

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

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

<sup>1</sup> Медико-генетический научный центр, Москва, Россия

<sup>2</sup> Биобанк Северной Евразии, Москва, Россия

<sup>3</sup> 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 РФФИ №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 Опубликовано онлайн: 29.05.2023

DOI: 10.24075/vrgmu.2023.017

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

Balanovska EV<sup>1,2</sup>, Gorin IO<sup>1</sup>, Petruschenko VS<sup>1</sup>, Ponomarev GYu<sup>1</sup>, Belov RO<sup>1</sup>, Pocheshkhova EA<sup>1,3</sup>, Salaev VA<sup>1</sup>, Iskandarov NA<sup>1</sup>, Pylev VYu<sup>1,2</sup> ✉

<sup>1</sup> Bochkov Research Centre of Medical Genetics, Moscow, Russia

<sup>2</sup> Biobank of North Eurasia, Moscow, Russia

<sup>3</sup> 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, Petruschenko 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 Committee 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

Генофонд народов Кавказа давно привлекает внимание популяционных генетиков. Этот небольшой по площади регион, расположенный между Европой и Азией, населяет более 60 народов, говорящих на языках трех лингвистических семей — кавказской, индоевропейской и алтайской. Из всех регионов Кавказа наименее генетически изучен Восточный Кавказ, где сосредоточен самый обширный спектр (более 30) народов, представляющих три лингвистические семьи. Поскольку западный Прикаспий в течение тысячелетий служил мостом между Европой и Азией, для понимания генофонда Восточного Кавказа надо привлекать данные и о степях на севере региона, и популяциях Иранского горного плато. Причудливая структура народов Восточного Кавказа и сплетение множества миграционных потоков крайне осложняют анализ его генофонда. Если по генетике популяций Дагестана имеется ряд публикаций, то население Азербайджана представляет одно из самых больших белых пятен на генетической карте. А ведь именно оно является ключевым для выявления многовекового влияния Персии на генофонды Восточного Кавказа: распространялось оно на весь Каспийский регион или концентрировалось только на южных приграничных генофондах? Генетическая история ряда малочисленных на Северном Кавказе народов тоже может быть реконструирована только при планомерном генетическом изучении Восточного Кавказа в контексте данных о народах Ирана.

Сложная структура полиэтничного региона Восточного Кавказа требует детального анализа. Поэтому в данной работе внимание сосредоточено только на его аутосомном генофонде. Параллельно будет опубликована статья по изменчивости Y-хромосомы в этих же популяциях Восточного Кавказа. Такое «бинокулярное зрение» позволит увидеть наиболее объемную и объективную картину изменчивости генофонда Восточного Кавказа.

По широкогеномным панелям — одной из наиболее востребованных за последние 10 лет и эффективных систем ДНК-маркеров — опубликовано крайне мало данных о популяциях Дагестана и Азербайджана. Почти во всех этих статьях данные по Восточному Кавказу не анализировали отдельно, они были лишь составной частью изучения много более обширных регионов — всего Кавказа или Евразии.

Популяции Дагестана изучены по панели Illumina (~600 тыс. SNP) в работе [1] о генофонде Кавказа в целом и по панели Human Origin (~600 тыс. SNP) [2].

Генофонд азербайджанцев Северо-Западного Ирана рассмотрен в работе о следах тюркской экспансии [3], где для них показан 5%-й вклад восточноевразийской предковой компоненты, что можно объяснить распространением тюркских языков на запад согласно модели «доминирования элиты» (смена языка без значительного изменения генофонда). Азербайджанцы Азербайджана по широкогеномной панели Human Origin представлены только в статье [2] при сравнении с другими генофондами.

В большинстве публикаций рассмотрен широкий круг проблем: формирование генофонда Кавказа на основе вклада ближне- и средневожочных миграций [1], наследие миграции тюркоязычных групп [3] или эко-географического районирования Северной Евразии [2]. При этом особенности генофонда Восточного Кавказа прицельно не анализируются. Отдельные немногочисленные выборки народов Восточного Кавказа представлены и в двух других крупных исследованиях,

посвященных совсем иным вопросам [4, 5]. К сожалению, часть из упомянутых выборок проанализирована только по панели HumanOrigin (Affimatrix), плохо сопоставимой с панелью Illumina.

Из исследований последних лет стоит отметить полногеномные и полноэкзомные исследования генофондов Ирана и Турции. Авторы одной из работ [6] изучают геномную изменчивость народов Ирана. Поскольку исследование нацелено на составление базы данных геномного разнообразия Ирана, то в нем уделено много внимания генетической структуре самого Ирана, а соседние регионы (в частности Азербайджан) затронуты лишь поверхностно. Похожая ситуация и в работе [7], где анализируется генетическая структура популяций Турции, однако влияние Восточного Кавказа описывается только кратким упоминанием об их смешении с некоторыми популяциями Турции.

В целом по народам Восточного Кавказа в опубликованных данных приведены генотипы только 43 образцов, изученных в основном по узким панелям Illumina, тогда как в рамках данной работы проанализировано 243 образца по широкой панели Illumina.

Таким образом, в мировой литературе по широкогеномным панелям не только отсутствуют выводы по генофонду Восточного Кавказа, но и сами данные крайне немногочисленны и охватывают лишь отдельные группы — тогда как в данной работе представлены популяции 22 народов Восточного Кавказа.

