	Persian	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
Iranian Plateau	31	IRN	Iran	Iranian North	Indo-European	Present study
	117	IRS	Iran	Iranian South	Indo-European	Present study
	37	GIL	Gilaki	Caspian Sea	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
	21	MAZ	Mazandarian	Caspian Sea	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
	32	KUR	Turkmenistan	Kurdish ^a	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
	41	TUM	Turkmenistan	Turkmen ^a	Altaic	Quintana-Murci <i>et al.</i> ¹⁵
Central Asia	249	TUK	Turkmenistan	Turkmenistan	Altaic	Irwin <i>et al.</i> ²⁸
	244	TAJ	Tajikistan	Tajikistan	Indo-European	Irwin <i>et al.</i> ²⁸
	256	KAZ	Kazakhstan	Kazakhstan	Altaic	Irwin <i>et al.</i> ²⁸
	328	UZB	Uzbekistan	Uzbekistan	Altaic	Irwin <i>et al.</i> ²⁸
	249	KYR	Kyrgyzstan	Kyrgyzstan	Altaic	Irwin <i>et al.</i> ²⁸
	100	SEP	Southeast Pakistan	Karachi	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
	38	SWP	Southwest Pakistan	Brahui	Dravidic	Quintana-Murci <i>et al.</i> ¹⁵
	44	NWP	Northwest Pakistan	Pathan	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
Southwest Asia	98	AFG	Afghanistan	Afghanistan	Indo-European	Irwin <i>et al.</i> ²⁸
	59	SHI	Uttar Pradesh	Shia	Indo-European	Terreros <i>et al.</i> ²⁶
	60	SUN	Uttar Pradesh	Sunni	Indo-European	Terreros <i>et al.</i> ²⁶
	5	UPR	Uttar Pradesh	Uttar Pradesh	Indo-European	Metspalu <i>et al.</i> ²³
	34	GUJ	Gujarat	Gujarat	Indo-European	Quintana-Murci <i>et al.</i> ¹⁵
	37	HIM	Himachal	Himachal	Tibeto-Burman	Metspalu <i>et al.</i> ²³
	109	PUN	Punjab	Punjab	Indo-European	Metspalu <i>et al.</i> ²³

^aThese populations were used for network analyses; however, they were not included in other phylogenetic analyses owing to resolution/overlap with the rest of the collections involved.

to most recent common ancestor of several mtDNA haplogroups was calculated according to Forster *et al.*⁶⁵ using an average mutation rate of one nucleotide change in 20 180 years for nucleotide positions 16 090–16 365.^{5,22,66,67}

Multidimensional scaling (MDS) plots were generated with the SPSS program version 16.0 for both mtDNA and Y-chromosomal data sets in order to infer genetic relationships among IN, IS and all reference populations. The MDS analyses were performed utilizing the observed frequency of haplogroups. The software package *Surfer* (<http://www.goldensoftware.com>) was employed to obtain contour plots to visualize regional frequency gradients of specific mtDNA and Y-chromosomal haplogroups.

A series of global (multipopulation) and Iranian (segregated into north and south collections) network analyses based on the most frequent Iranian

haplogroups were generated with the reduced median method⁶⁸ and the NETWORK 4.1 software program (<http://www.fluxus-engineering.com>).

Reanalyzed, previously published Y-chromosome haplogroup frequency data

The Y-chromosome haplogroup frequency data published in the study by Regueiro *et al.*¹ were utilized to perform admixture, MDS and contour plot analyses as described in the Phylogenetic/statistical analyses section of Materials and methods section (see above). In addition, we employed the Y-chromosome haplogroup frequency data from the study by Regueiro *et al.*¹ to generate a map of haplogroup distribution in the two Iranian collections and the 36 reference populations.

RESULTS

Iranian and global haplogroup distribution

Haplogroup frequencies and the associated 0.95 credibility region data for IN, IS and Iran total are presented in Table 2. A phylogeographic map, constructed on the basis of haplogroup distributions in the above-mentioned Iranian collections and the 44 reference populations, is provided in Figure 1a. Also, for a complete list of the high-resolution haplotypes observed in the Iranian populations surveyed in the current report, see Supplementary Table 1. Phylogeographic frequency distributions based on the major Y-chromosomal haplogroups are

presented in Figure 1b. References for all populations included in this projection are provided in Supplementary Table 2.

IN and IS possess the same three most prominent mtDNA haplogroups, H, J and U; however, the frequency distribution of these lineages varies between the northern and southern regions of the Plateau. For the northern collection, J is by far the most frequent haplogroup (0.355, 0.95 CR: 0.211–0.532) and is followed in decreasing order by U (0.161, 0.95 CR: 0.072–0.328) and H (0.131, 0.95 CR: 0.053–0.290) lineages. In contrast, haplogroup U (0.222, 0.95 CR: 0.157–0.306) predominates in the southern population, and

