Hum Genet (2017) 136:437–450
DOI 10.1007/s00439-017-1770-2

ORIGINAL INVESTIGATION

Genetic differentiation between upland and lowland populations shapes the Y-chromosomal landscape of West Asia

O. Balanovsky[1,2] · M. Chukhryaeva[1,2] · V. Zaporozhchenko[1,2] · V. Urasin[3] · M. Zhabagin[4] · A. Hovhannisyan[5] · A. Agdzhoyan[1,2] · K. Dibirova[1,2] · M. Kuznetsova[2] · S. Koshel[6] · E. Pocheshkhova[7] · I. Alborova[8] · R. Skhalyakho[1,2] · O. Utevska[9] · The Genographic Consortium · Kh. Mustafin[8] · L. Yepiskoposyan[5] · C. Tyler-Smith[10] · E. Balanovska[1,2]

Received: 24 May 2016 / Accepted: 20 February 2017 / Published online: 9 March 2017

Abstract Y-chromosomal variation in West Asian populations has so far been studied in less detail than in the neighboring Europe. Here, we analyzed 598 Y-chromosomes from two West Asian subregions—Transcaucasia and the Armenian plateau—using 40 Y-SNPs and 17 Y-STRs and combined them with previously published data from the region. The West Asian populations fell into two clusters: upland populations from the Anatolian, Armenian and Iranian plateaus, and lowland populations from the Levant, Mesopotamia and the Arabian Peninsula. This geographic subdivision corresponds with the linguistic difference between Indo-European and Turkic speakers, on the one hand, and Semitic speakers, on the other. This subdivision could be traced back to the Neolithic epoch, when upland populations from the Anatolian and Iranian plateaus carried similar haplogroup spectra but did not overlap with lowland populations from the Levant. We also found that the initial gene pool of the Armenian motherland population has been well preserved in most groups of the Armenian Diaspora. In view of the contribution of West Asians to the autosomal gene pool of the steppe Yamnaya archaeological culture, we sequenced a large portion of the Y-chromosome in haplogroup R1b samples from present-day East European steppe populations. The ancient Yamnaya samples are located on the "eastern" R-GG400 branch of haplogroup R1b-L23, showing that the paternal descendants of the Yamnaya still live in the Pontic steppe and that the ancient Yamnaya population was not an important source of paternal lineages in present-day West Europeans.

Introduction

West Asia, located between Africa, Asia and Europe, was a key region for human migrations, including the origin and spread of farming. However, populations from many areas within West Asia are underrepresented in studies of Y-chromosomal variation, particularly compared with the

O. Balanovsky and M. Chukhryaeva have equally contributed to this article.

Affiliations for The Genographic Consortium are moved to Acknowledgements.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-017-1770-2) contains supplementary material, which is available to authorized users.

✉ O. Balanovsky
balanovsky@inbox.ru

1 Vavilov Institute of General Genetics, Moscow, Russia

2 Research Centre for Medical Genetics, Moscow, Russia

3 YFull Service, Moscow, Russia

4 National Laboratory Astana, Nazarbayev University, Astana, Republic of Kazakhstan

5 Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia

6 Faculty of Geography, Lomonosov Moscow State University, Moscow, Russia

7 Krasnodar State Medical University, Krasnodar, Russia

8 Moscow Institute of Physics and Technology (State University), Moscow, Russia

9 V. N. Karazin Kharkiv National University, Kharkiv, Ukraine

10 The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK

well-studied neighboring Europe. These undersampled areas include, among others, Transcaucasia and the Armenian highlands. In this study, "West Asia" means the southwestern part of Eurasia and includes Anatolia, the Levant, Mesopotamia, the Arabian Peninsula, the Armenian and Iranian plateaus, Transcaucasia, and North Caucasus. The term "Transcaucasia", or South Caucasus, traditionally refers to Georgia, Azerbaijan, and sometimes to present-day Armenia, though in this study we consider the Armenian plateau as a separate region within West Asia.

In Transcaucasia, ancient DNA study has suggested genetic continuity since the Mesolithic (Jones et al. 2015). In the late Neolithic, the Kartvelian language family originated in situ (Ruhlen 1987) and split into a set of languages and dialects, which are now spoken by many ethnic and subethnic groups in Georgia. Y-chromosomal variation in these groups has been studied using a restricted set of markers in early papers (Semino et al. 2000; Rosser et al. 2000; Weale et al. 2001; Nasidze et al. 2003), by a wider set of haplogroups later (Battaglia et al. 2009; Yunusbayev et al. 2012), by Y-STR markers (Tarkhnishvili et al. 2014), and haplogroup R1 variation was addressed in (Myres et al. 2011; Underhill et al. 2010). However, none of these studies was especially dedicated to Georgians or Transcaucasia.

The next undersampled area, the Armenian highland, is the middle link in the chain of West Asian highlands: the Asia Minor, Armenian, and Iranian uplands (Lang 1970). The Armenian highlands were populated by multiple groups, but starting from middle of the first millennia BC, historical records indicate that Armenians predominated in this area. The Armenian language forms its own old branch of the Indo-European linguistic family, and there are two major theories explaining the appearance of the Indo-Europeans in this area. The first suggests migration from the Eastern Mediterranean and Balkans (Devoto 1962) while the second locates the Indo-European homeland itself very close to the Armenian highland (Gamkrelidze and Ivanov 1995).

The historical area of the Armenians was much larger than the current boundaries of the Armenian state, and included parts of present-day Turkey, Georgia, Azerbaijan, Iraq, and Iran (Hovannisian 1997). But starting from the 11th century AD, the expansion of the Turks led to intensive immigration into the region. Subsequently, many Armenian populations migrated in all directions to neighboring areas. As a result of this complex population history, most Armenians now live in Diasporas outside the initial area, while a less numerous part of the population remained in the area of the present-day Armenian state. Because of this, to study the genetic composition of the Armenian highland, one needs to sample both the present-day population of Armenia and the different groups of the Armenian Diaspora. The majority of these populations remember their geographic place of origin within the historical Armenia, and practice endogamy. Y-chromosomal variation in Armenian populations has been studied previously (Weale et al. 2001; Herrera et al. 2012; Hovhannisyan et al. 2014). Most of these studies, however, concentrated on populations from present-day Armenia, Turkey, and Levantine Diasporas, while Diasporas in North Caucasus, Transcaucasia and the Black Sea north coast have never been studied genetically. In addition to the Y-chromosomal pool, the autosomal variation in Armenians has been investigated using genome-wide arrays (Yunusbay ev et al. 2012; Haber et al. 2016) and even ancient DNA data from Armenian plateau populations were published recently (Allentoft et al. 2015).

The most striking finding from these ancient DNA studies was the suggested chain of Bronze Age migrations: from the Armenian plateau (Allentoft et al. 2015) or Transcaucasia (Jones et al. 2015) to the Yamnaya culture in the Pontic steppe, and from there to Central Europe (Haak et al. 2015; Allentoft et al. 2015). The Y-chromosomal pool of the Yamnaya population consisted almost exclusively of haplogroup R1b (Haak et al. 2015; Allentoft et al. 2015), though phylogenetic details were not explored in these studies. Such information could be retrieved by the complete sequencing of the Y-chromosome. This is a powerful approach to refine phylogenies and unravel population events (Rootsi et al. 2013; Karmin et al. 2015; Balanovsky et al. 2015; Bergström et al. 2016), and it might shed light on the aforementioned migration chain from the perspective of paternal lineages.

The central question of our study is the broader structuring of the West Asian paternal pool. It did not receive much attention in previous works, as West Asian populations were typically considered from a European perspective as a source for migration into Europe in the Neolithic (Cavalli-Sforza et al. 1988; Quintana-Murci et al. 2003; Haber et al. 2016) and Bronze Age (Allentoft et al. 2015), or as a recipient of migrations from Europe (Zalloua et al. 2008b). But the genetic pattern within West Asia itself was rarely considered as a subject of research in its own right, though recent autosomal data on ancient DNA have revealed the extreme variation between different areas of West Asia in pre-Neolithic and Neolithic times (Lazaridis et al. 2016; Broushaki et al. 2016). The Y-chromosomal patterns in present-day populations were so far reported mostly for one particular region within West Asia—the North Caucasus. Yunusbayev et al. (2012) revealed west-to-east differences within the North Caucasus. Balanovsky et al. (2011) reported the parallel evolution of Y-chromosomal pools and languages in the North Caucasus in general, while Karafet et al. (2016) demonstrated the same pattern for the Dagestani populations in particular and stressed reduced intra-population diversity for the upland Dagestan populations,

while lowland Dagestani groups appeared to be more susceptible to gene flow. Beyond the Caucasus, genetic differences were demonstrated between coastal and inland areas within the Levant (El-Sibai et al. 2009).

