УСПЕХИ СОВРЕМЕННОЙ БИОЛОГИИ, 2022, том 142, № 5, с. 442–450

УДК 575;575.17

ГЕНЕТИЧЕСКОЕ РАЗНООБРАЗИЕ И СТРУКТУРА ИЗОЛИРОВАННОЙ ПОПУЛЯЦИИ ЕВРОПЕЙСКОГО СЕВЕРА РОССИИ ПО ДАННЫМ ПОЛНОЭКЗОМНОГО СЕКВЕНИРОВАНИЯ

© 2022 г. Е. А. Гибитова[1, *], П. В. Добрынин[1, 2], О. Ю. Наумова[2], С. Ю. Рычков[2], О. В. Жукова[2], Е. Л. Григоренко[3]

[1]Национальный исследовательский университет ИТМО, Санкт-Петербург, Россия

[2]Институт общей генетики им. Н.И. Вавилова РАН, Москва, Россия

[3]Научно-технологический университет "Сириус", Сочи, Россия

**e-mail: e.gibitova@yandex.ru*

Поступила в редакцию 30.06.2022 г.
После доработки 30.06.2022 г.
Принята к публикации 30.06.2022 г.

Представлены результаты анализа экзомных вариаций в сельской популяции Европейского Севера России. Популяция, исходно возникнув как монастырское поселение, развивалась на протяжении своей полутысячелетней истории в относительной культурной и географической изоляции. Определенную уникальность популяции составляет установленная высокая превалентность в ней нарушений развития языка и речи. На основе данных полноэкзомного секвенирования жителей поселения одного поколения (дети), мы показываем, что популяция характеризуется высоким уровнем инбридинга и аутозиготности, а также значимо выраженной долей генетического груза в виде вредоносных однонуклеотидных вариантов и структурных вариаций генома. Причем, значимое большинство этих вариантов сосредоточено в локусах, контролирующих развитие ЦНС, что позволяет предполагать, что накопленный генетический груз может быть связан с высокой превалентностью нарушений речи в популяции.

Ключевые слова: Европейский Север России, изолированное сельское поселение, полноэкзомное секвенирование, мутационная нагрузка, аутозиготность

DOI: 10.31857/S0042132422050052

ВВЕДЕНИЕ

В последние годы развитие технологий полногеномных исследований позволило накопить обширные данные о генетической изменчивости в глобальных, континентальных и региональных популяциях и в различных этнических группах. Это, в свою очередь, в значительной мере стимулировало прогресс в исследованиях генетических детерминант, определяющих сложные фенотипические признаки, и генетических вариантов, ассоциированных с развитием этих признаков и его нарушениями. Примером этому могут служить исследования по поиску геномных ассоциаций с уровнем кровяного давления (Ehret et al., 2016; Surendran et al., 2016), ростом (Yang et al., 2010; Wood et al., 2014), индексом массы тела (Locke et al., 2015), глобальными и специфическими когнитивными способностями (Bearden, Glahn, 2017; Lam et al., 2017) и другими сложными признаками. На сегодня опубликовано более пяти тысяч работ по полногеномному поиску ассоциаций с более чем тремя тысячами фенотипов (Watanabe et al., 2019). Эти исследования используют различные стратегии для повышения статистической силы анализа геномных сигнатур, ассоциированных с полигенными признаками, – это исследования на семейных выборках, метаанализы, популяционные исследования и, отдельно стоящие в этом ряду, исследования популяционных изолятов (Uffelmann et al., 2021). Ключевым преимуществом последнего является то, что редкие функциональные варианты могут присутствовать с более высокой частотой в такой, как правило, генетически более однородной популяции, что значительно повышает вероятность их выявления (Hatzikotoulas et al., 2014).

В данной публикации мы представляем первые генетические данные об отдельном сельском изоляте Европейского Севера России; здесь и далее исследованный изолят обозначается как популяция AZ. Популяция AZ – это жители отдаленного русского сельского поселения Архангельской области – агломерации деревень, возникновение которой связано с основанием монастыря. Поселение было основано в XV в. несколькими семья-

ми из близлежащих деревень, а за последующее столетие численность его возросла до нескольких тысяч человек, в том числе за счет иммиграции из других деревень области. На протяжение последних нескольких веков, поселение, окруженное лесами и болотами, оставалось географически и культурно относительно изолированным. С начала XX в. и по наше время оно претерпевало драматическую депопуляцию, сокращалось число деревень агломерации и общая численность поселения. Так, согласно данным переписей населения 1905 и 2020 гг., к началу XX в. численность поселения AZ составляла около 3.5 тыс. жителей, а в наши дни число зарегистрированных жителей составляет 470 человек.

Согласно реконструкции родословных в AZ, проведенной группой исследователей в 2005–2010 гг. на основе опроса населения и записей в церковно-приходских книгах (Rakhlin et al., 2011, 2013), популяция характеризуется высоким уровнем достаточно близкородственных браков. По данным этой реконструкции, основанной на более чем 6 тыс. индивидуальных записей, было показано, что около 80% насельников были связаны сложной родословной из 11 поколений, при этом свыше 70% нынешних жителей поселения связаны с этой родословной (Rakhlin et al., 2013).

