Title

Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan.

Journal

Journal of human genetics, 61(3)

ISSN

1434-5161

Authors

Karafet, Tatiana M
Bulayeva, Kazima B
Nichols, Johanna
et al.

Publication Date

2016-03-01

DOI

10.1038/jhg.2015.132

Co-evolution of genes and languages and high levels of population structure among the highland populations of Daghestan

Authors' final version of manuscript.
Published as: Karafet, Tatiana M. et al. 2015. Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan. *Journal of Human Genetics* 2015.1-11.
(Senior author and corresponding author Michael F. Hammer. mfh@email.arizona.edu)

For tables, figures, and supplement see the published version.

Tatiana M. Karafet[1], Kazima B. Bulayeva[2], Johanna Nichols[3], Oleg A. Bulayev[3], Farida Gurgenova[3], Jamilia Omarova[3], Levon Yepiskoposyan[4], Olga V. Savina[1], Barry H. Rodrigue[5], and Michael F. Hammer[*,1]

[1]ARL Division of Biotechnology, University of Arizona, Tucson, AZ, 85721, USA

[2]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

[3]University of California, Berkeley, CA, 94720, USA

[4]Institute of Molecular Biology, National Academy of Sciences, Yerevan, Armenia

[5] Current Address: Lewiston-Auburn College, University of Southern Maine, Lewiston, ME, 04240, USA

Running title: Genetic and linguistic affinities among isolated populations in Daghestan.

Keywords: Daghestan/ Nakh-Daghestanian languages/ autosomal genome-wide SNPs/ Y chromosome/ mtDNA.

Many interesting questions about the history of the isolated populations of mountainous Daghestan remain unresolved. Previous genetic studies generally have been restricted to uniparental markers and have not included many of the key populations of the region. To improve our understanding of the genetic structure of Daghestani populations at a fine geographic scale, and to investigate possible correlations between genetic and linguistic variation, we analyzed ~550,000 autosomal SNPs from 21 ethnic Daghestani groups, along with Y chromosome and mtDNA markers from the same samples. We found high levels of population structure in Daghestan consistent with the hypothesis of long-term isolation among populations of the highland Caucasus. Highland Daghestani populations exhibit very low within- and high levels of between-population diversity, leading to some of the highest F_{ST} values observed for any region of the world. Historical patterns of social interaction among highland farmers at the community level largely explain the significant positive correlation between gene and language diversity assuming that language and genetic variation in highland Daghestan have actually evolved together. Our data are consistent with the scenario in which most Daghestanian-speaking groups descend from a common ancestral population (~6,000 – 6,500 years ago) that spread to the Caucasus by demic diffusion followed by population fragmentation and low levels of gene flow.

INTRODUCTION

The Caucasus region is characterized by extreme cultural and linguistic differentiation, with more than 50 autochthonous ethnic groups living in a small geographic area. The compact and very old branch of the Nakh-Daghestanian (ND) or East Caucasian linguistic family occupies most of Daghestan (the Russian republic comprising the eastern one-third of the Great Caucasus range). The Caucasus Mountains have long served as a crossroad connecting the Near East and the eastern European plains, and likely witnessed one of the initial expansions of agriculture from Mesopotamia to the north and northeast[1-3]. The Caucasus highlands were uninhabitable until after the end of glaciation. Archaeological sites in Daghestan appear at nearly 2,000 meters in the Mesolithic (~10,000 BP), and provide evidence of continuous human occupation afterward [2,4]. One of the earliest Neolithic sites outside of Mesopotamia is found at Chokh in the eastern Daghestan (~8,000 BP)[4,5]. Its cultural continuity with the earlier Mesolithic strata suggests that plant and animal domestication spread to this region by diffusion rather than by population replacement.

The ancestral ND protolanguage is about 8,000 years old and has a reconstructable vocabulary consistent with early Neolithic culture[1,6,7]. The ND linguistic family is extremely diversified with some 30-35 daughter languages, and has never been found beyond the Caucasus highlands and highland/lowland interface[1,8]. While specific language associations with Chokh cannot be demonstrated, the dispersal of ND likely dates to the early Neolithic in highland Daghestan. From the Mesolithic to historical times there is no archaeological or linguistic evidence pointing to migrations into the highlands. Thus, the great age and diversification among groups living in close proximity may well reflect expansion of the Neolithic through the highlands and its entrenchment there [9,10]. While the highland populations have likely lived for hundreds of generations in relative isolation in the same region[2,11-13], the North Caucasus plain has seen several spreads of steppe nomadic languages and cultures. The major known linguistic impacts on the Caucasus have been the Iranian arrival beginning in the second millennium BC, the movements of Bulgar and Khazar Turkic groups in the mid-first millennium AD, and the arrival of Kipchak Turkic groups in the late first millennium.

Daghestan, with its exceptional combination of linguistic, geographic, and cultural diversity, presents an excellent natural laboratory for tracking the influence of demographic processes on patterns of genetic variation. Its compact distribution of deeply divergent languages

provides a unique possibility to test gene-language co-evolution at a fine scale and to illuminate the genomic footprint of events such as migration and admixture on genetic variation. However, most previous genetic studies either did not perform dense sampling of the region, or were limited to autosomal *Alu* insertion, autosomal STR, Y-chromosomal, and mtDNA surveys[14-25].

In the present study, we gather dense SNP data from the autosomes, as well as from both haploid regions of the genome in 21 ethnic groups in Daghestan. This study was designed to (1) investigate the co-evolution of genes and languages, comparing and contrasting patterns of linguistic, genetic and geographic variation among Daghestani populations, 2) examine similarities and differences in population differentiation among Daghestanian ethnic groups and populations from Europe, the Near East, Central Asia, and South Asia, and (3) investigate congruence in patterns of genetic variation on the autosomes, Y chromosome, and mtDNA.