Table 2 Haplogroup frequencies and credible region of Iran

Haplogroup	Iran North		Iran South		Iran (total)	
	Number of samples	Frequency (credible region)	Number of samples	Frequency (credible region)	Number of samples	Frequency (credible region)
M*	2	0.065 (0.020–0.208)	7	0.060 (0.030–0.118)	9	0.0610 (0.033–0.112)
M1	1	0.032 (0.008–0.162)			1	0.0067 (0.002–0.037)
M total	3	0.097 (0.035–0.250)	7	0.060 (0.030–0.118)	10	0.0680 (0.037–0.120)
N*	1	0.032 (0.008–0.162)	10	0.085 (0.047–0.150)	11	0.0740 (0.042–0.128)
N1b			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
N1d			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
N total	1	0.032 (0.008–0.162)	12	0.103 (0.060–0.171)	13	0.0880 (0.052–0.145)
I	3	0.097 (0.035–0.250)	2	0.017 (0.005–0.060)	5	0.0340 (0.015–0.077)
W	1	0.032 (0.008–0.162)	3	0.026 (0.009–0.073)	4	0.0270 (0.011–0.067)
R*			3	0.026 (0.009–0.073)	3	0.0203 (0.007–0.058)
R2	1	0.032 (0.008–0.162)			1	0.0067 (0.002–0.037)
R6			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
R total	1	0.032 (0.008–0.162)	4	0.034 (0.014–0.085)	5	0.0340 (0.015–0.077)
J*	1	0.032 (0.008–0.162)	2	0.017 (0.005–0.060)	3	0.0203 (0.007–0.058)
J			3	0.026 (0.009–0.073)	3	0.0203 (0.007–0.058)
J1a			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
J1b	7	0.227 (0.115–0.400)	7	0.060 (0.030–0.118)	14	0.0946 (0.057–0.153)
J2	3	0.097 (0.035–0.250)	4	0.034 (0.014–0.085)	7	0.0470 (0.023–0.094)
J total	11	0.355 (0.211–0.532)	17	0.145 (0.093–0.221)	28	0.1890 (0.134–0.260)
T*			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
T1			2	0.017 (0.005–0.060)	2	0.0135 (0.004–0.048)
T3			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
T total			4	0.034 (0.014–0.085)	4	0.0270 (0.011–0.067)
U*	1	0.032 (0.008–0.162)	4	0.034 (0.014–0.085)	5	0.0340 (0.015–0.077)
U1	1	0.032 (0.008–0.162)	3	0.026 (0.009–0.073)	4	0.0270 (0.011–0.067)
U2			3	0.026 (0.009–0.073)	3	0.0203 (0.007–0.058)
U2a			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
U3			2	0.017 (0.005–0.060)	2	0.0135 (0.004–0.048)
U4			2	0.017 (0.005–0.060)	2	0.0135 (0.004–0.048)
U5	1	0.032 (0.008–0.162)			1	0.0067 (0.002–0.037)
U5a1			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
U5a1a	1	0.032 (0.008–0.162)	1	0.009 (0.002–0.046)	2	0.0135 (0.004–0.048)
U7	1	0.032 (0.008–0.162)	9	0.077 (0.041–0.140)	10	0.0680 (0.037–0.120)
U total	5	0.161 (0.072–0.328)	26	0.222 (0.157–0.306)	31	0.2090 (0.152–0.282)
K	1	0.032 (0.008–0.162)	9	0.077 (0.041–0.140)	10	0.0680 (0.037–0.120)
HV	1	0.032 (0.008–0.162)	6	0.051 (0.024–0.107)	7	0.0470 (0.023–0.094)
H	4	0.130 (0.053–0.290)	18	0.154 (0.100–0.230)	22	0.1490 (0.100–0.215)
H2			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
H2a1			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
H4			3	0.026 (0.009–0.073)	3	0.0203 (0.007–0.058)
H7			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
H total	4	0.130 (0.053–0.290)	24	0.205 (0.142–0.287)	28	0.1890 (0.134–0.260)
L1C2			1	0.009 (0.002–0.046)	1	0.0067 (0.002–0.037)
L0a2			2	0.017 (0.005–0.060)	2	0.0135 (0.004–0.048)
L total			3	0.026 (0.009–0.073)	3	0.0203 (0.007–0.058)
Total	31		117		148

Figure 1 (a) mtDNA haplogroup distributions. Population designations are followed as presented in Table 1. (b) Y-chromosomal haplogroup distributions. Population designations are followed as presented in Supplementary Table 2.

is followed closely by haplogroups H (0.205, 0.95 CR: 0.142–0.287) and J (0.145, 0.95 CR: 0.093–0.221). Yet, according to the results of the χ^2 tests (Supplementary Table 3), the only haplogroups that differ significantly between these two collections are J total (0.355 versus 0.145 for IN and IS, respectively) and J1b (0.226 versus 0.060 for IN and IS, respectively). Despite other notable differences between the IN and IS collections, including the complete absence of L haplogroups as well as H and U sub-haplogroups in IN, the significance of these dissimilarities was not supported by the Pearson's χ^2 statistical analysis. The lack of L haplogroups, and H and U diversification in the IN population may be merely the result of smaller sample size ($N=31$).

MDS

In the MDS plot based on mtDNA (Figure 2a), the southwest Asian populations are restricted to the left portion of the chart, the majority of which sequester in the lower left quadrant. The Afghanis group with the central Asians in the lower center of the graph is an expected association given that Afghanistan is frequently considered as a part of central Asia.⁴⁰ The Balkan Peninsula populations form a tight cluster at the right-most extreme of the lower right quadrant, whereas the populations from the Caucasus, Levant/Anatolia and North Africa conform to a ladder-like pattern that extends from the extreme right center of the chart into the upper right quadrant. The central and southern Iranians are close to each other and to the North Africa and Levant/Anatolia assemblages. The Peninsular Arabs partition to the left of above mentioned groups of populations; interestingly, IN (present study) is located within this cluster, specifically close to the Qatar collection. Two other North Iranian populations from the South Caspian region, the Gilaki and Mazandarian, are positioned between Arabian and Levant groups, closest to Saudi Arabia, Oman and Egypt.