Изолированность популяции AZ, ее демографическая история не являются уникальными, а, скорее, достаточно типичны для сельских регионов России и в особенности для Российского Севера. Однако, именно эта популяция вызывает особый интерес, т.к. в ней выявлена высокая превалентность нарушений развития языка и речи, которые проявляются как сложный многоаспектный наследуемый фенотип (Rakhlin et al., 2011, 2013, 2016). Это, в свою очередь, делает ее уникальным модельным объектом исследований генетических маркеров, ассоциированных с развитием языка и речи – одного из сложных когнитивных фенотипов со значимой, но до настоящего времени все еще слабо изученной, генетической компонентой (Mountford, Newbury, 2019).

В ходе работ последних трех лет, коллективом были собраны коллекция биологического материала и обширные поведенческие данные, включая детальные характеристики уровня развития языка и речи, от жителей поселения AZ. В данной публикации, в преддверие полногеномного поиска ассоциаций с целевым фенотипом, мы представляем результаты анализа экзомных вариаций в популяции AZ на выборке из одного поколения – детей, приводим сравнительные оценки генетического разнообразия популяции и накопленного в изоляте мутационного груза.

МАТЕРИАЛЫ И МЕТОДЫ

В данной работе была привлечена выборка из 108 детей из 71 семьи (все детское население AZ) в возрасте от 2 до 17 лет. В качестве источника геномной ДНК были собраны образцы слюны; образцы забирались с помощью специализированных наборов Oragene Discover OGR-500 производства DNAGenotek (Оттава, Канада). Выделение ДНК проводилось с использованием реагента prepIT.L2P с последующей преципитацией в этаноле, согласно протоколу производителя (DNAGenotek).

Полноэкзомное секвенирование и анализ данных секвенирования

Для установления геномных вариантов проводилось полноэкзомное секвенирование (WES, Whole-Exome Sequencing), в ходе которого использовалась панель и набор реагентов для экзомного обогащения TruSeq DNA Exome (hg19). Приготовление библиотек и секвенирование на платформе Illumina HiSeq4000 было проведено в Ресурсном Центре СПбГУ “БиоБанк”. При секвенировании 7–9-плексные пулы библиотек были объединены в одной дорожке HiSeq4000, что обеспечивало приблизительно стократное среднее покрытие генома (100X) — общепринятый стандарт глубины покрытия при WES (Rehm et al., 2013).

Проверка качества данных секвенирования проводилась с помощью FastQC v. 0.11.9 (Andrews, 2010) и оценки распределения k-меров (Marçais, Kingsford, 2011). Выравнивание фрагментов секвенирования и аннотирование к референсной последовательности GRCh37/hg19 проводились с использованием инструмента BWA (Burrows-Wheeler Alignment) v. 0.7.17-r1188 (Li, Durbin, 2009) с последующей сортировкой, дедупликацией и повторной калибровкой в соответствии со стандартами и рекомендациями GATK (van der Auwera et al., 2013); использовалась версия GATK 4.1.6.

Идентификация геномных вариантов

Обнаружение однонуклеоидных вариантов (Single Nucleotide Variant, SNV) проводилось с помощью HaplotypeCaller с последующей рекалибровкой вариантов с использованием VariantRecalibrator (GATK 4.1.6). Рекалибровка вариантов проводилась методом мягкой фильтрации при следующих уровнях чувствительности: 99.70 для нуклеотидных замен и 95.00 – для инсерций и делеций. Аннотация SNV к геному и референсным базам данных проводилась при помощи программ BCFtools и ANNOVAR (Wang et al., 2010).

Помимо SNV, на основе данных экзомного секвенирования и широко используемых алгоритмов сегментации были идентифицированы структурные геномные вариации или вариации числа копий

(Copy Number Variation, CNV). Выявление CNV выполнялось функцией exomecn.mops, реализованной в R-пакете cn.MOPS v. 1.32.0, при этом использовались два алгоритма сегментации: fastseg (Klambauer et al., 2012) и DNAcopy (Seshan, Olshen, 2021).

Оценка генетических характеристик популяции

Для оценки мутационной нагрузки (ML, Mutation Load) в популяции были проанализированы вредоносные миссенс-мутации и мутации с потерей функции (LoF, Loss-of-Function). Вредоносность миссенс-мутаций определялась при помощи шкалы Грантама (Grantham, 1974). Для этого набор вариантов был аннотирован программой ANNOVAR с добавлением аргумента “aamatrixfile”, мутации с индексом Грантама ≥ 150 оценивались как вредоносные. В качестве LoF рассматривались варианты в стоп-кодонах (нонсенс-мутации) и варианты, нарушающие сайты сплайсинга. Для выявления LoF-мутаций использовались алгоритм SnpEff (версия 5.1d) и встроенная база данных GRCh37.75. Индекс ML оценивался для каждого индивида раздельно для миссенс- и LoF-мутаций как отношение числа гомозиготных вариантов к общему числу альтернативных аллелей (Feng et al., 2019) по следующей формуле: ML = 2 × Nгом/(2 × × Nгом + Nгет), где Nгом – гомозиготные варианты мутаций, Nгет – гетерозиготные.

