THE ROLE OF CAUCASIAN, IRANIAN AND STEPPE POPULATIONS IN SHAPING THE DIVERSITY OF AUTOSOMAL GENE POOL OF THE EASTERN CAUCASUS

Balanovska EV[1,2], Gorin IO[1], Petrushenko VS[1], Ponomarev GYu[1], Belov RO[1], Pocheshkhova EA[1,3], Salaev VA[1], Iskandarov NA[1], Pylev VYu[1,2] ✉

[1] Bochkov Research Centre of Medical Genetics, Moscow, Russia
[2] Biobank of North Eurasia, Moscow, Russia
[3] Kuban State Medical University, Krasnodar, Russia

Eastern Caucasus is home to more than 30 peoples speaking Caucasian, Iranian and Turkic languages. Fusion of multiple migration flows together with the complex population structure of the Eastern Caucasus make it more difficult to analyze its gene pool: this is the most poorly studied one among all regions of the Caucasus. The study is aimed to identify the main patterns of the autosomal gene pool variation in this region. A total of 356 genomes of 29 ethnic groups were studied using the large panels of SNP markers: 243 genomes of 22 peoples of the Eastern Caucasus and 113 genomes of 7 peoples living in adjacent regions. The bioinformatics analysis involved the use of the ADMIXTURE ancestral component method and the gene pool variability principal component analysis (PCA). The hypothesis of three genetic strata, the interaction of which forms the structure of gene pool of the Eastern Caucasus, was put forward. The "Dagestan" stratum carries information about the gene pool of the ancient autochthonous population of the Eastern Caucasus. The "Iranian" stratum represents the legacy of ancient and middle-aged migrations surges of the Iranian-speaking population: it constitutes three quarters of the gene pool of modern Azerbaijan and about one third of the Dagestan peoples' gene pool. The "Steppe" stratum represents a negligible influence of the Eurasian steppe. Interaction of three genetic strata is only indirectly related to the peoples' linguistic affiliation, however, the association with linguistics is more obvious in the Caucasian-speaking peoples. Four genetically distinct groups of indigenous population of the Eastern Caucasus have been identified, the combination of which should be included in the characteristics of its autosomal gene pool: 1) Dargins, Laks; 2) Avars, Lezghins, Tabasarans, Aghuls, Rutul people, Tsakhur people; 3) Kumyks, Tat people and Azerbaijanis living in Dagestan; 4) Azerbaijanis and Talysh living in Azerbaijan. The directions of further research have been defined.

Keywords: gene geography, gene pool, population genetics, autosomal SNP markers, Eastern Caucasus, population

Funding: the study was supported by the RSF grant № 21-74-00156 (bioinformatics analysis of gene pools of the Eastern Caucasus and Transcaucasia), State Assignment of the Ministry of Science and Higher Education of the Russian Federation for the Research Centre of Medical Genetics (cartographic analysis, data interpretation), Biobank of North Eurasia (expedition survey).

Acknowledgements: the authors would like to thank all members of the expedition survey (sample donors) and the autonomous non-profit organization "Biobank of North Eurasia" for access to DNA collections and the genotyping results.

Author contribution: Balanovska EV — management, design, manuscript writing and expedition survey of the Dagestani peoples; Gorin IO, Petrushenko VS — bioinformatics analysis; Ponomarev GYu — work with DNA collections, cartographic analysis; Belov RO — work with DNA collections, manuscript formatting; Pocheshkhova EA — expedition survey of the Dagestani peoples; Salaev VA — organization and conducting the expedition survey of the Talysh living in Azerbaijan; Iskandarov NA — organization and conducting the expedition survey of the Azerbaijanis living in Azerbaijan; Pylev VYu — organization of genotyping, statistical analysis.

Compliance with ethical standards: the study was approved by the Ethics Commitee of the Bochkov Research Centre of Medical Genetics (protocol № 1 of 29 June 2020).

✉ **Correspondence should be addressed:** Vladimir Yu. Pylev
Moskvorechye, 1, 115522, Moscow, Russia; freetrust@yandex.ru

Received: 26.04.2023 **Accepted:** 13.05.2023 **Published online:** 29.05.2023

DOI: 10.24075/brsmu.2023.017

РОЛЬ КАВКАЗСКОГО, ИРАНСКОГО И СТЕПНОГО НАСЕЛЕНИЯ В ФОРМИРОВАНИИ МНОГООБРАЗИЯ АУТОСОМНОГО ГЕНОФОНДА ВОСТОЧНОГО КАВКАЗА

Е. В. Балановская[1,2], И. О. Горин[1], В. С. Петрушенко[1], Г. Ю. Пономарёв[1], Р. О. Белов[1], Э. А. Почешхова[1,3], В. А. Салаев[1], Н. А. Искандаров[1], В. Ю. Пылёв[1,2] ✉

[1] Медико-генетический научный центр, Москва, Россия
[2] Биобанк Северной Евразии, Москва, Россия
[3] Кубанский государственный медицинский университет, Краснодар, Россия

На Восточном Кавказе проживают более 30 народов, говорящих на кавказских, иранских и тюркских языках. Слияние многих миграционных потоков и сложная популяционная структура Восточного Кавказа затрудняют анализ его генофонда: из всех регионов Кавказа он наименее изучен. Цель работы — выявить основные закономерности в изменчивости аутосомных генофондов этого региона. По обширным панелям SNP-маркеров изучено 356 геномов 29 этносов: 243 генома 22 народов Восточного Кавказа и 113 геномов 7 народов окружающих регионов. Биоинформатический анализ проведен методами предковых компонент ADMIXTURE и главных компонент изменчивости генофонда (PCA). Выдвинута гипотеза трех основных пластов генофонда Восточного Кавказа, взаимодействие которых формирует его структуру. «Дагестанский» пласт несет информацию о генофонде древнего автохтонного населения Северного Кавказа. «Иранский» пласт отражает наследие древних и средневековых волн миграций ираноязычного населения: он составляет три четверти генофонда современного Азербайджана и около трети генофонда народов Дагестана. «Степной» пласт фиксирует слабое влияние евразийской степи. Взаимодействие трех генетических пластов лишь косвенно связано с языковой принадлежностью народов, но у кавказоязычных народов связь с лингвистикой проявляется ярче. Выявлены четыре генетически своеобразные группы коренного населения Восточного Кавказа, комплекс которых должен включаться в характеристику его аутосомного генофонда: 1) даргинцы, лакцы; 2) аварцы, лезгины, табасараны, агулы, рутульцы, цахуры; 3) кумыки, таты и азербайджанцы Дагестана; 4) азербайджанцы и талыши Азербайджана. Определены направления дальнейших исследований.