Here, we present data on the Y-chromosomal variation in seven West Asian populations and compiled the most up-to-date dataset of Y-chromosomal variation in the region. We analyzed geographic patterns within the gene pool of the West Asian metapopulation. We also compared Armenian Diaspora groups with the gene pool in the homeland Armenia. In view of the possible contribution of the Armenian plateau populations to the gene pool of the steppe Yamnaya culture and subsequent migration to West Europe, we sequenced a large portion of the Y-chromosome in haplogroup R1b samples and placed the ancient Yamnaya samples onto the emerging phylogenetic tree.

Materials and methods

Populations studied

We studied three Georgian and four Armenian populations with a total sample size 598 individuals. Supplementary Fig. 1 presents the locations and historical overview of these populations. A total of 446 Armenian and 152 Georgian saliva or blood samples were collected (Table S1). Sampled individuals in all the studied populations except the Don Armenians identified all four grandparents as members of the given ethnic group, and were unrelated at least up to the third degree. For the sampled Don Armenians, their grandfathers on the paternal line were from this population, while other grandparents in some cases were non-Armenians. DNA from both blood and saliva was extracted using an organic extraction method (Powell and Gannon, 2002).

Compliance with ethical standards

All sample donors gave written informed consent and the study was approved by the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia.

Genotyping

The following 40 Y-SNP markers were hierarchically examined in 598 samples: C-M130, C-M217, E-M35, E-M78, E-V13, G-M201, G-M285, G-M406, G-P15, G-P16, G-P303, H-M89, IJ-M429, I-M170, I-M223, I-M253, I-P215, I-P37, J1-M267, J1-P58, J2-M12, J2-M172, J2-M47, J2-M67, J2-M92, J-M304, K-M9, L-M317, O-M122, P1-M74, Q-M242, R1a-M198, R1a-M458, R-L23, R1b-M269, R1b-M343, R1b-Z2103, R2-M124, R-M207, T-M70. All markers were genotyped by real-time PCR using custom TaqMan assays (Applied Biosystems). All samples were additionally genotyped at 17 Y-STR loci using the Yfiler™ PCR Amplification Kit (Applied Biosystems). The amplified fragments were analyzed with the Genemapper v3.2 program. For all analyses, DYS389I and [DYS389II-DYS398I] loci were used, and the DYS385 loci were omitted.

Statistical and phylogenetic analyses

Nei's genetic distances between populations were calculated using the DJ software (Balanovsky et al. 2008) and visualized using MDS and tree diagrams constructed with Statistica 6.0 and 10.0 (StatSoft Inc. 2001). Data on 21 haplogroups were used for MDS analysis of the Armenian populations (C-M217, E-M35, E-M78, E-M123, G-M201, G-P15, I-M170, J-M267, J-L136, J-M172(xM67, M12), J-M67, J-M92, J-M12, L-M20, N-M231, O-M175, Q-M242, R-M198, R-M269, R-M124, T-M184), and a slightly different set of 21 haplogroups (C-M130, E-M35.1, G1-M285, G2-P287, H-M69, I*-M170, I-P37.2, I-M253, J-M267, J*-M172, J-M67, J-M92, L-M11, N-M231, O-M175, Q-M242, R-M448, R-M73, R-M269, R-L261, T-L206) was used for MDS analysis in the West Eurasian context. Data on 12 haplogroups (C-M130, E-M35, G-L116, H-M69, I-M170, J-M267, J-M172, L-M11, NO-M214, Q-M242, R-M173, T-L206) were used for MDS analysis of the West Asian populations.

Cartographic analysis was performed in the GeneGeo software using algorithms described previously (Balanovsky et al. 2011; Koshel 2012). When constructing haplogroup frequency distribution maps, the weight function was set to 3 and radius of influence to 10,000 km. The cartographic approach provides interpolated frequencies for areas where direct data on the given marker are missing, which allows genetic distance maps to be created. (Balanovskaia et al. 1999). We calculated the classical Nei's statistic of genetic distances (Nei 1975) between the haplogroup frequencies in the reference population and the haplogroup frequencies in the each node of the map grid, using the set of 12 haplogroups listed above.

Calculation of genetic diversity was conducted in Arlequin 3.5.1 using SNP haplogroup frequencies.

An Analysis of Molecular Variance (AMOVA) was also performed in Arlequin 3.5.1. (Schneider et al. 2000). Y-STR networks were constructed in Network 4.6 (Fluxus Engineering, http://www.fluxus.engineering.com) using the reduced median algorithm, with the reduction threshold set to 1, and visualized with Network Publisher (Fluxus Engineering, Clare, UK).

Expansion times for the haplotype clusters were calculated by the rho-estimator (Forster et al. 1996; Saillard et al. 2000) applying the genealogical mutation rate (Gusmao et al. 2005; Sanchez-Diz et al. 2008; Ge et al. 2009, Ravid-Amir and Rosset 2010; Goedbloed et al. 2009; Balanovsky et al. 2015) and a generation time of 30 years (Fenner 2005).

Deep phylogenetic analysis of the Y-chromosome

We sequenced 11 Y-chromosomes belonging to haplogroup R-L23. Five East European samples were obtained during field trips to indigenous populations, while the rest of the samples were from customers of the Yfull service (www.yfull.com). Indigenous sample donors gave informed consent as described above, and YFull customers consented for using their data in study of R1b phylogeography during email conversations with Yfull administrators. For sequencing and data analysis, we used the approach described previously (Balanovsky et al. 2015). Briefly, we used the commercially available "BigY" product (Gene by Gene, Ltd) capturing 11,383,697 bp of the so-called "Gold Standard regions" of the Y-chromosome to generate sequence and then obtain VCF files. The phylogenetic tree was constructed using the Phylomurka software (http://phylomurka.sourceforge.net) using SNPs with call rate above 90%. Seven ancient genomes from the Yamnaya culture representing haplogroup R1b (data published by Haak et al. 2015 and updated in Mathieson et al. 2015) and one ancient genome from Iron Age Iran (Broushaki et al. 2016) were checked for all SNPs which appeared on our resulting tree. The phylogenetic position of each ancient Y-chromosome was estimated to the degree possible given the sequencing coverage.

Results

The genetic structure of the West Asian paternal pool

We genotyped 40 Y-SNP and 17 Y-STR markers in 598 individuals from 4 Armenian and 3 Georgian populations (Fig. 1, Table S2) and compiled a dataset of Y-chromosomal variation in West Asia including data from both this study and 15 previous publications (Abu-Amero et al. 2009; Balanovsky et al. 2011; Cadenas et al. 2008; Cinnioglu et al. 2004; Di Cristofaro et al. 2013; El-Sibai et al. 2009; Flores et al. 2005; Karafet et al. 2016; Grugni et al. 2012; Haber et al. 2011; Herrera et al. 2012; Hovhannisyan et al. 2014; Sanchez et al. 2005; Zalloua et al. 2008a, b). The dataset (Table S3) included 6064 Y-chromosomes from 60 West Asian populations. Though some papers presented high-resolution data, with the number of haplogroups identified ranging from 26 to 84 (Di Cristofaro et al. 2013; Herrera et al. 2012; Grugni et al. 2012; this study), the combined phylogenetic analysis inevitably decreased it down to 12 haplogroups and 44 populations (Table S3) when data from the different studies were pooled.

The MDS plot based of this dataset (Fig. 2) revealed two clusters. The first included Jordanians, Lebanese, Mesopotamians (Iraqis), Palestinians, Syrians, and populations from the Arabian Peninsula. The second included Armenians, Azeri, Kurds, Iranians, and Turks. The latter cluster thus unites populations from the highlands, while the former is made up mostly of populations from the lowlands. Though some mountains are also present, for example in Syria, for this analysis we considered only the principal mountain systems of West Asia: the mountains of Asia Minor, Iran, and the Armenian highlands.