Анализ протяженных гомозиготных фрагментов в индивидуальных геномах (Runs of Homozygosity, ROH) проводился при помощи алгоритма BCFtools/RoH (Li, 2011), использующего скрытую марковскую модель для идентификации ROH – метод, широко применяемый для полноэкзомных данных (Narasimhan et al., 2016). Для статистик генетического груза и непрерывных гомозиготных участков генома была проведена стратификация популяции (из каждой семьи был оставлен только один сиблинг), и собственно анализ проводился на 71 образце. Перед анализом родственности данные были поданы на контроль качества при помощи программы plink, версия 1.9 (Purcell et al., 2007), в ходе которого был применен ряд метрик: maf 0.05, geno 0.05, hwe 0.001. Анализ степени родства индивидов на отфильтрованном наборе данных был проведен при помощи программы King, версия 2.2.4 (Manichaikul et al., 2010).

Для сравнения ряда генетических показателей в изученной популяции AZ с таковыми в генеральной популяции региона в анализ вовлекались данные полногеномного секвенирования 25 индивидов – жителей Архангельской обл., из базы данных проекта Human Genome Diversity Project (HGDP) (Bergström et al., 2020). Здесь и далее эта выборка обозначена как HGDP-группа. Для получения сопоставимых показателей полногеномные данные по HGDP-группе трансформировались в экзомные в соответствии с координатами панели, используемой для секвенирования AZ – TruSeq DNA Exome (hg19).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Степень родства и паттерны аутозиготности индивидов из популяции AZ

Согласно полученным оценкам, большая часть индивидов из популяции AZ – представителей одного поколения, являются достаточно близкими родственниками. Так, среднее значение попарного индекса родства в детской популяции AZ составило 0.105 ± 0.113, что сближается с соответствующей оценкой для родства четвертой степени или для кузенов – 0.125. Иерархический кластерный анализ индивидов на основе попарных индексов родства указывает на присутствие основного ядра наиболее близкородственных индивидов и ряд относительно дистанцированных семейных кластеров (рис. 1). Это, по-видимому, может отражать историю первичного заселения и последующих волн прироста численности популяции за счет семей – выходцев из других поселений области. Кроме того, это наблюдение согласуется с процитированными нами ранее данными исследований родословных (Rakhlin et al., 2013), согласно которым около 70% насельников AZ могут быть связаны единой родословной.

Отмеченный, условно двукомпонентный, состав генофонда AZ воспроизводится и по данным анализа протяженных участков гомозиготности (ROH) в индивидуальных геномах. На рис. 2 показано соотношение суммарной длины ROH и их количества в геномах индивидов из популяции AZ и архангельской группы сравнения HGDP. Как видно на графике, выборка HGDP, очевидно, относится к аутбредной популяции относительно большого размера, что отражается в соответствующих паттернах ROH – преобладание коротких фрагментов, суммарная длина которых не превышает 100MB (Ceballos et al., 2018). В то же время, паттерны ROH в популяции AZ разительно отличаются; здесь наблюдаются два четко выраженных кластера индивидов. Первый объединяет носителей более коротких ROH (N = 29), что обычно свойственно адмиксным популяциям. Второй, более многочисленный (N = 42), включает индивидов с преимущественно длинными ROH или большим бременем аутозиготности, что указывает на высокий уровень инбридинга в этой группе.

Мутационная нагрузка в популяции AZ

Для оценки ML в популяции мы проанализировали распределение вредоносных миссенс-мутаций (индекс Грантама ≥ 150) и LoF-мутаций (нонсенс- и сплайсинг-мутации) в экзомах индивидов из популяции AZ и из группы сравнения HGDP. В сравнении с выборкой HGDP, в популяции AZ обнару-

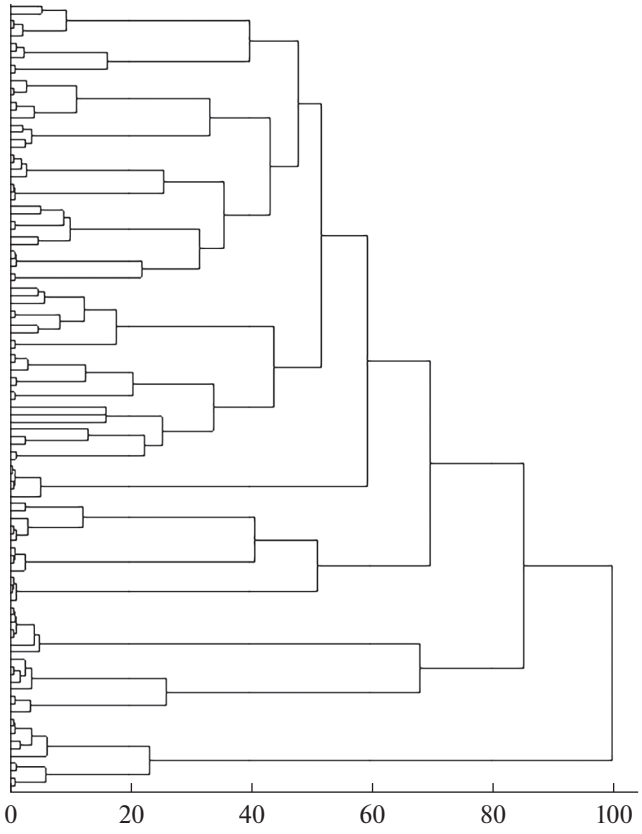

Рис. 1. Кластерный анализ индивидов из популяции AZ (N = 108 детей из 71 семьи) на основе попарных индексов родства. Для построения иерархического древа применялся метод Ворда, в качестве расстояний использовались логарифмированные значения индекса родства. Шкала расстояний дана в условных единицах.

