Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

На правах рукописи

СКИТЧЕНКО РОСТИСЛАВ КОНСТАНТИНОВИЧ

ВЛИЯНИЕ ЧАСТОТНОГО СПЕКТРА АЛЛЕЛЕЙ НА РИСКИ ЗАБОЛЕВАНИЙ В РАМКАХ КОГОРТНЫХ ИССЛЕДОВАНИЙ

Специальность:
1.5.7 – генетика

Диссертация
на соискание ученой степени
кандидата биологических наук

Артемов Н.Н.
Ph.D, Broad Institute/ Massachusetts General Hospital
Доцент-исследователь центра геномного разнообразия, Университет ИТМО

Санкт-Петербург – 2022

ОГЛАВЛЕНИЕ

СПИСОК ИСПОЛЬЗУЕМЫХ СОКРАЩЕНИЙ

1. ОШ – отношение шансов (OR – odds ratio);
2. CDCV – Common Disease Common Variant («распространенное заболевание – распространенный ДНК-вариант»);
3. CDRV – Common Disease – Rare Variant («распространенное заболевание – редкий ДНК-вариант»);
4. CV – частые аллельные варианты;
5. RV – редкие аллельные варианты;
6. AF – частота аллели (allele frequency);
7. ФСГС – фокально-сегментарный гломерулосклероз;
8. GWAS – genome-wide association study (полногеномного ассоциативного исследования);
9. ДНК – дезоксирибонуклеиновая кислота;
10. HLA – человеческий лейкоцитарный антиген (human leukocyte antigen);
11. кДНК – комплементарная ДНК;
12. РНК – рибонуклеиновая кислота;
13. OMIM – база данных менделевских наследуемых заболеваний и признаков (Online Mendelian Inheritance in Man);
14. dNTP – дезоксинуклеотид (deoxynucleotides);
15. ddNTP – дидезоксинуклеотид (dideoxynucleotides);
16. ПЦР – полимеразная цепная реакция;
17. АТФ – аденозинтрифосфат;
18. NGS – секвенирование следующего поколения;
19. emPCR – полимеразная цепная реакция в эмульсии;
20. SOLiD – система секвенирования путем олигонуклеотидного лигирования и детекции (sequencing by oligonucleotide ligation and detection);
21. SMS – секвенирования одной молекулы (single-molecule sequencing);

22. SMRT – SMS в реальном времени;
23. NIH – национальный институт здоровья (National Institutes of Health);
24. ПГЧ – проект «Геном человека»;
25. LD – неравновесие по сцеплению (linkage disequilibrium);
26. CVAS – ассоциативный анализ частых ДНК-вариантов (common variants association study);
27. RVAS – ассоциативный анализ редких ДНК-вариантов (rare variants association study);
28. SNP – однонуклеотидный полиморфизм (single-nucleotide polymorphism);
29. WGS – полногеномное секвенирование (Whole Genome Sequencing);
30. SV – структурные ДНК-варианты (structural variants);
31. WES – полноэкзомное секвенирование (whole-exome sequencing);
32. QTL – локус количественного признака (Quantitative trait locus);
33. PheWAS – феномные ассоциативные исследования (phenome-wide association studies);
34. LDSR – LD-регрессия (linkage disequilibrium score regression);
35. PCA – метод главных компонент (principal component analysis);
36. LMM – линейные смешанные модели (linear mixed models);
37. PRS – оценка полигенного риска (Polygenic Risk Score);
38. GCTA – геномный анализ сложных признаков (Genome-wide Complex Trait Analysis);
39. MR – менделевская рандомизация (mendelian randomization);
40. MAF – частота минорной аллели (minor allele frequency);
41. QQ-график – квантиль-квантиль график (QQ-plot);
42. CNV – вариация числа копий (copy number variation);
43. NWR – когорта жителей Северо-Запада России (Northwest Russia);
44. AFR – африканская популяция (Africans);
45. AMR – американская популяция, с примесью латиноамериканской (Admixed Americans);
46. EAS – восточно-азиатская популяция (East Asian);
47. FIN – финская популяция (finnish European);
48. NFE – европейская популяция без примеси финской народности (non-finnish European);
49. SAS – южно-азиатская популяция (South Asians);
50. CI – доверительный интервал (confidence interval);

51. UKB – UK Biobank;
52. OR – отношение шансов (odds ratio);
53. PTV – ДНК-варианты, приводящие к укорачиванию белка (protein truncating variants)

ВВЕДЕНИЕ

Актуальность работы. Важным аспектом общественного здравоохранения является прогнозирование индивидуального риска заболеваний, а также укрепление здоровья посредством использования персонифицированных профилактических стратегий. Ключевым для персонификации здравоохранения является понимание механизмов патогенеза заболеваний и соответствующих врожденных предрасположенностей, которые во многом определяют индивидуальную траекторию здоровья пациента. Информация о том, как ДНК-варианты связаны с рисками заболевания и как влияют на выживаемость пациента, является ключевой для развития персонализированной медицины и направленного поиска новых лекарственных препаратов [1,2].

С практической точки зрения, вариабельность отдельных генетических локусов может носить решающее значение в фармакологической адаптации пациента, т.е. способности человека отвечать на ту или иную терапию. Так, например, семейство ферментов *CYP450* отвечает за первую фазу метаболизма ксенобиотиков, и их активность может быть изменена генетическими ДНК-вариации, расположенными в соответствующих генах. Выявление таких генетических отклонений от референсного значения может помочь в прогнозировании фармакокинетики и фармакодинамики лекарств. Таким образом данный подход может напрямую оказывать влияние на выбор определенной терапии, которая с большей вероятностью обеспечит желаемый терапевтический эффект или снизит возможные побочные проявления [3,4]. Для лекарств с ограниченным диапазоном терапевтического воздействия, например, для препаратов антикоагулянтов, даже небольшое изменение функциональной активности может привести к либо недостаточному, либо чрезмерно высокому физиологическому эффекту. Это может создать риск осложнений для каждого отдельного пациента [5].

Определение патогенности генетических вариантов, выявленных с помощью генетического тестирования, в клиническом фенотипе является сложной задачей и требует дополнительных знаний в области генетики человека и клинической медицины. Так, ДНК-вариации с большим размером эффекта (ОШ≥2; отношение шансов) чаще

всего являются характерными для моногенных заболеваний, в то время как ДНК-вариации с умеренным и малым размером эффекта (ОШ<2) ответственны за олигогенные и полигенные заболевания, что согласуется с моделью бесконечно малых эффектов Фишера [6].

Однако это правило для некоторых случаев не является абсолютным из-за чего существует две противоположные модели, описывающие влияние частоты ДНК-варианта на патогенез заболевания: «распространенное заболевание – распространенный вариант» (Common Disease Common Variant; CDCV) и «распространенное заболевание – редкий вариант» (Common Disease – Rare Variant; CDRV). В контексте таких сложных заболеваний, как воспалительные заболевания кишечника (ВЗК) [7], гипертония [8], колоректальный рак [9], диабет [10], аутизм [11], шизофрения [12], риторика дискуссии ведется в двух направлениях. Во-первых, до сих пор проблемным является вопрос о фактическом количестве генов или генетических вариаций, вовлеченных в формировании того или иного признака. Во-вторых, хотя большинство генетиков утверждают, что генетическая основа большинства распространенных хронических заболеваний предполагает наличие нескольких генетических факторов, работающих в совокупности, существуют значительные различия в реальной частоте генетических вариаций, которые могут быть ответственны за конечный фенотип. Кроме того, значительные отличия в аллель-частотном спектре изучаемой когорты и другими независимыми когортами может привести к тому, что результаты исследования могут быть неприменимы к другим популяциям и ограничивают возможность масштабирования интерпретации [13–15].

Конфликт теорий CDCV и CDRV привел к тому, что сегодня медицинская генетика рассматривает сложную наследственную болезнь, как «аллельный спектр» и применяет более сложные модели для оценки генетического компонента [6,16,17]. Аллельный спектр болезни – совокупность вариаций различного характера, способствующих развитию заболевания, например, ДНК-вариации с низкой или высокой пенетрантностью, частые аллельные варианты (CV, т.е. вариации с частотой более 1 % в популяции) или редкие аллельные варианты (RV, т.е. вариации с частотой менее 1 %), полиморфизмы и мутации, соответственно [13]. Но, однако, в данном контексте в первую очередь необходимо говорить не столько о форме и этиологии болезни, сколько о функциональности конкретных вариаций и об их вкладе в развитие фенотипа.

Задача исследования врожденных рисков полигенных признаков усложняется неоднозначным соответствием между мутацией и фенотипом. Зачастую один и тот же ДНК-вариант оказывает влияние на предрасположенность к нескольким фенотипам одновременно, что может объяснять одновременное сосуществование моделей CDCV и

CDRV. Данный эффект широко распространен среди всех форм жизни и называется плейотропией. Ранее исследователи с использованием модельных организмов изучали плейотропные эффекты мутаций с потерей функции и делеции генов которые закономерно имели большее воздействие на родственные друг с другом признаки [18,19]. Можно выделить, например, исследования по оценке влияния естественного отбора на изменчивость сложных признаков, а также исследования по разработке показателей оценки риска комплексных заболеваний [20]. Классическими примерами плейотропии являются пары фенотипов: серповидно-клеточная анемия и резистентность к малярии [21], болезнь Хантингтона и снижение риска некоторых видов рака [22], а также другие виды взаимодействий признаков.

Актуальным примером сложного заболевания с неоднозначной этиологией, которая может объясняться плейотропией, является фокально-сегментарный гломерулосклероз (ФСГС) – нефропатия, наиболее распространенная в Африке. ФСГС – одно из немногих сложных заболеваний, для которого не проведено полногеномного ассоциативного исследования (GWAS). Это объясняется существующим предубеждением о том, что CV не несут значимого риска возникновения ФСГС. Однако последние исследования показывают, что полиморфизмы в гене *APOL1*, напротив, несут значимые риски болезни в африканской популяции. В то время как частоты данных аллелей (AF; allele frequency) и распространенность заболевания в европейской популяции на несколько порядков ниже. Подобный эффект, как и в примере с серповидно-клеточной анемией и малярией, наблюдается из-за влияния естественного отбора [23].

Таким образом, данное исследование является актуальным, так как представляет систематический анализ влияния частот аллелей и эффектов плейотропии на индивидуальную предрасположенность к проявлению дезадаптирующих признаков на примере конкретных когортных исследований.

Цели исследования. Целью настоящей работы является изучение свойств частотного спектра аллелей и кросс-популяционных рисков наследственных заболеваний с последующим выявлением антагонистической взаимосвязи признаков.

Задачи исследования:

1) оценить частотный спектр аллелей связанных с носительством аутосомно-доминантных наследственных заболеваний у жителей Северо-Западного региона Российской Федерации;

2) выявить генетические локусы с аллелями низкой и высокой популяционной частоты, ответственные за систематическое проявление множественных ассоциаций сразу с несколькими сложными признаками с использованием когорты из UK Biobank;
3) исследовать кросс-популяционные риски наследственных заболеваний связанные с носительством определенных аллелей и выявить возможные различия в распространенности и воздействии в разных популяциях.

Научная новизна и практическая значимость исследования. В данной работе впервые показан спектр генетических вариаций в России (в особенности для RV), и выявлены наиболее распространенные аллели риска аутосомно-рецессивных заболеваний, которые характерны для российской Северо-Западной когорты.

Новым научным достижением является систематический анализ плейотропных признаков в британском биобанке. В работе сделан вывод о том, как степень плейотропности зависит от частоты и эффекта генетической вариации, а также о том, каким образом конкретный генетический ДНК-вариант влияет на закрепление патогенных аллелей в популяции.

Для ФСГС настоящее исследование является новым наглядным примером того, как определенные локальные особенности эволюционного давления в африканской популяции способны через механизм плейотропии влиять на поддержание высокой популяционной частоты причинной аллели.

Теоретическая и практическая значимость исследования. Конкретные генетические вариации, найденные в данной работе, могут быть использованы в клинической практике для постановки молекулярных диагнозов пациентам с редкими менделевскими заболеваниями. Результаты и выводы из данной работы впоследствии были использованы в мета-исследовании при создании российского экзомного браузера RuSeq [24].

Выявленные генетические варианты высокой популяционной частоты, связанные с широким спектром заболеваний, позволяют уточнять модели индивидуального полигенного риска развития патологий.

Помимо вклада в понимание природы семейной и спорадической формы ФСГС, исследование кросс-популяционного переноса рисков на примере конкретного заболевания позволяют на механистическом уровне проследить за динамикой аллельных частот и эффектами эволюционного давления в различных популяциях. Найденный

ДНК-вариант и соответствующий ему ген вносят вклад в понимание коморбидности нефропатологий с точки зрения плейотропии.

Методология и методы исследования. В своей основе данная работа опирается на дизайн исследования типа «случай-контроль», где когорте людей с изучаемым признаком («случай») противопоставляется соответствующая ей когорта «контролей». Проводимый далее этап фильтрации генетических данных призван сократить долю ложноположительных результатов и увеличить степень достоверности выводов, полученных на этапе ассоциативных исследований [25].

В работе используется методология и методы обработки данных генотипирования и секвенирования [26]. Основной платформой при анализе популяционных данных служит программный продукт Hail 0.2 [27] для языка программирования Python, а также ряд биоинформатических библиотек для языка программирования R [26,28], с помощью которых осуществлялся поиск генетических ассоциаций и статистическая обработка найденных результатов. На этапе оценки плейотропности генетических вариаций использовалась кластеризация ДНК-вариантов по сходным фенотипам [25].

Основные положения, выносимые на защиту:

1) клиническая выборка российских жителей, проживающих на Северо-Западе России имеет риски моногенных аутосомно-рецессивных заболеваний, связанные с редкими генетическими ДНК-вариантами в кодирующей последовательности ДНК, которые во многом отличаются от ближайшей европейской популяции;
2) плейотропия характерна в большей степени для CV. Редкие ДНК-варианты, испытывающие большее эволюционное давление, как правило, селективно связаны с одним фенотипом;
3) риски заболевания в разных популяционных группах могут вызываться одними и теми же генетическими ДНК-вариантами, при этом их частота определяется факторами эволюционного давления, существующими в конкретной популяции.

Личный вклад автора в исследование

Автор настоящей диссертации принимал непосредственное участие в обработке результатов секвенирования экзомов российских пациентов, анализе качества данных, биоинформатическом анализе, получении результатов биоинформатического анализа. В

исследовании плейотропии автор принимал участие в анализе приоритизации ДНК-вариантов, обладающих плейотропным эффектом.

При изучении спорадической формы ФСГС автор настоящей диссертации лично участвовал в большинстве этапов исследования, таких как:

1) анализ данных секвенирования;
2) анализ контроля качества генотипических данных;
3) аннотация данных секвенирования;
4) кластеризация популяционной структуры когорты;
5) учет популяционной стратификации, т.е. подбор контрольных образцов для группы случаев;
6) ассоциативный анализ распространенных ДНК-вариантов в каждой популяции;
7) ассоциативный анализ редких ДНК-вариантов;
8) анализ перепредставленности специфических HLA типов в различных популяциях;
9) интерпретация результатов;
10) создание графиков и написание текста статьи.

Степень достоверности и апробация результатов

Достоверность полученных результатов подкрепляется выводами на основе статистически значимых наблюдений, а также репликацией результатов с использованием независимых когорт.

Основные положения диссертации доложены на внутрилабораторных семинарах в Университете ИТМО, в том числе в международной лаборатории «Компьютерные технологии» (2020 – 2021 г.) и ФГБУ «НМИЦ им. В. А. Алмазова» (2022 г.), а также на конференциях: 1) XLVIII научная и учебно-методическая конференция Университета ИТМО (29 января – 1 февраля 2019 года); 2) BGRS/SB-2020: 11th International Multiconference; 3) 2020 ESHG.

Публикации

Результаты исследования представлены в 4 научных публикациях, 4 из 4 индексируются системами цитирования Scopus и Web of Science:

1. Zlotina A. et al. A 300-kb microduplication of 7q36.3 in a patient with triphalangeal thumb-polysyndactyly syndrome combined with congenital heart disease and optic disc coloboma: a case report. // BMC Med. Genomics. 2020. Vol. 13, № 1. P. 175. [29]

2. Glotov O.S. et al. Whole‑exome sequencing in Russian children with non‑type 1 diabetes mellitus reveals a wide spectrum of genetic variants in MODY‑related and unrelated genes. // Mol. Med. Report. 2019. Vol. 20, № 6. P. 4905–4914. [30]
3. Skitchenko R.K.* & Barbitoff Y.A.* et al. Whole-exome sequencing provides insights into monogenic disease prevalence in Northwest Russia. // Mol. Genet. Genomic Med. 2019. Vol. 7, № 11. P. e964. [31] (* – совместное первое авторство)
4. Shikov A.E. et al. Phenome-wide functional dissection of pleiotropic effects highlights key molecular pathways for human complex traits. // Sci. Rep. 2020. Vol. 10, № 1. P. 1037. [32]

Структура и объем работы

Диссертация представлена на 157 страницах формата А4. Кегль шрифта основного текста – 12, тип шрифта – Times New Roman. В диссертации представлено 31 иллюстративный материал и 8 таблиц. Структура диссертации содержит 3 основные главы, а также введение, заключение, список используемых сокращений, список литературы и приложения. Количество наименований цитируемой литературы насчитывает 552 иностранных издания.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Становление генетики человека

Первые исследования наследственности фенотипических признаков получили распространение в 19 веке и были в основном сфокусированы на скрещивании растений. В частности, именно так были сформулированы основные принципы наследственности – законы Менделя. Применимость данных постулатов в исследованиях более сложных организмов – дрозофил, мышей и в конечном итоге, человека потребовала продолжительной и сложной работы.

В конце XIX века ученые обнаружили внутри клеточных ядер структуры, получившие название "хромосомы" (др.-греч. χρῶμα «цвет»), из-за способности специфических химических соединений окрашивать их. В 1902 году исследователи Уолтер Саттон (Колумбийский университет) и Теодор Бовери (Вюрцбургский университет) отметили соответствие поведения хромосом с теориями Менделя [33]. Их предположение состояло в том, что хромосомы несут наследственные факторы или генетический материал, что заложило основы для хромосомной теории Бовери-Саттона.

В 1904 году Томас Хант Морган начал исследовать процессы, воздействующие на наследственность и развитие, в университете Колумбии. Тем не менее, Морган был среди критиков хромосомной теории наследования, также как и другие ученые своего времени, и не желал принимать эту теорию. Морган утверждал, что часть научного сообщества склонна связывать такие явления как наследование признаков с известными структурами, такими как хромосома. Точно так же он утверждал, что если один ген (некоторые базовые термины «аллель» и «ген» будут введены лишь в 1909 году [34]) не объясняет признак, то ученые будут настаивать на том, что признак может объяснить любое другое количество генов.

В том же 1904 году ученые впервые обнаружили свидетельства полигенного характера наследования в результате скрещивания мышей с желтым и серым окрасом шерсти [35]. В ходе проведенной серии экспериментов наблюдалось нестандартное для

законов Менделя соотношение получаемых аллелей. Было обнаружено, что наличие сразу двух аллелей, отвечающих за желтую окраску шерсти, вызывало летальный фенотип. Как ген, который отвечал за окраску мог приводить к летальному исходу? Возможно, одна аллель вызывает пожелтение шерсти, но при двукратной экспрессии белка происходит критически важное нарушение, которое убивает животное. Таким образом, этот ген фактически оказывал влияние сразу на два фенотипа. Так впервые было обнаружено и задокументировано явление плейотропии.

Цитологические исследования половых хромосом начались еще в 1901 – 1905 гг. [36,37].

В 1910 году Томас Хант Морган провел эксперимент, который помог окончательно установить роль хромосом в наследственности. В процессе своего исследования Морган обнаружил, что генетический фактор, определяющий цвет глаз у *Drosophila melanogaster*, связан с фактором, определяющим пол. Этот результат явно указывал на связь цвета глаз и пола с хромосомами. Это открытие помогло Моргану и его коллегам утверждать, что именно хромосомы несут гены, которые обеспечивают передачу черт от родителей к потомству, а также открыл явление сцепленного наследования [38]. Сначала Морган взял белого мутанта и скрестил его с чистокровными красноглазыми самками мух. Затем он взял этих красноглазых самок и скрестил их с исходным белоглазым самцом-мутантом, чтобы определить, соответствует ли наследование цвета глаз законам Менделя. В полученном поколении мух, независимо от пола, на каждые три красноглазые мухи приходится одна белоглазая муха, независимо от пола. Белоглазые мухи были самцами, что указывало на то, что получаемое распределение частот признака не следовало соотношению Менделя в традиционном смысле.

Моргану было интересно, почему в его экспериментах у самок мух никогда не было белых глаз, и он рассмотрел несколько возможных причин этого явления. Одно из возможных объяснений заключалось в том, что белоглазые самки не рождались или умирали на раннем этапе развития, как и мыши с двойным количеством аллеля желтой окраски из вышеописанного исследования. Другими словами, эта гипотеза предсказывала летальность белоглазых самок мух, следовательно, среди потомства анализирующего скрещивания гетерозиготных (F1) красноглазых самок с белоглазыми самцами не должно быть белоглазых самок. Но получаемые результаты не совпали с его предсказаниями. Результаты скрещивания дали соотношение 1:1:1:1 для всех возможных комбинаций двух признаков. Главный вывод состоял в том, что белоглазость следует за паттернами наследования половых хромосом. Иронично, что именно эти результаты помогли основному критику хромосомной теории наследования Томасу Ханту Моргану, а также его

коллегам установить, что именно хромосомы несут гены, которые позволяют потомству наследовать черты своих родителей. Поскольку все сцепленные с полом признаки обычно наследуются вместе, Морган пришел к выводу, что Х-хромосома несет ряд дискретных наследственных единиц или факторов. Он принял и популяризировал термин «ген», который был введен датским ботаником Вильгельмом Йохансеном в 1909 году, и пришел к выводу, что гены, возможно, расположены в хромосомах линейным образом. В результате, в 1933 году Морган за свои открытия получил Нобелевскую премию по физиологии и медицине.

Переход от общей генетики к генетике человека характеризуется признанием значимости генетики в области здравоохранения. Медицинско-прикладная функция генетики человека не является ее единственным направлением. Например, палеогенетика помимо *Homo sapiens* также изучает предковые для современного человека виды. Но сегодня под современной генетикой человека чаще всего подразумевается именно медицинская генетика. В данной работе генетика человека, в первую очередь, будет рассматриваться в контексте ее медицинского применения.

Существует ряд отдельных ранних попыток связать знания генетики с наследственными заболеваниями человека. В 1940 г. в Великобритании появился первый учебник по медицинской генетике Фрейзера Робертса «Введение в медицинскую генетику» [39]. В 1934 году А. Фоллинг описал фенилкетонурию как причину умственной отсталости [40]. Попытки практического применения генетики в медицине можно найти даже в начале прошлого века [41].

1949 год стал знаковым для понимания роли наследственности в заболеваниях. Джеймс В. Нил описал серповидноклеточную анемию как аутосомно-рецессивный признак, а четыре месяца спустя в том же томе журнала Science Лайнус Полинг определил это заболевание как «молекулярное» [42,43]. В 1949 г. Дж. Б. С. Холдейн в первый раз оценил частоту мутаций у людей на основе анализа семи заболеваний человека примерно в 4×10^{-5} [44]. Также в 1949 г. в публикации под названием «Болезнь и эволюция» Дж. Б. С. Холдейн рассматривал инфекционные заболевания как потенциальные «агенты естественного отбора» у человека. В том же 1949 году был опубликован первый учебник по генетике человека, основан Американский журнал генетики человека (American Journal of Human Genetics, ASHG), а годом ранее — Американское общество генетики человека.

Современная генетика человека включает в себя медицинскую генетику, посвященную всем ее медицинским аспектам, и, более специфичный, прикладной раздел - клиническую генетику, включающую в себя практику диагностики и лечения генетических нарушений. Новые методы культивирования клеток и улучшенные

митотические хромосомные препараты для светового микроскопического анализа непосредственно привели к признанию в 1959-1960 гг. того, что некоторые заболевания человека возникают в результате определенных аберраций в числе или структуре хромосом, например, при трисомии 21, 18, 13 хромосом, а также из-за частичных хромосомных делеций или дупликаций. Поскольку каждое нарушение кариотипа было связано с отдельным заболеванием, можно было непосредственно определить связь между генотипом и фенотипом человека. Именно в 1959 году были проведены первые работы по определению хромосомного пола человека и как следствие остальных млекопитающих. Было выяснено, что именно наличие Y-хромосомы определяет мужской пол человека, независимо от количества X-хромосом [45,46]. Таким образом, 1959 год можно считать годом появления клинической генетики, а цитогенетику наукой внесшей наибольший вклад в ее развитие [47,48].

С 1960-х годов культивируемые клетки стали широко использоваться для исследования моногенных заболеваний человека (генетика соматических клеток). Клетки, гомозиготные по генетическому дефекту, можно отличить от гетерозиготных клеток. Путем слияния гомозиготных клеток от разных пациентов (получение гибридов клеток) ученые наблюдали возвращение мутантного признака к нормальному клеточному фенотипу, доказывая, что рассматриваемое заболевание является генетически обусловленным. С помощью биохимических анализов стали определять наследственные заболевания, связанные с обменом веществ человека, такие как нарушения аминокислот и лизосомные болезни накопления. Позднее в конце 1960-х годов была введена пренатальная диагностика.

1.2. Развитие секвенирования

1.2.1. Гены α- и β-глобина

В 1978-1979 гг. из-за обилия РНК-транскриптов и кодируемых белков в эритроцитах были впервые клонированы гены α- и β-глобина (*HBA* [49,50]; *HBB* [51]). Из-за того, что кровь является наиболее удобной и доступной для анализа тканью, а серповидно-клеточная анемия – широко распространенная среди всех популяций болезнь, сложилась благоприятная среда для исследований направленных на понимание патогенеза серповидно-клеточной анемии. Успехи в секвенировании пептидов привели к идентификации белковых субъединиц гемоглобина и патогенного ДНК-варианта серповидного гемоглобина за долго до клонирования генов [52]. Кроме того, обилие α- и β-глобина РНК в крови позволило клонировать кДНК генов *HBA* и *HBB* у человека и

мыши. Методы специфического расщепления ДНК эндонуклеазами рестрикции [53] и клонирования генов в лямбда-векторах [54], а также методы определения последовательности нуклеиновых кислот [55,56], дали возможность идентифицировать большинство патогенных вариаций β- и α-талассемии.

Кластер генов *HBB* был клонирован и секвенирован в 1979 [51], а в последующие годы было обнаружено множество патогенных ДНК-вариантов [51,57]. Изучение патогенных ДНК-вариаций в генах *HBA* и *HBB* дало обширные базовые знания о природе и последствиях мутаций в генах человека. Нонсенс-мутации, миссенс-мутации, замены кодонов терминации, ошибки сплайсинга различного рода (канонические динуклеотиды донорных и акцепторных сайтов, активация скрытых сайтов, новые сайты сплайсинга), промоторные области, дистальные регуляторные элементы, микроделеции и микродупликации, механизмы неравного скрещивания – первоначальное понимание этого было получено из исследований генов α- и β-глобинов [58]. Этот исторический момент дал первоначальное представление о значительной полиморфной изменчивости генома человека.

Помимо этого полиморфная изменчивость ДНК вокруг генов β-глобинов позволила узнать о структуре гаплотипов, неравновесии по сцеплению, горячих точках рекомбинации и спектрах мутаций, специфичных для популяции [59,60]. Гаплотипическая структура кластера генов бета-глобина оказала существенное влияние на выбор мутантных аллелей-кандидатов для секвенирования и открытие полного спектра патогенных ДНК-вариантов в данной популяции в эпоху до открытия цепной реакции (ПЦР).