Ключевые слова: геногеография, генофонд, популяционная генетика, аутосомные SNP-маркеры, Восточный Кавказ, народонаселение

Финансирование: исследование выполнено при поддержке гранта РНФ №21-74-00156 (биоинформационный анализ генофондов Восточного Кавказа и Закавказья), Государственного задания Министерства науки и высшего образования РФ для Медико-генетического научного центра им. академика Н. П. Бочкова (картографический анализ, интерпретация результатов), Биобанка Северной Евразии (экспедиционное исследование).

Благодарности: авторы благодарят всех участников экспедиционного обследования (доноров образцов), АНО «Биобанк Северной Евразии» — за предоставление коллекций ДНК и результатов генотипирования.

Вклад авторов: Е. В. Балановская — руководство, дизайн и написание статьи, организация и проведение экспедиционного обследования народов Дагестана; И. О. Горин, В. С. Петрушенко — биоинформатический анализ; Г. Ю. Пономарев — работа с ДНК-коллекциями, картографический анализ; Р. О. Белов — работа с ДНК-коллекциями, оформление статьи; Э. А. Почешхова — проведение экспедиционного обследования народов Дагестана; В. А. Салаев — организация и проведение обследования талышей Азербайджана; Н. А. Искандаров — организация и проведение обследования азербайджанцев Азербайджана; В. Ю. Пылёв — организация генотипирования, статистический анализ.

Соблюдение этических стандартов: исследование одобрено этическим комитетом Медико-генетического научного центра имени Н. П. Бочкова (протокол № 1 от 29 июня 2020 г.).

✉ **Для корреспонденции:** Владимир Юрьевич Пылёв
ул. Москворечье, д. 1, 115522, г. Москва, Россия; freetrust@yandex.ru

Статья получена: 26.04.2023 **Статья принята к печати:** 13.05.2023 **Опубликована онлайн:** 29.05.2023

DOI: 10.24075/vrgmu.2023.017

The gene pool of the Caucasian peoples has long attracted the attention of population genetic scientists. This region, small in area, that is located between Europe and Asia, is home to more than 60 peoples speaking languages of three linguistic families: Caucasian, Indo-European, and Altaic. Among all regions of the Caucasus, the Eastern Caucasus, where the largest number of peoples (more than 30) representing three language families is concentrated, is the most poorly studied region in terms of genetics. Since the Western Caspian Sea region served as a bridge between Europe and Asia over millennia, it is necessary to involve data on both steppes in the north of the region and populations of the Iranian plateau to understand the gene pool of the Eastern Caucasus. Bizarre structure of peoples of the Eastern Caucasus together with fusion of multiple migration flows make it extremely difficult to analyze its gene pool. While there are some papers on genetics of the populations of Dagestan, the population of Azerbaijan represents one of the largest blank spots in the genetic map. And it's the key to understanding the centuries-old influence of Persia on the gene pools of the Eastern Caucasus: did this influence extend throughout the Caspian Sea region or was it concentrated only on the southern border gene pools? The genetic history of a number of small peoples living in the North Caucasus can also be reconstructed only by the systematic genetic study of the Eastern Caucasus in the context of knowledge about the peoples of Iran.

The complex structure of the multi-ethnic region, the Eastern Caucasus, requires thorough analysis. That is why our study is focused on its autosomal gene pool only. In parallel, we will publish a paper on the Y-chromosome variation in the same populations of the Eastern Caucasus. Such “binocular vision” will make it possible to get the most impartial and fair picture of the gene pool variation in the Eastern Caucasus.

There are very little published data on the populations of Dagestan and Azerbaijan obtained using the genome-wide panels that are among the most popular and effective systems of DNA markers over the last 10 years. In almost all of these papers, the data on the Eastern Caucasus were not analyzed separately, these were just an integral part of research focused on much larger region, the entire Caucasus or Eurasia.

The populations of Dagestan were assessed using the Illumina panel (~600,000 SNPs) in the study [1] of gene pool of the Caucasus as a whole, and using the Human Origin panel (~600,000 SNPs) [2].

The gene pool of Azerbaijanis living in the Northwestern Iran is discussed in the paper on the traces of Turkic expansion [3], in which they showed a 5% contribution of the East Eurasian ancestral component that could be explained by the spread of Turkic languages westwards in accordance with the “elite dominance” model (language change without significant gene pool changes). Assessment of Azerbaijanis living in Azerbaijan in comparison with other gene pools using the Human Origin genome-wide panel is provided only in paper [2].

The majority of papers discuss a broad spectrum of issues: formation of the Caucasian gene pool based on the contribution of migration in the Middle East [1], legacy of the Turkic-speaking groups' migration [3] or the eco-geographical zoning of North Eurasia [2]. However, none of the papers is focused on assessing the features of gene pool of the Eastern Caucasus. Certain small samples of peoples of the Eastern Caucasus are represented in other two large studies focused on the completely different issues [4, 5]. Unfortunately, some of the above samples were assessed using only the HumanOrigin panel (Affimetrix) that was hardly comparable with the Illumina panels.

Among studies conducted in recent years, the whole genome and whole exome studies of the gene pools of Iran and Turkey should be noted. The authors of one of the papers [6] study genomic variation in peoples of Iran. Since the study is focused on compiling the database on genomic variation in Iran, there is much emphasis on the genetic structure of Iran itself, while adjacent regions (specifically Azerbaijan) are not addressed in depth. A similar situation in the paper [7], where the genetic structure of the populations of Turkey is assessed, however, the impact of the Eastern Caucasus is described only in a short comment about its admixture with certain populations of Turkey.

In general, the published data contain genotypes of only 43 samples obtained from peoples of the Eastern Caucasus. These samples were assessed using mainly small Illumina panels, while in our study we assessed 243 samples using a large Illumina panel.

Thus, the world literature on genome-wide panels provides no conclusions about the gene pool of the Eastern Caucasus and reports just a few data that are restricted to specific groups. In contrast, our paper discusses populations of 22 peoples living in the Eastern Caucasus.

In the paper of our team focused on the search for traces of Alans in the autosomal gene pools of the North Caucasus [8] the main attention is paid to the ethnic groups of the Central and Western Caucasus. The gene pool of peoples living in the Eastern Caucasus is represented by four Dagestani peoples. The results obtained highlight the need for the detailed review of this data on the gene pool of the Eastern Caucasus, as well as for the targeted analysis of autosomal gene pools of all peoples living in this region and identification of the main patterns underlying variation, which is the aim of our study.

METHODS

In this paper the whole population of Dagestan (17 ethnic groups), Azerbaijan (Azerbaijanis, Azerbaijanis-Karapapakhs, Talysh) together with the other Iranian-speaking Kurds and Yazidis studied in the Caucasus (including migrants from various populations of the Caucasus and Iran; in our study they are considered as representatives of the Iranian-speaking population of the Eastern Caucasus, which is part of their ethnic range) are referred to as “peoples of the Eastern Caucasus”. These 22 ethnic groups of the Eastern Caucasus are represented (Table 1) by original data on 243 genomes obtained by our team. These are assessed in the context of original data on the neighbouring peoples living in the Central Caucasus (Chechens), Transcaucasia (Georgians), Caspian Sea region (Astrakhan Nogais), Transcaspia (Karakalpaks, Turkmens), as well as the literature data on the peoples of Iran, using the same panel of SNP markers [3, 9].