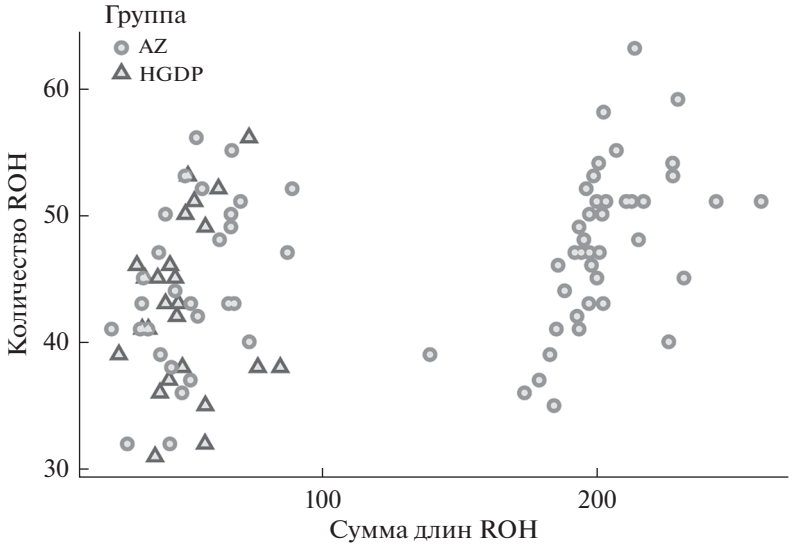

Рис. 2. Распределение индивидов (N = 71 ребенок, сиблинги исключены) из изученной популяции AZ и группы сравнения HGDP в пространстве статистик, описывающих бремя протяженных участков гомозиготности (ROH) в геномах: суммарная длина ROH (ось абцисс) и общее количество ROH (ось ординат).

Таблица 1. Статистика по распределению миссенс- и LoF-мутаций и оценки мутационной нагрузки в изученной популяции AZ и в группе сравнения HGDP

Тип мутаций	Популяция	Мутационная нагрузка, ML		Гомозиготные варианты, число		Гетерозиготные варианты, число	
		среднее ± SD	U-тест	среднее ± SD	U-тест	среднее ± SD	U-тест
Миссенс	AZ	0.725 ± 0.029	$W = 3$ $p < 0.001$	208.747 ± 9.014	$W = 1667.5$ $p < 0.001$	158.704 ± 25.471	$W = 1775$ $p < 0.001$
	HGDP	0.544 ± 0.018		229.240 ± 9.778		385.120 ± 19.424	
LoF	AZ	0.478 ± 0.041	$W = 875$ $p = 0.920$	56.409 ± 5.533	$W = 1741$ $p < 0.001$	123.113 ± 11.741	$W = 1750$ $p < 0.001$
	HGDP	0.477 ± 0.030		72.400 ± 5.745		158.520 ± 10.728	

жилось значимо более низкое абсолютное число гомозиготных и гетерозиготных вариантов миссенс-мутаций и мутаций с потерей функции (табл. 1). Это косвенно еще раз указывает на более низкое разнообразие геномных вариантов в изученной популяции AZ.

Мутационная нагрузка миссенс-вариантами для популяции AZ оказалась значимо выше, чем для HGDP-группы (U-критерий Манна–Уитни–Уилкоксона: $W = 3$, $p < 0.001$), что указывает на значимое накопление гомозиготных миссенс-мутаций в популяции AZ (табл. 1, рис. 3). В то же время, различий в нагрузке LoF-мутациями между популяциями не обнаружилось ($W = 875.0, p > 0.05$), что говорит о высоком селективном давлении против гомозиготных мутаций с потерей функции в обеих выборках, вне зависимости от иерархического статуса и демографической истории популяций, которые они представляют.

Для того чтобы оценить возможные функциональные последствия накопления в популяции вредоносных вариантов, для 897 генов, несущих такие SNV, был проведен анализ обогащения генных онтологий (GO, Gene Ontology). Анализ показал значимое обогащение в данной группе генов молекулярных процессов и биологических путей, вовлеченных в развитие клеток и морфогенез клеточных структур, и, в первую очередь, клеток ЦНС. Так, среди наиболее значимо перепредставленных онтологий отмечены биологические процессы, связанные с нейронной дифференциацией, морфогенезом нейронов и развитием нейрональ-