В 1980-х существует ряд других примеров идентификации причинных генов на основе знаний о белке. Например, ген *LDLR* для семейной гиперхолестеринемии [61], ген *HEXA* для болезни Тея-Сакса [62], ген *GBA* для болезни Гоше [63], ген *F8* для гемофилии А [64] и ген *PAH* для фенилкетонурии [65].

1.2.2. Картирование генов

Изучение гена *HBB* дало первоначальное представление о значительной полиморфной изменчивости генома человека. С момента открытия полиморфного сайта рестриктазы *HpaI* располагающимся примерно в 5 000 нуклеотидов от 3'-конца гена *HBB [66]* были идентифицированы тысячи таких сайтов. В 1980 г. исследователи представили теоретическую основу для описания того, как причинный локус может быть сцеплен с маркером в геноме человека [67]. На практике это означало, что в больших семьях с

достаточным количеством больных и наличии достаточного количества полиморфных маркеров можно было успешно картировать неизвестный ген болезни на небольшом участке генома человека. Впервые это было проверено на болезни Хантингтона, что сделало эту болезнь первым менделевским генетическим заболеванием, картированным с использованием полиморфизмов ДНК [68]. Успех этой истории мотивировал исследования по картированию генов и последующему их клонированию путем поиска в «окрестностях» сцепленного с ним маркера [69].

В 1990-х повышенный интерес к картированию генов стимулировал необходимость открытия большого количества информативных полиморфных маркеров и создание карт сцепления для каждой хромосомы в начале, чтобы можно было сузить поиск местоположения гена, связанного с болезнью, до примерно одной мегабазы последовательности ДНК [70]. Использование образцов консорциума CEPH (Centre d'Etude du Polymorphism Humain [71]), инициированного в 1984 г., сыграло важную роль в создании этих карт. Были созданы достаточно подробные карты сцепления для каждой хромосомы [72,73], в то время как проект HapMap, начатый в конце 1990-х годов, предоставил множество полиморфных маркеров и оценку блоков неравновесного сцепления генома человека [74]. Анализ сцепления широко использовался для размещения генов, ответственных за менделевские фенотипы, в небольшом геномном интервале примерно 1 Мб. Параллельно с этим внедрение технологии ПЦР [75] в 1986 г. значительно облегчило изучение последовательностей ДНК, не требуя трудоемкого клонирования в различных векторах.

Позиционное картирование генов болезней оказало большое влияние на дальнейшее развитие секвенирования. В этот период были клонированы гены, ответственные за наиболее распространенные менделевские расстройства. Успех был основан на базовом знании о структуре генома (в основном карты сцепления), которая была разработана, методом анализа сцепления, наличии множества общих полиморфных участков ДНК и изучении больших семей со значительным числом пораженных лиц. Кроме того, разработка физических карт из библиотек клонированных сегментов генома человека и данных о хромосомах и соматических клетках также облегчила поиск генов [76]. Эпоха позиционного клонирования продолжалась до начала 2000-х годов [77]. 2 февраля 2000 г. база данных менделевских наследуемых заболеваний и признаков OMIM преодолела отметку в 1000 генов [78]. Первыми двумя генами напрямую связанными с болезнями и клонированными с помощью позиционного картирования, был ген хронической гранулематозной болезни [79] и X-сцепленный ген мышечной дистрофии Дюшенна *DMD* [80,81]. Существует огромный перечень наследственных заболеваний, для

которых был успешно идентифицирован ген с помощью позиционного клонирования: ретинобластома – ген *RB1* [82], ген *CFTR* для муковисцидоза [83,84], ген *TP53* в раке, предрасположенность к синдрому Ли-Фраумени [85], ген опухоли Вильмса *WT1* [86], нейрофиброматоз – ген *NF1* [87–89], ген колоректального полипоза [90], синдром Марфана – ген *FBN1[91]*, ген *APP*, сцепленный с одной из форм болезни Альцгеймера [92] ген Fragile X *FMR1* [93], ген *PMP22* одной из форм болезни Шарко-Мари-Тута [94], ген *MECP2* синдрома Ретта [95], гены *MSH2* и *MLH1* связанные с наследственной формой рака толстой кишки [96,97], ген пресенилина 1 *PSEN1*, ответственный за одну из менделевских форм болезни Альцгеймера [98], гены рака молочной железы и яичников [99,100] *BRCA1* и *BRCA2*, ген *ATM* атаксии-телеангиэктазии [101], ген *FGFR3* ахондроплазии [102,103], ген *SMN1*, связанный со спинальной мышечной атрофией [104], ген *TSC1* для туберозного склероза [105], ген *PTPN11* для одной из форм синдрома Нунан [106], *NIPBL* ген для синдрома Корнелии де Ланге [107,108] и ген *CDH7* связанный с CHARGE-синдромом [109].

1.2.3. Развитие технологий секвенирования

Первоначальные усилия по секвенированию были сосредоточены на наиболее доступных видах РНК, таких как рибосомные или транспортные РНК бактерий, или геномы одноцепочечных РНК-бактериофагов. Популярность прокариотической и вирусной РНК была обусловлена тем, что ее молекулы могут быть легко получены в больших количествах микробиологическим путем. Еще одним преимуществом РНК является отсутствие комплементарной цепи, и длина нуклеотидной последовательности, которая значительно меньше длины эукариотической ДНК. Кроме того, уже были известны и доступны ферменты РНКазы, способные разрезать цепи РНК в определенных местах. В результате попыток комбинирования разных химических методов выделения и очистки с помощью селективной обработки рибонуклеазой были получены первые деградированные фрагменты РНК [110]. В 1965 г. Роберт Холли и его коллеги впервые смогли получить полную последовательность нуклеиновой кислоты, тРНК аланина из *Saccharomyces cerevisiae* [111].

Благодаря усилиям Фреда Сэнгера и его коллегам появился новый метод прочтения нуклеиновых кислот с помощью использования меченых нуклеотидов и двумерного фракционирования [112]. Это позволило исследователям существенно расширить свои возможности по секвенированию [113–117]. С помощью этого метода в 1972 году лаборатория Уолтера Фирса секвенировала первый ген - ген оболочки бактериофага MS2

[118]. Через четыре года в 1976 году был секвенирован уже его полный РНК-геном [119]. Термин геномика, происходящий от генома (введен Винклером в 1920 г.), был введен в 1987 г. [120]. Он относился не только ко всем генам, но и к молекулам, регулирующим их функции, а также к ядерным структурам.

Позднее оригинальный метод секвенирования по Сэнгеру существенно изменялся. Во-первых, было добавлено использование ДНК-полимеразы фага-λ *Enterobacteria* с целью заполнения «липких»-5`-концов одноцепочечной вирусной ДНК радиоактивными нуклеотидами по одному [121,122]. Во-вторых, была проведена замена двумерного фракционирования, которое на тот момент проводилось в два этапа (электрофорез и хроматография) на одномерное разделение с помощью электрофореза в полиакриламидном геле [123]. В-третьих, Максам и Гилберт отказались от использования ДНК-полимеразы. Вместо того, чтобы полагаться на ДНК-полимеразу для создания фрагментов, ДНК с радиоактивной меткой обрабатывается химическими агентами, которые разрывают цепь по определенным основаниям [55]. В дополнении с использованием электрофореза это позволяло различать длины расщепленных фрагментов, и соответственно восстанавливать последовательность нуклеотидов в цепи. Именно с помощью такой вариации метода Сэнгер и его коллеги секвенировали первый ДНК-геном — геном бактериофага *ϕX174* (или «*PhiX*», который сегодня занимает положение во многих лабораториях по секвенированию в качестве положительного контроля) [124]. Это был первый метод секвенирования, получивший широкое распространение, и поэтому именно его можно считать настоящим рождением секвенирования ДНК «первого поколения».

В 1977 году метод секвенирования по Сэнгеру был существенно доработан. Была разработана методика «обрыва цепи» [125]. Метод обрыва цепи использует химические аналоги дезоксирибонуклеотидов (dNTP), называемыми «флуоресцентными обратимыми терминаторами», которые являются мономерами цепей ДНК. В дидезоксинуклеотидах (ddNTP) отсутствует 3'-гидроксильная группа, необходимая для удлинения цепей ДНК, и поэтому они не могут образовывать связь с 5'-фосфатом следующего dNTP [126]. Выполняя четыре параллельные реакции, содержащие каждое отдельное основание ddNTP, и анализируя результаты со всех четырех электрофоретических дорожек полиакриламидного геля с использованием автоматизированной радиографии можно сделать вывод об исходной последовательности. Сегодня метод дидезокси-обрыва цепи — или просто секвенирование по Сэнгеру — стал наиболее распространенным и используемым методом секвенирования ДНК. В современных условиях секвенирование

по Сэнгеру используется как способ технического контроля качества прочтения отдельных участков ДНК другими методами секвенирования.

В последующие годы в секвенирование по Сэнгеру был внесен ряд улучшений. В первую очередь была произведена замена радиометрического обнаружения (с помощью мечения тритием и фосфором) на флуорометрическое обнаружение, сильно упрощающее практическую работу в лаборатории. Во вторую очередь начал широко использоваться капиллярный электрофорез, который дополнительно увеличил чувствительность метода. Оба эти усовершенствования способствовали разработке все более автоматизированным машинам для секвенирования ДНК [127–132]. Также существует пример первого коммерческого секвенатора [133].

Машины секвенирования первого поколения производят считывания длиной чуть менее одной килобазы (1000 пар нуклеотидов). Для анализа более протяженных участков ДНК/РНК исследователи прибегали к методам, таким как «секвенирование дробовиком», , где перекрывающиеся фрагменты ДНК клонировались и последовательно прочитывались по отдельности, а затем объединялись в единую консенсусную последовательность — «контиг» [134,135]. Внедрение техник полимеразной цепной реакции (ПЦР) (ПЦР) [136,137] и технологий, связанных с использованием рекомбинантной ДНК [138] дополнительно способствовало геномной революции, предоставив возможность достигать высоких концентраций ДНК, необходимых для новых методов секвенирования.

Основной активный фермент также претерпел ряд изменений. Например, ДНК-полимераза фрагмента Кленова, это фрагмент ДНК-полимеразы *Escherichia coli*, лишенный экзонуклеазной активности от 5' до 3' конца, который производился путем разделения исходного фермента с помощью протеазы [139]. Изначально он использовался для секвенирования из-за своей способности эффективно встраивать ddNTP. Однако этот фрагмент в последствии был заменен на альтернативные версии, которые лучше адаптировались к дополнительным химическим фрагментам все более модифицированных dNTP, используемых в новых секвенаторах [140]. В конце концов, более новые дидезокси-секвенаторы, например, представленные линейкой моделей ABI PRISM, позволяли одновременно секвенировать сотни образцов [141,142]. В дальнейшем они были использованы в проекте «Геном человека», помогая получить первые черновые результаты на несколько лет раньше запланированного срока [143].

Параллельно с развитием секвенирования по Сэнгеру разрабатывалась еще одна технология, на основных идеях которой в дальнейшем были созданы первые секвенаторы нового поколения. Этот метод заметно отличался от существовавших тем, что он не определял нуклеотиды с помощью радио- или флуоресцентно меченных dNTP. Вместо

этого исследователи использовали люминесцентный метод измерения синтеза пирофосфата: он состоял из двухферментного процесса, в котором АТФ-сульфурилаза используется для превращения пирофосфата в АТФ, молекула которого затем используется в качестве субстрата для люциферазы, производя таким образом количество света, пропорциональное количеству пирофосфата [144,145]. Метод пиросеквенирования, впервые предложенный Полом Ниреном и его коллегами, обладал рядом особенностей, которые считались полезными: его можно было проводить с использованием природных нуклеотидов (вместо модифицированных dNTP, используемых в протоколах метода Сэнгера) и наблюдать в режиме реального времени [146–148]. Позже лицензия на пиро-секвенирование была передана биотехнологической компании 454 Life Sciences, основанной Джонатаном Ротбургом, где данная технология превратилась в первый успешный коммерческий продукт в области "секвенирования следующего поколения" (NGS).

Секвенаторы, произведенные компанией 454 (впоследствии приобретенной Roche), существенно увеличили количество прочтений последовательностей ДНК за один запуск, благодаря возможности параллельного выполнения множества реакций секвенирования. [149]. Прибор работает по особому принципу. В начале ДНК-библиотеки прикрепляются к гранулам с помощью последовательностей-адаптеров. Каждая из гранул помещается в отдельную лунку, в которой затем проводится амплификация с помощью ПЦР в эмульсии (emPCR), в результате которой образуются кластеры клонированных ДНК последовательностей [150]. На следующем этапе в каждую лунку с кластерами ДНК добавляется ДНК-полимераза, которая осуществляет реакцию пиросеквенирования. Далее выполняется серия последовательных циклов, где к ДНК, закрепленной на твердой фазе, последовательно добавляются дезоксинуклеотидтрифосфаты всех четырех типов: A, T, G, C. Если на секвенируемой цепи ДНК есть комплементарный к добавленному нуклеотид, то при образовании фосфодиэфирной связи побочным продуктом станет пирофосфат. Высвобожденные молекулы пирофосфата измеряются с помощью датчика под лунками. Детекция пирофосфата осуществляется за счет каскада химических реакций, который заканчивается выделением кванта света. Эта установка была способна производить прочтения длиной около 400–500 пар оснований (п.н.) для миллиона заполненных гранулами лунок [149]. Первой машиной для высокопроизводительного секвенирования, широко доступной для потребителей, была оригинальная машина 454, названная GS20, которая позже была заменена на прибор 454 GS FLX, предлагающий не только еще большее количество считываний (за счет большего количества лунок), но и более качественные данные [151]. Этот принцип выполнения параллельных реакций

секвенирования — стал тем, что определило секвенирование ДНК второго поколения [152].

Успешный пример 454 индуцировал появление целого ряда методов параллельного секвенирования. Наиболее важным среди них, вероятно, является метод секвенирования Solexa, который позже был приобретен компанией Illumina [151]. Вместо параллелизации путем проведения emPCR, данная технология подразумевает следующую последовательность действий: 1) исследуемая двуцепочечная ДНК фрагментируется; 2) с двух концов секвенируемой молекулы ДНК с помощью ДНК-лигазы пришиваются последовательности ДНК-адаптеров, которые представляют собой пару частично комплементарных олигонуклеотидов; 3) с помощью одного из пары адаптеров молекула ДНК закрепляется в проточной ячейке прибора; 4) молекула ДНК изгибается, образуя мостик; 5) в ходе изотермической ПЦР молекула ДНК реплицируется в ограниченной области, образуя локальные кластеры клонированной ДНК [153,154]. Этот процесс был назван «мостовой амплификацией» из-за того, что реплицирующиеся нити ДНК должны перекрещиваться, чтобы вызвать следующий раунд полимеризации соседних олигонуклеотидов, связанных с поверхностью [151]. Сам процесс секвенирования также использует флуоресцентные "обратимые терминаторы" dNTP, аналогично технологиям, описанным ранее. Эти терминаторы не позволяют последующим нуклеотидам присоединяться, так как флуорофор занимает позицию 3′-гидроксильной группы [155]. Модифицированные dNTP и ДНК-полимераза за несколько циклов проходят через праймированные одноцепочечные кластеры, связанные с проточной ячейкой. В каждом цикле идентификация включенного нуклеотида осуществляется с помощью матрицы CCD (charge-coupled device), которая возбуждает флуорофоры при помощи соответствующих лазеров. Несмотря на то, что первые устройства для анализа генома (Genome Analyzer, GA) изначально могли выполнять только чтения очень коротких участков (длиной до 35 п.н.), их преимущество заключалось в возможности генерирования данных с двусторонними концами (PE, pair ends), что позволяло записывать последовательность на обоих концах каждого кластера ДНК. Это достигается путем получения сначала одного считывания одноцепочечной молекулы ДНК, связанной с проточной ячейкой, а затем выполнения еще одного раунда твердофазной амплификации ДНК за счет второго адаптера.

Смена ориентации ДНК-нитей относительно потока раствора в проточной кювете позволяет осуществить обратное считывание молекул с противоположного конца, что дает вторую последовательность после первоначального чтения. Учитывая приблизительную известную длину входных молекул, наличие парных концевых (PE) чтений обеспечивает

более обширную информацию. Это повышает точность при сопоставлении данных о прочтениях с эталонными последовательностями, особенно с последовательностями-повторами. Это улучшает точность при сопоставлении прочтений с эталонными последовательностями, особенно при обработке повторяющихся ДНК-последовательностей. Также это помогает распознавать сплайсированные экзоны и структурные изменения в ДНК, а также «слитые гены» (gene fusion). Вскоре после появления стандартной модели геномного анализатора (GAIIx) последовали секвенаторы HiSeq, которые обеспечивали еще более длинные и более глубокие считывания, а затем MiSeq, который хоть и был менее производительным, но представлял собой более доступный вариант секвенатора с возможностью более длинных чтений [156,157].

Многие компании, занимающиеся секвенированием, использовали свои собственные новые методики и оказывали различное влияние как на возможности проведения экспериментов, так и на рынок в целом.

В период становления второго поколения технологии секвенирования, кроме уже упомянутых секвенаторов 454 и Solexa/Illumina [158], альтернативным вариантом была система SOLiD (sequencing by oligonucleotide ligation and detection) от компании Applied Biosystems (позднее известной как Life Technologies после слияния с Invitrogen) [158]. В отличие от процесса синтеза, используемого в технологии Illumina, SOLiD производит секвенирование путем лигирования с помощью ДНК-лигазы, основываясь на принципах, заложенных в ранних методах, таких как «полони-секвенирование» [159]. Несмотря на то, что платформа SOLiD не способна предоставить такую же длину и глубину чтения, как Illumina [160], что делает процесс сборки более сложным, она все равно остается конкурентоспособной с точки зрения стоимости на один прочтенный нуклеотид [161].

Другим значительным технологическим достижением, использующим принцип лигирования, является технология Complete Genomic «DNA nanoballs». Этот процесс генерирует клонированную популяцию молекул ДНК, используя механизм «катящегося круга» (rolling circle) для создания длинных цепочек ДНК, которые образуют наношарики, прикрепленные к основанию стекла для последующего секвенирования [162]. Ещё одним примером технологий секвенирования второго поколения является разработанная Джонатаном Ротбургом платформа, которая появилась после его ухода из компании 454. Ion Torrent, также представляющий продукт компании Life Technologies, является первой технологией "постсветового секвенирования" (post-light sequencing), так как она не использует ни флуоресценцию, ни люминесценцию [163]. Сходно с методом 454, бусины, содержащие клональные популяции фрагментов ДНК, полученные с помощью emPCR, промываются на планшете. Однако, в отличие от метода 454, встраивание нуклеотидов

измеряется не через высвобождение пирофосфата, а через разницу в pH, вызванную высвобождением протонов (ионов H+) во время полимеризации. Это стало возможным благодаря применению технологии комплементарных металл-оксид-полупроводников, используемой в производстве микропроцессорных чипов [163]. Эта технология значительно быстрее проводит непосредственно секвенирование на этапе фактического обнаружения [161], хотя, также как и 454 (и все другие технологии пиросеквенирования), имеет проблему с интерпретацией гомополимерных последовательностей [164].

Долгое время не было единого консенсуса в отношении разделения технологий секвенирования на второе и третье [165–168]. Сейчас в третье поколение принято относить технологии «секвенирования одной молекулы» (Single-Molecule Sequencing; SMS).

Первая технология SMS была разработана в лаборатории Стивена Квейка [169,170]. В дальнейшем технология была коммерциализирована компанией Helicos BioSciences и работала, за исключением «мостиковой-амплификации», аналогично технологии Illumina – закрепленные последовательности ДНК-затравок с молекулами dNTP [171] последовательно промывались растворами каждого из четырех оснований и получали набор изображений. Хотя эта технология была относительно медленной и дорогой (и давала относительно короткие чтения), она стала первой технологией, позволяющей секвенировать неамплифицированную ДНК, что позволило избежать всех связанных с этим смещений и ошибок [165]. В начале 2012 года Helicos подала заявление о банкротстве [172] и эстафету третьего поколения подхватили другие компании.

Одной из заметных SMS-платформ реального времени (SMRT) являлись приборы PacBio от Pacific Biosciences [173]. В процессе SMRT ДНК-полимераза функционирует в специальных ячейках, которые представляют собой миниатюрные отверстия в металлической пленке, покрывающей чип. Эти ячейки используют свойства света, проходящего через отверстия меньшего диаметра чем определенная длина волны, что вызывает его затухание, освещая только дно лунок и позволяя визуализировать индивидуальные молекулы флуорофора. В каждую ячейку помещаются отдельные молекулы ДНК-полимеразы. Затем фиксированная молекула ДНК промывается dNTP, а процесс удлинения цепей молекулы ДНК отслеживается в реальном времени с помощью четкого разделения детектируемой флуоресценции [174]. Этот процесс обеспечивает возможность секвенирования индивидуальных молекул за кратчайший промежуток времени. Также приборы PacBio способны производить невероятно длинные чтения, до и более 10 т.п.н. в длину, что существенно упрощает множество исследований, например, сборку генома *de novo [173]*.

В настоящее время одной из самых многообещающих технологий секвенирования ДНК третьего поколения является секвенирование с использованием нанопор. Эта технология представляет собой отрасль более широкого применения нанопор для обнаружения и количественного анализа различных биологических и химических молекул [175].

Потенциал данного подхода был отмечен еще до развития секвенирования второго поколения, когда исследователи показали, что одноцепочечная РНК или ДНК могут проходить через липидный двойной слой через большие ионные каналы α-гемолизина с помощью электрофореза. Кроме того, прохождение через канал нуклеиновой кислоты приводит к блокировке ионного потока, что временно уменьшает концентрацию ионов H+ и прямо пропорционально зависит от длины биополимера [176].

Также возможно использование небиологических твердотельных технологий для создания соответствующих нанопор, что также способствует возможности последовательного чтения двухцепочечных молекул ДНК [177,178]. Компания Oxford Nanopore Technologies (ONT), первая на рынке предложившая нанопоровые секвенаторы, вызвала значительный интерес к своим нанопоровым платформам GridION и MinION [179,180]. MinION, в частности, представляет собой компактное USB-устройство размером с мобильный телефон, которое было выпущено впервые для конечных пользователей в рамках пробного раннего доступ в 2014 году [181].

Наблюдаемое в первое время низкое качество секвенирования компенсировалось ожиданием, что такие секвенаторы представляют собой действительно революционную технологию в данной области, производя невероятно длинные последовательности ДНК гораздо дешевле и быстрее, чем это было возможно ранее [182–185]. Применительно к секвенированию геномов, сопоставимых с человеческим, в начале MinION использовался для создания скаффолда, применяемого в паре с данными Illumina [186,187], сочетая сверхдлинное чтение технологии нанопор с высокой глубиной и точностью чтения, обеспечиваемой секвенированием короткими прочтениями. А уже сегодня, благодаря технологиям секвенирования третьего поколения, стало возможным добиться решения фундаментальных задач геномики [188,189].

Успех с секвенированием геномов целых организмов, развитие технологий секвенирования до последнего десятилетия XX века, а также положительный опыт широкого применения картирования генов привели к тому, что в 1990 году стало возможным запустить проект «Геном человека». Впервые идею проекта "Геном человека" (ПГЧ) публично представил Ренато Дульбекко в статье, опубликованной в 1984 году. В своем исследовании он отметил, что знание последовательности генома человека

значительно упростит понимание рака [190]. В мае 1985 года состоялась встреча, полностью посвященная перспективам и сложностям реализации проекта "Геном человека", на которой Роберт Синшаймер, ректор Калифорнийского университета в Санта-Круз (UCSC), собрал 12 экспертов [191]. По результатам совещания был сделан вывод о том, что технически проект является осуществимым, несмотря на свою сложность. В первые годы публичного обсуждения ПГЧ (с середины и до конца 1980-х) подавляющее большинство биологов было настроено против, что совпадало с официальной позицией Национального института здоровья (NIH). Министерство энергетики США (DOE) изначально настаивало на ПГЧ, используя аргумент, что знание последовательности генома поможет нам понять влияние радиации на геном человека в результате воздействия атомных бомб и других аспектов передачи энергии. Поддержка министерства энергетики имела решающее значение для принятия ПГЧ. Интересным наблюдением является то, что Конгресс США поддержал ПГЧ больше, чем большинство биологов. Члены Конгресса понимали привлекательность международной конкурентоспособности в области биологии и медицины, потенциал для промышленных побочных продуктов и экономических выгод, а также потенциал для более эффективных подходов к борьбе с болезнями. В отчете комитета Национальной академии наук проект был одобрен в 1988 г., а мнение научного сообщества по поводу целесообразности проекта поменялось [192]. В 1990 г. программа была начата.

Программа суммарной стоимостью 3 миллиарда долларов существенно развивалась по мере совершенствования технологий геномики. Первоначально ПГЧ намеревался определить генетическую карту человека, затем физическую карту генома человека и, наконец, собрать полноценный геном человека [193]. ПГЧ финансировался со стороны государства и общественности с помощью Национального Института Здравоохранения США и британского фонда Wellcome Trust соответственно. Руководителем проекта был назначен Джеймс Уотсон. Позднее в 1993 году его должность занял Френсис Коллинз. На старте ПГЧ в начале 1990-х существовал оптимизм в отношении того, что преобладавшая тогда технология секвенирования будет заменена. Исходный подход секвенирования методом Сегнера считался слишком громоздким и низкопроизводительным для эффективного прочтения генома. Как оказалось, исходная референсная последовательность генома человека была расшифрована с использованием 96-капиллярной версией технологии первого поколения. Были предприняты попытки использовать альтернативные подходы, такие как мультиплексирование и секвенирование путем гибридизации, но они не получили эффективного распространения [194,195].

Дополнительным стимулом оригинального ПГЧ стало, параллельное с общественной и государственной инициативой, создание аналогичного коммерческого ПГЧ под руководством Крейга Вентера и его компании Celera Corporation. Задачей коммерческого ПГЧ было использование наработок оригинального проекта с целью обогнать его и запатентовать результаты ради частных интересов компании. Ситуация с наличием сразу двух проектов привела к конфликту интересов и к противостоянию идей, технологий и методов. В отличии от секвенирования «первого поколения», использовавшегося в оригинальном проекте, частный проект в своей основе использовал «метод дробовика» [196]. Метод включал в себя фрагментацию больших кусков ДНК и прочтение их небольших фрагментов. В свою очередь, для расшифровки последовательности ДНК использовались бактериальные искусственные хромосомы (BAC).

1.3. Проект «Геном человека»

В 2000 году первые результаты двух независимых проектов были опубликованы одновременно в научных журналах «Science» и «Nature» [143,197]. ПГЧ произвел точную референсную последовательность для каждой хромосомы человека с небольшим количеством пробелов и исключением больших гетерохроматиновых областей [198]. Первые опубликованные данные ПГЧ инициировали каталогизацию частей большинства генов человека [143,197]. На основании этой информации было определено большинство белков человека, наряду с другими важными элементами, такими как некодирующие регуляторные РНК. Понимание всего многообразия открывшейся для научного сообщества информации потребовало очередной смены представления о науке [199]. В этот момент зарождается системная биология, которая изменит подход к биологии и медицине [200,201]. Важно отметить, что ПГЧ популяризировал идею немедленного предоставления общественности данных в удобных для пользователя базах, таких как GenBank и UCSC Genome Browser [202,203]. По сегодняшний день GenBank расширяет и уточняет свои данные о нуклеотидных и белковых последовательностях, а также добавляет данные об их функциональной и структурной аннотации.

На текущий момент ПГЧ близок к абсолютному завершению. Примерно десятая часть генома человека оставалась неизученной, когда в 2019 году был основан консорциум «От теломеры к теломере» (T2T). Теперь это число сократилось до нуля. В препринте, опубликованном в мае 2021 года, консорциум сообщил о первой полной

последовательности генома человека (T2T-CHM13), добавив почти 200 миллионов новых пар оснований к широко используемой референсной последовательности генома человека, известной как GRCh38, тем самым завершая ПГЧ [204].