В работе, посвященной поиску аланского следа в аутосомных генофондах Северного Кавказа [8], основное внимание уделено этносам Центрального и Западного Кавказа. Генофонд народов Восточного Кавказа был представлен четырьмя народами Дагестана. Полученный результат указал на необходимость как детального рассмотрения этих данных о генофонде Восточного Кавказа, так и прицельного анализа аутосомных генофондов всей совокупности народов этого региона и выявления основных закономерностей в его изменчивости, что стало целью предполагаемого исследования.

## МАТЕРИАЛЫ И МЕТОДЫ

В данной работе в понятие «народы Восточного Кавказа» включено все коренное население Дагестана (17 этносов), Азербайджана (азербайджанцы, азербайджанцы-карапапахы, талыши), а также другие ираноязычные народы (курды и езиды), изученные на Кавказе (они включают мигрантов из разных популяций Кавказа и Ирана; в данной работе рассмотрены как представители ираноязычного населения Восточного Кавказа, который входит в их этнический ареал). Эти 22 этнические группы Восточного Кавказа представлены (табл. 1) собственными данными коллектива о 243 геномах. По единой панели SNP-маркеров они проанализированы в контексте собственных данных о соседних народах Центрального Кавказа (чеченцы) и Закавказья (грузины), Прикаспия (астраханские ногайцы) и Закаспия (каракалпаки, туркмены) и литературных данных о народах Ирана [3, 9].

Общая анализируемая выборка составила 356 геномов 29 этносов: собственные данные (318 геномов 27 этносов) изучены по панелям Illumina4M и Illumina750K; литературные данные (38 геномов 2 этносов) — по панелям Illumina750K и HumanOrigin. Проведена фильтрация с помощью PLINK 1.9 [10]: по качеству прочтения геномов; родственности индивидов (менее 3-й степени по версии

программы KING 2.3.0 [11]); сцепленности ДНК-маркеров и их мономорфизму.

Анализ главных компонент изменчивости геномов (PCA) проведен с помощью утилиты smartpca из пакета EIGENSTRAT [12]. Конвертация формата plink (bed-bim-fam) в формат eigensoft (eigenstratgeno-snp-ind) осуществлена с помощью программы convertf из того же программного пакета (с параметрами по умолчанию). Расчет проведен для пяти главных компонент с пятью итерациями исключения outliers, результаты визуализированы на языке программирования Python 3 с использованием библиотек pandas [13], matplotlib [14] и seaborn [15]. Для каждой популяции на графике главных компонент определен (и обозначен более крупной точкой) ее центроид как средняя по всем рассчитанным компонентам и по всем образцам, входящим в эту популяцию.

Анализ предковых компонент методом ADMIXTURE проводили с помощью программного обеспечения ADMIXTURE v1.3.0 [16], значение количества моделируемых предковых компонент  $K$  варьировалось от 2 до 20. Для

каждого значения  $K$  была рассчитана кросс-валидация для его вероятностной оценки. Визуализацию результатов работы ADMIXTURE осуществляли на Python 3 с помощью библиотек pandas, matplotlib, seaborn.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

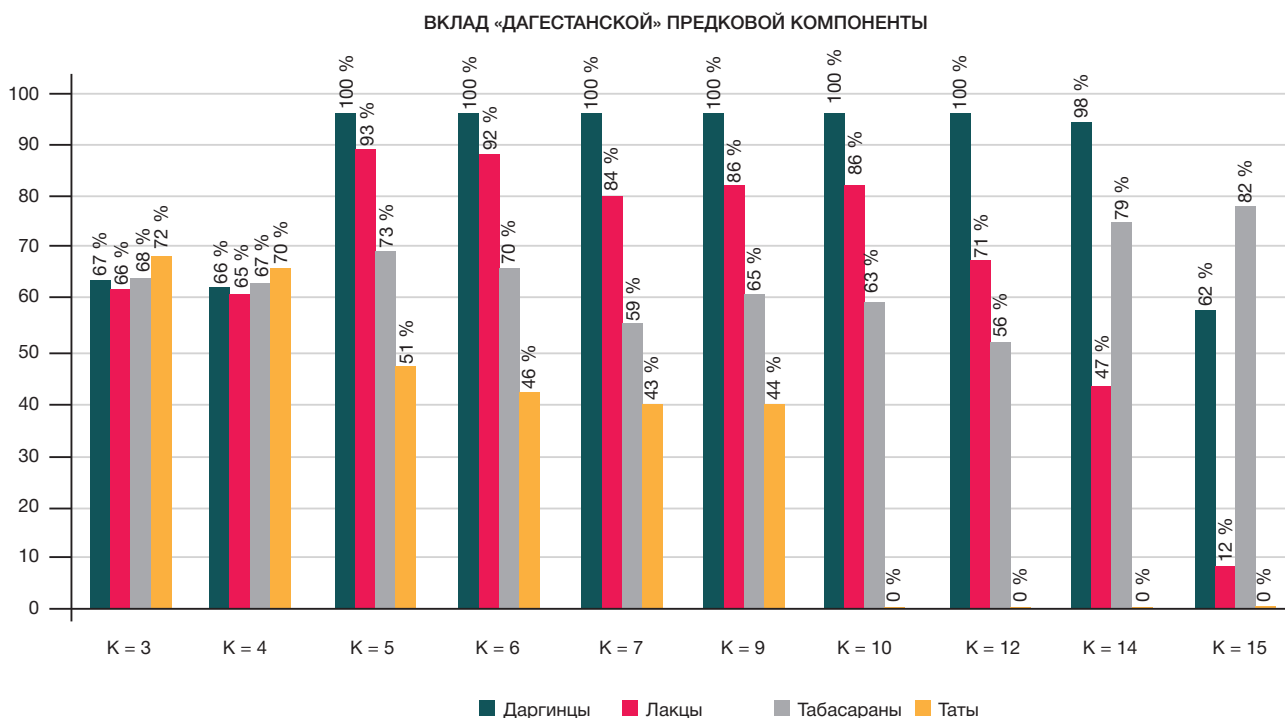
### Предварительная оценка вклада предковых компонент ADMIXTURE в генофонды четырех народов Дагестана