The overall sample size was 356 genomes of 29 ethnic groups: original data (318 genomes of 27 ethnic groups) were assessed using the Illumina4M and Illumina750K panels; the literature data (38 genomes of 2 ethnic groups) were assessed using the Illumina750K and HumanOrigin panels. PLINK 1.9 [10] was used for filtering by quality of the genome reads; kinship of individuals (below the 3rd degree according to KING 2.3.0 [11]); DNA markers' linkage and monomorphism.

The principal component analysis (PCA) of genomic variation was performed using the smartpca utility of the EIGENSTRAT software package [12]. Conversion from the plink format (bed-bim-fam) to the eigensoft format (eigenstratgeno-

Table 1. Linguistic affiliation and number of the studied genomes

LINGUISTIC CLASSIFICATION				Peoples	N of genomes
Family	Branch	Sub-branch	Group		
Caucasian	Nakho-Dagestanian	Lezgin-Dargin-Lak	Dargin	Kubachins	6
				Dargins	8
				Kaitags	8
			Lak	Laks	11
			Lezgin	Tabasarans	11
				Aghuls	1
				Rutul people	9
				Tsakhur people	8
				Lezgins	43
		Avar-Ando-Cesian	Avar	Avars	7
			Andian	Tindi people	6
			Cesian (Didoi)	Didoi people	5
				Hinukh people	5
		Nakh		Chechens	15
	Kartvelian	South-Kartvelian		Georgians	19
Altai	Turkic	Polovtsian-Kipchak	Caucasian	Kumyks	27
		Oguz	Essentially Oguz	Azerbaijanis: Dagestan,	9
				Azerbaijan,	13
				Iran*	18
				Karapapakhs	6
				Turkmens	19
		Kipchak	Kipchak-Nogai	Karanogai people	11
				Astrakhan Nogais	5
				Karakalpaks	17
Indo-European	Aryan	New Iranian	South-Western	Tat people living in Dagestan	13
				Iranian-speaking peoples of Iran**	20
			North-Western	Talysh	10
				Yazidis	10
				Kurds	16

Note: literature data on peoples of Iran [3, 9].

snp-ind) was carried out with the convertf tool of the same software package (using the default settings). Calculation was performed for five principal components with five iterations of outlier removal, the results were visualized in Python 3 using the pandas [13], matplotlib [14], and seaborn [15] libraries. The centroid for each population was defined (and designated by a larger dot) for each population on the principal component plot. It was determined as an average of all calculated components for all samples included in the population.

Analysis of the ancestral components by the ADMIXTURE method was performed using ADMIXTURE v1.3.0 [16], the number of the modelled ancestral components (K) varied between 2 and 20. Cross-validation is performed for each K-value in order to estimate error. The ADMIXTURE results were visualized in Python 3 using the *pandas*, *matplotlib*, and *seaborn* libraries.

RESULTS

Preliminary assessment of contributions of the ADMIXTURE ancestral components to the gene pools of four Dagestan peoples

In the earlier published report [8] the autosomal gene pool of the Eastern Caucasus was represented by genomes of the Dagestan peoples: Caucasian-speaking (Dargins, Laks, Tabasarans) and Iranian-speaking (Tat people). In the ADMIXTURE model of ancestral components (ACs) with K = 11 the contribution of the Dagestanian AC to these genomes was 77%. But the question remains, how valid is such pooling? Genomes of which large peoples of the Eastern Caucasus can represent it correctly in the study of the large regions of Eurasia?

Changes in the contribution of the "Dagestanian" component with increasing number of ACs (Fig. 1) reveals the differences

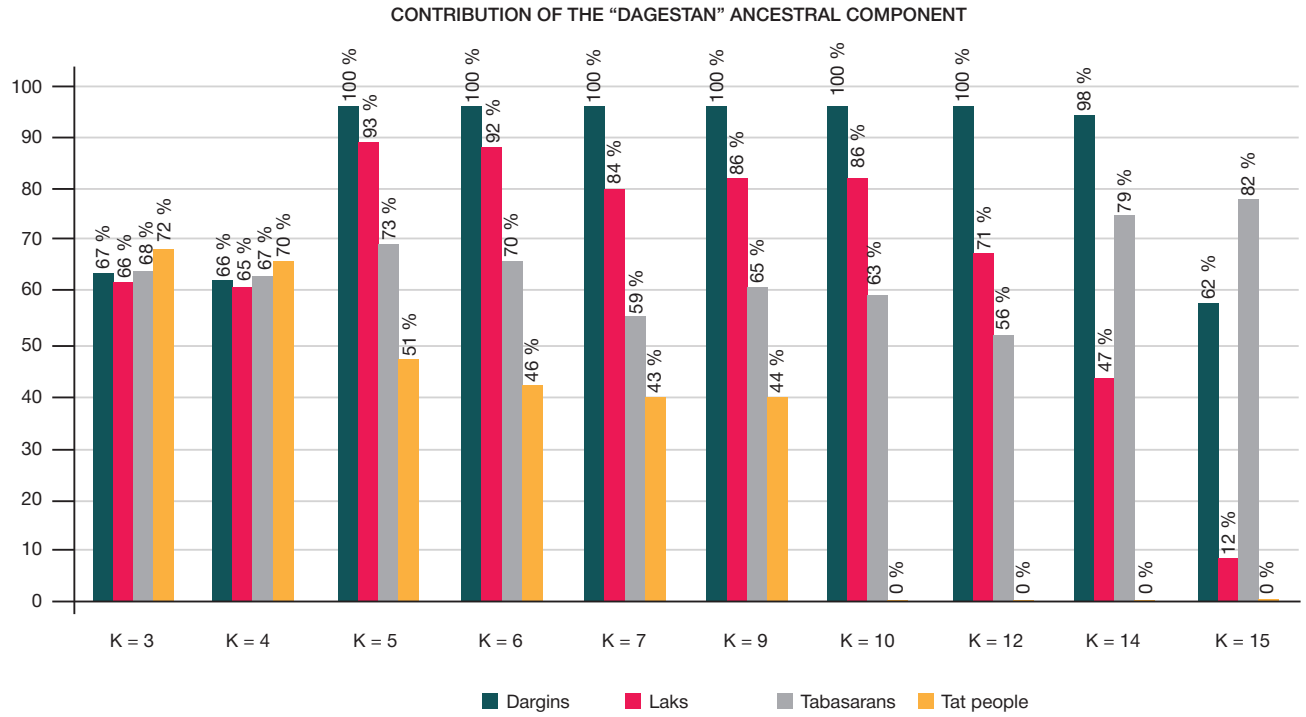

Fig. 1. Dynamic changes in the contribution of the "Dagestan" ADMIXTURE ancestral component to the genomes of Dargins, Laks, Tabasarans, and Tat people in the models with K-values between 3 and 15

