Применение популяционного биобанка для анализа частот клинически значимых ДНК-маркеров у населения России: биоинформатические аспекты

Горин И. О.[1], Петрушенко В. С.[1], Записецкая Ю. С.[1], Кошель С. М.[2,3], Балановский О. П.[1,2,4]

[1]ФГБУН Институт общей генетики им. Н. И. Вавилова, Москва; [2]ФГБНУ Медико-генетический научный центр им. академика Н. П. Бочкова, Москва; [3]ФГБОУ ВО Московский государственный университет им. М. В. Ломоносова, Москва; [4]АНО Биобанк Северной Евразии. Москва, Россия

Одной из задач популяционных биобанков является определение частот клинически значимых генетических полиморфизмов у населения. Население России характеризуется исключительно высокой гетерогенностью как в этническом, так и в генетическом плане, поэтому частоты генетических маркеров востребованы не в одной выборке, а в серии выборок, отражающих основное разнообразие генофонда различных народов и регионов.

Цель. Разделение народонаселения России и сопредельных стран на группы популяций, удовлетворяющие определенным условиям, а также имеющие репрезентативную выборку в существующих данных и биобанках.

Материал и методы. Разработан метод объединения популяций в более крупные группы с сохранением гомогенности внутри этих групп на основе главных компонент с кластеризацией K-средних, с последующей доработкой кластеризации для большей гомогенности и более равномерного распределения размеров групп с применением F_{ST} расстояний. Технология отлажена на примере Биобанка Северной Евразии, поэтому материалом для исследования послужили массивы широкогеномных данных по 4,5 млн генетических маркеров для 1883 образцов, представляющих 247 популяций России и сопредельных стран из выборок данного биобанка. Разработанный подход, результирующий набор популяций и их карта могут применяться при использовании других коллекций биоматериалов из российских популяций.

Результаты. Применение этого подхода позволило разделить все население России и сопредельных стран на 29 этногеографических групп (ЭГГ), характеризующихся относительной генетической гомогенностью. Этот набор популяций рекомендуется как базовый для популяционных скринингов на выявление частоты любых генетических маркеров среди населения России. Построена карта, демонстрирующая деление народонаселения на 29 территорий — ареалов ЭГГ.

Заключение. На основе надежного массива полногеномных данных проведено районирование генофонда населения России: выделены ЭГГ, обладающие контрастными частотами аллелей при сравнении друг с другом, но при этом относительно гомогенные внутри себя. Результирующая карта и реестр групп могут применяться в популяционно-генетических, медико-генетических и фармакогенетических исследованиях.

Ключевые слова: ДНК-маркеры, биоинформатика, популяционный биобанк, население, генофонд, районирование.

Отношения и деятельность. Исследование выполнено в рамках Государственного задания Министерства науки и высшего образования РФ для Медико-генетического научного центра им. академика Н. П. Бочкова и для Института общей генетики РАН им. Н. И. Вавилова.

Благодарности: благодарим всех доноров образцов. Коллекция образцов ДНК предоставлена АНО “Биобанк Северной Евразии”.

Поступила 05/11-2020
Получена рецензия 14/11-2020
Принята к публикации 10/12-2020

Для цитирования: Горин И. О., Петрушенко В. С., Записецкая Ю. С., Кошель С. М., Балановский О. П. Применение популяционного биобанка для анализа частот клинически значимых ДНК-маркеров у населения России: биоинформатические аспекты. *Кардиоваскулярная терапия и профилактика.* 2020;19(6):2732. doi:10.15829/1728-8800-2020-2732

Population-based biobank for analyzing the frequencies of clinically relevant DNA markers in the Russian population: bioinformatic aspects

Gorin I. O.[1], Petrushenko V. S.[1], Zapisetskaya Yu. S.[1], Koshel S. M.[2,3], Balanovsky O. P.[1,2,4]

[1]Vavilov Institute of General Genetics, Moscow; [2]N. P. Bochkov Research Center of Medical Genetics, Moscow; [3]Lomonosov Moscow State University, Moscow; [4]Biobank of Northern Eurasia. Moscow, Russia

One of the tasks of population-based biobanks is to determine the frequencies of clinically relevant genetic polymorphisms in the population. The population of Russia is very heterogeneous both ethnically and genetically. Therefore, the frequencies of genetic markers are in demand not in one sample, but in a series of samples reflecting the heterogeneity of the gene pool of different peoples and regions.

*Автор, ответственный за переписку (Corresponding author):

e-mail: balanovsky@inbox.ru

Тел.: +7 (985) 232-72-23

[Горин И. О. — м.н.с. лаборатории геномной географии, ORCID: 0000-0001-9532-8954, Петрушенко В. С. — м.н.с. лаборатории геномной географии, ORCID: 0000-0002-5763-5280, Записецкая Ю. С. — м.н.с. лаборатории геномной географии, ORCID: 0000-0002-8696-6302, Кошель С. М. — в.н.с. лаборатории популяционной генетики человека, к.г.н., в.н.с. кафедры картографии и геоинформатики, ORCID: 0000-0002-4540-2922, Балановский О. П.* — д.б.н., профессор РАН, в.н.с., зав. лабораторией геномной географии, г.н.с. лаборатории популяционной генетики человека, учредитель, ORCID: 0000-0003-4218-6889].

Aim. To divide the population of Russia and neighboring countries into population groups meeting certain conditions, as well as having a representative sample in existing data and biobanks.

Material and methods. We developed a method for combining populations into larger groups with maintaining intragroup homogeneity based on the principal components analysis with K-means clustering, followed by refinement of clustering for higher homogeneity and a more equal distribution of group sizes using FST distances. The technology has been adjusted using the example of the Biobank of Northern Eurasia. Therefore, the material was the genome-wide data on 4.5 million genetic markers for 1,883 samples representing 247 populations of Russia and neighboring countries from this biobank. The developed approach, the resulting set of populations and related map can be applied for other collections of biomaterials from Russian populations.

Results. Application of this approach made it possible to divide the entire population of Russia and neighboring countries into 29 ethnogeographic groups, characterized by relative genetic homogeneity. This set of populations is recommended as a baseline for population screenings to identify the frequency of any genetic markers among the population of Russia. A map has been constructed showing the division of population into 29 ethnogeographic areas.

Conclusion. On the basis of a reliable genome-wide data, the zoning of gene pool of the Russian population was carried out. We identified ethnogeographic groups with intergroup contrasting allele frequencies, but at the same time with relatively homogeneous intragroup parameters. The resulting map and register of groups can be used in population genetic, medical genetic and pharmacogenetic studies.

Key words: DNA markers, bioinformatics, population-based biobank, population, gene pool, zoning.

Relationships and Activities. The study was carried out within the State assignment of the Ministry of Science and Higher Education of the Russian Federation for the N. P. Bochkov Research Center of Medical Genetics and Vavilov Institute of General Genetics.