В более ранней публикации [8] аутосомный генофонд Восточного Кавказа был представлен геномами Дагестана — кавказоязычных (даргинцы, лакцы, табарасаны) и ираноязычных (таты) народов. В модели предковых компонент (ПК) ADMIXTURE при  $K = 11$  вклад дагестанской ПК в эти геномы составил 77%. Однако остается вопрос: насколько правомерно такое объединение? Геномы каких многочисленных народов Восточного Кавказа могут корректно представлять его при изучении крупных регионов Евразии?

Таблица 1. Лингвистическая принадлежность и численность изученных геномов

ЛИНГВИСТИЧЕСКАЯ КЛАССИФИКАЦИЯ				Народы	N геномов	
Семья	Ветвь	Субветвь	Группа			
Кавказская	Нахско-дагестанская	Лезгино-дарго-лакская	Даргинская	кубачинцы	6	
				даргинцы	8	
				кайтагцы	8	
			Лакская	лакцы	11	
				Лезгинская	табасараны	11
					агулы	1
		рутульцы	9			
		цахуры	8			
		Аваро-андо-цезская	лезгины	43		
			Аварская	аварцы	7	
	Андийская		тиндинцы	6		
	Нахская	Цезская (дидойская)	дидойцы	5		
гинухцы			5			
Картвельская	Южно-картвельская		чеченцы	15		
Алтайская	Тюркская	Половецко-кыпчакская	Кавказская	грузины	19	
				кумыки	27	
		Огузская	Собственно огузская	азербайджанцы:		
				Дагестана,	9	
				Азербайджана,	13	
				Ирана*	18	
				карапахахи	6	
		Кыпчакская	Кыпчакско-ногайская	туркмены	19	
				караногайцы,	11	
				ногайцы астраханские,	5	
		каракалпаки	17			
Индо-европейская	Арийская	Ново-иранская	Юго-западная	таты Дагестана	13	
				ираноязычные народы Ирана**	20	
		Северо-западная	талыши	10		
			езиды	10		
			курды	16		

Примечание: литературные данные по народам Ирана [3, 9].



**Рис. 1.** Динамика вклада «дагестанской» предковой компоненты ADMIXTURE в геномы каждого из даргинцев, лакцев, табасаран и татов в моделях от  $K = 3$  до  $K = 15$

Динамика вклада «дагестанской» компоненты по мере увеличения числа ПК (рис. 1) выявляет различия между этносами уже при малом числе ПК ( $K = 5$ ). С  $K = 10$  таты Дагестана формируют свою собственную ПК, отделяясь от кавказоязычных народов. При этом (рис. 1) вклад генофонда Дагестана в основном делают даргинцы: в их геномы вклад «дагестанской» ПК в интервале  $4 < K < 15$  составляет 98–100% и лишь при  $K = 15$  падает до 62%. Иную динамику можно наблюдать для лакцев: при увеличении значения  $K$  от 5 до 15 вклад в их геномы «дагестанской» ПК падает от 93 до 29% по мере проявления собственной ПК лакцев. У табасаран своеобразный паттерн: при увеличении числа ПК от  $K = 5$  до  $K = 12$  вклад в их геномы «дагестанской» ПК падает от 73 до 56%, но затем возрастает до 82% при  $K = 15$ .

Это сравнение позволило обнаружить неожиданный феномен: при разном числе ПК ( $3 < K < 15$ ) «дагестанская» отражает генофонды разных этносов Дагестана. Только при малом числе ПК ( $K = 3$  и  $K = 4$ ) их вклад одинаков, но при  $K > 4$  постепенно исчезает вклад татов и лакцев, а вклад даргинцев и табасаран меняется по величине. Такой результат говорит о необходимости создать генетические портреты аутосомных генофондов для каждого из народов Восточного Кавказа, а затем сформировать группы этносов, которые могут служить опорными при анализе аутосомных генофондов Кавказа и других крупных регионов Евразии.

#### Положение 22 народов Восточного Кавказа в пространстве главных компонент

Для корректного ответа на поставленные вопросы важно, во-первых, охватить исследованием максимально возможный спектр народов Восточного Кавказа, а во-вторых — использовать основные независимые методы анализа популяционной генетики.