between ethnic groups even with a small number of ACs (K = 5). When K = 10, the Tat people of Dagestan form their own AC, separating from the Caucasian-speaking peoples. Furthermore (Fig. 1), Dargins are the major contributors to the gene pool of Dagestan: the contribution of the "Dagestanian" AC to their genomes within the interval of 4 < K < 15 is 98–100%, it drops to 62% only when K = 15. Laks demonstrate different dynamics: when the K-value increases from 5 to 15, the contribution of the "Dagestanian" AC to their genomes drops from 93 to 29%, as the Laks' AC shows itself. Tabasarans demonstrate a specific pattern: when the K-value increases from 5 to 12, the contribution of the "Dagestanian" AC to their genomes drops from 73 to 56% and then increases to 82% when K = 15.

The comparison has revealed an unexpected phenomenon: the Dagestanian AC represents the gene pools of different ethnic groups of Dagestan with different number of ACs (3 < K < 15). Their contributions are equal only when the number of ACs is low (K = 3 and K = 4), however, when K > 4 the contributions of the Tat people and Laks gradually diminish, and the contributions of Dargins and Tabasarans change. Such results show that it is necessary to generate the genetic portraits of autosomal gene pools for each people of the Eastern Caucasus and then form clusters of ethnic groups to be used as the basis for analysis of autosomal gene pools of the Caucasus and other large regions of Eurasia.

Position of 22 peoples of the Eastern Caucasus in the principal component space

To answer the above questions correctly, it is important to extend the study to the broadest possible spectrum of peoples living in the Eastern Caucasus and to use basic independent methods for population genetics analysis.

Fig. 2 shows positions of 22 peoples of the Eastern Caucasus and six reference groups within the space of the gene pool variation principal components (PC) 1 and 2. Six clusters of genomes are clearly distinguished showing that similarity of peoples living in the Eastern Caucasus loosely matches classification of their languages. Almost all clusters include peoples speaking not only different branches of languages, but the languages of different linguistic families (Table 2). The first PC clearly divides all peoples into steppe peoples and all other peoples. In contrast, the second PC demonstrates a long chain containing all other genomes, from Kubachins to Kurds (Fig. 2).

The **Dargin-Lak-Ando-Cesian cluster** includes representatives of five ethnic groups of four groups speaking different sub-branches of the Nakho-Dagestanian languages (Table 2): Dargins, Kaitags, Kubachins, Laks, Tindi people, Didoi people, Hinukh people. It should be noted that the genomes of isolated populations differ sharply from the whole set by other PCs: the Didoi people and Hinukh people differ by PC 3, the ethnic group of Kubachins living in one mountain village differs by PC 4, and the Tindi people differ by PC 5.

The **Lezgin-Avar cluster** includes representatives of six ethic groups speaking both sub-branches of the Nakho-Dagestanian languages (Table. 2). In the Avar group only Avars have been included in the cluster. Despite the fact that their centroid is located among the genomes of Lezgin peoples, individual genomes of Avars are extremely diverse (Fig. 2; red dots): these stretch along the entire Lezgin-Avar cluster and extend to the Ando-Cesian one beyond its boundaries. As for Lezgin peoples, no differences between genomes of the East-Lezghin (Lezghins,Tabasarans, Aghuls) and the Rutul-Tsakhur (Rutul people, Tsakhur people) subgroups are observed. Furthermore, the genomes of Lezghins (purple dots in Fig. 2) extend to the other cluster that includes the Tat people and Azerbaijanis living in Dagestan.

The **Turkic-Iranian cluster of Dagestan** brings together the genomes of the Iranian-speaking Tat people and the Turkic-speaking Kumyks and Azerbaijanis living in Dagestan (Table 2). Similarity of genomes of the Tat people and Azerbaijanis of Dagestan is no surprise, since there was a tradition in Dagestan to register the Tat people as Azerbaijanis, and the boundary was drawn based on their compact settlement in Dagestan. The genomes of all three ethnic groups are extremely diverse and