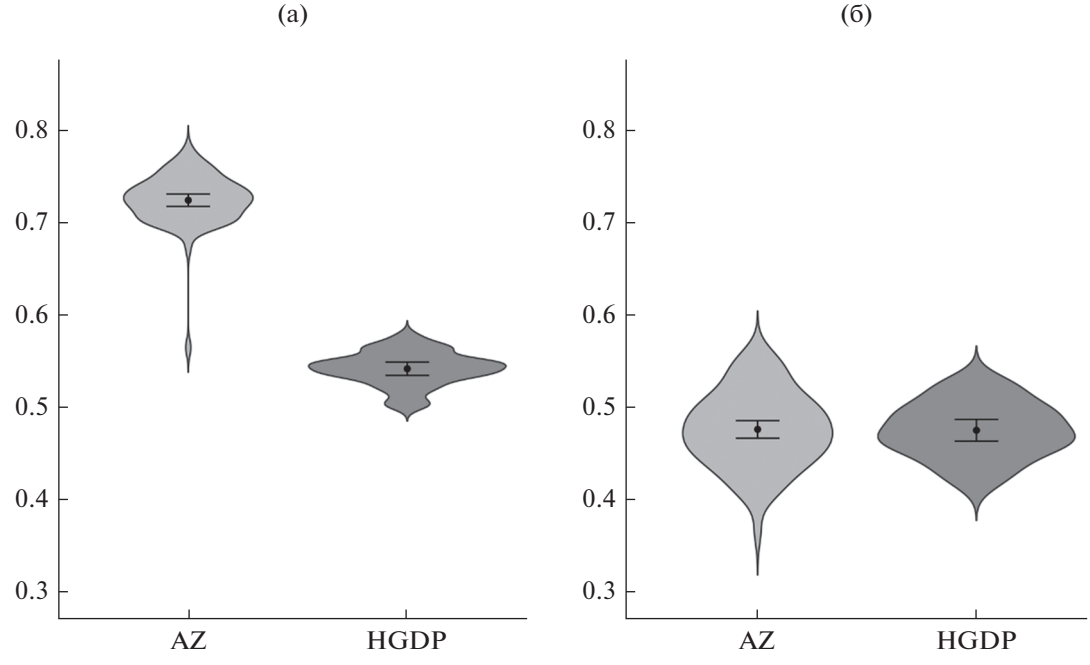

Рис. 3. Скрипичные графики, показывающие распределение индекса мутационной нагрузки, или доли гомозиготных вариантов, оцененного для вредоносных миссенс-мутаций (а) и LoF-мутаций (б) в экзомах индивидов из изученной популяции AZ и группы сравнения HGDP. Планки погрешностей показывают 95%-ные доверительные интервалы для средних.

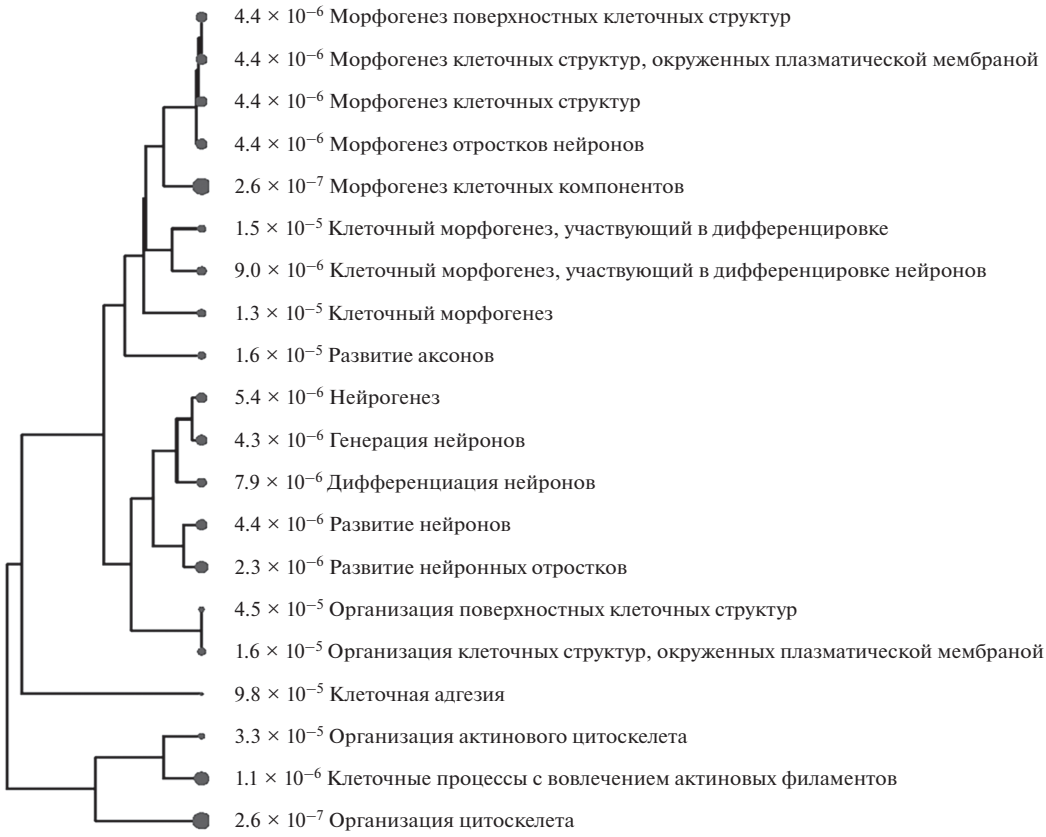

Рис. 4. Иерархическое древо, обобщающее результаты анализа обогащения по функциональной принадлежности пула генов, несущих вредоносные мутации в геномах индивидов из популяции AZ. Показаны 20 первых генных онтологий (GO) в классе “биологические процессы”, ранжированных по *p*-значениям теста перепредставленности. Размеры точек на древе отражают уровень статистической значимости теста для каждой GO; приведены *p*-значения, скорректированные на множественные сравнения. Функциональные категории сгруппированы в кластеры в соответствии с количеством общих генов. Древо построено с помощью аналитического инструмента ShinyGO (по: Ge et al., 2020, адаптировано).

