MBE Advance Access published August 27, 2011

Parallel Evolution of Genes and Languages in the Caucasus Region

Oleg Balanovsky,*†,[1,2] Khadizhat Dibirova,†,[1] Anna Dybo,[3] Oleg Mudrak,[4] Svetlana Frolova,[1] Elvira Pocheshkhova,[5] Marc Haber,[6] Daniel Platt,[7] Theodore Schurr,[8] Wolfgang Haak,[9] Marina Kuznetsova,[1] Magomed Radzhabov,[1] Olga Balaganskaya,[1,2] Alexey Romanov,[1] Tatiana Zakharova,[1] David F. Soria Hernanz,[10,11] Pierre Zalloua,[6] Sergey Koshel,[12] Merritt Ruhlen,[13] Colin Renfrew,[14] R. Spencer Wells,[10] Chris Tyler-Smith,[15] Elena Balanovska,[1] and The Genographic Consortium[16]

[1]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia

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

[3]Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

[4]Russian State University for Humanities, Institute of Oriental Cultures, Moscow, Russia

[5]Adygei State University, Maikop, Russia

[6]The Lebanese American University, Chouran, Beirut, Lebanon

[7]Computational Biology Center, IBM T.J. Watson Research Center

[8]Department of Anthropology, University of Pennsylvania

[9]Australian Centre for Ancient DNA, Adelaide, Australia

[10]National Geographic Society

[11]Evolutionary Biology Institute, Pompeu Fabra University, Barcelona, Spain

[12]Moscow State University, Faculty of Geography, Moscow, Russia

[13]Department of Anthropology, Stanford University

[14]McDonald Institute for Archaeological Research, Cambridge, United Kingdom

[15]The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom

[16]Consortium members are listed in the Appendix

†These authors contributed equally to this work.

***Corresponding author:** E-mail: balanovsky@inbox.ru

Associate editor: Sarah Tishkoff

Abstract

We analyzed 40 single nucleotide polymorphism and 19 short tandem repeat Y-chromosomal markers in a large sample of 1,525 indigenous individuals from 14 populations in the Caucasus and 254 additional individuals representing potential source populations. We also employed a lexicostatistical approach to reconstruct the history of the languages of the North Caucasian family spoken by the Caucasus populations. We found a different major haplogroup to be prevalent in each of four sets of populations that occupy distinct geographic regions and belong to different linguistic branches. The haplogroup frequencies correlated with geography and, even more strongly, with language. Within haplogroups, a number of haplotype clusters were shown to be specific to individual populations and languages. The data suggested a direct origin of Caucasus male lineages from the Near East, followed by high levels of isolation, differentiation, and genetic drift *in situ*. Comparison of genetic and linguistic reconstructions covering the last few millennia showed striking correspondences between the topology and dates of the respective gene and language trees and with documented historical events. Overall, in the Caucasus region, unmatched levels of gene–language coevolution occurred within geographically isolated populations, probably due to its mountainous terrain.

Key words: Y chromosome, glottochronology, Caucasus, gene geography

Introduction

Since the Upper Paleolithic, anatomically modern humans have been present in the Caucasus region, which is located between the Black and Caspian Seas at the boundary between Europe and Asia. Although Neolithization in the Transcaucasus (south Caucasus) was stimulated by direct example and possible in-migration from the Near East, in the North Caucasus, archaeologists have stressed the cultural succession from the Upper Paleolithic to the Mesolithic and Neolithic (Bader and Tsereteli 1989; Bzhania 1996). Neolithic cultures developed in the North Caucasus ~7,500 years before present (YBP) from the local Mesolithic cultures (microlithic stone industries indicate a gradual transition) and, once established, domesticated local barley and wheat species (Masson et al., 1982; Bzhania 1996). Only in the Early Bronze Age (5,200–4,300 YBP) did cultural innovations from the Near East become more intensive with the emergence of the Maikop archaeological culture (Munchaev 1994). These could have occurred alongside migratory events from the area between the Tigris River in the east and northern Syria and adjacent East Anatolia in the west (Munchaev 1994: 170). The Late Bronze Age Koban culture (3,200–2,400 YBP), which predominated across the North Caucasus until Sarmatian times (2,400–2,300 YBP), became the common cultural substrate for most of the present-day peoples in the North Caucasus (Melyukova 1989, p. 295).

After approximately 1,500 YBP, this pattern changed, and most migration came to the North Caucasus from the East European steppes to the north rather than from the Near East to the south. These new migrants were Iranian speakers (Scythians, Sarmatians, and their descendants, the Alans) who arrived around 3,000–1,500 YBP, followed by Turkic speakers about 500–1,000 YBP. The new migrants forced the indigenous Caucasian population to relocate from the foothills into the high mountains. Some of the incoming steppe dwellers also migrated to the highlands, mixing with the indigenous groups and acquiring a sedentary lifestyle (Abramova 1989; Melyukova 1989; Ageeva 2000). The defeat of the Alans by the Mongols, and then by Tamerlane in the 14th century, stimulated the expansion of the indigenous populations of the West Caucasus into former Alan lands (Fedorov 1983). A later expansion of Turkic-speaking Nogais took place about 400 YBP.

The genetic contributions of these three components (indigenous Upper Paleolithic settlement, Bronze Age Near Eastern expansions, and Iron Age migration from the steppes) is unknown. Physical anthropologists have traced the continuity of local anthropological (cranial) types from the Upper Paleolithic, but data reflecting the influences from the Near East and Eastern Europe are contradictory (Abdushelishvili 1964; Alexeev 1974; Gerasimova et al. 1987).

Linguistically, the North Caucasus is a mosaic consisting of more than 50 languages, most of which belong to the North Caucasian language family. Iranian languages (Indo-European family) are represented by Ossets and there are a variety of Turkic-speaking groups (Altaic family) as well (Ruhlen 1987). A number of studies of North Caucasian languages (Trubetzkoy 1930; Gigeneishvili 1977; Shagirov 1977; Talibov 1980; Bokarev 1981; Chirikba 1996) collected linguistic data and established regular phonetic correspondences between North Caucasian languages. The resulting classification (Nikolaev and Starostin 1994) became generally accepted with some modifications (Kuipers 1963; Comrie 1987; Ruhlen 1987; www.ethnologue.com).

This classification was based on the common innovation method and particularly on the glottochronological method (Starostin 1989), which is now widely used by the Evolution of Human Languages Project coordinated by the Santa Fe Institute (http://ehl.santafe.edu/intro1.htm, http://starling.rinet.ru/main.html). For example, application of this method to the modern Romance languages (Spanish, Italian, Romanian, etc.) produced a date for their split to about 1,600 YBP. This date corresponds to the time of the disintegration of the Roman Empire, when Latina Vulgata (the common language in the Empire's provinces) became subdivided into regional dialects (Blazhek and Novotna 2008). Thus, this result validates the methodology, at least for this example.

The present work employs Starostin's methodology, and we made special efforts to create the high-quality linguistic databases required for this analysis. Thus, based on significantly extended and revised linguistic databases, we have applied a glottochronological approach to the North Caucasian languages. As a result, our study provides a unique opportunity to make direct comparisons of linguistic and genetic data from the same populations. Lexicostatistical methods have also been applied to a number of language families using a Bayesian approach to increase the statistical robustness of language classification (Gray and Atkinson 2003; Kitchen et al. 2009; Greenhill et al. 2010). Using these methods with the Caucasus languages under study here will be the focus of future work.

Previous studies of genetic diversity in the Caucasus (Nasidze et al. 2003, 2004a, 2004b) noted that geography, rather than language, provides a better (but statistically nonsignificant) explanation for the observed genetic structure. However, although the sample sizes for Y-chromosomal markers in those studies were large for southern populations, they were substantially smaller for North Caucasus populations (average $n = 28$, with the exception of the Ossets). Several subsequent papers have explored the genetic composition of the Dagestan in the east Caucasus (Bulaeva et al. 2006; Tofanelli et al. 2009; Caciagli et al. 2009). A later survey of the Y-chromosomal composition of the Caucasus was published in Russian only, with phylogenetic resolution no deeper than the designation of haplogroups G-M201, J1-M267, and J2-M172 (Kutuev et al. 2010). Some data can also be retrieved from papers that focused on other regions (Rosser et al. 2000; Semino et al. 2000; Wells et al. 2001; Zerjal et al. 2002; Di Giacomo et al. 2004; Semino et al. 2004; Cruciani et al. 2007; Battaglia et al. 2009). To the best of our knowledge, no other Y chromosome data from the Caucasus have been published and, except for Georgians and the Ossets, reliable data sets are very few.

Our study presents a much more extensive survey of Y-chromosomal variation in the Caucasus. All geographic subregions are covered and all large ethnic groups are represented by large sample sizes ($n_{\text{avg}} = 109$). We did not include Turkic-speaking populations as their recent immigration from Eastern Europe and Central Asia could possibly blur the deeper genetic patterns. Instead, we subtyped previously analyzed samples from the Near East (El-Sibai et al. 2009; Haber et al. 2010) and new samples from Eastern Europe for comparative purposes. Our genotyping strategy included the deepest known level of phylogenetic resolution for the common Caucasus haplogroups, as well as 19 short tandem repeats (STRs) to facilitate the dating of these genetic lineages.

Table 1. Characteristics of the Study Populations.

Populations	N	Language Group	Geographic Position	Region (country, province, district)	Latitude	Longitude	Sampling Supervisor(s)
Shapsug	100	Abkhazo-Adyghian	West Caucasus	Adygei republic, Tuapsinskii, Lazarevskii district	44.15	39.12	Balanovska
Abkhaz	58	Abkhazo-Adyghian	West Caucasus	Abkhazian republic	43.10	41.10	Balanovska, Pocheshkhova
Circassians	142	Abkhazo-Adyghian	West Caucasus	Karachai-Cherkess republic Habezskii, Prikubanskii district	43.80	41.75	Balanovska, Pocheshkhova
Ossets-Iron	230	Iranian	Central Caucasus	North Ossetian republic Alagirskii, Irafskii, Prigorodnyi district	42.90	44.47	Balanovska
Ossets-Digor	127	Iranian	Central Caucasus	North Ossetian republic Digorskii district	43.12	43.55	Balanovska
Ingush	143	Nakh	East Caucasus	Ingushetia republic Nazran' Malgobekskii	43.12	45.04	Balanovska, Pocheshkhova
Chechen (Ingushetia)	112	Nakh	East Caucasus	Ingushetia republic Malgobekskii district	43.20	45.20	Balanovska, Pocheshkhova
Chechen (Chechnya)	118	Nakh	East Caucasus	Chechnya republic Achhoi-Martanovskii district	43.20	45.98	Balanovska, Pocheshkhova
Chechen (Dagestan)	100	Nakh	East Caucasus	Dagestan republic Hasavyurtovskii, Kazbekovskii, Novolakskii district	43.32	46.55	Balanovska, Radzhabov
Avar	115	Dagestan	East Caucasus	Dagestan republic Shamil'skii, Buinakskii, Uncukul'skii, Gunibskii district	42.47	46.88	Balanovska
Dargins	101	Dagestan	East Caucasus	Dagestan republic Akushinskii and Dahadaevskii districts	42.18	47.22	Balanovska
Kubachi	65	Dagestan	East Caucasus	Dagestan republic Dahadaevskii district, Kubachi	42.08	47.58	Balanovska, Radzhabov
Kaitak	33	Dagestan	East Caucasus	Dagestan republic Kaitakskii district	42.15	47.63	Balanovska, Radzhabov
Lezghins	81	Dagestan	East Caucasus	Dagestan republic Ahtynskii district	41.55	47.70	Balanovska

Overall, the present study sets out to draw a precise and reliable portrait of the Y-chromosomal and linguistic variation in the Caucasus and to use this information to generate a more comprehensive history of the peoples of this area.