Впервые выпущенный в 2013 году, GRCh38, заменив на своем посту GRCh37, стал ценным инструментом в медицинской генетике. Во многом из-за того, что использовалась технология секвенирования второго поколения, на тот момент в нем присутствовало множество «незакрытых» участков. Длины прочтений было недостаточно для однозначного картирования геномных последовательностей-повторов, включая теломеры, которые закрывают концы хромосом, и центромеры, которые координируют разделение вновь реплицирующейся ДНК во время деления клетки.

Технологии секвенирования третьего поколения PacBio и Oxford Nanopore с длинными прочтениями оказались решающим фактором. К тому времени, когда в 2020 году команда T2T реконструировала [188,189] первые отдельные хромосомы (X и 8) - секвенирование Pacific Biosciences продвинулось настолько, что ученые T2T смогли обнаружить крошечные вариации в длинных участках повторяющихся последовательностей. Эти тонкие различия сделали возможным разрешить длинные повторяющиеся сегменты хромосом, и остальная часть генома быстро встала на свои места.

ПГЧ, стал первым примером «большой науки» в биологии, и он прямо продемонстрировал актуальность этого подхода для решения ее интегрированных биологических и технологических задач. ПГЧ характеризовался четким набором амбициозных целей и планов их достижения, ограниченным числом финансируемых исследователей, обычно организованных вокруг центров или консорциумов; обязательством по предоставлению общедоступных данных/ресурсов; и потребностью в значительном финансировании для поддержки инфраструктуры проекта и разработки новых технологий. Большая и малая прикладная наука, ориентированная на отдельных исследователей, прекрасно дополняют друг друга, поскольку первая создает ресурсы, которые являются основополагающими для всех исследователей, а вторая добавляет подробные экспериментальные разъяснения к конкретным вопросам, а также аналитическую глубину и детализацию к данным, полученным благодаря большим исследованиям.

В 2003 году силами NIH запущен проект ENCODE (Энциклопедия элементов ДНК), который направлен на обнаружение и понимание функциональных частей генома [205]. Используя несколько подходов, многие из которых основаны на секвенировании второго поколения, проектный консорциум ENCODE получил объемные и ценные данные,

связанные с регуляторными сетями, которые управляют экспрессией генов [206]. «Большие данные», созданные ENCODE, поднимают вопросы относительно функциональности генома, например, как отличить настоящий биологический сигнал от статистического шума [207,208].

ПГЧ также привел к появлению протеомики, дисциплины, ориентированной на идентификацию и количественную оценку белков, присутствующих в дискретных биологических компартментах, таких как клеточная органелла, орган или кровь. Белки — независимо от того, действуют ли они как сигнальные устройства, молекулярные машины или структурные компоненты — определяют специфичную для клеток функциональность. ПГЧ упростил внедрение масс-спектрометрии, предоставляя референсные транслированные последовательности белков, а следовательно, и предсказанные массы множества пептидов в протеоме человека [209].

После завершения проекта по геномной характеристике (ПГЧ) было составлено более 4 000 готовых или высококачественных черновых геномных последовательностей, как прокариотов, так и эукариотов, что значительно обогатило наше понимание процессов эволюции [210]. Разнообразие этих геномов предоставляет уникальное представление о том, как разные организмы связаны на генеалогическом древе жизни, ясно подтверждая, что все современные виды произошли от общего предка [211]. Стало доступным решение таких фундаментальных вопросов, как происхождение новых генов, роль высококонсервативных участков ДНК у ортологичных видов, определение степени сложности генной организации, устойчивость или чувствительность областей генома к мутациям и реорганизации, эволюция регуляторных сетей и изменение паттерна экспрессии генов [212]. Последний вопрос представляет особый интерес сейчас, когда геномы приматов и гоминидов секвенированы или секвенируются в надежде пролить свет на эволюцию человека [213,214]. Так последовательность генома неандертальца оказала огромное влияние на понимание эволюции человека [214–216].

ПГЧ стимулировал интеграцию известных и разработку новых сложных вычислительных и математических подходов к большим данным и объединил специалистов из информационной отрасли, математиков, инженеров и физиков-теоретиков с биологами, потворствуя междисциплинарной культуре развития науки [217]. Более того, ПГЧ также продвигал идею распространения общедоступного исходного кода, для программ прикладной медико-биологической направленности с целью модификации кода для нужд всего сообщества [218,219]. Так операционные системы на базе ядра Linux с открытым исходным кодом стали основой для создания большей части имеющегося

программного обеспечения. Исключением из этого правила служат программы для протеомики, где подавляющее число программного обеспечения было создано на операционной системе Windows. Доступность данных стала важной концепцией для научного сообщества в целом, потому что «демократизация данных» имеет решающее значение для привлечения новых людей в отрасль, а также для стимулирования роста новых исследований [220]. Критически важным это стало для медицины [221].

1.4. Развитие популяционной генетики

1.4.1. Локусы генетической вариации

С завершением ПГЧ в 2000 году и развитием технологий секвенирования, стало известно, что для любой пары людей примерно 99.9 % участков в геноме идентичны. В рамках Международного консорциума HapMap, для нескольких сотен людей было проведено ресеквенирование десяти сегментов по 500 т.п.о. [222]. Было выяснено, что в неконсервативной части генома, на уровне отдельных людей существует множество одиночных нуклеотидных отличий от референсного генома, в дальнейшем известные, как однонуклеотидные полиморфизмы (SNP), которые в разной степени уникальны между индивидуумами [223].

С момента публикации оригинального ПГЧ технологии полногеномного секвенирования (WGS) значительно улучшились и стали значительно дешевле, что теперь позволяет идентифицировать абсолютное большинство всех генетических вариаций, характерных для конкретного индивидуума. Полученные после секвенирования прочтения выравниваются, т.е. сопоставляются с референсным геномом, а отличия от референса идентифицируются в процессе, известном как картирование вариаций (variant calling). За последнее время методы точного картирования вариаций значительно улучшились, в том числе методы для поиска SNP, коротких вставок или делеций (indels), а также более крупных генетических изменений, включая изменение числа копий вариантов и структурные перестройки (SV). В рамках диссертации в первую очередь учитывать точечные вариации и небольшие делеции или инсерции.

В рамках проекта «1000 геномов», путем секвенирования 2 504 человек, был составлен каталог ДНК-вариаций, охватывающий широкий частотный спектр аллелей. Это позволило создать карту генетической изменчивости ДНК-вариаций от очень распространенных CV, до чрезвычайно редких, в том числе *de novo* RV, то есть новых вариаций, возникающих у потомства и отсутствующих у родителей. Учитывая, что размер генома человека составляет приблизительно 3.2×10^9 пар оснований, частота ошибок

механизма репликации ДНК приводит к появлению приблизительно 50 точечных мутаций *de novo* и меньшего числа более крупных мутаций при каждой репликации генома [224–227]. Таким образом, каждое потомство отличается от родителей примерно 50 новыми SNP. Частота ошибок репликации не одинакова для всего генома человека и зависит от сложности генома, причем некоторые участки более подвержены мутациям [227–229].

Благодаря подсчету *de novo* вариаций и филогенетическому сравнению современного человека с приматами, ближайшими предками, стало возможно оценить частоту мутаций у человека (μ), которая составляет примерно 10^{-8} на нуклеотид в поколение [230,231]. В среднем на человека приходится примерно 65 SNV, 5 инделов и 0.35 SV *de novo* ДНК-вариантов [232,233]. В соответствии с нейтральной теорией эволюции, большинство ДНК-вариантов являются некодирующими и безвредными для развития и жизни человека полиморфизмами. Таким образом, такие генетические вариации могут передаваться следующим поколениям, и их частота может стохастически повышаться или понижаться в течение нескольких поколений, данный процесс известен как генетический дрейф. Вероятность того, что *de novo* мутация закрепится в популяции, обратно пропорционален эффективному размеру популяции, т.е. количеству особей, оставляющих потомство [234].

Помимо генетического дрейфа еще одним фактором, влияющим на распространенность ДНК-варианта является естественный отбор. Большая часть генетических вариаций нейтральна, в крайне редких случаях *de novo* мутации могут быть полезными в популяции и быть положительно отобранными, увеличивая частоту, давая преимущество в выживании. И наоборот, некоторые мутации могут привести к тяжелому заболеванию или даже к гибели эмбриона и, таким образом они являются летальными и могут не передаваться будущим поколениям. Между летальными и нейтральными ДНК-вариантами существуют те, которые оказывают негативное влияние на приспособленность организма, тем самым снижая количество потомства, которое произведут носители соответствующей ДНК-вариации. Подобные эффекты приспособленности приводят к снижению частоты мутации посредством механизма, называемого негативным отбором. Так из-за естественного отбора CV, как правило, менее вредны, чем RV.

Существует множество локусов, которые являются общими для разных народностей. Геном человека содержит около 4 миллионов ДНК-вариантов, популяционные частоты которых варьируются в зависимости от этнического происхождения. Очевидным выводом было то, что как минимум какая-то часть вариации

геномов должна объяснять различия как внутри, так и между человеческими популяциями. Рекомбинация дополнительно участвует в формировании генетической изменчивости, влияя на то, как наследуются близлежащие ДНК-вариации. Потомство обычно полностью наследует одну хромосому от матери и одну хромосому от отца. Передающиеся материнские и отцовские хромосомы претерпевают мейоз, в результате которого происходит один кроссинговер, чтобы физически поменять местами хромосомные сегменты, что привело к новой мозаике вариаций вдоль полной хромосомы, которая ранее не наблюдалась. Сегменты, которые были напрямую скопированы от каждого родителя, называются гаплотипами. Длины гаплотипов экспоненциально уменьшаются в зависимости от поколений из-за рекомбинации [222]. Интуитивно люди наследуют уменьшающееся количество вариаций от своих родителей к своим бабушкам и дедушкам и их более далеким предкам. Таким образом, ДНК-варианты, находящиеся далеко друг от друга на одной и той же хромосоме, как правило, не наследуются вместе, в то время как ДНК-варианты, расположенные близко друг к другу, более коррелированы, что приводит к блочной структуре гаплотипов. Подобное состояние, известное как неравновесие по сцеплению (LD), обычно измеряется как квадрат корреляции между генотипами пары SNP. События рекомбинации обычно происходят вблизи горячих точек: областей генома, которые с гораздо большей вероятностью подвергаются кроссинговеру. Эти горячие точки, как правило, являются общими для всей популяции и регулируются геном *PRDM9*.

Демографическая история популяций также влияет на популяционные частоты аллелей. Люди произошли из Африки, и та часть людей, которая мигрировала на Ближний Восток, а затем в другие части мира, унесла с собой определенную долю генетического разнообразия изначальной человеческой популяции. Эта концепция называется последовательным эффектом основателя [235]. На протяжении всей истории человечества имели место события как расширяющие популяцию людей, так и уменьшающие ее, которые, соответственно, увеличивали или уменьшали степень генетического разнообразия [236,237]. Подобные события также напрямую влияют на конечную аллельную частоту. Например, эффективный размер популяции влияет на силу естественного отбора [238]. Так была подробно изучена, с использованием ряда моделей, концепция мутационной нагрузки, которая описывает «бремя» (степень подверженности) вредных ДНК-вариантов, переносимых популяцией [239–241].

В связи с тем, что та часть изменчивости популяционных частот аллелей, которая может отвечать за проявление наблюдаемого признака, перемешана с демографическими особенностями популяций, то необходимо проводить корректировку данных. Первым

шагом такой корректировки является применение анализа главных компонент (PCA). PCA, применяемый к многомерным данным генотипов, позволяет кластеризовать генетические данные по степени различия их популяционной структуры [242,243].

Поскольку близлежащие области генома, как правило, наследуются вместе, образуя гаплотипную структуру, и поскольку было секвенировано большое количество людей в масштабах всего сообщества, то нет необходимости в глубоком секвенировании всего генома для оценки состояния генотипа во всех ДНК-вариантах. Развитие технологий микрочипов сделало возможным генотипирование или определение генотипического состояния (референсная гомозигота, гетерозигота или альтернативная гомозигота) известного местоположения в геноме от сотен тысяч до миллионов генетических ДНК-вариаций в геноме человека. Структура гаплотипов позволила развить технологию импутации, т.е. процесса заполнения генотипов, что позволило повысить рентабельность многих исследований. Однако сама технология генотипирования подразумевает, что существует ограниченное число известных локусов и, следовательно, вектор исследований смещен в сторону более изученных популяций.

1.4.2. Ассоциативные исследования

Основой человеческой генетики является определение степени, в которой генетические факторы по сравнению с факторами окружающей среды вносят вклад в фенотип, то есть в его наследуемость [244]. После вывода о доли наследуемости признака типичным следующим шагом является определение генетических вариаций, ответственных за конкретное фенотипическое проявление.

В любом эксперименте, включающим в себя ассоциативное исследование, обязательно присутствует, предшествующая основному этапу анализа ассоциаций, стадия контроля качества данных. Технические артефакты и ошибки могут возникать на каждом этапе, что обязывает проводить подробный анализ основных характеристик данных генотипирования или секвенирования. Существует ряд показателей, которые можно рассчитать, как индикаторы качества генетических данных, в том числе глубина прочтения и качество генотипа, баланс аллелей, коэффициент транзиций/трансверсий, количество известных и *de novo* ДНК-вариаций, коэффициент гетерозиготности, число одиночных полиморфизмов и инделов на один образец. Можно рассчитать дополнительные меры, такие как меры контроля качества на каждый ДНК-вариант или образец, включая показатели долю однозначно идентифицированных генотипов, показатели качества

картирования, оценки гаплотипов и многие другие. Успешное проведение этапа контроля качества носит определяющее значение для непосредственного анализа ассоциаций.

Исходно в ассоциативных исследованиях не определяют конкретный ген–мишень или механизм проявления фенотипа. Это отличается от обнаружения менделевских мутаций в семейных исследованиях, где ДНК-вариант, целевой ген и механизм проявления (конкретное изменение в белке) идентифицируются одновременно. Следует отметить, что, хотя величина эффекта отдельных полиморфизмов невелика в популяции, величина их влияния на молекулярные фенотипы может быть значительной.

Естественный отбор снижает частоту особенно патогенных мутаций, изменяя спектр вариаций, связанных с болезнью. Как правило, статистические ассоциативные тесты выполняются для каждого локуса в отдельности. Поскольку ДНК-вариаций в геноме сцеплены, многие из них, не являются независимыми. Стандартное количество независимых локусов при генотипировании равняется одному миллиону. Исходя из этого числа, считается, что результат тестов ассоциации ДНК-вариантов должен превосходить порог значимости в $0.05 \times 5 \times 10^{-6}$.

Со времени первого полногеномного ассоциативного исследования (GWAS) для возрастной дегенерации желтого пятна (AMD) в 2005 году, более 50000 ассоциаций полногеномной значимости ($p.value < 5 \times 10^{-8}$) были зарегистрированы между ДНК-вариантами и распространенными заболеваниями и признаками [245]. Эти ассоциации привели к пониманию архитектуры восприимчивости к заболеваниям (путем идентификации новых генов и механизмов, вызывающих заболевание), а также к достижениям в клинической помощи (например, выявлению новых мишеней для лекарств и биомаркеров заболеваний) и персонализированной медицине (например, прогнозирование риска заболевания и корректировка терапии на основе известного генотипа), например, как в описанном ранее случае со статинами [246].

Все виды анализов ассоциаций накладывают особые требования к размеру выборки. Если в ассоциативных исследованиях CV (CVAS) возникают сложности поиска сигнала из-за малого размера эффекта, то для ассоциативных исследований RV (RVAS), напротив, характерна проблема спорадической нерегулярной встречаемости ДНК-вариантов с сильным эффектом. При этом полиморфизмы с аллельной частотой более 0.01 суммарно, как правило, объясняют большую долю наследуемой изменчивости в популяции, за некоторыми исключениями [247]. Например, генетический компонент, лежащий в основе риска заболевания диабета II типа для подавляющего большинства населения, обычно определяется CV. [248] Тем не менее, у некоторых людей RV вызывают подтипы или другие типы заболевания, такие как диабет зрелого типа у молодых (MODY

диабет). В частности, в CVAS предполагается, что распространенные в популяции полиморфизмы мало влияют на риск заболевания, поэтому для надежного выявления связи генотип-фенотип необходимы особенно большие размеры выборок. К сожалению, большое количество частых вариантов имеет очень малый эффект на фенотип, требующий значительного размера выборки для обнаружения значимой ассоциации. Таким образом, известные к настоящему моменту ассоциированные варианты вносят небольшой вклад в объяснение фенотипической дисперсии. Однако, пример ассоциативного исследования такого сложного признака, как человеческий рост с достаточно большим размером выборки позволяет обнаружить подавляющее большинство полиморфизмов с отличным от статистического шума эффектом. Yengo и соавторы показали, что CV в определении роста объясняют порядка 40-50% дисперсии. [249] Несмотря на небольшой размер эффекта, для CV, терапевтическое воздействие препаратов, нацеленных на гены со значительной ассоциацией, не обязательно должно быть небольшим. Например, ингибиторы *PCSK9* могут резко снижать уровень холестерина ЛПНП (липопротеинов низкой плотности) в семьях с семейной гиперхолестеринемией, даже несмотря на то, что SNP в *PCSK9* не так сильно влияют на метаболизм ЛПНП [250].

В свою очередь RV могут быть агрегированы с использованием различных методов ассоциации, но для достижения статистической мощности требуются очень большие размеры выборки. Одним из первых открытий GWAS-исследований было то, что для большинства признаков даже самые важные локусы в геноме имеют небольшие размеры эффекта, подтвердив тем самым модель Фишера, и что совокупность значимых ассоциаций объясняет лишь скромную часть предсказанной генетической изменчивости, что первое время было известно как «загадка отсутствующей наследственности» [251]. С тех пор данный вопрос был в значительной степени раскрыт благодаря анализу, показывающему, что распространенные однонуклеотидные полиморфизмы (SNP) с размерами эффекта значительно ниже статистической значимости объясняют большую часть «отсутствующей наследуемости» многих признаков [252].

Как и предсказывалось ранее, большие размеры экспериментальной выборки приводят к новым открытиям, и именно это произошло за последнее время. Несколько лет назад это было показано на примере шизофрении и роста. Так в 2009 г. был обнаружен первый геномный локус, надежно связанный с предрасположенностью к шизофрении, 11 SNP в выборке из 3 000 [253] образцов; к 2014 г. это число увеличилось до 108 при размере выборки в 35 000 [254] образцов. Точно так же, было подчеркнуто, что в 2008 г. для исследования человеческого роста было идентифицировано только 40 значимых для всего генома SNP, и вместе они объясняли около 5 % [251] наследуемости, а в 2014 г.

число ассоциированных SNP увеличилось до ~700, что объясняет более 20 % [255] наследуемости.

Для полигенных признаков, в которых нет целевых генов, которые объясняли бы весомую часть наследуемости было показано, что можно использовать данные GWAS для создания генетических предикторов заболеваний и других сложных признаков путем оценки размера эффекта в нескольких локусах и расчета так называемого полигенного риска (PRS). Подобный анализ можно проводить также вне клинических исследований, рассчитывая PRS для здорового человека. Полигенные прогнозы не особенно информативны для индивидуума, но они объясняют достаточную долю изменчивости (от 1 % до 15 % в настоящее время для высокополигенных признаков без основного гена) для отдельных групп, например выборки с самым высоким и самым низким риском.

RV с большим эффектом также могут вносить вклад в генетическую изменчивость [256], особенно при заболеваниях с серьезными последствиями для функции приспособленности, таких как, например, аутизм [257,258].

Ассоциативные исследования в своем современном виде используют множество известных популяционных величин в том числе LD и корреляционную структуру, которая существует между ДНК-вариантами в геноме человека в результате эволюционных характеристик, в частности, размера популяции, мутаций, скорости рекомбинации и естественного отбора. Статистическая мощность для обнаружения ассоциаций между ДНК-вариацией и признаком зависит от размера экспериментальной выборки, распределения размеров эффектов причинных ДНК-вариантов, которые распространены в популяции, частоты этих ДНК-вариаций и LD между наблюдаемыми генотипами.

С целью более тонкой настройки анализа ассоциаций можно использовать дополнительную информацию о ДНК-вариантах, которую предоставляют различные биоинформатические инструменты. Так, стоит обратить внимание на природу эффекта, например, можно взять в анализ только конкретное подмножество ДНК-вариаций – синонимичные, миссенс ДНК-варианты или ДНК-варианты нарушающих целостность белка [259–263]. Многие такие инструменты также предоставляют информацию о гене, на который оказывает воздействие конкретный ДНК-вариант, информацию о популяционной аллельной частоте ДНК-варианта, обеспечивают различные численные оценки функционального воздействия на целевой белок, эволюционную консервативность биологической последовательности и многое другое [264]. Также была создана серия специализированных тестов, цель которых заключается в уточнении набора генов [265] направленных на выявление биологических путей и механизмов, связанных с сложными

признаками. Эта информация может быть использована для изучения ассоциаций и объяснения результатов генетических исследований.

Таким образом, успех ассоциативного исследования, проводимого для лечения определенного заболевания или выявления связи с конкретным признаком, зависит от распространенности в популяции локусов, влияющих на данный признак, его разнообразности и генетической структуры, а также от мощности экспериментальных данных и используемой для анализа дополнительной информации. Гетерогенность относится как к биологии признака, так и к способности точно диагностировать или измерить его. Если генетическая архитектура определенного признака или заболевания известна, то можно спланировать оптимальные эксперименты для обнаружения конкретных ДНК-вариантов, что сильно уменьшит требуемые затраты на эксперимент. Помимо проведения самого GWAS для увеличения уверенности в результатах следует проводить репликации в независимых когортах, а также функционально подтверждать новые данные *in vivo*.

Еще большее распространение GWAS получил благодаря разработке относительно недорогих SNP-таргетных панелей. Обычно такие SNP-панели различаются по своему содержанию, но в целом они содержат около миллиона ДНК-вариантов. В настоящее время опубликованы результаты GWAS для тысяч сложных признаков в широком диапазоне областей, включая распространенные заболевания, количественные признаки, которые являются факторами риска заболевания, геномные показатели, такие как экспрессия генов и метилирование ДНК, а также социальные и поведенческие черты, например, такие как финансовое благополучие и уровень образования. При условии достаточных размеров выборки ассоциации GWAS доказали высокую воспроизводимость как внутри, так и между популяциями.

Новым этапом развития ассоциативных исследований стало создание множества крупных популяционных биобанков с широким доступом для множества исследователей. Биобанки содержат данные тысяч генотипированных людей, которые подробно фенотипируются с помощью методов анкетирования, лабораторных измерений и/или с помощью централизованного связывания с электронными медицинскими записями. Ярким примером является Биобанк Великобритании, который включает данные о примерно 500 000 человек [266] и уже позволил провести GWAS для множества количественных признаков, включая антропометрические признаки [267], признаки клеток крови [268], метаболиты [269], когнитивные признаки [270], и психические симптомы [271], а также увеличить размер выборки для ассоциативных исследований многих других заболеваний [272–274].

Хотя биобанки исторически были сосредоточены на популяциях с европейским происхождением, в настоящее время создаются крупные биобанки данных лиц с неевропейским происхождением [275]. В большинстве биобанков для распространенных ДНК-вариантов используются вмененные данные о генотипах, хотя данные WES уже доступны для 50 000 участников Биобанка Великобритании [276]. В ближайшие несколько лет данные WES и WGS будут получены для всех участников Биобанка Великобритании, что значительно увеличит возможности оценки роли RV.

1.4.3. Менделевские ДНК-варианты

Наиболее изученными для медицинской генетики являются ДНК-варианты напрямую взаимосвязанные с физиологическими процессами, которые нарушаются при болезненных состояниях. В то время как подавляющее большинство ДНК-вариантов имеют ограниченную способность оказывать влияние на такие процессы, влияние некоторых из них настолько велико, что их простое присутствие в отдельно взятом геноме может надежно предсказать наблюдаемое фенотипическое проявление. Такие ДНК-варианты называются менделевскими ДНК-вариантами, а соответствующие болезни — менделевскими болезнями. Полноэкзомное (whole-exome sequencing; WES) и полногеномное (whole-genome sequencing; WGS) секвенирование сыграло важную роль в выявлении причинных ДНК-вариантов у пациентов с менделевскими заболеваниями [277]. Поскольку каждый человек несет в себе порядка 12 000 – 14 000 ДНК-вариантов [278], прямо изменяющих генетический продукт, отличить ДНК-варианты, вызывающие заболевания от доброкачественных («нейтральных») является, пожалуй, главной задачей современной клинической генетики. Как правило, низкая популяционная частота ДНК-варианта в референсных базах данных или его отсутствие в них признаются необходимым, но недостаточным критерием патогенности ДНК-варианта [279,280].

Исходя из сильного влияния отдельно взятого менделевского ДНК-варианта частота их встречаемости в популяции ожидаемо небольшая, но из-за непосредственной прямой связи между менделевскими ДНК-вариантами и болезненными состояниями такие ДНК-варианты гораздо легче обнаружить, чем ДНК-варианты, которые просто изменяют риск иногда в бесконечно малой степени. Четкая наблюдаемость этой связи также играет важную роль в принятии ключевых медицинских и репродуктивных решений. Менделевские гены являются непосредственной мишенью для таргетной терапии, и многие из самых продаваемых лекарств от распространенных заболеваний нацелены на белковые продукты таких генов [248]. Непереносимость статинов является

распространенной проблемой среди пациентов, получающих статины. У таких пациентов возникают осложнения, связанные с мышцами. Так в исследовании 2018 года было выяснено, что для некоторых пациентов возможно снизить дозировку статинов при соблюдении определенной диеты в случаях их непереносимости, что уменьшает токсический эффект терапии [249]. Было показано, что ингибиторы *PCSK9* приводят к снижению уровня холестерина до 70 %, в связи с чем были одобрены для пациентов с первичной гиперхолестеринемией, для пациентов с непереносимостью статинов и для пациентов с атеросклеротическими сердечно-сосудистыми заболеваниями [231]. Другим недавним примером служат доклинические исследования, где монотерапия энкорафенибом при метастатической меланоме с мутацией *BRAF-V600* показала хороший клинический профиль без значительной токсичности [250].

1.4.4. Генетика сложных признаков

Как было сказано во введении, несмотря на успех инфинитезимальной модели Фишера 1918 года в описании наследования, особенно при селекции растений и животных, на протяжении всего 20-го века было неясно, насколько большой разброс генов на самом деле будет важен для управления сложными признаками. Генетики ожидали, что у человека сложные признаки будут определяться лишь ограниченным числом локусов с небольшим, но наблюдаемым эффектом, что привело к большому количеству картографических исследований, которые в ретроспективе оказались недостаточного качества и приводили к сомнительным выводам [281]. Знаний и данных на тот момент примитивного секвенирования было недостаточно для понимания этиологических и патогенетических процессов при сложных заболеваниях.

Современная генетическая эра предоставила возможность проверить классические теории и создать на их основе новые. Обилие новых методов и данных пересмотрело наше понимание многих компонентов, определяющих структуру генома – полигенность, естественный отбор, распределение мутационных эффектов, плейотропия, – но при этом в этой сфере все еще не хватает моделей способных согласовать все эти элементы [282].

Еще одним важным открытием стало то, что, в отличие от менделевских болезней, которые в значительной степени вызваны изменениями в кодировании белков [77], сложные признаки в основном обусловлены некодирующими ДНК-вариантами, которые предположительно влияют на регуляцию генов, что на первый взгляд сбивает с толку [283–285]. Действительно, многие исследования показали, что значимые ДНК-варианты

сильно обогащены в областях активного хроматина, таких как промоторы и энхансеры, в соответствующих типах клеток. Например, ДНК-варианты риска аутоиммунных заболеваний обнаруживают особое обогащение активных областей хроматина иммунных клеток [286–288]. Исходя из этого можно сделать вывод, что сложное заболевание обусловлено накоплением слабых эффектов на ключевые гены и регуляторные пути, которые определяют риск заболевания [289].