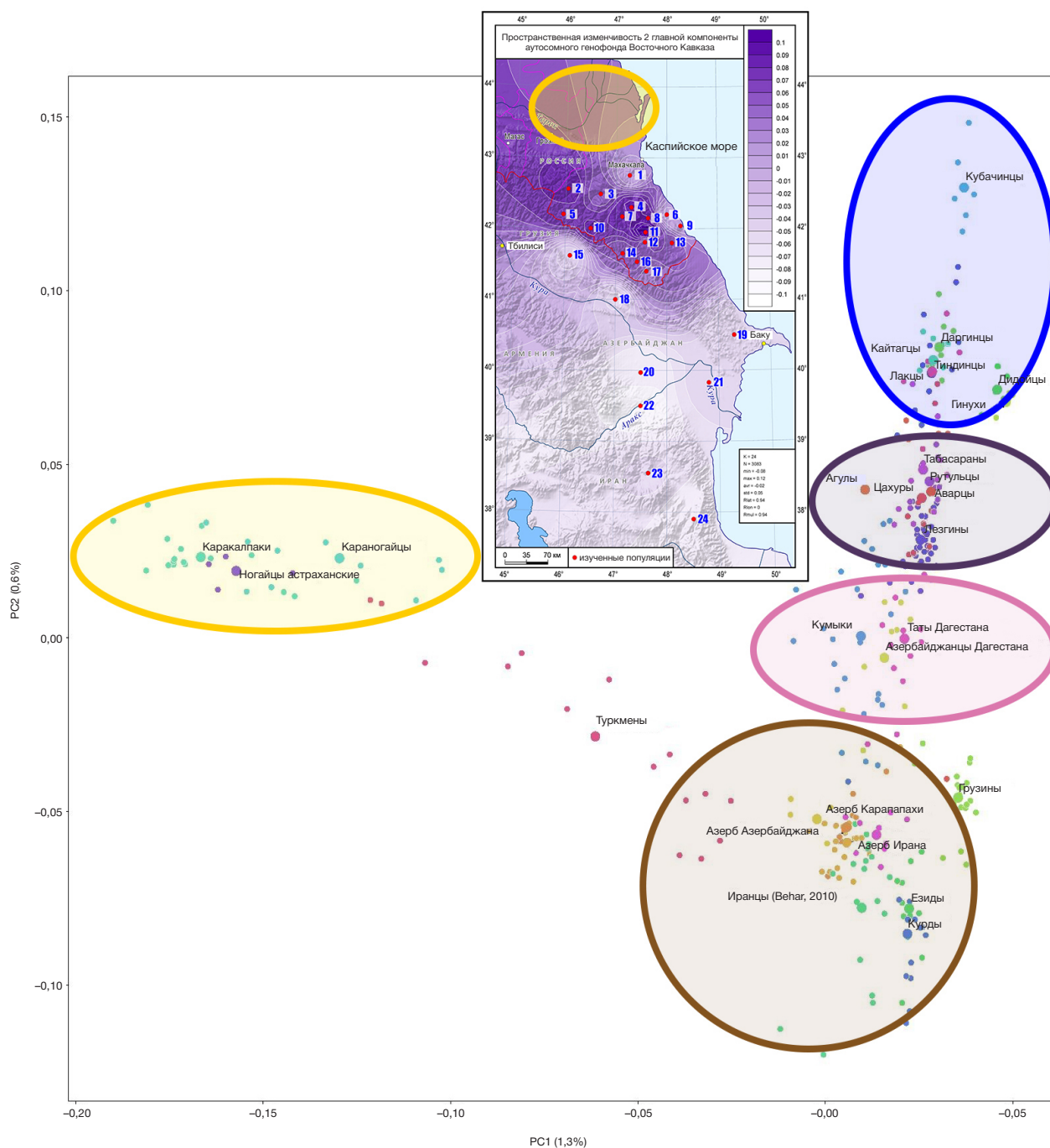
На рис. 2 приведено положение 22 народов Восточного Кавказа и шести групп сравнения в генетическом

пространстве 1 и 2 главных компонент (ГК) изменчивости генофонда. Четко выявляются шесть кластеров геномов, которые показывают, что сходство аутосомных геномов народов Восточного Кавказа очень слабо соответствует классификации их языков. Почти все кластеры включают народы, говорящие не только на разных ветвях языков, но даже на языках разных лингвистических семей (табл. 2). Первая ГК четко делит все народы на степные и все остальные. Зато вторая ГК демонстрирует длинную цепь — от кубачинцев до курдов — всех остальных геномов (рис. 2).

**Дарго-лакско-андо-цезский кластер** включил представителей пяти этносов четырех групп разных субветвей нахско-дагестанских языков (табл. 2): даргинцев, кайтагцев, кубачинцев, лакцев, тиндинцев, дидойцев, гинухцев. Отметим, что по другим ГК от всего массива резко отличаются геномы изолированных популяций: дидойцы, гинухцы по третьей 3 ГК, одноаульный этнос кубачинцев по 4 ГК, тиндинцы по 5 ГК.

**Лезгино-аварский кластер** включил представителей шести этносов из обеих субветвей нахско-дагестанских языков (табл. 2). Из аварской группы в кластер вошли лишь аварцы. Но хотя их центроид и расположен среди геномов лезгинских народов, отдельные геномы аварцев крайне разнообразны (рис. 2; красные точки): они тянутся вдоль всего лезгино-аварского кластера и уходят за его пределы к андо-цезам. Для лезгинских народов мы не видим различий между геномами восточно-лезгинской (лезгины, табасараны, агулы) и рутульско-цахурской (рутульцы, цахуры) подгрупп. При этом геномы самих лезгин (фиолетовые точки на рис. 2) тянутся в другой кластер — к татам и азербайджанцам Дагестана.

**Тюрко-иранский кластер Дагестана** объединил геномы ираноязычных татов и тюркоязычных кумыков и азербайджанцев Дагестана (табл. 2). Сходство геномов татов и азербайджанцев Дагестана не удивляет, поскольку татов в Дагестане было принято записывать как «азербайджанцев», и разделение проведено по местам их компактного проживания в Дагестане. Геномы всех трех



**Рис. 2.** График 1–2 главных компонент изменчивости (PCA) геномов коренного населения Восточного Кавказа в контексте окружающих популяций. На врезке приведена карта 2-й главной компоненты. 1 — кумыки; 2 — тиндинцы; 3 — аварцы; 4 — даргинцы; 5 — гинухцы; 6 — азербайджанцы Дагестана; 7 — лакцы; 8 — кайтагцы; 9 — таты Дагестана; 10 — дидойцы; 11 — кубачинцы; 12 — агулы; 13 — табасараны; 14 — цахуры; 15 — грузины; 16 — рутульцы; 17 — лезгины; 18 — азербайджанцы-карапахахи; 19 — азербайджанцы Азербайджана; 20 — езиды; 21 — талыши; 22 — курды; 23 — азербайджанцы Ирана [3]; 24 — ираноязычные иранцы [9]

этносов весьма многолики и тяготеют к азербайджано-иранскому кластеру. Но разнообразие геномов кумыков (рис 2; синие точки) зашкаливает — они тоже входят и лезгино-аварский кластер.