Whereas distinct geographical-based clustering is apparent within the mtDNA plot, the Y-chromosomal counterpart contains overlapping regional groups as opposed to clear subdivisions (Figure 2b). The left portion of the chart, toward the center and in the upper quadrant, consists of the southwest and central Asian populations. The IN collection and the Georgian group lie toward the center of the lower left portion of the chart. The IS population plots proximal to the Azerbaijan collection, suggestive of a connection between Iran and Caucasia. The IS, Turkey and Armenia groups partition toward the vertical axis of the lower portion of the plot. The Central (Esfahān) Iranian population lies on the dividing line between the right and left upper quadrants, whereas the Balkan populations form a loose assemblage in the upper right quadrant. The Somalis and Ethiopians are sequestered to the right extreme of the plot, whereas the other North African group from Egypt is adjacent to a closely intertwined Levant/Peninsular Arab grouping. The Yemenis are the only population from the Arabian Peninsula that deviates from this spatial pattern, likely due to their previously described geographical isolation from the rest of the Peninsular Arabs.⁵⁸

Contour plots

Contour plots based on the geographical clines of mtDNA haplogroups J, I, H, U, U2, U4, U5 and U7 are presented in Figure 3a, while Y-chromosomal haplogroups J and R are presented in Figure 3b. Haplogroups J and I (Figure 3a) are widespread throughout the Arabian Peninsula and in the region of the Fertile Crescent from the Levant to Iran. Interestingly, both haplogroups exhibit frequency pinnacles in IN as well as within the Arabian Peninsula. Haplogroup H displays a clear demic gradient from Europe into the Near/Middle East, with frequencies dwindling in the east of the Indus Valley.

When all branches of haplogroup U are considered together, there are no well-defined frequency clines observed except for the obvious lack of the haplogroup within the African continent (Figure 3a). Upon sub-dividing the branches of the aforementioned haplogroup (only the most highly represented branches within the Iranian domain were further explored), clear region-specific gradients are detected. For example, sub-haplogroups U2 and U7 are widely distributed throughout Asia and the Arabian Peninsula, exhibiting their highest frequencies in the southwest Asian collections and displaying east-to-west frequency clines. It is noteworthy that both haplogroups are found in the Arabian Peninsula. U7 specifically exists at considerable levels (10.8% in the Gilaki and 3.2% in IN) in north Iranian territories beyond the Indus Valley where U2 dwindles. The opposite scenario is observed for U4 and U5, which instead exhibit west-to-east demic dispersal, with high frequencies in the Balkan Peninsula and increasingly lower frequencies toward the Near and Middle East. Neither of these sub-haplogroups is present to any appreciable level in the southwest Asian groups, with the exception of Afghanistan for U4. U5 presents an additional high-frequency focus in northern Iran, exhibiting a decreasing gradient frequency distribution toward the Arabian Peninsula. A second high-density point for both haplogroups is apparent in central Asia (Supplementary Table 4).

Y-chromosomal haplogroup J is present in high frequencies throughout the Arabian Peninsula and the Levant, dissipating considerably in all directions (Figure 3b). Haplogroup R, on the other hand, presents very high frequencies in the central Asian/southwest Asian regions, with levels decreasing immediately beyond the Indus Valley area. A slight increase in frequencies is observed in the Balkan Peninsula (Supplementary Table 5).

Admixture analyses

Clear differences are observed in the maternal versus paternal gene pools of each specific Iranian region, as well as when these are compared with each other (Table 3). The IN collection exhibits a 92.1% influence from the Peninsular Arabs when mtDNA is examined while this impact diminishes to 11.2% when Y-chromosomal data are examined. Similarly, the north Iranian Caspian populations of Gilaki and Mazandarian as well as central Iran and IS exhibit considerable proportions of mtDNA from the Arabian Peninsula (43.5 and 64.3%, 53.3 and 52.1%, respectively), whereas no apparent effect is seen in the Y-chromosomal component for central Iran and only 7.3% is observed for IS. Unfortunately, the Y-chromosome haplogroup counterparts were not reported at the resolution required for these analyses in the north Iran/Caspian populations. Balkan inputs are observed in the mtDNA pool of both IN (7.9%) and IS (23.1%), but are absent in Central Iran and in the other two north Iran collections. Whereas the Balkan region impacts the central Iran group at 28.7% via Y-chromosomal inputs, no Y-influence is detected in either the IN or IS populations. Imprints from the Levant and southwest Asia are mostly of Y-chromosomal origin, but are seen in the mtDNA of the central Iranian population and in the Gilaki. Central Asian impacts are only detected at the Y-chromosomal level and are absent from IS, whereas influences from Caucasia are observed in all instances except via mtDNA in IN despite its close geographical proximity to the region. No north African effects were detected for any of the Iranian populations using either mtDNA or Y-markers.

Neutrality tests and nucleotide diversity index

Control-sequence variation of both the IN and IS collections conforms to neutral expectations according to the outcome of the Ewens-Watterson test. The nucleotide diversity indices for 336 nucleotides

Figure 2 (a) MDS plot based on observed frequency of mtDNA haplogroup distributions (stress=0.28852). Population designations are followed as presented in Table 1. (b) MDS plot based on observed frequency of Y-chromosome haplogroup distributions (stress=0.12492). Population designations are followed as presented in Supplementary Table 2.

Figure 3 (a) Contour plots based on regional frequency distributions of specific mtDNA haplogroups. Populations included, latitudes and longitudes employed, as well as frequencies are presented in Supplementary Table 3. (b) Contour plots based on regional frequency distributions of specific Y-chromosomal haplogroups. Populations included, latitudes and longitudes employed, as well as frequencies are presented in Supplementary Table 4.

