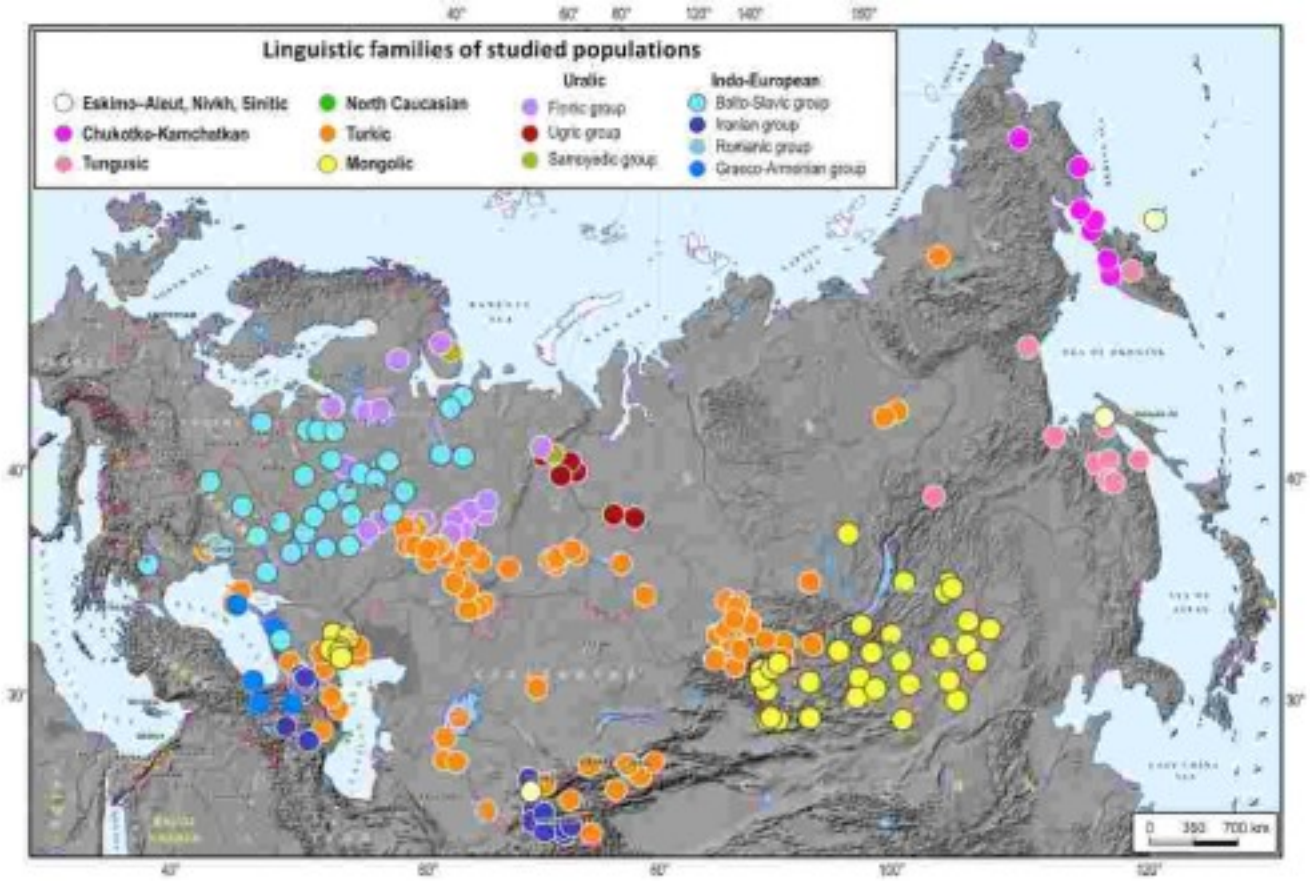
Создана система предсказания этногеографического происхождения индивида по ДНК для Северной Евразии

Российские исследователи разработали систему, позволяющую по ДНК определить этногеографическое происхождение индивида. Система основана на анализе около 5000 SNP, ассоциированных с происхождением в популяциях Северной Евразии. Она была создана в результате исследования 1883 образцов ДНК из 266 популяций, которые были объединены в 29 этногеографических групп (ЭГГ). После отбора 5229 SNP, отличающих каждую ЭЭГ от других, была создана модель для предсказания происхождения индивида из какой-либо ЭЭГ. В среднем точность предикции для 29 ЭЭГ составила 71%, при этом абсолютную точность модель продемонстрировала для Южной и Центральной Сибири, Дальнего Востока и Камчатки. Предложенный метод может быть использован с целью определения происхождения индивида для популяций России и сопредельных стран, области его применения – криминалистическая наука и генетическая генеалогия.