**Азербайджано-иранский кластер Дагестана** вновь объединил геномы тюрков (азербайджанцев Азербайджана и Ирана) с ираноязычными народами (талыши, курды, езиды и суммарная группа Ирана). По 1 ГК только одна группа азербайджанцев — карапахахи — обнаруживает небольшой сдвиг к народам Евразийской степи. Зато геномы туркмен (рис. 2; розовые точки) обозначили целый мост между азербайджано-иранским и степным кластерами. Другой «полюс» азербайджано-иранского

кластера фиксируют грузинские геномы, которые в других вариантах ГК сближаются с народами Северного Кавказа.

**Степной кластер** объединил караногайцев Дагестана с другими степными народами Прикаспийской степи (астраханские ногайцы) и Закаспия (каракалпаки). Резкое отличие геномов степного кластера формирует различия по 1 ГК, отражающей наибольшую изменчивость анализируемых геномов.

#### Пространственная изменчивость главных компонент

На карте 2 ГК (рис. 2; врезка) на севере Дагестана мы разместили в ареале караногайцев желтый овал,

Таблица 2. Кластеры в генетическом пространстве 1–2 главных компонент (PCA) и этнолингвистическая принадлежность геномов каждого кластера

Кластер	ЛИНГВИСТИЧЕСКАЯ КЛАССИФИКАЦИЯ				Народы
	Семья	Ветвь	Субветвь	Группа	
Дарго-лакско-андо-цезский	Кавказская	Нахско-дагестанская	Лезгино-дарго-лакская	Даргинская	кубачинцы, даргинцы
				Лакская	кайтагцы
			Аваро-андо-цезская	Андийская	лакцы
				Цезская (дидойская)	тиндинцы
					дидойцы
Лезгино-аварский	Кавказская	Нахско-дагестанская	Лезгино-дарго-лакская	Лезгинская	табасараны
					агулы
					рутульцы
					цахуры
			Аваро-андо-цезская	Аварская	лезгины
Тюрко-иранский Дагестана	Алтайская	Тюркская	Половецко-кыпчакская	Кавказская	кумыки
			Огузская		Собственно огузская
	Индо-европейская	Арийская	Ново-иранская	Юго-западная	таты Дагестана
Азербайджано-иранский	Алтайская	Тюркская	Огузская	Собственно огузская	азербайджанцы Азербайджана
					азербайджанцы Ирана
					карапахахи
	Индо-европейская	Арийская	Ново-иранская	Северо-западная	талыши
					езиды
Юго-западная	ираноязычные народы Ирана**				
Степной тюркский	Алтайская	Тюркская	Кыпчакская	Кыпчакско-ногайская	караногайцы, ногайцы астраханские
					каракалпаки
Отдельно	Алтайская	Тюркская	Огузская	Собственно огузская	туркмены
Отдельно	Кавказская	Картвельская	Южно-картвельская		грузины

указывающий на влияние евразийской степи по 1 ГК. Карта позволяет увидеть взаимодействие трех генетических пластов в ареале Восточного Кавказа. С юга распространяется влияние Ирана, захватывающее весь Азербайджан и пробирающееся в Дагестан по кромке Каспийского моря. Дагестан сохраняет генетическое своеобразие автохтонного населения восточной окраины Северного Кавказа. С севера распространяется наиболее позднее влияние евразийской степи. И эти три генетических пласта слабо связаны с тремя крупными лингвистическими подразделениями — ираноязычных, кавказоязычных и тюркоязычных народов. Для проверки гипотезы «трех пластов» далее был использован независимый метод биоинформатики.

#### Вклад предковых компонент ADMIXTURE в генофонды народов Восточного Кавказа

Рассмотрев положение геномов Восточного Кавказа в пространстве главных компонент изменчивости, перейдем к анализу более информативного метода — к моделированию предковых компонент методом ADMIXTURE. Напомним, что каждая модель рассчитана по совокупности всех тех же геномов, что и в предыдущем анализе (дополнены лишь чеченцами, которые в анализе

ГК ушли в аутлайеры), и каждая модель отличается лишь тем, какое число предковых компонент (K) мы задаем для данной модели. Из всех моделей с K от 2 до 20 для детального рассмотрения отобраны три (K = 3, K = 10, K = 20), позволяющие проследить динамику по мере увеличения числа гипотетических предковых популяций (рис. 3).