Table 3 Admixture analysis for populations within the Iranian Plateau

Parental group	Hybrid population							
	North Iran		North Iran/Caspian populations ^a		Central Iran		South Iran	
	mtDNA	Y-chromosome	Gilaki	Mazandarian	mtDNA	Y-chromosome	mtDNA	Y-chromosome
North Africa	0.000 ± 0.000	0.000 ± 0.101	0.000 ± 0.224	0.000 ± 0.170	0.000 ± 0.156	0.000 ± 0.117	0.000 ± 0.092	0.000 ± 0.049
Arabian Peninsula	0.921 ± 0.403	0.112 ± 0.827	0.435 ± 0.437	0.643 ± 0.333	0.533 ± 0.304	0.000 ± 0.953	0.521 ± 0.180	0.073 ± 0.402
Levant/Anatolia	0.000 ± 0.715	0.277 ± 1.249	0.256 ± 0.934	0.000 ± 0.712	0.070 ± 0.651	0.000 ± 1.439	0.000 ± 0.386	0.456 ± 0.607
Balkan Peninsula	0.079 ± 0.342	0.000 ± 0.233	0.000 ± 0.371	0.000 ± 0.282	0.000 ± 0.258	0.287 ± 0.269	0.231 ± 0.153	0.000 ± 0.113
Caucasus	0.000 ± 0.485	0.180 ± 0.152	0.275 ± 0.632	0.357 ± 0.482	0.372 ± 0.440	0.435 ± 0.175	0.177 ± 0.261	0.161 ± 0.074
Central Asia	0.000 ± 0.175	0.197 ± 0.376	0.000 ± 0.188	0.000 ± 0.143	0.000 ± 0.131	0.278 ± 0.433	0.000 ± 0.078	0.000 ± 0.183
Southwest Asia	0.000 ± 0.134	0.234 ± 0.365	0.034 ± 0.147	0.000 ± 0.112	0.026 ± 0.102	0.000 ± 0.420	0.071 ± 0.061	0.310 ± 0.177

Abbreviation: mtDNA, mitochondrial DNA.

^aY-chromosomal DNA counterparts were not reported at the resolution needed for these analyses.**Table 4** Selective neutrality test and standard diversity indices of Iran populations

	Iran North (IN)	Iran South (IS)	Iran total (IN+IS)
No. of sequences	28	110	138
No. of haplotypes	28	77	100
Observed F-value	a	0.0190	0.0149
Expected F-value	a	0.0176	0.0133
Ewen-Watterson P-value	a	0.8810	0.9730
Slatkin's exact P-value	a	0.6380	0.8210
Nucleotide diversity	0.0219 ± 0.0117	0.0217 ± 0.0114	0.0213 ± 0.0112
Gene diversity ± s.d.	1.0000 ± 0.0095	0.9900 ± 0.0036	0.9923 ± 0.0027

^aThe test is impossible because all gene copies are different.**Table 5** Time-to-most recent common ancestor

Network	TMRCA (kyr)	Phylogeny
Global H	41.6 ± 5.4	Starlike
H IS	24.0 ± 4.4	Starlike
H IN	25.2 ± 11.2	Non-starlike
Global I	38.9 ± 7.3	Starlike
Global J	43.7 ± 9.9	Starlike
Global J1b	17.8 ± 3.5	Starlike
J1b IS	25.1 ± 8.6	Non-starlike
J1b IN	26.4 ± 10.2	Non-starlike
Global U	39.2 ± 9.1	Non-starlike
Global U7	25.9 ± 5.5	Starlike
U7 IS	23.1 ± 4.6	Non-starlike

Abbreviations: IN, Iran North; IS, Iran South; kyr, kiloyears; TMRCA, time to most recent common ancestor.

of the mtDNA control region (nucleotide positions) is 0.021935 ± 0.011783, 0.021769 ± 0.011434 and 0.021388 ± 0.011224 for IN, IS and Iran total, respectively (Table 4).

Network analyses

Time to most recent common ancestor estimates for H, I, J, J1b, U and U7 Iranian-specific and global networks are provided in Table 5 along with the corresponding topological shapes of each. In addition, network diagrams for each of the aforementioned lineages

are available online in Supplementary Figures 1 to 11. Of these phylogenies, all global networks with the exemption of haplogroup U exhibit star-shaped morphologies, whereas the H network for IN (Supplementary Figure 10) is the only individual population projection to do so.

DISCUSSION

Differences between northern and southern Persians: relationship with the Arabian Peninsula

A comparison of the mtDNA pools of IN and IS (see Supplementary Figure 12 for exact location of collection of each sample) populations reveals contrasting frequencies of haplogroups H, J and U. Although haplogroup J constitutes the majority (35.5%) of the maternal component in the north, it is considerably lower (14.5%) in IS. Haplogroup U accounts for the majority (22.2%) of the mtDNA lineages in the south, a pattern consistent with that presented by Quintana-Murci *et al.*¹⁵ The large percentage of J-derived samples in Iran, specially in the north, contrasts with the more modest frequencies observed by Houshmand *et al.*,⁶⁹ who also reported a greater proportion of J haplotypes for the northern (9.8%) versus the southern (5.9%) regions of Iran. Differences in collection sites may account for the higher frequencies of J in the collections reported in the present study as compared to that by Houshmand *et al.*⁶⁹ The IN and IS also differ with respect to haplogroups T*, T1 and T3 (middle eastern- and lower Arabian Peninsula-specific), and L0 and L1 (characteristic of sub-Saharan Africans). In IS, haplogroups T and L are detected at frequencies of 3.4 and 2.56%, respectively, whereas both lineages are completely absent from the northern sample set. These findings, however, contradict the data published by Quintana-Murci *et al.*,¹⁵ where L lineages are reported for the northern but not southern groups, and haplogroups T* and T1 are observed in both regions of the Plateau but are higher in the north than in the south. These differences could be due to the small sample size of the North Iranian collection. The presence of both haplogroups in the Iranian populations may be indicative of gene flow from the Middle East and Africa.