MATERIALS AND METHODS

Populations and samples

A total of 842 cheek swab samples from 21 ethnic Daghestani groups and cosmopolitan Chechens were collected from the volunteers with informed consent and approval of the IRB of the University of Arizona. Armenian samples were collected by LY in Ararat Region, Armenia with a written consent form approved by the Institute of Molecular Biology, Yerevan, Armenia. All additional non-Daghestani samples were included in previous studies[26,27]. Fifteen highland Daghestani populations speak distinct languages that belong to the Daghestanian branch of the Nakh-Daghestanian (ND) language family. Six lowland populations speak languages that are not members of ND (non-ND). Three ethnic groups of Daghestan (Kumyks, Nogais, and Azerbaijans) speak languages ~~that are closely related to~~ of the Turkic language family. Ethnic Tats, Mountain Jews and a group of Azerbaijans speak languages belonging to the Iranian language branch of the Indo-European language family. **Figure 1** shows the population locations. Additional information on sampling locations, sample and population sizes, language classification is available in **Supplementary Table 1**. Two populations of Laks were combined for analyses because the lowland population in Novo-Churtakh represents recent migrants (~70 years) from the higland village Churtakh[28].

Autosomal SNP data analyses

A total of 314 samples from Daghestan and 261 samples from the Near East (NEA), Caucasus (CAU), Europe (EUR), South Asia (SAS) and Central Asia (CAS) (**Supplementary Table 1**) were genotyped for 567,096 single nucleotide polymorphisms (SNPs) on Affymetrix (Axiom) platform using standard protocols. Details on the curation and public availability of these data are presented elsewhere[29]. After removing close relatives, the total number of samples in our Axiom dataset was 480. For several analyses we used SNPs from the intersection of our data with publicly available samples (**Supplementary Table 1**). The final merged data set resulted in 104,519 SNPs for 1,430 individuals across 59 populations. The merged autosomal data set was used for Principle Components Analysis (PCA)[30] , MDS plots[31] and ADMIXTURE analysis[32], for estimating diversity parameters and genetic differentiation indices by ARLEQUIN 3.5 software[33] and SMARTPCA[30], and to infer population splits and migration events with the TreeMix (version 1.12) program[34].

To estimate the effective population sizes and divergence time between populations we evaluated the decay of LD with recombination distance for each chromosome using the genotypic-based r^2 statistic estimated in PLINK[35]. Analyses of LD and IBD were performed on full data set of 549,008 SNPs. See details in the recent study by Karafet *et al*.[29] Divergence time between populations was estimated as $T = 2N_eF_{ST}$, where N_e is effective population size as the harmonic means between the two populations[36]. We ran GERMLINE 1.5.1[37] on the phased unpruned data with default parameters (-min_m 3 –bits 128 –err_hom 4 –err_het 1) to detect IBD (identically by descent) pairwise segments sharing for all pairs of study samples (N=480). We divided the genome into non-overlapping 1Mb blocks, removed blocks with <100 SNPs, and kept only the shared IBD segments whose length exceeded 3 Mb. We computed the mean length of IBD sharing among ND-speaking populations in the same way as Behar et al.[38] Genetic distances based on IBD sharing were evaluated as $D_{ij} = 1 – (X_{ij}/X_{Max})$, where X_{ij} is the total length of shared IBD segments between populations i and j, X_{Max} is the maximum total length of shared IBD segments among ND populations.

Analyses of Y chromosome and mtDNA.

A total of 2,461samples belonging to 60 populations from Daghestan, CAU (Chechens and Armenians), NEA, EUR, CAS and SAS were analyzed for 140 polymorphic sites from the nonrecombining portion of the human Y chromosome (NRY) by allele-specific PCR or RFLP

(**Supplementary Table 2**)**.** NRY polymorphic sites included 137 published binary polymorphisms together with a set of three new SNPs: P323, P354, and P369. Mapping information as well as primer sequences for these SNPs can be found in the **Supplementary Table 3 and Supplementary Figure 1**. The information was submitted to the International Society of Genetic Genealogy (http://www.isogg.org/tree/index.html). We use the mutation-based naming system that keeps the major haplogroup information followed by the name of the terminal mutation that defines a given haplogroup[39]. Genotyping data resulted in 101 Y haplogroups which are presented in **Supplementary Table 4.** We also analyzed 13 short tandem repeats (STRs): *DYS19, DYS385a, DYS385b, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS426*, *DYS438*, and *DYS439* as described by Redd et al[40]. Y chromosome STR data are provided in **Supplementary Table 5**.

A total of 2,163 samples were sequenced for mtDNA HVSI and typed for 45 coding region markers (**Supplementary Table 6)**. Mitochondrial DNA hypervariable region I sequences are available in GenBank (accession numbers: KP883308 - KP885623). For further genetic analyses we included published data on additional samples from five EUR populations: Austrian, Bulgarian, Greek, French, and Italian [41-45] (**Supplementary Table 1**). mtDNA haplogroup results are presented in **Supplementary Table 7**. Population structure analyses were based on mtDNA haplotypes combining diagnostic SNPs in the coding regions and HVS-I variable sites except highly recurrent 16182C, 16183C, 16193.1C(C) and 16519 mutations.