Materials and Methods

Samples

A total of 1,525 blood samples from 14 Caucasus populations (table 1) were collected in 1998–2009 under the supervision of Elena Balanovska using a standardized sampling strategy. All sampled individuals identified their four grandparents as members of the given ethnic group and were unrelated at least up to the third degree of relation. Informed consent was obtained under the control of the Ethics Committee of the Research Centre for Medical Genetics, Russia.

We also included 254 samples from the Near East (El-Sibai et al. 2009; Haber et al. 2010), which were further subtyped here for single nucleotide polymorphisms (SNPs) within haplogroups J2-M172 and G2a-P15.

Molecular Genetic Analysis

DNAs were extracted from white cells using an organic extraction method (Powell and Gannon 2002). DNA concentration was evaluated via quantitative real-time polymerase chain reaction (PCR) using the Quantifiler DNA Quantification Kit (Applied Biosystems) and normalized to 1 ng/µl.

Samples were SNP-genotyped using the Applied Biosystems 7900HT Fast Real-Time PCR System with a set of 40 custom TaqMan assays (Applied Biosystems). The samples were additionally amplified at 19 Y-chromosomal STR loci in two multiplexes and read on an Applied Biosystems 3130xl Genetic Analyzer. The first multiplex was the 17 STR loci Y-filer PCR Amplification Kit (Applied Biosystems). The remaining two STR loci, DYS388 and DYS426, along with six insertion–deletion polymorphisms (M17, M60, M91, M139, M175, and M186) were genotyped in a separate custom multiplex also provided by Applied Biosystems. Quality control procedures included checking SNP genotypes for phylogenetic consistency, comparing with the haplogroup predicted from STR profiles (http://www.hprg.com/hapest5/index.html), and independent replication of 20 samples at the University of Arizona.

Statistical Analysis

Haplogroup frequency maps were created with the GeneGeo software using algorithms described previously (Balanovsky et al. 2008). Nei's genetic distances between populations were calculated using the DJ software (Balanovsky et al. 2008) and visualized using multidimensional scaling plots and tree diagrams constructed with Statistica 6.0 (StatSoft Inc. 2001). Geographic distances between populations were obtained from geographic coordinates in DJ software using spherical formulae.

To estimate correlation and partial correlation coefficients between matrices of genetic, geographic, and linguistic distances, we conducted Mantel tests using Arlequin 3.11 (Schneider et al. 2000). The same software was used for the hierarchical analysis of molecular variation (AMOVA). Genetic boundaries were identified by Barrier 2.2 software (Manni and Guerard 2004).

Network Analysis

The phylogenetic relationships between haplotypes within a haplogroup were estimated with the Reduced Median (RM) network algorithm in the program Network 4.1.1.2 (Bandelt et al. 1995), with the reduction threshold equal to 1. RM networks were visualized with Network Publisher (Fluxus Engineering, Clare, UK). No available algorithm automatically identifies haplotype clusters within the network; doing this by hand (the common practice) is inevitably arbitrary to some degree. Therefore, we applied the following rules to minimize variation between individuals when identifying the clusters: 1) Because all the networks used had a clear center (the probable root), we considered as clusters only those groups of haplotypes that were linked to the root via an individual nodal haplotype (put differently, monophyletic branches in the network); 2) this cluster-specific shared node was considered to be the founder haplotype (selecting the founder is important for the age calculation using the ρ estimator); 3) to avoid using small sample sizes, we only considered clusters consisting of 10 or more samples; and 4) finally, we required that a cluster should be highly (above 80%) specific to a given population or group of closely related populations.

Dating Genetic Lineages

To estimate the age of particular Y-chromosomal lineages in Caucasus populations, we applied four commonly used methods. First, the ρ (rho) estimator (Forster et al. 1996; Saillard et al. 2000) was used to date haplotype clusters. Second, BATWING (Wilson et al. 2003) was used to obtain independent dates for these haplotype clusters. Third, BATWING was also used to estimate the possible sequence and dates of population splits. Fourth, the standard deviation (SD) estimator (Sengupta et al. 2006) was used to estimate the age required to accumulate the observed diversity within populations for entire haplogroups (not for clusters within the haplogroup as in the first and second analyses).

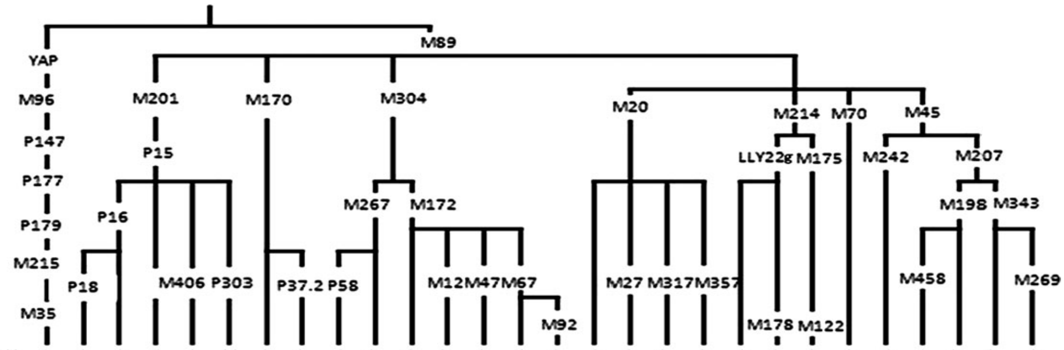
The Time to the Most Recent Common Ancestor (TMRCA) of the clusters of STR haplotypes that appeared to have evolved within specific populations, and which were identified in the networks, was estimated with the ρ statistic according to Saillard et al. (2000). Because haplotype clusters are population specific, the resulting age estimations serve as lower bounds for the time that a population may have been isolated following a split.

BATWING provides a mechanism to identify bounds, established by coalescent events, between which a unique event polymorphism (UEP) may have emerged. UEPs are typically identified by SNPs but can be generalized to include "virtual UEPs" that mark clusters of phylogenetically related STR haplotypes identified by Network, following the methods of Cruciani et al. (2004, 2006). Populations within which the clusters are observed were identified, and BATWING computations included all samples representing those populations in order to estimate UEP dates. Markers DYS385a and DYS385b were excluded from calculations. We used prior distribution parameters obtained from genealogical mutation rates reported in Ge et al. (2009). BATWING was configured to start with an ancestral effective population size that began an exponential expansion at a date that BATWING estimated.

BATWING was also employed to model the sequence of the population splits and to estimate the split times among populations. These estimates did not employ the virtual UEPs identified using Network. As such, this procedure provides an independent check for the times of the population splits.

SDs of microsatellite variances for four major haplogroups, G2a3b1-P303, G2a1a-P18, J2a4b*-M67(xM92), and J1*-M267(xP58), were calculated for each population with a sample size of at least five individuals from a given haplogroup. The confidence interval was estimated based on the standard error of the SD. This method is based on the average squared difference (ASD) in STR variation. It does not estimate the population divergence time but instead the relative age (amount of time) required to produce the observed microsatellite variation within the haplogroup at each population under the assumptions of limited gene flow and the same general mutation scheme for all the loci. All haplotypes of the haplogroup were used in the calculations and not just specific clusters. Calculations were carried out both including and excluding DYS385a and DYS385b because these markers

Table 2. Frequencies of Y-chromosomal Haplogroups (percent).



Shapsug	100	1	0	0	1	0	86	0	0	0	0	6	0	0	0	0	2	0	0	0	0	0	0	0	0	1	3	0	0
Abkhaz	58	1.7	12.1	0	24.1	0	20.7	0	0	0	0	5.2	0	0	6.9	1.7	0	0	3.4	0	0	0	0	1.7	0	0	10.3	0	12.1
Circassians	142	1.4	9.2	0	1.4	0	29.6	0	0.7	0.7	4.9	12	0.7	1.4	5.6	2.1	0	0.7	0	0	1.4	2.1	0	0.7	0.7	6.3	13.4	2.1	2.8
Ossets-Iron	230	0.4	72.6	0.4	0	0	1.3	0	0	0	1.3	7	0	0	11.3	0	0	0	0.9	0	0	0.4	0	0	0.9	0	0.4	0.4	2.6
Ossets-Digor	127	0.8	55.9	0	0	0	4.7	0	0	0	3.9	6.3	0	0	5.5	0	0	0.8	0	0	0	0	0	0.8	3.9	0	0.8	0.8	15.7
Ingush (Ingushetia)	143	0	1.4	0	0	0	0	0	0.7	0	2.8	1.4	0	0	87.4	0	0	0	0	2.8	0	0	0	0	0	0	3.5	0	0
Chechen (Ingushetia)	112	0	4	0	0	0	5	0	0	0	21	2	1	0	51	0	0	0	0	1	0	0	0	0	6	0	8	0	1
Chechen (Chechnya)	118	0	0	0	0	0	1	1	0	0	25	1	0	0	57	0	0	0	0	7	0	0	0	0	3	0	3	0	3
Chechen (Dagestan)	100	0	6	0	0	0	1	0	0	0	16	1	0	0	58	0	0	0	0	14	1	1	0	0	0	0	1	0	1
Avar	115	0	1	0	0	1	9	1	1	1	58	5	0	0	1	0	0	0	3	0	2	0	0	0	1	0	2	0	15
Dargins	101	0	0	0	1	0	1	0	0	1	69	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	22	0	2
Kubachi	65	0	0	0	0	0	0	2	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kaitak	33	0	0	0	0	0	0	0	0	0	85	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	3	0	6
Lezghins	81	1.2	1.2	0	12.3	0	0	0	0	0	44.4	2.5	0	0	0	0	0	0	1.2	0	0	1.2	0	2.5	0	0	3.7	0	29.6
Total	**1525**	**E1b1b1**	**G2a1a**	**G2a1**	**G2a***	**G2a3a**	**G2a3b1**	**I***	**I2a**	**J1e**	**J1***	**J2***	**J2b**	**J2a4a**	**J2a4b***	**J2a4b1**	**L***	**L1**	**L2**	**L3**	**N1***	**N1c1**	**O3**	**T**	**Q**	**R1a1a7**	**R1a1a***	**R1b**	**R1b1b2**

were not typed in an order-specific manner in our data. In the majority of the cases, the inclusion or the exclusion of these markers did not influence the estimates suggesting that these markers shared a common ancestor and a strong founder effect for the majority of the populations. In order to avoid overestimations of the relative ages, any population-specific locus behavior exhibiting an associated microsatellite variance 10-fold higher than the locus average variance among haplogroups, and the overall loci variance within population were excluded from calculations. The underlying logic here is to avoid including any locus variation that were not ultimately generated by the same general mutation scheme. Only in a few specific cases, the DYS385 and the DYS388 loci exhibited a 10-fold higher variances and therefore were excluded from the reported values.