ных отростков, как это отражено на рис. 4, суммирующем результаты тестов на обогащение.

Структурные геномные вариации в популяции AZ

Мы выявили свыше 20 структурных геномных вариаций в популяции AZ; для ряда локусов, несущих эти CNV, известны ассоциации с различными клиническими фенотипами (табл. 2). Среди CNV наиболее масштабных в контексте размера делеций и дупликаций и с наиболее высокой частотой в популяции AZ были отмечены вариации в следующих геномных локусах. Это – CNV в локусе 11q14.3, затрагивающие полностью белоккодирующий ген с неизвестной на сегодня функцией *TRIM49L2*, отмеченные у порядка 70% AZ-индивидов с преобладанием делеций ($p_{дел}$ = 0.465) над дупликациями ($p_{дуп}$ = 0.235). Делеции в этом локусе известны в ассоциации с абнормальными показателями ЭЭГ головного мозга (табл. 2). Два блока распространенных CNV наблюдались на хромосоме 17. Первый – в районе гена *LGALS9C* ($p_{дел}$ = 0.265 и $p_{дуп}$ = 0.015) в локусе 17p11.2 – районе, ассоциированном с синдромом Смит–Магениса, а второй – в районе гена *KANSL1* в локусе 17q21.31 ($p_{дел}$ = 0.485 и $p_{дуп}$ = 0.103), ассоциированном с синдромом Кулена–де Вриза. Свыше 85% индивидов AZ являлись носителями CNV, преимущественно делеций ($p_{дел}$ = 0.456 и $p_{дуп}$ = 0.029), в локусе 19q13.42 – кластере высокогомологичных генов иммуноглобулин-подобных рецепторов клеток-киллеров *KIR2DL1*, *KIR3DP1*, *KIR2DL4* и *KIR2DS4*. Последние играют важную роль в регуляции иммунного ответа, как следствие, мутации в этих генах ассоциированы с широким спектром клинических фенотипов и врожденных пороков развития, таких как синдром Дуэйна и синдром Ваарденбурга.

Помимо вышеперечисленных вариаций, которые частично перекрываются (в плане генного контента) с казуальными для фенотипов CNV, был обнаружен ряд более редких в популяции и неперекрывающихся с “клиническими” CNV событий,

Таблица 2. Структурные геномные вариации (CNV) и частоты их распространения в популяции AZ

Локализация CNV		Частота CNV		Ассоциированный клинический фенотип
хромосома	координаты (hg19)	делеция	дупликация	
1p36.11	25294163-25294643	0.147	0.015	–
1p13.3	109687813-109709039	0.353	0.000	–
5q31.3	140794851-141479625	0.044	0.000	Синдром делеции 5q31.3
5q35.3	180967188-180982611	0.191	0.000	–
6p22.1	29887751-29943067	0.206	0.000	–
6p21.33	31164336-31402250	0.529	0.000	–
6p21.32	32549939-32731695	0.529	0.029	–
8p23.1	7967343-7997521	0.324	0.015	Синдром делеции 8p23.1 Синдром дупликации 8p23.1
	12115766-12182719	0.044	0.000	
	12362018-12425000	0.559	0.000	
8p11.22	39018614-39920890	0.074	0.250	–
11q14.3	89944034-89949402	0.456	0.235	Абнормальная ЭЭГ*
14q11.2	18614387-18980742	0.162	0.015	–
14q11.2	19485335-19921659	0.103	0.015	–
15q13.3	32175246-32387484	0.118	0.000	–
17p11.2	18432050-18605356	0.265	0.015	Синдром Смит–Магенис* Синдром Потоки–Лупски
17q12	38003975-38257192	0.000	0.074	Синдром дупликации 17q12
17q21.31	45435899-46337794	0.485	0.103	Синдром Кулена–де Вриза*
19p13.2	7030577-7058640	0.000	0.118	Синдром дупликации 19p13.2
19p13.2	42752685-43248646	0.074	0.000	–
19q13.42	54738512-54753052	0.456	0.029	Синдром Дуэйна Синдром Ваарденбурга*
22q11.21	18773664-18884682	0.044	0.147	Синдорм дупликации 22q11.21
	21141623-21284161	0.029	0.088	–
Xq28	154884971-154885558	0.074	0.000	Синдром делеции Xq28

Примечание: показаны аппроксимированные границы CNV. Данные об ассоциациях клинических фенотипов с CNV в хромосомных бэндах приведены по материалам базы данных MalaCards (Rappaport et al., 2013); при отсутствии ассоциированного фенотипа поставлены прочерки. При перекрывании границ казуальных перестроек с CNV, обнаруженными в AZ, фенотипы отмечены звездочкой.