Molecular diversity, population structure estimates, and genetic distances between populations for NRY and mtDNA markers were computed using Arlequin v. 3.11[33].
To standardize for different mutation rates we applied a measure of interlocus differentiation G'_{ST}[46]. Nonmetric multidimensional scaling (MDS)[31] was performed on the F_{ST} distances using the software package NTSYS[47]. The program Network v. 4.5.1.6 (Fluxus Engineering; http://www.fluxus-engineering.com)[48] was used to build Median-Joining network and to estimate the approximate age of paragroup J-M267* in Daghestan with ρ statistic[49]. To evaluate the correlation among linguistic, geographic and genetic distances, Mantel tests were performed in Arlequin.

Phylogenetic analyses.

We used the SplitsTree program[50] to calculate language distances and to perform phylogenetic

analyses of linguistic and genetic data. Trees were constructed with different techniques. While the language tree was built on linguistic entries across the 21 languages in Daghestan, the autosomal, mtDNA and Y-chromosome trees were built from the matrices of population pairwise F_{ST} genetic distances using the neighbor joining method. Language distances were calculated on 85 characters: most of the words are from items 1-40 and 56-100 in the stability-ranked Swadesh list of the Automated Similarity Judgment Program[51]; and 12 other words were cultural terms. For Mountain Jews and Iranian-speaking Azerbaijanis we have used Tat and Persian languages, respectively, as proxies.

RESULTS

Relationships among Daghestani and neighboring European populations

To explore regional relationships, multidimensional scaling based on autosomal, Y-chromosome STR and mtDNA data was performed on Daghestani ethnic groups along with populations from CAU, CAS, EUR, NEA, and SAS (**Figure 2 a, b, c**). Although populations are roughly clustered according to their geographic regions, the patterning within and between groups is quite distinctive for three different sets of markers. Autosomal and Y chromosome markers reveal relatively distinct geographic clusters with partial overlap between Daghestani, CAU and NEA populations. In contrast, MDS plot based on mtDNA shows that European populations are intermingled with a group of Daghestani, CAU and NEA populations. In all three plots Daghestani non-ND populations (Kumyks, Nogais, Azebajanians, Tats, and Mountain Jews) are generally found within NEA and/or CAU clusters, while ND populations always form very loose but distinct clusters with several outlier populations.

We applied PCA on the merged autosomal data set using a 'drop one in' procedure for incorporating populations [52] (**Supplementary Figure 2)**. Specifically, PCA analysis was performed for each individual from a Daghestani population isolate along with all other samples. Each population isolate sample's resultant PC coordinates for the first two components were then plotted together along with the average PC co-ordinates for other samples across all runs. This procedure helps to avoid the potential effect of high relatedness among the individuals in Daghestani isolated populations and uneven sample sizes[52]. In general, the resulting PCA plot separates regional populations according to their geographical location. The first PC splits NEA, CAU and EUR populations from Asians, whereas the second component subdivides NEA, CAU

and EUR groups. Daghestani groups (other than Nogais and Mountain Jews) are intermingled with other populations from the Caucasus. Consistent with their origin, Nogais demonstrate a genetic resemblance with CAS populations, while Mountain Jews are extended towards NEA populations. Turks and Iranians are drawn towards populations from Daghestan, in particular, to Azerbaijanis and Tats. To investigate the internal structure of Daghestani populations 'drop one in' PCA was constructed on full set of autosomal SNPs in 18 ethnic groups from Daghestan (**Supplementary Figure 3a and 3b**). The plots reveal a structure that was not apparent from the PCA on the global populations. Tsezic-speaking groups (Hinukh, Hunzib and Tsez) cluster together; Laks, Akvakhs, Avars, and Tindi form their own loose groups, while non-ND-speaking Kumyks and Nogais are drawn to Azebaijans, Tats, and Mountain Jews.

We employed an unsupervised STRUCTURE-like approach[32] to estimate individual ancestry in K hypothetical ancestral populations. Yoruba and Han Chinese were included in this analysis. The best projecting accuracy was observed for a model with K = 8. Consistent with PCA analysis, Daghestani and Caucasus groups are nearly indistinguishable at K =3 except Nogais, who share a higher Asian ancestral component (**Figure 3).** At K = 5,6 ND populations differentiate from non-ND and other Caucasus groups. With K = 7,8 three ND-speaking populations (Hinuq, Hunzib, and Tsez) became dominated by one single ancestry component and three other populations (Akhvakh, Ratlub, and Tindi) by another component.

To infer the history of population splitting and mixing in the ancestry of ND groups we built a tree using TreeMix[34]. With no admixture events the maximum-likelihood population tree places all ND populations in one cluster accompanied by non-ND populations and other ethnic groups from the Caucasus (**Supplementary Figure 4a)**. Allowing 10 admixture events (**Supplementary Figure 4b)** finds evidence of admixture mostly between and within CAS and SAS regions. No sign of migration edges was observed from any population to ND ethnic groups. We applied the 3-population f3-test to each of the 56 populations (**Supplementary Table 8**). This analysis was introduced by Reich et al[53] to determine evidence of admixture for the *Test* population. A significantly negative value of the f3 statistic implies that population is admixed. Among Daghestani populations the most negative statistics were found for Nogais and Kumyks taking as the reference populations Kazakhs and Georgians. The majority of ND populations do not produce any negative f3 assuming low admixture or substantial post-admixture drift [54].

