6-4-2014

Human Paternal Lineages, Languages and Environment in the Caucasus

David Tarkhnishvili
1Center of Biodiversity Studies, Institute of Ecology, Ilia State University, Tbilisi, Georgia

Alexander Gavashelishvili
1Center of Biodiversity Studies, Institute of Ecology, Ilia State University, Tbilisi, Georgia

Marine Murtskhvaladze
1Center of Biodiversity Studies, Institute of Ecology, Ilia State University, Tbilisi, Georgia

Mariam Gabelaia
1Center of Biodiversity Studies, Institute of Ecology, Ilia State University, Tbilisi, Georgia

Gigi Tevzadze
Faculty of Arts and Sciences, Ilia State University, Tbilisi, Georgia

Human Paternal Lineages, Languages and Environment in the Caucasus

David Tarkhnishvili,[1] Alexander Gavashelishvili,[1] Marine Murtskhvaladze,[1] Mariam Gabelaia,[1] Gigi Tevzadze[2]

[1]Center of Biodiversity Studies, Institute of Ecology, Ilia State University, Tbilisi, Georgia.

[2] Faculty of Arts and Sciences, Ilia State University, Tbilisi, Georgia.

Suggested running head: "Y-DNA haplogroups and ecology in Caucasus"

Keywords: Y-DNA haplogroup, paternal lineage, Caucasus, Glacial Refugia, human ecology, landscape genetics, ethnogenesis, language

Acknowledgments: Author contributions DT and AG analyzed the data and wrote the text. MM coordinated molecular genetic work and provided primary data analysis; MG did much of genetic analysis and identification of the haplogroups; GT initiated human genetic studies at Ilia State University, discussed the results from the standpoint of history and social sciences, and made much effort to provide the research with all necessary resources.

Abstract. Publications that describe the human Y-DNA haplogroup composition in different ethnic or linguistic groups and geographic regions provide no explicit explanation of the distribution of human paternal lineages in relation to specific ecological conditions. Our research attempts to address this topic for the Caucasus – a geographic region that encompasses a relatively small area but harbors high linguistic, ethnic, and Y-DNA haplogroup diversity. 224 men that identified themselves as ethnic Georgian were genotyped for Y-chromosome 23 STR markers and assigned to their geographic places of origin. The genotyped data were supplemented with the published data on the haplogroup composition and location of the other ethnic groups of the Caucasus. We used multivariate statistical methods to see if linguistics, climate and landscape accounted for geographical difference in frequencies of the Y-DNA haplogroups G2, J2, R1b, J1, and R1a. The analysis showed significant associations of (1) haplogroup G2 with well forested mountains; (2) haplogroup J2 with warm areas or poorly forested mountains; (3) haplogroup J1 with poorly forested mountains. R1b showed no association with environment. Unlike haplogroups J1 and R1a significantly associated with Daghestanian and Kypchak speakers, respectively, the other haplogroups showed no such simple associations with languages. Climate and landscape in the context of competition over productive areas among different paternal lineages, arriving in the Caucasus in different times, have

played an important role in shaping the present-day spatial distribution of patrilineages in the Caucasus. This spatial pattern had formed before linguistic subdivisions were finally shaped, probably in Neolithic to Bronze Age. Later historical turmoil had little influence on the patrilineage composition and spatial distribution. Based on our results, the plausible scenario of post-glacial expansions of humans and their languages to the Caucasus from the Middle East, western Eurasia and the East European Plain is discussed.

Y-DNA haplogroup diversity is most commonly used for the analysis of the ancestry of individual ethnic groups or linguistic families (Kayser et al. 1997; Brisighelli, 2012). The reason is that Y-DNA haplogroups generally show more distinct ethno-geographic patterns than matrilineally inherited mt-DNA (Comas et al. 2000; Nasidze et al. 2003, 2004b). This is most likely due to higher dispersal rates of women (Seielstad et al. 1998; Oota et al. 2001; Nasidze et al. 2004a), the effects of selective pressures on the mitochondrial genome (Mishmar *et al*. 2003) and/or sex ratio in favor of women, causing more genetic drift in males (Dupanloup *et al*. 2003). Moreover, there is a popular nomenclature of the haplogroups linked to a well-established phylogenetic pattern (Underhill et al. 2001; Y Chromosome Consortium, 2002; Karafet et al. 2008; Chiaroni et al. 2009).

The Caucasus is among the most linguistically and culturally diverse regions of Eurasia (Comrie, 2008; Nasidze et al. 2004b; Marchani et al. 2008; Balanovsky et al. 2011; Yunusbayev et al. 2012). Currently, the region hosts dozens of languages that are grouped into three language families (Comrie, 2008): Caucasian, Indo-European and Turkic. The Caucasian language family traditionally includes Adyghean, Vainakh, Daghestanian, and Kartvelian languages (Catford, 1977), although the common origin of these languages is disputed (Starostin, 1989). Recent study by Pagel *et al.* (2013) shows that the Kartvelian and Dravidic language families are the most basal in relation to the other Eurasian language families. Diakonoff and Starostin (1988) suggest that Vainakh and Daghestanian (i.e. Northeast Caucasian languages) are related to extinct Hurro-Urartian. Armenian and Ossetian languages belong to the Indo-European language family, and Oghuz and Kypchak subgroups of the Turkic language group are spoken in the Caucasus as well (Catford, 1977; Comrie, 2008). Linguistic differences, along with the differences in political history, influence (but do not determine) the ethnic identities of the people inhabiting the region. Some ethnic boundaries (e.g. that of Armenians or Ossetians) coincide with the linguistic boundaries but those of the other ethnic groups only partly do so. Some groups speaking several mutually unintelligible but related languages consider themselves to be part of a single ethnos – e.g. Georgians or Avarians. Simultaneously, language rather than religion accounts for ethnic identity in the

Caucasus – e.g. Ossetians, Abkhazians and Georgians maintain ethnic integrity in spite of different religions practiced within each of these ethnic groups. Recent molecular genetic studies (Bulayeva et al. 2003; Yunusbayev et al. 2012) demonstrated that most of the Caucasian ethnic groups are more closely related to one another than to the neighboring populations of Western Eurasia. Genetic differences between populations of the Caucasus and the Eastern European Plain are much greater than those between the Caucasus and the Middle East.

The human population of the region predominantly consists of descendants of the patrilinear haplogroups G2, J2, J1, R1b, and R1a (Y Chromosome Consortium, 2002). Other haplogroups widely spread in Western Eurasia, including I2, E1b1b and L, are present but rare, and the rest are very rare (Nasidze et al. 2004b; Battaglia et al. 2008; Balanovsky et al. 2011; Yunusbayev et al. 2012). Haplogroup G2 dominates in the Western and Central Northern Caucasus and in Georgians (Balanovsky et al. 2011; Yunusbayev et al. 2012; Teuchezh et al. 2013). High frequency of the haplogroup R1b is found in Armenians (Battaglia et al. 2008; Yunusbayev et al. 2012; Herrera et al. 2012), and that of the haplogroup J1 occurs in Daghestan (Balanovsky et al. 2011; Yunusbayev et al. 2012). Relatively high frequencies of haplogroup R1a are found in Turkic-speaking peoples of the Northern Caucasus (Yunusbayev et al. 2012). High frequencies of the haplogroup J2 (over 20%), presumably having descended from the early agricultural populations of the Middle East (Cinnioğlu

et al. 2004; Battaglia et al. 2008; Grugni et al. 2012) are documented for the entire Western Asia and the Caucasus, peaking in Vainakh language speakers (Nasidze et al., 2004b).

In spite of the patterns described in previous paragraph, it's difficult to explicitly link the current distributions of paternal lineages to ethnicity or linguistics in the Caucasus probably due to phenomena of language or gene replacement.

In the spread of anatomically modern humans, many groups formed due to heterogeneous terrain and climate; limited migration among these groups - especially those stranded within glacial refugia - triggered the formation of distinct genetic lineages and languages (i.e. parallelism between genetic and linguistic evolution); e.g. the groups that survived a number of glacial cycles, expanding from and contracting into their refugia, evolved into distinct Y-chromosome populations (Underhill et al. 2001; Wei et al., 2012); over time during periods favorable for migration, genetic and cultural admixture caused full language or partial gene replacement in some of these groups, though the correlation between the trees of current genetic lineages and linguistic families generally remains high (Cavalli-Sforza, 1997). The formation of the current ethnic, linguistic and paternal layers of the Caucasus involved post-glacial human expansions into the region from different parts of Eurasia. Divergence among the paternal lineages found in the Caucasus happened before their expansions into the

Caucasus as a result of spatial isolation (i.e. isolation by distance) between 50 KYA and early Holocene (Wei et al., 2012). Some major genetic admixture events in the region happened 1000–1500 years ago (Hellenthal et al., 2014) and, hence, one should suggest that the correlation between the haplogroups and ethno-linguistic groups here was much higher than now until relatively recent historical time.

Variation in human adaptations to different environments is rarely taken into account in attempts to explain current genetic diversity of human populations. Of the ancient populations, spatially separated from one another during glacial periods at least until the end of the Last Glacial Maximum (hereafter, LGM), some were probably adapted to differential climatic and ecological conditions. Although humans easily adapt to completely new environments, it is likely that, when expanding into new areas during warm periods, they initially settled in the environments similar to those of their origin. For this reason, one can expect the composition and diversity of the human gene pool to vary with environment (i.e. climate, terrain, vegetation cover) within a geographic region so diverse in climate and geography as the Caucasus, but too small to cause isolation by distance among human populations given the high levels of human mobility and gregariousness. Here, we hypothesize that geographic gradients in combination with terrain and climate might influence the current distribution of the Y-DNA lineages in the Caucasus. In addition to linguistics we include ecological

environment as a predictor for the Y-DNA haplogroup distribution. To the best of our knowledge, this study is the first of its kind to explicitly consider Y-haplogroup distributions in relation to ecology and environment.