#### Модель трех предковых компонент

Вклад в каждый геном каждой из трех предковых компонент (ПК) окрашен своим цветом. И полученная картина ADMIXTURE может служить подтверждением гипотезы «трех пластов» генофонда Восточного Кавказа: рыжий цвет, характерный для караногайцев, отражает гипотетический степной генетический пласт; зеленый цвет, ярко выраженный у курдов и ираноязычных иранцев, — иранский пласт; желтый цвет, преобладающий в геномах народов Дагестана, — вклад кавказоязычного населения региона. Если мы примем такую интерпретацию цветовой гаммы, то можем оценить вклад каждого из трех генетических пластов в геномы народов Восточного Кавказа и проверить гипотезу трех пластов его генофонда. Для этого объединим геномы в соответствии с кластерами (табл. 2) и представим вклады ПК на столбчатой

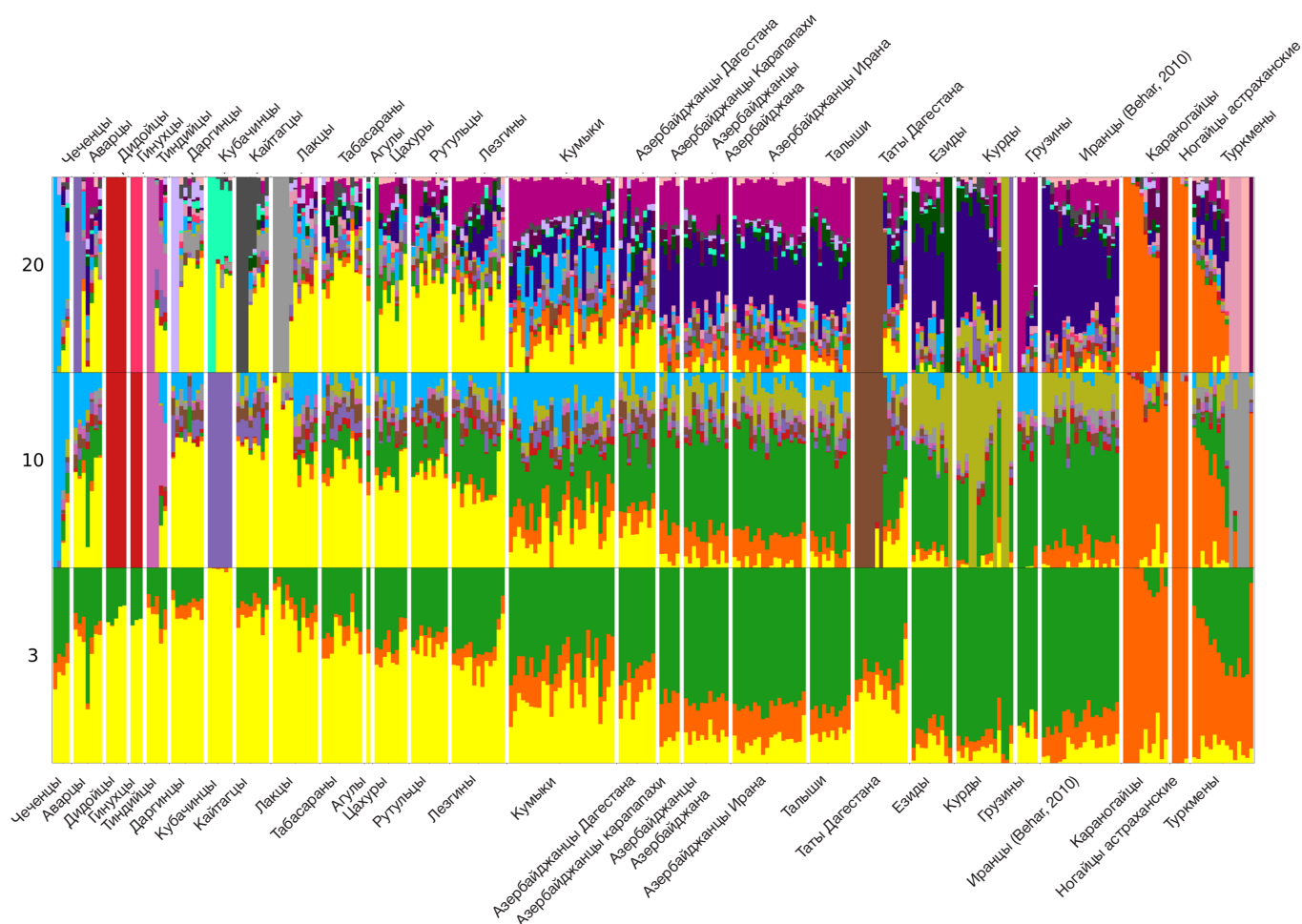


Рис. 3. Вклад предковых компонент ADMIXTURE (в %) в геномы народов Восточного Кавказа при числе предковых компонент  $K = 3$ ,  $K = 10$ ,  $K = 20$

диаграмме (рис. 4; количественные оценки см. на рис. А в Приложении).

По соотношению «трех пластов» геномные профили ПК оказались в очень хорошем соответствии с кластерами ГК. «Кавказский» вклад в геномах дарго-лакского кластера составляет 75%; в лезгино-аварском — чуть более половины генофонда; в тюрко-иранском — около трети, в азербайджанско-иранском — около 10%. По мере убывания «кавказского» пласта столь же резко возрастает «иранский»: 17%, 35%, 50%, 75%. «Степной» пласт оказывается мощным только в степном кластере (91%): у кавказоязычных народов он составляет всего 7–8%, а у тюркоязычных — 14–19%.

#### Модель 10 предковых компонент

При моделировании 10 предковых компонент для той же совокупности геномов мы видим сохранение трех исходных ПК («Степь», «Иран», «Кавказ»). Однако их соотношение несколько меняется за счет появления новых ПК. Те новые компоненты, которые вносят небольшой вклад в геномы других народов (рис. 3), объединены в одну ПК «Прочие» (рис. 4). К ним отнесены ПК малочисленных этносов (кубачинцы, дидойцы, гинухцы, тиндинцы), обычно отражающие близкородственные геномы в их пределах, а также особые варианты ПК у части геномов татов и туркмен.