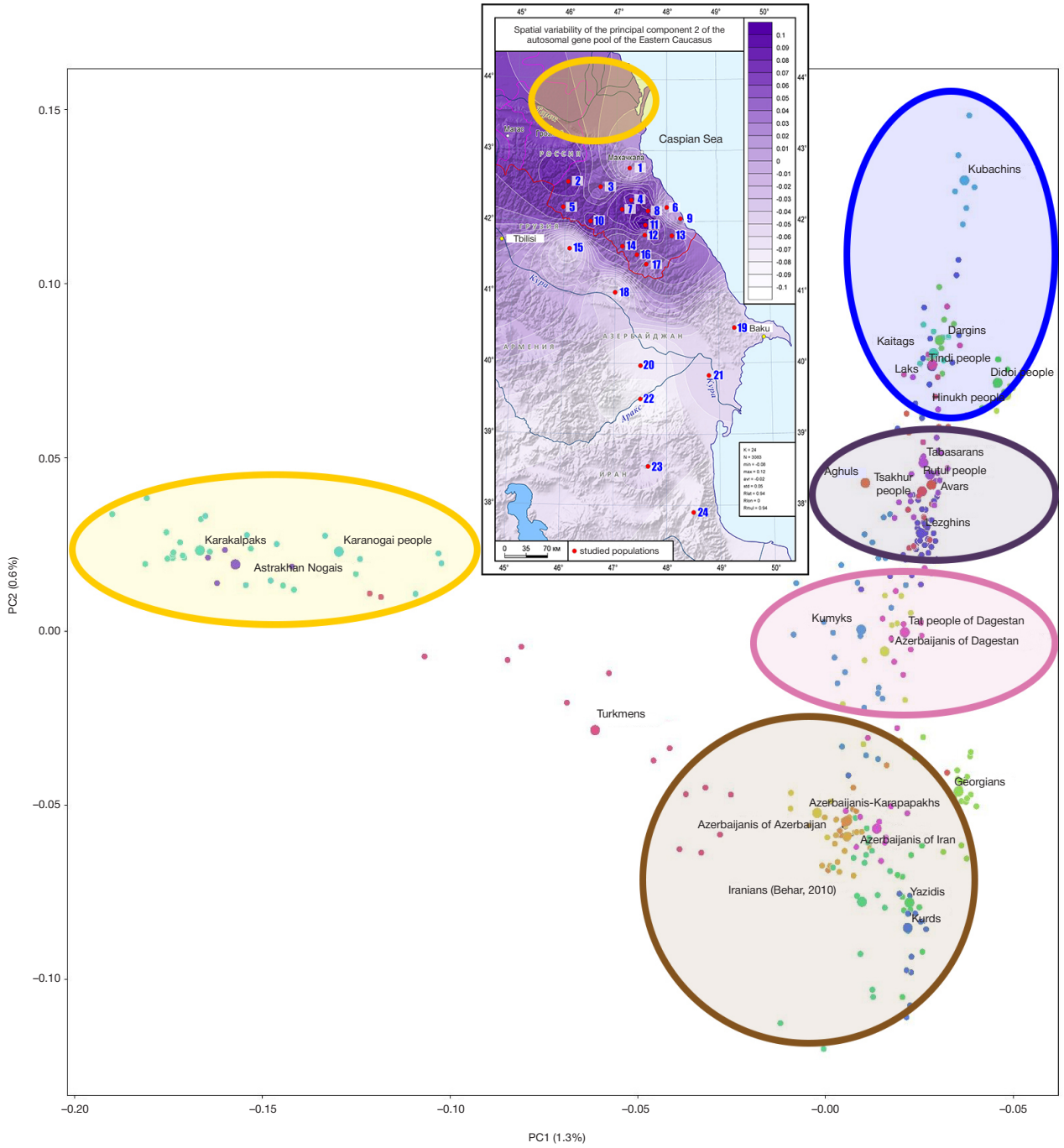

Fig. 2. Plot of principal components 1–2 (PCA) of the genome variability in peoples of the Eastern Caucasus in the context of neighboring populations. The map of the principal component 2 is provided in the inset. 1 — Kumyks; 2 — Tindi people; 3 — Avars; 4 — Dargins; 5 — Hinukh people; 6 — Azerbaijanis living in Dagestan; 7 — Laks; 8 — Kaitags; 9 — Tat people living in Dagestan; 10 — Didoi people; 11 — Kubachins; 12 — Aghuls; 13 — Tabasarans; 14 — Tsakhur people; 15 — Georgians; 16 — Rutul people; 17 — Lezghins; 18 — Azerbaijanis-Karapapakhs; 19 — Azerbaijanis living in Azerbaijan; 20 —Yazidis; 21 — Talysh; 22 — Kurds; 23 — Azerbaijanis living in Iran [3]; 24 — Iranian-speaking Iranians [9]

gravitate towards the Azerbaijan-Iranian cluster. However, Kumyks, that are also included in the Lezgin-Avar cluster, demonstrate the excess diversity of genomes (Fig 2; blue dots).

The **Azerbaijan-Iranian cluster of Dagestan** brings together the genomes of the Turkic-speaking (Azerbaijanis living in Azerbaijan and Iran) and Iranian-speaking (Talysh, Kurds, Yazidis and the aggregate group of Iran) peoples too. Based on PC 1, only one group of Azerbaijanis (Karapapakhs) shows a slight shift towards peoples of the Eurasian steppe. However, the genomes of Turkmens (Fig 2; pink dots) have become a bridge between the Azerbaijan-Iranian and steppe clusters. The other pole of the Azerbaijan-Iranian cluster is fixed by the Georgian genomes that move closer to peoples of the North Caucasus in other PC variants.

The **steppe cluster** brings together the Karanogai people living in Dagestan and other steppe peoples of the Caspian steppe (Astrakhan Nogais) and Transcaspia (Karakalpaks). The sharp difference shown by the steppe cluster genomes generates the differences by PC 1 reflecting the greatest variability of the assessed genomes.

Spatial variability of principal components

On the map of PC 2 (Fig. 2; inset) we have placed a yellow oval within the range of the Karanogai people in the northern

Table 2. Clusters in the genetic space of principal components 1–2 (PCA) and ethno-linguistic affiliation of genomes in each cluster

Cluster	LINGUISTIC CLASSIFICATION				Peoples
	Family	Branch	Sub-branch	Group	
Dargin-Lak-Ando-Cesian	Caucasian	Nakho-Dagestanian	Lezgin-Dargin-Lak	Dargin	Kubachins, Dargins,
					Kaitags
				Lak	Laks
			Avar-Ando-Cesian	Andian	Tindi people
				Cesian (Didoi)	Didoi people
					Hinukh people
Lezgin-Avar	Caucasian	Nakho-Dagestanian	Lezgin-Dargin-Lak	Lezgin	Tabasarans
					Aghuls
					Rutul people
					Tsakhur people
					Lezgins
			Avar-Ando-Cesian	Avar	Avars
Turkic-Iranian of Dagestan	Altai	Turkic	Polovtsian-Kipchak	Caucasian	Kumyks
			Oguz	Essentially Oguz	Azerbaijanis living in Dagestan
	Indo-European	Aryan	New Persian	South-Western	Tat people living in Dagestan
Azerbaijan-Iranian	Altai	Turkic	Oguz	Essentially Oguz	Azerbaijanis living in Azerbaijan,
					Iran
					Karapapakhs
	Indo-European	Aryan	New Iranian	North-Western	Talysh
					Yazidis
					Kurds
				South-Western	Iranian-speaking peoples of Iran**
Steppe Turkic	Altai	Turkic	Kipchak	Kipchak-Nogai	Karanogai people, Astrakhan Nogais
					Karakalpaks
Separate	Altai	Turkic	Oguz	Essentially Oguz	Turkmens
Separate	Caucasian	Kartvelian	South-Kartvelian		Georgians

Dagestan suggesting the influence of the Eurasian steppe based on PC 1. The map makes it possible to see interaction of three genetic strata within the range of the Eastern Caucasus. The influence of Iran covering the entire Azerbaijan and wading into Dagestan along the edge of the Caspian Sea extends from the south. Dagestan retains genetic specificity of the autochthonous population of the Eastern Caucasus eastern outskirts. The most recent influence of the Eurasian steppe extends from the north. And these three genetic strata are loosely related to three linguistic divisions: Iranian-speaking, Caucasian-speaking, and Turkic-speaking peoples. An independent bioinformatics method was further applied to test the “three strata” hypothesis.

Contribution of the ADMIXTURE ancestral components to the gene pools of peoples living in the Eastern Caucasus

After considering positions of the Eastern Caucasus genomes in the variation principal component space, let's move to analysis by more informative method, i.e. to the ADMIXTURE ancestral component modeling. Recall that each model was calculated based on the same set of genomes as was used in the previous analysis (only Chechens were added, who became outliers in the PC analysis). The models differed only in the number of ancestral components (K) set for each model. Among all models with K-values between 2–20 three models (K = 3, K =10, K = 20) allowing one to trace the changes with increasing number of hypothetical ancestral populations were selected for thorough investigation (Fig. 3).