Современные генетические методы позволяют с определенной вероятностью предсказать этногеографическое происхождение человека по его ДНК. Описаны сотни тысяч полиморфных участков — SNP маркеров, специфичных для больших субконтинентальных регионов, таких как Восточная или Южная Азия, Океания, Северная Африка, Ближний Восток, Европа. В криминалистической практике применяются десятки и сотни таких маркеров. Северная Евразия, в которую входят страны постсоветского пространства плюс Монголия, занимает треть самого крупного континента и отличается культурно-языковым (200 коренных народов, принадлежащих к 10 языковым семьям) и генетическим разнообразием. Вместе с тем, она до сих пор остается недостаточно представленной в базе генетических маркеров, информативных для предсказания географического происхождения индивида. Имеющиеся в настоящее время панели SNP маркеров, созданные для Западной Европы и для Центральной и Восточной Азии, лишь частично перекрываются с генофондом Северной Евразии.

Большой шаг в этом направлении сделан в статье, опубликованной в журнале Frontiers in Genetics коллективом под руководством О.П.Балановского и Е.В.Балановской, представляющим Институт общей генетики РАН и Медико-генетический научный центр, первый автор Игорь Горин; в исследовании также участвовали специалисты Московского физико-технического института, МГУ им. М.В.Ломоносова, Федеральный исследовательский и клинический центр физико-химической медицины и Национальный центр биотехнологии в Нурсултане, Казахстан.

В работе было исследовано 266 популяций из 12 стран Северной Евразии (Армения, Азербайджан, Грузия, Казахстан, Киргизстан, Литва, Молдова, Монголия, Россия, Турция, Украина и Узбекистан), представляющих 92 этнические группы. Все образцы для исследований были предоставлены Биобанком Северной Евразии. Они были генотипированы на панели Illumina Infinium Omni5Exome-4 v1.3 BeadChip, включающей 4,5 млн SNP маркеров.