When using ρ and SD estimators, we applied both an evolutionary mutation rate (6.9×10^{-4} per locus per generation; Zhivotovsky et al. 2004) and a genealogical rate (2.1×10^{-3}; Gusmao et al. 2005; Sanchez-Diz et al. 2008; Ge et al. 2009) to convert the observed variation into a number of generations. The results obtained by using these two rates were compared with linguistic and historical evidence. The BATWING prior distribution parameters were based on only the genealogical rate because BATWING models mutations in each generation of a genealogy and the genealogical rate seems therefore to be most suitable for this analysis. In all methods, when converting the number of generations into calendar years, we had to use a different generation time for each: 25 years with the evolutionary rate because this rate was initially estimated in years and converted into generations using 25 years/generation (Zhivotovsky et al. 2004), and 30 years for the genealogical rate because this is the approximate male generation time measured in demographic and anthropological studies (Fenner 2005) and shown for the Caucasus populations in the genetico-demographic study (Pocheshkhova 2008).

Dating Languages

Linguistic dates were calculated from the number of word substitutions that have accumulated after a language split (Starostin 1989; Starostin 2000; Embleton 2000). The basic principles of this method, its applications, and the formulas used in our study are described in the supplementary note 1 (Supplementary Material online). This glottochronological approach was first used by Starostin with North Caucasus languages (Nikolaev and Starostin 1994). The present study continues this analysis with updated linguistic databases (word lists). The Caucasian word lists used in our study were significantly modified by Mudrak, whereas the Ossetian word lists were recorded by Ershler and analyzed by Dybo (this study).

Note that both the linguistic and genetic dating methods used in our study provide a most recent (lower) estimate of the population split time (supplementary note 1, Supplementary Material online).

Results

Structuring of the Caucasus Y-chromosomal Gene Pool

We analyzed 1,525 Y-chromosomal haplotypes from 14 Caucasus populations. Table 2 presents the haplogroup frequencies, whereas the Y-STR haplotypes are provided in the supplementary table 1 (Supplementary Material online). We additionally subtyped 121 haplogroup G-M201 samples and 133 haplogroup J-M304 samples from previously analyzed Near Eastern populations (supplementary table 2,

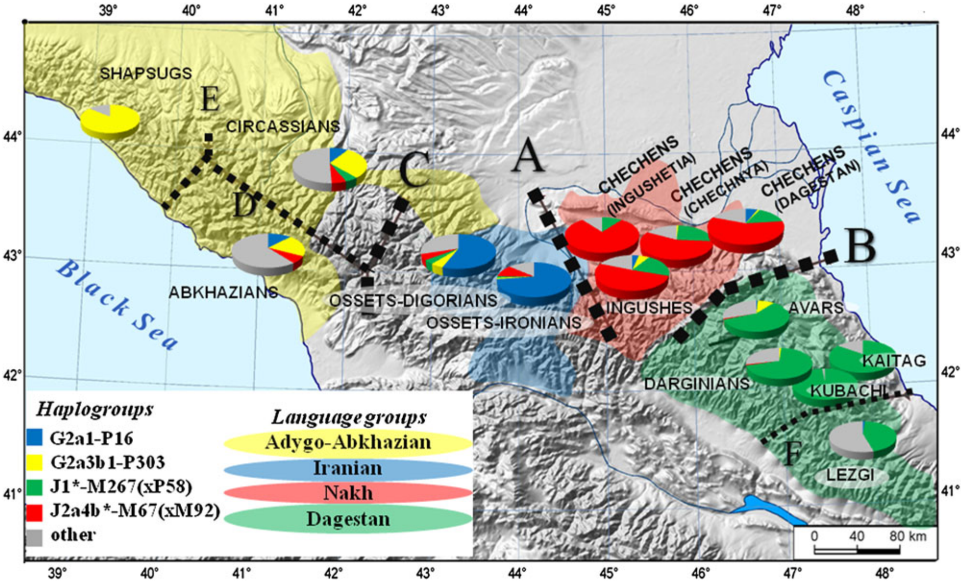


Fig. 1. Geographic location, linguistic affiliation, and genetic composition of the studied populations. Each population is designated by a pie chart representing frequencies of the major haplogroups in it. Areas of the linguistic groups of the Caucasus (except for Turkic groups) are shown by semitransparent color zones. Black dotted lines indicate genetic boundaries identified in the barrier analysis (thick lines: most important boundaries A, B, and C; thin lines: other boundaries D, E, and F).

Supplementary Material online). Overall, the most frequent haplogroups in the Caucasus were G2a3b1-P303 (12%), G2a1a-P18 (8%), J1*-M267(xP58) (34%), and J2a4b*-M67(xM92) (21%), which together encompassed 73% of the Y-chromosomes, whereas the other 24 haplogroups identified in our study comprise the remaining 27% (table 2).

However, these average frequencies masked the real pattern that became apparent when regional populations were considered (fig. 1). Each of these four haplogroups had its own focus within the Caucasus. More specifically, haplogroup G2a3b1-P303 comprised at least 21% (and up to 86%) of the Y chromosomes in the Shapsug, Abkhaz, and Circassians (fig. 1 and table 2). These populations live in the western part of the Caucasus and linguistically belong to the Abkhazo-Adyghe language group. The frequency of this haplogroup was below 10% (average 2%) in all other populations investigated from the Caucasus. Similarly, haplogroup G2a1a-P18 comprised at least 56% (and up to 73%) of the Digorians and Ironians (both from the Central Caucasus Iranic linguistic group), while not being found at more than 12% (average 3%) in other populations. Again, haplogroup J2a4b*-M67(xM92) comprised 51–79% of the Y chromosomes in the Ingush and three Chechen populations (North-East Caucasus, Nakh linguistic group), whereas, in the rest of the Caucasus, its frequency was not higher than 9% (average 3%). Finally, haplogroup J1*-M267(xP58) comprised 44–99% of the Avar, Dargins, Kaitak, Kubachi, and Lezghins (South-East Caucasus, Dagestan linguistic group) but was less than 25% in Nakh populations and less than 5% in the rest of Caucasus.

A genetic boundary analysis revealed the same pattern (fig. 1). This methodology (Womble 1951; Rosser et al. 2000) identifies zones of abrupt changes in haplogroup frequencies. The first (most significant) boundary A separated Nakh-Dagestan–speaking populations of the east Caucasus from other populations of the region. Boundary B separated the four Nakh-speaking populations from the five Dagestan-speaking ones, whereas Boundary C separated the Iranian-speaking Ossets (Central Caucasus) from Abkhazo-Adyghe speakers of the west Caucasus. Overall, the first three boundaries divided the Caucasus into four regions, each of which coincided with areas of prevalence of one of the four major Caucasus haplogroups and with areas of different major linguistic groups. The other genetic boundaries D, E, and F subdivided Abkhazo-Adyghe speaking populations and separated the Lezghins from populations speaking Dargin languages.

Geography Versus Linguistic Diversity

The observed pattern of genetic variation can be interpreted in two ways. On the one hand, it shows excellent correlation with geography as each of four major haplogroups is prevalent in a specific region of the Caucasus. On the other hand, this pattern also closely fits the classification of Caucasus languages (with no exceptions).

To test these patterns, we computed three matrices of pairwise distances between all studied populations: genetic (from haplogroup frequencies), geographic (in kilometers), and linguistic (percentage of words in common). The Mantel test results showed a significant ($P \leq 0.002$) correlation both between genetics and language ($r = 0.64$) and between genetics and geography ($r = 0.60$) (table 3). Partial correlations were calculated between genetics and both factors (language or geography) separately, holding the alternative factor constant. Unfortunately, the analysis was complicated by the fact that linguistics and geography in Caucasus are closely linked with each other ($r = 0.78$). For this reason, genetic distances exhibited insignificant partial correlation with

Table 3. Correlation between Genetic[a], Linguistic[b], and Geographic Distances.

Distance Considered	Correlation Coefficient	*P* value
Genetics and language	0.64	0.002
Genetics and geography	0.60	0.001
Genetics and language. Geography held constant	0.34	0.120
Genetics and geography. Language held constant	0.21	0.180

[a] Populations of the North Caucasian linguistic family (and their respective languages) from table 1 were considered, except for the genetic isolates Shapshug, Kubachi, and Kaitak with census size less than 10,000 persons as genetic drift may have caused substantial random fluctuation of the haplogroups frequencies in these populations.

[b] Three Chechen populations speaking the same Chechen language were pooled for this analysis. Ossets were not considered because they belong to the Indo-European linguistic family, and linguistic distances cannot be estimated between such divergent languages by this method.

both, although the correlation with linguistics was almost twice as strong (table 3).

AMOVA was used to further investigate which factor might be the major driving force behind this degree of differentiation (table 4). When populations were grouped geographically, the proportion of variation in haplogroup frequencies between geographic groups was 0.146. Linguistic classification of the same populations provided nearly two times the extent of variation between linguistic groups (0.268). Therefore, linguistics explained a larger part of Y-chromosomal variation in the Caucasus.

These analyses indicated that linguistic diversity is at least as important as geography in shaping the Y-chromosomal landscape and suggested that the pronounced genetic structure of the Caucasus might have evolved in parallel with the diversification of the North Caucasus languages.