Materials and Methods

Sampling. To answer the question raised in our study, we collected hair or blood samples from 224 ethnically Georgian men throughout Georgia. Each man represented a unique exogamous clan/surname. Georgians living in Georgia have a high variety of patrilinearly inherited family names, most of which are associated with specific geographic areas (places of origin). Thus, we used the online database on Georgian family names and their places of origin (www.geogen.ge) to link our genetic data to geography. If the data on geographic origin were not available from the online *geogen* database, then we made geographic assignments based on (a) the information provided by the sampled person about his historic place of origin or (b) the area where the sample was taken from unless the person knew his place of origin. We supplied our dataset with the genetic profiles of 87 individuals from the Georgian DNA Project at Family Tree DNA (www.familytreedna.com/public/georgia/). These profiles are provided with surnames that we used to link each individual to a specific geographic area of origin.

Our dataset of patrilineages was supplemented with the published data on haplogroup composition in other ethnic groups of the Caucasus region, which belong to different linguistic families and maintain different ethnic identities. We used the proportions of Y-DNA haplogroups in 20 ethnic groups from Armenia (www.familytreedna.com/public/armeniadnaproject/), Turkey (Cinnioğlu et al. 2004), Azerbaijan (Nasidze et al. 2004b), the Northern Caucasus and Georgia (Yunusbayev et al. 2012) (Appendix 1).

DNA extraction, STR genotyping, and Y-DNA haplogroup identification. DNA was extracted from blood (65 samples) and hair (159 samples). For extraction procedures 200 μl of whole blood and 0.5–1 cm piece of base of the hair (10–12 follicles), was used. Extraction was performed using Qiagen DNeasy Blood and tissue kit following the manufacturer's recommendations (QIAGEN, Valencia, California, USA). All samples were genotyped for 23 loci, including DYS576, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385a/b, DYS456 and Y-GATA-H4. PCR was conducted with PowerPlex® Y23 System Amplification kit (Promega), recommended for similar studies by Davis *et al.* (2013). DNA samples were amplified according to the following protocol: in 10 μl total volume, with 4–6 μl template DNA, PowerPlex. Y23 5X Master Mix 2ml, PowerPlex. Y23 10X

Primer Pair Mix 1 μl. Thermal cycling was performed at 95 °C for 2 min, 30 cycles of 94 °C for 20 s, 61 °C for 1 min, 72 °C for 45 s, followed by final extension at 60 °C for 20 min. Amplicons were run on an ABI 3130 automated Genetic analyzer with Hi-di Formamide and CC5 Internal Lane Standard 500 Y23. Genotypes were screened using Genemapper v. 3.5 software package (Perkin- Elmer, Waltham, MA, USA). The updated recommendations of the DNA Commission of the International Society of Forensic Genetics for analysis of Y-STR systems were followed (Gusmão et al. 2006).

Y-STR haplotypes were grouped into Y-DNA "major" haplogroups as defined by the classification of Y chromosome consortium (2002), using two methods: (1) Athey's Bayesian approach to estimate the probability of assignment to a particular haplogroup (Athey, 2005, 2006) and (2) the haplogroup predictor (y-predictor) by Vadim Urasin (available at http://predictor.ydna.ru/). Both of the online calculators grouped our sample into haplogroups G2, J2, J1, R1b, R1a, L, I2, T, and E1b1b. It is shown that probability of wrong assignment of the haplogroups based on the online software could be high enough (Muzzio et al., 2011). For this reason, we tested the accuracy of the online calculators, using individual Y-STR profiles of eight Caucasian ethnic groups, Turks, Iranians, and Russians with SNP-typed Y-DNA haplogroup assignments (source: www.familytreedna.com/). 100 Y-STR profiles of each major patrilineage (G, J1, J2, R1a, R1b) were downloaded from the FamilyTree database. Haplogroups

inferred from the Y-STR profiles through both online calculators were checked against SNP-typed haplogroup assignments.

Location and calculation of Y-DNA haplogroup frequencies and environmental variables. To locate Y-DNA haplogroup proportions throughout the Caucasus, first we masked out unpopulated or poorly populated areas in the Caucasus--that is, areas above 2200 m asl or areas where annual rainfall was below 250 mm. We used (a) SRTM 90x90m digital elevation data (Jarvis et al. 2008) to mask terrain and (b) climatic data from WorldClim v. 1.4 (Hijmans et al. 2005) to mask rainfall. Then, we developed a dataset of Y-DNA haplogroup proportions for 38 populated areas (hereafter, population units) in the Caucasus (18 in Georgia and 20 in an area encompassing Armenia, Azerbaijan, NE Turkey, and the Northern Caucasus). In our dataset, the number of population units per ethnic group turned out to be highest for Georgians because obtaining genetic samples and relatively accurate information on the places of origin of clans (see above) was only possible in Georgia at the time of our study. The population units in Georgia were identified as the country's smallest administrative units (districts). We grouped our sample of places of origin into districts. If the sample within a district was too small (i.e. < 10 places of origin), then the district was merged with neighboring district(s). Each of the units formed this way consisted of 1 to 5 districts. Further merging neighboring population units would greatly decrease

geographic resolution for environmental predictors (see below). Besides, there was no need for a further increase in the minimum sample size because the Wilcoxon signed-rank test procedure via the software SPSS v. 21.0 (IBM corp., Armonk, NY, USA) indicated that a sample of 10 was as representative of haplogroup frequencies within each population unit as larger samples. For this procedure, we randomly selected ten individuals from each of the Georgian population units, recalculated haplogroup frequencies and compared the frequency distributions with those of the original dataset via the Wilcoxon test. We repeated randomly selecting 10 individuals and the test procedure three times in order to increase the confidence of the results. At every run the Wilcoxon tests showed no significant differences in haplogroup distributions between the original dataset and the dataset where sample size per population unit was reduced to 10 (Table S2).

The population units of the Northern Caucasus were separated based on the dominating ethno-linguistic groups, following the data of Yunusbayev *et al.* (2012) and ethnic map of the Caucasus (www.zonu.com). Two population units of NE Turkey were separated based on the publication of Cinnioğlu *et al.* (2004). Neither Armenia nor Azerbaijan was further subdivided, and northern parts of each of these countries were treated as a single population unit (Fig. 1). We calculated proportions of "major" Y-DNA haplogroups within each population unit, using our and existing genetic data (see above). We used this sample of 38

population units and Y-DNA haplogroups G2, J2, R1b, J1, and R1a for further analyses. We focused on these paternal lineages because they overwhelmingly dominated in our dataset (Appendix 1, Table S1). We managed and mapped the proportion of each of these five haplogroups in the population units using ArcGIS Desktop 9.3 software package (ESRI Inc., Redlands, CA, USA).

To check to see if linguistics and environment accounted for geographical difference in frequency of each of the Y-DNA haplogroups, we used dominant language or linguistic group (Fig. 1) and average values of three environmental predictors measured within each of the population units: Bailey's effective temperature (Bailey, 1960), slope, and percent of tree canopy cover. Bailey's effective temperature reflects the variability of the local climate and provides a simple comparative measure of sunlight, warmth, and the length of the growing season. Data to calculate Bailey's effective temperature were downloaded from WorldClim v.1.4 with a resolution of 1km^2 cells (Hijmans et al. 2005). We derived slope from a SRTM elevation grid of 90x90m cells (Jarvis et al. 2008). Percent of tree canopy cover was extracted from 500-m MODIS data 'MOD44B' (NASA: www.echo.nasa.gov/reverb/about_reverb.htm).

Canonical Correlation Analysis. Nonlinear canonical correlation analysis or OVERALS was conducted for exploring the associations among major Y-DNA haplogroups of the Caucasus region, geographic variations in landscape and

climate, and linguistic divisions within the region. OVERALS is a multivariate statistical method that helps to find the best-fit equations among more than two sets of variables that can be scaled as either nominal, ordinal, or numerical--e.g. Manly (2004) demonstrates this method in finding the relationship of frequencies of several allozyme alleles with multiple ecological variables. This method allows assigning scores to objects and categories of variables, which can be used to plot a geometrical representation of the dependencies in the data in a low dimensional Euclidean space. OVERALS is equivalent to (1) principal components analysis if each set contains one variable, (2) multiple correspondence analysis if each of these variables is multiple nominal, and (3) categorical multiple regression if two sets of variables are involved, and one of the sets contains only one variable.

We used the OVERALS procedure via the software SPSS v. 21.0 (IBM corp., Armonk, NY, USA). The fit and loss values, loadings and weights of individual variables and relative importance of the OVERALS dimensions were estimated as described in the user's manual of the software. Prior to the analyses, the variables were transformed according to the procedure's requirements. Haplogroup frequencies were treated as discrete numeric variables varying between 1 and 100; effective temperature, canopy cover and slope were converted into ordinal variables by dividing each into three equal-size intervals and categorizing these intervals as "low", "medium" and "high". We divided languages spoken in the Caucasus into 8 linguistic groups (KA, VA, DA, AD,

KY, OG, OS, AR, see Fig.1 for details). Each of these 8 linguistic groups was treated as an ordinal variable by assigning 2 to a linguistic group if dominant within a population unit or 1 if otherwise.

General linear modeling. We applied Multivariate General Linear Modeling (MGLM) for testing significance and power of the effect of the environmental variables and linguistic differences on the haplogroup composition in the Caucasus. We performed three runs of the analysis estimating the effect of (1) environment only, (2) linguistics only, and (3) both linguistics and environment on geographical difference in frequency of each of the Y-DNA haplogroups. The software used was IBM SPSS Statistics v.21 (IBM corp., Armonk, NY, USA).

Results

Accuracy of Y-DNA haplogroup predictors. Athey's (2005) calculator correctly predicted 83% (J1) to 98% (R1a) of the SNP-typed haplogroups (92.4% of the total number of the validated individuals), with the highest misidentification rates between J1 and J2 (Table 1). Y-predictor by Vadim Urasin correctly identified 90% (J) to 97% (R1a) of the SNP-typed haplogroups (93.4% of the total number of the validated individuals). 99.2% of those individuals that were assigned to the same haplogroup by both calculators were in agreement with

the SNP-typed assignments; hence the probability of misclassification did not exceed 1%. Thus, for further analyses we only used those Georgian Y-STR profiles, whose haplogroup assignments both Athey's and Urasin's predictors were in agreement on (Table S1), reducing the total sample size of Georgians (incl. the ones from the *familytreedna* database) from 311 to 295 men.