To investigate whether or not the highland/lowland contrast was indeed the main predictor for clustering, we ran

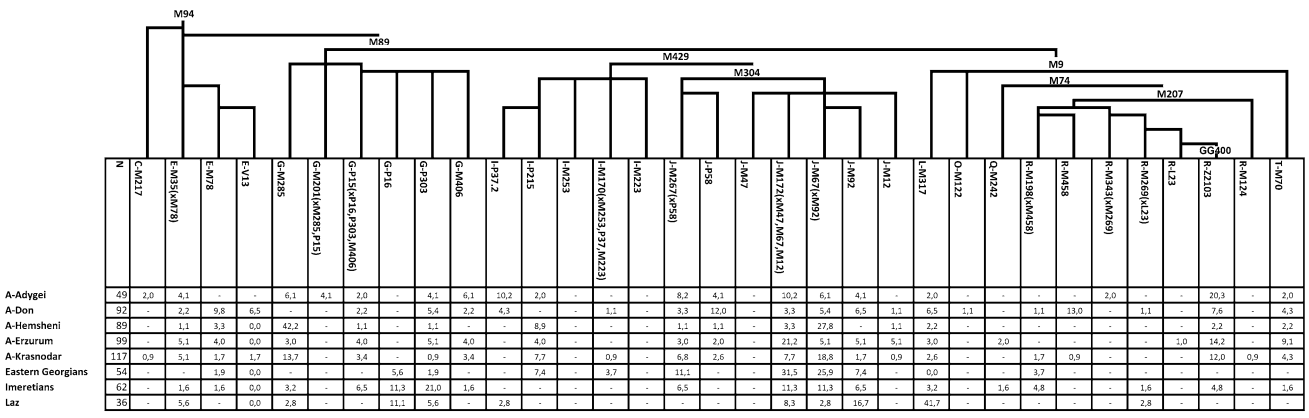


Fig. 1 Frequencies of Y-chromosomal haplogroups (percent) in the Armenian and Georgian populations studied. Population names follow Table S1

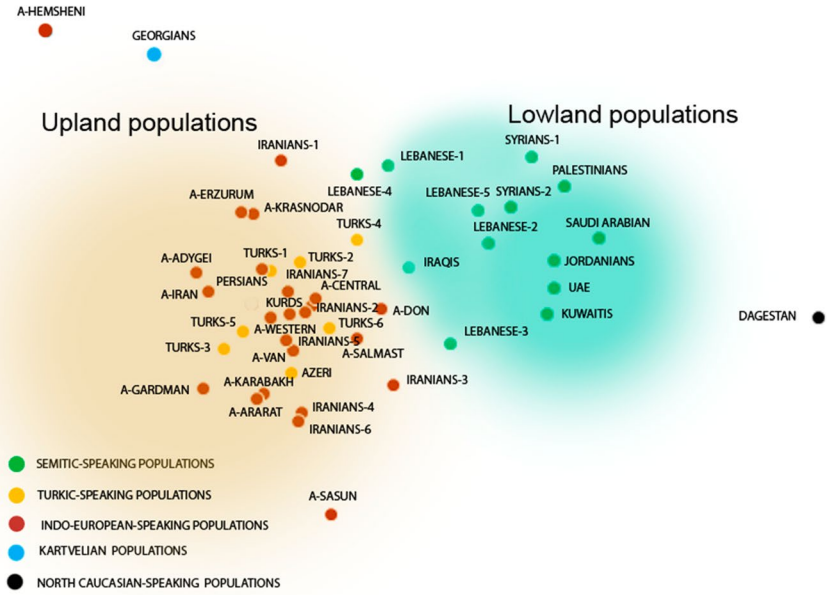


Fig. 2 Multidimensional scaling plot of West Asian populations

Table 1 Genetic variation within and between groups of West Asian populations (AMOVA)

Grouping model	Among groups	Within groups	Within populations
Mountain\plain populations	5.1*	3.0*	91.9*
Western\eastern populations	0.55*	5.7*	93.8*
Populations from different linguistic families (Indo-Europeans. Semitic. Turkic)	4.2*	3.1*	92.7*

Numbers are percentages

* p value <0.001

AMOVA. When the West Asian populations were grouped into highland-dwellers and lowland-dwellers (Table S2), the highest genetic differentiation, 5.17%, was found (Table 1). When the same set of populations was divided into western and eastern (using the 43.5° meridian to subdivide them into nearly equal areas), the level of differentiation was ten times lower (0.56%). When the populations were grouped according to the language families (Indo-European, Semitic, and Turkic), the variation (4.2%) was higher than the west *vs* east grouping, but lower than the highland vs lowland groups.

While MDS analysis (Fig. 2) visualizes the general pattern of interrelations between populations, plotting genetic distances on a geographic map can visualize differences between a single reference population and all the others. We therefore created a set of 12 maps showing genetic distances from 12 principal reference populations (Fig. 3). As expected, an area of high genetic similarity to the reference population (shown in green) on every map was centered on the geographic location of the chosen population, but these areas vary considerably in size on different maps. The smallest area of genetic similarity is seen on the map of the Georgian population (Fig. 3a). This finding agrees with the small area of the Kartvelian linguistic family and the lack of relationship between it and surrounding languages. A large area of genetic similarity is seen on the map of the Turks (Fig. 3b). The area partly corresponds with the territory of the Ottoman Empire from which Anatolian Turks might receive gene flow. The map of averaged Armenian populations demonstrates that the area of highest genetic similarity fits the territory of ancient Armenia (Fig. 3e). This might indicate that the gene pool has been preserved in this area, though much of it is currently populated by Turks: the Anatolian substratum assimilated by Turkish populations was likely represented by western Armenians. Overall, each map followed one of the two patterns. Maps of the first pattern (Fig. 3a–e) exhibited areas of genetic similarity on the Iranian, Anatolian, and Armenian uplands, while the Arabian Peninsula and Levant were marked as genetically distant. Maps of the second pattern (Fig. 3g–l), in contrast, exhibited areas of similarity in the Arabian Peninsula while the uplands were genetically distant. Maps of the upland populations (Turks, Armenians, Iranians, Kurds, and Georgians) followed the first pattern. Maps of the lowland populations (Saudi Arabs, UAE Arabs, Syrians, Jordanians, Lebanese

and Palestinians) followed the second pattern. The map of the Iraqis (Fig. 3f) is intermediate, though closer to the lowland pattern. We note that Syrians, being mainly a lowland group (Fig. 3h), are genetically much closer to geographically distant lowland populations like Arabs than to their immediate geographical neighbors from the uplands.

The male-line genetic structure of populations from the Armenian plateau and Transcaucasia

Our newly generated data (Fig. 1, Table S2) reveal new patterns of genetic structure in the Armenian plateau and Transcaucasia.

The genetic diversity within all Armenian and Georgian populations was much higher than in the North Caucasus (Table S4). However, high-frequency haplogroups could still be identified, including G-M285 for the Hemsheni Armenians, R-L23 and J-M67 for other Armenian groups, while the South Asian haplogroup L-M317 is surprisingly frequent (42%) in the Laz population. The phylogenetic analyses of these haplogroups are summarized in Supplementary Fig. 2 (R-L23), Supplementary Fig. 3 (J-M67), and Supplementary Fig. 4 (L-M317).

Supplementary Fig. 5 shows a comparison of the Armenian and Transcaucasian paternal pools with other West Eurasian populations. Populations from Georgia group together with the West/Central Caucasus, while the Azeri (another Transcaucasian population) and the Armenians belong to the Middle Eastern cluster. Note that, many Armenian populations group together, forming a subcluster of their own within the Middle Eastern cluster.

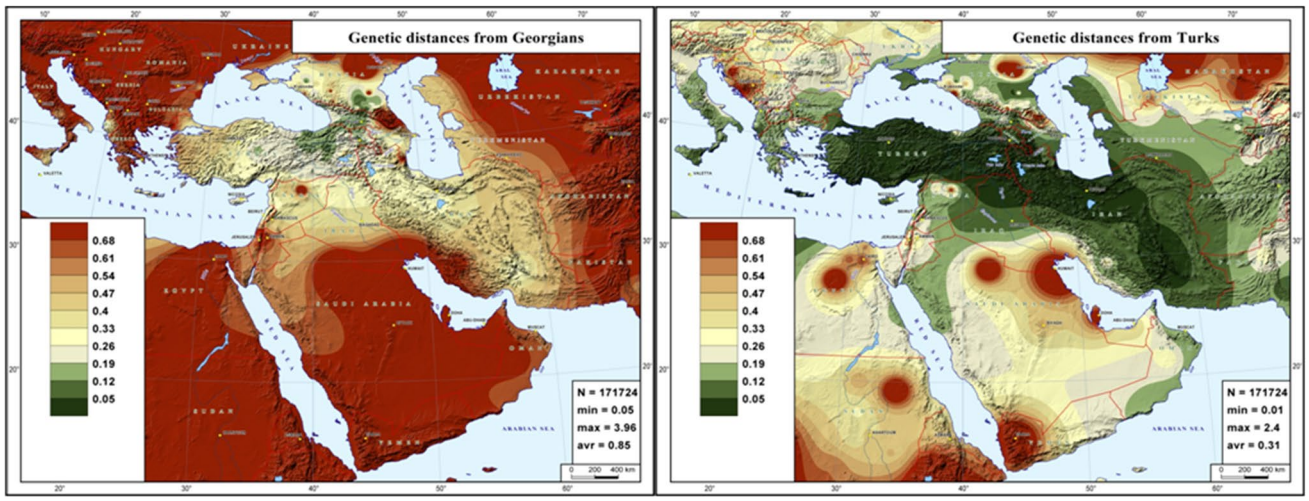
Supplementary Fig. 6 focuses on variation within Armenians, including both our new results and published data (Table S5). Most Armenian populations are genetically close to each other while the Hemsheni Armenian population—speaking a specific dialect (Simonian 2007)—and to a lesser degree Krasnodar, Sasun, and Don Armenians are genetic outliers. It is worth noting that the main cluster includes both Diaspora populations and populations from the Armenian highland; similarly, outlier populations comprise both Diaspora and motherland groups. AMOVA results confirmed that the difference between the gene pools of motherland and Diaspora Armenians is small (Table 2).