Caucasus in a Eurasian Context

Figure 2 compares the Y-chromosomal pool of the Caucasus with its neighboring regions (Balkans, South-East Europe, and the Near East) by presenting frequency maps for haplogroups, which predominate in any of these regions. Four haplogroups, G2a3b1-P303, G2a1a-P18, J2a4b*-M67(xM92), and J1*-M267(xP58), exhibit their highest documented frequencies in the Caucasus. Haplogroup G2a3b1-P303 predominates in the West Caucasus (table 2), although we also found it in the Near East (table 2) and in one Russian population (data not shown), and it has been reported in Western Europeans (www.ysearch.org). The second haplogroup G2a1a-P18 is almost absent outside of the Caucasus (fig. 2), although its ancestral clade, G2a1-P16, is present in the Near East (Cinnioglu et al. 2004; Flores et al. 2005). Similarly, J1*-M267(xP58) was found mainly in the Caucasus (fig. 2) with the ancestral J1-M267 being common in the Near East. The fourth haplogroup J2a4b*-M67(xM92) is prevalent in a region spanning the Near East and the Caucasus (fig. 2). Note that only a few Near Eastern haplogroups are represented in the Caucasus, and other major components of the Near Eastern Y-chromosomal pool (e.g., J2b-M12) are virtually absent there.

Table 4. AMOVA Results: Linguistic Versus Geographic Grouping of Populations.

	Linguistic Grouping[a]	Geographic Grouping[a]
Variation among groups[b]	0.268*	0.146*
Variation among populations within groups	0.099*	0.235*
Variation within populations	0.633*	0.619*

[a] The linguistic and geographic affiliation of each population is listed in table 1.

[b] All populations from table 1 were used. Thus, AMOVA used a larger data set than the Mantel test because it was possible to use information from the Indo-European speaking Ossetian populations, which were too linguistically distant to obtain lexicostatistical distances for the Mantel test.

**P* value < 0.001.

When haplogroups common in Europe were examined, we observed that the typical Balkan haplogroup I2a-P37.2 was virtually absent in the Caucasus (fig. 2). The recently defined subbranch R1a1a7-M458 (Underhill et al. 2010) was found among only the Circassians and Shapsug (table 2). R1a*-M198(xM458) has an average frequency in the Caucasus as low as 5% but was found in 20% of the Circassians and 22% of the Dargins, two populations that occupy opposite parts of the Caucasus. STR haplotypes from these Circassian and Dargins samples formed distinct clusters in a network (supplementary fig. 1, Supplementary Material online). Similarly, two different haplotype clusters within R1b1b2-M269 (supplementary fig. 1, Supplementary Material online) were found in the Lezghins (30%) and in Ossets–Digor (16%). These concentrations of (presumably European) haplogroups R1a*-M198(xM458), and R1a1a7-M458 found in few locations in the Caucasus might indicate independent migrations from Europe that were too small to make any significant impact on Caucasus populations.

The multidimensional scaling (MDS) plot of pairwise genetic distances (fig. 3) showed strong regional clustering separating populations from Europe, the Near East, and the Caucasus, with samples from the Caucasus grouping closer to the Near East samples than to the European ones. Of those three clusters, the Caucasus appeared to be the most diverse, and three subgroups could be seen within it. The first subgroup included Dagestan speakers (Avar, Dargins, Kubachi, and Kaitak), the second Nakh speakers (three Chechen populations and the Ingush), and the third Abkhazo-Adyghe speakers (Abkhaz, Circassians, Shapsug). Ossets also joined this cluster because we combined their predominant haplogroup G2a1a-P18 with Abkhazo-Adyghe predominant haplogroup G2a3b1-P303 to achieve compatibility with the less phylogenetically resolved European and Near Eastern data. This plot illustrates both the common Near Eastern background of all Caucasus populations and the pronounced intra- Caucasus genetic differentiation into groupings that corresponded well with their linguistic affiliation.

More insights into the relationships between Caucasus populations were obtained from a tree based on haplogroup frequencies (fig. 4, left). Another tree (fig. 4, right), representing the linguistic classification, had the same topology except for the Dargins, who joined the Kubachi/Kaitak cluster before the Avar did. The Indo-European–speaking Ossets were outliers in the Caucasus linguistic tree, and the genetic tree also placed them separately, with

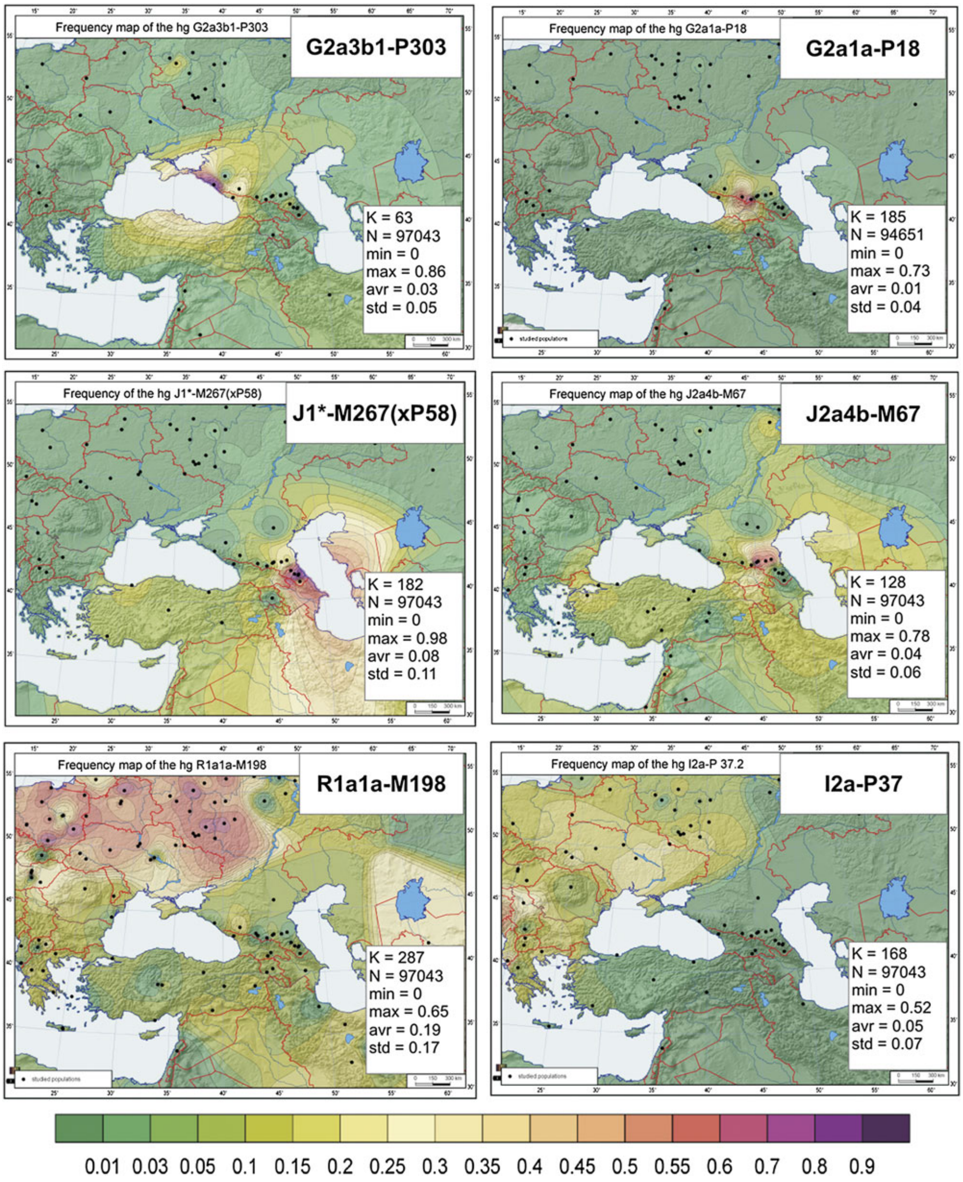


FIG. 2. Frequency maps of major Caucasus, Near Eastern, and East European haplogroups.

slight similarity to the Abkhaz. Generally, the tree based on genetic distances mirrored the linguistic tree in its overall pattern and in most details.

Haplotype Networks and Age Estimates

To analyze this parallelism further, we constructed phylogenetic networks for Caucasus haplogroups (fig. 5, supplementary fig. 1, Supplementary Material online). While all the results presented above were obtained using only haplogroup frequencies, in this and the following section, we analyze the variation of STR haplotypes within haplogroups.

Haplogroup G2a1a-P18 (fig. 5), which was found almost exclusively in the Caucasus, consisted of distinct branches of STR haplotypes rooted in the central reticulated zone. The larger cluster α includes mainly Ossets-Iron and demonstrated a star-like pattern with a central founder haplotype and a few subfounders. The cluster β comprised many samples of Ossets-Digor, whereas smaller cluster γ comprised both Ossetian populations. The diverse cluster in the upper part of the network comprised different Caucasus populations and a few non-Caucasus G2a1a-P18 samples also belonged to this branch.

Generally, haplogroup G2a1a-P18 seemed to have a long history in the Caucasus, being spread across the region and forming many branches. However, two of them (clusters α and β, found in the Ironians and Digorians, respectively) showed in their star-like structure signs of relatively recent expansion. The average number of mutation steps (ρ estimator) was similar for both clusters (1.46 for α and 1.41 for β), indicating that both clusters had expanded at around the same time, possibly because of the same event.