Distribution of NRY and mtDNA haplogroups

NRY haplogroups are presented in **Supplementary Table 3**. Y chromosome haplogroup distributions and frequencies differ strikingly between the highland ND and lowland non-ND populations. As a whole, the highland ND group exhibits 18 NRY haplogroups, but only two haplogroups are observed with frequencies greater than 7%. Haplogroup J-M267(xL136) is found in all ND populations, ranging from 40% in Lezgi to 100% in Hunzib and Tsez populations with an average of 58%. Interestingly, this haplogroup is rare in major geographic regions (0.2-2%), achieving noticeable occurrence only in lowland non-ND populations (16.3%), in Chechens (8.3%) and Armenians (7.5%) from the CAU, Assyrians (7.1%) and Iranians (6.9%) from the NEA. Haplogroup R-L23(xP310) is present in 9 out of 15 ND populations with the average incidence of 7.8%. Haplogroup R-L23 was found at low frequencies (4-10%) in NEA, EUR, and non-ND populations with the highest frequencies in Assyrians (29%), Tats (29%), Turks (15%), and Russians (13%). Contrary to the NRY, the distributions of mtDNA haplogroups are similar in highland ND and lowland non-ND populations except for relatively high frequency of the U4 haplogroup in ND populations (9.69%) (**Supplementary Table 7)**. Both ND and non-ND populations also resemble our samples from NEA and EUR in their frequencies of common haplogroups H and T.

Congruence in patterns of genetic variation on the NRY, mtDNA, and autosomes

We assessed associations between autosomal, Y-chromosomal and mitochondrial population structure of Daghestan by correlating matrices of genetic distances for two population sets: 19 Daghestani populations and 13 ND-speaking groups. Botlikh and Godoberi were omitted from these analyses since they were not genotyped for autosomal SNPs. For Y-chromosomal data we employed distances based on Y-STR frequencies because a single Y-chromosome haplogroup is prevalent in Daghestan. No significant correlations were found between Y-chromosome and mtDNA structure for both data sets (**Supplementary Table 9)**. We identified significant simple and partial correlations between distances based on autosomal versus Y-chromosome and autosomal versus mtDNA population structure for 19 populations. The highest correlation was observed between autosomal and Y-STR data ($r = 0.53$, $p = 0.005$). When only ND populations were considered, correlations continued to be positive; however, it was significant for autosomal

and Y-STR data (simple correlation: r =0.48, p = 0.024, partial correlation: r = 0.45, p = 0.032) and only marginally significant for autosomal versus mtDNA data (simple correlation: r =0.27, p = 0.083).

Population differentiation and genetic diversity

We investigated parameters of genetic diversity in Daghestani populations and compared the values with those for CAU, NEA, EUR, CAS, and SAS populations. Diversity statistics based on autosomal, NRY, and mtDNA data are given in **Supplementary Tables 10, 11, 12 and 13.** Average gene diversity for three systems exhibited a similar pattern with the lowest values in Daghestan, particularly in populations of highland ND language speakers (highly significant for mtDNA and Y chromosome, marginally significant for autosomal SNPs).

To address the question of population differentiation we employed AMOVA analyses. The F_{ST} values for the Daghestani populations were 0.017, 0.146, 0.155, and 0.075 for autosomal, Y-SNP, Y-STR and mtDNA data, correspondingly, indicating a significant degree of population differentiation within Daghestan (**Table 1**). When only ND populations were included in analyses, the F_{ST} estimates increased by 7-18% for different genetic systems except Y chromosome haplogroups. These values are higher than F_{ST} in NEA, EUR, CAS, SAS, approaching our global F_{ST} values of 0.018, 0.112, and 0.069 based on autosomal data, Y-chromosome STRs and mtDNA polymorphisms typed in ~1,100 – 2,400 individuals from 55-60 global populations. An analysis of molecular variance illustrates that the Y-chromosome STRs (F_{ST} = 0.174) have markedly higher variation among ND populations than mtDNA (F_{ST} = 0.081). The trend holds, when distances are standardized for the different mutation rates. Y-chromosome STRs (G'_{ST} = 0.978) showed noticeably higher population structure than mtDNA (G'_{ST} = 0.435) with an intermediate value for autosomal markers (G'_{ST} = 0.627).

Associations between linguistic, genetic and geographic distances.

The language tree (**Figure 4a**) and genetic tree (**Figure 4b)** based on autosomal data have several structural similarities. They separate the ND-speaking from the Turkic and Iranian-speaking Daghestani groups. Both trees show Lezgi as the first branch off the ND languages, strong clustering is observed for the Tsezic-speaking populations Hinuq, Tsez and Hunzib, and a close relationship is found between Bagwalal and Tindi. Trees constructed on

NRY and mtDNA data show a general lack of correspondence with the language tree (**Supplementary Figure 5a and b**).

To assess the effect of geography and languages on the genetic structure of ND populations we applied partial and multiple Mantel tests. In view of the mountainous Daghestan landscape we explored the correspondence between genetics based on autosomal SNPs, language and geographic distances computed as a) a great circle distance based on GPS coordinates, b) the distance based on altitude, latitude and longitude, and c) the shortest distance by existing automobile roads (**Supplementary Table 14**). All three geographic distances showed no significant correlation with genetics but highly significant association with languages. Further Mantel tests were performed with geographic distances as a great circle distance (**Table 2**). Genetic distances calculated with NRY and mtDNA markers uncovered no significant full or partial correlation with linguistic or geographic distances either for 21 Daghestani populations or for ND-speaking groups However, genetic distances among ND populations based on autosomal SNPs show a marginally statistically significant association with languages ($r = 0.343$, $p = 0.061$), but not with geography ($r = 0.029$, $p = 0.434$). Moreover, partial correlation of genetics with languages revealed a strong significant positive association among ND populations ($r = 0.428$, $p = 0.015$) when controlling for geography, and a negative correlation with geography after removing the effects of the linguistic variables. When we combined ND and non-ND populations of Daghestan, no correlation of genetics with geography or languages was observed.