Acknowledgments: the authors are grateful to all sample donors. The collection of DNA samples is provided by the Biobank of Northern Eurasia.

Gorin I. O. ORCID: 0000-0001-9532-8954, Petrushenko V. S. ORCID: 0000-0002-5763-5280, Zapisetskaya Yu. S. ORCID: 0000-0002-8696-6302, Koshel S. M. ORCID: 0000-0002-4540-2922, Balanovsky O. P.* ORCID: 0000-0003-4218-6889.

*Corresponding author: balanovsky@inbox.ru

Received: 05/11-2020
Revision Received: 14/11-2020
Accepted: 10/12-2020

For citation: Gorin I. O., Petrushenko V. S., Zapisetskaya Yu. S., Koshel S. M., Balanovsky O. P. Population-based biobank for analyzing the frequencies of clinically relevant DNA markers in the Russian population: bioinformatic aspects. *Cardiovascular Therapy and Prevention*. 2020;19(6):2732. (In Russ.) doi:10.15829/1728-8800-2020-2732

ДНК — дезоксирибонуклеиновая кислота, ЭГГ — этногеографическая группа, SNP — Single Nucleotide Polymorphism (однонуклеотидный полиморфизм), FST — fixation index, index F-statistics (индекс фиксации, индекс F-статистики).

Введение

Определение частот клинически значимых генетических полиморфизмов у населения является обычной задачей для популяционных биобанков [1]. Данная информация полезна для фармакогенетических работ, поскольку разные этнические группы значительно различаются по частоте ключевых аллелей [2, 3]. При большом объеме выборок таких данные важны и для диагностики наследственных болезней, поскольку одним из основных критериев патогенности/непатогенности обнаруженных у пациента мутаций являются данные о частотах этих полиморфизмов в популяции.

Таким образом, для решения целого ряда задач медицинской генетики, популяционной генетики, фармакогенетики и других направлений, связанных с изменчивостью генофонда, часто возникает необходимость определения частот того или иного генетического маркера в популяциях России. При этом население России характеризуется исключительно высокой гетерогенностью генофонда [4, 5]. Поэтому данные о частоте генетических маркеров востребованы не в одной российской выборке, а в серии выборок, отражающих основное разнообразие народов и регионов страны.

Для решения этой, часто возникающей, задачи разработана технология популяционного скрининга, состоящая из молекулярно-биологического, биоинформатического и популяционного блоков. Биоинформатический блок используется для решения вычислительных и аналитических задач. В частности, предметом исследования, представленного в данной статье, является выделение в народонаселении России конкретных популяций, которые будут включены в стандартизованную панель скрининга. Этот набор популяций должен удовлетворять следующим требованиям:

1) охватывать население всей территории России;
2) включать территории сопредельных государств, демографически тесно связанных с Россией;
3) на каждой территории быть ориентированным на коренное население;
4) каждая выделенная популяция должна быть генетически достаточно гомогенна, чтобы определенные для нее частоты маркеров были репрезентативны для каждого из народов или регионов, входящих в состав популяции;
5) число популяций должно быть минимизировано, чтобы типичный объем выборки составлял не менее 50 образцов.

Поясним, что изучение каждого из коренных народов России (требование 3) необходимо как по смыслу популяционного скрининга и по практике популяционно-генетических исследований, так и с целью соблюдения его права на полноценное медицинское обслуживание (например, в тех случаях, когда скрининг ориентирован на фармакогенетически значимые маркеры, определяющие особую реакцию на лекарства). Поэтому на каждой территории популяции для скрининга представляют генофонд ее коренного населения, даже для территорий, где коренное население составляет меньшинство (например, на большей части Сибири). Чтобы оценить частоту фармакогенетического маркера, например, среди представителей русской национальности, проживающих в Южной Сибири, следует использовать значения, полученные не для популяции "Южная Сибирь", а для источника миграции — популяции "Русские Центральной России"; при этом для коренного населения Южной Сибири (хакасов, алтайцев, шорцев) следует использовать данные, полученные для популяции "Южная Сибирь".

Для формирования набора популяций ключевыми являются требования (4) и (5), однако требование гомогенности увеличивает число выделяемых групп, тогда как требование объемов выборки их уменьшает. Поэтому необходимо математически корректное выявление оптимального баланса между ними. На практике (например, в проекте "1000 Genomes") чаще всего применяют выборки ~50 человек (100 хромосом). Полученные к настоящему времени широкогеномные массивы данных по российским популяциям включают от 1 до 3 тыс. образцов [4-6]. Проект "Российские геномы" также ставит своей целью секвенировать 2-3 тыс. полных геномов представителей коренных народов России. Такие объемы доступных общих выборок (2-3 тыс. образцов) и требования к среднему объему одной выборки (~50 образцов) определяют, что максимальное число выделяемых популяций может быть 40-60. При этом для удобства использования и для дополнительного увеличения объемов выборок число популяций желательно минимизировать. Основные проведенные к настоящему времени геномные исследования российских популяций включают порядка 50-100 популяций [4, 5], причем большинство из них изучено по выборке всего в несколько образцов. Таким образом, возникает задача объединить несколько сотен популяций в 3-6 десятков более крупных групп, которые были бы гомогенны внутри себя и отражали бы основное разнообразие генофонда народонаселения страны.

В связи с вышесказанным целью работы являлось разделение народонаселения России и сопредельных стран на группы популяций, удовлетворяющие вышеперечисленным условиям, а также имеющие репрезентативную выборку в существующих данных и биобанках.

Материал и методы

Выбор методического подхода. Если переформулировать цель исследования на языке математических моделей, возникает задача кластеризации популяций, представленных в имеющихся широкогеномных массивах данных, на основе генетического разнообразия популяций. Существует много различных подходов к кластеризации, в т.ч. специально для генетических данных [7]. Методы глобально делятся на два подхода — параметрический и непараметрический. Параметрические методы опираются на статистические модели, в основе которых лежат несколько генетических допущений. К таким методам относятся STRUCTURE, ADMIXTURE и др. Однако данные методы являются ресурсоемкими, поэтому трудно применимы к разностороннему анализу больших массивов данных с большим количеством однонуклеотидных полиморфизмов — SNP (Single Nucleotide Polymorphism).

К непараметрическим методам относятся методы либо на основе уменьшения размерности и дальнейшей кластеризации [8], либо на основе расчета попарных расстояний между образцами и дальнейшей кластеризации. Непараметрические методы менее ресурсоемки и не требуют допущений.

Для данной работы был выбран непараметрический подход с уменьшением размерности методом главных компонент с применением программного пакета EIGENSTRAT/smartpca [9]. В качестве метода кластеризации использовали K-means [10] на главных компонентах, который дает результаты не хуже, чем STRUCTURE [11].