The admixture analysis results indicate that the majority of Iran's mitochondrial pool is derived from Arabia (Table 3). The Persian groups obtained from previous studies also display high degrees of similarity with the Peninsular Arabs; however, they all exhibit greater contribution from adjacent populations especially with groups from Caucasia. These genetic affinities are also evident in the MDS projection (Figure 2a) in which all the Iranian populations plot between the Arab collections, and the Levant-Anatolia and the

northeast Africa assemblages. The three north Iranian populations partition nearest to the Arab cluster, whereas the central and south Iranian populations segregate closest to the Levant–Anatolia and the north African groups. The genetic affinities between the Arabian Peninsula and Iranian groups may stem from gene flow at various points during the time continuum since the initial out-of-Africa dispersal including: (1) ancient migrations during the initial out-of-Africa exodus in which the Strait of Hormuz is believed to have played a major role (see ref. 1), (2) during Neolithic times as a conduit for pastoral nomads and/or (3) during the Arab expansions of the third to the seventh centuries AD.^{70,71}

Another plausible explanation for the closeness between Persia and Arabia may be the result of dispersals emanating out of central Asia into the Arabian Peninsula via Persia. It has been documented that military incursions during Sassanian times have left Persian communities deep behind in the heart of the Arabian Peninsula (as far south as Yemen).⁷² Similarly, the Balochis under the leadership of Iranian Gedrassians^{73,74} are believed to have traversed the Strait of Hormuz into the Arabian range and are considered responsible for introducing Asian/Indian-specific β -thalassaemia mutations into the Peninsula.^{73,75} These military campaigns are Y-chromosome-driven migrations of minimal mtDNA impact. However, the effect of these migrations is not well understood, and the degree of similarities between the Peninsular Arabs and the Iranians suggests widespread (involving the movement of large numbers of individuals) rather than discreet (a few scattered communities) migratory waves (Figure 2a; Table 3).

In the current study, the MDS plot (Figure 2b) based on Y-chromosomal profiles portrays a close genetic relationship between IN and IS that is reflected in their similar admixture profiles (Table 3). As previously reported by Regueiro *et al.*,¹ the two collections have been thoroughly affected by the migratory waves that have impacted the region with seemingly higher proportions of admixture from the western perimeter than from the east, though over 37% (average of IN and IS) of their Y-chromosomal component is of central or southwest Asian origin (Table 3). It should be noted that the degree of genetic flow from Arabia, as seen in the admixture analysis results, is much lower for the Y-chromosome than it is for the mtDNA (Table 3). It is possible that this is the result of a larger male dissemination from other territories into Persia. This is apparent in the high frequencies of Y-haplogroup R1a1 (M198) of central Asian descent, which is believed to be a tell-tale marker for the expansion of the Kurgan horse culture and Indo–European languages (Figure 3b).¹ It is widely accepted that Iranians are Aryans who migrated from the central Asian steppes around 4000 years before present.¹⁰

Haplogroups J and I: autochthonous to north Persian or the result of genetic drift?

Although the χ^2 analyses performed (Supplementary Table 3) reveal genetic homogeneity between IN and IS for most mtDNA haplogroups, notable differences in the distribution of haplogroups J and J1b, both of which are more abundant in IN, exist (X^2 P -value < 0.005 for both). J and J1b lineages have been associated with the Neolithic diffusion of agriculture and domestication from its proposed geographical origin, the Tigris–Euphrates river valley, westward into Europe.³⁵ The river valley, also known as the Fertile Crescent, was the physical location of ancient Mesopotamia and encompasses present day eastern Iraq, northeastern Syria, southeastern Turkey and western Iran. The J1b sub-haplogroup is abundant in the Mediterranean and southern Atlantic regions.^{35,76} Interestingly, the frequency of this

marker in IN is significantly (with the Bonferroni adjustment for 11 comparisons) higher (Supplementary Table 3) than that of any of the surrounding regions surveyed (panel J in Figure 3a portrays mtDNA haplogroup J as a whole), including those from the Levant (Palestine, Syria, Egypt and Jordan), west central Asia (Armenia), the Near East (Iraq and Iran Kurdish) and the Arabian Peninsula (Oman, UAE, Qatar and Yemen). Although it is tempting to conclude that this distribution pattern suggests a North Iranian origin for this lineage, genetic drift may be responsible. Although IN and IS individuals form part of the ancestral core in the global J1b network (Supplementary Figure 2), most of the remaining Iranian J1b haplotypes are located individually along the branch harboring the 16222 transition. If the J1b source lies within northern Iran, it seems logical to expect more haplotype sharing or, at least, more integration of the IN and reference collections J1b sequences. The significance of the Iranian J1b frequency distribution and lineage pattern is not clear at this point. Denser sampling within and around Iran may provide added insight with respect to the phylogeographic history of J1b within this region.

The asymmetrical partitioning of mtDNA haplogroups J (IN 35.5% and IS 14.5%) and J1b (IN 22.7% and IS 6%) between the two study populations parallels that of the Y-lineage R1b1a-M269, also found at a substantially higher frequency in the northern portion of the Plateau (15 versus 6% for IN and IS, respectively). Furthermore, as was observed with the J and J1b mtDNA haplogroups, this Y-specific marker is substantially more abundant in IN than in most of the surrounding Middle East, Near East and Levantine groups examined, with the exception of Turkey (14.5%).^{1,59} The M269 mutation is observed at elevated levels throughout Europe⁷⁷ and declines in frequency along a southeast trajectory from Europe toward Pakistan (14.5–2.8%). The significance with respect to the enrichment of this European Y-chromosome marker in IN remains unclear. It is not known whether the presence of M269 in north Persia is associated with the northwest Neolithic agricultural movement from the Near East to Europe or if it signals a subsequent back migration eastward from Europe.

The distribution of haplogroup I also differs between the northern (9.7%) and southern (1.7%) regions of Iran. This incongruence is significant at $\alpha=0.05$ ($P<0.03$) but not following the application of the Bonferroni adjustment (Supplementary Table 3). It is noteworthy that, with the exception of its northern neighbor Azerbaijan, IN is the only population in which haplogroup I exhibits polymorphic levels. Also, a contour plot based on the regional phylogeographic distribution of the I haplogroup exhibits frequency clines consistent with an Iranian cradle (panel I in Figure 3a). Moreover, when compared with other populations in the region, those from the Levant (Iraq, Syria and Palestine) and the Arabian Peninsula (Oman and UAE) exhibit significantly lower proportions of I individuals (1–2%; Supplementary Table 3). It should be noted that this haplogroup has been detected in European groups (Krk, a tiny island off the coast of Croatia (11.3%),⁷⁸ and Lemko, an isolate from the Carpathian Highlands (11.3%)⁷⁹) at comparable frequencies to those observed in the North Iranian population. However, the higher frequencies of the haplogroup within Europe are found in geographical isolates and are likely the result of founder effects and/or drift.⁷⁹ In addition, several studies^{5,34,36,80} report the Middle East as the origin of this haplogroup, but for unknown reasons, the prevalence of this lineage in the region has been lost. Thus, it is plausible that the high levels of haplogroup I present in IN may be the result of a localized enrichment through the action of genetic drift or may signal geographical proximity to the location of origin.