Distribution of human paternal lineages in the Caucasus.

Fig. 2 shows frequencies of paternal lineages G2, J2, R1b, J1, and R1a throughout the Caucasus. G2 reached the highest proportions (> 30%) in the western and central parts of the Greater Caucasus, peaking at 80% in Svan speakers of the Kartvelian linguistic group. J2 had the highest proportion in the eastern part of the Greater Caucasus, peaking at 82% in Ingush speakers of the Vainakh linguistic group, and was also common in the lowland areas both in the east and in the west of the region. Haplogroup R1b had considerably high frequencies (>25%) in the southern part of Georgia, in Armenia, and at the Caspian Sea Coast, peaking at 50% in Georgian speakers of the Kartvelian linguistic group in southern Georgia in close proximity to Armenia. At the national and ethnic level, R1b was most frequent in Armenia. R1a was relatively frequent in the north-western Greater Caucasus (peaking at 30% in the Kypchak linguistic group of the Turkic languages), while J1 dominated in the eastern Greater Caucasus, peaking at 92% in Darghin speakers of the Daghestanian linguistic group.

OVERALS ordination. The two first dimensions of OVERALS explained 92.4 % of the total variation in the data (Table 2). Fig. 3 shows ordination of the included variables along the first two axes. Variables, showing the highest loadings along the first axis, were haplogroup J1 having the highest frequencies in Daghestanian language speakers, and poorly forested areas. Variables showing the lowest loadings along both the first and the second axis were Y-DNA haplogroup R1a most associated with speakers of Kypchak subgroup of Turkic languages, and cold areas. Y-DNA haplogroup G2, dense forest and Kartvelian and Ossetian speakers from the mountains of the Central-Western Greater Caucasus were strongly associated with one another, having the lowest loadings along the first axis and intermediate loadings along the second axis (Fig. 3, left panel).

Removal of variables with the highest absolute ordination scores (J1, R1a, DA, and KY) from the analysis (Fig. 3, right panel) distinguished some additional groupings. High values along the 1st axis and low values along the 2nd axes (West Caucasus Mountains) identified high proportion of the haplogroup G2, speakers of Ossetian and Adyghean languages and Kartvelian speakers that inhabit well forested mountains. Speakers of the Vainakh languages as well as Kartvelian speakers from mountains of eastern Georgia (high values along both the 1st and the 2nd axis) were associated with a high proportion of the haplogroup J2, but not

with any specific environmental variable. The area with low values along the 1st axis (the Southern Caucasus) clustered a high proportion of the haplogroup R1b and medium slope with Armenian, Oghuz, and Georgian speakers from Armenia, Azerbaijan, southern Georgia and eastern lowland Georgia. This area is associated with high effective temperature and sparse canopy cover. There was a large cluster in the central part of the plot, not clearly associated with any of the Y-DNA haplogroups and occupied by populations with comparable proportions of the haplogroups G2, J2, and to less extent R1b. This cluster mostly encompassed lowland areas in Georgia.

Multivariate General Linear Modeling. Initially we visually checked frequencies of Y-DNA haplogroups against landscape types. This procedure suggested an obvious link of well forested *and* poorly or non-forested mountains to the distribution of some paternal lineages. Consequently, we used derivative variables such as the product of canopy cover and slope, and the product of [1-canopy cover] and slope, accounting for well forested mountains and poorly forested mountains, respectively. Models including these variables performed much better than those considering no interaction between slope and canopy cover. The outputs of MGLM are shown in Table 3. Frequencies of the major Y-DNA haplogroups, with the exception of R1b, were significantly associated not with individual linguistic groups but with different sets of these groups. The

frequency of G2 did not respond significantly to any individual linguistic group, although it was significantly linked to the Adyghean, Ossetian, and Kartvelian linguistic groups. The frequency of J2 was significantly associated with the Vainakh linguistic group. R1b, more common in Armenian than in the other linguistic groups, did not show significant association with any linguistic groups including Armenians. R1a was significantly associated with both Kypchak and Adyughean speakers. Only J1 was clearly associated with Daghestanian linguistic group.

Association with ecological conditions was significant for G2 and J1. The frequency of G2 significantly increased in areas where the portion of well forested mountains was dominant, and decreased in poorly forested mountains and warmer lowland areas. The frequency of J1 was associated with poorly forested mountains.

Running MGLM on both linguistic and environmental predictors slightly modified the outputs. G2 was significantly associated with either well forested mountains or the linguistic groups such as Adyghean or Ossetian. Patrilineage J2 had significant positive response to either effective temperature or poorly forested mountains and significant negative response to all linguistic groups but Vainakh. Multivariate effect of the predictors on the individual haplogroups remained significant for all major haplogroups except for R1b.

Discussion

In the 2000s, hundreds of publications appeared that describe the Y-DNA haplogroup composition in different ethnic or linguistic groups and geographic regions. However, there is lack of studies that would explicitly attempt to analyze human genotype distribution in relation to specific ecological conditions, and our research attempts to be pioneering in this respect. The analysis provided here suggests that there is significant impact of physical environment on the spatial distribution of patrilineages in the Caucasus. Ecological environment contributes to the paternal distribution pattern no less than the ethnic or linguistic boundaries and, hence we conclude that this pattern had formed before these subdivisions appeared, probably in Neolithic to Bronze Age. Later historical turmoil had less influence on the patrilineage composition and spatial distribution than it is traditionally thought.

Multiple studies conducted in the Caucasus (Nasidze et al. 2004a,b; Balanovsky et al. 2011, 2013; Yunusbayev et al. 2012) showed substantial differences in Y-DNA haplogroup proportions among the ethnic groups populating the region. Balanovsky *et al.* (2011) showed significant association among the patrilineal differences and linguistic differences in the Northern Caucasus. However, more comprehensive study of Yunusbayev *et al.* (2012), which covered both non-recombinant and autosomal markers, showed high

genetic similarity among all ethnic groups of the Caucasus and suggested that, in Western Asia, geography is far more important to genetic structure than linguistics. Our explicit findings are in line with this conclusion.

Paternal lineages, glacial refugia, and ethno-linguistic divisions in the Caucasus. Currently, Western Asian population, including Turkey, Iran, and the Caucasus, is largely composed of the following patrilineages: J2, J1, G2, R1b, R1a, E1b (www.eupedia.com; www. familytreedna.com). In general, it forms a distinct genetic cluster, closest to but different from European population (Nasidze et al. 2004b; Yunusbayev et al. 2012). Early split among the patrilineages G, JI (ancestral to J and I), and K (ancestral to R1) happened as early as 43,000–51,000 years ago (Wei et al. 2012)--i.e. shortly after the expansion of their common ancestral group F from Africa (Karafet et al. 2008). The genetic split must have been caused by limited or no gene flow among human refugia--that is, climatically suitable areas where humans survived during shorter 1,000–2,000 year long glacial episodes periodically taking place prior to the LGM (Clement and Peterson, 2006). The gene flow rate must have been at its lowest during the LGM. The final split of major patrilineages such as split between R1a and R1b most likely happened in Holocene (Wei et al. 2012) and could have been caused by dispersal rather than vicariance as it is traditionally defined (Mayr, 1970).

The expansion of people from the human refugia in post-LGM times (Banks et al. 2008) caused admixture among the patrilineages, which further increased after the development of early agriculture 10,000–9,000 years ago (Cavalli-Sforza, 1997; Pinhasy and Stock, 2011; Thomas et al. 2013). The expanding tribes were probably adapted to differential climatic conditions and used different technologies. Patrilineages that our study focuses on come from different refugia. Most researchers place the origin of the haplogroup J2 in the northern part of the Fertile Crescent (Battaglia et al. 2008)--that is, the location of the earliest Neolithic agrarian cultures (Diamond, 1997; Abbo, Lev-Yadun and Gopher, 2010). Both high frequencies and phylogenetic diversity of G2 in the Caucasus (Balanovsky et al. 2011) suggest that the Western Caucasus, well-known as a glacial refugium (Tarkhnishvili, Gavashelishvili and Mumladze, 2012), is indeed the ancestral area to this patrilineage. The place of origin of people of patrilineage R1b (most likely early speakers of the Western Indo-European languages (in sense of Renfrew, 1987) is either the Atlantic coast of Europe (Wilson et al. 2001; Busby et al. 2012) or the western part of Anatolia (Balaresque et al. 2010). R1a stemmed from grassland areas north of the Black Sea and the Caucasus (Keyser et al. 2009; Underhill et al. 2009), and J1 originated somewhere near the Caspian Sea Coast of Iran (Grugni et al.2012) or in the Zagros mountains.

The inferred pattern linking individual languages of the Caucasus and Y-DNA haplogroups with their geographic areas of origin is shown in Fig. 4. Ancestral areas of languages currently spoken in the Caucasus do not necessarily coincide with those of the paternal lineages. Paternal lineage J2 originates from the area where Hurro-Urartian languages were spoken in Bronze Age. These languages are related to Vainakh and Daghestanian languages (Diakonoff and Starostin, 1988) spoken by people of predominantly J (J2+J1) origin. Lineages R1b and R1a are usually associated with the western and the eastern Indo-European languages, respectively (Underhill et al. 2009; Balaresque et al. 2010; Thomas et al. 2013). Archaeologists suggest that the first Indo-Europeans expanded to the Southern Caucasus from the south-west in 3rd Millennium BC (Melikishvili, 1959; Melaart, 1970). Currently, relatively high proportion of the haplogroup R1b is found in Indo-European speaking Armenians, but also in Georgians from the areas south of the Lesser Caucasus (this study) and in some Daghestanian ethnic groups (Yunusbayev et al. 2012). The presence of the haplogroup R1a in the northern Caucasus is associated with Turkic-speakers that most likely descend from the inhabitants of the Eastern European Plain. Ancient DNA research suggests that this lineage was dominating throughout Eurasian steppe in Bronze Age (Keyser et al. 2009) and probably people of this lineage spoke Scytho-Sarmatian that later got replaced by Turkic and Slavic languages from Central Asia and Eastern Europe, respectively. Although Ossetian is a

language closely related to Scytho-Sarmatian (Lubotsky, 2002; Nasidze, 2003, 2004a), speakers of this language have a very high frequency of G2, not R1a. Adyghean languages spoken by G2-dominated people are linked to extinct languages of Anatolia (Ivanov, 1985; Kassian, 2010), spoken by ancient people of paternal lineage J2 that is rare in Adygheans. Thus, the languages spoken by present-day Ossetians and Adygheans, who genetically descend from the glacial-time population (patrilineage G2) of the Caucasus, have probably been adopted from paternally unrelated populations of the Eastern Europe and the Middle East, respectively. The Kartvelian and Dravidic language families hold the most basal position in a tree of Euroasiatic languages (Bomhard and Kerns 1994; Pagel et al. 2013). Y-DNA haplogroups G and H dominant in speakers of these two linguistic groups: Kartvelian (this study; Yunusbayev et al. 2012) and Dravidic (Sengupta et al. 2006), respectively, similarly hold the most basal position in a tree of patrilineages descending from superhaplogroup F widespread in Eurasia (Karafet et al. 2008). This fact may indicate correlated evolution of the G and H patrilineages and the Kartvelian and Dravidic languages, respectively. This logic suggests that the Kartvelian languages take origin from people dominated by G lineage.