Location of ancient Yamnaya samples on an updated haplogroup R1b phylogenetic tree

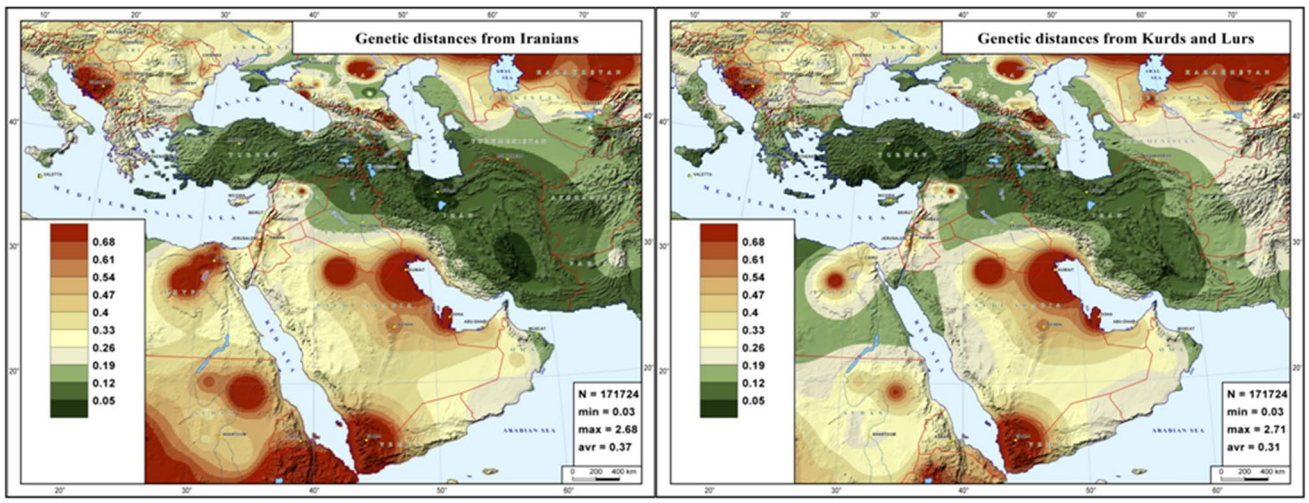
Haplogroup R1b is found at high frequencies in many European populations, and available phylogenetic resources are based on full sequences of hundreds of Y-chromosomes from West Europeans (www.isogg.org, www.yfull.com). However, R1b variation in Eastern Europe is poorly studied. We focused on this area and sequenced 11 Mb of five Y-chromosomes from different East European populations, balanced them with an equal number of West European samples, and revealed 452 polymorphic SNPs with call rates above 90% (Table S6). The resulting phylogenetic tree (Fig. 4a) demonstrates that haplogroup R-L23 splits into two main branches, R-L51 and R-GG400. The former includes West Europeans, while the latter comprises exclusively representatives of East European populations. Both branches are of similar age: around 6 thousand years (Fig. 4a). Note that members of this eastern branch R-GG400 came mainly from the steppe area of East Europe.

We then placed seven ancient Yamnaya genomes (Haak et al. 2015; Mathieson et al. 2015) on this phylogenetic tree. These genomes belong to haplogroup R1b, as they are derived for M269 or its phylogenetic equivalents PF6434 and PF6431: I0231 (M269+, L23+; "+" indicates the derived allele), I0370 (M269+), I0429 (M269+, L23+), I0438 (PF6434+, L23+), I0439 (PF6434+), I0443 (M269+, L23+), I0444 (PF6431+). To place these ancient genomes more precisely on the R1b tree, we examined the status of each SNP present on the tree in each ancient genome. To retain the maximum information, we did not apply any filters apart from phred >30, and used even SNPs called from a single read. The number of reads is shown in Fig. 4b, so one can assess the reliability of each genotype. Since coverage of the ancient samples was low and their sequencing was targeted to SNPs extracted from an old version of the ISOGG database (ISOGG version 8.22 as of April 22, 2013), only eight of our SNPs of interest were genotyped in the ancient samples (Fig. 4). However, this dataset was enough to show that five out of seven Yamnaya genomes do not belong to the West European branch (they have the ancestral state of L51). The remaining two ancient samples were not successfully genotyped by this marker, but they are ancestral for markers of major sub-branches (P310 and P312). One may conclude that all the Yamnaya genomes analyzed here do not belong to the western branch. The alternative possibility that they belong to the eastern branch is directly supported for at least five samples. Namely, I0231 is GG400 derived, Z2103 derived and GG625 derived; I0370 is Z2103 derived; I0429 is GG625 derived and Z2105 derived (this Z2105 SNP was not typed in our modern samples but

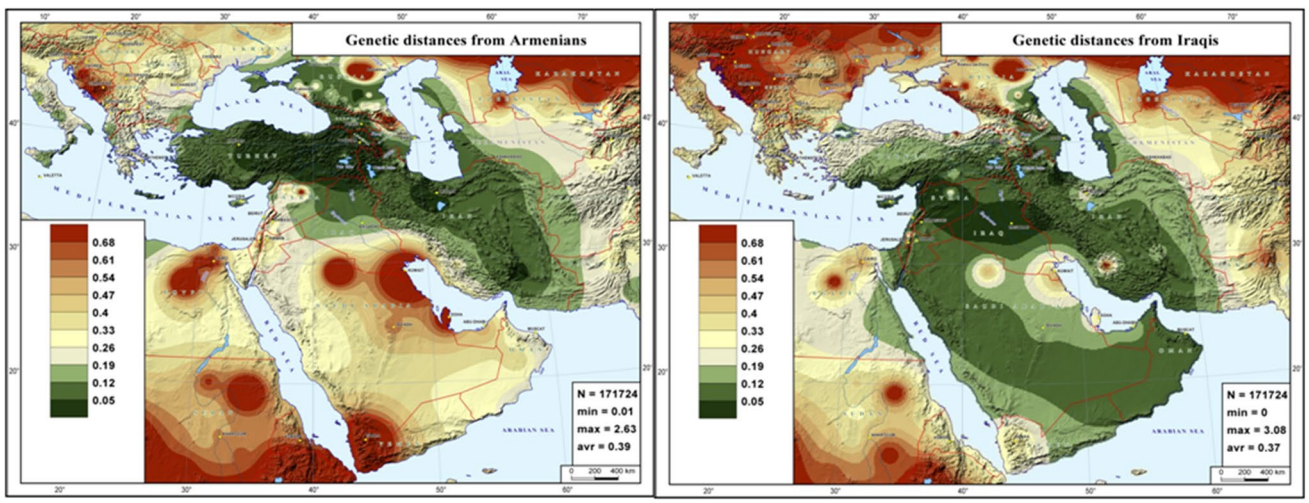
Fig. 3 Maps of genetic distances from reference West Asian populations. **a** Map of genetic distances from Georgians. **b** Map of genetic distances from Turks. **c** Map of genetic distances from Iranians. **d** Map of genetic distances from Kurds and Lurs. **e** Map of genetic distances from Armenians. **f** Map of genetic distances from Iraqis. **g** Map of genetic distances from Lebanese. **h** Map of genetic distances from Syrians. **i** Map of genetic distances from Saudi Arabs. **j** Map of genetic distances from UAE Arabs. **k** Map of genetic distances from Palestinians. **l** Map of genetic distances from Jordanians


Genetic distances from Georgians
0.68
0.61
0.54
0.47
0.4
0.33
0.26
0.19
0.12
0.05
N = 171724
min = 0.05
max = 3.96
avr = 0.85
Genetic distances from Turks
N = 171724
min = 0.01
max = 2.4
avr = 0.31