Gene flow and topogeographical barriers

Although haplogroup H and its subclades are found in highest frequencies in Europe and Caucasia, the presence of these haplogroups in Iran may reflect gene flow from neighboring southwest Asia where they are present at moderate frequencies. Furthermore, considering the substantial frequency of H2a1 (12.5%) in central and inner Asia, its low frequency in eastern Europe and its absence in western Europe,³⁹ it is likely that its presence in Iran may be due to gene flow from Asia. The fact that sub-haplogroups H2, H2a1, H4 and H7 are seen only in IS (absent in IN), and at relatively low frequencies, may stem from the low number of individuals collected in IN ($n=31$).

mtDNA haplogroup T is common in eastern and northern Europe, and is found as far as the Indus Valley and the Arabian Peninsula.⁵ Thus, the presence of sub-haplogroups T*, T1 and T3 in IS, and their absence in IN, may be associated with gene flow from the Arabian Peninsula to southern Iran.

The best examples of barriers to gene flow are observed in the contour plots of haplogroups J and I and of sub-haplogroups U2 and U7. Both haplogroups J and I are found in high frequencies in northern Iran and exhibit a dwindling cline toward the Levant/Anatolia region, Europe and Asia. A sharp decline is observed beyond the Dasht-e Kavir desert, with some resurgence of haplogroup I in central Asia but no similar presence in southwest Asia, suggesting that this desert could have deterred migrants from traversing from one region to the other. On the other hand, haplogroups U2 and U7 exhibit the opposite demic pattern with high frequencies in southwestern Asia and in the Indus Valley, experiencing a sharp decline/disappearance (it is especially notable for U7) upon arrival at the Zagros mountains/Dasht-e Kavir region. These mtDNA results are mirrored in the distribution of Y-chromosomal haplogroup R, which exhibits a dramatic drop in the Dasht-e Kavir zone (Figure 3b).

The presence of haplogroups/sub-haplogroups J, I, U2 and U7 in the Arabian Peninsula²⁵ again attests to the close genetic affiliation between Persia and Arabia and may suggest gene flow between the two regions. In connection with the putative migratory link between Persia and Arabia, it is noteworthy that high-frequency foci for haplogroups J and I, in the contour plots, are observed in the Arabia Peninsula, again possible signals of gene flow between the two regions (Figure 3a).

ACKNOWLEDGEMENTS

We thank Tanya M Simms and Dr Gabriel E Novick for their useful discussions and suggestions.

- Regueiro, M., Cadenas, A. M., Gayden, T., Underhill, P. A. & Herrera, R. J. Iran: Tricontinental nexus for Y-chromosome driven migration. *Hum. Hered.* **61**, 132–143 (2006).
- Roman, A. & Cleveland, C. J. *Geography of Energy, Energy and Transportation*. http://www.eoearth.org/article/Strait_of_Hormuz (2007).
- Forster, P. Ice ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 255–264 (2004).
- Richards, M., Côte-Real, H., Foster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A. *et al.* Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59**, 185–203 (1996).
- Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V. *et al.* Tracing European founder lineages in the Near Eastern mtDNA Pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).
- Boyle, J. H. *The Mongol World Empire 1206–1370* (Variorum reprints, London, 1977).
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E., Martinez-Arias, R. *et al.* Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. *Am. J. Hum. Genet.* **63**, 1824–1838 (1998).
- Bosworth, C. E. in *History of Civilizations of Central Asia* Vol. IV: the age of achievement: AD 750 to the end of the fifteenth century, part one: the historical, social & economic setting (eds Asimov, M. S. & Bosworth, C. E.). Multiple History Series (Motilal Banarsidass/UNESCO, Paris, 1999).