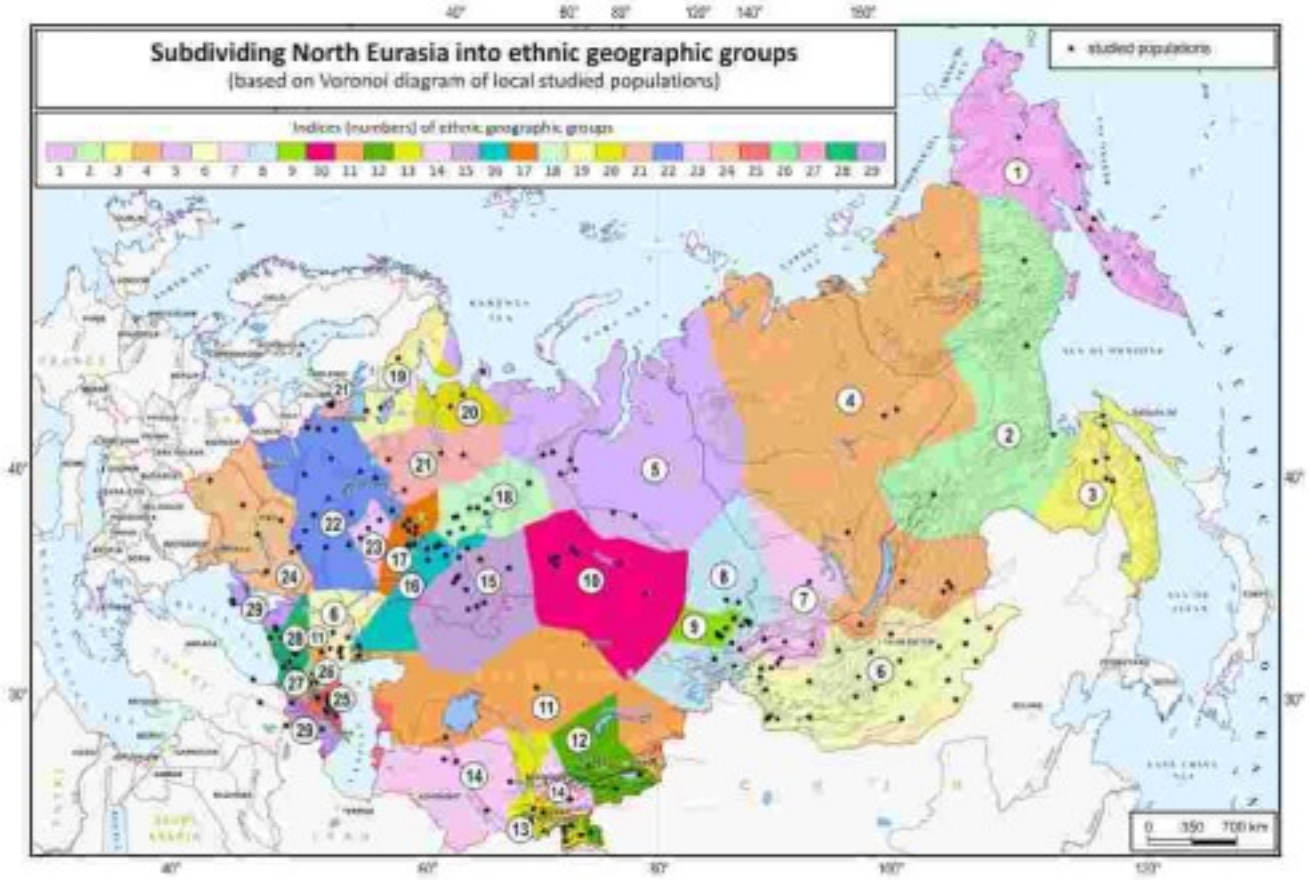


Карта 266 популяций Северной Евразии, использованных для анализа. Точками разных цветов обозначены лингвистические группы, к которым относятся исследованные популяции (Gorin et al., 2022).

Работа по созданию платформы для идентификации этногеографического происхождения включала пять этапов. Первый этап состоял в том, чтобы разделить 266 популяций на кластеры, генетически различающиеся между собой, но внутренне относительно гомогенные; в результате из популяций были сформированы 29 этногеографических групп (ЭГГ). Каждая ЭГГ включала не менее 25 образцов. Их список приведен в таблице.

№	ЭГГ	Популяции	Число образцов
1	Амурские нанайцы&Нивхи&Орочи&Ульчи	Нанайцы, нивхи, ульчи, орочи	55
2	Башкиры	Башкиры	44
3	Буряты&Хамнегане&Якуты	Буряты, хамнегане, якуты	59
4	Чеченцы&Ингуши	Чеченцы, ингуши	39
5	Чукчи&Коряки&Ительмены	Коряки, чукчи, ительмены, камчадалы,	75
6	Дагестанцы	Аварцы, кубачи, даргинцы, табасараны, лакцы, лезгины, рутулы	74
7	Эвенки&Эвены	Эвены, эвенки	49
8	Карелы&Вепсы	Карелы, вепсы	38
9	Казахи&Каракалпаки&Уйгуры&Ногайцы	Каракалпаки, астраханские ногайцы, ставропольские ногайцы, уйгуры, казахи	33
10	Хакасы&Южные алтайцы	Хакасы, алтайцы	46
11	Ханты&Манси&Ненцы	Ханты, ненцы, манси	53
12	Коми&Удмурты	Коми-пермяки, коми-зыряне,удмурты, бесермяне	84
13	Киргизы	Киргизы	43
14	Мари&Чуваши	Чуваши, мари	53

15	Монголы&Калмыки	Монголы, калмыки	127
16	Мордва	Мордва мокша, мордва эрзя,мордва шокша	41
17	Осетины	Осетины	36
18	Северные русские	Русские, ижора, води	81
19	Южные русские	Русские, белорусы	240
20	Русский Север	Русские	35
21	Шорцы&Северные алтайцы	Шорцы, алтайцы	37
22	Сибирские татары	Сибирские татары	68
23	Таджики&Памирцы&Ягнобцы	Памирцы, таджики, ягнобцы	72
24	Татары	Татары кряшены, казанские татары, татары мишари, Татары из Башкирии, астраханские татары	60
25	Закавказье&Крым	Армяне, азербайджанцы, крымские татары, караимы, турки, курды, езиды, грузины	113
26	Тувинцы&Тофалары	Тувинцы, монголы, тофалары	64
27	Украинцы	Украинцы	79
28	Узбеки&Туркмены	Туркмены, узбеки	55
29	Западный Кавказ	Адыгейцы, кабардинцы, шапсуги, карачаевцы, абхазы, черкесы,абазины, балкарцы	87



Этногеографические группы (ЭГГ) на карте Северной Евразии, обозначены зонами разных цветов. Черные точки указывают на географическое расположение популяций (Gorin et al., 2022).