Ожидается, что естественный отбор будет поддерживать популяцию вблизи оптимального значения количественных признаков и снижать распространенность потенциально дезадаптивных фенотипов, например, заболеваний. Такие оптимумы и дезадаптивные фенотипы определяются в контексте влияния окружающей среды [290]. Отбор обычно действует путем снижения частоты фенотипически значимых аллелей, хотя он может стимулировать увеличение частоты аллелей, когда значение признака выходит за рамки оптимальных показателей. Эта базовая логика привела к вопросу о том, можно ли в принципе наблюдать эффект естественного отбора в данных GWAS. Недавние исследования пришли к твердому консенсусу, что размеры фенотипических эффектов отрицательно коррелируют с частотой аллелей [291–294]. Эти результаты не согласуются с классическими «чисто нейтральными» моделями, но различные модели естественного отбора, влияющие на вариативность признаков, остаются правдоподобными. На сегодняшний день существует неразрешенный вопрос – насколько отбор является стабилизирующим и имеет ли выраженную направленность. В итоге, в основе современного представления о фенотипической изменчивости лежит наблюдаемое огромное количество мутаций со слабым эффектом в сочетании с неэффективным отбором против таких аллелей.

По мере того, как секвенирование нового поколения сделало доступными большие когортные исследования стало возможным изучать полигенные болезни, которые предъявляют дополнительные условия к размеру выборки и покрытию данных. Стало очевидным, что лишь данные секвенирования недостаточны для полного объяснения связи между генотипом и фенотипом при многофакторных заболеваниях. Это характерное свойство многофакторных медицинских признаков характеризуются сложной этиологией, где генетические факторы и воздействие окружающей среды вносят переменный вклад.

Считается, что сети регуляции генов достаточно взаимосвязаны, из чего следует, что все гены, экспрессируемые в клетках, связанных с заболеванием, могут влиять на функции критически важных для дезадаптивного фенотипа генов. Следовательно большая часть наследуемости может быть объяснена эффектами на гены вне основных метаболических и регуляторных путей.

Эмпирические результаты GWAS по распределению размеров эффектов и частот аллелей все еще ставят вопрос о том, какие классические и новые модели из теоретической популяционной генетики способны наилучшим образом объяснить появляющиеся наблюдения. Существующие теории представляют разнообразные модели воздействия отбора: от тех, которые предполагают прямое воздействие на основной признак, до моделей, в которых отбор на генетическую изменчивость возникает в результате одновременного влияния на другие признаки (плейотропия). И даже до теорий о полностью «мнимом» (apparent) отборе, когда основной признак, как предполагается, не подвергается никаким селективным ограничениям. [295].

Моделирование и измерение плейотропии является ключом к эмпирическим вопросам о том, находится ли дезадаптивный признак под значимым прямым отбором и как коэффициенты отбора зависят от фенотипических эффектов отдельных ДНК-вариантов. Современные оценки полигенности косвенно свидетельствуют о высокой плейотропности генетической архитектуры для большинства сложных признаков [292,296]. Действительно, было подсчитано, что 2 % генетических вариаций вовлечены в рост и аналогичная доля (1 %) вовлечена в риск развития диабета 2 типа [294]. Сравнительная и функциональная геномика обычно оценивают долю функционально значимой части генома в 0.1 [297,298]. Если 10 % генома имеет какое-либо функциональное значение, и мутации, влияющие на признаки, происходят именно из нее, то очевидно, что геном не может содержать независимые локусы количественных признаков (QTL; Quantitative trait locus) для абсолютного большинства сложных признаков даже с минимальной оценкой полигенности [299].

Изучение ДНК-вариантов в геноме человека, ассоциированных сразу с несколькими признаками привлекает значительное внимание в области изучения плейотропии [300,301]. Несколько попыток проанализировать плейотропные эффекты отдельных ДНК-вариантов или генов было предпринято в масштабе изучения всего фенома человека (phenome-wide association studies, PheWAS) [302]. Хотя до сих пор не было предложено единой методологии для интерпретации феномных ассоциаций. Для выявления плейотропии в данных GWAS также были предложены подходы, основанные на менделевской рандомизации. Глобальная картина плейотропии у человека, однако, оказалась неоднозначной из-за расхождений в опубликованных исследованиях, с наблюдаемыми ошибками при сборе данных о фенотипах и генотипах. Данные UK Biobank представляют собой исключительную возможность для систематического анализа общих составляющих генетической структуры единообразно обработанных фенотипов

человека в рамках всего генома. Первые глобальные обзоры свидетельствуют о широком распространении сигналов плейотропии в данных UK Biobank [299,303].

1.4.5. Популяционная стратификация в данных генотипирования

Генетические и фенотипические наблюдения, используемые в ассоциативных исследованиях, часто получают из популяционной когорты, где предполагается, что индивидуумы являются случайной выборкой из популяции. Межпопуляционные ассоциативные исследования основаны на предположении, что ДНК-варианты с высокой ассоциацией с большей вероятностью будут общими для разных популяций, в то время как отличия в LD-паттернах обеспечивают более сильное различие между причинными и непричинными ДНК-вариантами.

Межпопуляционная изменчивость возникает из-за нарушения принципа случайного скрещивания. Если имеющиеся в данных паттерны коррелируют с факторами окружающей среды, они могут привести к ложным ассоциациям и смещенным оценкам размера эффекта в GWAS [304,305]. Такие подходы, как геномный контроль [304], метод главных компонент (PCA) [243], линейные смешанные модели (LMM) [306–308] и LD-регрессия (LDSR) [309] были разработаны для обнаружения и нормализации данных с учетом этой стратификации. Однако эти подходы не обязательно должны избавляться от нее целиком, особенно при проведении мета-анализа из нескольких исследований [310,311]. В крупных GWAS с относительно однородными популяциями, таких как, например, Британский биобанк (UKB) [266], должны устраняться многие из этих опасений. Однако такие проекты по-прежнему продолжают демонстрировать локальную структуру популяций [312–316]. Степень, в которой популяционная стратификация влияет на результат GWAS на практике, в значительной степени неизвестна, и успешность применения корректирующих методов сильно зависит от конкретных данных. Данное наблюдение становится все более важным в свете растущего внимания к полигенным показателям для прогнозирования риска заболевания [317,318]. Полигенные оценки многих антропоморфных и поведенческих признаков, даже после строгой поправки на популяционную стратификацию, демонстрируют географическую кластеризацию в пределах Великобритании [315,319]. Хотя некоторые из этих вариаций могут быть связаны с недавними моделями миграции [319], они также могут отражать остаточную стратификацию в оценках размера эффекта.

Фенотипы, соответствующие непрерывным или бинарным переменным, проверяются на ассоциацию с ДНК-вариантами. Распространенным дизайном

ассоциативных исследований является исследование типа «случай-контроль» (case-control), в котором «случаи» и «контроль» определяются на основе наличия или отсутствия определенного фенотипа, соответственно. Успешное применение подхода межпопуляционных ассоциативных исследований зависит от двух факторов. Во-первых, мощность данных для конкретного эксперимента должна быть обеспечена достаточным размером выборки для каждой народности. Во-вторых, методология проводимого ассоциативного исследования должна соответствовать генетической архитектуре разных популяций, лежащей в ее основе.

Для преодоления возникающих из-за популяционной структуры ошибок широко используются классические методы, такие как анализа главных компонент (PCA; principal component analysis) и линейные смешанные модели (LMM; linear mixed models) [320]. PCA является статистическим методом, направленным на выявление линейных осей наибольшей изменчивости в данных. Его главные компоненты часто отражают географическую удаленность между предками популяций [243]. Преимущество LMM по сравнению с PCA заключается в возможности одновременной коррекции популяционной структуры, семейной структуры и скрытых генеалогических связей [307,320]. Для подбора линейных смешанных моделей, применимых к количественному, непрерывному фенотипу, был разработан ряд вычислительно эффективных методов, включая EMMAX [307], Fast-LLM, [321,322] и GEMMA [323].

Несмотря на то, что методы PCA и смешанные линейные модели успешно корректируют стратификацию популяции для распространенных ДНК-вариантов, эффективность этих методов для RV до сих пор остается недостаточно изученной. PCA и смешанные модели предполагают плавное распределение частот минорных аллелей (MAF; minor allele frequency) по географическому (или родовому) пространству. Поскольку RV зачастую отражают относительно недавние эволюционные изменения, они локализованы внутри современных популяций. В связи с этим PCA и LMM могут оказаться не способными скорректировать стратификацию популяции, если распределение риска заболевания также резко специфично для определенной этнической группы [324]. Известно, что Fast-LLM-Select может снизить частоту ошибок первого типа, но при этом данный подход также склонен отрицательно влиять на итоговую мощность анализа, если причинные RV неравномерно распределены между популяциями [325,326]. Уже опубликовано несколько исследований на тему использования PCA для коррекции стратификации популяции в тестах на ассоциации RV [327–329]. Эффективность PCA сильно зависит от базового распределения риска развития фенотипа и структуры популяции, и могут быть полезны альтернативные стратегии использования PCA в

качестве ковариат для непосредственного ассоциативного теста или же, например, для использование PCA для помощи в подборе «случаев» и «контролей» [330]. Сейчас доподлинно известно, что проведение PCA с использованием только RV не является эффективной стратегией для контроля стратификации популяции. Наилучшей практикой стало совместное использование всех доступных для анализа ДНК-вариантов [327,328]. Если имеется достаточно большой пул контрольных индивидуумов, то для учитывания расслоения популяции можно использовать оценочные показатели частот предковых аллелей [330].

С точки зрения практики дизайн первых чипов генотипирования для GWAS был в значительной степени смещен в сторону CV, присутствующих в европейских популяциях, что делало исследование других популяций затруднительным. Обычно в исследованиях случай-контроль когорты случаев и контролей отбираются без учета стратификации по народностям, из-за чего аллельные частоты не соответствуют средним популяционным показателям. Разрешение подобного дисбаланса в данных необходимо отражать в статистическом анализе, например, стоит уделить дополнительное внимание корректировке ковариат [331,332]. Использование «контролей» из когорты населения с неизвестным статусом заболевания позволяет учесть наличие «случаев» с популяционной частотой в «контрольной» популяции, хотя это не будет иметь значительного эффекта для заболеваний с популяционной частотой менее 1 %. В качестве альтернативы группы «контролей» могут быть напрямую сопоставлены с группами «случаев» по этническому происхождению. При низких частотах заболевания такой подход обладает достаточной мощностью [333]. В некоторых современных больших таргетных панелях генотипирования эта проблема в значительной степени была решена, хотя вместо этого присутствовали другие трудности, такие как отсутствие высококачественных неевропейских референсных панелей импутации и менее точная информация о частотах популяционных аллелей.

Если случаи и контроли не генотипируются или секвенируются вместе в едином дизайне эксперимента, необходимо приложить дополнительные усилия на этапе контроля качества для нормализации данных. Следует отметить, что хотя предполагается, что образцы являются случайной выборкой из популяции, это предположение не соответствует действительности при наличии предвзятости участия и несопоставимых социально-демографических факторов [334,335].

1.4.6. Методы оценки полигенных эффектов

Как было упомянуто ранее, понимание генетической основы сложных признаков является давней проблемой в области генетики. Хотя очевидно, что многочисленные генетические варианты вносят свой вклад в наследственность этих признаков, количественная оценка коллективного эффекта этих вариантов представляет собой сложную задачу. Для оценки полигенных эффектов генетических вариантов широко используются три метода: оценка полигенного риска (PRS), геномный анализ сложных признаков (GCTA) и регрессия по шкале неравновесия связей (LDSR), упомянутая ранее.

Оценка полигенного риска (PRS; Polygenic Risk Score) – это статистический метод, используемый для прогнозирования генетического риска развития сложных признаков или заболеваний у человека. В основе PRS лежит предположение о том, что множество генетических вариантов, каждый из которых оказывает незначительное влияние, в совокупности вносят вклад в фенотипическую вариативность, наблюдаемую в популяции. Вычисление PRS включает два основных этапа: этап обнаружения и этап подтверждения. На этапе обнаружения проводятся ассоциативные исследования с целью выявления генетических вариантов, ассоциированных с интересующим признаком. Эти варианты, а также размеры их эффектов и частоты аллелей используются для расчета PRS на этапе валидации. PRS – это, по сути, взвешенная сумма аллелей риска, которые несет индивидуум, причем каждый аллель взвешивается по величине его эффекта, оцененного на этапе обнаружения. Из преимуществ PRS стоит выделить простоту и интерпретируемость метода, а также возможность персонализированной оценки риска. [336] Однако PRS имеет и ряд недостатков, таких как необходимость использования данных значительного размера для увеличения точности оценки риска, т.к. интерпретирование PRS изначально исходит из предположения, что все значимые генетические варианты были учтены на этапе обнаружения. [337]

Геномный анализ сложных признаков (GCTA; Genome-wide Complex Trait Analysis) представляет собой методологию и одноименный инструмент, позволяющие оценить долю дисперсии фенотипа, которая объясняется аддитивными генетическими эффектами, без явного использования генетических вариантов. Для проведения оценки генетической схожести между индивидуумами и фенотипической дисперсии, которая объясняется совокупными генетическими эффектами, GCTA использует однонуклеотидные полиморфизмы (SNP), полученные из большой выборки несвязанных особей. GCTA особенно ценна в тех случаях, когда отсутствуют крупномасштабные данные GWAS по интересующему признаку, или при изучении признаков со сложной генетической структурой. К преимуществам GCTA можно отнести возможность оценки наследственности непосредственно по геномным данным и применимость к широкому

кругу признаков, даже к тем, для которых отсутствуют подробные данные о генетических ассоциациях. Однако GCTA не дает информации о конкретных генетических вариантах, способствующих формированию признака, и может оказаться не совсем подходящим для признаков, на которые оказывают влияние RV. [338]

LDSR (LDSR; Linkage Disequilibrium Score Regression) – это метод, использующий корреляционную структуру ДНК-вариантов для оценки различных генетических показателей, в т.ч. наследственности сложных признаков. В основе LDSR лежит идея о том, что генетические варианты, ассоциированные с признаком, скорее всего наследуются совместно с близлежащими вариантами и информация о неравновесном сцеплении может быть использована для оценки общей полигенной наследственности. В LDSR суммарная статистика ассоциативных исследований используется для расчета полигенной оценки для каждого индивидуума в наборе данных. Затем полигенные оценки вместе с информацией о неравновесии связей из референсной панели регрессируют на суммарную статистику, что позволяет оценить наследственность. К преимуществам LDSR относится возможность оценки наследственности только на основе сводной статистики GWAS, что позволяет обойтись без персональной генотипической информации на индивидуальном уровне. LDSR эффективен с точки зрения вычислений и успешно используется для оценки наследуемости широкого круга признаков. Однако, как и GCTA, LDSR не позволяет выявить конкретные варианты, вносящие вклад в формирование признака, и может быть ограничена в улавливании эффектов редких вариантов. [339]

1.4.7. Менделевская рандомизация и генетические корреляции

Менделевская рандомизация (MR; mendelian ranomization) – метод, используемый для вывода причинно-следственных связей между фактором риска и интересующим фенотипом. MR использует принципы менделевской генетики, где генетические варианты, выступающие в качестве инструментальных переменных, используются для имитации влияния фактора риска на фенотип. Предполагается, что эти инструментальные переменные распределяются случайным образом в процессе формирования гамет и связаны с фенотипом только через их влияние на интересующее воздействие. [340,341]

Ключевым допущением в MR является то, что генетические варианты, используемые в качестве инструментальных переменных, тесно связаны с интересующим воздействием и не зависят от сопутствующих ковариирующих факторов, которые в противном случае могли бы исказить оценки, полученные в ходе наблюдений. Использование генетических вариантов в качестве инструментальных переменных дает

MR ряд преимуществ по сравнению с традиционными обсервационными исследованиями и рандомизированными контролируемыми испытаниями (РКИ). Во-первых, MR позволяет получить более надежные доказательства причинно-следственных связей, поскольку генетические варианты фиксируются в момент зачатия и, следовательно, не подвержены обратной причинно-следственной связи, которая часто характерна для обсервационных исследований. Во-вторых, MR позволяет изучать долгосрочные воздействия, которыми нецелесообразно или неэтично манипулировать в РКИ. В-третьих, MR позволяет определить приоритетность потенциальных терапевтических мишеней, оценив, приведет ли изменение степени воздействия к изменению результата. [340,341]

Генетическая корреляция – это понятие, тесно связанное с MR и направленное на оценку степени общности генетической архитектуры между двумя или более признаками или заболеваниями. Она определяет степень влияния одних и тех же генетических вариантов сразу на несколько фенотипов, т.е. плейотропию. Коэффициент генетической корреляции варьирует от -1 до 1, где отрицательные значения указывают на противоположные генетические влияния, нулевые - на отсутствие общих генетических эффектов, а положительные - на общую генетическую архитектуру. [342]

Используя данные крупномасштабных ассоциативных исследований, можно оценить генетические корреляции между широким спектром сложных признаков, начиная от распространенных заболеваний и заканчивая поведенческими чертами и физиологическими изменениями. Понимание генетических корреляций имеет важное значение по нескольким причинам. Во-первых, оно может раскрыть новые биологические аспекты, выявив общие пути или механизмы между, казалось бы, не связанными между собой признаками. Во-вторых, это может помочь в расчете PRS, когда информация, полученная по одному признаку, может быть использована для прогнозирования риска другого фенотипа. В-третьих, генетические корреляции могут помочь в выявлении потенциальных плейотропных эффектов генетических вариантов, когда варианты влияют на несколько признаков, что позволяет лучше понять биологию, лежащую в их основе. [342]

Существует несколько методов расчета генетических корреляций между сложными признаками. Некоторые из широко используемых подходов это: Bivariate Genetic Analysis, LDSR и GREML.

Двумерный генетический анализ (Bivariate Genetic Analysis) оценивает генетическую дисперсию и ковариацию между двумя признаками с помощью данных близнецовых или семейных исследованиях. [343]

LDSR (LD Score Regression) – популярный метод оценки генетических корреляций с использованием сводной статистики геномных исследований ассоциаций (GWAS). LDSR был описан в предыдущем разделе как метод для решения задачи оценки полигенных эффектов сложных признаков. В контексте решения задачи определения генетических корреляций LDSR использует информацию о неравновесии связей между SNPs для оценки генетической ковариации между признаками. [344]

GREML (Genomic RElatedness-based Mixed Model): GREML – метод, способный оценивать генетическую ковариацию между различными признаками, используя информацию о геномных полиморфизмах однонуклеотидных позиций (SNP), полученных от неродственных индивидуумов. [345]

1.4.8. Ограничения ассоциативных исследований

Одной из самых больших проблем, стоящих перед интерпретацией результатов GWAS, является разделение сигнала от причинных ДНК-вариантов и других ДНК-вариантов, которые коррелируют с ними. Коррелированные ДНК-варианты могут демонстрировать статистически значимую связь, но не дают представления о биологии, лежащей в основе состояния или признака. Методы статистического точного картирования определяют вероятные причинно-следственные ДНК-варианты путем оценки вероятности того, что каждый ДНК-вариант в коррелированном наборе является причинным по отношению к другим в наборе. Байесовский подход был введен для обработки простейшего случая единственного исхода болезни против единственного причинного ДНК-варианта (то есть один ДНК-вариант имеет истинное биологическое действие, а ассоциативные сигналы во всех других ДНК-вариантах обусловлены исключительно их корреляцией с причинным ДНК-вариантом). Поскольку стало очевидным, что этот простой случай часто не соответствует действительности, эта схема была расширена, чтобы позволить одновременное тестирование нескольких заболеваний и нескольких независимых причинных ДНК-вариантов в одном и том же локусе. [346]

Например, во многих крупнейших мета-анализах сложных признаков и заболеваний ни один человек не имеет доступа к данным на уровне выборки для всех когорт. По этой причине в последнее время было разработано несколько методов, которые требуют только сводной статистики (например, величины эффекта и «p-значения» для каждого однонуклеотидного ДНК-варианта) в качестве входных данных. Для сигналов только с одним причинным ДНК-вариантом эти подходы должны давать идентичные результаты, но точное отображение нескольких причинных ДНК-вариантов из сводной

статистики добавляет две сложности. Во-первых, исчерпывающий условный анализ с заданным максимальным числом предполагаемых причинных сигналов требует значительных вычислительных ресурсов. Более новые методы, такие как FINEMAP, выполняют стохастический поиск, значительно сокращая время поиска и, таким образом, позволяя проверить, обусловлен ли сигнал множественными причинно-следственными ДНК-вариантами [347]. Во-вторых, для этого псевдоусловного анализа требуется LD-матрица попарных взаимодействий ДНК-вариантов либо из исходного GWAS, либо из справочной панели, такой как проект 1 000 геномов. Важно, чтобы эта LD-матрица соответствовала этногенетическому происхождению населения когорты GWAS и была достаточно большой, чтобы обеспечить достаточную точность в оценках LD. Так было показано, что LD-матрица, полученная на 1 000 человек, релевантна для когорты до 10 000 человек, но плохо подходит для GWAS из 50 000 человек (что обеспечивает очень хорошую точность при расчете LD из 10 000 человек) [348].

Получение соответствующих LD-матриц для менее изученных популяций может быть осложнено их меньшей представленностью. Эта проблема актуальна для редковариантного картирования. В принципе, из-за низкой LD с соседними ДНК-вариантами, RV должны быть легче картированы, но эти ДНК-варианты могут не быть включены в эталонную популяцию. Отсутствие информации о LD для конкретного RV сделает невозможным его точную картографию. Эта проблема побуждает исследователей делиться информацией о LD наряду со сводной статистикой проводимых ими ассоциативных исследований. Исследователи предложили ряд инструментов для «эффективной оценки, хранения и беспрепятственного обмена информацией о LD», чтобы упростить этот процесс [349,350].

Обычно ассоциативные исследования проводятся с учетом наличия определенной структуры популяций в данных, концентрируясь либо на выборках сходных по генетическому происхождению или включая отдельные характеристики выборки в качестве ковариат. Однако по мере того, как когорты с мультипопуляционным происхождением были генотипированы или секвенированы, стало возможным использовать знание о структуре популяций для проведения мета-анализа, который повышает статистическую мощность и помогает точнее определять вариации, влияющие на проявление фенотипа.

1.4.9. Стратегия выбора платформы для ассоциативных исследований

Самым очевидным, на первый взгляд, решением является секвенирование полного генома. Действительно, WGS уже имеет очевидные преимущества в диагностике пациентов с редкими менделевскими нарушениями [351,352]. Так, канадский консорциум Care4Rare неоднократно доказывал, что раннее применение WGS может сократить дорогостоящий и психологически обременительный диагностирующий процесс, который вынуждены проходить более 50 % пациентов и семей с неизученными или атипичными редкими заболеваниями [352]. Кроме того, секвенирование всего генома помогает генотипировать соматически приобретенные ДНК-варианты в опухолевых тканях; это изменяет подход к терапии онкологии от лечения, основанного на морфологических особенностях локализации и/или гистологии, к лечению, учитывающему молекулярно-генетические особенности опухоли. Однако основным ограничением такого подхода все еще остается высокая стоимость. Следствием из этого является то, что WGS не может широко применяться в исследовании широких популяционных когорт с пациентами и для рутинной клинической практики во многих странах. В свою очередь это вызывает рост популярности альтернативных платформ, которые остаются актуальными только из-за своей значительно более низкой стоимости.

WGS с низкой глубиной секвенирования предлагает экономически эффективную альтернативу полноценному WGS [353,354]. Принцип полногеномного секвенирования с низким покрытием основан на использовании информации о структуре популяции, т.е. LD и карту гаплотипов.

С практической точки зрения, экономическая целесообразность данного подхода проявляется в возможности проведения секвенирования семи-восьми индивидуумов с глубиной прочтения 4X вместо традиционного секвенирования одного индивидуума со средней глубиной прочтения 30X при одном и том же бюджете. Проект «1 000 геномов» иллюстрировал, что широкомасштабное секвенирование всего генома с низкой степенью покрытия может быть использовано для обнаружения и генотипирования клинически значимых вариантов [355]. Ожидалось, что в сравнении с полноценным WGS, полногеномное секвенирование с низким покрытием приводит к более высокому уровню ошибок генотипирования, что, в свою очередь, приводит к снижению мощности. Однако, первые исследования показали, что полногеномное секвенирование с малой глубиной для большей выборки приводит к большей статистической мощности, чем при использовании полноценного WGS на меньшем количестве образцов. Так, например, было продемонстрировано, что данные с ДНК-вариантами с MAF>0.002 при секвенировании 3 000 образцов со средним покрытием 4X имеют мощность, аналогичную мощности полноценного WGS 2 000 человек при 30X в тестах ассоциации для одного ДНК-варианта

[353]. В настоящее время эмпирические исследования подтверждают эти выводы, построенные с использованием моделирования [356].

Если не экономить на покрытии ДНК-варианта, как в случае WGS с низким покрытием, то более практичным подходом на сегодняшний день является секвенирование всех кодирующих и фланкирующих регионов (WES), которое охватывает от 1 до 6 % области всего генома, а стоимость коммерческого использования может составлять всего 400 долларов США [357]. Обычно WES нацелено на консенсусную кодирующую последовательность (проект CCDS) [358], которая составляет порядка 30 миллионов оснований, но точные регионы, на которые нацелено секвенирование, различаются у разных поставщиков услуг. Для клинических целей считается достаточной целевое покрытие (глубина) секвенирования равной 100X для приборов Illumina. На практике секвенирование экзома обычно проводится с средней глубиной покрытия 60-80X, что позволяет в целевых регионах достичь высокой вероятности покрытия >20X для 80 %-90 % регионов, кодирующих белки [359]. Поскольку технология обогащения целевых регионов несовершенна, секвенирование экзома производит некоторое количество чтений в нецелевых регионах. Эти «внецелевые» (off-target) чтения могут быть полезны для проверки качества последовательности и определения структуры популяции [330,359–362]. Секвенирование всего экзома привлекательно для клинического применения, главным образом потому, что оно охватывает все белок кодирующие области генома для определения вариаций в областях экзонов вызывающих заболевание мутаций [363–367]. Впервые этот метод проявил свою эффективность в определении участков генома, связанных с повышенным риском развития рака предстательной железы (8q24 и 17q21) у мужчин, происходящих из африканских популяций [368–370]. Эти открытия способствовали объяснению увеличения вероятности заболевания раком простаты на 50 процентов [371].

Если в первую очередь интересуют CV, которые, как было установлено, связаны со сложными неменделевскими заболеваниями, то секвенирование таргетных панелей регионов представляет собой экономически эффективный подход для дальнейшего исследования наиболее приоритетных геномных регионов. Ситуация такова, что во многих ассоциативных исследованиях с помощью таргетных панелей не удается выявить RV, связанные с фенотипом [372,373]. Но это правило не абсолютно и из него есть ряд исключений, которые демонстрируют потенциал для выявления причинных RV. Например, в одном исследовании провели секвенирование 56 генов-кандидатов и обнаружили несколько RV, связанных с болезнью Крона, включая ДНК-варианты сплайсинга с

протективным эффектом в *CARD9* [7]. В другом примере, аналогичным образом обнаружили большое количество RV среди лиц с гипертриглицеридемией [374].