таких как вариации в хромосомных бэндах 5q31.3, 8p23.1, 17q12, 19p13.2, 22q11.21 и Xq28. Все эти локусы известны в ассоциации с одноименными синдромами (табл. 2), связанными прежде всего с нарушениями и задержками психофизического развития, включая развитие речи.

ЗАКЛЮЧЕНИЕ

Подводя итоги проведенного нами исследования экзомных вариаций у жителей поселения AZ, следует отметить, что данная популяция может достаточно однозначно определяться не только как относительный географический изолят, но и как, в какой-то мере, изолят генетический с относительно высоким уровнем инбридинга и значимо выраженной долей генетического груза. В пользу этого свидетельствует, во-первых, распределение и высокие оценки попарного индекса родства в популяции; во-вторых, в сравнении с условно аутбредной популяцией из того же географического региона, генофонд популяции AZ отличается значимым накоплением гомозиготных миссенс-мутаций с высоким показателем предсказанной вредоносности; и, в-третьих, большинство индивидов в популяции (дети из 60% обследованных семей) характеризуются высоким уровнем аутозиготности, их индивидуальные геномы содержат относительно небольшое количество длинных фрагментов гомозиготности. Такое бремя ROH соответствует демографическому сценарию развития популяции от относительно небольшой исходной группы и при достаточно высоком уровне близкородственных связей. В целом, этот сценарий хоро-

шо соотносится с историческими свидетельствами об образовании данного поселения, в истоках которого предполагаются несколько семей из окружающих поселений.

В то же время, детальное рассмотрение индивидуальных паттернов ROH, таких статистик как отношение суммарной длины фрагментов гомозиготности к их количеству, указывает на гораздо более сложный состав генофонда популяции. Наряду с описанным выше ядром из условных потомков первопоселенцев, в популяции выделяется группа (около 40% семей) с иным паттерном ROH – носители преимущественно коротких фрагментов гомозиготности. Такой паттерн обычно связан с адмиксией, которая сопровождается объединением различных гаплотипов и уменьшением количества ROH, преимущественно коротких по размеру. Возможно, это является следствием притока в AZ, в периоды роста численности поселения за счет иммиграции, групп из популяций относительно большего размера и генетически относительно отдаленных от первопоселенцев.

В контексте предполагаемого в дальнейшем исследования геномных ассоциаций с нарушениями развития языка и речи, выявленные в AZ на основе паттернов ROH генетические кластеры, несомненно, будут нами учтены. Не исключено, что данные “субпопуляции” могут различаться в плане превалентности нарушений в целом и/или распространения форм проявления этого сложного полидоменного фенотипа. В этом же ключе, многообещающими для поиска ассоциаций видятся наблюдения, сделанные по результатам геномного аннотирования однонуклеотидных потенциально вредоносных вариантов и структурных вариантов, CNV. Первые, как показал анализ генного функционального обогащения значимо перепредставлены в генах, вовлеченных в контроль развития ЦНС. Вторые – распространенные в популяции CNV, зачастую расположены в геномных районах, структурные пертурбации в которых известны в ассоциации с целым рядом синдромов, связанных с нарушениями развития, в спектре которых и нарушения психоречевого развития.

БЛАГОДАРНОСТИ

Авторы хотели бы выразить особую признательность детям и родителям – жителям исследованного поселения, за участие в исследовании и предоставлении биологических образцов для генетического анализа. Также, мы искренне благодарим сотрудников и студентов Лаборатории междисциплинарных исследований развития человека СПбГУ за помощь в сборе материалов в ходе экспедиционных работ.

ФИНАНСИРОВАНИЕ

Работа была выполнена при поддержке гранта РНФ № 18-18-00451.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют, что у них нет конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Все процедуры, выполненные в исследовании с участием людей, соответствуют этическим стандартам институционального и/или национального комитета по исследовательской этике и Хельсинкской декларации 1964 г. и ее последующим изменениям или сопоставимым нормам этики. Для каждого из включенных в исследование несовершеннолетних участников было получено информированное добровольное согласие от родителей или официальных опекунов. Исследование было одобрено Этическим Комитетом СПбГУ (Протокол № 02-155 от 20 июня 2018 г.).

СПИСОК ЛИТЕРАТУРЫ

Andrews S. FastQC: A Quality control tool for high throughput sequence data [Online] // Babraham Bioinformatics: website. 2010. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (access date: 08.06.2022)

Bearden C.E., Glahn D.C. Cognitive genomics: searching for the genetic roots of neuropsychological functioning // Neuropsychology. 2017. № 31 (8). P. 1003–1019.

Bergström A., McCarthy S.A., Hui R. et al. Insights into human genetic variation and population history from 929 diverse genomes // Science. 2020. № 367 (6484). P. eaay5012.

Ceballos F.C., Joshi P.K., Clark D.W. et al. Runs of homozygosity: windows into population history and trait architecture // Nat. Rev. Genet. 2018. № 19. P. 220–234.

Ehret G.B., Ferreira T., Chasman D.I. et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals // Nat. Genet. 2016. № 48. P. 1171–1184.

Feng S., Fang Q., Barnett R. et al. The genomic footprints of the fall and recovery of the crested ibis // Curr. Biol. 2019. № 29. P. 340–349.

Ge S.X., Jung D., Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants // Bioinformatics. 2020. № 36. P. 2628–2629.