При составлении базы данных исследователи столкнулись с отсутствием некоторых генотипов и использовали прием импутации – введения вместо отсутствующих генотипов, наиболее частых в данной популяции. Второй этап состоял с том, чтобы отобрать те SNP, которые будут специфичны для каждой из 29 этногеографических групп. Начальный список в 817 120

тысяч кандидатных SNP после фильтрации и биоинформатического анализа сначала сократился до 50 – 60 тысяч SNP, а в ходе дальнейшего отбора авторы получили финальный список из 5 229 SNP, информативных для определения происхождения. Иными словами, этот набор SNP позволял отличить каждую ЭГГ от других.

Чтобы проверить, адекватно ли отражают выбранные SNP популяционную структуру, авторы провели анализ главных компонент, основанный на 5 229 SNP из финального списка. Другой график PCA построили для тех же популяций, по данным о 4,5 млн SNPs из панели Illumina. Эти графики демонстрировали сходный паттерн распределения популяций.

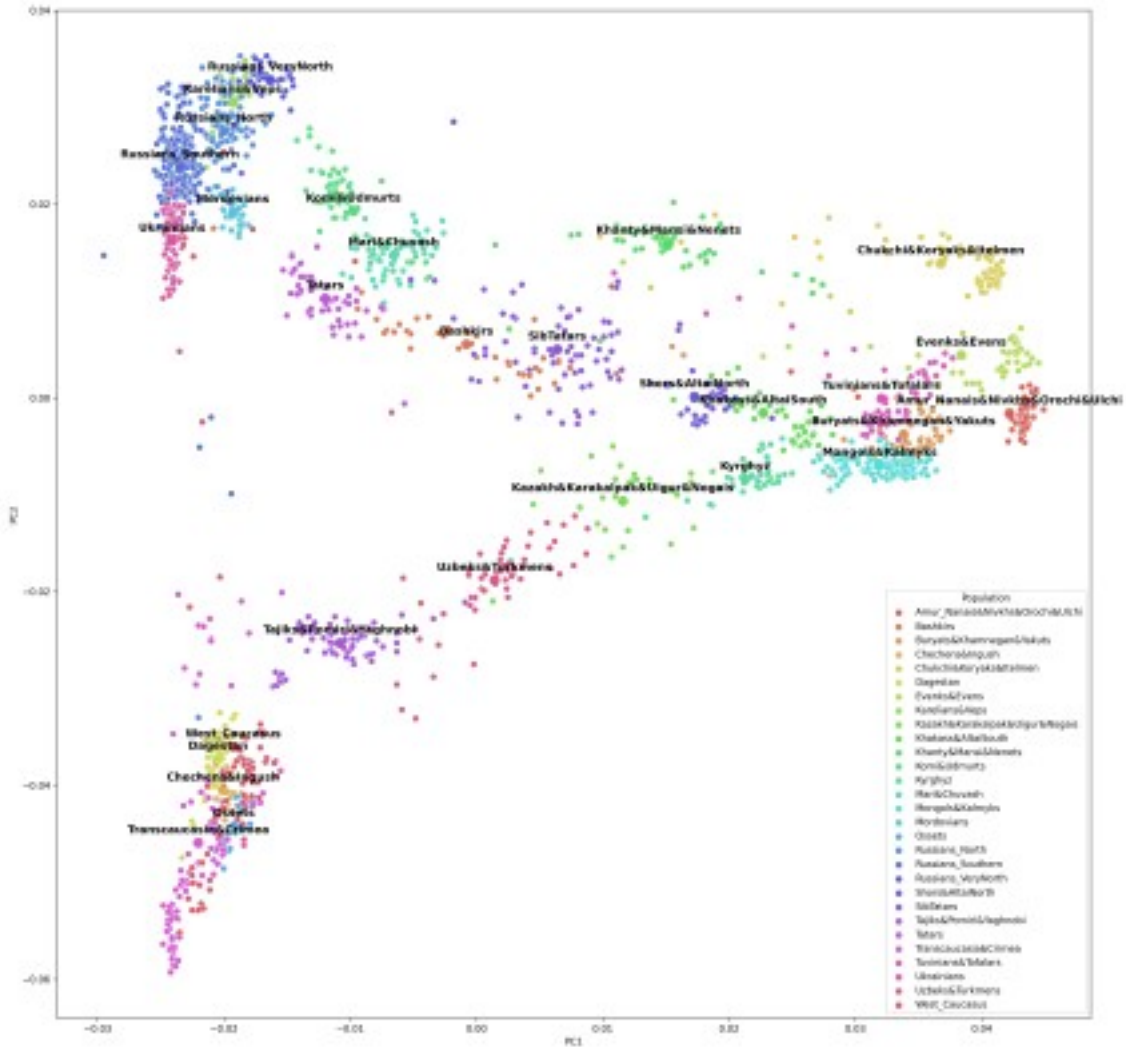


График анализа главных компонент, основанный на полной 4.5 M SNP панели (Gorin et al., 2022).