Чтобы избежать терминологической путаницы при использовании понятий "население", "народонаселение", "народ", "субэтническая группа", "этнографическая группа", "регион" и т.д. предлагаем термин "этногеографические группы" (ЭГГ) для обозначения групп популяций коренного населения, совокупность которых охватывает все население изучаемого крупного региона таким образом, что каждая ЭГГ генетически сравнительно гомогенна внутри себя, но при этом обладает генофондом, отличающимся от других ЭГГ.

Материал. Для работы использовались данные генотипирования широкогеномных панелей SNP-маркеров в различных популяциях Северной Евразии. Основной анализ проведен на массиве данных, полученном при генотипировании панели производства Illumina (Infinium OmniExome BeadChip Kit, Illumina; США), включающей 4,5 млн ДНК-маркеров. При этом выделяемые группы зависят от структуры генофонда самого населения, а не от используемой панели маркеров, поэтому очень близкие результаты могут быть получены и при использовании других широкогеномных панелей (например, Human Origin производства Affimetrix, США) или полногеномного секвенирования.

Общий объем используемого в анализе массива данных: 1883 образца, представляющие 247 популяций России и сопредельных стран. Минимальным размером ЭГГ был выбран порог в 25 образцов, оптимальным признавалась ЭГГ размером от 50 до 100 образцов. Таким образом, задачей было разделение данных 1883 образцов на гомогенные ЭГГ размером не <25 образцов каждая.

Фильтрация. Фильтрация данных проводилась в соответствии с описанием в разделе Результаты. Для фильтрации использовалось программное обеспечение PLINK [12]. Для первичной фильтрации применялась команда mind 0.1. Для фильтрации перед применением метода главных компонент последовательно использовались команды geno 0.05, maf 0.01, mind 0.1, indep-pairwise 1500 150 0.2. Для фильтрации перед расчетом значений FST (fixation index, index F-statistics, индекс фиксации, индекс F-статистики) последовательно применялись команды geno 0.1, maf 0.01, mind 0.05, indep-pairwise 1500 150 0.5.

Исключение родственников. Для поиска родственников ближе второй ступени применялось программное обеспечение KING 2.2.4 [13] в режиме related с настройками по умолчанию.

Построение графиков главных компонент. Для расчета значений главных компонент всех образцов применялась программа smartpca, входящая в пакет eigensoft. Данные конвертировались из формата plink (bed-bim-fam) в формат eigensoft (eigenstratgeno-snp-ind) с помощью программы convertf из того же пакета с параметрами по умолчанию. Конвертированные данные подавались на вход программе smartpca с параметрами по умолчанию, кроме параметра, задающего количество итераций исключения outliers (numoutlieriter). Данный параметр имел значение 3 при построении промежуточных графиков, и значение 0 при построении итоговых. Результаты работы smartpca визуализировались с помощью языка программирования Python 3, в т.ч. с использованием библиотек pandas, matplotlib и seaborn.

Кластеризация. Кластеризация проводилась на результатах работы smartpca на трех первых компонентах. Значения этих компонент для всех образцов считывались библиотекой pandas и подавались на вход методу sklearn.cluster.KMeans [14] с фиксированным количеством кластеров и фиксированным генератором псевдослучайных чисел для воспроизводимости результатов. Для итогового деления было выбрано количество кластеров, равное 30.

Расчет FST. Для расчета значений генетических расстояний между парами популяций FST применялась программа smartpca, входящая в пакет eigensoft. Данные конвертировались из формата plink (bed-bim-fam) в формат eigensoft (eigenstratgeno-snp-ind) с помощью программы convertf из того же пакета с параметрами по умолчанию. Конвертированные данные подавались на вход программе smartpca с параметром “fstonly: YES” для расчета FST. Результаты работы smartpca визуализировались с помощью языка программирования Python 3, в т.ч. с использованием библиотек pandas, matplotlib и seaborn.

Картографирование. Пространственное распределение выделенных этногеографических групп было картографировано в программе GeneGeo [15] по данным о географических координатах 247 исходных популяций, каждая из которых была отнесена к одной из 29 результирующих ЭГГ. Области на карте, раскрашенные в индивидуальный цвет для каждой группы, получены как объединение ячеек диаграммы Вороного с одинаковым номером группы, построенной на множестве точек, соответствующих исходным популяциям (на карте отмечены звездочками). Расстояния между точками при построении диаграммы Вороного вычислялись на сфере с использованием их географических координат. Важно понимать, что полученные таким образом области условны и границы между ними не следует трактовать как ареалы расселения соответствующих популяций. Однако можно говорить о том, что ареалы содержатся внутри этих областей, что позволяет наглядно представить географические закономерности в их пространственном распределении.

Результаты

Первым этапом работы было проведение фильтрации данных с помощью программного обеспечения PLINK. Был отфильтрован 51 плохо прочитанный образец (с покрытием <90%). Далее были исключены 19 образцов, для которых анализ с помощью программного обеспечения KING показал родство ближе второй степени с другим образцом из выборки. После данной фильтрации осталось 1813 образцов.

Для проведения всех анализов методом главных компонент применялся следующий алгоритм. Изначально выбирались образцы, для которых хотели применить метод (либо все образцы, прошедшие фильтрацию, либо только образцы из определенного региона). Затем исключались полиморфизмы, прочитанные у <95% образцов, и полиморфизмы, частота минорного аллеля которых была <1%. После этого снова исключались образцы, у которых прочитано <90% позиций из уже отфильтрованного набора. Затем исключались сцепленные полиморфизмы с коэффициентом r2 >0,2. Заключительным этапом построения было применение smartpca с тремя итерациями исключения outliers (кроме итоговых графиков, на которых pca проводился без исключения outliers).

Изначально были рассчитаны главные компоненты по всем отфильтрованным 1813 образцам. Затем были рассчитаны координаты центроидов популяций, как среднее значение по компонентам всех образцов данной популяции. Центроидам был присвоен вес, пропорциональный размеру выборки образцов из данной популяции. К имеющим вес центроидам по первым 3 компонентам был применен метод кластеризации k-means из программного пакета sklearn. Таким образом была выполнена кластеризация, имитирующая кластеризацию по отдельным образцам, однако обязательно относящая популяции к ЭГГ целиком.