We also explored the effect of recent ancestry and migration on the association among genetic, linguistic and geographic distances employing genetic distances calculated on IBD sharing. Genetic distances were calculated on IBD segments > 3Mb (assuming common ancestry approximately 400 years ago) shared among ND populations. Genetic distances based on IBD sharing have a very strong positive correlation with languages and geography ($r = 0.516$, $p = 0$; $r = 0.480$, $p = 0$, respectively). Partial correlation remained significant with languages ($r = 0.306$, $p = 0.012$), but only marginally significant with geography ($r = 0.222$, $p = 0.059$) (**Table 2**).

Time of ND population divergence and TMRCA for J-M267* haplotypes

We calculated population divergence time T_F based on N_e and F_{ST} information. Inter-population T_F values represented in **Supplementary Table 15** were used to construct a neighbor-joining phylogenetic tree for indigenous populations from Daghestan, EUR, NEA, and CAS (**Figure 5**).

The tree provides clear separation of geographic groupings with ND populations as a distinctive cluster. The average TF estimates of the most diverged ND populations – Hinuq and Hunzib from the closest branch of the combined EUR and NEA populations – is ~6 (KYA). We also estimated the age of paragroup J-M267* in Daghestan. Though very rare outside Daghestan, paragroup J-M267* is by far the major Y-haplogroup in ND males. This paragroup can be likely associated with very early population movements into the Daghestani highlands. We constructed a network for haplogroup J-M267* in Daghestani populations (**Supplementary Figure 6**). The genealogy of the Y chromosome genetic pool shows a star-like pattern with an abundance of reticulations. This feature supports a demic expansion from ancestral haplotypes currently shared by people of Daghestan. We obtained the time estimate for the radiation of the paragroup J-M267* of 6,650 years (+/- 1,430 years) using ρ statistic and a mutation rate 6.8×10^{-5} [55].

DISCUSSION

Daghestan, particularly its mountainous area, is one of the few places on earth with exceptionally high linguistic density and diversity. Typical highland social structure is rooted in the mountainous topography and land scarcity[12]. The traditional economy was always dominated by sheep and goat pastoralism, terrace agriculture and horticulture. Daghestani clans (tukhums) typically consist of some 60-80 related families living in the same village. Small clans sometimes are united into larger settlements commonly on a linguistic or ethnic basis[12]. To keep land and property in the community, marriages traditionally arranged by families were usually clan- and village-endogamous. Both men and women inherited a portion of land as well as movable possessions. Small villages were typically associated with an adjacent town that was essentially a city-state with its own language, traditional constitution, customary law, and leadership[12,13]. Impoverishment of the highland economy and settlement in the lowlands began during the peak of the Little Ice Age (17th-18th centuries). Resettlement, both forced and economically driven, increased after the Russian conquest of the Caucasus in 19th century and, particularly, in the 20th century. However, relocation rarely led to assimilation or cultural amalgamation. Highlanders moving to lowland rural areas tended to remain compactly settled, and typically maintained close ties with their traditional villages and clans[58,59].

The working-age male population in highland Daghestan was transhumant, with men spending several months in lowland winter pastures or working in cities. Highlanders regularly learned lowland languages for economic purposes, and often also intervening foothill languages, but lowlanders rarely traveled uphill and almost never learned highland languages. As a consequence, language influence spread uphill, with highlanders sometimes shifting to lowland dialects or languages but almost never vice versa. This asymmetrical vertical bilingualism was universal in Daghestan[1,7, 8,56].

In this genetic study, which is the largest such study of Daghestani populations to date, we investigated biparental and uniparental genetic markers from 21 ethnic groups to measure the extent of genetic differentiation and isolation among Daghestani ethnic groups, as well as associations between genetic, linguistic, and geographic variation. We found that reduced genetic diversity and strong differentiation prevail among ND populations relative to non-ND and other continental groups for all genetic markers: autosomal, NRY and mtDNA (**Table 1**, **Supplementary Tables 10, 11, 12 and 13**). The large genetic distances among ND populations are also apparent in all three PC plots (**Figure 2a, b, c**). These results are consistent with study of autosomal STR data in small number of Daghestani populations (4 ND and 2 non-ND populations)[19]. Our recent work demonstrated that ND-speaking populations are characterized by exceptionally elevated coefficients of inbreeding, very high numbers and long lengths of Runs of Homozygosity, and elevated linkage disequilibrium compared with surrounding groups from the CAU, NEA, EUR, CAS and SAS[60]. It was also shown that inbreeding and long-standing small effective population sizes have most likely been a common feature in Daghestan over a sustained period. Consistent with long-term isolation we observed no signal of admixture in ND highland populations today (**Supplementary Figure 4, Supplementary Table 8**).

A previous study of Daghestani populations reported a reduction of genetic diversity in the NRY haplogroups among highland populations compared to mtDNA diversity, suggesting the effects of a patrilocal mating system[21]. However, under patrilocality the same pattern of reduced diversity would be expected for Y-STR haplotypes. Nevertheless, heterozygosity based on Y-STR haplotypes (0.911) proves to be only slightly and insignificantly lower than mtDNA heterozygosity (0.916) in highland ND populations. A high frequency of a single haplogroup J-M267(xL136) limits the accuracy of any parameters based on NRY haplogroups and explains drastic reduction of Y-chromosome diversity in Daghestan (**Supplementary Tables 4 and 11)**.