Grantham R. Amino acid difference formula to help explain protein evolution // Science. 1974. № 185. P. 862–864.

Hatzikotoulas K., Gilly A., Zeggini E. Using population isolates in genetic association studies // Brief. Func. Genom. 2014. № 13. P. 371–377.

Klambauer G., Schwarzbauer K., Mayr A. et al. cn.MOPS: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate // Nucl. Acid. Res. 2012. № 40 (9). P. e69.

Lam M., Trampush J.W., Yu J. et al. Large-scale cognitive GWAS meta-analysis reveals tissue-specific neural expression and potential nootropic drug targets // Cell Report. 2017. № 21. P. 2597–2613.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data // Bioinformatics. 2011. № 27. P. 2987–2993.

Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform // Bioinformatics. 2009. № 25. PP. 1754–1760.

Locke A.E., Kahali B., Berndt S.I. et al. Genetic studies of body mass index yield new insights for obesity biology // Nature. 2015. № 518. P. 197–206.

Manichaikul A., Mychaleckyj J.C., Rich S.S. et al. Robust relationship inference in genome-wide association studies // Bioinformatics. 2010. № 26. P. 2867–2873.

Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers // Bioinformatics. 2011. № 27. P. 764–770.

Mountford H.S., Newbury D.F. International handbook of language acquisition. London: Routledge, 2019. 586 p.

Narasimhan V., Danecek P., Scally A. et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data // Bioinformatics. 2016. № 32. P. 1749–1751.

Purcell S., Neale B., Todd-Brown K. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses // Am. J. Hum. Genet. 2007. № 81. P. 559–575.

Rakhlin N., Kornilov S.A., Kornilova T.V. et al. Syntactic complexity effects of russian relative clause sentences in children with and without developmental language disorder // Lang. Acquisition. 2016. № 23. P. 333–360.

Rakhlin N., Kornilov S.A., Palejev D. et al. The language phenotype of a small geographically isolated Russian-speaking population: implications for genetic and clinical studies of developmental language disorder // Appl. Psycholing. 2013. № 34. P. 971–1003.

Rakhlin N., Kornilov S.A., Reich J. et al. The relationship between syntactic development and theory of mind: evidence from a small-population study of a developmental language disorder // J. Neuroling. 2011. № 24. P. 476–496.

Rappaport N., Nativ N., Stelzer G. et al. MalaCards: an integrated compendium for diseases and their annotation // Database (Oxford). 2013. P. bat018.

Rehm H.L., Bale S.J., Bayrak-Toydemir P. et al. ACMG clinical laboratory standards for next-generation sequencing // Genet. Med. 2013. № 15 (9). P. 733–747.

Seshan V.E., Olshen A. DNAcopy: DNA copy number data analysis. R package version 1.66.0. [Online] // Bioconductor: website. 2021. URL: https://bioconductor.org/packages/release/bioc/html/DNAcopy.html (access date: 08.06.2022)

Surendran P., Drenos F., Young R. et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension // Nat. Genet. 2016. № 48. P. 1151–1161.

Uffelmann E., Huang Q.Q., Munung N.S. et al. Genome-wide association studies // Nat. Rev. Meth. Prim. 2021. № 1. P. 59.

van der Auwera G.A., Carneiro M.O., Hartl C. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline // Curr. Prot. Bioinform. 2013. № 43. P. 11.10.11–11.10.33.

Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data // Nucl. Acid. Res. 2010. № 38. P. e164.

Watanabe K., Stringer S., Frei O. et al. A global overview of pleiotropy and genetic architecture in complex traits // Nat. Genet. 2019. № 51. P. 1339–1348.

Wood A.R., Esko T., Yang J. et al. Defining the role of common variation in the genomic and biological architecture of adult human height // Nat. Genet. 2014. № 46. P. 1173–1186.

Yang J., Benyamin B., McEvoy B.P. et al. Common SNPs explain a large proportion of the heritability for human height // Nat. Genet. 2010. № 42. P. 565–569.

The Genetic Diversity and Structure of an Isolated Population from Northern European Russia Based on Whole-Exome Sequencing Data

E. A. Gibitova[a, *], P. V. Dobrynin[a, b], O. Yu. Naumova[b], S. Yu. Rychkov[b], O. V. Zhukova[b], and E. L. Grigorenko[c]

[a] *University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia*

[b] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia*

[c] *Sirius University of Science and Technology, Sochi, Russia*

**e-mail: e.gibitova@yandex.ru*

The analysis of exome variations in an isolated rural population from the North of European Russia is presented. This population originated as a monastic settlement and has a 500-year history of relative cultural and geographical isolation. A unique feature of this population is the high prevalence of speech and language disorders. Here, we perform a whole-exome study (WES) of one generation of individuals (children), where we show that the population is characterized by high levels of inbreeding, autozygosity, and genetic load including potentially deleterious single nucleotide variants (SNV) and copy number variation (CNV). Moreover, a significant majority of the SNV were found in loci controlling the central nervous system (CNS) development, and the majority of CNV were found in loci associated with developmental impairments. We assume that such a genetic load may in part underlie the high prevalence rate of speech and language disorders in the population.

Keywords: Northern European Russia, isolated rural population, whole-exome sequencing, mutation load, autozygosity