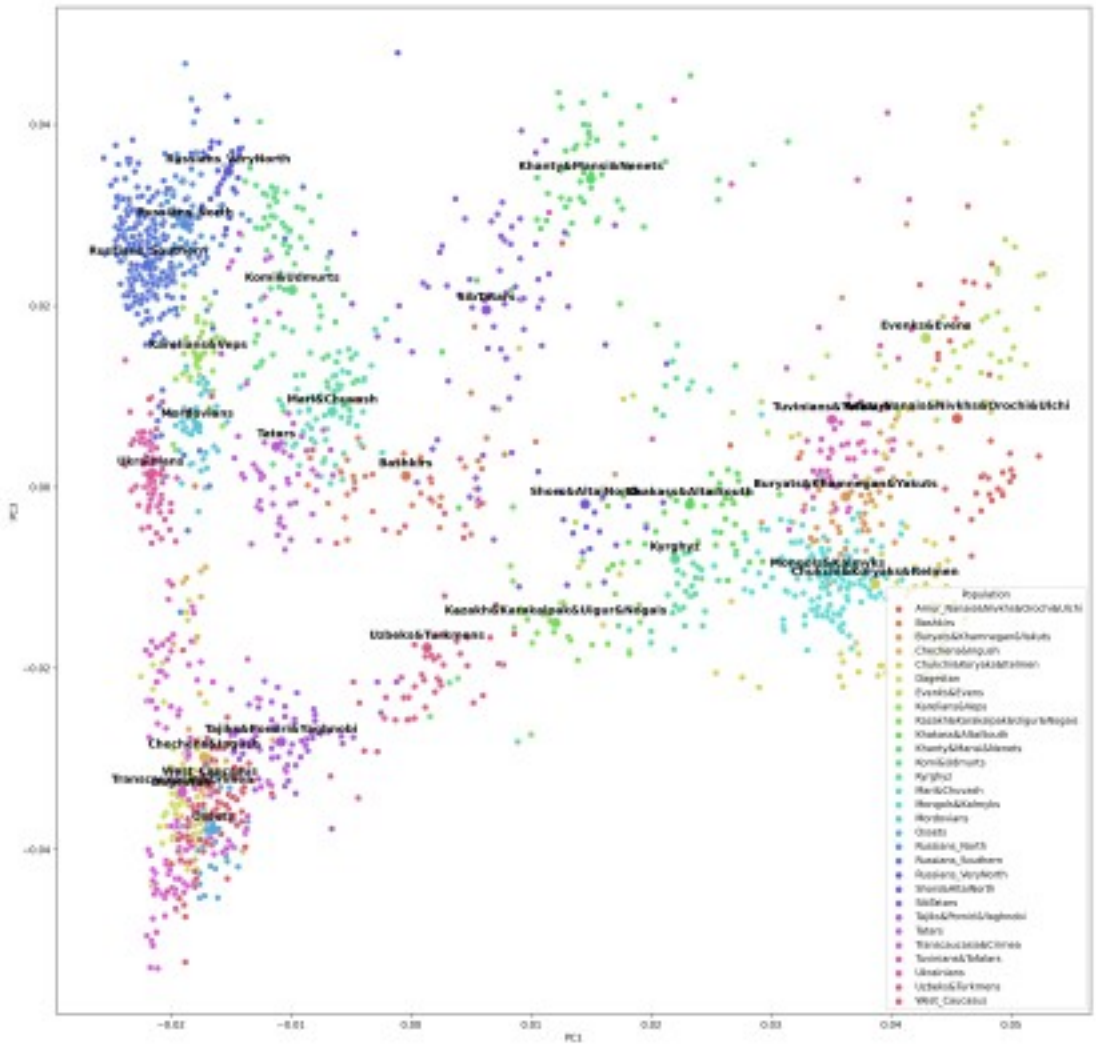


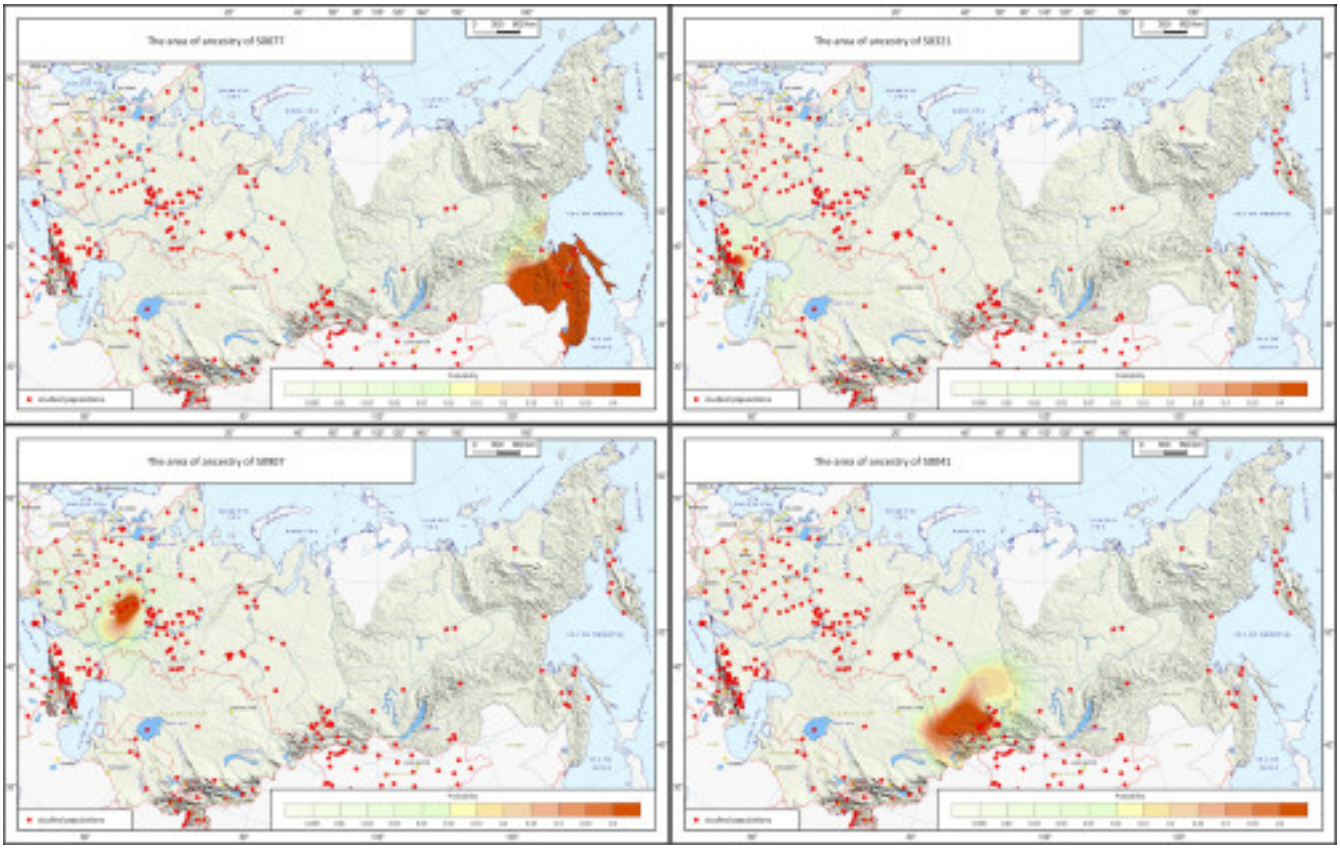
График анализа главных компонент, основанный на отобранных 5000 SNP (Gorin et al., 2022).