A B


Genetic distances from Iranians
0.68
0.61
0.54
0.47
0.4
0.33
0.26
0.19
0.12
0.05
N = 171724
min = 0.03
max = 2.68
avr = 0.37
Genetic distances from Kurds and Lurs
N = 171724
min = 0.03
max = 2.71
avr = 0.31

C D


Genetic distances from Armenians
0.68
0.61
0.54
0.47
0.4
0.33
0.26
0.19
0.12
0.05
N = 171724
min = 0.01
max = 2.63
avr = 0.39
Genetic distances from Iraqis
N = 171724
min = 0
max = 3.08
avr = 0.37

E F

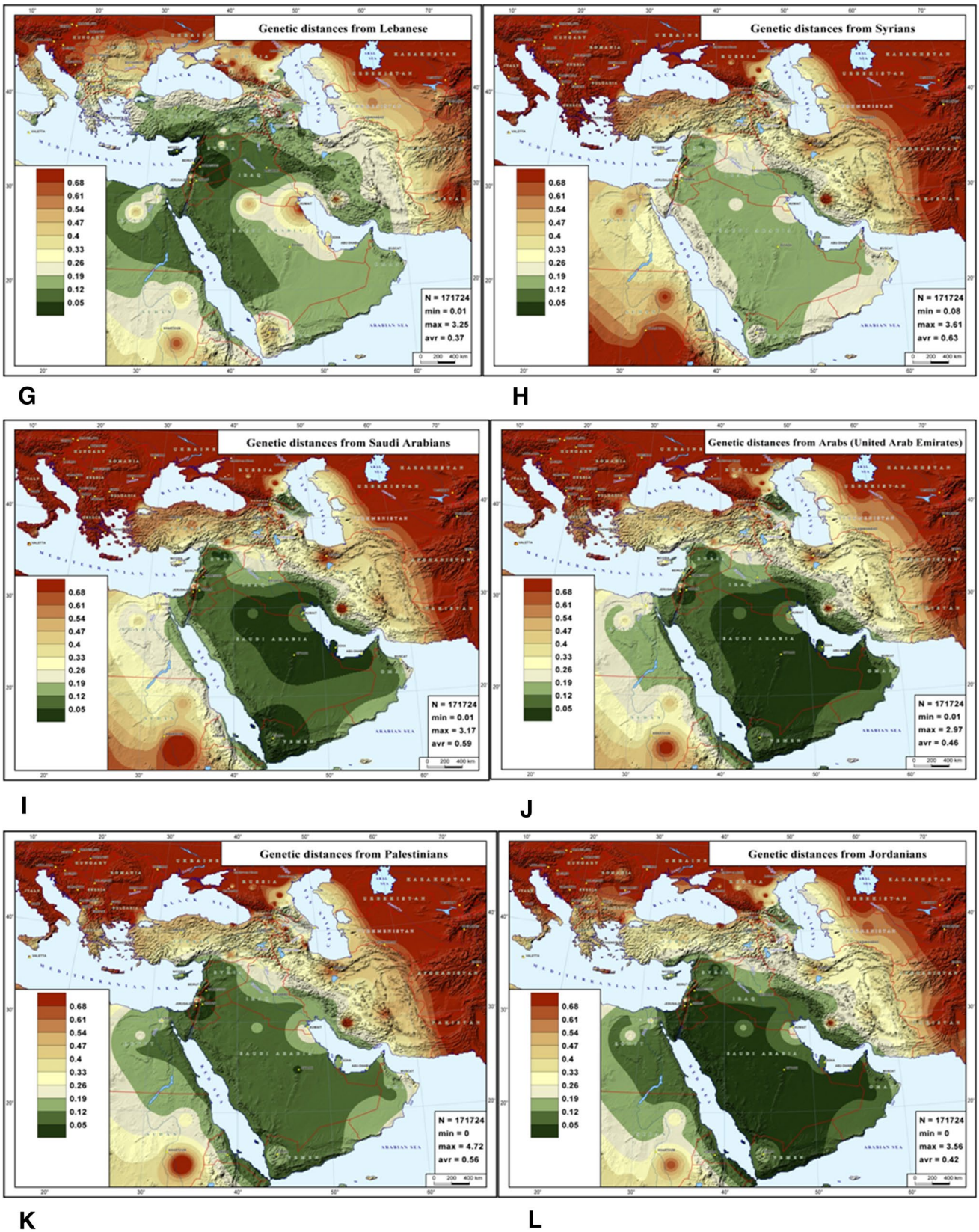


Fig. 3 continued

Table 2 The genetic significance of different groupings of Armenian populations (AMOVA)

Percent of variation	Among groups	Within groups	Within populations
Hemsheni\not Hemsheni	11.8*	2.2	86.0
Diaspora\Motherland	2.6*	2.7	94.7
Random grouping in two sets	−0.3	4.2	96.1

* p value <0.001

is known to be equivalent to Z2103); I0438 is also Z2105 derived; I0444 is Z2103 derived. Among the two remaining samples, the phylogenetic position of I0439 remained unresolved, while I0443 does not belong to either the western or the eastern branch, and might represent a third branch within R-L23 which is now rare or extinct. Three Yamnaya samples belong not only to the eastern branch in general, but to a specific sub-branch identified in a present-day sample from the Crimean Tatar population: I0231 and I0429 are GG625 derived, and I0444 is 17146508 derived (this SNP was identified in the our Crimean Tatar sample but not shown on the tree as the call rate was less than 90%). To summarize, the ancient Yamnaya genomes published in (Haak et al. 2015; Mathieson et al. 2015) do not belong to the main Western European branch R-L51, but most do belong to the eastern branch R-GG400, which we identified in present-day East Europeans.

Discussion

The Y-chromosomal landscape of West Asia

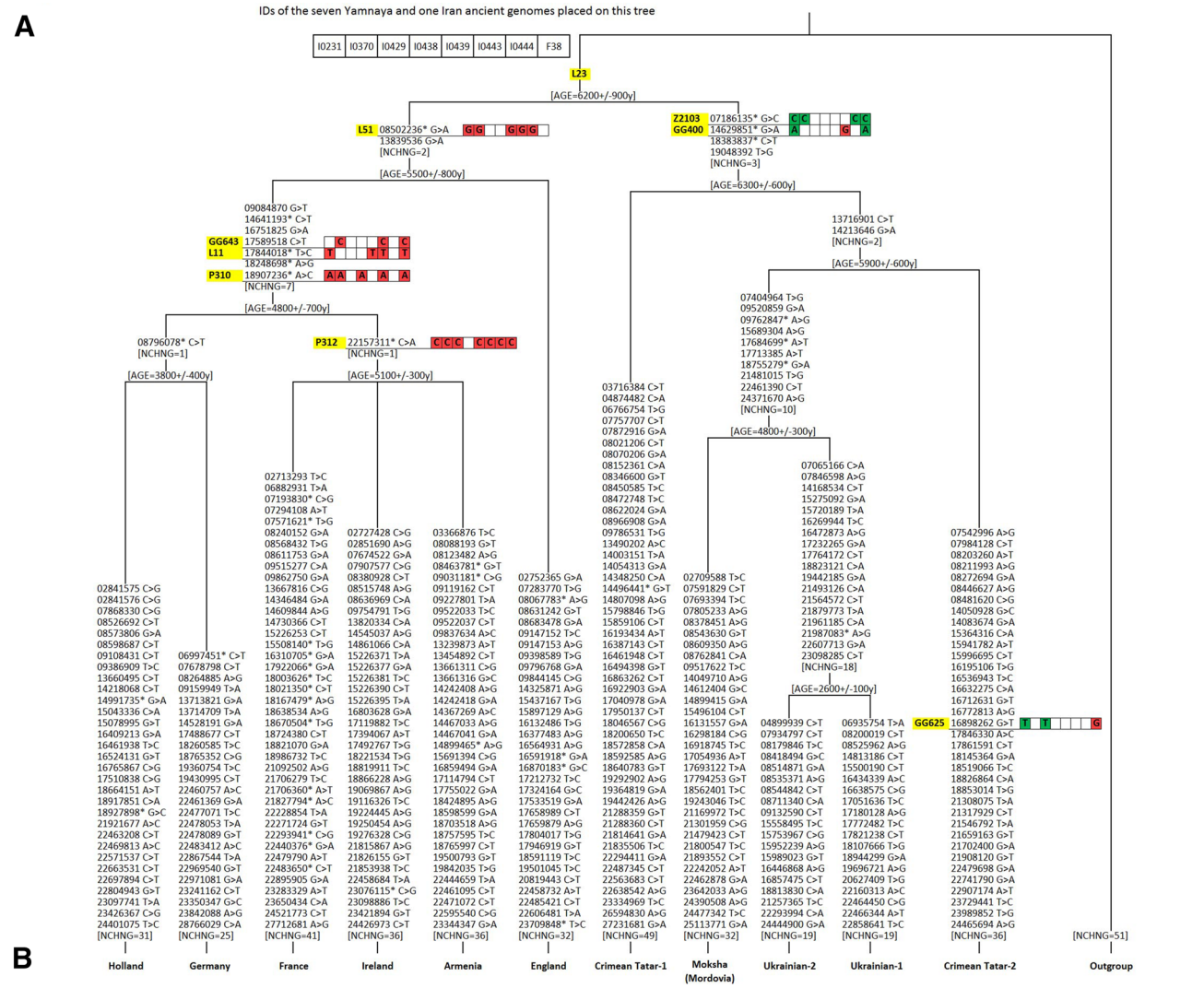
The West Asian populations that were substantially underrepresented in earlier studies of Y-chromosomal variation have received greater attention in the last five years, and new data have been published (Di Cristofaro et al. 2013; Grugni et al. 2012; Herrera et al. 2012; Karafet et al. 2016; the present study). This increased dataset allowed us to analyze the West Asian Y-chromosomal pool in a more systematic way.