Ecological associations of the haplogroups and interaction among the expanding populations. Association of patrilineage G2 with forested mountains

may be a result of both similarity of this landscape to the refugial area they survived the LGM, and competition with invading lineages from distant human refuguia in post-LGM times. Grasslands or sparsely forested areas provide a greater percentage of production accessible to hunter-gatherers (Kelly, 1983) and better starting conditions for agriculture than forests (Diamond, 1997). Expanding tribes (most likely dominated by haplogroups J1 and J2) that had survived the Ice Age south of the Caucasus probably started settling in the Caucasus both before emerging early agricultural settlements in the Fertile Crescent about 9.5 KY ago (Allaby et al. 2008) and after that. The earliest invaders could have been settled in mountain areas of the Eastern Caucasus, which were relatively distant from the LGM refugial area of the West Caucasus and probably less populated. The Neolithic or later invaders, already familiar with agricultural technologies, would prefer lowland, less forested areas. They could have been more successful (compared to the older settlers dominated by Y-DNA lineage G2) in populating the most productive lowlands of both the Eastern and the Western Caucasus. Less productive forested mountain areas remained dominated by the local tribes. The invaders turned out more successful in the most productive areas probably because they outnumbered the locals, possessed better weaponry or introduced contagious diseases to which the locals were less resistant. Later invaders (e.g. R1b from Western Europe or Western Anatolia and R1a from the Eastern European Plain) further increased competition in productive and agriculturally

suitable areas, making J2 and J1 concentrate in poorly forested or non-forested mountains currently inhabited by Vainakh and Daghestanian speakers in the eastern part of the Greater Caucasus.

Our results and reasoning are in line with that of Yunusbayev *et al.* (2012) who suggest that the core of the genetic structure of the Caucasian populations formed long before its present-day linguistic pattern. This particularly applies to the patrilineages of the earliest settlers, which still dominate in most of the region: G2 and J2. It appears that historical religious, linguistic and political expansions have had less influence on the current geographic distribution of the dominant paternal lineages in the Caucasus than territoriality established in pre-historical times in the context of local ecological adaptations.

Ethnogenetic processes that include major linguistic and cultural expansion occurred in the relatively recent past (Geary, 2002) and rarely caused full or substantial displacement of rural communities by invaders. Rural communities typically changed their identities and spoken languages as a result of political circumstances without changing preferred environments and paternal genetic structure. Our results are in line with such a vision. There is a clear geographic pattern for each of patrilineages G2 and J2, and R1b making up 75% of ethnic Georgian population. This pattern is associated neither with historical provinces of Georgia speaking different dialects and even different languages such as Megrelian and Svan, nor even with ethnic boundaries between Georgians

and Ossetians, or Georgians and Armenians. Frequency of patrilineage G2 significantly correlates with forested mountains, and its frequency declines in both eastern and western parts of what is usually called Transcaucasian Depression (Zimina, 1978)--that is, plains in both the Black and Caspian Sea basins, where agriculturally productive lands are concentrated and human population has been dense since early post-glacial times (Murtskhvaladze, Gavashelishvili and Tarkhnishvili, 2010). Proportion of G2 also declines in forestless mountains of the southern and the western Caucasus. Population of the Transcaucasian Depression is the most diverse patrilineally and comprises of comparable frequencies of G2, J2, and to less extent R1b and J1. Even though inhabitants of lowlands of western and eastern Georgia speak mutually unintelligible languages of the Kartvelian linguistic group (Megrelian and Georgian, respectively), they are paternally identical and dominated by lineage J2.

Away from the Caucasus, our modeled relationship between G2 frequency and environment seems to be applicable to at least two areas covered with forested mountains and characterized with a reasonably high rainfall level: in the Alps (Berger et al. 2013) and Iranian provinces of Gilan and Mazenderan at the southern Caspian coast (Grugni et al. 2012). In regions where haplogroup J2 is frequent (Battaglia et al. 2008; King et al. 2008; Myres et al. 2011), either poorly forested mountains or lowlands dominate, which also agrees with our model. Patrilineage J1, which is found in dry mountain areas of the Caucasus, is common

in even drier parts of the Middle East. Our models obtained for R1b and R1a are not transferable outside the Caucasus, which might suggest that these paternal lineages were the last of the five studied haplogroups to reach the Caucasus and had to adapt to new niches different from their places of origin.

Ethnogenesis in the southern Caucasus: general remark. It is likely that the formation of the major ethno-linguistic groups of the Southern Caucasus followed shaping of the current pattern of the major haplogroup distribution rather than preceded it. Historical records suggest that the first Georgian political state, comprising parts of the ancient states of Colchis and Iberia, emerged in the 3rd century BC at the latest (Suny, 1994; Rapp, 2003). This state probably expanded over several adjacent geographic areas with different proportions of the major paternal lineages J2, G2, R1b. Therefore, proto-Georgian ethnos included at least three genetically and ecologically distinct units with long-established economic and cultural interactions. This might have been reflected in writings of Strabo, who says "The plain of the Iberians is inhabited by people who are rather inclined to farming and to peace… but the major, or warlike, portion occupy the mountainous territory" (in: Suny, 1994). It is likely that the segregation of the mountain and lowland rural populations marked by different Y-DNA haplogroups was stronger in Strabo's time than now. It was recently shown that gene pool of present-day Georgians is a result of a major admixture event in 11th century that

involved, on one side, population genetically similar to the present-day inhabitants of the West Greater Caucasus, and, and on the other side, population genetically similar to the rest of West Asia (Hellenthal et al., 2014). This was the time of consolidation of Georgian state in Medieval time under the rule of kings of Bagratid dynasty (Suni, 1994).

Most likely, similar ethno-genetic processes took place in Armenia and current Azerbaijan, although the spatial-genetic structure of these countries was different due to different proportions of major landscape types.

Acknowledgements. The research was financed from the budget of Ilia State University and implemented at the Center of Biodiversity studies, established within the CoRE framework of a GRDF/GNSF joint project. MSc and BSc students Ardashel Latsusbaia, Giorgi Iankoshvili and Levan Kalatozishvili assisted in the processing of genetic data.

Literature Cited

Abbo, S., S. Lev-Yadun and A. Gopher. 2010. Agricultural origins, centers and noncenters; a near eastern reappraisal. *Critic. Rev. Plant. Sci.* 29:317–328.

Allaby, R.G., D.Q. Fuller, and T.A. Brown. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl. Acad. Sci. USA* 105:13982–13986.

Athey, T.W. 2005. Haplogroup prediction from Y-STR values using an allele-frequency approach. *J. Genet Genealogy* 1:1–7.

Athey, T.W. 2006. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *J. Genet. Genealogy*, 2, 34 –39.

Bailey, H.P. 1960. A Method of Determining the Warmth and Temperateness of Climate. *Geografiska Annaler* 42:1–16.

Balanovsky, O., Kh. Dibirova, A. Dybo et al. 2011. Parallel Evolution of Genes and Languages in the Caucasus Region. *Mol. Biol. Evol.* 28:2905–2920.

Balaresque, P., G. Bowden, S. Adams et al. 2010. A predominantly Neolithic origin for European paternal lineages. *PLoS Biol*. 8: e1000285.

Banks, W., F. D'errico, A. Peterson et al. 2008. Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. *J. Archaeol. Sci.* 35:481–491.

Battaglia, V., Fornarino, S., Al-Zahery, N. et al. 2008. Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. *Eur. J. Hum. Genet.* 17:820–830.

Berger, B., Niederstatter, H., Erhart, D. et al. 2013. High resolution mapping of Y haplogroup G in Tyrol Austria. *Forensic Science International, Genetics* 7:529–536.

Bomhard, A., and J. Kerns. 1994. *The Nostratic Macrofamily*. Mouton de Gruyter, Amsterdam,The Netherlands.

Brisighelli, F. 2012. *Genetic analysis of uniparental and autosomal markers in human populations*. Universidade de Santiago de Compostela, Facultade de Medicina e Odontoloxía Instituto de Ciencias Forenses “Luís Concheiro”.

Bulayeva, K., L.B. Jorde, C. Ostler et al. 2003. Genetics and population history of Caucasus populations. *Hum. Biol.* 75:837–853.

Busby, G., F. Brisighelli, P. Sanchez-Diz et al. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc. Royal. Soc. B: Biol. Sci.* 279:884–892.

Catford, J.C. 1977. Mountain of tongues. the languages of the Caucasus. Ann Rev Anthropol 6:283–31.

Cavalli-Sforza, L.L. 1997. Genes, peoples, and languages. *Proc. Natl. Acad. Sci. USA* 94:7719–7724.

Chiaroni, J., P.A. Underhill, and L.L. Cavalli-Sforza. 2009. Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc. Natl. Acad. Sci. USA* 48:20174–20179.

Cinnioğlu, C., R. King, T. Kivisild et al. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* 114:127–148.

Clement, A.C. and L.C. Peterson. 2008. Mechanisms of abrupt climate change of the last glacial period. *Rev. Geophys.* 46, 1–39.