Для анализа, наоборот, RV, технологии генотипирования могут подойти лучше, чем технология секвенирования. Хотя современные генотипирующие чипы не могут предоставить для анализа достаточное число ячеек, которое было бы способно охватить большинство RV в популяции, они все еще представляют собой экономически эффективную альтернативу секвенированию таргетных регионов. Были разработаны микрочипы для генотипирования, такие как Metabochip [375] для изучения метаболических и сердечно-сосудистых заболеваний, а также Immunochip [376] для исследования аутоиммунных и воспалительных заболеваний. Эти чипы основаны на высокоприоритетных вариантах ДНК, выявленных в результате GWAS. Они включают как частые ДНК-варианты, отобранные для воспроизведения первоначальных сигналов GWAS, так и наборы CV и RV для более детального анализа нескольких сотен регионов, связанных с соответствующими фенотипами, выявленными в ходе GWAS. Дополнительно стоит обозначить чипы Illumina и Affymetrix, сделанные на базе экзомов, став недорогой альтернативой для экзомного секвенирования, о котором подробно пойдет речь в следующем разделе [359]. Экзомные чипы были разработаны на основе 12 000 секвенированных экзомов (в основном европейского происхождения). В их состав входит порядка около 250 000 целевых несинонимичных ДНК-вариантов, около 12 000 целевых сплайсинговых ДНК-вариантов и более 7 000 целевых ДНК-вариантов со «stop gained» эффектом, а также несколько дополнительных категорий ДНК-вариантов, включая маркеры предковых ДНК-вариантов, SNP-сетку для интерполяции, митохондриальные SNP и SNP с информацией об HLA типе человека [377]. Из-за своей относительно невысокой стоимости экзомный чип позволяет проводить исследования на широкой выборке людей, что существенно увеличивает статистическую мощность для ДНК-вариантов, присутствующих на чипе. Примером результатов применения таких чипов может служить исследование METSIM, в котором участвовало приблизительно 8 000 финских индивидуумов. В рамках этого исследования были обнаружены связи между низкочастотными ДНК-вариантами в генах *SGSM2* и *MADD* (MIM#231680) с процессом обработки и выделения инсулина [378].

При значительном превышении количества образцов, доступных для секвенирования или генотипирования, по сравнению с исследовательским бюджетом, увеличение силы ассоциации может быть достигнуто путем предпочтительного отбора лиц для секвенирования, которые с наибольшей вероятностью будут информативными. Один из таких подходов включает отбор лиц с экстремальными фенотипами в надежде

выявить причинно-следственные RV. [379–383]. При изучении количественных признаков можно отобрать лиц с экстремальными значениями признака после поправки на известные ковариаты. В качестве альтернативы, в исследованиях, ориентированных на заболевание, выбор лиц с экстремальными фенотипами часто может осуществляться на основе известных факторов риска. Например, к таким факторам риска могут быть отнесены индекс массы тела, семейный анамнез сахарного диабета 2 типа, ожирение, продолжительный период курение. Для количественных характеристик требуемый размер выборки для сбора экстремального фенотипа может быть существенно меньше, чем для случайной выборки. Например, при отборе образцов из верхних и нижних 10 % «хвостов распределения» конкретного фенотипа количество объектов, необходимых для секвенирования с целью достижения заданной мощности, часто может быть уменьшено более чем в 2 раза [381,382]. Простой подход к анализу данных предполагает рассмотрение экстремальных фенотипов как бинарных величин. В качестве альтернативы, экстремальные фенотипы могут быть описаны с использованием усеченного нормального распределения [381,382]. Такой подход позволяет добиться большей мощности при том же числе образцов, но может быть чувствительным к предположению о нормальности распределения лежащего в основе непрерывного признака.

1.4.10. Практические аспекты работы с данными экзомного секвенирования

Произошло значительное увеличение производства новых данных WES в масштабе всего населения. Многие причинные ДНК-варианты для менделевских расстройств были выявлены с помощью секвенирования экзома. Яркими примерами являются *DHODH* для синдрома Миллера [384], *MLL2* для синдрома Кабуки [385], другими менделевскими фенотипами [277,384], а также некоторыми спорадическими формами расстройств [386,387]. В настоящее время все большее число исследований направлено на использование секвенирования экзома для выявления генов и ДНК-вариантов, связанных со сложными заболеваниями.

С 2011 года WES регулярно внедряется в качестве инструмента для молекулярной диагностики в лабораториях клинической генетики [388,389]. В последующие годы WES стало ключевым компонентом в рамках проектов, таких как «1000 геномов» [355], исследования экзомов в проекте NHLBI «Grand Opportunity» (GO-ESP) [390,391] и в базе данных ExAC [278] для обширного каталогизирования популяционных вариантов ДНК и обнаружения заболеваний, связанных с редкими вариантами. Консорциум T2D-GENES секвенировал экзомы около 10 000 индивидов из пяти семейных групп с целью

обнаружения генетических ДНК-вариантов, связанных с *T2D* и связанными с метаболическими фенотипами. В рамках проекта UK10K были секвенированы экзомы более 6 000 человек с неврологическими расстройствами, ожирением или одним из редких заболеваний с целью выявления генетической основы заболеваний. С помощью экзомного секвенирования были выявлены некоторые RV предрасположенности к заболеваниям. Например, было продемонстрировано, что RV в *PLD3 [372]* связаны с поздней стадией болезни Альцгеймера (LOAD) путем секвенирования 14 больших семей, страдающих LOAD. В другом исследовании были выявлены ассоциации между RV в гене *PNPLA5* и холестерином липопротеинов низкой плотности путем секвенирования 2 005 экзомов [392].

Геномные данные, в том числе данные WES очевидным образом относятся к так называемым «большим данным». Ожидается, что в 2025 году геномика превзойдет трех крупнейших игроков в области больших данных: Twitter, астрономию и YouTube [393]. Были определены ключевые технологии, необходимые для развития геномики с точки зрения работы с большими данными, в том числе сбора, хранения, распространения и анализа данных. В геномике данные обладают следующими параметрами, характерными для больших данных в порядке возрастания их важности [394]: объем, скорость, разнообразие и достоверность.

Объем данных от секвенирования экзома может изменяться в зависимости от количества образцов и покрытия каждого из них. Для выборки с 100X покрытием полноэкзомное секвенирование обычно генерирует около 5–6 ГБ данных, что является значительно более низким значением по сравнению с ~90 ГБ для полногеномного секвенирования с таким же покрытием. Однако, объем данных может значительно увеличиться при учете большого числа образцов. Например, анализ вариантов ДНК, полученных из данных WES в ExAC v0.3.1 от 60 706 человек, охватывал 540 ГБ [395]. В настоящее время многие исследования, включающие десятки тысяч образцов, используют WES для экономической эффективности, однако, генерация данных не является основной проблемой, а узким местом является обработка и анализ данных.

Скорость генерации и накопления данных WES также зависит от покрытия и охвата выборки, а также от используемого оборудования. Например, в 2013 году крупный центр по секвенированию, оснащенный примерно 50 системами Illumina HiSeq 2000 и 2500, мог секвенировать четыре экзома на каждый полный геном и достигал производительности около 2 000 экзомов в неделю [396]. К 2018 году система Illumina NovaSeq 6000 была способна секвенировать геном человека (с покрытием 30X, >120 Гб) каждые 55 минут и экзом (с покрытием 100X, ~8 ГБ) каждые ~5 минут [397]. Это предоставляет

пользователям возможность высокопроизводительного секвенирования до 48 геномов человека или почти 500 экзомов за цикл менее чем за 45 часов.

Разнообразие данных WES сильно варьируется от сложности исследования. Выражается это в наличии дополнительных факторов, учитываемых при сборе данных, например, пол, возраст или этническая принадлежность, разделение данных на «болезненные» и контрольные образцы, учтении тканевой специфичности данных, а также времени отбора пробы.

На достоверность данных секвенирования могут повлиять различные источники ошибок. К ним относятся неточности референсной последовательности, мозаицизм, нарушения в протоколе пробоподготовки, а также ошибки секвенирования. В связи с этим этап поиска ДНК-вариантов может в конечном итоге предсказать почти в семь раз больше ДНК-вариантов, чем возможно [398]. Трудно отличить RV от случайных ошибок, возникающих во время секвенирования [399]. Кроме того, основным недостатком WES является неравномерное покрытие прочтениями последовательностей экзомных мишеней, что способствует множеству низкопокрытых областей, которые влияют на последующий анализ и, таким образом, препятствуют точному определению ДНК-вариантов. Например, в выборке со средней глубиной секвенирования более 75X некоторые регионы все еще остаются недостаточно покрытыми из-за ограничений технологий (охват всего 10X) [400]. Это ведет к увеличению количества пропущенных вариантов ДНК и, следовательно, приводит к формированию пустых генотипов [401].

Нарушения стандартной структуры данных WES могу свидетельствовать о различных аномалиях и выброса, которые могут быть определены при использовании аналитических методов. Например, у пациента может быть полностью или частично пересортированный геном, из-за крупных структурных перестроек или образования новых аберрантных или трисомии функциональных хромосом, как это наблюдается у людей в одном из крайних проявлений расстройств аутичного спектра [402]. Несоответствия могут носить технический характер, например, несколько измерений данных могут быть результатом различных типов данных и источников, а также образовываться в ходе непостоянства скорости загрузки данных в репозиторий.

В 2017 году точность различных протоколов обработки данных секвенирования с финальным этапом в виде поиска ДНК-вариантов была оценена в ходе участия различных алгоритмов в конкурсе PrecisionFDA Hidden Treasures — Warm Up, проводимого американским управлением по контролю за продуктами и лекарствами (FDA) в целях стимулирования разработки более совершенных методов генетического скрининга [403]. Edico Genome's DRAGEN продемонстрировала наивысший общий результат, в то время

как Saphetor занял второе место. Помимо выбора протокола обработки данных WES, на надежность результатов поиска ДНК-вариантов могут влиять различные артефакты секвенирования, особенно при определении свойств, которые отличают ложноположительные ДНК-варианты от истинных ДНК-вариантов. Для преодоления указанной проблемы применялась стратегия трио-дизайна, включающая отца, мать и ребенка, с целью фильтрации артефактов секвенирования и сохранения истинных мутаций [404]. Что касается скорости обработки данных на этапе вызова вариантов, то, например, использование DeepVariant в комбинации с Google Cloud требует приблизительно 70 минут (время оценено без учета картирования исходных данных прочтений) для обработки всего генома при покрытии 30X и примерно 25 минут для аналогичного покрытия экзома [405].

Стоит упомянуть такую важную черту как уязвимость данных генотипирования. Любые данные генотипирования человека, в том числе и WES, потенциально могут выдать конфиденциальную информацию о пациенте путем их повторного использования. В нескольких исследованиях сообщалось об уязвимости геномных данных человека, что позволяет повторно идентифицировать пациентов из «анонимной базы данных» [406–408]. Исследователи продемонстрировали, что человека можно идентифицировать, повторно запрашивая наборы данных через базы с открытым доступом к аллелям. Подобная информация может быть использована, например, страховыми компаниями с целью профилирования людей на основе их медицинских данных. Кроме того, существуют опасения связанные с политикой и практикой возврата последовательностей генома участникам исследования и затратами на безопасность данного процесса [409,410].

Многие WES данные в относительно ближайшем будущем неизбежно будут считаться устаревшими или неактуальными. На сегодняшний день одним из основных мотивирующих факторов при выборе платформы секвенирования в пользу WES является его более доступная стоимость по сравнению с WGS. Однако анализ на основе данных WGS демонстрирует большую стабильность характеристик качества по сравнению с WES [411]. Кроме того, анализ WGS также является более комплексным и полезным в случае, когда мутации, повышающие риск развития заболевания, находятся вне охватываемых областей WES, как, например, при пороках развития конечностей из-за мутации в энхансере гена sonic hedgehog (SHH) [412]. Следовательно, с уменьшением стоимости WGS до уровня, сопоставимого или даже более низкого, чем текущая стоимость WES, целесообразность использования WES как метода сильно ограничится. В итоге, привлекательность WES для клинического применения имеет ограниченный срок

действия, поскольку WGS в конечном счете станет доступным для более широкого круга исследований.

Визуализация данных о геномных последовательностях представляет собой важный инструмент для исследователей и клиницистов, особенно для тех, кто не обладает высоким уровнем навыков в области информационных технологий. В настоящее время экзомные данные подвергаются визуализации с использованием разнообразных популярных геномных браузеров, которые обеспечивают отображение генно-ориентированных вариаций и транскриптов [395], предоставляя обширные возможности для сравнительного анализа и агрегации доступных знаний. Однако визуализация данных секвенирования нового поколения с использованием графических инструментов требует значительных затрат на обслуживание серверов, включая высокие вычислительные требования (потребление ресурсов процессора, оперативной памяти и дискового пространства) и стабильное интернет-соединение высокой скорости. Кроме того, многие онлайн-решения визуализации генома основаны на более старых версиях баз данных аннотаций по сравнению с локально установленными, что может представлять существенное ограничение. Своевременная актуализация базы данных также затратная задача. Сложность обработки данных становится все более проблематичной, приводя к дополнительным потребностям к ресурсам хранилища данных. Например, использование браузера 3D Genome Browser требует не менее 10 ГБ для сжатых данных и до 1 ТБ для несжатых. В свете этого, появляется растущий интерес к новым инструментам просмотра генома, основанным на облачных вычислениях. Такие инструменты становятся популярными благодаря их способности эффективно управлять ресурсами, обеспечивать высокую производительность и доступность для тех, кто нуждается в коммерческой лицензии, как, например, DNAnexus.

Данные WES могут быть получены с применением разнообразных технологических платформ. Первоначальные методы секвенирования, такие как секвенирование по Сэнгеру, опирались на дробление цепи нуклеиновых кислот и электрофоретическое разделение для обнаружения вновь интегрированных нуклеотидов. Эти методы, хоть и обладали высокой точностью, но сильно замедляли процесс исследования и значительно увеличивали затраты, и чаще всего применялись для подтверждения генетических изменений, найденных другими методами [125].

При проведении WES ключевым фактором является выбор специализированного набора реактивов для захвата экзомных участков ДНК, а не выбор платформы. Доступны различные коммерческие решения, такие как Agilent SureSelect XT, Agilent SureSelect QXT, NimbleGen SeqCap EZ и Illumina Nextera Rapid Capture Exome. Используются

биотинилированные ДНК или «РНК-приманки», которые гибридизуются с библиотеками геномных фрагментов. Однако они отличаются по выбору целевой области, длине пробника, плотности пробника, молекуле, используемой для захвата, и методу фрагментации генома. Можно выделить платформу NimbleGen, если целью является обнаружение SNV и инделов в нетранслируемых областях (UTR), в то же время Agilent XT и Illumina одинаково эффективно работают для обнаружения SNV и инделов в кодирующих областях [413].

Фенотипы медицинской значимости, полностью обусловленные генетическими факторами или имеющие значительный генетический компонент, возникают в результате различных изменений в структуре ДНК. Эти молекулярные события включают одиночные нуклеотидные полиморфизмы (SNP), если они проявляются в заметной частоте (> 1 %) в популяции, а также структурные изменения в ДНК, такие как вариации числа копий (CNV), короткие вставки и делеции, повторы, большие вставки и делеции, транслокации (которые могут вызвать так называемое "слияние генов"), инверсии, и мутации, влияющие на кариотип, например, анеуплоидия. [414]. Исходя из покрываемой методом области генома WES в основном используется для обнаружения SNV/SNP и вставок в кодирующих областях генома.

Множественное параллельное секвенирование коротких прочтений с помощью методов секвенирования следующего поколения генерирует большие данные, которые необходимо выровнять (сопоставить с референсным геномом или создать последовательность генома *de novo*) для анализа. Когда референсный геном доступен, первым шагом в анализе данных является сопоставление прочтений с ним [415]. Цель данного шага состоит в том, чтобы «сложить» каждое отдельное прочтение в единый консенсус, сопоставляемый с референсным геномом. В случае, когда референсный геном недоступен, приходится использовать *de novo* сборку, с последующей его разметкой и аннотацией.

Сборка *de novo* основана на предположении, что каждое чтение может перекрываться и может использоваться для создания сборки контигов (т.е. достаточно длинных прочтений дохромосомного порядка) [416,417], во многом это схоже с секвенированием методом дробовика [134,418]. После сборки контига его можно временно использовать как часть референсного генома, в котором определяются и аннотируются функциональные элементы. С другой стороны, аннотация генома может использоваться без непосредственной сборки [419,420] и представляет собой прямой анализ прочтений в два этапа. На первом этапе каждое прочтение аннотируется с использованием таких инструментов, как BLAST [421], функциональных аннотаций с

использованием таких инструментов, как InterProScan [422], или с помощью сравнения последовательностей с другими известными базами данных. Это иногда называют прочитанной аннотацией. Затем следует второй шаг, во время которого отдельные сопоставляются с некоторым референсным элементом; например, геном или специальными метаболитическими и регуляторными картами.

Broad Institute разработал набор инструментов для NGS анализа геномных данных или GATK [423], для анализа прочтений с возможностью объединения различных инструментов GATK в единый рабочий протокол обработки данных для улучшения документации и воспроизводимости результатов исследований. GATK и инструменты, входящие в его состав, имеют свободный доступ для некоммерческого использования. По мере добавления новых инструментов для выполнения GATK возможности рабочих процессов практически безграничны. Например, был объединен GATK и MuTect [424], еще один инструмент Broad Institute для анализа соматических мутаций.

Можно предвидеть, что комбинации эффективных инструментов могут дать результаты более высокого качества, чем результаты каждого инструмента в отдельности, что также демонстрирует преимущество новых разрабатываемых протоколов обработки геномных данных на базе объединения существующих инструментов. В ряду некоторых исследований [425–429] GATK использовался для анализа мутаций/SNP с использованием WES панелей. Например, в исследовании 2017 года ученые провели анализ WES данных у более чем 10 000 пациентов и проанализировали данные с помощью GATK для выявления RV наследственной меланомы. Из этого исследования был выявлен мутационный ландшафт меланомы кожи и глаз, а также определено участие *EBF3* в качестве потенциального гена, предрасполагающего к кожной меланоме.

После публикации инструментов GATK и MuTect появилось множество других программных средств, среди которых были такие, которые базировались на функционале GATK. Например, OTG-snpcaller [430] объединил TMAP и GATK для вызовов SNP. Этот инструмент был использован в анализе WES, что привело к обнаружению миссенс-мутации в гене *SCN8A*, отвечающего за восьмую альфа-субъединицу натриевого канала, при клинических проявлениях ранней детской эпилептической энцефалопатии 13 типа [431]. ASEQ [432] разработан для анализа аллель-специфичной экспрессии на уровне гена на основе геномных и транскриптомных данных NGS для выявления специфических особенностей аллелей. Он был использован в анализе резистентности уротелиальной карциномы к химиотерапии с целью извлечения информации, которая может быть использована для разработки новых методов лечения [433]. Halvade-RNA [434] повторно адаптирует методику работы GATK, с целью использования параллельной обработки для

сокращения времени обработки и достижения 93,8% сходства при выявлении вариаций в ДНК.

Помимо анализа SNP, активно разрабатываются инструменты для обнаружения структурных вариаций в геноме. Например, CNNdel [435] использует сверточные нейронные сети для объединения выходных данных различных инструментов анализа признаков, с целью выявления структурных вариаций. GT-WGS [436] использует вычислительные мощности облачных сервисов Amazon для эффективной обработки данных нового поколения с последующим обеспечением высокой консистентности (99,9%) с передовой практикой GATK в отношении вызовов SNP и InDel.

Для обнаружения CNV и других значительных структурных изменений, ограниченных кодирующими геномными областями, используются биоинформатические алгоритмы, способные точно измерять глубину покрытия и дисбаланс аллелей в файле выравнивания последовательностей (файл BAM). Примеры таких методов включают EXCAVATOR2 и KaryoScan, которые способны обнаруживать CNV и крупные структурные изменения хромосом [437,438].

Не рекомендуется использовать WES для обнаружения транслокаций и повторов (например, тандемных повторов) из-за их склонности к наличию точек разрыва или к выходу за пределы геномной области [411].

Не смотря на то что генерация данных не является проблемой с появлением NGS и существуют инструменты и базы данных биоинформатики для обработки полученных больших данных, грядущее развитие технологии непрерывного секвенирования молекул ДНК, таких как технология Oxford Nanopore, предполагается как возможный способ справиться с растущим объемом данных. Тем не менее, для обработки «омиксных» данных, необходимых для персонализированной медицины, потребуются качественно новые подходы по хранению биоинформатических данных. Например, в апреле 2016 года компания AstraZeneca объявила об интегративной геномной инициативе, направленной на разработку лекарств путем получения новых знаний о биологии болезней, выявления новых мишеней для лекарств, поддержки отбора пациентов для клинических исследований и выбора наиболее эффективного лечения для каждого пациента, подразумевающей персонализированную медицину [439]. Инициатива включала сотрудничество с Human Longevity, The Wellcome Trust Sanger Institute (Соединенное Королевство), и Институтом молекулярной медицины (Финляндия). Чтобы реализовать эту смелую инициативу, AstraZeneca создала собственный Центр геномных исследований, который к 2026 году секвенирует и анализирует до двух миллионов геномных

последовательностей (WGS и WES), включая 500 000 образцов, полученных в ходе клинических испытаний [440].

Успехи в области персонализированной медицины были достигнуты также с помощью технологий WGS и WES. Помимо прогнозирования реакции пациента на обычные лекарства, генетическая информация также используется для подбора таргетных противораковых препаратов. Johannessen C.M. и соавторы обращают внимание на следующее: «В то время как фармакогенетика обычных лекарств выявляет ДНК-варианты зародышевой линии, фармакогенетика рака предназначена для подбора низкомолекулярных ингибиторов и анализа соматических ДНК-вариантов из опухолевых клеток. Поскольку рак является преимущественно генетическим заболеванием, анализ ДНК опухоли обычно используется для молекулярной характеристики раковых клеток, а также для прогнозирования лечения и мониторинга. Получение образцов опухоли для генетического анализа может быть проблемой, если опухоль небольшая или недоступна. В последние годы биопсия не твердых опухолей успешно применялась для получения опухоли, циркулирующей в свободной ДНК. Теперь такую биопсию можно использовать для ранней диагностики рака, прогнозирования, выбора лечения и мониторинга. К сожалению, стоимость генетических онкологических тестов и таргетных методов лечения все еще высока, что делает их недоступными в менее развитых странах».[441]

Репродуктивное здоровье — еще одна область, в которой извлекли пользу WGS и WES. Так, например, неглубокий WGS с 3X (3х-кратным) покрытием выполняется для преимплантационной оценки эмбрионов, а также для выбора будущего пола. Неинвазивный пренатальный тест (NIPT) представляет собой комбинацию жидкой биопсии и WGS для обнаружения трисомий или других крупных хромосомных перестроек в клетках плода.

Из-за наличия значительной доли сигнала в некодирующей области генома, клиническая полезность геномной информации для многофакторных заболеваний при использовании WES все еще не имела достаточной прогностической силы. Тем не менее, последние исследования в области технологий биоинформатики, наряду с дополнительными практиками, увеличивающими степень уверенности в полученных результатах, успешно достигают поставленных целей. В настоящий момент есть исследования, свидетельствующие о том, что шкала полигенного риска для сложных заболеваний имеет такую же прогностическую силу, что и оценка генетического риска для моногенных заболеваний [20].

1.5. Будущее развитие технологий в области медицинской генетики

1.5.1. Пангеном человека

В предыдущих разделах данного литературного обзора было упомянуто, что в 2021 году геном человека стал на 100 % завершенным усилиями консорциума T2T. Процесс внедрения новой референсной последовательности генома человека (T2T-CHM13) должен занять какое-то время, но очевидно, что это позитивно повлияет на точность диагностики в клинических исследованиях. Важным результатом новой последовательности T2T является возможность более точной оценки генетических ДНК-вариантов. Команда исследователей ДНК-вариантов T2T зафиксировала значительные улучшения в выявлении и интерпретации генетических ДНК-вариантов с использованием новой последовательности T2T по сравнению со стандартным эталонным геномом человека.

Недавно консорциум T2T объединился с консорциумом Human Pangenome Reference Consortium (HPRC), целью которого является создание нового «эталонного пангенома человека» на основе полных последовательностей геномов 350 человек. Нынешняя структура генома человека представляет собой линейную композицию слитых гаплотипов более чем 20 человек, причем большая часть последовательности приходится на одного человека. Она содержит смещения и ошибки в рамках, которые не отражают глобальную геномную вариативность человека. Пангеномика заключается в том, чтобы наиболее полно охватить разнообразие человеческой популяции. T2T и HPRC ставят перед собой задачу просмотреть сотни геномов от теломеры к теломере. Проект «Пангеном человека» позволит сравнить новые секвенированные геномы с несколькими полными геномами, представляющими различные родословные человека и переосмыслить понимание эталонной информации для ресеквенирования. Карты сложных для секвенирования областей генома для нескольких особей способны дать более точное и разнообразное представление глобальной геномной вариативности, улучшить ассоциативные исследования различных фенотипов, в т.ч. медицинских, в разных популяциях, расширить сферу исследований геномики до наиболее повторяющихся и полиморфных регионов генома и послужить основным генетическим ресурсом для будущих биомедицинских исследований и точной медицины [442].

1.5.2. Применение длинных прочтений в медицинской генетике

Еще одним вектором решения задачи «недостающей наследственности» является целенаправленное применения технологии длинных прочтений. Было много предположений об источнике этой «недостающей наследственности», часто указывая на SV [443]. На долю SV приходится большее общее число нуклеотидных изменений в геномах человека, чем на гораздо более многочисленные однонуклеотидные ДНК-варианты [444]. До настоящего времени такие популяционные исследования в основном опирались на высокопроизводительные технологии секвенирования с короткими чтениями, которые дают чтения длиной от 25 п.н. до 400 п.н. [445]. Однако короткие чтения имеют важные ограничения при характеристике повторяющихся регионов [446,447]. Повторы ДНК часто способствуют образованию SV [448], и в то же время затрудняют их обнаружение SV из-за неточностей выравнивания чтений.

Секвенирование с длинным прочтением стало превосходить секвенирование с коротким прочтением и другие методы (например, массивы) для выявления структурных вариаций, что было продемонстрировано консорциумами Genome in a Bottle (GIAB) и Human Genome Structural Variation (HGSV), которые объединили несколько технологий для всесторонней характеристики структурных вариаций в геномах человека [449,450]. Эти исследования показали, что значительная часть скрытых вариаций может быть обнаружена с помощью секвенирования с длинным прочтением. Действительно, недавние исследования по секвенированию исландской и китайской популяций уже выявили ранее не обнаруженные ДНК-варианты, связанные с ростом, уровнем холестерина и анемией [451,452]. Секвенирование с длинным прочтением полезно для улучшения качества фазирования ДНК-вариантов [453], а также применяется для поиска аллелей, связанных с заболеваниями [454–456].

1.5.3. Внедрение искусственного интеллекта в медицинскую генетику

Высокая стоимость и технологические ограничения являются основными препятствиями для более глубокой интеграции медико-генетических подходов. Уже существуют различные решения с использованием искусственного интеллекта, которые внедряются для сокращения расходов, особенно в направлении преодоления огромного объема собираемых данных о пациентах. Так, например, применение инструмента Exomiser, представленного компанией Congenica в их продукте Sapientia, способствует ускорению аннотации и приоритизации вариантов ДНК-последовательностей всего экзома при диагностике редких заболеваний [457].

В настоящее время ведущие технологические компании, включая Google, усердно интегрируют возможности машинного обучения в свои облачные платформы, с целью стимулировать принятие сообществом новейших методов искусственного интеллекта. Одной из таких практик является применение глубокого обучения компанией Deep Genomics, для выявления генетических факторов заболеваний и подбора потенциальных лекарственных средств для целевых молекул. Аналогично, компания Nextcode, принадлежащая Wuxi, активно инвестирует в методы машинного обучения, чтобы участвовать в таких инициативах [458].

Также, команда Google Brain совместно с компанией Verily Life Sciences, еще одной дочерней компании Alphabet, фокусируются на разработке инструментов для биологических наук. Они разработали инструмент под названием DeepVariant, использующий передовые методы искусственного интеллекта для более точного анализа генома человека на основе NGS данных. Этот инструмент представляет собой альтернативу существующему GATK. В 2016 году DeepVariant занял первое место в PrecisionFDA Truth Challenge в категории "Лучшая эффективность определения SNP", что свидетельствует о высочайшей точности инструмента. Он также обладает быстротой, надежностью, гибкостью и простотой в использовании, и интегрирован в Google Cloud Platform [459].