Model of three ancestral components

The contribution of each of three ancestral components (AC) to each genome is highlighted in specific color. The resulting ADMIXTURE image can serve as confirmation of the hypothesis of three genetic strata in the gene pool of the Eastern Caucasus: the orange color that is typical for the Karanogai people represents the hypothetical steppe genetic stratum; the green color that is pronounced in Kurds and Iranian-speaking Iranians represents the Iranian stratum; the yellow color that predominates in genomes of the peoples of Dagestan represents the contribution of the Caucasian-speaking population of the region. It we accept this interpretation of the color scheme, we will be able to estimate the contribution of each of three genetic strata to the genomes of peoples of the Eastern Caucasus and test the hypothesis of three strata in its gene pool. For that let's merge the genomes in accordance with the clusters (Table 2) and represent the AC contributions as a bar graph (Fig. 4; for quantitative data see fig. A in Appendix).

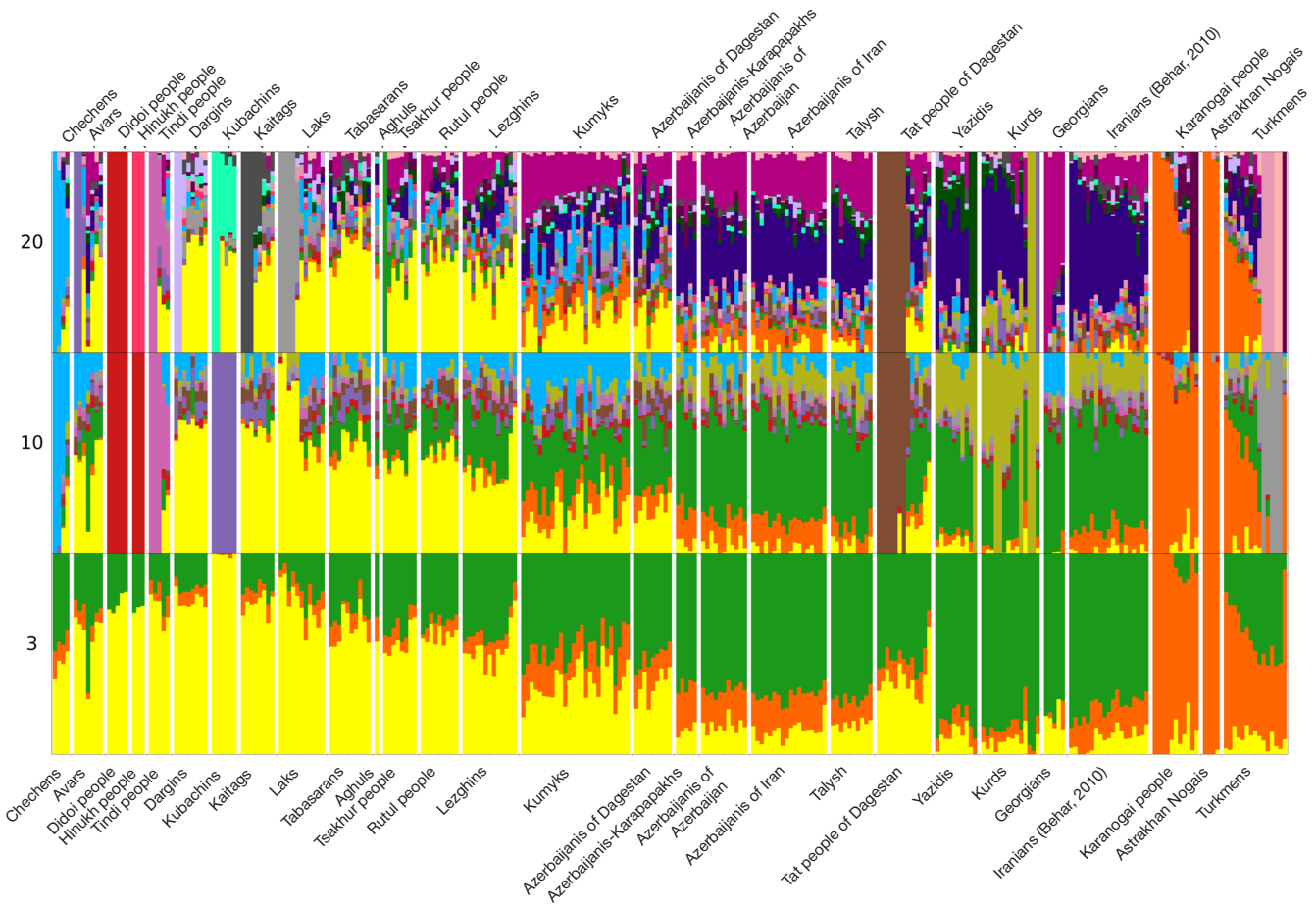

Fig. 3. Contributions of the ADMIXTURE ancestral components (%) to the genomes of peoples of the Eastern Caucasus, K = 3, K = 10, K = 20

The AC genomic profiles show very close agreement with the PC clusters based on the correlation of "three strata". The "Caucasian" contribution to the genomes of the Dargin-Lak cluster is 75%; the contribution to the Lezgin-Avar cluster is slightly more than a half of the gene pool; the contribution to the Turkic-Iranian cluster is about one third, and the contribution to the Azerbaijan-Iranian cluster is about 10%. As the "Caucasian" stratum descends, the Iranian stratum dramatically increases: 17%, 35%, 50%, 75%. The "Steppe" stratum turns out to be potent in the steppe cluster only (91%): it constitutes only 7–8% in the Caucasian-speaking peoples and 14–19% in the Turkic-speaking ones.

Model of 10 ancestral components

When modeling 10 ancestral components for the same set of genomes, preservation of three original ACs ("Steppe", "Iran", "Caucasus") is observed. However the correlation of strata changes due to the emergence of new ACs. Specific new components that make little contribution to the genomes of other peoples (Fig. 3) are merged into one AC, referred to as "Other" (Fig. 4). These include ACs of small ethnic groups (Kubachins, Didoi people, Hinukh people, Tindi people) that usually reflect the closely related genomes within ethnic groups, and specific AC variants found in some Tat people and Turkmens.

Such merging into models with K = 10 (Fig. 4; also see fig. B in Appendix) results in the emergence of only two new components reflecting the impact of peoples living in other regions of the Caucasus: the first one predominates among Georgians, and the second one prevails among Chechens. The "Caucasian" stratum is the major contributor to the "Nakhi" AC, while the "Iranian" stratum is the main contributor to the "Transcaucasia" AC. Since the contribution of the "Transcaucasia" AC to the genomes of the Iranian-speaking population of Iran (54%) is almost equal to its contribution to the genomes of Georgians (58%), it can be assumed that the "Transcaucasia" AC represents an ancient Southwest Asian stratum in the gene pools of the Caucasus and Transcaucasia.