Reduced median networks for other haplogroups (supplementary fig. 1, Supplementary Material online) revealed

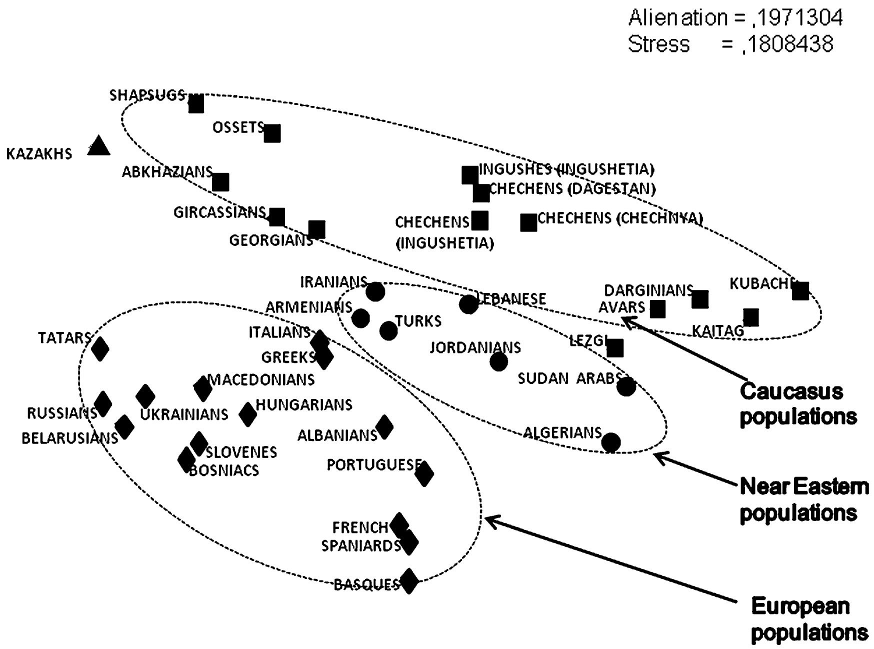


Fig. 3. Multidimensional scaling plot depicting genetic relationships between Caucasus, Near Eastern, and European populations. The plot is based on Nei's pairwise genetic distances calculated from frequencies of 13 Y-chromosomal haplogroups (C, E, G, I, J1, J2, L, N1c, O, R1a1, R1b1, Q, and other) in populations of North Caucasus (this study), Transcaucasus (Georgians, Battaglia et al. 2009), Near East (this study; Cinnioglu et al. 2004; Flores et al. 2005), and some other European, African, and Asian populations (data from Y-base, compiled in our lab from published sources). Caucasus populations are shown by squares, Near Eastern populations by circles, and European populations by diamonds.

similar patterns of branching. Some branches were shared between different Caucasus populations (and often included also Near Eastern samples), whereas many others were absent from the Near East and, moreover, specific to individual Caucasus populations (e.g., clusters α, β, and γ in fig. 5 are specific to the Ossets).

Applying the formal criteria, we identified 18 population-specific clusters (average specificity 95%). Table 5 lists these clusters with their ages, suggests the population event relevant to each, and indicates the linguistic date of this event (the tree of the North Caucasian languages obtained in our study is presented in supplementary fig. 2, Supplementary Material online).

A Tree of Population Splits

Because of the controversy between "evolutionary" and "genealogical" mutation rates, we set out to reconstruct the population history of the Caucasus in two phases. The first was based solely on genetic evidence without consideration of any mutation rates, whereas the second converted genetic diversity into time using both rates and then compared them with the linguistic times.

In phase 1, we grouped populations according to the predominant haplogroup, which yielded four branches (supplementary fig. 3, Supplementary Material online). To explore the intra-branch relationships, we examined the population-specific STR clusters (table 5). Cluster P303-β was shared

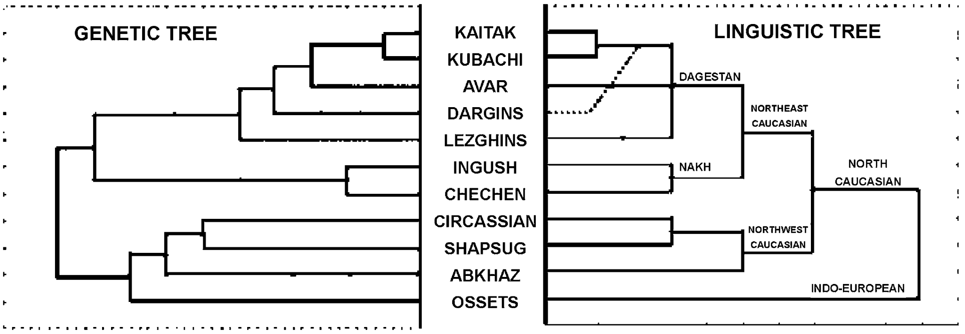


Fig. 4. Comparison of the genetic and linguistic trees of North Caucasus populations. The genetic tree was constructed from frequencies of 28 Y-chromosomal haplogroups in North Caucasus populations (data from table 2). Populations speaking the same language (three Chechen populations and two Ossetian ones) were pooled to make the genetic data set compatible with the linguistic classification. The weighted pair-group method was used as a clustering algorithm. The linguistic tree represents the classification of the North Caucasian languages from classical work (Ruhlen 1987). Kubachi and Kaitak (languages of small populations) were not listed in Ruhlen's classification, but most linguists agree that they are most related to the Dargin language.

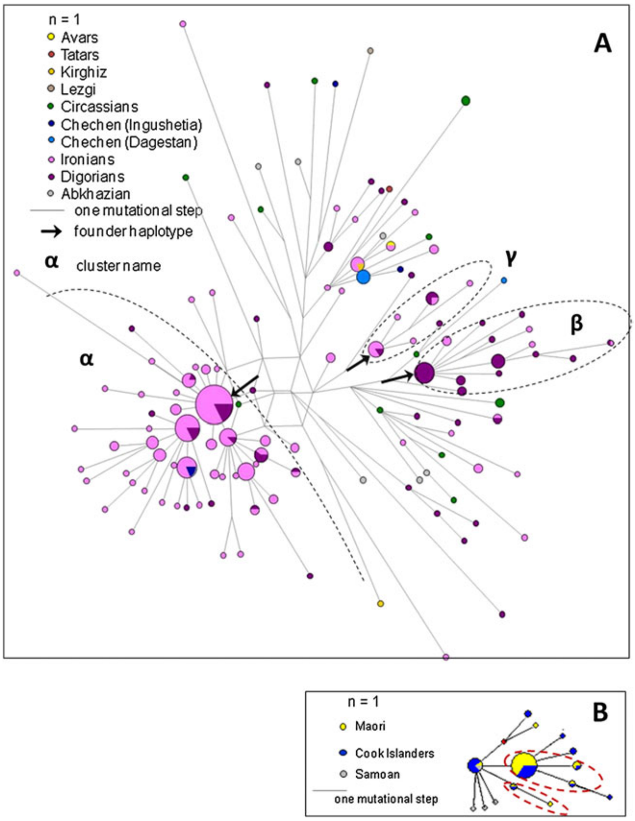


Fig. 5. Phylogenetic networks of the haplogroup G2a1a-P18 in the Caucasus and haplogroup C-M208 in Polynesia. (*A*) Reduced median network of haplogroup G2a1a-P18 was constructed using all available (worldwide) STR haplotypes for this haplogroup, with a reduction threshold $r = 1.00$ based on nonweighted data from 15 STRs (DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and GATA_H4). Black dotted lines designate clusters selected in our study for age estimations. (*B*) Reduced median network of haplogroup C-M208 in Polynesia (modified from Zhivotovsky et al. 2004). Red dotted lines designate clusters selected by Zhivotovsky et al. (2004) for estimating the evolutionary effective mutation rate.

between Shapsug and Circassians, whereas no cluster linked any of these populations with Abkhaz. We concluded that Abkhaz separated first, whereas Shapsug and Circassians maintained a shared ancestry for a longer period, during which the P303-β cluster originated. The cluster P303-α, which is present in Shapsug but absent in Circassians, marks the next split on the population tree. In terms of mutation steps, the first split (Abkhaz vs. Shapsug-Circassians) occurred 1.5 "mutations before present," whereas the second split (Shapsug vs. Circassians) took place 0.6 "mutations before present" (table 5).

Applying the same methodology to another three branches of populations led to a tree of population splits (supplementary fig. 3, Supplementary Material online). This tree is based solely on genetic data and yet shows good agreement with the topology of both linguistic trees: the classical way of grouping North Caucasian languages (fig. 4, right) and the quantitative lexicostatistical tree (supplementary fig. 2, Supplementary Material online). Up to this point, we have avoided using any mutation rate. As a result, we compare the topology of the trees (sequence of splitting events) without reference to a time scale.

In phase 2, to introduce time estimates, we used both "genealogical" and "evolutionary" mutation rates. The ρ estimator using the genealogical rate provided a good fit between genetics and linguistics as the genetic dates were similar to, or younger than, the linguistic dates. Because clusters can expand at any time after the split, they are expected to be younger than the respective languages; however, they should not be older, as described in supplementary note 1 (Supplementary Material online). Estimates based on the "evolutionary" mutation rate were too old to be in agreement with the linguistic dates (table 5).

BATWING computations of the ages of the same clusters based on genealogical rates showed similar results to those indicated by the linguistic analysis. The age for the four major haplogroups in individual populations obtained by using SD estimator (supplementary table 3, Supplementary Material online) are close to the Neolithic epoch and might be interpreted as signs of population expansion due to the shift to a farming economy.

The BATWING tree of population splits (supplementary fig. 4, Supplementary Material online) was based on all STR data from supplementary table 1 (Supplementary Material online) disregarding the haplotype clusters to which they belong. This tree therefore provides an independent test of the other genetic trees presented in this study (the tree on the left of fig. 4 is based on the haplogroup frequencies, whereas trees in fig. 6 and supplementary fig. 3 (Supplementary Material online) are based on the ages of haplotype clusters). This tree again showed a striking resemblance to the linguistic tree: one observes an initial split into west and east Caucasus populations and then the separation of Abkhaz from Circassian–Shapsug on the western branch and Nakh populations from Dagestan ones on the eastern branch; disagreements could be found only within the Dagestan group. Ossets (linguistic outliers) showed a slight similarity to Abkhaz as they did on the haplogroup-based tree, as well (fig. 4 left). Although the topology of this tree is similar to the linguistic one, the BATWING dates were on average 1.5 times younger. If the evolutionary mutation rate were applied (data not shown), the topology of the BATWING tree would remain the same but the dates would become on average 1.5 times older than corresponding linguistic dates.

Discussion

Origin and Structuring of the North Caucasus Paternal Pool

Four haplogroups are predominant in the Caucasus (table 2), and each of them has its own domain (recognizable geographically and also linguistically), where it represents the lion's share of the regional gene pool. In all other domains, the given haplogroup is infrequent or absent. The robustness of this conclusion is enhanced by the fact that each domain is occupied by more than one population in our data set

Table 5. Genetic and Linguistic Dates of the Populations Splits.