Существует фундаментальное отличие классической статистики от статистики с использованием машинного обучения. Машинное обучение может улучшить статистические вычисления, но ему нужно гораздо больше данных, чтобы делать свои предположения достаточно точными [460]. Хотя рост объема данных NGS стал значительно замедляться благодаря временной замене WGS на WES, а также внедрению секвенирования отдельных молекул с помощью технологии Oxford Nanopore. Для последующего анализа с использованием методов машинного обучения требуются экспоненциальные объемы данных для того, чтобы максимально точно определять связь генотип-фенотип [461,462]. Пока алгоритмы по определению такой связи совершенствуются, параллельно с этим был разработан ряд инструментов биоинформатики, направленных на соотнесение ДНК-вариантов последовательностей с широким спектром биологических мета-данных и фенотипов. Эти инструменты нового поколения обеспечивают *in silico* оценку «омиксных» данных, полученных из WGS или WES, и аналитические возможности (часто с использованием ИИ) для приоритизации ДНК-вариантов или фенотипов [463,464]. Уже доказано, что этот подход обеспечивает достаточную прогностическую силу, которую можно сравнить с уровнем предсказательной способности при анализе менделевских заболеваний [20].

Методы машинного обучения чаще всего используются на двух уровнях клинической биоинформатики:

1. Оценка повреждения генов *in silico* (в основном с использованием скрытых марковских моделей) [464,465];
2. Приоритизация ДНК-вариантов и оценка ассоциации фенотипов, где используются различные методы машинного обучения при анализе текста, на базе алгоритмов анализа естественных языков.

Еще одним из основных достоинств использования методов машинного обучения в клинической практике является их эффективное применение в области обработки изображений [466,467]. Многие исследования включают использование методов машинного обучения для обработки изображений, связанных с патологическими новообразованиями и клинической визуализацией, с целью дополнительной поддержки принятия диагностических решений.

1.5.4. Расширение доступности прикладных биоинформатических инструментов

Стоит отметить, что параллельно разработке основных фундаментально важных алгоритмов и программ ведется разработка инструментов для удобства специалистов без профессиональных навыков программирования. Можно отметить что, такие инструменты могут быть направлены на адаптацию научных выводов в удобные медицинские форматы, аналогичные тем, которые применяются для представления результатов тестов на патологии. Доступные для врачей шаблоны, специально разработанные для отчетности о результатах исследований на основе NGS уже существуют. Однако, время, затраченное на внедрение новых технологий в рутинную медицинскую практику, было значительным [468]. Отчетность, удобная для врачей, определенно ускоряет процесс взаимодействия биоинформатиков и медицинских работников [469]. Не так давно некоторые компании предложили свои решения для клинической отчетности NGS, используя аналитику омиксных данных, основанную на технологиях искусственного интеллекта. Уже существует ряд хорошо зарекомендовавших себя коммерческих решений:

- Qiagen (Ingenuity Variant Analysis and Ingenuity Pathway Analysis) [470];
- Golden Helix (VarSeq, VSCkinical) [471];
- Advaita (iVariant/iPatway/iBio Guides) [472];
- Lifemap Sciences [473].

Все четыре решения доступны через веб-интерфейс и предлагают клиническую расстановку приоритетов с использованием в качестве входного файла стандартный

формат VCF. Приложения Qiagen (частью пакета является известный аннотатор ДНК-вариантов Annovar) являются лидерами у клиницистов, поскольку традиционно большинство компаний, занимающихся геномными лабораториями, используют их предложения [474]. Альтернативой Qiagen служит набор инструментов для анализа клинических экзомов (TGex) входящих в Lifemap Sciences [475,476], Это также удобный для врачей инструментарий анализа и отчетности WES. Он также является одним из самых доступных на рынке и объединяет более сотни различных биологических баз данных, начиная от базы данных генной онтологии и биологических путей (GO) и баз данных экспрессии, заканчивая специфическими базами данных по некоторым фенотипам.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Сведения об анализируемых когортах

2.1.1. Получение референсной информации по частотам аллелей для изучения редких ДНК-вариантов

Для проведения метаисследования по секвенированию экзомов жителей северо-западного региона России было использовано 694 образца, секвенированных с помощью секвенаторов Illumina HiSeq 2500 и HiSeq 4000. Данные получены в рамках проектов таких организаций, как ФГБНУ «НИИ АГиР им.Д.О.Отта» и СПБ ГБУЗ Городской Больницы №40. Данные частично были описаны в [30,477,478]. Набор данных содержал секвенированные образцы людей, агрегированных из различных исследовательских и клинических проектов (как контрольные группы, так и люди с заболеваниями (основные фенотипы: диабет взрослого типа у молодых (MODY), диабет 2 типа (T2D), ожирение, расстройства аутистического спектра (ASD), заболевания соединительной ткани (CTD) и нейрофиброматоз)). Большинство образцов принадлежало людям русской (~80 %) национальности (по самоотчету) и европеоидной расы; участники исследования преимущественно проживали в Северо-Западном регионе России.

Было проведено сравнение спектра генетических ДНК-вариантов с данными dbSNP (версия 151) и провели оценку распространенности аутосомно-рецессивных аллелей заболеваний на основе ClinVar по сравнению с gnomAD r. 2.1. В рамках этого исследования были рассмотрены одиночные точечные мутации и небольшие инсерции и делеции. Среди основных используемых инструментов следует выделить языки программирования (и их программные пакеты) python (numpy, scipy.stats) и R (ggplot2, reshape2, dplyr). С используемыми инструментами и программным кодом более подробно можно ознакомиться на публично доступной странице репозитория [479].

2.1.2. Получение результатов ассоциативных исследований для исследования генетических локусов ответственных за систематическое проявление множественных ассоциаций сразу с несколькими сложными признаками

Данные были получены с сайта лаборатории Бенджамина Нила (UK Biobank GWAS results imputed v2, загружено в 2017; [480]). В анализ взяты только те данные, которые соответствовали фенотипам со значимыми ненулевыми оценками наследуемости ($p<0.05$, оценки предоставлены авторами набора данных). После получения списка наследуемых фенотипов были сохранены геномно-значимые ДНК-варианты (одиночные точечные мутации и небольшие инсерции и делеции) для каждого фенотипа (p-значение$<5\times10^{-9}$). Сводная статистика ассоциаций для каждого ДНК-варианта по каждому признаку была объединена в единую матрицу для дальнейшего анализа.

С используемым программным кодом можно ознакомиться на публично доступной странице репозитория исследования [481].

2.1.3. Исследование плейотропии для объяснения дисбаланса частоты аллелей на примере когортного исследования фокального сегментарного гломерулосклероза

WES контрольных образцов представлены пациентами, данные которых собраны из нескольких когорт dbGAP, с отсутствующими нефропаталогиями в истории болезни (табл. 1).

Таблица 1. Доступ к данным

Когорта	Количество образцов	Доступ к данным
1000 Genomes	2074	https://www.internationalgenome.org
ATVB	2144	dbGAP phs000814.v1.p1
Autism_Daly	200	-
ESP	5182	http://evs.gs.washington.edu /EVS/
Ottawa_heart	787	dbGAP phs000806.v1.p1
T2GENES	3538	http://www.type2diabetesgen etics.org/projects/t2dGenes

Образцы ДНК, из подтвержденных методом биопсии ФСГС, были получены от пациентов, участвующих в многоцентровом исследовании NIH, а также от пациентов с

диагнозом, поставленным в Вашингтонском университете. Генетические данные были получены с помощью секвенирования «подоцитного экзома» – генетической панели из 2 482 генов, что согласуется с результатами предыдущего исследования. Из отобранных генов 5 напрямую связаны с семейной формой ФСГС, а 200 генов функционально связаны с 5 основными генами. Большая часть генов была отобрана по уровню экспрессии – 677 высоко экспрессировались в микропрепарированных гломерулах человека, а другие 1600 генов являются человеческими ортологами высоко экспрессированных генов подоцитов мыши. В рамках этого исследования были рассмотрены одиночные точечные мутации и небольшие инсерции и делеции.

2.2. Обзор методов использованных в исследовании

2.2.1. Анализ силуэта

Анализ силуэта применяется для нахождения оптимального числа кластеров в многомерных данных, для интерпретации данных кластеризации и для выявления степени связанности отдельных наблюдений внутри кластера [482]. В данном анализе каждому наблюдению присваивается численное значение, а именно мера того, насколько хорошо классифицирован образец, когда он отнесен к конкретной группе и в соответствии как с плотностью кластеров, так и в соответствии с разделяемостью групп. Взяв среднее значение по всем образцам, можно получить среднее значение силуэта. Оно варьируется от 1.0 до –1.0, что характеризует высокую или низкую степень разделения между кластерами, соответственно. Затем силуэты успешно используются после кластеризации в качестве меры достоверности кластеров [20, 29-31].

В медико-биологических областях анализ силуэта применяется, например, для оценки кластеризации на данных экспрессии генов [483] и генетических вариаций [484].

2.2.2. Метод главных компонент

Метод главных компонент (PCA) является одним из старейших и наиболее широко используемых статистических методов, направленных на снижение размерности данных. Метод PCA старается наиболее полно сохранить информацию об изменчивости переменных в виде их линейных функций. Новые переменные должны максимизировать дисперсию в данных и не коррелировать друг с другом. Такие переменные называются принципиальными компонентами. [485]

В генетике человека PCA используется в трех основных направлениях:

1) выводы о миграционной истории человека;
2) выявление субструктуры популяций;
3) коррекция популяционной стратификации в исследовании наследственных заболеваний [486].

В контексте медицинской генетики в первую очередь для нас актуальна возможность коррекции популяционной стратификации. Для решения этой задачи PCA применяется на когортных данных генотипирования человека. В анализе используют различные отсечки по частоте ДНК-варианта исходя из эмпирических наблюдений. За разделение «супер-популяций», таких как европейская и африканская в большей степени ответственны CV. В свою очередь более RV отвечают за локальное разделение внутри популяции, например, за различия между западными и восточными европейцами. Обычно первые несколько главных компонент, объясняющие большую часть дисперсии в данных, отвечают за разделение «супер-популяций», а последующие минорные принципиальные компоненты способны показывать локальную изменчивость внутри популяций [487].

2.2.3. Смешанные гауссовские модели

Для решения задачи кластеризации многомерных линейных данных применяются различные статистические модели. Одной из хорошо зарекомендовавших себя моделью является смешанные Гауссовские модели (Gaussian mixture model, GMM). Модель Гауссовской смеси является композицией нескольких нормальных распределений, соответствующих исследуемым измерениям в данных. На основе Гауссовской смеси распределений вычисляется комбинированная вероятность принадлежности отдельного наблюдения, представленного в виде многомерного вектора, конкретному кластеру [488].

В медико-биологической области Гауссовские смешанные модели нашли применение для кластеризации многомерных данных генотипирования. На практике это означает, что данные, полученные в ходе PCA анализа, объединяются с помощью Гауссовских смешанных моделей в группы исходя из генетической близости образцов. Подобную информацию применяют для исключения ложного сигнала ассоциации, возникающего из-за популяционной стратификации данных [488].

2.2.4. Одно-вариантные тесты ассоциации

В ассоциативных исследованиях стандартным подходом к тестированию ассоциации между генетическими ДНК-вариантами и сложными признаками является

одно-вариантный тест в рамках аддитивной генетической модели. Ассоциация между каждым ДНК-вариантом и признаком обычно оценивается с помощью линейной регрессии для непрерывных признаков и логистической регрессии для бинарных признаков. В тестах GWAS для одного ДНК-варианта обычно используется порог значимости 5×10^{-8}, что соответствует 5 % в масштабах всего генома, если провести ~1 миллион независимых тестов ассоциации [222]. С помощью этой простой процедуры были выявлены тысячи локусов, связанных с признаками. Классические тесты также могут выявить ассоциацию с ДНК-вариантами низкой частоты, при условии достаточного размера выборки. Например, как отмечалось ранее, одно-вариантные тесты в выборке из около 8 000 человек выявили ассоциации между процессингом инсулина и ДНК-вариантами в *SGSM2* (MAF=1.4 %, $p=8.7\times10^{-10}$) и *MADD* (MAF=3.7 %, $p=7.6\times10^{-15}$) [378].

Однако, простые тесты гораздо менее эффективны для анализа RV, чем для анализа CV с одинаковым размером эффекта [489]. Например, при отношении шансов ОШ = 1.4, размеры выборки, необходимые для достижения 80 % мощности, составляют 6 400, 54 000 и 540 000 для MAF=0.1, 0.01 и 0.001, соответственно, если предположить 5 % распространенность заболевания и уровень значимости 5×10^{-8}. Поскольку количество RV намного больше, чем количество распространенных ДНК-вариантов, могут потребоваться более строгие уровни значимости, что еще больше снизит итоговую мощность.

Тем не менее Ма С. и коллеги замечают: «Одно-вариантные тесты могут быть полезным инструментом для анализа RV, если объем выборки достаточно велик, эффекты очень велики или ДНК-варианты не слишком редки. Кроме того, в сочетании с такими инструментами, как «квантиль-квантиль»-график (QQ-plot) и Манхэттен-график, классические тесты могут использоваться для оценки качества данных и стратификации популяции. Следует отметить, что оценки уровней значимости числа p на основе одного ДНК-варианта, основанные на стандартных методах регрессии, могут быть неточными, если число испытуемых с данным ДНК-вариантом невелико, и решение этого вопроса потребует дальнейших методических разработок» [490].

2.2.5. Анализ редких ДНК-вариантов в ассоциативных исследованиях

RVAS является более сложной задачей, чем анализ обычных ДНК-вариантов. Во-первых, в случае, когда для одного отдельно взятого RV размеры эффекта и выборки не столь велики – статистическая мощность классических ассоциативных тестов становится недостаточной для разрешения подобного рода сигнала. Поэтому для простого

наблюдения RV с высокой вероятностью необходим большой объем выборки. Например, выборка аллелей с частотой 0.5 % или 0.05 % с вероятностью 99 % требует секвенирования не менее 460 или 4 600 человек соответственно. Во-вторых, стандартный анализ одно-вариантных ассоциаций недостаточно эффективен для выявления RV. Для решения этой проблемы разработана специальная методика проведения ассоциативных исследований, с использованием особых тестов. Данных подход оценивает ассоциацию нескольких ДНК-вариантов в биологически значимой области, вместо проверки эффектов отдельных ДНК-вариантов, как это происходит в классических ассоциативных исследованиях. В последние годы было предложено множество мультимаркерных тестов на основе регионов или генов.

2.2.6. Виды агрегирующих тестов

Вместо того чтобы тестировать каждый ДНК-вариант по отдельности, тесты агрегации оценивают совокупный эффект нескольких генетических ДНК-вариантов в гене или регионе, увеличивая мощность, когда несколько ДНК-вариантов в группе связаны с данным заболеванием или признаком. К настоящему моменту разработано множество методов, но в первую очередь стоит рассмотреть тесты, основанные на регрессии, которые позволяют легко корректировать ковариаты. В целом существует разделение методов на три класса: 1) тесты мутационной нагрузки (burden tests); 2) адаптивные тесты мутационной нагрузки (adaptive burden tests); 3) тесты на основе дисперсионных компонент (variance-component tests). Отдельно от этих трех групп существует ряд методов, направленных на объединение результатов – комбинированные тесты мутационной нагрузки (combined burden) и метод экспоненциальной комбинации (exponential-combination, EC).

Эти методы основаны на различных предположениях о базовой генетической модели, и мощность каждого теста в каждом конкретном случае, напрямую зависит от того, насколько он приближен к истинной модели исследуемого фенотипа. Поскольку истинная модель признака или болезни априорно неизвестна и изменчива, желательно использовать комбинацию тестов.

2.2.6.1. Тесты мутационной нагрузки

Тесты мутационной нагрузки, к которым относятся ARIEL, CAST, CMC, MZ, WSS исходя из названия численно описывают «обремененность» гена/региона множеством

генетических ДНК-вариантов, давая кумулятивную оценку гену одним числом [491–496]. Данный простой подход обобщает информацию о генотипе путем подсчета количества минорных аллелей по всем ДНК-вариантам в наборе ДНК-вариантов для конкретного гена или региона, после чего возвращает суммарный генетический балл. Узким местом при использовании этих тестов является то, что все RV в целевом гене или регионе должны иметь ненулевое и однонаправленное влияние на фенотип. Нарушение этого принципа ведет к значительной потере мощности.

2.2.6.2 Адаптивные тесты мутационной нагрузки

Для устранения ограничения обычных тестов мутационной нагрузки было разработано несколько адаптивных методов, которые устойчивы в присутствии нулевых ДНК-вариантов и ДНК-вариантов «разнонаправленно» влияющих на фенотип.

Тест ASUM [497] сначала оценивает направление эффекта для каждого ДНК-варианта, а затем проводит тест мутационной нагрузки уже с учетом имеющейся оценки. Тест Step-up [498] дополнительно уточняет процедуру, избавляясь от ДНК-вариантов, которые имеют незначительное направление эффекта.

Тест на расчетный коэффициент регрессии EREC [499] использует более прямой подход; он оценивает коэффициент регрессии каждого ДНК-варианта и использует его в качестве веса. Тест основан на ожидании, что истинный коэффициент регрессии является оптимальным весом для максимизации мощности. EREC нестабилен, когда количество минорных аллелей что снижает его оптимальность.

2.2.6.3. Тесты на основе дисперсионных компонент

Тесты на основе дисперсионных компонент исходят из модели со случайными эффектами. Эти методы проверяют наличие ассоциации, оценивая распределение генетических эффектов для группы ДНК-вариантов. В частности, дисперсионно-компонентные тесты, включают в себя тест C-alpha [500], тест SKAT [501,502] и тест SSU [503], которые оценивают распределение агрегированной статистики теста баллов отдельных ДНК-вариантов. SKAT представляет проблему в виде смешанных линейных моделей. В отсутствие ковариаций SKAT сводится к тесту С-альфа. SKAT также может учитывать взаимодействие SNP-SNP.

2.2.6.4. Комбинированные тесты мутационной нагрузки

Тесты на основе дисперсионных компонент более эффективны, чем тесты мутационной нагрузки, если в регионе много непричинных ДНК-вариантов или если причинные ДНК-варианты имеют разное направление ассоциации. Напротив, тесты мутационной нагрузки более эффективны, чем тесты дисперсионных компонентов, если в регионе высока доля причинных ДНК-вариантов с одинаковым направлением ассоциации. Оба сценария могут возникнуть, поэтому желательно объединить эти два подхода, что и реализуется в комбинированных тестах мутационной нагрузки.

Было предложено несколько методов для объединения тестов мутационной нагрузки и дисперсионно-компонентного теста. Наиболее простым способом для объединения p-значений этих двух тестов является метод Фишера [504,505]. Альтернативой является метод SKAT-O [506] и метод Саймса (Simes) [507].

2.2.6.5. Метод экспоненциальной комбинации

Тесты мутационной нагрузки и дисперсионно-компонентные тесты основаны на линейных и квадратичных суммах. Тест экспоненциальной комбинации [508] исходит из предположения, что только один ДНК-вариант в гене или регионе является причинным ДНК-вариантом.

2.2.6.6. Сравнение одно-вариантных и агрегирующих статистических тестов

Как упоминалось ранее, тесты на основе генов и регионов предназначены для увеличения мощности путем объединения сигналов ассоциации по нескольким RV. Действительно, если несколько ассоциированных ДНК-вариантов можно сгруппировать вместе, эти подходы могут привести к существенному увеличению мощности. Однако по сравнению с классическими тестами на основе одного ДНК-варианта, тесты агрегирующие статистику на ген или регион могут привести к потере мощности, в случае, когда один или несколько вариантов в гене ассоциированы с признаком, когда многие варианты не имеют эффекта, и когда причинные ДНК-варианты являются низкочастотными ДНК-вариантами.

Существуют примеры сравнительного применения тестов на основе генов и «повариантных» тестов, когда результаты «погенных» тестов могут превосходить результаты классических тестов [509]. В частности, с помощью генного теста была

выявлена ассоциация между болезнью Альцгеймера и геном *PLD3* (p-значение 1.4×10^{-11}), при этом ни один отдельный ДНК-вариант в гене не имел p-значения $< 10^{-6}$. Многие RV в *PLD3* встречались среди больных людей, но ассоциация не была столь значимой из-за очень низкого MAF, в связи с чем, «погенный» тест обеспечил лучшую мощность за счет объединения этих RV. В другом примере, изучая ассоциацию между липидами крови и генами *BCAM* и *CD300LG*, было обнаружено, что одиночные ДНК-варианты демонстрируют четкие доказательства ассоциации, но при этом на уровне целых генов статистические тесты показывают более слабый сигнал [510]. Объяснением может служить то, что соответствующие гены содержат очень небольшое число, не слишком RV, связанных с липидами крови.

2.2.7. Составление выборки для анализа редких ДНК-вариантов

Одним из важных вопросов при проведении статистических тестов на основе генов или регионов является выбор ДНК-вариантов для проверки ассоциации. Можно использовать все ДНК-варианты в регионе или подмножество ДНК-вариантов, отобранных на основе минорной частоты аллели (MAF), влияния на аминокислотную последовательность (например, несинонимичные SNP) или другой аннотации на основе последовательности. Были разработаны методы биоинформатики для предсказания функциональной роли ДНК-вариантов, такая информация также может быть использована для уточнения подмножеств ДНК-вариантов для проведения ассоциативного анализа. Например, PolyPhen-2 [261] предсказывает, является ли ДНК-вариант «доброкачественным», «возможно повреждающим» или «вероятно повреждающим». Можно провести тест на ассоциацию с ДНК-вариантами «возможно повреждающими» и «вероятно повреждающими» или только с ДНК-вариантами «вероятно повреждающими». В качестве альтернативы можно назначить весовые коэффициенты для различных классов ДНК-вариантов, повысив вес функционально повреждающих или низкочастотных ДНК-вариантов.

Существующие методы биоинформатики не совершенны и могут давать неточные прогнозы, поэтому их следует рассматривать лишь как один из возможных ДНК-вариантов для уточнения подмножеств ДНК-вариантов для проведения ассоциативного анализа.

2.2.8. Выбор статистического теста для анализа редких ДНК-вариантов

Было разработано множество методов для проверки связи заболевания с наборами RV. Относительная эффективность этих методов зависит от превалирующей и обычно неизвестной архитектуры заболевания. Если существует предварительная информация, можно выбрать тест ассоциации, включив эту информацию. Например, если предполагается, что в регионе имеется большая доля причинно-следственных RV, и большинство из них увеличивает риск заболевания, тесты мутационной нагрузки, вероятно, будут более мощными. Если предполагается, что в регионе существуют как повышающие, так и снижающие риск ДНК-варианты (причинные и протективные) или, что большинство ДНК-вариантов обладают нулевым эффектом, дисперсионно-компонентные тесты, вероятно, будут более мощными. Если нет предварительной информации, можно попробовать несколько методов и скорректировать p-значения с учетом использования нескольких методов, чтобы избежать завышенных ошибок первого рода, или использовать комплексный тест, который, вероятно, будет иметь надежную мощность для ряда моделей заболеваний.

2.2.9. Мета-анализ

Мета-анализ представляет собой эффективный способ объединения данных нескольких исследований [511–513]. Мета-анализ может быть проведен с использованием простой сводной статистики по конкретным исследованиям для построения тестовой статистики по большому количеству выборок как для CV, так и для RV. Для нахождения ассоциации в RV мета-анализ играет особенно важную роль, из-за повышенных требований к объему выборки. Простейшим методом мета-анализа является объединение p-значений или Z-баллов с помощью методов Фишера или Стауффера [504,505,514,515]. Однако хорошо известно, что такой подход менее эффективен, чем совместный анализ данных индивидуального уровня и мета-анализ с фиксированными эффектами [514].

Есть примеры алгоритмов проведения мета-анализа RV, в которых вместо p-значений используются балльно-ранжированная статистика с фиксированными размерами эффектов [510,516–518]. Эти подходы сильно отличаются по вычислительной эффективности в сравнении с традиционным мета-анализом на основе, например, теста Вальда. Это достигается за счет учета только нулевой модели и численной стабильности, т.е. отсутствия необходимости оценивать коэффициенты регрессии и их стандартную ошибку, что затруднительно для RV в условии малых выборок. Мета-анализ с фиксированными эффектами может использовать данные индивидуального уровня для достижения мощности, практически идентичной мощности единого совместного анализа

этих исследований [510,516]. Подобный эффект достигается благодаря тому, что каждое исследование, принимающее участие в мета-анализе, помимо того, что предоставляет подробную статистику баллов для отдельных ДНК-вариантов, дополнительно рассчитывает межвариантные ковариационные матрицы, которые отражают информацию о LD между ДНК-вариантами в зависимости от популяционной структуры данных. Эти матрицы впоследствии позволяют рассчитать асимптотические p-значения. Ранее рассмотренные тесты мутационной нагрузки, SKAT, SKAT-O и VT были разработаны в рамках подобных мета-анализов.

Генетические эффекты могут быть неоднородными в разных исследованиях, что приводит к общей потери мощности. Например, это происходит из-за описанной ранее межпопуляционной стратификации, которая может быть устранена внутри каждого исследования, но отличаться между исследованиями. Если устранить меж-исследовательскую неоднородность в данных, то можно избежать потери мощности итогового мета-анализа [519,520]. Так, был разработал метод мета-анализа одно-вариантных трансэтнических исследований с использованием байесовской модели разбиения, которая учитывает ожидаемую неоднородность между различными группами предков [519]. Авторы другого проекта разработали метод мета-анализа RV, который позволяет учитывать различные уровни неоднородности между отдельными исследованиями или отдельными группами предков путем введения различной структуры корреляции между параметрами генетического эффекта [516].

Помимо влияния генетических различий, возникающей из-за несоответствия в генетической структуре отдельных исследований, существует опасность накопления технических неточностей. Различные платформы и стратегии секвенирования могут создавать различные типы ошибок секвенирования, артефактов и смещений [521]. Тщательная фильтрация ДНК-вариантов и контроль качества важны для предотвращения выявления ассоциаций, обусловленных межплатформенной гетерогенностью в данных [490].

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Получение информации о редких причинных ДНК-вариантах в менделевских заболеваниях на примере русской этнической группы

Частота минорной аллели в популяции является одним из важнейших факторов, влияющих на интерпретацию риска, связанного с генетическим ДНК-вариантом. В течение предыдущего десятилетия несколько крупномасштабных проектов были направлены на характеризацию частот ДНК-вариантных аллелей, как в глобальных, так и региональных популяционных группах. К ним относятся, например, проект 1000 геномов [522] и консорциум Exome Aggregation Consortium (ExAC)/Genome Aggregation Database (gnomAD) [523,524]. Несмотря на чрезвычайно большое количество образцов в крупнейшей базе данных gnomAD (125 748 для версии 2.1), генетическая изменчивость во многих регионах земного шара до сих пор плохо изучена.

Многие страны пытаются восполнить этот пробел, запуская национальные геномные проекты (например, исследования населения Катара [525,526]. Одной из таких инициатив является проект «Геномы России», запущенный в 2015 г. Основной целью проекта является сбор генетической информации для характеризации спектра генетической изменчивости различных этнических групп в России [527,528]. Однако текущая судьба проекта неизвестна, а количество проб, включенных в первую фазу проекта, недостаточно для того, чтобы делать предположения о распространенности моногенных расстройств. Помимо целенаправленного проведения мета-исследования генетики российской популяции существует большое количество данных из клинических, и исследовательских проектов по всей России.

Данные о частотах аллелей, распространенных среди определенных популяций представленных, например, в таких базах данных, как gnomAD, имеют первостепенное значение для интерпретации результатов ресеквенирования в клинических исследованиях. Из-за вариативности генома частоты аллелей, в том числе в его кодирующей части, могут значительно различаться в малоизученных популяциях, недостаточно представленных в крупномасштабных проектах. Наиболее интерпретируемыми с точки зрения рисков,

являются кодирующие RV, поэтому настоящую работу было принято начать с описания генетического ландшафта моногенных заболеваний для одной из специфических популяций. В частности, Российская популяция обладает недостатком референсной информации об аллель-частотном спектре.