При таком объединении в модели при  $K = 10$  (рис. 4; см. также рис. Б в Приложении) появляются только две новые компоненты, отражающие влияние народов других

регионов Кавказа: одна преобладает у грузин, другая — у чеченцев. ПК «Нахи» забирает на себя в основном вклад от «кавказского» пласта, а ПК «Закавказье» — от «иранского» пласта. Поскольку вклад ПК «Закавказье» в геномы ираноязычного населения Ирана (54%) практически равен ее вкладу в геномы самих грузин (58%), можно предполагать, что ПК «Закавказье» отражает древний переднеазиатский пласт в генофондах Кавказа и Закавказья.

#### Модель 20 предковых компонент

При моделировании 20 предковых компонент для той же совокупности геномов мы видим — кроме обнаруженных еще при  $K = 3$  двух исходных ПК («Степь», «Иран») — четыре новых ПК, которые выделились из ПК «Кавказ» и отражают генетическое своеобразие народов даргинской и лезгинской групп, аварцев и лакцев (рис. 4; см. также рис. В в Приложении). Из них наибольший вклад в геномы других народов Дагестана вносит «Лезгинская» ПК — она описывает треть генофондов дарго-лакского и аварского кластеров, половину лезгинского генофонда и пятую часть генофонда кумыков.

Но крайне важно, что наряду с выделением собственных ПК в геномах всех народов Дагестана сохраняется общий генетический пласт, отраженный желтым тоном на рис. 3 и 4. Раньше мы называли этот пласт «Кавказ». Но поскольку в геномах других кавказских народов преобладают иные ПК, то мы имеем полное право дать этому генетическому пласту более точное название — ПК «Дагестан».

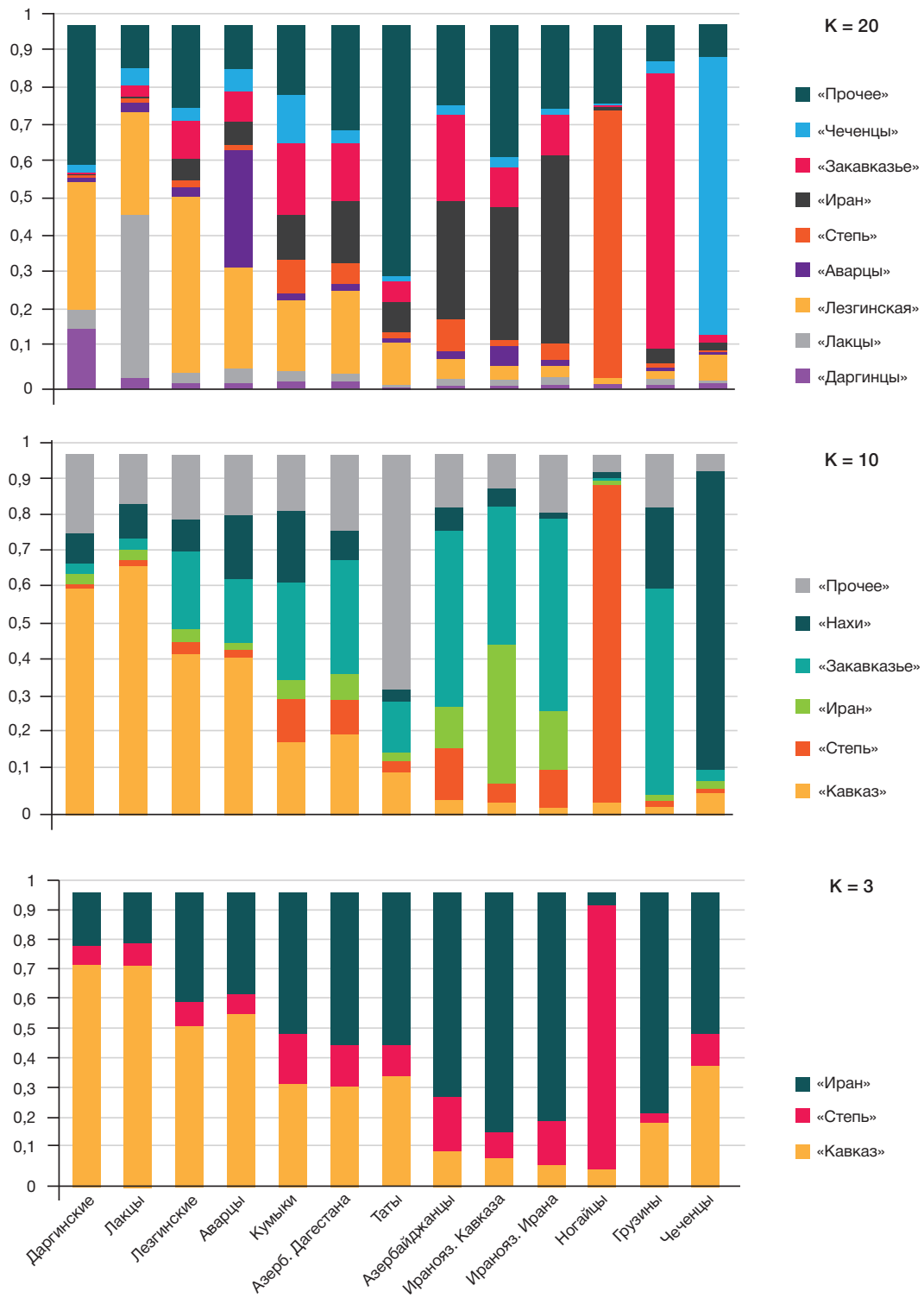


Рис. 4. Относительный вклад предковых компонент ADMIXTURE (в %) в геномы народов Восточного Кавказа при  $K = 3$ ,  $K = 10$ ,  $K = 20$

#### ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Два независимых метода биоинформатического анализа изменчивости геномов народов Восточного Кавказа позволяют выявить схожие закономерности и дополняют друг друга.