Model of 20 ancestral components

When modeling 20 ancestral components for the same set of genomes, we see four new ACs that have diverged from the "Caucasus" AC in addition to two original ACs found when K = 3 ("Steppe", "Iran"). These new ACs reflect the genetic diversity of peoples of the Dargin and Lezgin groups, as well as of Avars and Laks (Fig. 4; also see fig. C in Appendix). Among them the "Lezgin" AC is the major contributor to the genomes of other peoples: it defines one third of gene pools of the Dargin-Lak and Avar cluster, a half of the Lezgin gene pool, and one fifth of the Kumyk gene pool.

It is extremely important that a common genetic stratum that is highlighted in yellow in Fig. 3 and 4 is preserved in the genomes of all peoples of Dagestan, along with the diverged inherent ACs. Previously we called this stratum "Caucasus". However, other ACs predominate in the genomes of other Caucasian peoples, that is why we have every right to propose a more accurate name for this genetic stratum: "Dagestan" AC.

DISCUSSION

Two independent methods for bioinformatics analysis of genomic variation in peoples of the Eastern Caucasus make it possible to reveal similar patterns and complement each other.

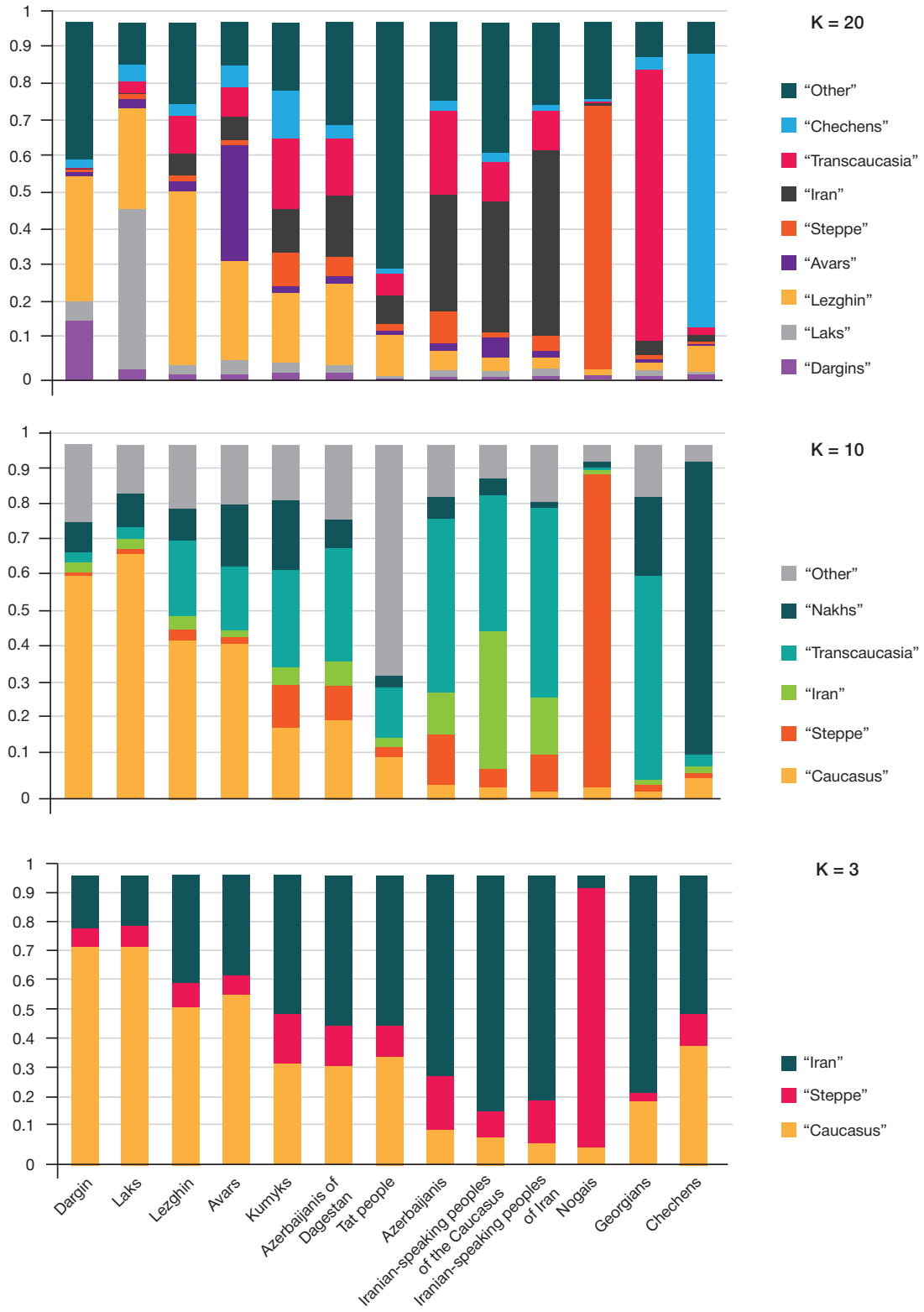

Fig. 4. Relative contributions of the ADMIXTURE ancestral components (%) to the genomes of peoples of the Eastern Caucasus, K = 3, K = 10, K = 20

Three genetic strata

Both methods identify three genetic strata: the first one is associated with multiple surges in the Iranian population, the second one with the recent influence of the Eurasian steppe, and the third one with the ancient population formed within the range of modern Dagestan. The strength of these genetic strata among different groups of population of the Eastern Caucasus varies.

In the Eastern Caucasus, the "Steppe" genetic stratum constitutes almost the entire gene pool only in the Karanogai people, and in the other genomes of the region the share varies between 7% in peoples of Dagestan and 19% in Azerbaijan. However, the findings show that this "Steppe" stratum forms the basis of the gene pools of many peoples living in the Caspian Sea region (Astrakhan Nogais, Karakalpaks, Turkmens), representing a strong genetic component in the vast area.

The "Iranian" genetic stratum of the Eastern Caucasus plays an important role, since it permeates all gene pools of Azerbaijan and "fades" as it travels through Dagestan from its south border (Lezgin peoples) and the Caspian Sea region (Kumyks, Tat people) to the center and north.

The "Dagestan" genetic stratum forms the basis of all Dagestan peoples. Despite the fact that original ancestral component is found in almost all the assessed Dagestan peoples when the K-values are high, all ethnic groups of Dagestan are united by the common ancient genetic stratum. This is an important conclusion of the study.

All three genetic strata are well placed within the Eastern Caucasus geographical space. The "Steppe" stratum is only limited by the small area of the Eurasian steppe in the north of the region. In contrast, the south "Iranian" stratum that flows from the Iranian plateau as a powerful stream constitutes three quarters of the gene pool of modern Azerbaijan and one third of the gene pools of peoples of Dagestan on average. The "Dagestan" genetic stratum is geographically located between other strata, it constitutes more than a half of the gene pool of Caucasian-speaking peoples (52–100%). The Caspian Sea region is a crossroads and a meeting place of all three strata: the "Iranian" stratum constitutes a half of the gene pool of Kumyks, Azerbaijanis living in Azerbaijan, and the Tat people, the "Dagestan" stratum makes up one third, and the "Steppe" stratum make up about 15%.