Comas, D., F. Calafell, N. Bendukidze et al. 2000. Georgian and Kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *Amer. J. Phys. Anthropol.* 112:5–16.

Comrie, B. 2008. Linguistic Diversity in the Caucasus. *Ann. Rev. Anthropol.* 37:131–143.

Davis, C., J. Ge, C. Sprecher et al. 2013. Prototype PowerPlex1 Y23 System. A concordance study. Forensic Science International: *Genetics* 7:204–208.

Diamond, J. 1997. Location, location, location. the first farmers. *Science* 278:1243–1244.

Diakonoff, I.M. and S.A. Starostin. 1986. Hurro-Urartian as an Eastern Caucasian Language. *Münchener Studien zur Sprachwissenschaft*. 12. Munich.

Donnelly, P., and S. Tavaré. 1986. The ages of alleles and a coalescent. *Advances in Applied Probability* 18:1–19.

Dupanloup, I., Pereira, L., Bertorelle, G., Calafell, F., Prata, M.J., Amorim, A., and Barbujani, G. 2003. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J. Mol. Evol. 57(1):85–97.

Geary, P.J. 2002. *The Myth of Nations: The Medieval Origins of Europe*. Princeton University Press, Oxford.

Grugni, V., V. Battaglia, B. Hooshiar Kashani 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

Gusmão, L., J.M. Butler, A. Carracedo et al. 2006. DNA Commission of the International Society of Forensic Genetics. Commission of the International Society of Forensic Genetics ISFG.. an update of the recommendations on the use of Y-STRs in forensic analysis. *Internat. J. Legal. Med.* 120:191–200.

Hellenthal, G., G.B.J. Bushby, G. Band, J.F. Wilson et al. 2014. A genetic atlas of human admixture history. *Science* 343: 747–751.

Herrera, K.J., R.K. Lowery, L. Hadden et al. 2012. Neolithic patrilineal signals indicate that the Armenian plateau was repopulated by agriculturalists. *Eur. J. Hum. Genet.* 20:313–320.

Hijmans, R.J., S.E. Cameron, J.L. Parra et al. 2005. Very high resolution interpolated climate surfaces for global land areas. *Internat. J. Climatol.* 25:1965–1978.

Ivanov, V.V. 1985. *Ob otnoshenii khattskogo jazyka k severozapadnokavkazskim* [On the relations of Hatti and North-West Caucasian languages]. pp. 26–59 In. Piotrovsky, B.B. et al. eds., Drevnjaya Anatolia Ancient Anatolia. Moscow, Nauka in Russian.

Kassian, A. 2010. Hattic as a Sino-Caucasian Language. pp. 309–448 In. I*nternationales Jahrbuch für die Altertumskunde Syrien-Palästinas Herausgegeben von Manfried Dietrich and Oswald Loretz*. Band 41. Ugarit-Verlag, Münster.

Kayser, M., P. de Knijff, P. Dieltjes et al. 1997. Applications of microsatellite-based Y-chromosome haplotyping. *Electrophoresis* 18:1602–1607.

Keyser, C., C. Bouakaze, E. Crubézy et al. 2009. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum. Genet.* 126:395–410.

Karafet, T.M., F.L. Mendez, M.B. Meilerman et al. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838.

Kelly, R.L. 1983. Hunter-Gatherer Mobility Strategies. *J. Anthropol. Res.* 39:277–306.

King, R.J., S. Ozcan, T. Carter, E. Kalfoglu, and S. Atasoy. 2008. Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. *Ann. Hum. Genet.* 72:205– 214.

Jarvis, A., H.I. Reuter, A. Nelson, and E. Guevara. 2008. *Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture CIAT*., available from http.//srtm.csi.cgiar.org. Downloaded 20 Nov. 2013.

Lubotsky, A. 2002. Scythian elements in Old Iranian. *Proc. Brit. Acad.* 116:189–202.

Manly, B.F.J. 2004. *Multivariate Statistical Methods: A Primer, Third Edition*. Chapman and Hall, NY.

Marchani, E.E., W.S. Watkins, K. Bulayeva et al. 2008. Culture creates genetic structure in the Caucasus. autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. *BMC Genetics* 9:47.

Masson, V.M., N.Y. Merpert, R.M. Munchaev et al. 1994. 4 co-authors. Chalcolithic of the USSR. Series *Archaeology of the USSR*, B.A. Rybakov, ed. Moscow, Nauka.

Mayr, E. 1970. *Populations, Species, and Evolution*. The Belknap Press of Harvard University Press, Cambridge, MA.

Melaart, J. 1970. The Earliest Settlements in Western Asia from the Ninth to the end of the Fifth Millenium B.C. In. *The Cambridge Ancient History, 3rd ed*. vol. 1. Cambridge, Cambridge University press.

Melikishvili, G.A. 1959. *K Istorii Drevnei Gruzii* [on the history of Ancient Georgia]. Tbilisi, Metsniereba in Russian.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, et al. 2003. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* 100:171–176.

Muzzio M, Ramallo V, Motti JMB, Santos MR, López Camelo JS, and G Bailliet. 2011. Software for Y-haplogroup predictions: a word of caution. Intl. J. Legal Med. 125:143–147.

Murtskhvaladze, M., A. Gavashelishvili, and D. Tarkhnishvili. 2010. Geographic and genetic boundaries of brown bear Ursus arctos. population in the Caucasus. *Mol. Ecol.* 19:1829–1841.

Myres, N.M., S. Rootsi, A.A. Lin et al. 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* 19:95–101.

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

Nasidze, I., D. Quinque, I. Dupanloup et al. 2004a. Genetic evidence concerning the origins of South and North Ossetians. *Ann. Hum. Genet.* 68:588–599.

Nasidze, I., E.Y. Ling, D. Quinque et al. 2004b. Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Ann. Hum. Genet.* 68:205–221.

Oota, H., W. Settheetham-Ishida, D. Tiwawech et al. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics* 29:20–21.

Oppenheimer. S. 2006. *The Origins of the British: A Genetic Detective Story*. Constable, London.

Pagel. M., Q.D. Atkinson, A.S. Calude, and A. Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proc. Natl. Acad. Sci. USA* 110:8471–8476.

Pinhasi, R. and J. Stock. 2011. *Human Bioarchaeology of the Transition to Agriculture*. John Wiley and Sons LTD, Chichester.

Rapp, S.H. 2003. *Studies In Medieval Georgian Historiography: Early Texts And Eurasian Contexts*. Peeters Bvba ISBN 90-429-1318-5.

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge University Press, Cambridge.

Seielstad, M.T., E. Minch, and L.L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20:278–280.

Sengupta, S., L. Zhivotovsky, R. King et al. 2006. Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists. *Amer. J. Hum. Genet.* 78:202–221.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47:264–279.

Starostin, S. 1989. *Nostratic and Sino-Caucasian Explorations in Language Macrofamilies*. pp. 42–,67, Universitaetsverlag Dr. Norbert Brockmeyer, Bochum.

Suny, R.G. 1994. *The Making of the Georgian Nation, 2nd edn*. Indiana University Press, Bloomington.

Tarkhnishvili, D., A. Gavashelishvili, and L. Mumladze. 2011. Palaeoclimatic models help to understand current distribution of Caucasian forest species. *Biol. J. Linn. Soc.* 105:231–248.

Teuchezh, I.E., E.A. Pocheshkova, R.A. Skhalyakho et al. 2013. Gene pools of Abkhaz-Adyghe, Georgian and Armenian populations in their Eurasian context. *Vestnik Moskovskogo Universiteta. Antropologia* 2/2013:49–62.

Thomas, M.G., T. Kivisild, L. Chikhi, and J. Burger. 2013. Europe and western Asia. genetics and population history . *The Encyclopedia of Global Human Migration* ed. by I. Ness., pp. 1–11. Blackwell.

Underhill, P.A., G. Passarino, A.A. Lin, P. Shen, M. Mirazon Lahr, R.A. Foley, P.J. Oefner and L.L. Cavalli-Sforza et al. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* 65:43–62.

Underhill, P.A., N.M. Myres, S. Rootsi et al. 2009. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a". *Eur. J. Hum. Genet.* 18:479–484.

Wei, W., Q. Ayub, Y. Chen et al. 2012. A calibrated human Y-chromosomal phylogeny based on re-sequencing. *Genome Res.* 18:830–838.

Wilson, J., D. Weiss, M. Richards et al. 2001. Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc. Natl. Acad. Sci. USA* 98, 5078–5083.

Y Chromosome Consortium. 2002. A nomenclature system for the tree of human Y chromosomal binary haplogroups. *Genome Res*. 12:339–348.

Young, K.L., G. Sun, R. Deka, and M.H. Crawford. 2011. Paternal Genetic History of the Basque Population of Spain. *Human Biology* 83:455–475.

Yunusbayev, B., M. Metspalu, M. Jarve et al. 2012. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* 29:359–365.

Zimina, R.P. 1978. The Main Features of the Caucasian Natural Landscapes and Their Conservation, USSR. *Arctic and Alpine Research* 10:479–488.

Table 1. Accuracy of STR-based Y-DNA haplogroup predictors, checked against SNP-typed Y-DNA haplogroup assignments (N=100 for each individual haplogroup).

Haplogroups	Athey's calculator, % correct	Urasin's calculator, % correct	Both calculators in agreement, % correct
G2	94	96	100
J2	96	90	99
R1b	91	94	99
J1	83	90	98
R1a	98	97	100
All	92.4	93.4	99.2

Table 2. Output of the Nonlinear Canonical Correlation Analysis, exploring association among the Y-DNA haplogroups, ecological conditions, and linguistic units in the Caucasus *.