После того, как был завершен отбор SNP маркеров, авторам предстояло выбрать математическую модель для предсказания этногеографического происхождения. Из множества проверенных моделей они выбрали модель логистической регрессии. Эту модель они применили для машинного обучения на 1 773 образцах с учетом 5 229 отобранных SNP. 1241 образец использовали для обучения модели и оставшиеся 532 образца – для тестирования. Задачей модели было правильно отнести индивидуальный образец к той или иной этногеографической группе. Модель точно предсказала принадлежность индивидов к следующим ЭГГ: Мари&Чуваши, Украинцы, Ханты&Манси&Ненцы, Чукчи&Коряки&Ительмены, Эвенки&Эвены, Амурские нанайцы&Нивхи&Орочи&Ульчи, Тувинцы&Тофалары, Шорцы&Северные алтайцы. Ошибки были допущены для следующих ЭГГ: Русский Север, Северные русские, Карелы&Вепсы, Южные русские, Коми&Удмурты, Мордва, Буряты&Хамнегане&Якуты, Монголы&Калмыки, Хакасы&Южные алтайцы, Таджики&Памирцы&Ягнобцы, Осетины, Западный Кавказ&Крым. Чаще всего ошибки, снижающие точность модели, возникали в случаях генетической и географической близости ЭГГ. Тем не менее, ошибочно предсказанные ЭГГ практически всегда являлись соседними с истинной ЭГГ. Подобные ошибки могли быть вызваны большой долей генотипов, унаследованных от общей предковой популяции и распространенных по широкому региону.

Снижение чувствительности модели наблюдалось в тех случаях, когда образцы были отнесены к неверной ЭГГ внутри одного

географического региона происхождения. Такие ошибки наиболее часто возникали для популяций Урала, Западной Сибири, Центральной Азии и Кавказа. Как было ранее показано, эти территории генетически очень разнообразны. Для улучшения качества предсказания в этих регионах необходимо более дробное деление на ЭГГ, что, в свою очередь, требует большие размеры выборок. Вместе с тем наибольшее число абсолютно точных предикций ЭГГ было достигнуто для Южной и Центральной Сибири, Дальнего Востока и Камчатки.

С помощью оригинального программного обеспечения «Прародина» результаты картографировали. Представленные карты показывают вероятность происхождения четырех индивидов из того или иного региона Северной Евразии. Вероятность соответствует яркости красного цвета, как указано на шкале.



Примеры карт, построенные с помощью программы «Прародина» (Gorin et al., 2022).

Доля правильных предикций (совпадений между реальной ЭГГ и предсказанной ЭГГ) в среднем составляла 71%. На карте доля правильных предикций (совпадений между реальным и предсказанным регионом) составляла 61% для наиболее вероятных регионов происхождения (вероятность более 0,4) и 81% для объединения наиболее вероятных и менее вероятных регионов происхождения (вероятность более 0,2). Объединение ЭГГ в большие кластеры или расширение географического ареала происхождения увеличивает точность модели (долю правильных предикций), но снижает информативную ценность метода (географическую точность). Отсюда возникает необходимость поиска правильного баланса между точностью и информационной ценностью. Абсолютная точность предикции была достигнута для большинства ЭГГ Сибири, Дальнего Востока и Камчатки. Хорошая точность была достигнута для популяций Восточной Европы, Урала, Западной Сибири, Кавказа и Центральной Азии, небольшие отклонения можно, вероятно, объяснить высокой генетической гетерогенностью этих регионов. Повысить точность, как предполагают авторы, можно будет при увеличении размера выборки.

Авторы полагают, что разработанный ими метод в его текущем состоянии может быть применим для предсказания происхождения индивида для популяций России и сопредельных стран. В этом качестве он может использоваться в криминалистике и в генетической генеалогии. Его ограничением они считают два фактора: первый — высокая стоимость генотипирования, второй – то что метод пока не был испытан на индивидах смешанного происхождения.

текст Надежды Маркиной

Источник:

Gorin I.O., Balanovsky O.P., Kozlov O.V., Koshel S.M., Kostryukova E.S., Agdzhoyan A.T., Pylev V.Y., Balanovska E.V. Determining the area of ancestral origin for individuals from North Eurasia based on 5,000 SNP markers // Frontiers in Genetics. 16 May 2022. DOI: 10.3389/fgene.2022.902309