Cluster	Specific to Population	STRs[a]	N_S[b]	N_H[c]	S[d]	ρ[e] $\pm\sigma$[f]	Age_G[g]	Age_E[h]	Age_B[i]	Linguistic Date	Population Event

P303-β	Shapsug and Circassians	15	12	7	100	1.5 ± 0.5	1400±500	4300 ± 1400	MIN(1.1–4.8)
									MAX(1.8–7.8)
										3600	Separation of Shapsug-Circassians branch from Abkhaz

P303-α	Shapsug	15	27	8	92.6	0.55 ± 0.25	500 ± 200	1600 ± 700	MIN(1.0–2.6)
									MAX(1.4–3.9)
										800	Separation of Shapsug from Circassians

P18-α	Ossets-Iron	15	211	49	88.6	1.46 ± 0.48	1400 ± 500	4200 ± 1400	MIN(2.5–5.2)
									MAX(2.9–6.7)
										1300	

P18-β	Ossets-Digor	15	28	12	85.7	1.41 ± 0.48	1300 ± 500	4100 ± 1400	MIN(1.1–2.5)
									MAX(1.4–3.3)
										1300	

R1b1b2-β	Ossets-Digor	17	24	12	100	0.91 ± 0.31	800 ± 300	2300 ± 800	MIN(1.1–3.0)
									MAX(1.9–5.3)
										1300	Split of Ossets into Iron and Digor

P18-γ	Ossets-Iron and Digor	15	10	5	100	1 ± 0.52	1000 ± 500	2900 ± 1500	—	—	—

M67-β	Chechen and Ingush	14	45	12	93	1.56 ± 0.81	1600 ± 800	4800 ± 2500	MIN(1.0–4.0)
									MAX(1.8–6.7)
										5600	Separation of Nakh populations from Dagestan ones

M67-γ	Chechen and Ingush	14	81	22	96	1.84 ± 0.69	1900 ± 700	5700 ± 2100	MIN(0.9–2.5)
									MAX(1.5–4.1)
										5600	

M67-α	Ingush	14	53	9	81	1.96 ± 1.02	2000 ± 1000	6100 ± 3200	MIN(1.4–3.2)
									MAX(1.9–4.2)
										1400	

M67-δ	Chechen	14	22	3	100	0.14 ± 0.1	100 ± 100	400 ± 300	—	1400	

L3	Chechen	17	24	13	95.8	1.13 ± 0.42	900 ± 400	2900 ± 1100	MIN(0.9–3.3)
									MAX(11.9–20.4)
										1400	Split of Nakh branch into Chechen and Ingush

Q-α	Chechen	17	10	6	100	2.18 ± 1.02	1800 ± 900	5600 ± 2600	MIN(0.6–2.2)
									MAX(1.2–4.6)
										1400	

M267(xP58)α	Dargins. Kubachi	15	16	6	100	1.06 ± 0.5	1000 ± 500	3100 ± 1400	MIN(0.4–1.7)
									MAX(0.7–3.0)
										3400	

M267(xP58)β	Dargins Kubachi	15	15	8	93.3	1.44 ± 0.63	1400 ± 600	4200 ± 1800	MIN(0.7–2.3)
									MAX(1.2–3.3)
										3400	Separation of Dargins. Kubachi and Kaitak from other Dagestan populations

M267(xP58)γ	Dargins. Kubachi. Kaitak	15	11	6	100	1.09 ± 0.68	1000 ± 600	3200 ± 2000	MIN(0.5–1.9)
									MAX(1.0–3.1)
										3400	

R1a1a*-α	Dargins	17	13	5	100	1.77 ± 0.99	1500 ± 800	4500 ± 2500	MIN(0.4–1.4)
									MAX(0.6–2.7)
										1900	Separation of Kubachi from Dargins

R1b1b2-α	Lezghins	17	21	9	100	2.38 ± 0.89	2000 ± 700	6100 ± 2300	MIN(1.0–2.7)
									MAX(1.3–3.7)
										4300	Separation of Lezghins from Dargins populations

P15*-α	Lezghins	15	11	8	90.9	2.45 ± 0.76	2300 ± 700	7100 ± 2200	MIN(1.2–3.7)
									MAX(1.7–5.6)
										4300	

[a] Number of STRs used for the network.
[b] Number of samples in cluster.
[c] Number of haplotypes in cluster.
[d] Specificity of the clusters (proportion of samples from the indicated population among all samples within the cluster, in percent).
[e] Average distance to the founder haplotype (Forster et al. 1996).
[f] Estimator for the variance (Saillard et al. 2000).
[g] Age of the cluster, obtained from the ρ estimator using the genealogical mutation rate (YBP).
[h] Age of the cluster, obtained from the ρ estimator using the evolutionary mutation rate (YBP).
[i] 95% confidence interval of age of the NETWORK-identified STR haplotype cluster within the population obtained in the BATWING analysis using the genealogical rate (ky BP).

whose characteristic haplogroup is prevalent in each population of the domain (table 1). This pronounced structuring of the Y-chromosomal pool of the Caucasus has not previously been reported. In the study of Nasidze et al. (2004), AMOVA computation revealed a lack of correlation with language, whereas the correlation with geography did not reach the significance threshold. The increased phylogenetic resolution and large sample sizes used in our study were necessary to reveal the links between haplogroups, regions, and language groups (fig. 1). Methodologically, the analysis of correlations between geography, language, and genetics (table 3) parallels an earlier study performed with European populations (Rosser et al. 2000), where geography was shown to be the leading factor in Europe. In our analysis, the correlation was much stronger (maximum value 0.64) than in the European study (0.39).

Strikingly, language rather than geography tended to have a larger influence on the genetic structuring in the Caucasus (tables 3 and 4). Language and geography are also closely linked with each other, probably because of the mountainous nature of the Caucasus region where languages are often restricted to a few valleys. In this context, the slightly higher dependence of genetic structure on language could be explained by marriage and individual migration practices, linking linguistically similar populations in preference. For example, Circassians, who are geographically situated between Adyghes and Ossets, might receive more gene flow from Adyghes, who speak a similar

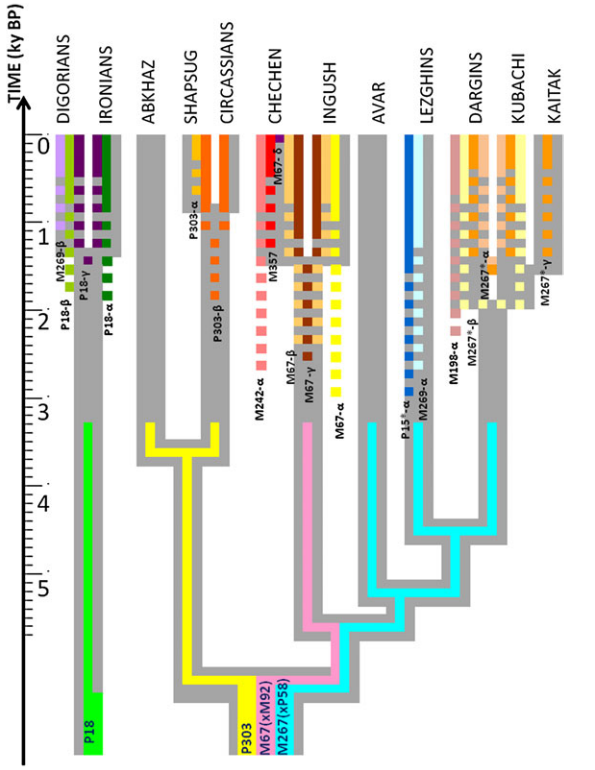


Fig. 6. Model of the evolution of Caucasus populations combining genetic and linguistic evidence. The gray background outlines the linguistic tree, obtained by lexicostatistical method. Each colored line near the tips of the tree marks a haplotype cluster that is specific to a given population. If the cluster is shared between two populations, then both populations carry this color on their branches. Standard errors of a cluster's age are shown by dotted colored lines. Each colored line near the root of the tree marks one of four major haplogroups. These lines stop 3,300 YBP. The root of the population tree indicates an initial migration from the Near East carrying four major haplogroups. This proto-population then separated into the West Caucasus, proto-Ossets, Nakh, and Dagestan branches, differing by language and predominant haplogroup. The subsequent evolution (occurring independently in each of these four groups) consisted in the diversification of their languages and emergence of branch-specific or population-specific haplotype clusters.

language, than from Ossets who differ in their language and culture.

Ossets, who speak an Indo-European language, find their place among populations of the North Caucasus language family. This genetic association is consistent with the physical anthropological evidence (Abramova 1989; Melyukova 1989) that Ossets are mainly descendants of indigenous Caucasus populations, who were assimilated by Alans and received from them the present language. Little is known about the language that these populations initially spoke. Note that, on the genetic trees, Ossets join the western (Abkhaz-Adyghe) branch of the North Caucasus family.

Although occupying a boundary position between Europe and the Near East, all four major Caucasus haplogroups show signs of a Near Eastern rather than European origin (fig. 2, supplementary fig. 1, Supplementary Material online). These four haplogroups reach their maximum (worldwide) frequencies in the Caucasus (table 2, fig. 2). They are either shared with Near East populations (G2a3b1-P303 and J2a4b*-M67(xM92)) or have ancestral lineages present there (G2a1*-P16(xP18) and J1*-M267(xP58)). Typical European haplogroups are very rare (I2a-P37.2) or limited to specific populations (R1a1a-M198) in the Caucasus.

This pattern suggests unidirectional gene flow from the Near East toward the Caucasus, which could have occurred during the initial Paleolithic settlement or the subsequent Neolithic spread of farming. Archaeological data do not indicate a Near Eastern influence on the Neolithic cultures in the North Caucasus (Bader and Tsereteli 1989; Bzhania 1996; Masson et al. 1982), whereas Neolithization in the Transcaucasus was part of a Neolithic expansion that perhaps paralleled those occurring in Europe (Balaresque et al. 2010) and North Africa (Arredi et al. 2004). However, the current genetic evidence does not allow us to distinguish between Paleolithic and Neolithic models in shaping the genetic landscape of the North Caucasus.

All of these genetic findings are based solely on Y-chromosomal data. This choice was prompted by the high interpopulation variation in the data set (and therefore the best detection of the differences) compared with mitochondrial DNA and autosomal markers. However, one may wonder if the pattern of the entire gene pool is different from its Y-chromosomal subset. In the context of this study, languages are typically learned from the maternal side (we say "mother tongue"; Beauchemin et al. 2010). Thus, the observed similarity between the distributions of languages and genes might become even more evident if full-genome data, incorporating maternally inherited information as well, become available; this possibility may be explored in future studies.

Challenges of Genetic Dating

To estimate the ages of population splits, we employed both "evolutionary" and "genealogical" mutation rates for calibration and used four different methods, namely the ρ estimator, BATWING dating of the clusters and the population splits, and ASD microsatellite variation.