		D1a	D2a	sum_a	D1b	D2b	sum_b
Loss	Set 1	0.033	0.028	0.062	0.054	0.124	0.178
	Set 2	0.089	0.168	0.257	0.181	0.268	0.450
	Set 3	0.049	0.089	0.138	0.103	0.093	0.195
	Mean	0.057	0.095	0.152	0.113	0.162	0.274
Eigenvalue		0.943	0.905		0.887	0.838	
Fit				1.848			1.726
set	variable	Load_D1a	Load_D2a		Load_D1b	Load_D2b	
1	G2	-0.524	-0.291		0.487	-0.534	
	J2	-0.163	0.609		0.501	0.672	
	R1b	0.334	-0.231		-0.582	0.461	
	J1	0.651	0.167				
	R1a	-0.256	-0.685				
	X	-0.138	-0.107		0.026	-0.050	
2	Can1	0.529	0.304		-0.570	-0.041	

	Can2	-0.090	-0.215	0.255	0.597
	Can3	-0.429	-0.295	0.287	-0.62
	ET1	-0.356	0.354	0.063	-0.093
	ET2	0.253	-0.071	0.225	0.147
	ET3	0.115	0.047	-0.332	-0.064
	slope1	-0.228	-0.396	0.219	-0.039
	slope2	0.267	0.414	-0.573	0.011
	slope3	-0.062	-0.088	0.430	0.030
	AD	-0.029	0.357	0.112	-0.113
	VA	-0.198	0.177	0.399	0.216
	DA	0.732	-0.093	0.269	-0.253
	OS	-0.330	-0.141	-0.195	0.006
3	AR	0.055	-0.042	-0.126	-0.091
	OG	0.053	-0.570	-0.149	-0.031
	KY	-0.156	-0.019	-0.017	0.274
	KA1	0.063	0.127	-0.051	-0.346
	KA2	0.281	-0.281	0.299	0.641

KA3	-0.307	0.481	0.129	0.027
KA4	0.048	0.077	-0.069	0.174
KA5	-0.075	-0.157	-0.100	-0.042
KA6	-0.014	0.042	0.083	-0.087
KA7	-0.027	0.040	-0.005	0.010
KA8	0.042	0.087	0.041	-0.030
KA9	-0.099	-0.115	0.411	-0.414
KA10	0.023	0.038	0.112	-0.113
KA11	-0.340	0.033	0.399	0.216

*Upper panel: Summary of the OVERALS analysis. Set 1 – Y-DNA haplogroup frequencies; X – haplogroups that do not belong to the target five haplogroups. Set 2 – environmental variables: effective temperature (ET), average slope (slope), percent of tree canopy cover (Can); Set 3 – linguistic groups. Lower panel: loadings of individual variables along the first 2 dimensions. D1, D2 – analysis dimensions. a – the analysis run for the entire dataset, b – the analysis run with some variables (R1a, J1, DA and KY) excluded from the dataset. Y-DNA haplogroups: G2, J2, R1b, R1a, J1 (see text for details). Environmental conditions: ET1 – 10-12^{0}C; ET2 – 12-13^{0}C; ET3 – 13^{0}C; slope1 – < 10^0; slope2 – 10-20^0; slope3 - >20^0; Can1 – < 20%; Can2 – 20-40%; Can3 - >40%. Languages: DA – Daghestanian; VA – Vainakh; AD – Adyghean; AR – Armenian; OS – Ossetian; KY – Turkic Kypchak from the Northern Caucasus; OG – Oghuz (Azeris and Turks); KA – Kartvelian (KA1 – lowlands of eastern Georgia; KA2 – low mountains of eastern Georgia;

KA3 – mountainous forests of Eastern Georgia; KA4 – uplands of Eastern Georgia; KA5 – floodplain areas of eastern Georgia; KA6 – uplands of southern Georgia; KA7 – lowlands of Western Georgia, Georgian-speakers KA8 – mountainous forests of western Georgia, Georgian-speakers; KA9 – lowlands of Western Georgia, Megrelian-speakers; KA10 – forest mountains of Western Georgia, Megrelian speakers; KA11 – uplands of Western Georgia, Svan speakers).

Table 3.The outputs of MGLM analysis, linking frequencies of the Y-DNA haplogroups to (1) environment only, (2) linguistics only, and (3) both linguistics and environment. Effective temperature (ET), well forested mountains (FM = canopy cover * slope), poorly forested mountains (NFM = [1 - canopy cover] * slope), R^2 – variation in the haplogroup frequency explained by the model; AD – Adyghean, AR – Armenian, DA – Daghestanian, KA – Kartvelian, KY – Kypchak, OG – Oghuz, OS – Ossetian, VA – Vainakh. Significant (P<0.05) coefficients are shaded and in boldface.

		G2	***J2***	***R1b***	***J1***	***R1a***
(1)Environment	INTERCEPT	**40.043**	21.974	10.118	3.637	13.537
	ET	0.011	-0.004	0.020	-0.008	-0.023
	FM	**0.033**	0.004	-0.006	**-0.025**	-0.003
	NFM	**-0.033**	0.002	0.001	**0.027**	-0.002
	R^2	0.419	0.020	0.082	0.312	0.060
(2)Linguistic groups	INTERCEPT	32.535	25.302	10.698	11.953	7.302
	AD	**18.873**	-5.824	-6.332	**-9.942**	**8.301**
	AR	-15.127	-2.224	14.268	-1.542	-4.699
	DA	**-22.377**	**-22.224**	5.018	**51.208**	**-5.699**
	KA	**11.301**	5.490	-.923	**-10.970**	**-5.509**
	KY	2.473	-10.624	2.868	-5.542	**18.101**
	OG	-11.794	.776	2.601	-4.542	0.301
	OS	**32.873**	-5.724	-7.732	**-13.042**	-4.699
	VA	**-22.627**	**40.276**	-9.732	-2.042	-5.699
	R^2	0.479	0.525	0.211	0.858	0.779
(3)	INTERCEPT	**32.535**	**25.302**	10.698	11.953	7.302

ET	-0.718	**1.125**	-.0277	0.046	-0.271
FM	**0.025**	-0.007	-0.005	-0.002	-0.004
NFM	**-0.038**	**0.041**	-0.010	-0.002	-0.002
AD	**21.796**	-13.775	-4.972	-11.452	**12.329**
AR	1.271	-19.121	17.073	-2.293	-2.706
DA	10.113	**-60.639**	14.131	**51.250**	-1.243
KA	17.391	-14.025	6.675	-10.355	0.088
KY	9.876	**-24.322**	6.977	-5.647	**21.418**
OG	0.598	**-22.879**	9.935	-4.564	6.006
OS	**39.986**	**-19.689**	-3.249	**-12.940**	-1.378
VA	-9.503	**23.429**	-6.606	-3.243	-1.987
R^2	0.659	0.689	0.343	0.866	0.803

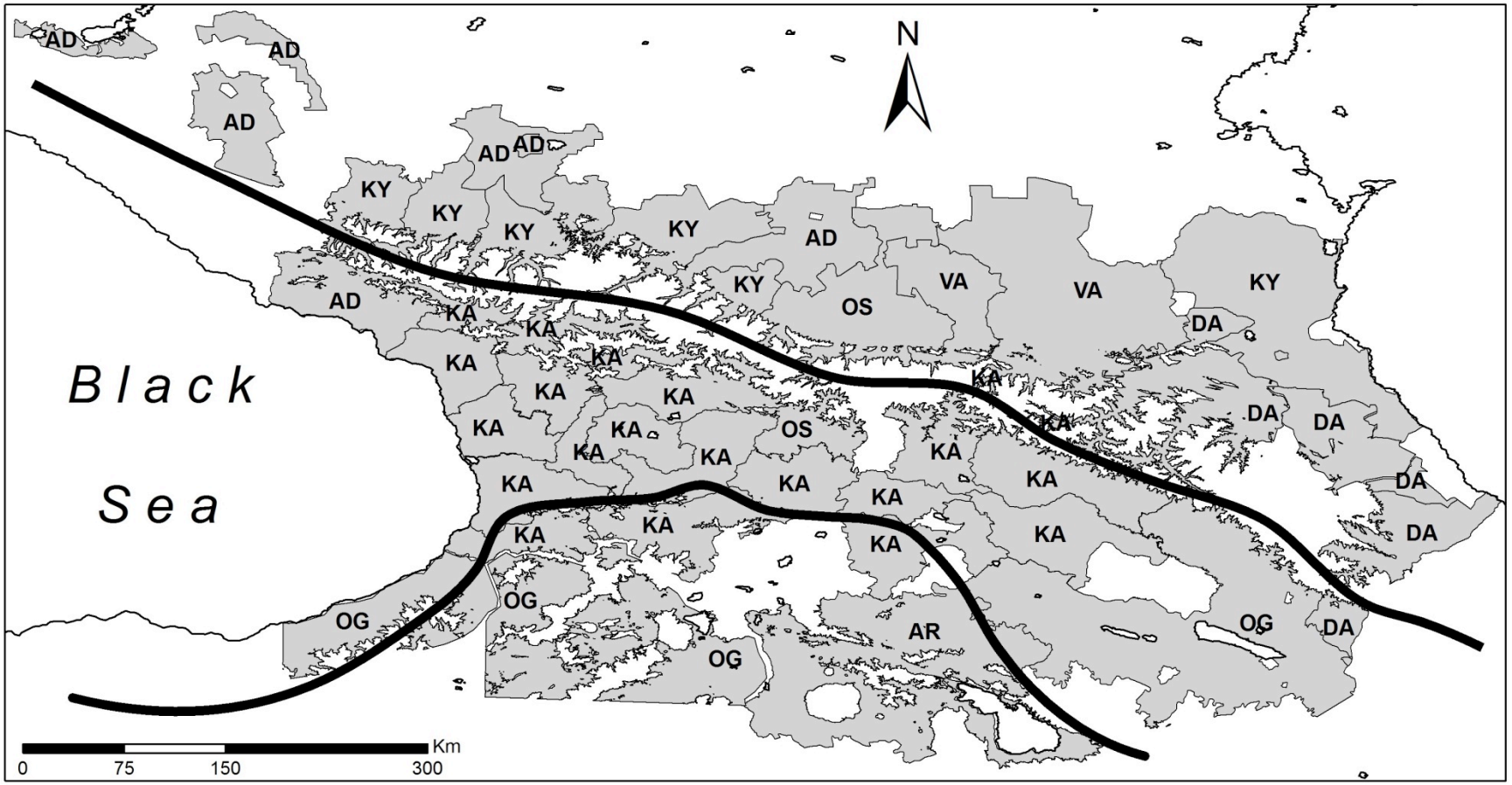

Fig. 1. The study area (the Caucasus) showing population units (gray polygons), no data or unpopulated areas (white polygons) and abbreviations of individual polygons that indicate linguistic groups: KA – Kartvelian (Georgian, Megrelian, Svan); AD – Adyghean (Circassian, Kabardin, Abkhazian etc); VA – Vainakh (Chechen, Ingush); DA – Daghestanian (Avar, Lezgi, Darghin, etc.); OS – Ossetian (eastern Indo-European); AR – Armenian (basal or western Indo-European); OG – Turkic Oghuz subgroup; KY – Turkic Kypchak subgroup.Upper thick line – the Greater Caucasus; lower thick line – the Lesser Caucasus. All maps shown in this manuscript have been projected to Mollweide; False Easting: 0; False Northing: 0; Central Meridian: 45; WGS: 1984.