- Farjadian, S., Mogadam, F. A. & Ghaderi, A. HLA class II gene polymorphism in Parsees and Zoroastrians of Iran. *Int. J. Immunogenet.* **33**, 185–191 (2006).
- Farjadian, S. & Ghaderi, A. HLA class II similarities in Iranian Kurds and Azeris. *Int. J. Immunogenet.* **34**, 457–463 (2007).
- Farjadian, S., Ota, M., Inoko, H. & Ghaderi, A. The genetic relationship among Iranian ethnic groups: an anthropological view based on HLA class II gene polymorphism. *Mol. Biol. Rep.* **36**, 1943–1950 (2009).
- Shepard, E. M. & Herrera, R. J. Iranian STR variation at the fringes of biogeographical demarcation. *Forensic Sci. Int.* **158**, 140–148 (2006).
- Mohyuddin, A. & Mehdi, S. Q. HLA analysis of the Parsi (Zoroastrian) population in Pakistan. *Tissue Antigens* **66**, 691–695 (2005).
- Wells, R. S., Yuldashova, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J. *et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl Acad. Sci. USA* **98**, 10244–10249 (2001).
- Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R. *et al.* Where west meets east: the complex mtDNA landscape of the southwest and central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004).
- Nasidze, I., Quinque, D., Rahmani, M., Alemohamad, S. A. & Stoneking, M. Close genetic relationship between semitic-speaking and Indo-European-speaking groups in Iran. *Ann. Hum. Genet.* **72**, 241–252 (2008).
- Nasidze, I., Quinque, D., Rahmani, M., Alemohamad, S. A., Asadova, P., Zhukova, O. *et al.* mtDNA and Y-chromosome variation in the Talysh of Iran and Azerbaijan. *Am. Phys. Anthropol.* **138**, 82–89 (2009).
- Nasidze, I., Sarkisian, T., Kerimov, A. & Stoneking, M. Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. *Hum. Genet.* **112**, 255–261 (2003).
- Tambets, K., Kivisild, T., Metspalu, E., Parik, J., Kaldma, K., Laos, S. *et al.* in *Archaeogenetics: DNA and the Population Prehistory of Europe* (eds Renfrew, C. & Boyle, K.). 219–235 (Mc Donald Institute Monographs Series, Cambridge, 2000).
- Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J. *et al.* Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann. Hum. Genet.* **67**, 312–328 (2003).
- Cvjetan, S., Tolc, H. V., Lauc, L. B., Colak, I., Dordevic, D., Efremovska, L. *et al.* Frequencies of mtDNA haplogroups in southeastern Europe. *Coll. Antropol.* **28**, 193–198 (2004).
- Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E. *et al.* Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am. J. Hum. Genet.* **75**, 752–770 (2004).
- Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K. *et al.* Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet.* **5**, 1–25 (2004).
- Bosch, E., Calafell, F., Gonzalez-Neira, A., Flaiz, C., Mateu, E., Scheil, H. G. *et al.* Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann. Hum. Genet.* **70**, 459–487 (2006).
- Rowold, D. J., Luis, J. R., Terreros, M. C. & Herrera, R. J. Mitochondrial DNA gene flow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. *J. Hum. Genet.* **52**, 436–447 (2007).
- Terreros, M. C., Rowold, D., Luis, J. R., Khan, F., Agrawal, S. & Herrera, R. J. North Indian Muslims: enclaves of foreign DNA or Hindu converts? *Am. J. Phys. Anthro.* **133**, 1004–1012 (2007).
- Abu-Amero, K. K., Larruga, J. M., Cabrera, V. M. & Gonzalez, A. M. Mitochondrial DNA structure in the Arabian Peninsula. *BMC Evol. Biol.* **8**, 45 (2008).
- Irwin, J. A., Ikramov, A., Saunier, J., Bodner, M., Amory, S., Rock, A. *et al.* The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. *Int. J. Legal Med.* **124**, 195–204 (2010).
- Ausubel, F. M. *Current Protocols in Molecular Biology* (John Wiley & Sons Inc, New York, 1987).
- Antunez de Mayolo, G., Antunez de Mayolo, A., Antunez de Mayolo, P., Papita, S. S., Hammer, M., Yunis, J. J. *et al.* Phylogenetics of worldwide human populations as determined by polymorphic *Alu* insertions. *Electrophoresis* **23**, 3346–3356 (2002).
- Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L. & Erlich, H. A. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence specific oligonucleotide probes. *Am. J. Hum. Genet.* **48**, 370–382 (1991).
- Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
- Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciano, F., Guida, V. *et al.* The emerging tree of West Eurasian mtDNAs: a synthesis of control region sequences and RFLPs. *Am. J. Hum. Genet.* **64**, 232–249 (1999).
- Richards, M. B., Macaulay, V., Bandelt, H. J. & Sykes, B. C. Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.* **62**, 241–260 (1998).
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R. *et al.* Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850 (1996).
- Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R. *et al.* The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian

- glacial refuge was a major source of the European gene pool. *Am. J. Hum. Genet.* **75**, 910–918 (2004).
- 38 Loogväli, E. L., Roostalu, U., Malyarchuk, B. A., Derenko, M. V., Kivisild, T., Metspalu, E. *et al.* Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol. Biol. Evol.* **21**, 2012–2021 (2004).
 - 39 Roostalu, U., Kutuev, I., Loogväli, E. L., Metspalu, E., Tambets, K., Reidla, M. *et al.* Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in west Eurasia: the Near Eastern and Caucasian perspective. *Mol. Biol. Evol.* **24**, 436–448 (2007).
 - 40 Gayden, T., Regueiro, M., Martinez, L., Cadenas, A. M. & Herrera, R. J. Human Y-chromosome haplotyping by allele-specific polymerase chain reaction. *Electrophoresis* **29**, 2419–2423 (2008).
 - 41 Long, J. C., Williams, R. C., McAuley, J. E., Medis, R., Partel, R., Tregellas, M. *et al.* Genetic variation in Arizona Mexican Americans: estimation and interpretation of admixture proportions. *Am. J. Phys. Anthropol.* **84**, 141–157 (1991).
 - 42 Perez-Miranda, A. M., Alfonso-Sanchez, M. A., Peña, J. A. & Herrera, R. J. Qatari DNA variation at a crossroad of human migrations. *Hum. Hered.* **61**, 67–79 (2006).
 - 43 Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E. *et al.* The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* **29**, 1155–1159 (2000).
 - 44 Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L. & Underhill, P. A. Ethiopians and Khoisans share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* **70**, 265–268 (2002).
 - 45 Barac, L., Perici, M., Klaric, I. M., Rootsi, S., Janicijevic, B., Kivisild, T. *et al.* Y chromosomal heritage of Croatian population and its island isolates. *Eur. J. Hum. Genet.* **11**, 535–542 (2003).
 - 46 Al-Zahery, N., Semino, O., Benuzzi, G., Magri, C., Passarino, G., Torroni, A. *et al.* 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 (2003).
 - 47 Luis, J. R., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioli, C., Roseman, C. *et al.* The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migration. *Am. J. Hum. Genet.* **74**, 532–544 (2004).
 - 48 Naside, I., Ling, E. Y. S., Quinque, D., Dupanloup, R., Cordaux, R., Rychkov, O. *et al.* Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Ann. Hum. Genet.* **68**, 205–221 (2004).
 - 49 Flores, C., Maca-Meyer, N., Larruga, J. M., Cabrera, V. M., Karadsheh, N. & Gonzalez, A. M. Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. *J. Hum. Genet.* **50**, 435–441 (2005).
 - 50 Marjanovic, D., Fornarino, S., Montagna, S., Primorac, D., Hadziselimovic, R., Vidovic, S. *et al.* The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups in the three main ethnic groups. *Ann. Hum. Genet.* **69**, 757–763 (2005).
 - 51 Sanchez, J. J., Hallenberg, C., Borsting, C., Hernandez, A. & Morling, N. High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *Eur. J. Hum. Genet.* **13**, 856–866 (2005).
 - 52 Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. T. *et al.* Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* **78**, 202–221 (2006).
 - 53 King, R. J., Ozcan, S. S., Carter, T., Kalfoglou, E., Atasoy, S., Triantaphyllidis, C. *et al.* Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. *Ann. Hum. Genet.* **72**, 205–214 (2008).
 - 54 Abu-Amero, K. K., Hellani, A., Gonzalez, A. M., Larruga, J. M., Cabrera, V. M. & Underhill, P. A. Saudi Arabian Y-chromosome diversity and its relationship with nearby regions. *BMC Genet.* **22**, 10–59 (2009).
 - 55 Noveski, P., Trivodalieva, S., Efremov, G. D. & Plaseska-Karanfilska, D. Y chromosome single nucleotide polymorphisms typing by SNaPshot minisequencing. *BJMG* **12**, 3–10 (2009).
 - 56 Zhao, Z., Khan, F., Borkar, M., Herrera, R. J. & Agrawal, S. Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Ann. Hum. Biol.* **36**, 46–59 (2009).
 - 57 Ferri, G., Tofanelli, S., Alú, M., Taglioli, L., Radheshi, E., Corradini, B. *et al.* Y-Str variation in Albanian populations: implications on the match probabilities and the genetic legacy of the minority claiming an Egyptian descent. *Int. J. Legal Med.* **124**, 363–370 (2010).
 - 58 Cadenas, A. M., Zhivotovsky, L. A., Cavalli-Sforza, L. L., Underhill, P. & Herrera, R. J. Y-chromosome diversity characterizes the Gulf of Oman. *Eur. J. Hum. Genet.* **16**, 374–386 (2008).
 - 59 Cinnioglu, C., King, R., Kivisild, T., Kalfoglou, E., Atasoy, S., Cavalleri, G. L. *et al.* Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* **114**, 127–148 (2004).
 - 60 Zalloua, P. A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Platt, D. E. *et al.* Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am. J. Hum. Genet.* **82**, 873–882 (2008).
 - 61 Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
 - 62 Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
 - 63 Watterson, G. The homozygosity test of neutrality. *Genetics* **88**, 405–417 (1978).
 - 64 Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
 - 65 Forster, P., Harding, R., Torroni, A. & Bandelt, H. J. Origin and evolution of native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59**, 935–945 (1996).
 - 66 Saillard, J., Forster, P., Lynnerup, N., Bandelt, H. L. & Norby, S. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* **67**, 718–726 (2000).
 - 67 Salas, A., Richards, M., De la Fe, T., Lareu, M. V., Sobrino, B., Sanchiz-Diz, P. *et al.* The making of the African mtDNA landscape. *Am. J. Hum. Genet.* **71**, 1082–1111 (2002).
 - 68 Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753 (1995).
 - 69 Houshmand, M., Sanati, M.-H., Vakilian, M., Akuchekian, M., Babrzadeh, F., Teimori, M. *et al.* Investigation of the mitochondrial haplogroups M, BM, N, J, K and their frequencies in five regions in Iran. *Iranian J. Biotechnol.* **2**, 44–48 (2004).
 - 70 Dussaud, R. *La penetration des Arabes en Syrie avant l'Islam* (P. Geuthner, Paris, 1955).
 - 71 Ricci, L. in *L'Arabie du Sud* Vol. 1 (ed. Chelhod, J.) 239–248 (Maisonneuve et Larose, Paris, 1984).
 - 72 Cerny, V., Mulligan, C. J., Ridl, J., Zaloudkova, M., Edens, C. M., Hajek, M. *et al.* Regional differences in the distribution of the Sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. *Am. J. Phys. Anthropol.* **136**, 128–137 (2008).
 - 73 Quaipe, R., al-Gazali, L., Abbas, S., Fitzgerald, P., Fitches, A., Valler, D. *et al.* The spectrum of beta thalassaemia mutations in the UAE national population. *J. Med. Genet.* **31**, 59–61 (1994).
 - 74 Dawson, K. P. & Frossard, P. M. A hypothesis regarding the origin and spread of the cystic fibrosis mutation deltaF508. *QJM* **93**, 313–315 (2000).
 - 75 Zahed, L. The spectrum of beta-thalassemia mutations in the Arab populations. *J. Biomed. Biotechnol.* **1**, 129–132 (2001).
 - 76 Côte-Real, H. B., Macaulay, V., Richards, M., Hariti, G., Issad, M. S., Cambon-Thomsen, A. *et al.* Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.* **60**, 331–350 (1996).
 - 77 Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A. *et al.* 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 (2002).
 - 78 Pericic, M., Lauc, L. B., Klaric, I. M., Janicijevic, B. & Rudan, P. Review of Croatian genetic heritage as revealed by mitochondrial DNA and Y chromosomal lineages. *Croat. Med. J.* **46**, 502–513 (2005).
 - 79 Nikitin, A. G., Kochkin, I. T., June, C. M., Willis, C. M., McBain, I. & Videiko, M. Y. Mitochondrial DNA sequence variation in the Boyko, Hutsul, and Lemko populations of the Carpathian Highlands. *Hum. Biol.* **81**, 43–58 (2009).
 - 80 Richards, M., Rengo, C., Cruciani, F., Gratix, F., Wilson, J. F., Scozzari, R. *et al.* Extensive female lineage-mediated gene flow from sub-Saharan into Near Eastern Arab populations. *Am. J. Hum. Genet.* **72**, 1058–1064 (2003).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)