#### Три генетических пласта

Оба метода выявляют наличие трех генетических пластов, один из которых связан с многочисленными волнами

иранского населения, второй — с поздним влиянием евразийской степи, третий — с древним населением, сформировавшимся в ареале современного Дагестана. Мощности этих генетических пластов различна в разных группах населения Восточного Кавказа.

«Степной» генетический пласт на Восточном Кавказе составляет почти весь генофонд лишь у караногайцев, а в остальных геномах региона варьирует в пределах от 7% у народов Дагестана до 19% в Азербайджане. Однако результаты указывают, что этот «степной» пласт составляет основу генофондов многих народов Каспийского региона



(астраханских ногайцев, каракалпаков, туркмен), отражая мощную генетическую составляющую на огромном пространстве.

«Иранский» генетический пласт на Восточном Кавказе играет очень важную роль, пронизывая все генофонды Азербайджана и «тая» по мере движения по Дагестану от его южных пределов (лезгинские народы) и Прикаспия (кумыки, таты) к центру и северу.

«Дагестанский» генетический пласт составляет основу всех народов Дагестана. Хотя практически для всех изученных народов Дагестана при высоких значениях  $K$  и выявляется собственная предковая компонента, все этносы Дагестана объединяет мощный общий древний генетический пласт. Это важный вывод из данной работы.

Все три генетических пласта хорошо ложатся на географическое пространство Восточного Кавказа. «Степной» ограничен только небольшим ареалом евразийской степи на севере региона. Зато южный «иранский» пласт мощным потоком затекает с Иранского плато, составляя три четверти генофонда современного Азербайджана и в среднем треть генофонда у народов Дагестана (кроме караногайцев). «Дагестанский» генетический пласт географически расположен между другими пластами, составляя больше половины генофонда кавказоязычных народов (52–100%). А Прикаспий является перепутьем и местом встречи всех трех пластов: у кумыков, азербайджанцев Дагестана и татов половину генофонда составляет «иранский» пласт, треть — «дагестанский» и около 15% «степной».

#### Генетика и лингвистика

В отличие от географии, связь между генетикой и лингвистикой очень слаба. В основном это обусловлено переходом на тюркские языки народов Восточного Кавказа (кумыков и азербайджанцев), во многом сохранивших при этом «дотюркский» генофонд. Но мы видим явное влияние окружающих народов и на геномы народов, сохранивших свой язык. Так, «дагестанский» генетический пласт у татов Дагестана достигает 38%, тогда как у других ираноязычных народов Восточного Кавказа (курдов, езидов и тальшей) — всего 9%. Мощность «дагестанского» пласта составляет у кавказоязычных народов Дагестана в среднем две трети генофонда (66%), варьируя в зависимости от контактов с другими народами. И народы Восточного Кавказа не одиноки в нарушении связи между генетикой и лингвистикой. Например, у грузин собственная «грузинская» предковая компонента появляется только при  $K = 19$ , а на всех моделях при  $K < 19$  их генофонд в среднем на три четверти формируется «иранским» пластом.

#### 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, et 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, et 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,

#### Выбор «модельных» генофондов

Одной из целей работы был поиск таких групп этносов, которые могут надежно представлять Восточный Кавказ в евразийском масштабе. Два независимых метода дружно указывают на необходимость включения четырех таких групп: 1) даргинцы, лакцы; 2) аварцы, лезгины, табасараны, агулы, рутульцы, цахуры; 3) кумыки, таты и азербайджанцы Дагестана; 4) азербайджанцы и талыши Азербайджана. Их комплекс создает корректное представление многообразия Восточного Кавказа. Поскольку «степной» вклад геномов караногайцев в данном случае невелик, их геномы правильнее рассматривать в контексте генофонда Евразийской степи или же ее Каспийского сегмента.

#### Планирование дальнейших исследований

Данная работа охватывает основной спектр геномов Восточного Кавказа и в целом отражает полифонию его геномов, но может быть рассмотрена лишь как общий абрис архитектоники его многообразия. Результаты указывают, что необходимо специальное и детальное исследование аутосомных генофондов каждой из четырех групп коренного населения: Дагестана, Азербайджана, ираноязычного населения и населения степей шести стран Каспийского региона и Предкавказья.

#### Выводы

Выдвинута гипотеза трех генетических пластов, взаимодействие которых формируют «единство в многообразии» генофонда Восточного Кавказа. «Иранский» пласт сформирован многочисленными волнами иранского населения в ареале Восточного Кавказа; «степной» пласт связан с поздним и незначительным влиянием евразийской степи, «дагестанский» — отражает генофонд древнего населения Северного Кавказа. При моделировании большого числа предковых компонент у большинства народов Дагестана выявляется собственная предковая компонента, но все этносы Дагестана объединяет мощный общий древний генетический пласт. Взаимодействие этих трех генетических пластов обусловлено географическими особенностями региона и лишь косвенно связано с лингвистикой. Выявлены четыре группы коренного населения Восточного Кавказа, комплекс которых необходимо включать в характеристику его аутосомного генофонда.

5. 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.
6. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanov K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. 2014; 513 (7518): 409–13.
7. 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.
8. 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. In *Proceedings 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, Gempertline 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.

## 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, et 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, et 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, Mitnik 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, Agdzhojan AT, Gorin IO, Petrusenko 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. In *Proceedings 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, Gempertline 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.