Значение количества кластеров k подбиралось исходя из условия на размеры кластеров (не <25 образцов в каждом кластере и минимальное количество крупных кластеров). Оптимальным значением k было 30, т.е. метод давал разделение популяций на 30 ЭГГ. При данном k из 30 групп всего 8 имели <25 образцов, и всего 3 группы имели >100 образцов. Увеличение k увеличивало количество малых групп, не разделяя крупные, а уменьшение k увеличивало размеры крупных групп, не умень-

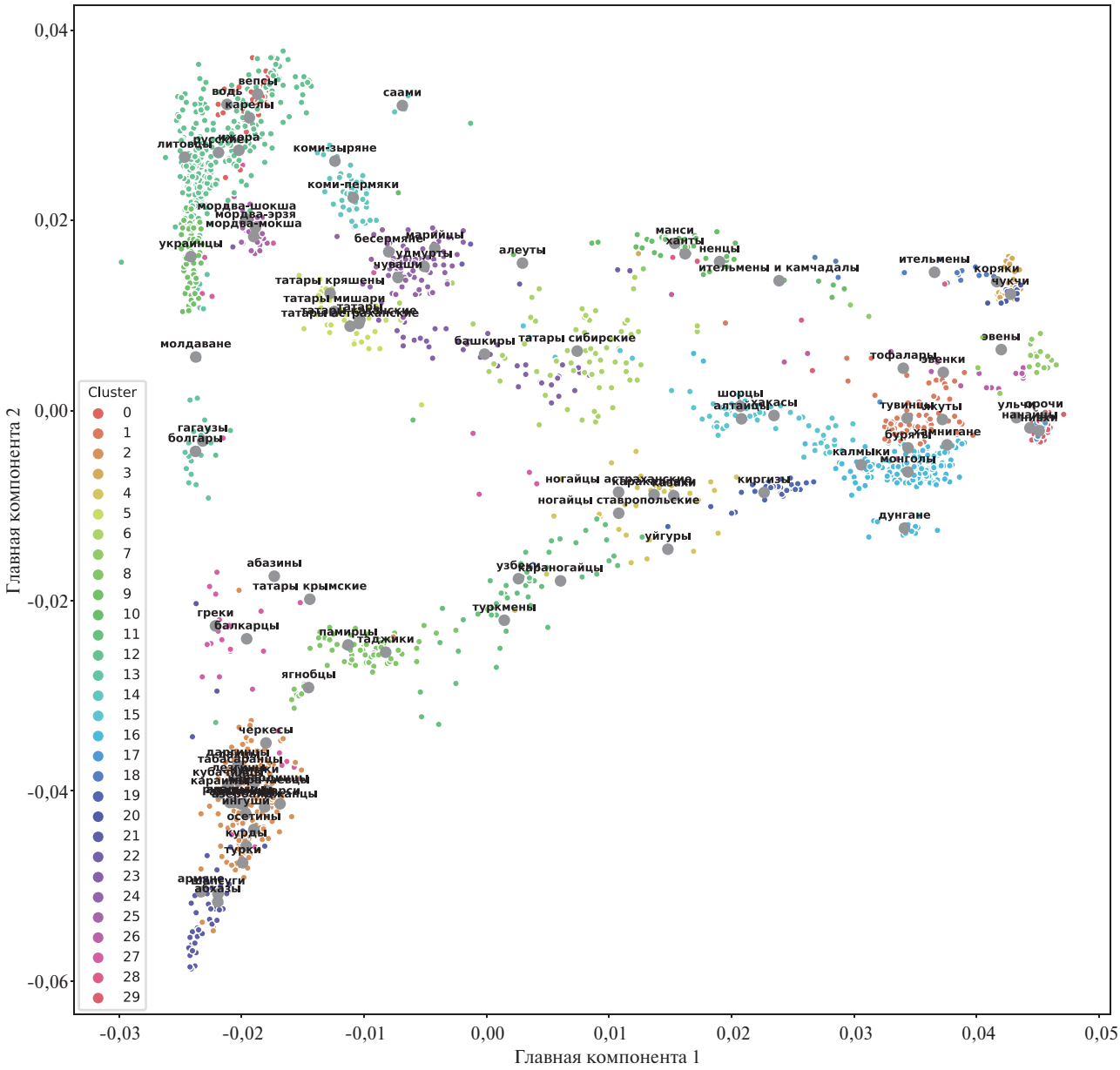

Рис. 1 График главных компонент 1 и 2 с разделением на ЭГГ методом K-means. Цветное изображение доступно в электронной версии журнала.

шая количество малых. Полученное деление на графике первых двух главных компонент приведено на рисунке 1.

Неравномерность разделения далее была исправлена вручную с применением дополнительных построений графиков главных компонент на подмножествах образцов.

Малые группы включали в себя группы, состоящие из следующих популяций: “алеуты”, “болгары, гагаузы и молдаване”, “ительмены”, “ительмены-камчадалы”, “нанайцы и орочи”, “нивхи и ульчи”, “чукчи”, “эвенки” и “эвены”. Учитывая внутреннее разнообразие данных популяций, объединение “эвенов” с “эвенками”, “нивхов и ульчей” с “нанайцами и орочи”, “ительменов” с “ительменами-камчадалами” и “чукчами” практически не нарушило гомогенность данных групп. Также к группе “ительмены и чукчи” были добавлены “коряки”, т.к. после объединения этой группы их внутреннее разнообразие было достаточно большим, чтобы включение “коряков” не нарушало гомогенность этой группы.

Малый размер и генетическое разнообразие “болгар, молдаван и гагаузов” и “алеутов” не позволил выделить их в отдельные группы или присоединить к какой-либо существующей группе. Тем не менее, группу из болгар и гагаузов стоило бы выделить отдельно, если бы она имела достаточное количество образцов, в то время как молдаване имеют слишком большое внутреннее разнообразие, и поэтому могут относиться как