Genetics and linguistics

In contrast to geography, the correlation between genetics and linguistics is very weak. This is mainly due to the fact that peoples of the Eastern Caucasus (Kumyks and Azerbaijanis) started speaking Turkic languages, but their gene pool still remained mostly "pre-Turkic". However, we clearly see that the neighboring peoples also have an impact on the genomes of peoples that have retained their languages. Thus, the share of the "Dagestan" genetic stratum in the Tat people of Dagestan reaches 38%, while in other Iranian-speaking peoples of the Eastern Caucasus (Kurds, Yazidis, and Talysh) it reaches only 9%. The "Dagestan" stratum strength in the Caucasian-speaking peoples of Dagestan is two-thirds of the gene pool (66%) on average, it varies depending on the contacts with other peoples. And these are not the only ones showing a broken relationship between genetics and linguistics. For example, original "Georgian" ancestral component is found in Georgians only when $K = 19$, and all models with $K < 19$ show that three quarters of their gene pool come from the "Iranian" stratum.

Selection of "model" gene pools

One of the objectives of the study was to search for such groups of ethnicities that could credibly represent the Eastern Caucasus in the Eurasian space. Two independent methods suggest the need to include four such groups: 1) Dargins, Laks; 2) Avars, Lezghins, Tabasarans, Aghuls, Rutul people, Tsakhur people; 3) Kumyks, Tat people and Azerbaijanis living in Dagestan; 4) Azerbaijanis and Talysh living in Azerbaijan. The combination of those provides a correct representation of the Eastern Caucasus diversity. Since in this case the "steppe" contribution of the Karanogai genomes is small, it is more appropriate to consider genomes of the Karanogai people in the context of gene pool of the Eurasian steppe or its Caspian part.

Planning further research

The study covers the main range of genomes of the Eastern Caucasus and generally reflects the polyphony of its genomes, however, it should be considered only as a general contour of its diversity architectonics. The findings suggest that there is a need to perform a specific thorough assessment of autosomal gene pools of each of four groups of the indigenous population: Dagestan, Azerbaijan, Iranian-speaking population, steppe populations of six countries of the Caspian Sea region and the Ciscaucasia.

CONCLUSIONS

The hypothesis of three genetic strata, the interaction of which forms "unity in diversity" of gene pool of the Eastern Caucasus, is put forward. The "Iranian" stratum is formed by multiple surges of Iranian population within the range of the Eastern Caucasus; the "Steppe" stratum is related to the recent negligible influence of the Eurasian steppe; the "Dagestan" stratum represents the gene pool of ancient population of the Eastern Caucasus. Modeling of a large number of ancestral components reveals the original ancestral component in the majority of Dagestan peoples, however, all ethnic groups of Dagestan are united by a strong common ancient genetic stratum. The interaction of these three genetic strata results from geographic features of the region and is only indirectly related to linguistics. Four groups of indigenous population of the Eastern Caucasus have been identified, the combination of which should be included in the characteristics of its autosomal gene pool.

References

1. Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Molecular biology and evolution. 2012; 29 (1): 359–65.
2. Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, at al. The genetic history of admixture across inner Eurasia. Nature ecology & evolution. 2019; 3(6): 966–76.
3. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, at al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genetics. 2015; 11 (4): e1005068.
4. Behar DM, Metspalu M, Baran Y, Kopelman NM, Yunusbayev B, Gladstein A, et al. No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. Human biology. 2013; 85 (6): 859–900.
5. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513 (7518): 409–13.
6. Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: A catalog of genomic variations in the Iranian population. Human mutation. 2019; 40 (11): 1968–84.
7. Kars ME, Başak AN, Onat OE, Bilguvar K, Choi J, Itan Y, et al. The genetic structure of the Turkish population reveals high levels of variation and admixture. Proceedings of the National Academy of Sciences. 2021; 118 (36): e2026076118.
8. Balanovska EV, Agdzhoyan AT, Gorin IO, Petrushenko VS, Pylev VY, Kulemin NA, i dr. V poiskax alanskogo sleda: geneticheskaya istoriya Severnogo Kavkaza po polnogenomnym dannym ob

autosomnom genofonde. Vestnik Moskovskogo universiteta. Seriya 23. Antropologiya. 2022 (3): 4–62. Russian.
9. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, et al. The genome-wide structure of the Jewish people. Nature. 2010; 466 (7303): 238–42.
10. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 1–16.
11. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010; 26 (22): 2867–73.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38 (8): 904–9.
13. McKinney W. Data structures for statistical computing in python. InProceedings of the 9th Python in Science Conference. 2010; 28–445 (1): 51–56.
14. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007; 9 (03): 90–5.
15. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. Mwaskom/Seaborn: V0. 8.1. Zenodo. 2017.
16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009; 19 (9): 1655–64.

Литература

1. Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Molecular biology and evolution. 2012; 29 (1): 359–65.
2. Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, at al. The genetic history of admixture across inner Eurasia. Nature ecology & evolution. 2019; 3(6): 966–76.
3. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, at al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genetics. 2015; 11 (4): e1005068.
4. Behar DM, Metspalu M, Baran Y, Kopelman NM, Yunusbayev B, Gladstein A, et al. No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. Human biology. 2013; 85 (6): 859–900.
5. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513 (7518): 409–13.
6. Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: A catalog of genomic variations in the Iranian population. Human mutation. 2019; 40 (11): 1968–84.
7. Kars ME, Başak AN, Onat OE, Bilguvar K, Choi J, Itan Y, et al. The genetic structure of the Turkish population reveals high levels of variation and admixture. Proceedings of the National Academy of Sciences. 2021; 118 (36): e2026076118.
8. Балановская Е. В., Агджоян А. Т., Горин И. О., Петрушенко В. С., Пылёв В. Ю., Кулемин Н. А., и др. В поисках аланского следа: генетическая история Северного Кавказа по полногеномным данным об аутосомном генофонде. Вестник Московского университета. Серия 23. Антропология. 2022 (3): 48–62.
9. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, et al. The genome-wide structure of the Jewish people. Nature. 2010; 466 (7303): 238–42.
10. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 1–16.
11. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010; 26 (22): 2867–73.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38 (8): 904–9.
13. McKinney W. Data structures for statistical computing in python. InProceedings of the 9th Python in Science Conference. 2010; 28–445 (1): 51–56.
14. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007; 9 (03): 90–5.
15. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. Mwaskom/Seaborn: V0. 8.1. Zenodo. 2017.
16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009; 19 (9): 1655–64.