Unlike other populations in the Caucasus, marriages in highland Daghestan are traditionally endogamous along both parental descent lines and also by village and often by social class[12,19]. The marriages in highland Daghestan can be called patrilocal only at the family level because a bride moves to husband's house within the same village. Marriages with outsiders happened, but they were rare. F_{ST} and G_{ST} parameters reveal twofold higher variation for NRY STRs than for mtDNA among ND populations (**Table 1**). The larger variation among populations for the Y chromosome in many geographic regions including Caucasus is usually attributed to a higher female than male migration rate due to patrilocality[21,22,61]. On the other hand, males and females can differ not only in their pattern of migration, but also in their effective population sizes. Several studies suggest that sex-specific processes throughout paternal and maternal history indicate consistently larger effective population sizes for females than for males, which are roughly half that of females[62-64]. Higher variance of male reproductive success, existence of polygyny, and warfare with high male mortality rates may have produced different male *versus* female demographic histories in the Daghestani highlands. Thus, it is not surprising that no significant correlation was observed between the mtDNA and Y-chromosome distance matrices (**Supplementary Table 9**). Genetic structure on autosomal data is associated with male and female ancestry, although mtDNA data shows weaker evidence of correspondence (P=0.083).

Genetic, geographic and linguistic associations were previously investigated in the Caucasus based on NRY, mtDNA, and autosomal (Alu or STRs) data [15,16,21,22,24]. These data produced very inconsistent and controversial results. Some studies have shown that neither geography nor linguistics have had a strong influence on the genetic structure[23]. Geography, rather than language, was claimed to provide a better explanation for the observed genetic structure by the majority of studies [15,20,22,24], while parallel evolution of Y chromosome and language variation was supported by Balanovsky *et al.*[16]. We suggest that part of the explanation for these diverse conclusions came from differences in sampling schemes. Sampling of a few geographically sparse populations speaking languages from different linguistic families is not sensitive to the recent demographic history of the population and prevents successful fitting of linguistic and genetic structure. To model interactive historical processes such as the developmental cycles of villages, and language speciation, our genetic and linguistic sampling has focused at the community scale. We found no correlation between genetic and geographic distances (vertical, linear, or along car roads) (**Table 1, Supplementary Table 14).** The latter

result is expected given that isolation by distance can hardly be achieved in Daghestan due to high isolation and the extremely low gene flow between communities resulting from endogamous marriage rules and social structure in highland Daghestan. The observed correlation between geography and genetic distances calculated on the basis of IBD segments > 3Mb (and the assumption of common ancestry ~400 years ago) might reflect the beginning of forced resettlements of traditional highland villages[58,59].

We also did not observe a gene-language association when all Daghestani populations were taken into account (**Table 2**). However, we found a marginally statistically significant positive correlation between linguistic and genetic distances based on autosomal data (r = 0.343, P = 0.061) when only highland ND populations were considered. Human genetic and linguistic diversity can be correlated either through a direct link, when linguistic and genetic affiliations reflect the same historical population processes, or an indirect one, where the evolution of genetic and linguistic diversity is independent but conditioned by another factor (e.g., the same geographical factors). By controlling for geography, we can test for a residual relationship between linguistic and genetic affiliations[65]. Our finding of a stronger significant correlation between linguistic and genetic variation when geography was held constant (r = 0.428, P = 0.015) for ND populations provides evidence that language and genetic variation in highland Daghestan have actually evolved together.

To verify that the language-gene association could have emerged within the time frame since farming spread to pre-existing Mesolithic populations in Daghestan, we determined the divergence time for the Daghestani branch of the ND linguistic family based on autosomal and NRY data; our calculation is ~ 6-6.5 KYA, which is consistent with the timing of the Proto-Daghestanian language dispersal[1]. Interestingly, ~~that~~ despite the great age of the ND language family and some of the highland villages, the internal ages of the major branches of ND are generally shallow (**Figure 5**), probably reflecting the latest phase of uphill language spreading.

In summary, our study revealed that Daghestanian-speakers are most likely descendants of the earliest farming communities in the Caucasus. Proto-Nakh-Daghestanian appears to have diversified and taken root in the eastern Caucasus foothills and highlands as an early consequence of the initial spread of agriculture from Mesopotamia. Linguistic and genetic correlation are consistent with the scenario that most Daghestanian-speaking groups descend from a common ancestral population that spread into the Caucasus by demic diffusion with

subsequent relative sedentism and low levels of gene flow in the last few thousand years[65]. The combination of geography and endogamy in the highland Caucasus has produced a highly structured population exhibiting great linguistic diversity, with genetically isolated societies existing more or less autonomously on within a relatively small geographic territory.

Figure legends

Figure 1. Approximate geographic location of sampling sites. Of the 22 Daghestani populations sampled here, 16 speak unique languages that are separate branch of the Nakh-Daghestanian (ND) language family (in black circles), three speak languages that are closely related to the Turkic branch of Altaic language family (in black triangles), and three speak languages belonging to the Iranian language branch of the Indo-European language family (in open triangles). Chechens speak a language that belongs to Nakh branch of ND linguistic family (in grey circle). See **Table S1** for population codes.

Figure 2. MDS plot constructed on a) 107,079 autosomal SNPs, stress = 0.17, r = 0.95; b) 13 Y-chromosome STRs, stress = 0.16, r = 0.94; c) mtDNA SNPs from coding and HVS1 regions, stress 0.15, r = 0.94

Figure 3. ADMIXTURE plots. Clustering of 1,141 individuals (104,519 SNPs) assuming K3 to K7 clusters. Individuals are shown as vertical bars colored in ratio to their estimated ancestry within each cluster.

Figure 4. Phylogenetic trees based on a) language, b) autosomal SNP data. The numbers at the branches on the language tree are confidence values based on bootstrap method (N = 10,000).

Figure 5. Neighbor-joining tree (NJ) constructed on *TF* divergence time. *TF* is estimated from genetic distance classes 0.005 – 0.25 cM. Branch length is proportional to divergence times in thousand years ago. EUR, NEA, CAS and indigenous Daghestani populations were included in the NJ tree. NJ tree was generated using SplitsTree program.