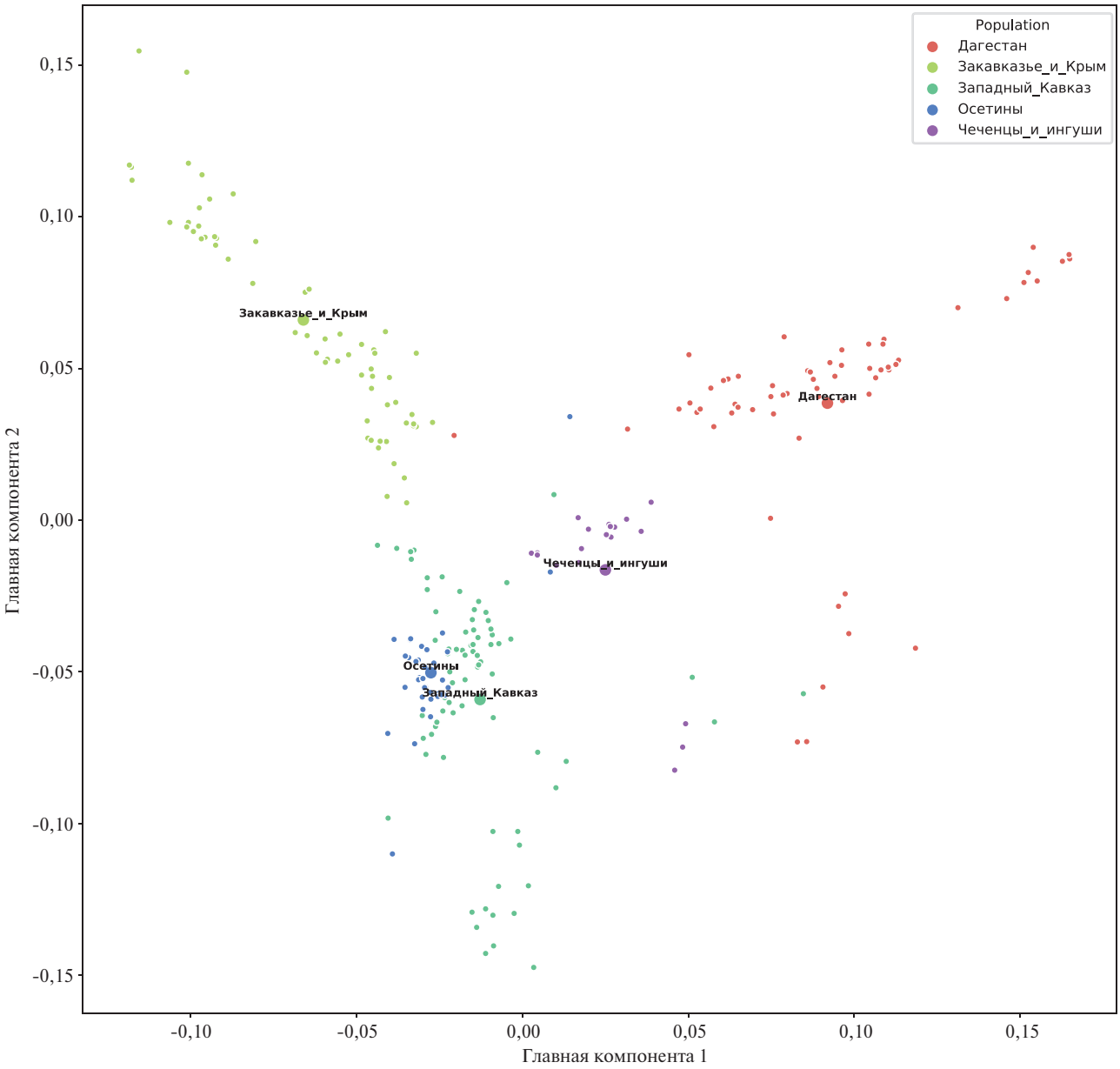

Рис. 2 График главных компонент 1 и 2 с разделением на ЭГГ для региона Кавказ. Цветное изображение доступно в электронной версии журнала.

к группе с болгарами и гагаузами, так и к группе с украинцами.

В свою очередь крупные группы включали в себя группы “популяции Кавказа”, “русские” и “буряты, дунганы, калмыки, монголы и хамнигане”. Все эти группы включают в себя достаточно разнообразные субпопуляции, поэтому они были разделены на более мелкие группы на основании графиков главных компонент для субпопуляций из данных групп. В качестве примера приведем график первых двух главных компонент с итоговым разделением на ЭГГ для региона Кавказ (рисунок 2).

В результате группа “буряты, дунганы, калмыки, монголы и хамнигане” была разделена на “калмыки и монголы” и “буряты, хамнигане и якуты” (якуты перемещены сюда, т.к. в результате этого деления они оказались ближе к этой группе, чем к изначальной “тофалары, тувинцы и якуты”), а популяция “дунгане” была исключена, т.к. сильно отличалась от остальных популяций этих групп. При наличии достаточного количества образцов “дунгане” составили бы отдельную группу.

Группа “русские” была разделена на три группы: “южные русские”, “северные русские” и “русские севера Архангельской области”. После этого разделения популяция водь оказалась ближе к группе “северные русские”, чем к карелам и вепсам, к которым ее отнесла кластеризация k-means изначально, в связи с чем она была перенесена.

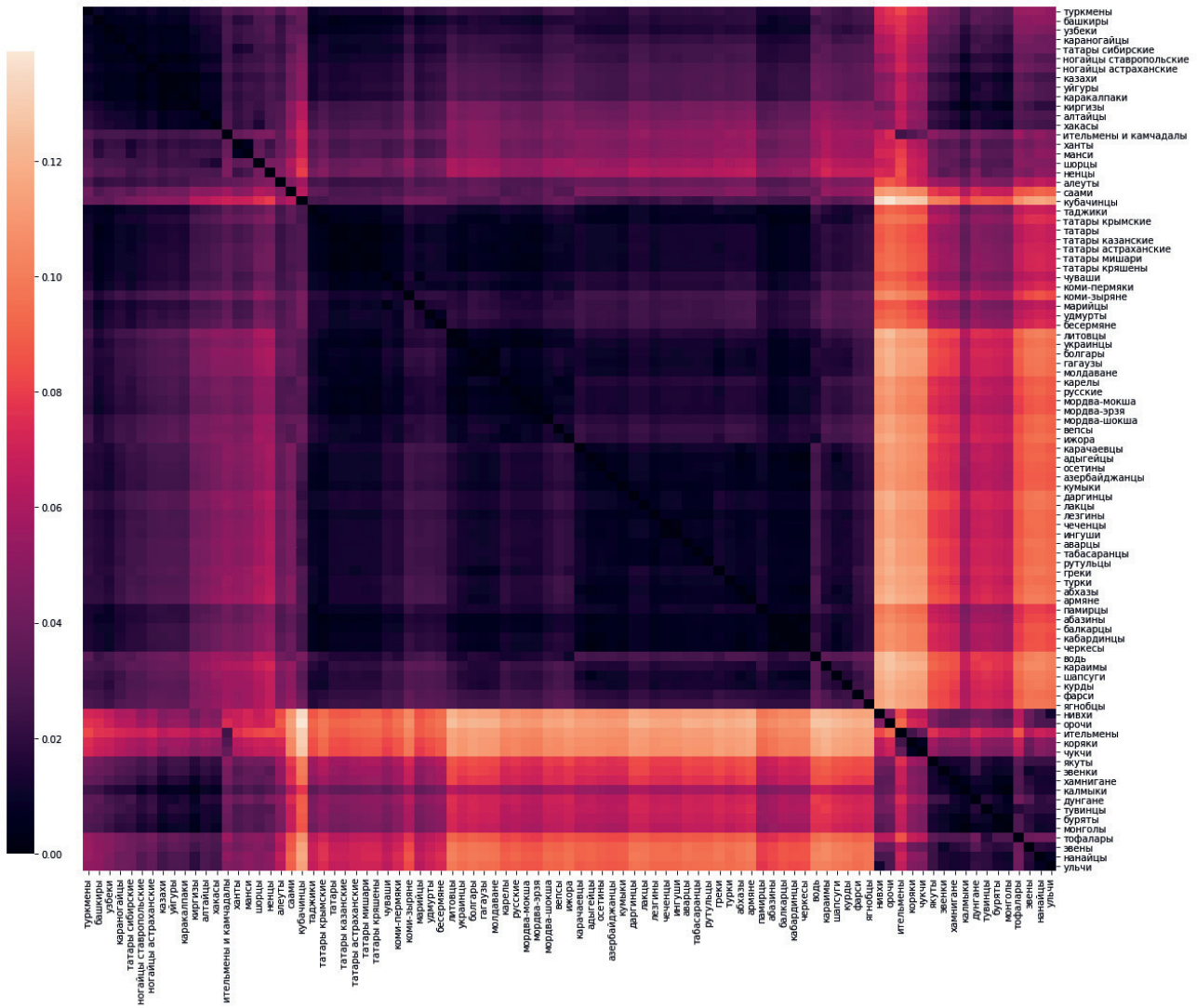

Рис. 3 Тепловая карта значений FST между парами этнических групп. Цветное изображение доступно в электронной версии журнала.