The reliability of the ρ estimator has been explored by Cox (2008) using simulations under a number of demographic models. He found that the mean age is biased only slightly, but the confidence intervals might not contain the true value in 34% cases for the simplest model (constant effective population size, $N_e = 1{,}000$). In some demographic conditions, the error rate increases, particularly when samples sizes are below 25 or when N_e is large, unstable (bottlenecks), or growing. The N_e estimates available for the Caucasus populations are small ($N_e = 187$ on average; Pocheshkhova 2008), which might indicate that the error rate should be low. However, Caucasus populations did grow and bottlenecks could not be excluded. In fact, the most pronounced demographic feature of the Caucasus populations is their high degree of subdivision; fortunately, "error rates of molecular dating with the ρ statistic are unaffected by simple population subdivision" (Cox 2008). Therefore, we might expect ~34% of our clusters (table 5, fig. 6) to have

Table 6. Endogamy Levels[a] for Caucasus and Some European Populations.

Populations	Sample Size (N_T)	Level of Endogamy (%)	Random Inbreeding[b] (f_r ● 10^2)	Method for Estimating Inbreeding	References
Shapsug	5,928	95.2[c]	2.86	Isonymy	Balanovska et al. (2000); Pocheshkhova (2008)
Abkhaz	253	98[d]	3.13	Isonymy	Pocheshkhova (2008)
Circassians	4,438	—	0.61	Isonymy	Pocheshkhova (2008)
Kubachi	182	99.0[e]	1.21	Demography	Bulaeva et al. (2004)
Dargins	350	92.2[e]	1.19	Demography	Bulaeva et al. (2004)
Avar	299	86.7[e]	1.03	Demography	Bulaeva et al. (1990)
Botlikh	248	75.6[e]	0.50	Demography	Bulaeva et al. (1990)
Andi	171	87.0[e]	1.20	Demography	Bulaeva et al. (1990)
Tindal	374	91.0[e]	1.12	Demography	Bulaeva et al. (1990)
Lak	349	81.3[e]	0.011	Demography	Bulaeva et al. (1990)
Lezghins	175	—	0.99	Demography	Bulaeva et al. (1990)
Mormon	625	—	0.19	Demography	Vogel and Motulsky (1986)
Kuban Cossacks	16,056	—	0.06	Isonymy	Pocheshkhova (2008)
French	530,000	—	0.02	Demography	Vogel and Motulsky (1986)
Irish	190,547	—	0.02	Demography	Vogel and Motulsky (1986)
Italians	1,646,612	—	0.07	Demography	Vogel and Motulsky (1986)
Switzerland (four mountain villages)	538	—	0.51	Demography	Vogel and Motulsky (1986)

[a] The table summarizes traditional marriage practices in Caucasus. These data indicate high ethnic endogamy and concomitantly high levels of inbreeding as a consequence.

[b] The random inbreeding f_r was estimated as the proportion of isonymy marriages expected in the panmictic population (Crow and Mange 1965).

[c] The proportion of endogamous marriages was estimated for ethnic group.

[d] The proportion of endogamous marriages was estimated for administrative districts (intra-ethnic level).

[e] The proportion of endogamous marriages was estimated for villages (local level). Most estimates were obtained for villages because inter-ethnic marriages for these groups are typically less than 1%.

actual ages falling outside the indicated confidence intervals. This factor, which randomly affects only one-third of the clusters, would not eliminate the overall agreement with linguistics seen from the figure 6, although it highlights the fact that genetic dates for each particular branch should be taken with caution.

We found that evolutionary estimates of most clusters fall far outside the range of the respective linguistic dates, whereas genealogical estimates gave a good fit with the linguistic dates. At least two population events in the Caucasus are documented archaeologically, which allows additional comparison with these "historical" dates. In both cases, the historical (archaeological) date is similar to a genetic estimate based on the genealogical mutation rate (supplementary note 2, Supplementary Material online). In this regard, a study of the link between Y chromosomes and British surnames working with time intervals close to those analyzed here obtained a mutation rate of 1.5×10^{-3} (King and Jobling 2009). This rate is similar to the genealogical rather than the evolutionary rate and provided good agreement with the historical dates for the surname ages.

The evolutionary rate (Zhivotovsky et al. 2004) was calibrated using two contrasting populations (Maori and Roma). The fact that, in the Caucasus, the genealogical rate provides a better fit with history and linguistics might be partly explained by the dependence of the estimated intralineage variance (and therefore age) on the way that clusters are selected in the network. We selected larger clusters, containing 11 haplotypes on average. However, when the evolutionary calibration was performed (Zhivotovsky et al. 2004), large data sets were not available and clusters contained only two to four haplotypes. For example, the Zhivotovsky et al. (2004) study chose two founders in the Polynesian network (fig. 5*B*), whereas, in our study, we considered similar topologies as a single cluster (see cluster γ in fig. 5*A* for comparison). These subclusters (fig. 5*B*) might be justified in the case of the peopling of New Zealand because they could originate in the homeland before the migration to New Zealand. In our case, however, we considered clusters within the networks because we were interested in the whole history of the clusters that had arisen *in situ* within the Caucasus. To avoid arbitrarily identifying the clusters, we followed a set of formal rules, as described above.

It should be mentioned here that the abundance of population-specific clusters, caused by extremely high endogamy (Table 6) and isolation, seems to be a peculiarity of the Caucasus region which may also have (yet unexplored) effects on the age calculations. Finally, for the BATWING tree (which does not require identifying the clusters), applying the genealogical rate underestimates the dates, whereas applying evolutionary rates overestimates the dates. These comparisons were made with 14 linguistic dates, but more sophisticated modeling and calibrations in other regions are needed to find the most appropriate way to incorporate mutation rate estimates into population-genetic applications. Our study shows that the results could be affected by the method of identifying the clusters and particularly by the chosen methods of dating (ρ, BATWING, SD).

Conclusions

Combining genetic and linguistic findings, we now propose a model of the evolution of the Caucasus populations. The final tree (fig. 6) was obtained by merging the genetic clusters with the background linguistic tree. We conclude that the Caucasus gene pool originated from a subset of the Near Eastern pool due to an Upper Paleolithic (or Neolithic) migration, followed by significant genetic drift, probably due to isolation in the extremely mountainous landscape. This process would result in the loss of some haplogroups and the increased frequency of others. The Caucasus meta-population underwent a series of population (and language) splits. Each population (linguistic group) ended up with one major haplogroup from the original Caucasus genetic package, whereas other haplogroups became rare or absent in it. The small isolated population of the Kubachi, in which haplogroup J1*-M267(xP58) became virtually fixed (99%, table 2), exemplifies the influence of genetic drift there. During population differentiation, haplotype clusters within haplogroups emerged and expanded, often becoming population specific. The older clusters became characteristic of groups of populations. Many younger clusters were specific to individual populations (typically speaking different languages).

We note that the method of inferring the topology of the genetic tree of the Caucasus populations does not require the inference of any mutation rates, and the result is strikingly concordant with the topology of the linguistic tree. Mutation rates are required only for adding a time scale to both trees. Based on the topologies of trees generated from both the genetic and linguistic data, the inference of the parallel evolution of genes and languages in Caucasus is supported, despite controversies about the mutation rates. This study of Caucasus Y-chromosomal variation demonstrates that genetic and linguistic diversification were two parallel processes or, perhaps more precisely, two sides of the same process of evolution of the Caucasus meta-population over hundreds of generations.

Supplementary Material

Supplementary notes 1 and 2, tables 1–3, figures 1–4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/)

Acknowledgments

We thank the people from the Caucasus populations who provided their DNA for the present analysis, Andrey Pshenichnov and Roman Sychev for help in compiling Y-base, Vadim Urasin for collecting STR data from DNA-genealogical websites, Valery Zaporozhchenko for a pilot study of the STR variation of haplogroup G-M201 in the West Caucasus, Kazima B. Bulaeva for consultation about endogamy levels in the East Caucasus, Vladimir Kharkov for help with first version of the phylogenetic network, Matt Kaplan for replication of 20 samples as part of the quality control procedures, Janet Ziegle for technical assistance, and two anonymous reviewers for helpful comments. The first author thanks Richard Villems for instruction in phylogenetic methods and helpful discussions.

This work received primary support from the Genographic Project and additional support from the Russian Foundation of Basic Research (grants 10-07-00515, 10-06-00451, 10-04-01603), and from the Presidium RAS programme "Molecular and Cell Biology". C.T.S. was supported by The Wellcome Trust.

References

Abdushelishvili MG. 1964. Antropology of the ancient and contemporary population of Georgia. Tbilisi (Georgia): Metsniereba

Abramova MP. 1989. The Central Caucasus in the Sarmatian epoch. In: Rybakov BA, editor. The steppes of the European part of the USSR in the Scythian-Sarmatian time. Series Archaeology of the USSR. Moscow (Russia): Nauka. p. 268–281.

Ageeva RA. 2000. Which tribe we are? Ethnic groups of Russia: ethnonims and fortunes. Ethnolinguistic dictionary. Moscow (Russia): Academia Press.

Alexeev VP. 1974. The origin of Caucasus peoples. Moscow (Russia): Mysl.

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet*. 75:338–345.

Bader NO, Tsereteli LD. 1989. Mesolithic in the Caucasus. In: Rybakov BA, editor. Mesolithic of the USSR. Series Archaeology of the USSR. Moscow (Russia): Nauka. p. 93–105.

Balanovsky O, Rootsi S, Pshenichnov A, et al. (11 co-authors). 2008. Two sources of the Russian patrilineal heritage in their Eurasian context. *Am J Hum Genet*. 82:236–250.

Balaresque P, Bowden GR, Adams SM, et al. (16 co-authors). 2010. A predominantly neolithic origin for European paternal lineages. *PLoS Biol*. 8:1–9.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. *Genetics*. 141:743–753.

Battaglia V, Fornarino S, Al-Zahery N, et al. (18 co-authors). 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur J Hum Genet*. 17:820–830.

Beauchemin M, González-Frankenberger B, Tremblay J, et al. (11 co-authors). 2010. Mother and stranger: an electrophysiological study of voice processing in newborns. *Cereb Cortex*. [439810]. doi: 10.1093/cercor/bhq242

Blazhek V, Novotna P. 2008. Retoromanske jazyky: prehled a klasifikace. Linguistica Brunensia. Brno (France): Sbornik praci filozoficke fakulty brnenske univerzity, Masarykova univerzita. Vol. A56, no 1, pp. 15–32.

Bokarev EA. 1981. Comparative-historical phonetics of the East Caucasian languages. Moscow (Russia): Nauka

Bulaeva KB, Isaichev SA, Pavlova TA. 1990. Population-genetics approach to the genetics of human behaviour. *Biomed Sci*. 1:417–424.

Bulaeva KB, Jorde L, Ostler C, Bulaev OA, Pavlova TA, Harpending H. 2004. STR polymorphism in populations of indigenous Daghestan ethnic groups. *Genetika* 40:691–703.