ДНК-варианты с риском доминантных заболеваний обычно предельно редкие и испытывают большое давление естественного отбора. Очевидно, что оценка популяционной частоты ДНК-вариантов с риском подобных заболеваний обычно требует огромного массива данных, что не всегда возможно без формирования специальных научных консорциумов.

Такие заболевания можно рассматривать не в контексте популяционных данных, а в контексте эпидемиологических, т.е. на небольших когортах в рамках «случай»-исследований (case studies), желательно с применением семейной информации по каждому пациенту. В качестве примера можно привести исследования синдрома трехфаланговой полисиндактилии большого пальца [29,30] и диабета зрелого типа у молодых [29,30]. Поэтому, с точки зрения популяционной частоты, наиболее реалистичной и релевантной является ее приближенная оценка с помощью частоты носительства рецессивных аллелей.

3.1.1. Составление русской этнической когорты

В рамках проводимого мета-исследования была составлена когорта из 694 образцов, собранных из различных независимых небольших клинических исследований [477,478,529]. Все образцы были секвенированы с помощью секвенаторов Illumina HiSeq 2500 и HiSeq 4000. Набор данных содержал людей, агрегированных из различных исследовательских и клинических проектов (как контрольные группы, так и люди с заболеваниями (основные фенотипы: диабет молодого возраста с началом зрелости (MODY), диабет 2 типа (T2D), ожирение, расстройства аутистического спектра (ASD), заболевания соединительной ткани (CTD) и нейрофиброматоз)). Большинство людей были русской (~80 %) национальности (по самоотчету) и европеоидной расы. Участники исследования преимущественно проживали в Северо-Западном регионе России.

3.1.2. Сравнение распределений аллелей между исследуемой когортой и открытыми базами данными

Был охарактеризован спектр генетических ДНК-вариантов, обнаруженных в используемой в исследовании выборке. Для полного набора из 694 участников исследования было определено в общей сложности 463 100 генетических ДНК-вариантов внутри целевых областей экзома. Из них 420 187 (90.7 %) имели rsID в соответствии с последней сборкой 151 dbSNP, остальные 42 913 ДНК-вариантов ранее не были учтены в базе данных. Между ДНК-вариантными сайтами в dbSNP и не-dbSNP не наблюдалось существенных различий в общей глубине секвенирования или качестве картирования. Стоит отметить, что присутствовало небольшое снижение частоты альтернативной аллели для гетерозиготных генотипов в новых сайтах (рис. 1).

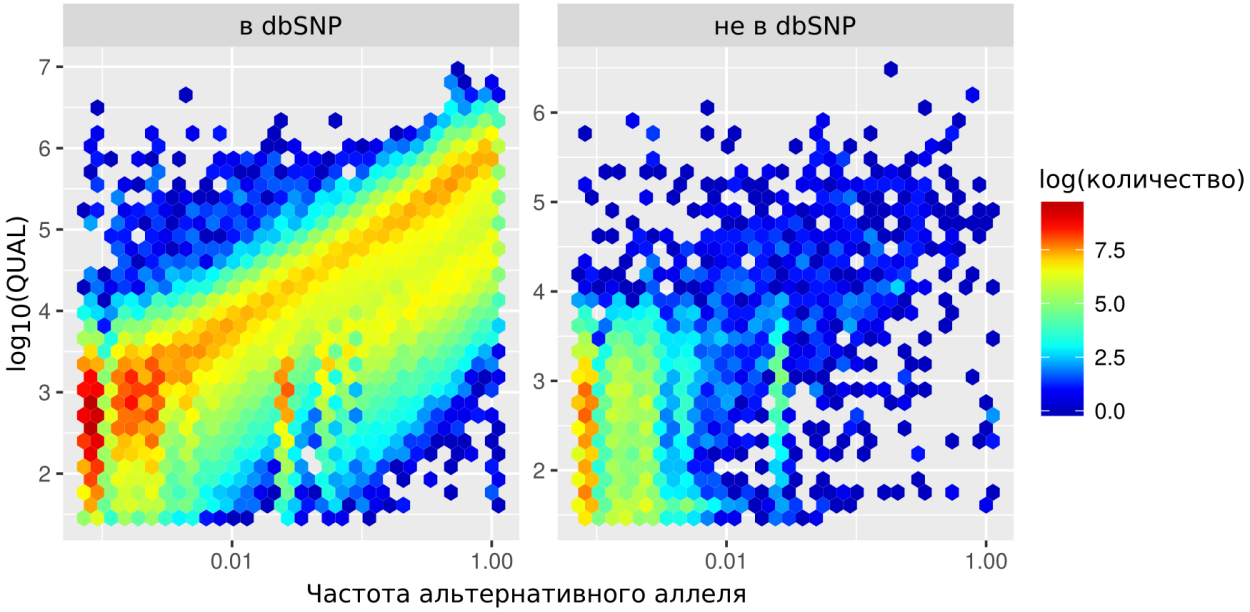

Рисунок 1. Сравнение распределений баллов качества для ДНК-вариантов, о которых сообщается или не сообщается в базе данных dbSNP (версия 151). Цвет шестиугольника обозначает количество сайтов ДНК-вариантов в логарифмической шкале

Была проведена оценка распределения ДНК-вариантов по функциональным категориям патологического воздействия в соответствии с классификацией SnpEff [530]. Как и ожидалось, подавляющее большинство ДНК-вариантов оказалось синонимичными или некодирующими ДНК-вариантами (300 117, 64.8 %); 14 106 ДНК-вариантов были классифицированы как ДНК-варианты с высоким уровнем воздействия (большинство ДНК-вариантов индуцируют укорачивание белкового продукта), в среднем 622.5

ДНК-варианта с высоким уровнем воздействия на образец относятся к WES популяции (134 к CES) (рис. 2).

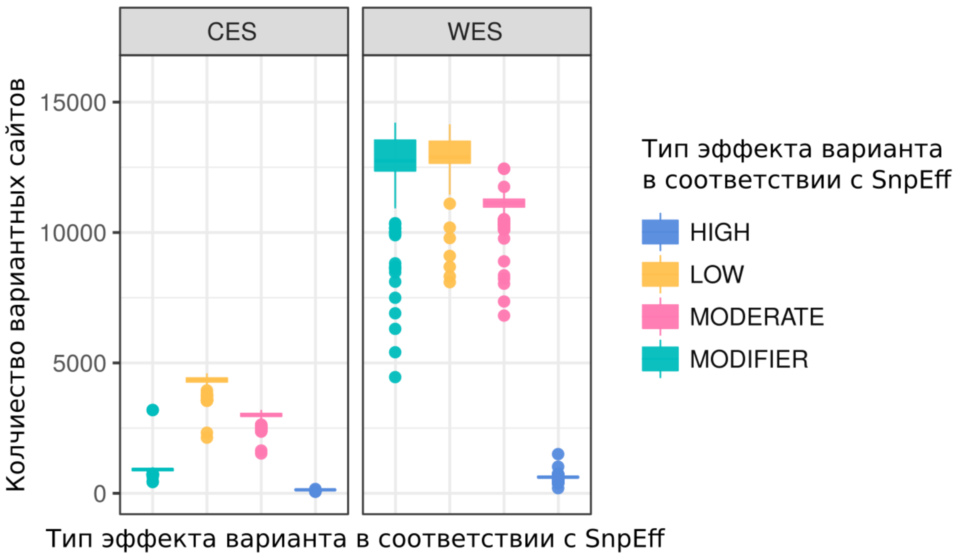

Рисунок 2. Распределение числа ДНК-вариантов, вызванных в каждом образце цельного экзома (WES, справа) и образце Illumina TruSight One (CES, слева), разделенных по типу эффекта ДНК-вариантов

Интересно, что большая часть ДНК-вариантов, которые принадлежали к группе с высоким уровнем воздействия, не были зарегистрированы в dbSNP (5 143/14 106, хи-квадрат p-значения по отношению к общему распределению < 2.2×10^{-16}). Это согласуется с наблюдением, что большинство идентифицированных новых ДНК-вариантов имели более низкую частоту альтернативного аллеля (AAF; alternative allele frequency) (рис. 3).

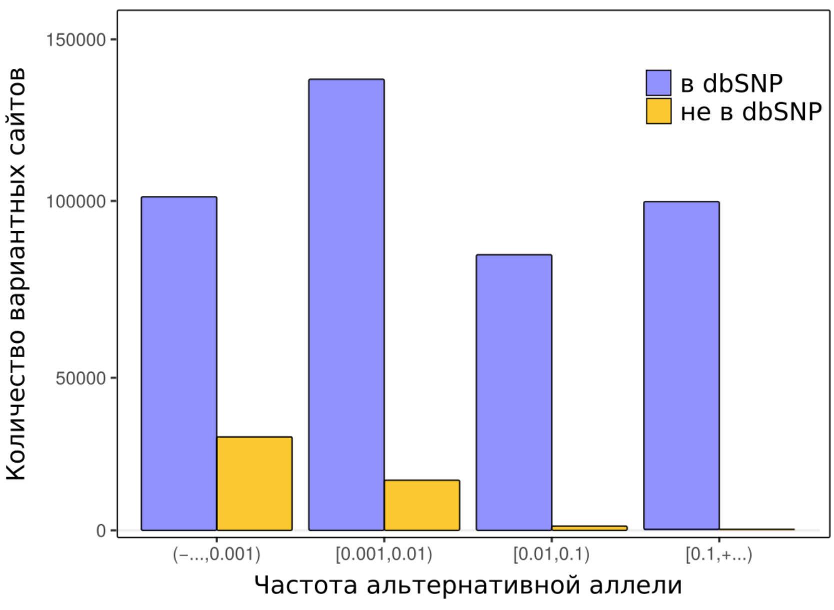

Рисунок 3. Пропорции сайтов dbSNP и не-dbSNP в наборе данных в зависимости от частоты альтернативной аллели. Обратите внимание, что подавляющее большинство новых вариантов имеют AAF<0.01.

3.1.3. Корреляция аллельных частот между исследуемой когортой и открытыми базами данными

На следующем этапе была проведена оценка соответствия между частотами альтернативных аллелей в исследуемой выборке и в экзомах из базы данных gnomAD r. 2.1. С этой целью набор данных был сужен до сайтов, которые имеют среднее покрытие не менее 15X в gnomAD для, а затем произведена оценка коэффициентов линейной регрессии для аллельных ДНК-вариантов. В итоге была обнаружена сильная корреляция между аллельными частотами в gnomAD и исследуемым набором данных (R^2=0.96; рис. 4).

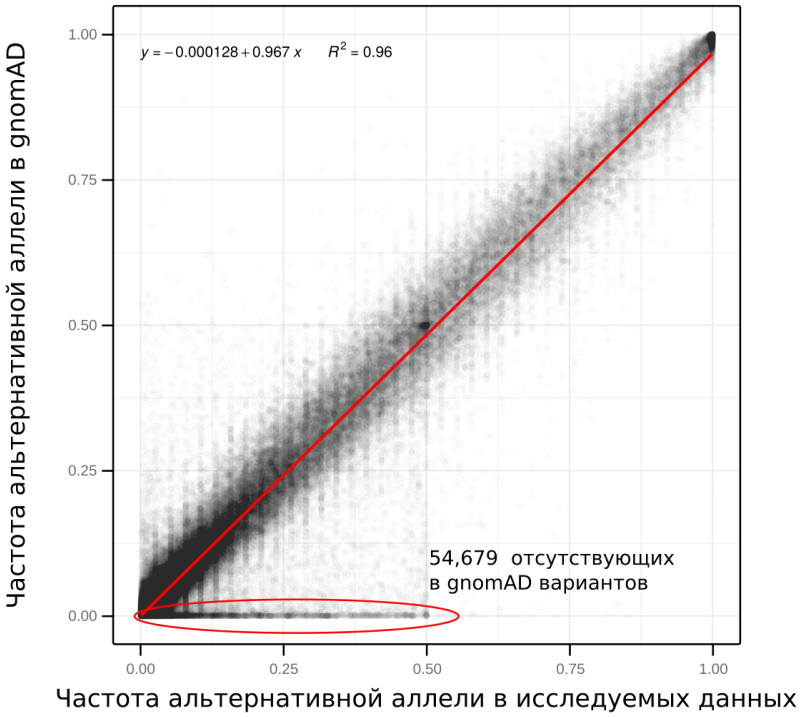

Рисунок 4. Диаграмма разброса частот альтернативных аллелей в наборе данных в сравнении с частотами на основе gnomAD; исключены сайты, использующие только gnomAD, а также мульти-аллельные записи и плохо покрытые регионы

Также стоит отметить, что в gnomAD не присутствовало 54 679 ДНК-вариантов. Большинство таких ДНК-вариантов также отсутствовало в dbSNP build v. 151 (37 338, 68.2 %), что свидетельствует о том, что эти ДНК-варианты составляют специфический компонент генетической структуры населения Северо-Запада России.

Следующим этапом исследования было изучение взаимосвязи между когортой, составленной из жителей Северо-Запада России (NWR) и основными популяциями, присутствующими в gnomAD. Сначала была изучена корреляция между частотами альтернативных аллелей в NWR и в пяти основных глобальных популяциях: африканской (AFR), смешанной американской (AMR), восточноазиатской (EAS), нефинской европейской (NFE) и южноазиатской (SAS). Метод главных компонент аллельных частот в 121 171 экзомных вариантах, представленных в исследуемом наборе данных, и во всех пяти популяциях gnomAD также показал, что когорта NWR находятся в непосредственной близости от нефинских европейцев (рис. 5).

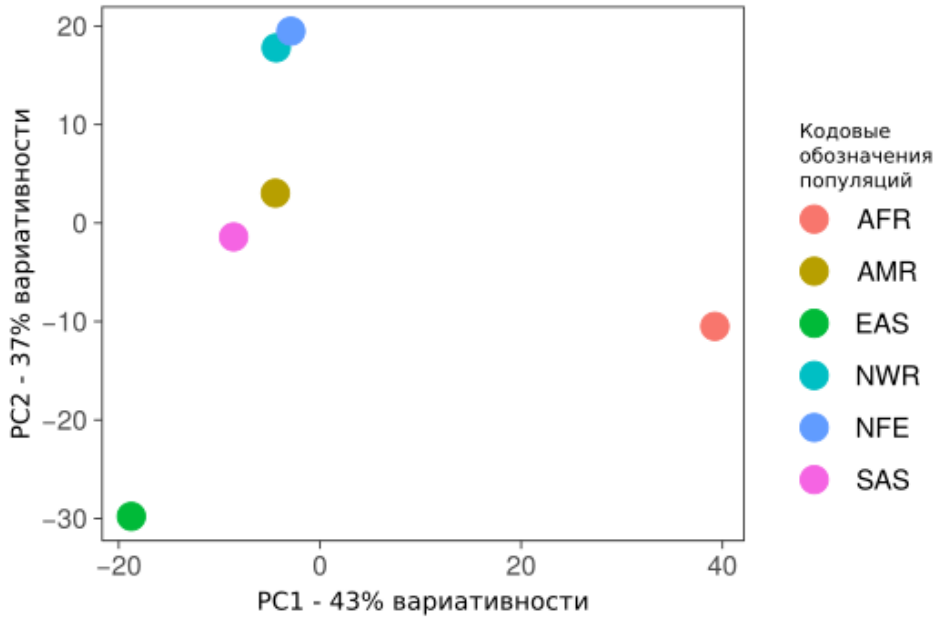

Рисунок 5. Анализ частот аллелей 121 171 WES вариантов, представленных в Северо-Западной России и во всех популяциях gnomAD, на основе анализа главных компонент. AFR – африканская; AMR – смешанная американская; EAS – восточноазиатская; NFE – нефинская европейская; NWR – Северо-Запад России; SAS – южноазиатская.

Чтобы дополнительно подтвердить эти наблюдения, было рассчитано среднеквадратичное отклонение частот аллелей между исследуемым набором данных и популяциями gnomAD. Наименьшая разница наблюдалась с популяцией NFE (рис. 6). Эти результаты хорошо согласуются с первичными данными проекта «Геномы России» [528].

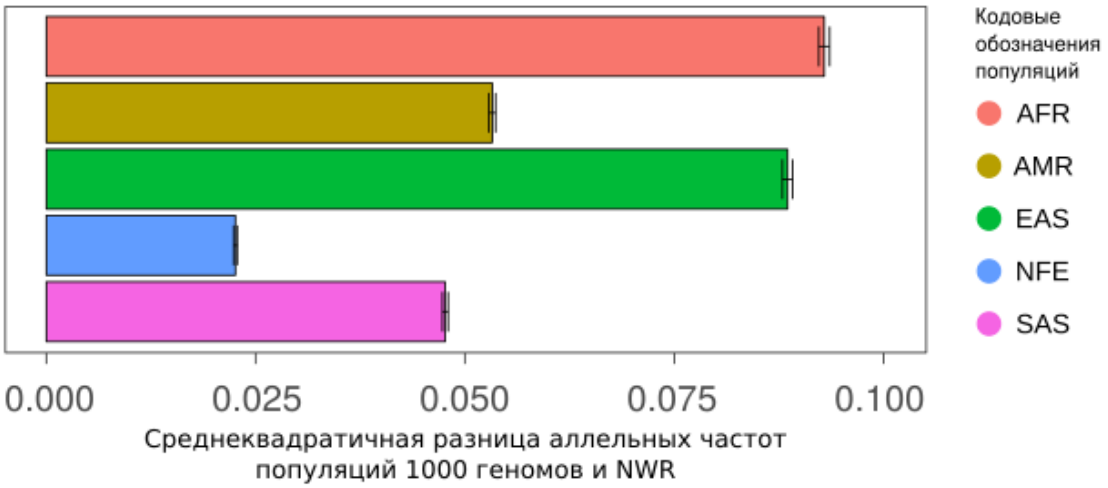

Рисунок 6. Среднеквадратичная разница в частотах аллелей между NWR и различными популяциями gnomAD. Диапазоны ошибок представляют собой границы 95 % доверительного интервала для среднего значения

3.1.4. Оценка частот патогенных аллелей в исследуемой когорте

Далее в исследуемой выборке была проведена оценка частот аллелей, вызывающих заболевания. Во-первых, набор данных был сужен до набора с хорошо установленным фенотипом без тяжелых заболеваний. В результате такой фильтрации был получен набор из 372 не родственных между собой лиц, который содержал 314 902 из 463 100 (68.0 %) вариантов. Далее анализ был сосредоточен на вариантах, которые являются патогенным согласно базе данных ClinVar и при этом имеют относительно низкую (<0.5 %) частоту в gnomAD и являются строго гетерозиготными в исследуемом наборе данных.

Было обнаружено несколько примеров широко распространенных известных патогенных вариантов аутосомно-рецессивных заболеваний в исследуемой выборке (табл. 2). Наиболее часто встречался вариант rs5030858 в гене *PAH* (MIM#612349; gnomAD_NFE_AF=0.0015, $p=7.9\times10^{-4}$), хорошо известный и один из наиболее патогенных CV связанный с фенилкетонурией [531]. Носители аллели для этого варианта наблюдались суммарно 17 раз в наборах данных различного размера, что соответствует оценочной частоте аллели 0.0081. Второй распространенной аллелью заболевания был вариант rs36209567 в гене фактора свертываемости крови VII (*F7*, MIM#613878) (gnomAD_NFE_AF=0.001, p=0.001). Заболевание чрезвычайно редко встречается в общей мировой популяции (1:500 000) [532]. Другим примером широко распространенного

причинного менделевского варианта в исследуемом наборе данных был вариант rs61754365 в гене *TYR*, связанный с тирозиназо-отрицательным альбинизмом (gnomAD_NFE_AF=3.2×10^{-4}, p=1.1×10^{-4}) [533]. Интерес представляет более высокая частота патогенного варианта rs76151636 (gnomAD_NFE_AF=0.0013, p=0.016) в гене *ATP7B* (MIM#606882), вызывающем болезнь Вильсона (WD). Мутация H1069Q, с идентификатором rs76151636, считается наиболее распространенной мутацией WD в Европе и Северной Америке [534], что согласуется с наблюдаемой частотой. Выявилась необычно высокая распространенность варианта rs542489955 в гене *FKBP14* (MIM#614505; gnomAD_NFE_AF=0.001, p=0.0061), мутации сдвига рамки считывания, связанной с кифосколиозным типом синдрома Элерса-Данлоса (EDS) [535]. Этот вариант является наиболее частой мутацией EDS, что необычно, учитывая, что EDS, связанный с *FKBP14*, по-видимому, является одной из редких форм заболевания. Среди менее распространенных патогенных вариантов наблюдалось статистически значимое преобладание аллелей болезни Шарко-Мари-Тута типа 4B3 (*SBF1*, MIM#603560; rs200488568, p=1.2×10^{-4}); Лейциноз (*BCKDHB*, MIM#248611; rs386834233, p=0.003), комбинированный дефицит окислительного фосфорилирования 10 (*MTO1*, MIM#614667; rs201544686, p=4.9×10^{-4}) (по 3 наблюдения каждое) и ряд других рецессивные патологии. Примечательно, что не было выявлено распространенных вариантов укорочения белка, отсутствующих в ClinVar (в генах заболеваний, с аутосомно-рецессивными типом наследования). Что позволяет предположить, что большинство аллелей заболевания являются общими для населения России и Европы. Также была проанализирована частота каждой выявленной аллели распространенного заболевания в популяциях gnomAD, отличных от NFE. Интересно, что один из вариантов, rs38683423 в гене *BCKDHB*, также чрезмерно представлен в финской популяции. Это может указывает на поток генов между NWR и финской популяцией.

Таблица 2. Распространенные аллели риска заболеваний в NWR когорте, значительно отличающиеся по частоте от NFE популяции

Локус	ID в dbSNP	Ген	Частота аллели в gnomAD	Частота аллели в gnomAD в NFE	N аллелей	Estimated AF (lower/upper CI)	Степень различия частот SNV между NWR и NFE		Заболевание или признак
							p.value	OR	
12:103234271	rs5030858	*PAH*	7.6×10^{-4}	0.0015	6	0.0081 (0.0037/0.0175)	7.9×10^{-4}	2.52	фенилкетонурия
13:113772982	rs36209567	*F7*	5.6×10^{-4}	0.001	5	0.0067 (0.0029/0.0157)	0.0010	3.51	дефицит фактора свертывания крови VII (гипопроконвертинемия)
7:30058726	rs542489955	*FKBP14*	5.5×10^{-4}	0.001	4	0.0054 (0.0021/0.0138)	0.0061	2.59	Синдром Элерса Данлоса, кифосколиотический 2 типа
11:88911771	rs61754365	*TYR*	2.9×10^{-4}	3.2×10^{-4}	4	0.0054 (0.0032/0.0139)	1.1×10^{-4}	4.71	Глазокожный альбинизм
13:52518281	rs76151636	*ATP7B*	9.2×10^{-4}	0.0013	4	0.0053 (0.0021/0.0138)	0.0159	1.89	Болезнь Вильсона — Коновалова (гепатолентикулярная дегенерация)
22:50893287	rs200488568	*SBF1*	3.3×10^{-4}	1.4×10^{-4}	3	0.0045 (0.0015/0.0133)	1.2×10^{-4}	13.25	моторно-сенсорная нейропатия, 4B3 типа (болезнь Шарко-Мари-Тута)
2:152357937	rs549794342	*NEB*	2.7×10^{-4}	4.7×10^{-4}	3	0.0040 (0.0014/0.0118)	0.0050	3.71	Немалиновая миопатия
6:74191932	rs201544686	*MTO1*	1.7×10^{-4}	2.0×10^{-4}	3	0.0040 (0.0014/0.0118)	4.9×10^{-4}	9.32	Комбинированный дефицит оксида фосфора 10
6:80910740	rs386834233	*BCKDHB*	5.5×10^{-4}	3.9×10^{-4}	3	0.0040 (0.0014/0.0118)	3.0×10^{-3}	4.57	Болезнь «Кленового сиропа»

									(лейциноз)
5:54527618	rs775051461	*CCNO*	9.8×10^{-5}	4.7×10^{-5}	2	0.0030 (8.3×10^{-4}/0.0110)	4.6×10^{-4}	30.75	Первичная цилиарная дискинезия
17:8015495	rs121434233	*ALOXE3*	1.5×10^{-4}	2.8×10^{-4}	2	0.0027 (7.4×10^{-4}/0.0098)	0.018	4.9	Аутосомно-рецессивный врожденный ихтиоз
18:21119369	rs543206298	*NPC1*	7.6×10^{-5}	1.1×10^{-4}	2	0.0027 (7.4×10^{-4}/0.0098)	0.0033	14.3	Болезнь Ниманна–Пика
8:75276240	rs104894080	*GDAP1*	3.2×10^{-5}	7×10^{-5}	2	0.0027 (7.4×10^{-4}/0.0097)	0.0013	18.63	Наследственная моторно-сенсорная нейропатия (болезнь Шарко-Мари-Тута)
2:73677806	rs1307458231	*ALMS1*	2.0×10^{-5}	4.4×10^{-5}	2	0.0027 (7.4×10^{-4}/0.0097)	5.0×10^{-4}	30.87	Синдром Альстрема

Интересно, что для генов, связанных с риском развития аутосомно-рецессивных заболеваний, не было выявлено высоко распространенных патогенных или вероятно патогенных вариантов, отсутствующих в базах данных ClinVar или dbSNP, в генах, связанных с аутосомно-рецессивными заболеваниями. В этом случае можно предположить, что для рецессивных патологий, большая часть генетических детерминант является общей для России и других популяций.

При этом, риски, связанные с носительством RV, не обязательно ограничиваются менделевскими заболеваниями. Так, в частности, известны примеры RV, вызывающих моногенные формы полигенных заболеваний [7,10]. На следующих этапах работы были рассмотрены эффекты взаимодействия между RV и CV и их влияние на риски широкого спектра фенотипов.

3.2. Анализ результатов ассоциативных исследований для идентификации генетических локусов, ответственных за систематическое проявление множественных ассоциаций сразу с несколькими сложными признаками.

Известно, что варианты с большой популяционной частотой, обычно несут низкие риски заболеваний. По этой причине для изучения связи генотипов таких вариантов с заболеваниями обнаруживается потребность в когортах большого размера. Что оказывается чрезвычайно затратным для методов NGS, но доступным для микрочипового генотипиирования. К примеру, UK Biobank (UKB) собрал результаты ассоциативных исследований об около 500 000 людей.

С появлением ассоциативных исследований гипотеза CDCV утратила свое единоличное главенство из-за проблемы «отсутствующей наследственности» [536]. Результаты GWAS почти без исключения объясняют меньшинство предполагаемой генетической дисперсии [251]. Подавляющее большинство сложных дезадаптирующих признаков нельзя объяснить несколькими десятками локусов с умеренным эффектом, каждый из которых объясняет несколько процентов наследуемости. Одной из проблем в данной сфере является факт того, что CV не обязательно имеют уникальную специфичность к одному фенотипу, из-за чего при оценке риска наследственных заболеваний многие исследователи все чаще фокусируют свои усилия на изучении плейотропии.

На углубление понимания плейотропии повлияли национальные биобанки. Данные о генетических ассоциациях из UKB, привлекли внимание ученых во всем мире, и уже было предпринято несколько попыток по проведению масштабных мета-исследований [537–540]. Эти исследования выявили и оценили сложность человеческого фенома и обнаружили известные локусы с высоким плейотропными ассоциациями (например, локус MHC (HLA)).

В рамках настоящего исследования были проанализированы все наследственные признаки с использованием сводной статистики GWAS из набора данных UK Biobank. Также была проведена оценка распространенности и функциональных последствий различных уровней плейотропии.