Группа “Кавказ” была разделена на четыре группы: “Дагестан”, “западный Кавказ”, “ингуши и чеченцы” и “осетины”, некоторая часть популяций была перенесена в группу “Закавказье и Крым”. Группа “ингуши и чеченцы” включила в себя 23 образца, однако должна была быть больше, т.к. часть образцов была исключена из нее во время фильтраций. В связи с этим мы оставили ее, как отдельную ЭГГ, хотя формально она не подходит из-за размера выборки.

Достаточно крупной оказалась и группа “алтайцы, хакасы и шорцы” (73 образца). При отдельном рассмотрении она отлично разделилась на “хакасы и южные алтайцы” и “шорцы и северные алтайцы”.

Также удмурты и бесермяне были перенесены из группы с мари и чувашами в группу с коми, а саамы исключены из этой группы, как вносящие излишнюю гетерогенность.

В связи с невозможностью определенно отнести к какой-то группе или выделить в отдельную ЭГГ по причине малых объемов выборок также были исключены греки, литовцы, караногайцы, фарси и кумыки.

При распределении популяций по группам в случае спорных моментов об отнесении популяции к той или иной группе использовались также значения FST между популяциями. Тепловая карта значений FST между парами этнических групп приведена на рисунке 3.

Итоговое разделение включило в себя 29 групп. Их состав приведен в таблице 1.

Итоговый график первых двух главных компонент после дополнительного анализа с исправлением неравномерности ЭГГ приведен на рисунке 4. По сравнению с рисунком 1 (на котором показан первоначальный вариант кластеризации) групп стало 29, а не 30, некоторые популяции исключены, а размер ЭГГ стал равномернее. Можно видеть, что выделенные ЭГГ охватывают все генетическое разнообразие населения России и сопредельных

Таблица 1

Итоговый набор ЭГГ

ЭГГ	Народы, включенные в ЭГГ	Количество образцов (в использованном массиве данных)
Башкиры	башкиры	43
Буряты и хамнигане	буряты, хамнегане, якуты	57
Дагестан	табасаранцы, аварцы, кубачинцы, даргинцы, лакцы, лезгины, рутульцы	68
Закавказье и Крым	армяне, азербайджанцы, караимы, турки-месхетинцы, курды, татары крымские	83
Западный Кавказ	карачаевцы, абхазы, адыгейцы, кабардинцы, шапсуги, балкарцы, черкесы, абазины	87
Казахи, каракалпаки, уйгуры и ногайцы	уйгуры, казахи, каракалпаки, ногайцы астраханские, ногайцы ставропольские	33
Карелы и вепсы	карелы, вепсы, карелы тверские	38
Киргизы	киргизы	35
Коми и удмурты	коми-пермяки, удмурты, бесермяне, коми-зыряне	84
Марийцы и чуваши	чуваши, марийцы	53
Монголы и калмыки	калмыки, монголы	126
Мордва	мордва-мокша, мордва-эрзя, мордва-шокша	40
Нанайцы, нивхи, орочи и ульчи	нанайцы, нивхи, ульчи, орочи	39
Осетины	осетины	36
Северные русские	русские, ижора, водь	76
Русские севера Архангельской области	русские	35
Сибирские татары	татары сибирские	68
Таджики, памирцы и ягнобцы	таджики, памирские народы, ягнобцы	72
Татары	татары, татары мишари, татары астраханские, татары кряшены, татары казанские	52
Тувинцы и тофалары	тофалары, тувинцы, монголы-тувинцы	56
Узбеки и туркмены	узбеки, туркмены	45
Украинцы	украинцы	79
Хакасы и Южный Алтай	алтайцы, хакасы	42
Ханты, манси и ненцы	ненцы, ханты, манси	53
Чеченцы и ингуши	чеченцы, ингуши	28
Чукчи, коряки и ительмены	коряки, чукчи, ительмены, камчадалы	67
Шорцы и Северный Алтай	алтайцы, шорцы	35
Эвенки и эвены	эвены, эвенки	45
Южные русские	русские	198

Примечание: ЭГГ — этногеографическая группа.

стран, при этом перекрывание между генетической изменчивостью групп минимальное.

Обсуждение

В результате применения методов биоинформатического анализа к массиву широкогеномных данных о популяциях Северной Евразии, был сформирован набор из 29 ЭГГ. Эти группы характеризуются генетической гомогенностью внутри себя, генетически различаются друг от друга, охватывают все генетическое разнообразие населения России и сопредельных стран, т.е. соответствуют требованиям, сформулированным в начале нашего исследования. На рисунке 5 представлена карта расположения выделенных групп.

Построенная карта фактически представляет собой схему районирования территории России и сопредельных стран по признаку генетического разнообразия населения. Важно, что карта получена по результатам наиболее подробного массива данных, охватывающего большое число конкретных групп населения (247 исходных популяций) и основанного на типировании большого числа маркеров, покрывающих весь геном. Поэтому проведенное районирование является если не финальным и не единственно возможным, то приближается к этой цели, поскольку опирается на репрезентативную выборку популяций и маркеров и получено путем применения объективных математических методов. В дальнейшем можно ожидать дополни-