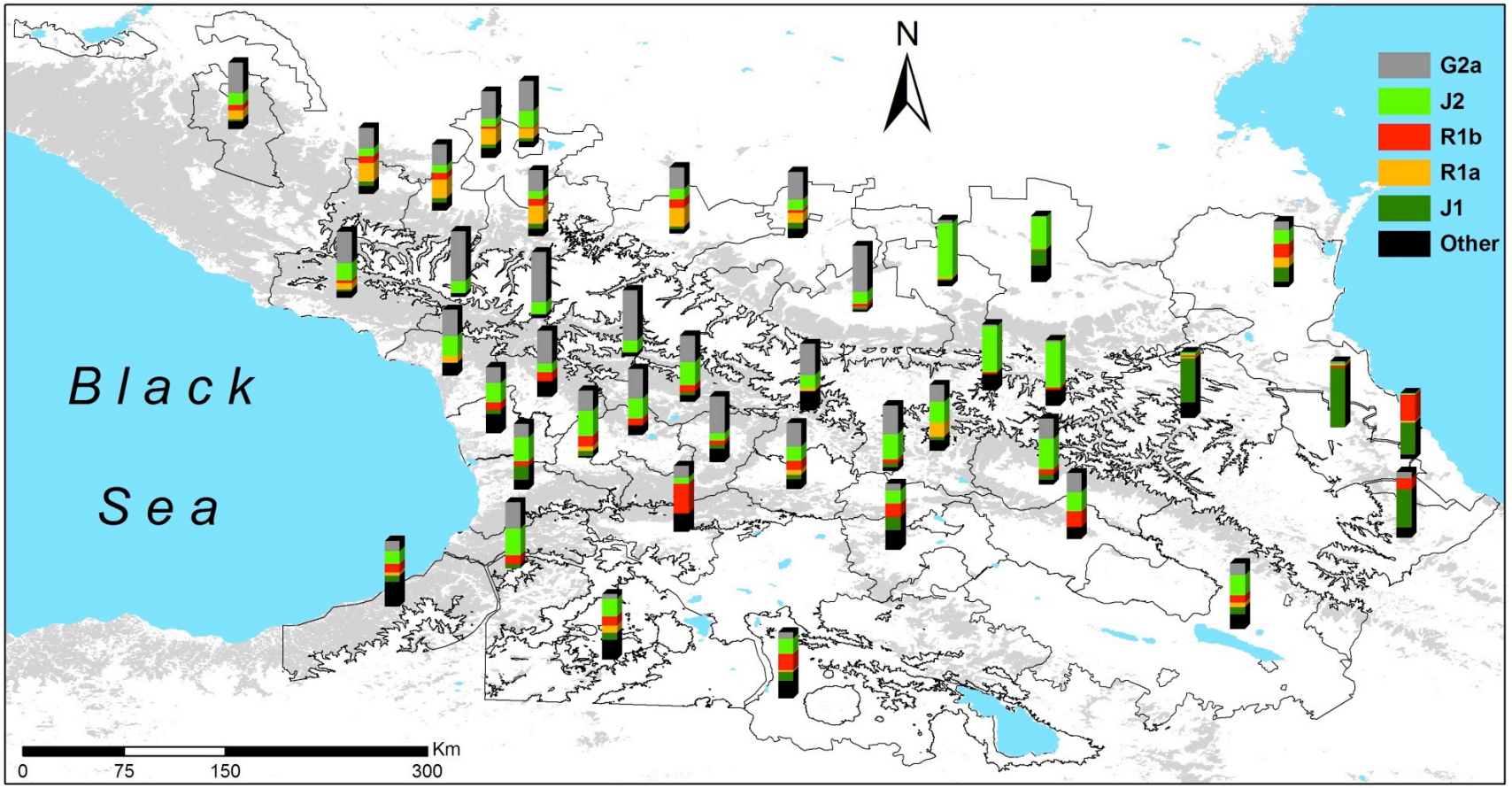

Fig. 2. Geographic distribution of major Y-DNA haplogroups in the Caucasus. The figure shows forest cover (shaded areas) and bars of haplogroup frequencies in the population units: gray sections = G2, green = J2, red = R1b, yellow = R1a, dark green = J1, black = other haplogroups: X=I2, L, E1b1b,T.

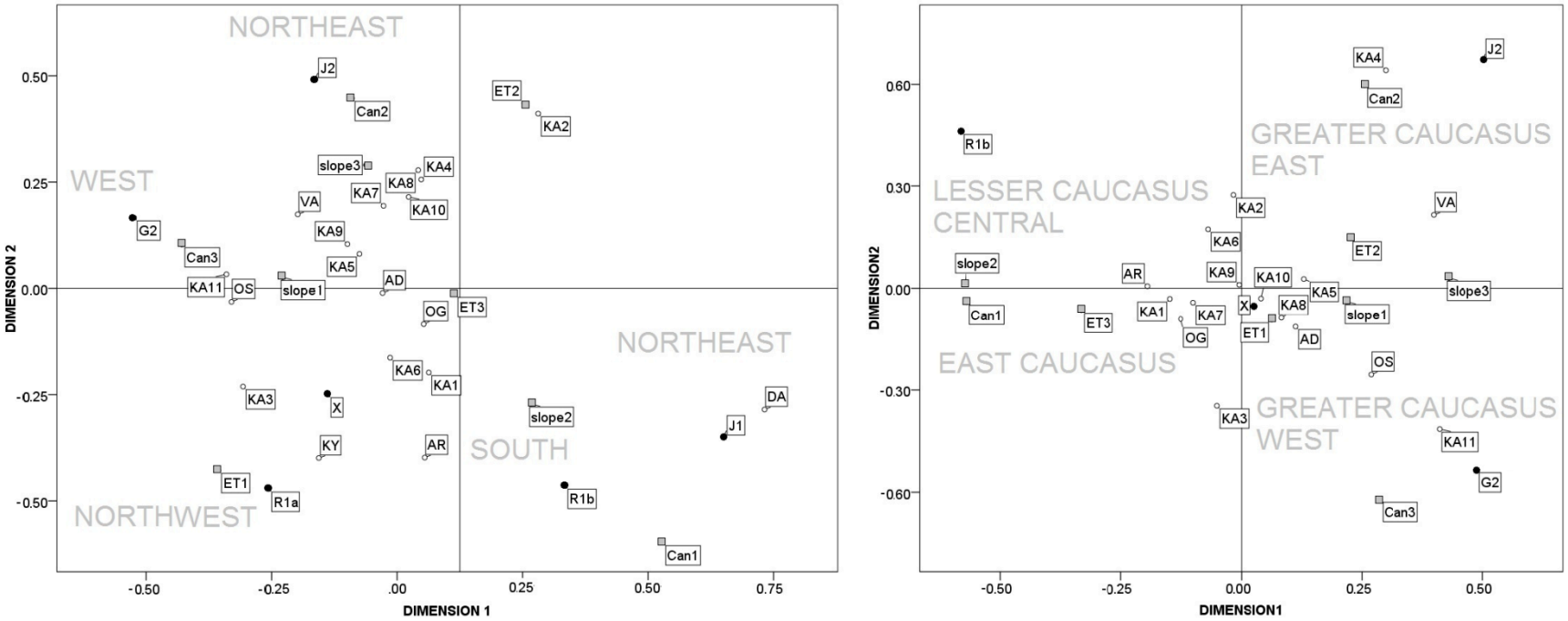

Fig. 3. Ordination of the major Y-DNA haplogroups (black circles), environmental conditions (gray squares), and linguistic units (open circles) in the Caucasus along the first two dimensions of Nonlinear Canonical Correlation Analysis. Left panel: the entire dataset, right panel: the dataset excluding outliers such as haplogroups J1 and R1a and linguistic groups DA and KY. See the legends of Table 2 and Fig. 1 for other abbreviations.