Several analyses, including two main patterns of genetic distance maps, two clusters on the MDS plot, and direct AMOVA testing, revealed that the West Asian pool of paternal lineages is structured most strongly by the genetic difference between upland and lowland groups. The first group includes populations of the Anatolian, Armenian and Iranian uplands, while the latter includes populations from lowlands of Mesopotamia, the Levant, and the Arabian Peninsula. It should be noted, however, that genetic clustering of populations into upland and lowland groups is unlikely to be explained by environmental differences directly, because the markers analyzed are thought to be neutral. Instead, we hypothesize indirect influences: gene flow occurred mainly between populations which live in similar environments and speak related languages.

Thus, the link between geographic and genetic landscapes might be mediated by differentiation into linguistic groups and indeed, the inter-group variation between language families is the second largest after variation between upland and lowland groups (Table 1). Note that, areas of language families themselves follow the upland/lowland contrast. It has been hypothesized that Indo-European languages might have originated in the Anatolian mountains (Renfrew 1987; Bouckaert et al. 2012), and populations of the Anatolian, Armenian, and Iranian uplands evidently spoke Indo-European languages for the last few millennia. Some of them shifted to Turkic languages in medieval times, forming the present-day Turks and Azeri. In contrast, the West Asian lowlands were probably the homeland for Semitic languages (Kitchen et al. 2009), and these languages are currently spoken in the Levant, Mesopotamia, and the Arabian Peninsula.

This structure of the language map was not reshaped even by such pronounced historical events as the expansion of the (Indo-European speaking) Persians, whose empire covered most of West Asia in the 3rd–7th centuries BC and the expansion of (Semitic speaking) Arabs whose Caliphate included most of West Asia in the 7th century AD. One may hypothesize that this stability of the areas of language families, and the stable genetic contrast between upland and lowland populations, reflects the stability of the West Asian population structure which originated in the Neolithic and survived in its principal features till today.

This view is supported by evidence from recent studies of ancient DNA in West Asia (Allentoft et al. 2015; Jones et al. 2015; Mathieson et al. 2015; Lazaridis et al. 2016; Broushaki et al. 2016) which, along with extensive autosomal data, provided Y-haplogroup spectra for key populations of the region (Y-chromosomal results from these studies are summarized in the Table S7). Among then, the Neolithic populations of Anatolia (ancestral to Neolithic Europeans) carry around 50% of haplogroup G2, one out of the 15 Y-typed samples represented haplogroup J, but no sample belonged to haplogroup E. Similarly, among six Neolithic and Chalcolithic individuals from the Iranian plateau (Broushaki et al. 2016; Lazaridis et al. 2016), three samples represented haplogroup G, one sample represented haplogroup J, and no sample belonged to haplogroup E. In great contrast, in the



POS	SNP name	REF	ALT	ANC	DER	I0231	I0370	I0429	I0438	I0439	I0443	I0444	F38
7186135	Z2103/CTS1078	G	C	G	C	Ccc	cc					C	CCC
8502236	L51/M412/PF6536/S167	A	G	G	A	GGGG	g			g	GGGgg	G	
14629851	GG400/Y4371/Z8128	G	A	G	A	A					G		AA
16898262	GG625	G	T	G	T	TTTt		T					G
17589518	GG643	T	C	C	T		CC				CCC		CC
17844018	L11/PF6539/S127	C	T	T	C	Tttt				t	TTTt		T
18907236	P310/PF6546/S129	C	A	A	C	AAAAaa	A		a		AA		A
22157311	P312/PF6547/S116	A	C	C	A	CCCccccccccc	c	Cccc		Ccc	CCCCCCCCCCCccccc	CC	CCC

Fig. 4 Phylogenetic tree of Y-chromosomal haplogroup R1b based on resequencing. **a** Tree based on equal numbers of West European and East European samples. The *colored squares* to the right of some SNPs show genotypes of seven ancient Yamnaya genomes and one ancient genome from Iron Age Iran. A *red square* means that the given genome has the ancestral genotype at this SNP and therefore does not lie on the branch defined by this SNP. A *green square* means that the given genome has the derived genotype at this SNP and therefore does lie on the branch defined by this SNP. [NCHNG] indicates number of changes along the given part of the tree. **b** Summary of raw genotypes of ancient genomes. Each *letter* indicates single read in the published BAM files of the ancient genomes (Haak et al. 2015; updated in Mathieson et al. 2015; Broushaki et al. 2016). *Upper case letters* designate reads from the forward primer while lower case mean reverse primer reads. For example, "Ccc" for the IO231 sample at SNP Z2103 shows that this SNP in this genome was read three times, once on one strand and twice on the other strand, and all three reads identified cytosine (color figure online)

Levant, haplogroup E comprised around 50% of Neolithic and ancestral hunter–gatherer populations, while haplogroups G and J were not found (Lazaridis et al. 2016).

The extreme genetic diversity in Neolithic West Asia was highlighted by autosomal data (Lazaridis et al. 2016; Broushaki et al. 2016). While autosomal markers

demonstrated that Neolithic Iranians were equally distant from Neolithic Anatolians and Neolithic Levantines, the aforementioned Y-chromosomal spectrum of Neolithic Iranians is partly similar to Neolithic Anatolians but does not overlap with Neolithic Levantines reported to date. Starting from the Bronze Age, most haplogroups tend to appear in most subregions of West Asia (Allentoft et al. 2015; Lazaridis et al. 2016; Broushaki et al. 2016), in agreement with autosomal evidence of intensive mixing starting from this epoch (Lazaridis et al. 2016). One may conclude that the upland vs. lowland contrast in paternal lineages has existed in West Asia at least since the Neolithic epoch, and, despite it became much less pronounced due to multiple migrations, this pattern remains the principal feature in structuring the present-day West Asian paternal pool.

Preserving the motherland gene pool in diaspora populations

The overall similarity between motherland and Diaspora Armenians allowed us to conclude that the Diaspora was established from large representative samples, and that preserving ethnic identity for centuries has resulted in preserving the genetic legacy of ancient Armenians in their emigrant Diaspora populations. Note, however, that this conclusion is drawn from paternal lineages, which do not trace possible female-mediated gene flow from host populations; further autosomal data from Diaspora populations would be needed to address this question.

The substantial frequency of haplogroup R1a-M198 in Don Armenians is a clear sign of admixture with the surrounding host populations. This is not surprising, because this Armenian population has for the last few generations lived in the suburban area of Rostov city with about 1 million ethnic Russians.

The phylogenetic analysis revealed Armenian-specific haplotype clusters in each of the three aforementioned haplogroups. Ages of the clusters R-L23-β (3000 ± 1200 YBP) and J-M67-α (2000 ± 500 YBP) coincide with the time of formation of Armenian people in the area, and the age of the Hemsheni-specific cluster G-GG265 (1150 YBP, Balanovsky et al. 2015) agrees with the splitting of this Armenian subpopulation (Torlakian 1981).

The genetic relationship of the North Caucasus and Transcaucasian populations

North Caucasian populations are genetically much closer in their male line to West Asian groups than to any other neighboring group, and received their initial gene pool from West Asia (Balanovsky et al. 2011; Yunusbayev et al. 2012). However, within West Asia, the North Caucasian groups have the most atypical genetic composition. The question arises where Transcaucasian and Armenian gene pools find their places, as they are geographically intermediate between the North Caucasus and other West Asian regions.