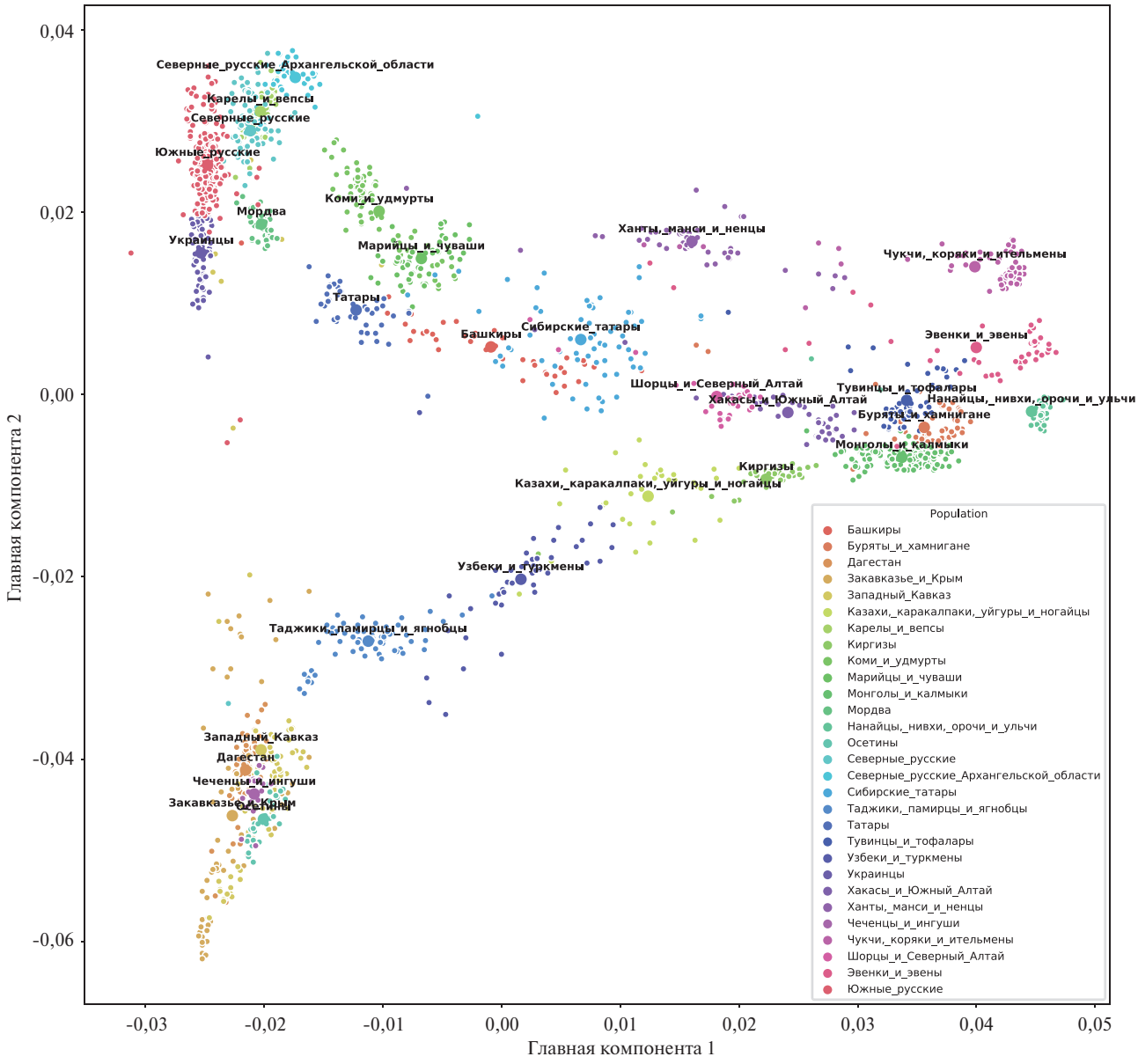

Рис. 4 График главных компонент 1 и 2 с итоговым разделением на ЭГГ. Цветное изображение доступно в электронной версии журнала.

тельного подразделения лишь среди ЭГГ Сибири, а также новых ЭГГ на периферии Европейской части, для выделения которых в текущем массиве данных оказалось недостаточно лишь объемов выборок (литовцы, болгары, фарси и т.д.).

Полученная схема деления генофонда имеет сходство с географическими и лингвистическими закономерностями. Хотя критерий географического соседства не применялся при выделении ЭГГ, практически все выделенные группы имеют целостные, а не разорванные ареалы.

Несмотря на то, что данные ЭГГ были получены на конкретном наборе образцов, выделенные группы могут применяться для анализа любых наборов данных о популяциях России, т.к. охватывают значительную часть популяций на рассматриваемой территории и имеют выборки, сходные по размеру с обычно используемыми в данной области исследований.

В частности, полученные группы могут использоваться в дальнейших исследованиях в области популяционной генетики для расчета частот ДНК-маркеров, применения методов машинного обучения или любых других подходов, требующих значительного числа образцов из каждой популяции. Полученная карта ЭГГ может использоваться при определении происхождения человека по его ДНК в генетической генеалогии и криминалистике. Наиболее существенным является тот факт, что сформированный набор ЭГГ, оптимизированный под существующие массивы широкогеномных данных и коллекций Биобанка Северной Евразии, мо-