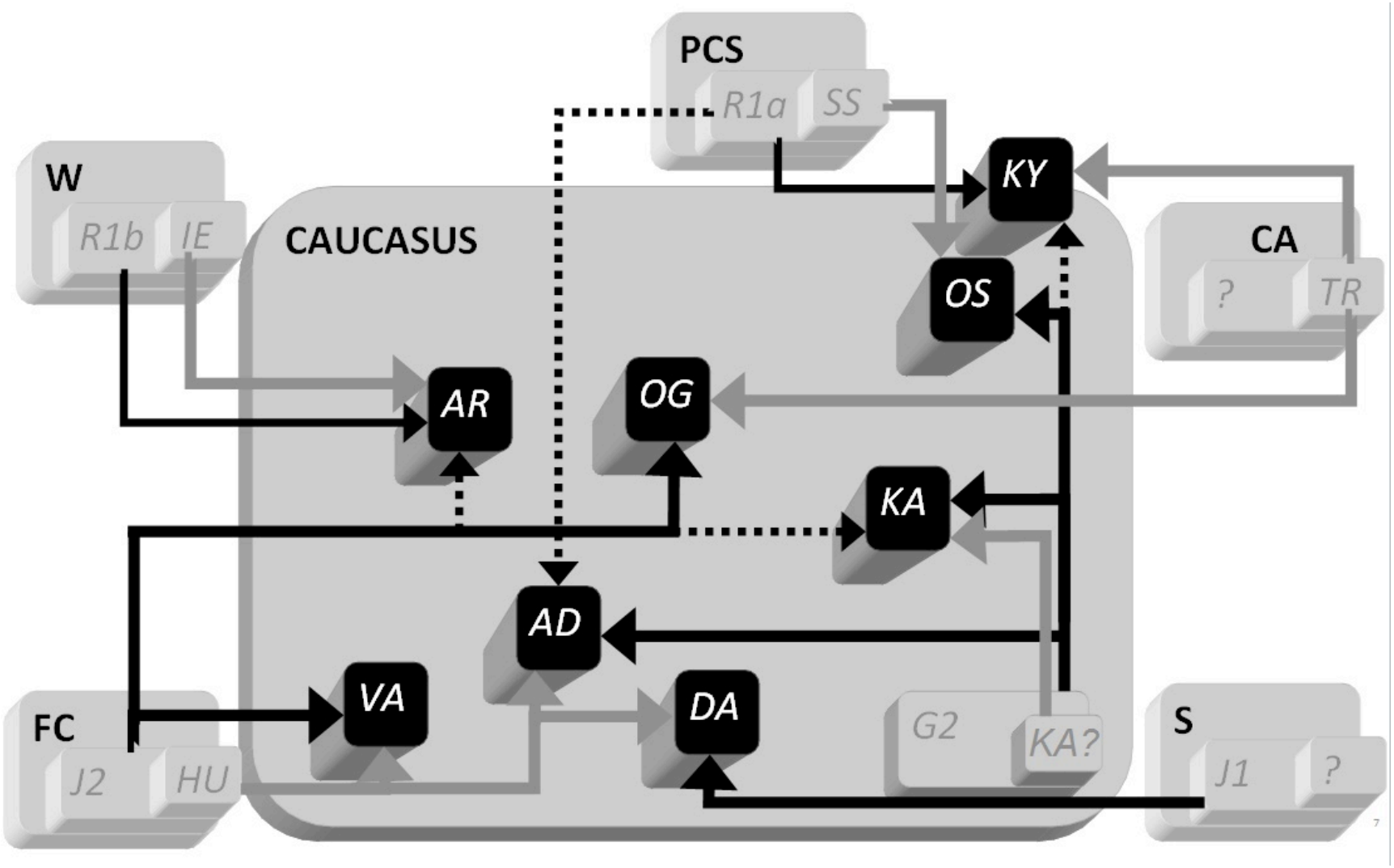

Fig. 4. Ancestral geographic areas, linguistic groups, and Y-DNA haplogroups found in the Caucasus. Ancestral geographic areas (gray rectangles): W – west (West Anatolia or Europe) associated with haplogroup R1b and proto-Indo-European (IE) linguistic group. PCS – Ponto-Caspian Steppe associated with haplogroup R1a and Scytho-Sarmatian (SS) linguistic group. CA – Central Asia associated with undefined Y-DNA haplogroup (?) and proto-Turkic (TR)

linguistic group. CAUCASUS – the Caucasus associated with haplogroup G2. FC – Fertile Crescent associated with haplogroup J2 and Hurro-Urartian (HU) linguistic group. S – south associated with the Zagros or the Alborz, haplogroup J1 and undefined (?) linguistic group. Current ethno-linguistic units (black rectangles): OS – Ossetian, KY – Kypchak, KA – Kartvelian, AD – Adyghean, AR – Armenian, VA – Vaynakh, OG – Oghuz, DA – Daghestanian. Black arrows show inferred genetic ancestry: solid lines – the most frequent Y-DNA haplogroup, dashed line – the second most frequent haplogroup with frequency exceeding 0.2. Gray arrows show inferred linguistic ancestry.

Appendix 1. Haplogroup samples used in the analyses.

#	description	G2	J2	R1b	J1	R1a	others	N	source
1	Georgians/Megrelian speakers. West Georgia. Districts Abasha, Senaki, Zugdidi	4	5	2	1	0	5	17	our data+DNA Family Tree Project
2	Georgians/Megrelian speakers. West Georgia, Abkhazia. Districts Gali, Ochamchire	4	3	0	0	1	2	10	our data+DNA Family Tree Project
3	Georgians/Megrelian speakers. West Georgia. Districts Martvili, Tsalenjikha, Chkhorotsku	8	2	2	0	0	4	16	our data+DNA Family Tree Project
4	Georgians/Georgian speakers. West Georgia. Districts Chokhatauri, Lanchkhuti, Ozurgeti, Kobuleti	3	5	1	3	0	2	14	our data+DNA Family Tree Project
5	Georgians/Georgian speakers. West Georgia. Districts Khelvachauri, Keda, Shuakhevi, Khulo	6	6	2	1	0	0	15	our data+DNA Family Tree Project
6	Georgians/Georgian speakers. West Georgia. Districts Khoni, Samtredia, Vani	4	5	2	1	1	0	13	our data+DNA Family Tree Project
7	Georgians/Georgian speakers. West Georgia. Districts Terjola, Tkibuli, Zestaphoni	9	6	2	0	0	3	20	our data+DNA Family Tree Project
8	Georgians/Georgian speakers. West Georgia. Districts Bagdadi, Kharagauli, Chiatura, Sachkhere	10	2	1	1	0	4	18	our data+DNA Family Tree Project

9	Georgians/Georgian speakers. West Georgia. Districts Ambrolauri, Oni, Tsageri	8	7	2	1	0	2	20	our data+DNA Family Tree Project
10	Georgians/Svan speakers. West Georgia. Districts Lentekhi, Mestia. Kodory Valley, Abkhazia	13	3	0	0	0	1	17	our data+DNA Family Tree Project
11	Georgians/Georgian speakers. Southern Georgia. Districts Adigeni, Aspindza, Akhaltsikhe, Borjomi	2	1	5	0	0	3	11	our data+DNA Family Tree Project
12	Georgians/Georgian speakers. East/Central Georgia. Districts Khashuri, Kareli, Gori, Tskhinvali	5	3	2	1	1	2	14	our data+DNA Family Tree Project
13	Georgians/Georgian speakers. East Georgia. Districts Kaspi, Mtskheta	7	6	1	1	0	1	16	our data+DNA Family Tree Project
14	Georgians/Georgian speakers. East Georgia. Districts Tianeti, Dusheti	5	6	0	1	4	3	19	our data+DNA Family Tree Project
15	Georgians/Georgian speakers. South-east Georgia. Districts Gardabani, Bolnisi, Dmanisi, Tetritskaro	1	2	2	2	0	3	10	our data+DNA Family Tree Project
16	Georgians/Georgian speakers. East Georgia. Districts Kvareli, Telavi, lagodekhi	8	8	1	1	0	1	19	our data+DNA Family Tree Project
17	Georgians/Georgian speakers. East Georgia. Districts Gurjaani, Signaghi, Sagarejo	5	5	4	0	0	3	17	our data+DNA Family Tree Project
18	Georgians/Georgian speakers. East Georgian mountaineers, historical province Tusheti	1	20	1	0	0	7	29	our data+DNA Family Tree Project

19	Ossetians, Scytho-Sarmatian linguistic group, Indoeuropean family, south of the Great Caucasus Mt. range	10	4	1	0	1	6	21	Yunusbayev et al. 2012
20	Ossetians, Scytho-Sarmatian linguistic group, Indoeuropean family, north of the Great Caucasus Mt. range	92	24	6	5	1	5	132	Yunusbayev et al. 2012
21	Armenians, Armenian linguistic group, Indoeuropean family, Armenia.	55	120	135	73	14	144	591	DNA Family Tree Project/ Armenians
22	Azeris, Oghus subgroup of Turkic linguistic group, Azerbaijan.	13	22	8	11	5	13	72	Nasidze et al. 2004b
23	Turks, Oghus subgroup of Turkic linguistic group, Karadeniz Mountains/ Black Sea area, Turkey.	8	16	11	8	4	36	83	Cinnioğlu et al. 2004
24	Turks, Oghus subgroup of Turkic linguistic group, Kars/ Ardagan area, Eastern Turkey	6	22	11	10	9	24	82	Cinnioğlu et al. 2004
25	Kumyks, Kypchak subgroup of Turkic linguistic group, Daghestan, NE Caucasus	10	13	17	16	11	17	73	Yunusbayev et al. 2012
26	Karachays, Kypchak subgroup of Turkic linguistic group, Karachay-Cherkesian authonomy, NW Caucasus	22	8	7	5	19	27	69	Yunusbayev et al. 2012
27	Balkars, Kypchak subgroup of Turkic linguistic group, Kabardin-Balkarian authonomy, NW Caucasus	44	21	18	5	38	47	135	Yunusbayev et al. 2012
28	Avarians, Daghestanian group of NE Caucasian family, Daghestan, NE Caucasus	0	2	2	28	1	10	42	Yunusbayev et al. 2012
29	Lezgi, Daghestanian group of NE Caucasian family, Daghestan, NE Caucasus	3	0	5	18	0	5	31	Yunusbayev et al. 2012
30	Dargin, Daghestanian group of NE Caucasian family, Daghestan, NE Caucasus	2	2	2	61	0	0	67	Yunusbayev et al. 2012
31	Tabasaran, Daghestanian group of NE Caucasian family, Daghestan, NE Caucasus	0	1	17	21	1	4	43	Yunusbayev et al. 2012
32	Chechen, Vainakh group of NE Caucasian family, Chechen Republic, NE Caucasus	2	80	1	40	1	42	165	Yunusbayev et al. 2012

33	Ingush, Vainakh group of NE Caucasian family, Ingush Republic, NE Caucasus	5	86	0	2	3	12	105	Yunusbayev et al. 2012
34	Shapsug, NW Caucasian (Adyghean) family, NW Caucasus	72	26	12	5	21	39	154	Yunusbayev et al. 2012
35	Circassians, NW Caucasian (Adyghean) family, NW Caucasus	57	31	3	6	19	29	126	Yunusbayev et al. 2012
36	Kabardin, NW Caucasian (Adyghean) family, NW Caucasus	60	22	5	13	20	40	140	Yunusbayev et al. 2012
37	Abaza, NW Caucasian (Adyghean) family, NW Caucasus	36	10	3	5	21	13	88	Yunusbayev et al. 2012
38	Abkhazians, NW Caucasian (Adyghean) family, Abkhazia, NW Georgia	77	43	6	4	16	16	162	Yunusbayev et al. 2012