Bulaeva KB, Jorde L, Watkins S, et al. (9 co-authors). 2006. Ethnogenomic diversity of Caucasus, Daghestan. *Am J Hum Biol*. 18:610–620.

Bzhania VV. 1996. Caucasus. In: Rybakov BA, editor. Neolithic of the Northern Eurasia. Series Archaeology of the USSR. Moscow (Russia): Nauka p. 73–86.

Caciagli L, Bulayeva K, Bulayev O, Bertoncini S, Taglioli L, Pagani L, Paoli G, Tofanelli S. 2009. The key role of patrilineal inheritance in shaping the genetic variation of Dagestan highlanders. *J Hum Genet*. 54:689–694.

Chirikba VA. 1996. Common West Caucasian: the reconstruction of its phonological system and parts of its lexicon and morphology. Leiden (The Netherlands): CNWS Publications.

Cinnioglu C, King R, Kivisild T, et al. (15 co-authors). 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet*. 114:127–148.

Comrie B. 1987. The world's major languages. New York: Oxford University Press.

Cox MP. 2008. Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. *Hum Biol*. 80:335–357.

Crow JF, Mange AP. 1965. Measurement of inbreeding from frequency of marriages between person of the same surname. *Eug Quart*. 12:199–203.

Cruciani F, La Fratta R, Santolamazza P, et al. (19 co-authors). 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet*. 74:1014–1022.

Cruciani F, La Fratta R, Torroni A, Underhill PA, Scozzari R. 2006. Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. *Hum Mutat*. 27:831–832.

Cruciani F, La Fratta R, Trombetta B, et al. (24 co-authors). 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol*. 24:1300–1311.

Di Giacomo F, Luca F, Popa LO, et al. (27 co-authors). 2004. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum Genet*. 116:529–532.

El-Sibai M, Platt DE, Haber M, et al. (38 co-authors). 2009. Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. *Ann Hum Genet*. 73:568–581.

Embleton S. 2000. Lexicostatistics. Glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In: Renfrew C, McMahon A, Trask L, editors. Time depth in historical linguistics. Cambridge (UK): The McDonald Institute for Archaeological Research Press. p. 143–165.

Fedorov YA. 1983. Historical ethnography of the North Caucasus. Moscow (Russia): Moscow University Press.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol*. 128:415–423.

Flores C, Maca-Meyer N, Larruga JM, Cabrera VM, Karadsheh N, Gonzalez AM. 2005. Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. *J Hum Genet*. 50:435–441.

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

Ge J, Budowle B, Aranda XG, (6 co-authors). 2009. Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci Int Genet*. 3:179–184.

Gerasimova MM, Rud' NM, Yablonskij LT. 1987. Anthropological data to the issue of ethnic relations in the North-East Black Sea (Bosporus kingdom). In: Anthropology of ancient and medieval populations of Eastern Europe. Moscow (Russia): Nauka p. 79–82.

Gigeneishvili BK. 1977. Comparative phonetics of the Dagestan languages. Tbilisi (Georgia): Tbilisi University Press.

Gray RD, Atkinson QD. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature* 426:435–439.

Greenhill SJ, Atkinson QD, Meade A, Gray RD. The shape and tempo of language evolution. *Proc Biol Sci*. 277:2443–2450.

Gusmao L, Sánchez-Diz P, Calafell F, et al. (42 co-authors). 2005. Mutation rates at Y chromosome specific microsatellites. *Hum Mutat*. 26:520–528.

Haber M, Platt DE, Badro DA, et al. (13 co-authors). Forthcoming 2010. Influences of history, geography and religion on genetic structure: the Maronites in Lebanon. *Eur J Hum Genet*. 19:334–340.

King TE, Jobling MA. 2009. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Mol Biol Evol*. 26:1093–1102.

Kitchen A, Ehret C, Assefa S, et al. (4 co-authors). 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc Biol Sci*. 276:2703–2710.

Kuipers AN. 1963. Caucasian. In: Sebeok T, editor. Current trends in linguistics. Soviet and East European Linguistics. The Hague (The Netherlands): Mouton. p. 315–344.

Kutuev IA, Litvinov SS, Yusunbaev BB, Khusainova RI, Valiev RR, Villems R, Khusnutdinova EK. 2010. The genetic structure and molecular phylogeography of Caucasus populations based on Y chromosome data. *Medicinskaya Genetika*. 9:18–25.

Manni F, Guerard E. 2004. Barrier vs. 2.2. (computer program). Paris (France): Population Genetics Team, Museum of Mankind (Musee de l'Homme).

Masson VM, Merpert NY, Munchaev RM, Chernysh EK. 1982. Chalcolithic of the USSR. In: Series Archaeology of the USSR. Rybakov BA, editor. Moscow (Russia): Nauka

Melyukova AI. 1989. Conclusions. In: Rybakov BA, editor. The steppes of the European part of the USSR in the Scythian-Sarmatian time. Series Archaeology of the USSR. Moscow (Russia): Nauka. p. 292–295.

Munchaev RM. 1994. Maikop culture. In: BA Rybakov, editor. The bronze age of the Caucasus and Central Asia, Series Archaeology of the USSR. Moscow (Russia): Nauka. p. 158–225.

Nasidze I, Ling EY, Quinque D, et al. (17 co-authors). 2004a. Mitochondrial DNA and Y-chromosome variation in the caucasus. *Ann Hum Genet*. 68:205–221.

Nasidze I, Quinque D, Dupanloup I, et al. (7 co-authors). 2004b. Genetic evidence concerning the origins of South and North Ossets. *Ann Hum Genet*. 68:588–599.

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

Nikolaev SL, Starostin SA. 1994. A North Caucasian Etymological Dictionary. Moscow (Russia): Asterisk Publishers.

Pocheshkhova EA. 2008. Structure of migrations and gene drift in populations of Adyghes-Shapsugs. *Medicinskaya Genetika*. 7:30–38.

Powell R, Gannon F. 2002. Purification of DNA by phenol extraction and ethanol precipitation. New York: Oxford University Press.

Rosser ZH, Zerjal T, Hurles ME, et al. (62 co-authors). 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet*. 67:1526–1543.

Ruhlen MA. 1987. Guide to the World's Languages. Classification. V. 1. Stanford (CA): Stanford University Press.

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Nørby S. 2000. MtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet*. 67:718–726.

Sánchez-Diz P, Alves C, Carvalho E, et al. (18 co-authors). 2008. Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study. *Int J Legal Med*. 122:529–533.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin vers. 2.000: a software for population genetics data analysis. Geneva (Switzerland): Genetics and Biometry Laboratory, Department of Anthropology and Ecology. University of Geneva.

Semino O, Magri C, Benuzzi G, et al. (16 co-authors). 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet*. 74:1023–1034.

Semino O, Passarino G, Oefner PJ, et al. (17 co-authors). 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* 290:1155–1159.

Sengupta S, Zhivotovsky LA, King R, et al. (15 co-authors). 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. *Am J Hum Genet*. 78:202–221.

Shagirov AK. 1977. The etymological dictionary of Adyghe (Circassian) languages. Moscow (Russia): Nauka

Starostin SA. 1989. Comparative-historical linguistics and lexicostatistics. In: Renfrew C, McMahon A, Trask L, editors. Linguistic reconstruction and the ancient history of the East. Moscow (Russia): Nauka. p. 2–39.

Starostin SA. 2000. Comparative-historical linguistics and lexicostatistics. In: Renfrew C, McMahon A, Trask L, editors. Time depth in historical linguistics. Cambridge (UK): The McDonald Institute for Archaeological Research Press. p. 223–265.

StatSoft Inc, Tulsa, OK. 2001. STATISTICA (data analysis software system), version 6. Available from: www.statsoft.com.

Talibov BB. 1980. Comparative phonetics of Lezghins languages. Moscow (Russia): Nauka

Tofanelli S, Ferri G, Bulayeva K, et al. (23 co-authors). 2009. J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur J Hum Genet.* 17:1520–1524.

Trubetzkoy NS. 1930. Nordkaukasische Wortgleichungen. Wiener Zeitschrift für die Kunde des Morgenlandes. Bd XXXVII, Heft 2. Wien. p. 76.

Underhill PA, Myres NM, Rootsi S, et al. (34 co-authors). 2010. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur J Hum Genet.* 18:479–484.

Vogel F, Motulsky AG. 1986. Human genetics. Problems and approaches. Berlin (Germany): Springer-Verlag.

Wells RS, Yuldasheva N, Ruzibakiev R, et al. (27 co-authors). 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A.* 98:10244–10249.

Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J R Stat Soc.* 166:155–201.

Womble WH. 1951. Differential systematics. *Science* 114:315–322.

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. 2002. A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet.* 71:466–482.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, et al. (17 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet.* 74:50–61.

Appendix

Genographic Consortium members: Syama Adhikarla (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Christina J. Adler (University of Adelaide, South Australia, Australia), Danielle A. Badro (Lebanese American University, Chouran, Beirut, Lebanon), Jaume Bertranpetit (Universitat Pompeu Fabra, Barcelona, Spain), Andrew C. Clarke (University of Otago, Dunedin, New Zealand), David Comas (Universitat Pompeu Fabra, Barcelona, Spain), Alan Cooper (University of Adelaide, South Australia, Australia), Clio S. I. Der Sarkissian (University of Adelaide, South Australia, Australia), Matthew C. Dulik (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Christoff J. Erasmus (National Health Laboratory Service, Johannesburg, South Africa), Jill B. Gaieski (University of Pennsylvania, Philadelphia, Pennsylvania, USA), ArunKumar GaneshPrasad (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Angela Hobbs (National Health Laboratory Service, Johannesburg, South Africa), Asif Javed (IBM, Yorktown Heights, New York, USA), Li Jin (Fudan University, Shanghai, China), Matthew E. Kaplan (University of Arizona, Tucson, Arizona, USA), Shilin Li (Fudan University, Shanghai, China), Begoña Martínez-Cruz (Universitat Pompeu Fabra, Barcelona, Spain), Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand), Marta Melé (Universitat Pompeu Fabra, Barcelona, Spain), Nirav C. Merchant (University of Arizona, Tucson, Arizona, USA), R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia), Amanda C. Owings (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Laxmi Parida (IBM, Yorktown Heights, New York, USA), Ramasamy Pitchappan (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Lluis Quintana-Murci (Institut Pasteur, Paris, France), Daniela R. Lacerda (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Ajay K. Royyuru (IBM, Yorktown Heights, New York, USA), Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa), Pandikumar Swamikrishnan (IBM, Somers, New York, USA), Kavitha Valampuri John (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), Janet S. Ziegle (Applied Biosystems, Foster City, California, USA).