Tables:

Table 1. Amova analyses

Table 2. Mantel tests

Supplementary Tables:

Supplementary Table 1. Geographic sampling location, population name, language affiliation, number of subjects and source of genotype data

Supplementary Table 2. Y chromosome markers used to determine Y chromosome haplogroups

Supplementary Table 3. Primer information, reference SNP ID and Y position for new polymorphic markers included in this work

Supplementary Table 4. Frequencies of Y-Chromosome Haplogroups

Supplementary Table 5. Frequencies of Y-Chromosome STRs

Supplementary Table 6. mtDNA mutations used to determine mtDNA haplogroups

Supplementary Table 7. Frequencies of major mtDNA haplogroups

Supplementary Table 8. The lowest f3 statistics for 56 populations

Supplementary Table 9. Correlation and partial correlation coefficients between genetic distances based on autosomal, Y-chromosome STRs and mtDNA distances

Supplementary Table 10. Genetic diversity indices based on autosomal SNPs frequencies in 56 populations

Supplementary Table 11. Genetic diversity indices based on Y chromosome haplogroup frequencies in 60 populations

Supplementary Table 12. Genetic diversity indices based on 12 Y-STR* haplotype frequencies in 60 populations

Supplementary Table 13. Genetic diversity indices based on mtDNA haplogroup and HVS1 frequencies in 55 populations

Supplementary Table 14. Correlation coefficients between genetic, linguistic and three measures of geographic distances in ND populations

Supplementary Table 15. TF divergence time estimates amomg populations from CAS, EUR, NEA and indigenous Daghestanian populations.

Supplementary Figures:

Supplementary Figure 1. Phylogenetic trees for Y-chromosome haplogroup C (a) and haplogroup J (b). Mutations shown in red define new branches on the phylogenetic tree.

Supplementary Figure 2. PCA plot on autosomal SNPs of 56 populations from Daghestan, Caucasus, Near East, Europe, Central Asia, and South Asia. 'Drop one in' procedure was used for analysis. PC1 and PC2 coordinates for each population were calculated as median coordinate values for individuals within one population.

Supplementary Figure 3. a) PCA analysis based on autosomal data for 18 Daghestanian populations PC1 and PC2, b) PC1 and PC3.

Supplementary Figure 4. A maximum likelihood tree and residual fit. a) no migration; b) with ten migration events.

Supplementary Figure 5. Genetic trees constructed on FST distances: a) Y-chromosome STR, b)mtDNA.

Supplementary Figure 6. Network for Y-chromosome haplogroup J-M267* in Daghestanian populations. Area is proportional to frequencies.

1. Nichols, J. Language dispersal from the Black Sea region. in *The Black Sea Flood Question: Changes in Coastline, Climate, and Human Settlement* (eds. Yanko-Holmbach, V., Gilbert, A.S., Panin, N. & Dolukhanov, P.M.) 775-796 (Springer, Dordrecht, 2007).
2. Gadzhiev, M.G., Davudov, O.M. & Shikhsaidov, S.M. *[The History of Dagestan]*, (Nauka, Moscow, 1996).
3. Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World*, (Oxford University Press, Oxford, 2012).
4. Kotovich, V.G. & Sheikhov, N.B. Arkheologicheskoe izuchenie Dagestana za 40 let (Archeological Study of Dagestan during Last 40 Years). in *Uchenye Zapiski IIYaL Dagestan Filial AN SSSR (IIYaL Scientific Transections Dagestan Filial Acad. Nauk SSSR*, Vol. 8 (IIYaL, Makhachkala, 1964).
5. Gadzhiev, M.G., Davudov, O.M. & Shikhsaidov, S.M. *Istoriya Dagestana (The History of Dagestan)*, (Nauka, Moscow, 1996).
6. Nichols, J. The Nakh-Daghestanian consonant correspondences. in *Current trends in Caucasian, East European, and Inner Asian linguistics: Papers in honor of Howard I. Aronson* (eds. Tuite, K. & Holisky, D.A.) 207-251 (Benjamins, Amsterdam, 2003).
7. Nichols, J. The origin of the Chechen and Ingush: A study in alpine linguistic geography. *Anthropological Linguistics* **46**, 129-155 (2005).
8. Nichols, J. Causativization and contact in Nakh-Daghestanian. in *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society: Special Session on Languages of the Caucasus* (eds. Cathcart, C., Kang, S. & Sandy, C.S.) 68-80 (Berkeley, Berkeley, 2013).

9. Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. & Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci U S A* **85**, 6002-6 (1988).
10. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597-603 (2003).
11. Bulayeva, K.B., Dubinin, N.P., Shamov, I.A., Isaichev, S.A. & Pavlova, T.A. [Population genetics of Dagestan highlanders]. *Genetika* **21**, 1749-58 (1985).
12. Aglarov, M.A. *[Rural society in Mountainous Daghestan in XVII - the beginning XIX centuary]* (Nauka, Moscow, 1988).
13. Aglarov, M.A. Ethnogenesis in terms of polystructural (federal) political layout in Daghestan. in *12th International Congress of Anthropological and Ethnological Sciences* (Zagreb, 1988).
14. Xing, J. *et al.* Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res* **19**, 815-25 (2009).
15. Yunusbayev, B. *et al.* The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol Biol Evol* **29**, 359-65 (2011).
16. Balanovsky, O. *et al.* Parallel evolution of genes and languages in the Caucasus region. *Mol Biol Evol* **28**, 2905-20 (2011).
17. Bulayeva, K. *et al.* Genetics and population history of Caucasus populations. *Hum Biol* **75**, 837-53 (2003a).
18. Bulayeva, K.B. *et al.* [Genetic subdivision of Dagestan ethnic populations]. *Genetika* **39**, 83-92 (2003b).
19. Bulayeva, K.B. *et al.* Ethnogenomic diversity of Caucasus, Daghestan. *Am J Hum Biol* **18**, 610-20 (2006).
20. Caciagli, L. *et al.* The key role of patrilineal inheritance in shaping the genetic variation of Dagestan highlanders. *J Hum Genet* **54**, 689-94 (2009).
21. Marchani, E.E., Watkins, W.S., Bulayeva, K., Harpending, H.C. & Jorde, L.B. Culture creates genetic structure in the Caucasus: autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. *BMC Genet* **9**, 47 (2008).
22. Nasidze, I. *et al.* Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Ann Hum Genet* **68**, 205-21 (2004).
23. Nasidze, I. *et al.* Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* **9**, 267-72 (2001).
24. Nasidze, I., Sarkisian, T., Kerimov, A. & Stoneking, M. Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. *Hum Genet* **112**, 255-61 (2003).
25. Tofanelli, S. *et al.* J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur J Hum Genet* **17**, 1520-4 (2009).
26. Karafet, T.M. *et al.* Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet* **64**, 817-31 (1999).
27. Hammer, M.F. *et al.* Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* **18**, 1189-203 (2001).
28. Bulayeva, K.B. *et al.* [Genetic-demographic study of mountain populations from Dagestan and their migrants to the lowlands. Comparison of basic parameters of fitness]. *Genetika* **31**, 1300-7 (1995a).

29. Karafet, T.M. *et al.* Extensive genome-wide autozygosity in the population isolates of Daghestan. *Eur J Hum Genet* (2015).
30. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
31. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1-27 (1964).
32. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-64 (2009).
33. Excoffier, L. & Lischer, H.E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**, 564-7 (2010).
34. Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).
35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
36. McEvoy, B.P., Powell, J.E., Goddard, M.E. & Visscher, P.M. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* **21**, 821-9 (2011).
37. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318-26 (2009).
38. Behar, D.M. *et al.* No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews. *Human Biology Open Access Pre-Prints. Paper 41* (2013).
39. Karafet, T.M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* **18**, 830-8 (2008).
40. Redd, A.J. *et al.* Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int* **130**, 97-111 (2002).
41. Dubut, V. *et al.* mtDNA polymorphisms in five French groups: importance of regional sampling. *Eur J Hum Genet* **12**, 293-300 (2004).
42. Irwin, J. *et al.* Mitochondrial control region sequences from northern Greece and Greek Cypriots. *Int J Legal Med* **122**, 87-9 (2008).
43. Karachanak, S. *et al.* Bulgarians vs the other European populations: a mitochondrial DNA perspective. *Int J Legal Med* **126**, 497-503 (2012).
44. Kloss-Brandstatter, A. *et al.* Somatic mutations throughout the entire mitochondrial genome are associated with elevated PSA levels in prostate cancer patients. *Am J Hum Genet* **87**, 802-12 (2010).
45. Ottoni, C. *et al.* Human mitochondrial DNA variation in Southern Italy. *Ann Hum Biol* **36**, 785-811 (2009).
46. Hedrick, P.W. A standardized genetic differentiation measure. *Evolution* **59**, 1633-8 (2005).
47. Rohlf, F.J. *NTSYS-pc: Numerical taxonomy and multivariate analysis system*, (Sekauket, New York, 1998).
48. Bandelt, H.J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37-48 (1999).
49. Cox, M.P. Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. *Hum Biol* **80**, 335-57 (2008).

50. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254-67 (2006).
51. Holman, D.W. *et al.* Explorations in automated language classification. *Folia Linguistica* **42**, 331-54 (2008).
52. Veeramah, K.R. *et al.* Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur J Hum Genet* **19**, 995-1001 (2011).
53. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-94 (2009).
54. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-13 (2014).
55. Zhivotovsky, L.A. *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* **74**, 50-61 (2004).
56. Lavrov, L.I. [Some results from the field work in Dagestan 1950-52]. *Kratkie soobschenija Instituta Ètnografii* **19**, 3-7 (1953).
57. Eldarov, E.D., Holland, E.C., Aliyev, S.M., Abdulagatov, Z.M. & Atayev, Z.V. Resettlement and Migration in Post-Soviet Dagestan. *Eurasian Geogr Econ* **48**, 226-48 (2007).
58. Karpov, J.J. & Kapustina, E.L. *Gorcy posle gor: Migracionnye processy v Dagestane v XX-nachale XXI vv.: ix social'nye i etnokul'turnye posledstvija i perspektivy. [Mountaineers down from the mountains: Migration processes in Daghestan, 20th and early 21st centuries: Social and ethnocultural consequences.]* (Rossijskaja AN, Muzej antropologii i etnografii, St. Petersburg, 2011).
59. Kurtsikidze, S. & Chikovani, V. *Ethnography and folklore of the Georgia-Chechnya border*, (LINCOM, Munich, 2009).
60. Karafet, T.M. *et al.* Extensive genome-wide autozygosity in the population isolates of Daghestan. *Eur J Hum Genet* (in press).
61. Seielstad, M.T., Minch, E. & Cavalli-Sforza, L.L. Genetic evidence for a higher female migration rate in humans. *Nat Genet* **20**, 278-80 (1998).
62. Wilder, J.A., Mobasher, Z. & Hammer, M.F. Genetic evidence for unequal effective population sizes of human females and males. *Mol Biol Evol* **21**, 2047-57 (2004).
63. Lippold, S. *et al.* Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig Genet* **5**, 13.
64. Balaresque, P. *et al.* Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. *Eur J Hum Genet* (2015).
65. Nettle, D. & Harriss, L. Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Hum Biol* **75**, 331-44 (2003).