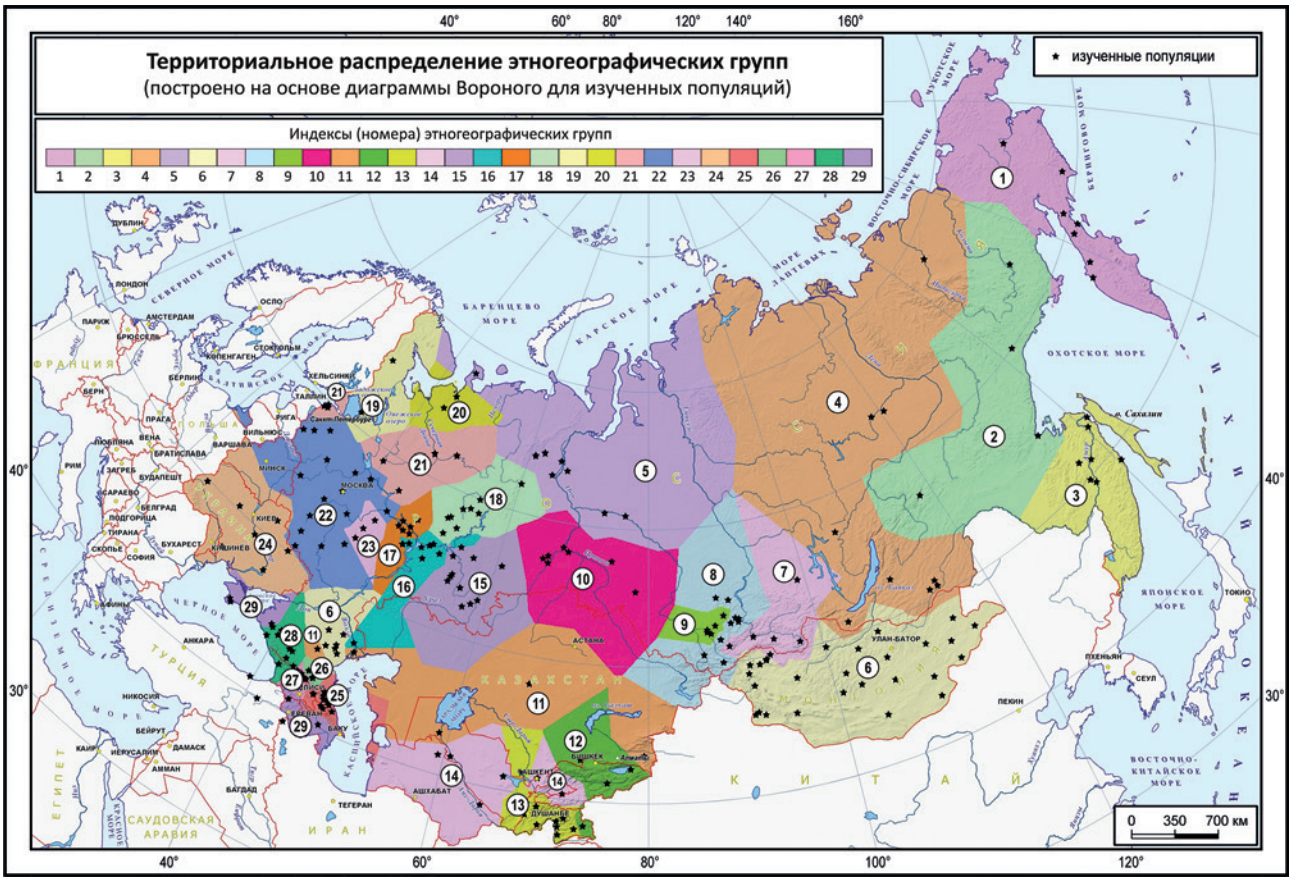

Рис. 5 Итоговая карта выделенных ЭГГ.
Примечание: цифрами в кружках на карте подписаны номера групп (совпадают с номерами в таблице 1), поскольку ввиду их большого количества определение конкретной группы по цвету может вызывать затруднения. Цветное изображение доступно в электронной версии журнала.

жет использоваться для систематического скрининга самых разных наборов генетических маркеров, имеющих медицинское значение. С применением этих коллекций и сформированного набора популяций нашим коллективом в 2020г уже проведены три исследования: распространение генетических маркеров, ассоциированных с тяжелым течением COVID-19; эпидемиология генетических полиморфизмов в гене TP53; распространение 43 фармакогенетически значимых маркеров. В ближайшие годы можно ожидать массового применения разработанной технологии популяционного скрининга и выделенного набора популяций в серии других работ.

Заключение

Популяционные биобанки позволяют определять частоты клинически значимых генетических полиморфизмов среди населения. Поскольку население России характеризуется высокой генетической гетерогенностью, необходимо объективное выделение конкретных популяций, для которых должны быть получены такие данные. Это число популяций не может быть ни слишком большим (это затруднит практическое применение, а главное, не позволит достичь репрезентативных наборов выборок), ни слишком малым (в этом случае выделенные популяции будут гетерогенны внутри себя, поэтому нерепрезентативны в отношении входящих в их состав этнических популяций). В результате объективного биоинформатического и математического анализа выделены 29 ЭГГ. Эти группы относительно генетически гомогенны внутри себя, их совокупность охватывает все генетическое разнообразие населения России и сопредельных стран, при этом их наложение на существующие массивы широкогеномных данных дает выборки порядка 30-70 образцов из каждой ЭГГ, что достаточно для оценки частот генетических маркеров. Построена карта, демонстрирующая деление народонаселения России и сопредельных стран на 29 территорий — ареалов ЭГГ.

Обоснованное разделение популяции на ЭГГ позволяет увеличить объем данных и улучшить качество результатов как для фармакогенетических работ, так и для ассоциативных исследований. Исследователи смогут обоснованно объединять данные по разным популяциям, входящим в одну ЭГГ, если эксперимент не требует популяционной дета-

лизации. Кроме того, это разбиение играет важную роль для увеличения объема данных, необходимого для повышения точности алгоритмов машинного обучения.

В целом, результирующая карта и реестр ЭГГ могут применяться и уже применяются в популяционно-генетических, медико-генетических, генетико-генеалогических и фармакогенетических исследованиях.

Благодарности: благодарим всех доноров образцов. Коллекция образцов ДНК предоставлена АНО “Биобанк Северной Евразии”.

Отношения и деятельность. Исследование выполнено в рамках Государственного задания Министерства науки и высшего образования РФ для Медико-генетического научного центра им. академика Н. П. Бочкова и для Института общей генетики РАН им. Н. И. Вавилова.

Литература/References

1. Balanovskaya EV, Zhabagin MK, Agdzhoyan AT, et al. Population biobanks: Organizational models and prospects of application in gene geography and personalized medicine. Russian Journal of Genetics. 2016;52(12):1371-87. (In Russ.) Балановская Е. В., Жабагин М. К., Агджоян А. Т. и др. Популяционные биобанки: принципы организации и перспективы применения в геногеографии и персонализированной медицине. Генетика. 2016;52(12):1371-87. doi:10.7868/S001667581612002X.
2. Jing L, Haiyi L, Xiong Y, et al. Genetic architectures of ADME genes in five Eurasian admixed populations and implications for drug safety and efficacy. J Med Genet. 2014;51(9):614-22. doi:10.1136/jmedgenet-2014-102530.
3. Mirzaev KB, Fedorinov DS, Ivashchenko DV, et al. ADME pharmacogenetics: future outlook for Russia. Pharmacogenomics. 2019;20(11):847-65. doi: 10.2217/pgs-2019-0013.
4. Triska P, Chekanov N, Stepanov V, et al. Between Lake Baikal and the Baltic Sea: genomic history of the gateway to Europe. BMC Genet. 2017;18(Suppl 1):110. doi:10.1186/s12863-017-0578-3.
5. Jeong C, Balanovsky O, Lukianova E, et al. The genetic history of admixture across inner Eurasia. Nat Ecol Evol. 2019;3:966-76. doi:10.1038/s41559-019-0878-2.
6. Balanovsky OP, Gorin IO, Zapisetskaya YS, et al. Interaction of the gene pools of the Russian and Finnish-speaking population of the Tver region: analysis of 4 million SNP markers. Vestnik RSMU. 2020;(6). (In Russ.) Балановский О. П., Горин И. О., Записецкая Ю. С. и др. Взаимодействие генофондов русского и финноязычного населения Тверской области: анализ 4 млн SNP-маркеров. Вестник РГМУ. 2020;(6). doi:10.24075/vrgmu.2020.072.
7. Alhusain L, Hafez AM. Nonparametric approaches for population structure analysis. Hum Genomics. 2018;12(1):25. doi:10.1186/s40246-018-0156-4.
8. Liu N, Zhao H. A non-parametric approach to population structure inference using multilocus genotypes. Hum Genomics. 2006;2(6):353-64. doi:10.1186/1479-7364-2-6-353.
9. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12):e190. doi:10.1371/journal.pgen.0020190.
10. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc. 1979;28:100-8. doi:10.2307/2346830.
11. Lee C, Abdool A, Huang C. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics. 2009;10 Suppl 1(Suppl 1):S73. doi:10.1186/1471-2105-10-S1-S73.
12. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. doi:10.1186/s13742-015-0047-8.
13. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867-73. doi:10.1093/bioinformatics/btq559.
14. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30. https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python.
15. Koshel SM. Geoinformation technologies in genogeography. Modern geographic cartography. 2012;158-66. (In Russ.) Кошель С. М. Геоинформационные технологии в геногеографии. Современная географическая картография. 2012;158-166. https://www.researchgate.net/publication/294848419_Geoinformacionnye_tehnologii_v_genogeografii.