We found that the Armenian and Azeri populations resembled the main corpus of West Asian populations rather than the North Caucasian cluster (Supplementary Fig. 5). The average haplogroup frequencies in the Armenian populations were much more similar to Turks (genetic distance $d = 0.08$) and some Iranian populations ($d = 0.14$) than to neighboring Georgians (genetic distance $d = 0.59$). In contrast, most Transcaucasian populations—Kartvelian-speaking Laz, Imeretins, eastern Georgians, and North Caucasian-speaking Abkhazians—were similar to the North Caucasus cluster (Supplementary Fig. 5). Note, however, that although these Transcaucasian populations are similar to North Caucasian ones according to average haplogroup frequencies, they contrast in their haplogroup diversity. The Y-chromosomal pool in Transcaucasia is heterogeneous (Supplementary Table 4) while most North Caucasus populations are characterized by a single predominant haplogroup (Balanovsky et al. 2011).

The newly identified eastern branch of haplogroup R1b links West Asian and ancient Yamnaya gene pools

The carriers of the Yamnaya archaeological culture that spread in the Bronze Age to the East European steppes were recently shown to be key element in formation of present-day European and Central Asian gene pools (Haak et al. 2015; Allentoft et al. 2015; Mathieson et al. 2015). The autosomal gene pool of the Yamnaya was modeled as mix of aboriginal East European populations and migrants from West Asia (Haak et al. 2015; Allentoft et al. 2015). Since the Y-chromosomal gene pool of the Yamnaya is represented mainly by haplogroup R1b (as shown for both Yamnaya subpopulations studied to date), the question arises of whether Yamnaya Y-chromosomes also originated from West Asia.

Our phylogenetic tree of haplogroup R1b clearly shows that, apart from well-known R-L51 branch predominant in West Europe, there is a distinct R-GG400 branch in East Europe (Fig. 4a). Markers defining this eastern branch were revealed in a body of published Y-chromosomal sequences (Karmin et al. 2015; Hallast et al. 2015; Batini et al. 2015; the present study). Though it was noted that Yamnaya Y-chromosomes do not belong to the predominant European R1b-L11 branch (Poznik et al. 2016) and that an Iranian Iron Age individual carried the same R1b-branch as Yamnaya individuals (Broushaki et al. 2016), the geographic distribution of this branch has not been further described. So, here, we wish to underscore that this eastern branch R-GG400 is

of equal phylogenetic level and similar age to the branch R-L51, but has a contrasting geographic distribution. The distribution of the western R-L51 branch is well known: it comprises more than half of the West European paternal pool and is found at lower frequencies in East Europe (Myres et al. 2011). Data which we collected on the geographic distribution of the eastern R-GG400 branch are scarce, but place its main area on the East European steppes and West Asia (Fig. 4; see also Figure S38 in Karmin et al. 2015). The frequency of this eastern branch has so far been estimated only in a few populations: it comprises 5% in Greeks from Cyprus (Voskarides et al. 2016), 0–5% in different Georgian and 2–20% in different Armenian populations (this study). Large-scale population genotyping surveys have yet to be reported to reconstruct the phylogeography of R-GG400 and its subbranches in detail.

The currently available dataset does not contradict the hypothesis that R-GG400 marks a link between the East European steppe dwellers and West Asians, though the route and even direction of this migration is disputable. It does, however, demonstrate that present-day West European R1b chromosomes do not originate from the Yamnaya populations analyzed in (Haak et al. 2015; Mathieson et al. 2015) and raises the question of their origin. A Bronze Age origin is more likely than a Neolithic one (Balaresque et al. 2010), but further ancient DNA studies may be necessary to identify this source.

Acknowledgements We thank all sample donors whose participation made this study possible, and the Georgian church which approved and assisted in sampling in Georgia.

The Genographic Consortium includes: Li Jin, Hui Li, & Shilin Li (Fudan University, Shanghai, China); Pandikumar Swamikrishnan (IBM, Somers, New York, United States); Asif Javed, Laxmi Parida & Ajay K. Royyuru (IBM, Yorktown Heights, New York, United States); R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia); Pierre A. Zalloua (Lebanese American University, Chouran, Beirut, Lebanon); Syama Adhikarla, ArunKumar, GaneshPrasad, Ramasamy Pitchappan, Arun Varatharajan Santhakumari, Kavitha Valampuri (Madurai Kamaraj University, Madurai, Tamil Nadu, India); R. Spencer Wells and Miguel G. Vilar (National Geographic Society, Washington, District of Columbia, United States); Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa); Elena Balanovska & Oleg Balanovsky (Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia); Chris Tyler-Smith (The Wellcome Trust Sanger Institute, Hinxton, United Kingdom); Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil); Jaume Bertranpetit, Marc Haber, Marta Melé, & David Comas(Universitat Pompeu Fabra, Barcelona, Spain); Christina J. Adler, Alan Cooper, Clio S. I. Der Sarkissian & Wolfgang Haak (University of Adelaide, South Australia, Australia); Matthew E. Kaplan & Nirav C. Merchant (University of Arizona, Tucson, Arizona, United States); Colin Renfrew (University of Cambridge, Cambridge, United Kingdom); Andrew C. Clarke & Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand); Jill B. Gaieski & Theodore G. Schurr (University of Pennsylvania, Philadelphia, Pennsylvania, United States).

Compliance with ethical standards

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Funding This study received primary support from the Russian Science Foundation Grant 14-14-00827 (to OB, MC, AA, VZ), including data analysis and completing the paper. Y-chromosomal resequencing was done mainly using funding from the Historical Genetics lab in the Moscow Institute of Physics and Technology. Erzurum Armenian samples were genotyped with support from the Russian Foundation for Basic Research (Grant 16-36-00122 to MC), while the Hemsheni and Krasnodar Armenian samples were genotyped with support from the Russian Foundation for Basic Research (Grant 16-06-00364 to EP) and from the Genographic project. CTS was supported by The Wellcome Trust (098051).

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abu-Amero KK, Hellani A, González AM, Larruga JM, Cabrera VM, Underhill PA (2009) Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BMC Genet 10:59. doi:10.1186/1471-2156-10-59

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S et al (2015) Population genomics of Bronze Age Eurasia. Nature 522(7555):167–172. doi:10.1038/nature14507

Balanovskaia EV, Nurbaev SD, Balanovskiĭ OP, Pocheshkhova EA, Borovinskikh AA, Ginter EK (1999) Genogeographic analysis of subdivided population. The Adyge gene pool in the Caucasian gene pool system. Russian J Genet 35(6):818–830

Balanovsky O, Rootsi S, Pshenichnov A et al (2008) Two sources of the Russian patrilineal heritage in their Eurasian context. Am J Hum Genet 82(1):236–250. doi:10.1016/j.ajhg.2007.09.019

Balanovsky O, Dibirova K, Dybo A et al (2011) Parallel evolution of genes and languages in the caucasus region. Mol Biol Evol 28(10):2905–2920. doi:10.1093/molbev/msr126

Balanovsky O, Zhabagin M, Agdzhoyan A et al (2015) Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of iranic speakers. PLoS One 10(4):e0122968. doi:10.1371/journal.pone.0122968

Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previderè C, Wilson IJ, Tyler-Smith C, Jobling MA (2010) A predominantly neolithic origin for European paternal lineages. PLoS Biol 19 8(1):e1000285. doi:10.1371/journal.pbio.1000285

Batini C, Hallast P, Zadik D et al (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. Nature Commun 6:7152. doi:10.1038/ncomms8152

Bergström et al (2016) Deep roots for aboriginal australian Y chromosomes. Current Biology. http://dx.doi.org/ 10.1016/j.cub.2016.01.028

Battaglia V, Fornarino S, Al-Zahery N et al (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. Eur J Hum Genet 17(6):820–830. doi:10.1038/ejhg.2008.249

Bouckaert R, Lemey P, Dunn M et al (2012) Mapping the origins and expansion of the Indo-European language family. Science (New York, NY) 337(6097):957–960. doi:10.1126/science.1219669

Broushaki F, Thomas M, Link V, et al (2016) Early neolithic genomes from the eastern fertile crescent. Science. 2016; Jul 14. Published online. doi: 10.1126/science.aaf7943

Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, Herrera RJ (2008) Y-chromosome diversity characterizes the Gulf of Oman. Eur j hum genet 16(3):374–386. doi:10.1038/sj.ejhg.5201934

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85(16):6002–6006

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL et al (2004) Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114(2):127–148

Devoto G (1962) Origini indeuropee. Sansoni, Florence, p 250

Di Cristofaro J, Pennarun E, Mazières S et al (2013) Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS One 8(10):e76748. doi:10.1371/journal.pone.0076748

El-Sibai M, Platt DE, Haber M et al (2009) Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. Ann Hum Genet 73(Pt 6):568–581. doi:10.1111/j.1469-1809.2009.00538.x

Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128(2):415–423. doi:10.1002/ajpa.20188

Flores C, Maca-Meyer N, Larruga JM, Cabrera VM, Karadsheh N, Gonzalez AM (2005) Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. J Hum Genet 50(9):435–441. doi:10.1038/sj.ejhg.5201225

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of native American mtDNA variation: a reappraisal. Am J Hum Genet 59(4):935–945

Gamkrelidze TM, Ivanov VVS (1995) Indo-European and the Indo-Europeans: a reconstruction and historical analysis of a proto-language and a proto-culture, 2nd edn. In: Trends in linguistics: studies and monographs, vol 80

Ge J, Budowle B, Aranda XG, Planz JV, Eisenberg AJ, Chakraborty R (2009) Mutation rates at Y chromosome short tandem repeats in Texas populations. Forensic sci int Genet 3(3):179–184. doi:10.1016/j.fsigen.2009.01.007

Goedbloed M, Vermeulen M, Fang RN et al (2009) Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpF/STR® Yfiler® PCR amplification kit. Int J Legal Med 123(6):471–482. doi:10.1007/s00414-009-0342-y

Grugni V, Battaglia V, Hooshiar Kashani B et al (2012) Ancient migratory events in the middle East: new clues from the Y-chromosome variation of modern Iranians. PLoS One 7(7):e41252. doi:10.1371/journal.pone.0041252

Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez F et al (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26(6):520–528. doi:10.1002/humu.20254

Haak W, Lazaridis I, Patterson N, Rohland N, et al (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. (7555):207–11. doi:10.1038/nature14317

Haber M, Platt DE, Badro DA et al (2011) Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. Eur J Hum Genet 19(3):334–340. doi:10.1038/ejhg.2010.177

Haber M, Mezzavilla M, Xue Y, Comas D et al (2016) Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations. Eur J Hum Genet 24:931–936. doi:10.1038/ejhg.2015.206

Hallast P, Batini C, Zadik D et al (2015) The Y-Chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol 32(3):661–673. doi:10.1093/molbev/msu327

Herrera KJ, Lowery RK, Hadden L et al (2012) Neolithic patrilineal signals indicate that the Armenian plateau was repopulated by agriculturalists. Eur J Hum Genet 20(3):313–320. doi:10.1038/ejhg.2011.192

Hovannisian RG (ed) (1997) The Armenian people. St Martin's, New York

Hovhannisyan A, Khachatryan Z, Haber M et al (2014) Different waves and directions of Neolithic migrations in the Armenian Highland. Investiga Genet 5:15. doi:10.1186/s13323-014-0015-6

Jones ER, Gonzalez-Fortes G, Connell S et al (2015) Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nature Commun 6:8912. doi:10.1038/ncomms9912

Karafet TM, Bulayeva KB, Nichols J, Bulayev OA, Gurgenova F, Omarova J, Yepiskoposyan L, Savina OV, Rodrigue BH, Hammer MF (2016) Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan. J Hum Genet 61(3):181–191

Karmin M, Saag L, Vicente M et al (2015) A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res 25(4):459–466. doi:10.1101/gr.186684.114

Kitchen A, Ehret C, Assefa S, Mulligan CJ (2009) Bayesian phylogenetic analysis of semitic languages identifies an early Bronze Age origin of semitic in the near East. Proc Royal Soc B 276(1668):2703–2710. doi:10.1098/rspb.2009.0408

Koshel SM (2012) Geoinformation technologies in genogeography. In: Lure IK, Kravtsova VI (eds) Modern geographic cartography. Data + p, Moscow, pp 158–166 **(In Russian)**

Lang DM (1970) Armenia: cradle of civilization. Allen & Unwin, London. ISBN 0-04-956007-7

Lazaridis I, Nadel D, Rollefson G et al (2016) Genomic insights into the origin of farming in the ancient Near East. Nature 536:419–424. doi:10.1038/nature19310

Lowery R, Herrera K, Uribe G, Reguiero M, Herrera RJ (2013) Subpopulation structure evident in forensic Y-STR profiles from Armenian geographical groups. Leg Med (Tokyo) 15(2):85–90. doi:10.1016/j.legalmed.2012.10.003

Mathieson I, Lazaridis I, Rohland N, Mallick S, Llamas B, Pickrell J, Meller H, Rojo Guerra M, Krause J, Anthony D, Brown D, Fox CL, Cooper A, Alt KW, Haak W, Patterson N, Reich D (2015) Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528(7583):499–503. doi:10.1038/nature16152

Myres NM, Rootsi S, Lin AA et al (2011) A major Y-chromosome haplogroup R1b holocene era founder effect in central and Western Europe. Eur J Hum Genet 19(1):95–101. doi:10.1038/ejhg.2010.146

Nasidze I, Sarkisian T, Kerimov A, Stoneking M (2003) Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. Hum Genet 112(3):255–261

Nei M (1975) Molecular population genetics and evolution. North-Holland Pub. Co, Amsterdam

Poznik GD, Xue Y, Mendez FL et al (2016) Punctuated bursts in human male demography inferred from 1244 worldwide Y-chromosome sequences. Nat Genet 48(6):593–599. doi:10.1038/ng.3559

Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES (2003) Genetic structure of Mediterranean populations revealed by

Y-chromosome haplotype analysis. Am J Phys Anthropol 121(2):157–171. doi:10.1002/ajpa.10187

Ravid-Amir O, Rosset S (2010) Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. Bioinformatics 26(18):i440–i445. doi:10.1093/bioinformatics/btq367

Renfrew C (1987) Archaeology and language: the puzzle of the Indo-European Origins. Jonathan Cape, London

Rootsi S, Behar DM, Järve M et al (2013) Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites. Nature Commun 4:2928. doi:10.1038/ncomms3928

Rosser ZH, Zerjal T, Hurles ME et al (2000) Y-Chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67(6):1526–1543

Ruhlen M (1987) A guide to the world's languages. vol I: classification. Stanford University Press, Stanford, p 463

Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67(3):718–726

Sanchez JJ, Hallenberg C, Borsting C, Hernandez A, Morling N (2005) High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. Eur J Hum Genet 13(7):856–866. doi:10.1038/sj.ejhg.5201390

Sanchez-Diz P, Alves C, Carvalho E, Carvalho M, Espinheira R, Garcia O et al (2008) Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study. Int J Legal Med 122(6):529–533. doi:10.1007/s00414-008-0265-z

Schneider S, Roessli D, Excoffier L (2000) Arlequin vers. 2.000: a software for population genetics data analysis. User man ver 2:2496–2497

Semino O, Passarino G, Oefner PJ et al (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155–1159. doi:10.1126/science.290.5494.1155

Simonian H (ed) (2007) The Hemshin. History, society and identity in the Highlands of Northeast Turkey, New York: Routledge, p 472

StatSoft, Inc. STATISTICA (data analysis software system) (2001) version 62001. www.statsoft.com

Tarkhnishvili D, Gavashelishvili A, Murtskhvaladze M, Gabelaia M, Tevzadze G (2014) Human paternal lineages, languages, and environment in the Caucasus. Hum Biol 86(2):113–130. doi:10.3378/027.086.0205

Tololyan K (2001) Armenian diaspora. Diaspora 9(1):107–136

Torlakian BG (1981) Ethnography of Hamshen Armenians. Ethnography and Folklore. vol. 13. Yerevan. pp. 24–111 **(In Armenian)**

Underhill PA, Myres NM, Rootsi S et al (2010) Separating the post-glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet 18(4):479–484. doi:10.1038/ejhg.2009.194

Voskarides K, Mazières S, Hadjipanagi D et al (2016) Y-chromosome phylogeographic analysis of the Greek-Cypriot population reveals elements consistent with Neolithic and Bronze Age settlements. Investig Genet 7:1. doi:10.1186/s13323-016-0032-8

Weale ME, Yepiskoposyan L, Jager RF, Hovhannisyan N, Khudoyan A, Burbage-Hall O (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethnonational group. Hum Genet 109(6):659–674. doi:10.1007/s00439-001-0627-9

Yunusbayev B, Metspalu M, Jarve M, Kutuev I, Rootsi S, Metspalu E et al (2012) The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol 29(1):359–365. doi:10.1093/molbev/msr221

Zalloua PA, Xue Y, Khalife J et al (2008a) Y-Chromosomal diversity in lebanon is structured by recent historical events. Am J Hum Genet 82(4):873–882. doi:10.1016/j.ajhg.2008.01.020

Zalloua PA, Platt DE, El Sibai M et al (2008b) Identifying genetic traces of historical expansions: phoenician footprints in the Mediterranean. Am J Hum Genet 83(5):633–642. doi:10.1016/j.ajhg.2008.10.012