3.2.1. Оценка соотношения фенотипической и генетической информации в популяционных данных UK Biobank

Для проведения геномного сканирования на предмет плейотропных локусов сначала были получены наборы значительно ассоциированных SNP для всех фенотипов в данных UK Biobank. На данном этапе была использована предварительно рассчитанная сводная статистика GWAS, предоставленная лабораторией Бенджамина Нила (релиз 1, 2018-02-25). Набор данных включал как стандартные локусы GWAS, так и импутированные варианты, в общей сложности 10 894 597 вариантов. Анализ был сосредоточен только на 543 сложных признаках, которые имеют значительные ненулевые оценки наследуемости (h^2). Оказалось, что общей сложности 469 013 (4.27 %) SNP имели по крайней мере один фенотип, связанный с геномным уровнем значимости, т.е. в среднем 4.34 фенотипа, связанных с каждым вариантом. Интересно, что в результатах анализа наблюдались многочисленные множественные ассоциации во всем наборе данных: 230 296 (49.21 %) SNP имели >1 ассоциированного фенотипа, а для 57 856 (12.34 %) SNP - более 10 ассоциированных фенотипов.

3.2.2. Кластеризация фенотипической информации

Хотя многие из этих множественных ассоциаций могут быть настоящими плейотропными вариантами, большая часть наблюдаемого сигнала, вероятно, возникает из-за нескольких высококоррелированных фенотипов. Поэтому была продолжена кластеризация тех признаков, которые имеют значительную долю общей генетической архитектуры. В качестве меры расстояния для кластеризации была использована восстановленная фенотипическая корреляция. Была применена иерархическая кластеризация в сочетании с силуэт-анализом для определения оптимального числа независимых кластеров. Было определено 308 кластеров сложных признаков, включающих в среднем 1.76 фенотипа. Среди них 48 кластеров включали три и более фенотипа (рис. 14, 15).

Рисунок 14. Диаграмма Вороного (верхняя часть) для мульти-фенотипических кластеров, построенная на иерархической кластеризацией значимых SNPs. Цветные блоки соответствуют отдельным кластерам. Числа равны количеству кластеров, объединенным внутри более общего фенотипа.

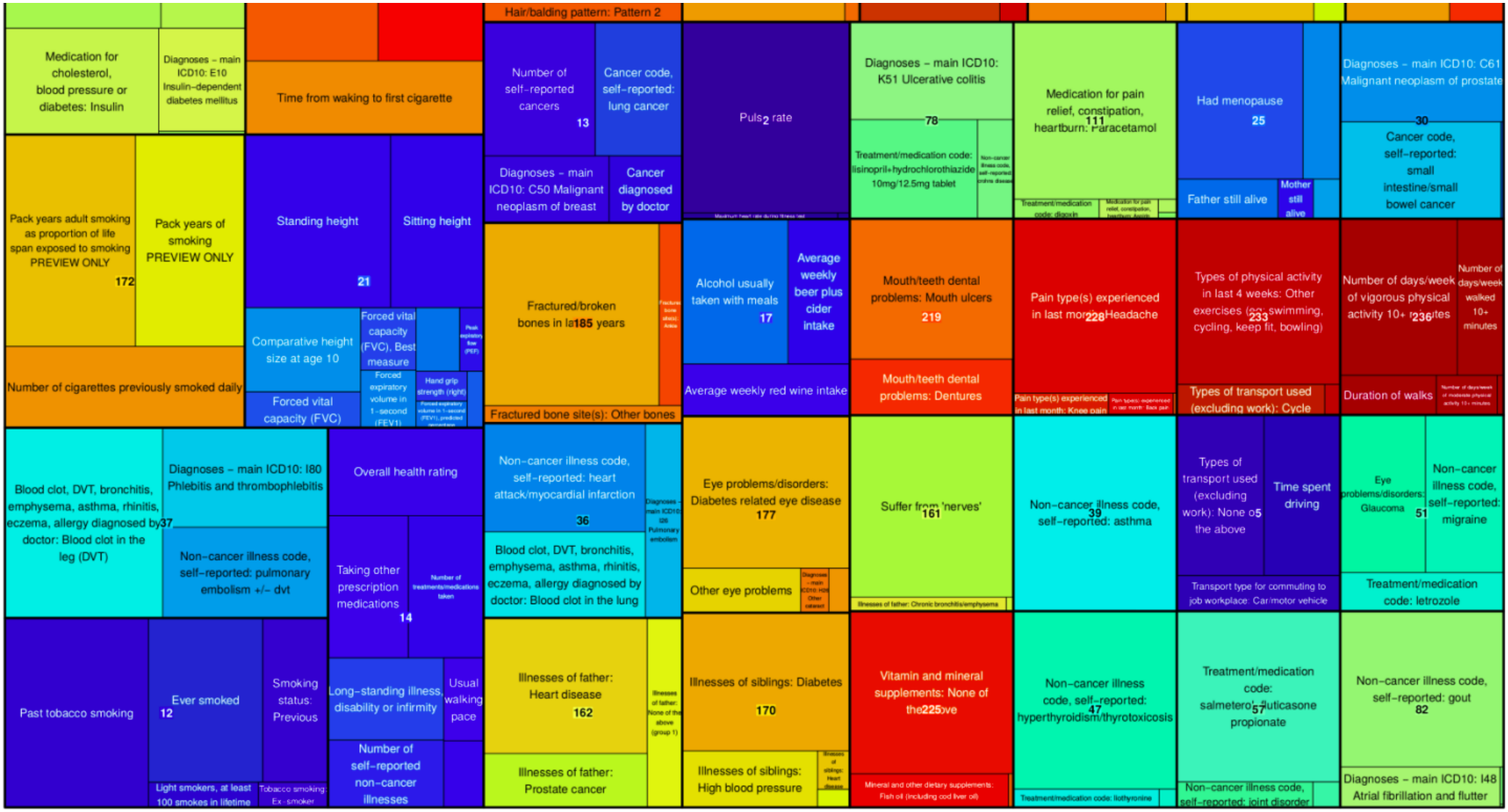

Рисунок 15. Диаграмма Вороного (нижняя часть) для мульти-фенотипических кластеров, построенная на иерархической кластеризацией значимых SNPs. Цветные блоки соответствуют отдельным кластерам. Числа равны количеству кластеров, объединенным внутри более общего фенотипа.

Процедура кластеризации существенно уменьшила количество SNP, имеющих множественные ассоциации (149 345 (31.1 % из всех ассоциированных SNP) SNP с более чем одной ассоциацией по сравнению с 230 296 в некластеризованных данных) (рис. 7); а также среднее число кластеров ассоциированных признаков (1.77 кластера на вариант).

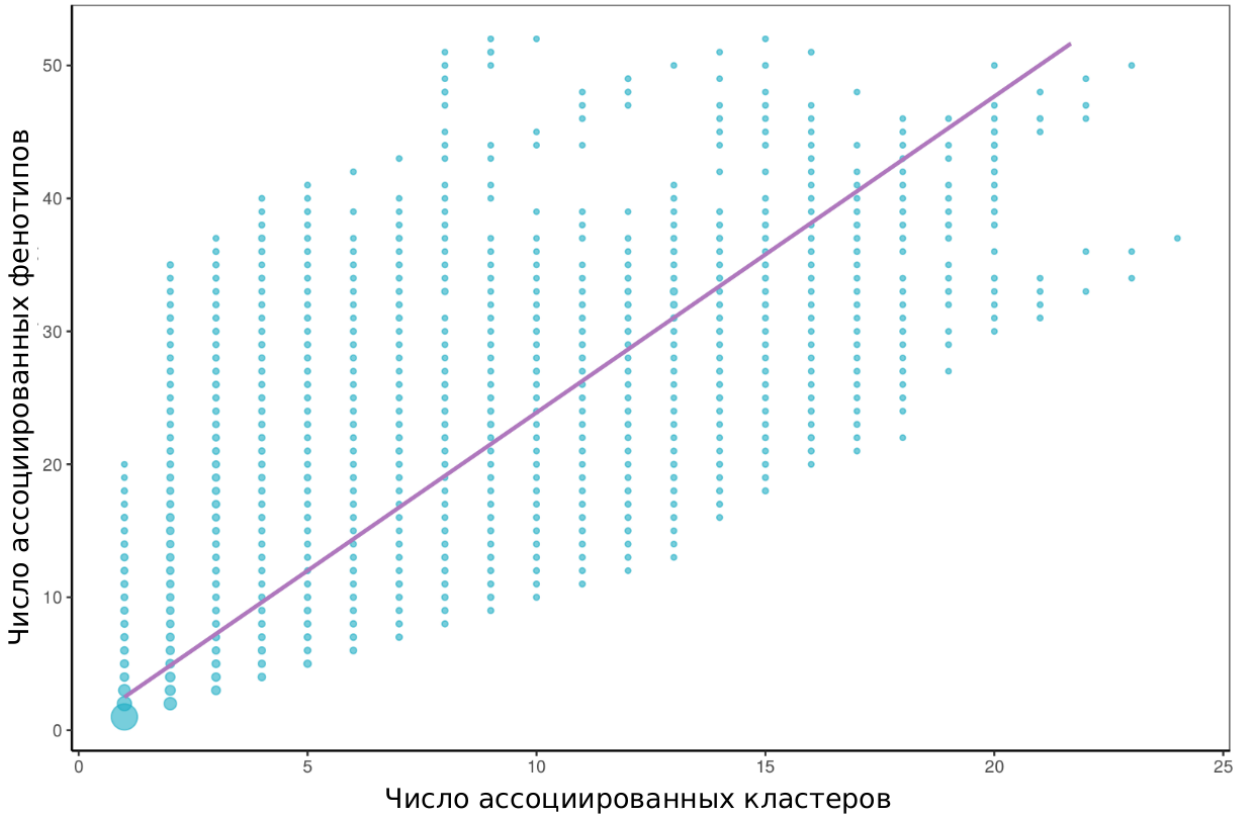

Рисунок 7. Диаграмма рассеяния количества ассоциированных фенотипов и кластеров фенотипов на один SNP. Размер точки пропорционален log10 (количество SNP).

Количество плейотропных SNP с более чем 10 ассоциациями также снизилось до 1.5 % (7 072 варианта) после проведения процедуры кластеризации (рис. 8, 9).

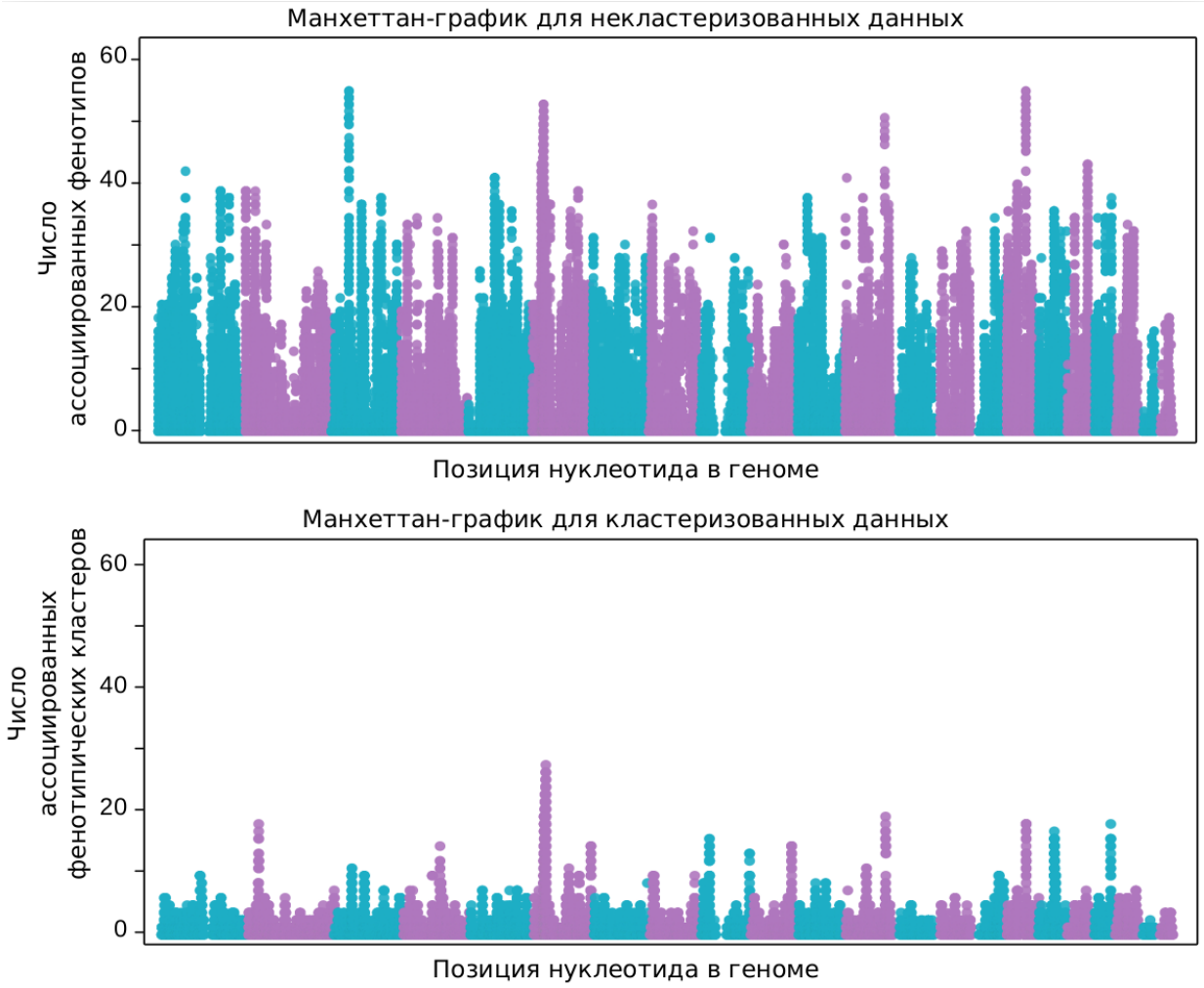

Рисунок 8. Кластеризация схожих признаков значительно снижает количество множественных ассоциаций. (а) График Манхэттена для числа ассоциаций на один SNP в некластеризованных (вверху) и кластеризованных (внизу) данных.

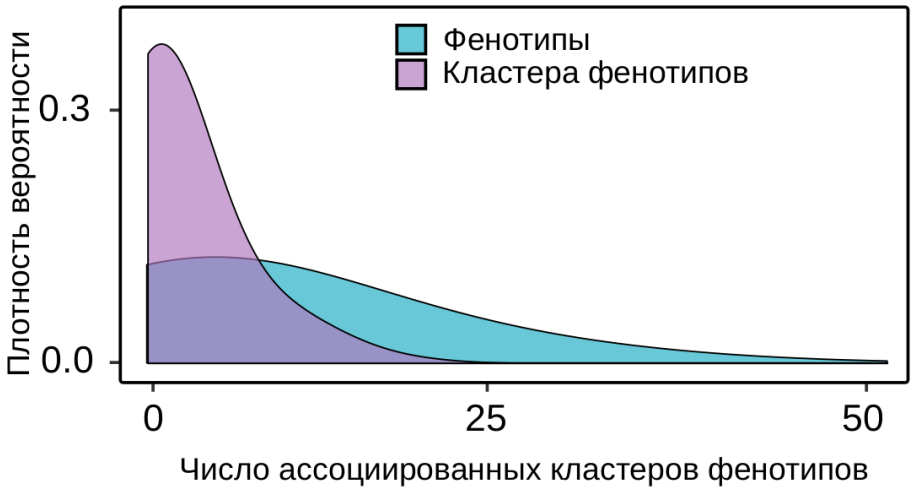

Рисунок 9. Сравнение числа ассоциаций на один SNP до и после кластеризации признаков по фенотипической корреляции.

3.2.3. Закономерность степени плейотропности относительно функционального эффекта варианта и аллельной частоты

Далее была проведена оценка распространенности различных функциональных типов вариантов в плейотропных наборах. Было обнаружено значительное преобладание миссенс-вариантов (651/64 544 по сравнению с 3 073/469 013; p.value Фишера = 3.1×10^{-11}) среди плейотропных вариантов по сравнению с вариантами, связанными хотя бы с одним признаком (рис. 10).

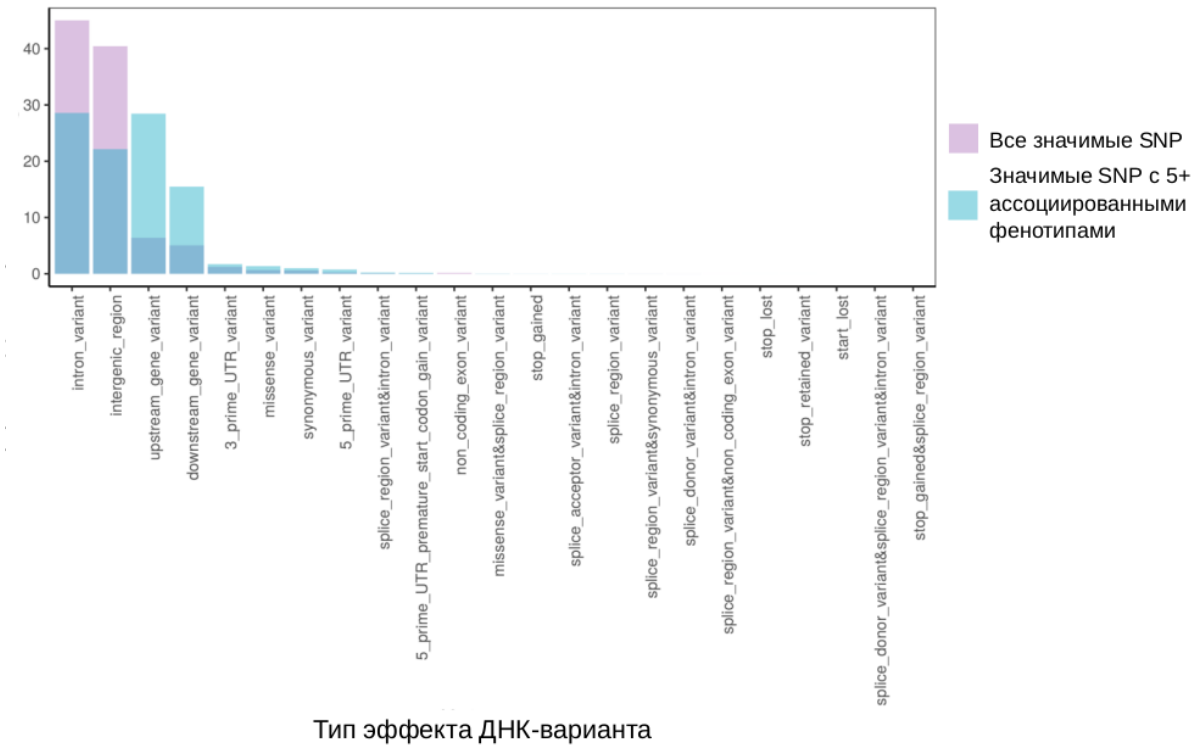

Рисунок 10. Гистограмма доли SNPs в каждом функциональном классе для плейотропных SNPs и всех SNPs с по крайней мере 1 значимой ассоциацией. Варианты перед 3` и 5` нетранслируемыми областями, а также миссенс-варианты перепредставлены в плейотропной группе.

Интересным наблюдением стало то, что присутствовало относительно небольшое количество RV с более чем 5 ассоциированными кластерами, несмотря на гораздо более высокую распространенность RV в исследуемом наборе данных (рис. 11). Другими словами, RV имеют тенденцию быть менее плейотропными по сравнению с распространенными.

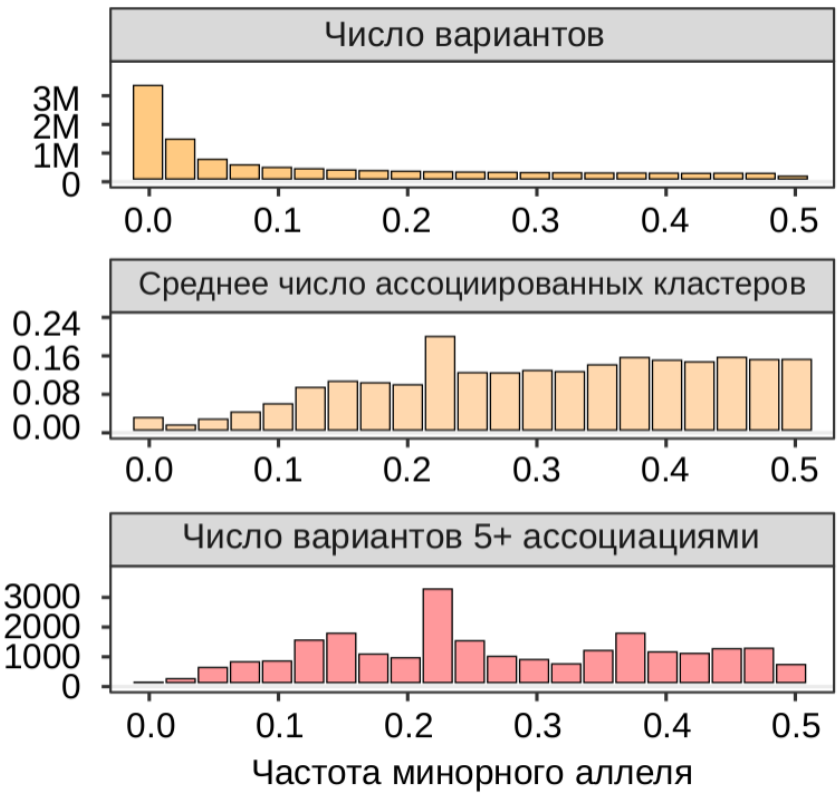

Рисунок 11 Сводная статистика ассоциаций для вариантов с различной частотой минорных аллелей. Значения агрегированы по бинам размером 0.025.

Это наблюдение можно объяснить тремя возможными способами. Во-первых, отсутствие плейотропных RV может быть вызвано ограниченной статистической мощностью для обнаружения ассоциации для RV. Во-вторых, более высокая распространенность плейотропных CV может говорить о том, что большая часть их плейотропных эффектов на самом деле ложно-ассоциирована и объясняется высоким LD у нескольких отдельных причинных вариантов [541]. Наконец, более низкая степень плейотропии RV может быть вызвана сильным очищающим естественным отбором, действующим против высоко плейотропных вариантов с большими эффектами, в результате чего все плейотропные варианты имеют более низкие размеры эффектов и более высокую частоту.

Предположение о наличии очищающего естественного отбора согласуется с наблюдениями, что естественный отбор против дезадаптирующих вариантов действует на некоторые формы сложных признаков [291,296]. Таким образом, если плейотропные варианты, влияющие на заболевания человека, имеют тенденцию быть пагубными, то можно ожидать, что высоко плейотропные варианты будут удаляться из популяции или сохраняться при низких аллельных частотах [542]. Чтобы добавить убедительности

данной теории было проведено кросс-популяционное когортное исследование на примере фокального сегментарного гломерулосклероза.

3.3. Вклад плейотропии в объяснение дисбаланса частоты аллелей на примере когортного исследования фокального сегментарного гломерулосклероза

Фокальный сегментарный гломерулосклероз (ФСГС) является основной причиной нефротического синдрома [543] и встречается в 20-30 % случаев хронической почечной недостаточности [544]. Уверенно оценить распространенность ФСГС сложно из-за сложности получения биопсии почек, необходимой для подтверждения диагноза. В настоящее время частота случаев ФСГС оценивается как 1.9 среди европейцев и 6.8 пациентов на миллион среди африканского и афроамериканского населения.

Генетические исследования ФСГС проводились с использованием как анализа родословных, так и с применением когортных ассоциативных исследований. Результатом стал ряд успешно выявленных генов, ассоциированных с увеличением риска развития ФСГС. Первые генетические исследования первичной ФСГС были основаны на наблюдении более высокой распространенности ФСГС в африканских и афроамериканских популяциях и гипотезе о наличии положительного отбора на гены предрасположенности к ФСГС. Открытие вариантов G1/G2 в гене *APOL1*, приводящих к риску развития ФСГС и защите от трипаносомоза, дало одно из возможных объяснений дисбаланса распространенности заболевания среди популяций. Кроме того, G1/G2 аллели были обнаружены как в семейных, так и в спорадических случаях ФСГС европейского происхождения, однако с гораздо меньшей частотой. Это показывает актуальность современного подхода к изучению сложных наследственных признаков, когда следует в первую очередь обращать внимание не на о форму и этиологию болезни, а на функциональность конкретных вариантов.

3.3.1. Составление когорты и методика проведения случай-контроль исследования

В рамках проводимого исследования были собраны крупномасштабные генетические данные первичного ФСГС вместе с когортой контрольных образцов. Для секвенирования использовалась специальная генетическая панель, содержащая порядка 2 500 генов, связанных с подоцитами. Было проведено ассоциативное исследование, в котором воспроизвели ранее ассоциированные гены и выявили новый ген-кандидат *CR1* на предрасположенность к ФСГС. Важно отметить, что, как и в случае с *APOL1*, варианты риска ФСГС в гене *CR1* находятся под положительным отбором в африканской популяции.

Образцы ДНК, из подтвержденных методом биопсии ФСГС, были получены от пациентов, участвующих в многоцентровом исследовании NIH, а также от пациентов с диагнозом, поставленным в Вашингтонском университете. Генетические данные были получены с помощью секвенирования «подоцитного экзома» – генетической панели из 2 482 генов, что согласуется с результатами предыдущего исследования. Из отобранных генов 5 напрямую связаны с семейной формой ФСГС, а 200 генов функционально связаны с 5 основными генами. Большая часть генов была отобрана по уровню экспрессии – 677 высоко экспрессировались в микропрепарированных гломерулах человека, а другие 1 600 генов являются человеческими ортологами высоко экспрессированных генов подоцитов мыши (рис. 12).

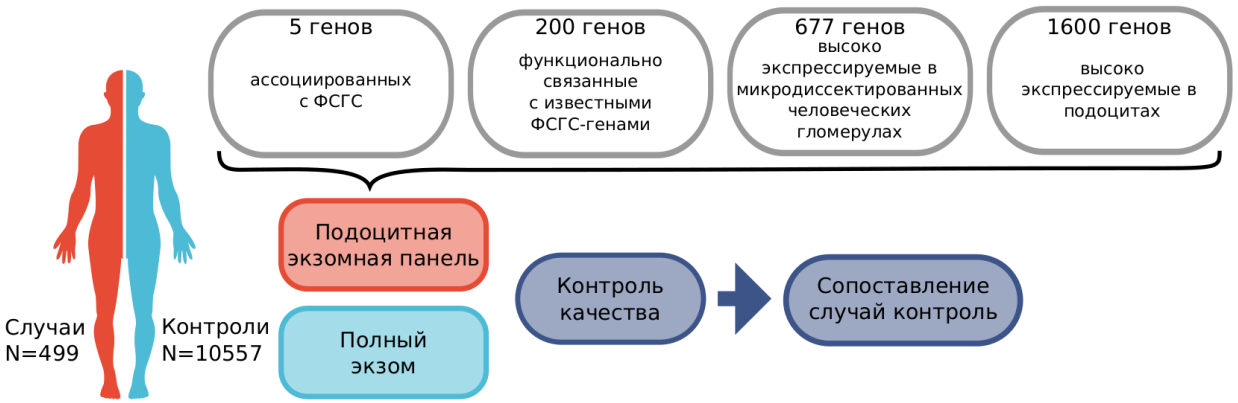

Рисунок 12. Дизайн исследования и анализ качества сопоставление образцов типа случай-контроль

Данные WES прошли совместный анализ поиска вариантов для создания набора данных случай-контроль. Получившийся набор генетических данных был подвергнут процессу обработки (рис. 13). Окончательные отфильтрованные данные включали 499 образцов случаев, 10 557 контрольных образцов, 131 179 вариантов.

3.3.2. Контроль качества и популяционная стратификация генотипических данных

Рисунок 13. Анализ контроля качества. Выбор порогового значения фильтров.

Для учета стратификации популяции был проведен совместный PCA-анализ генотипов случаев и контролей с последующей кластеризацией с использованием смешанных гауссовых моделей с помощью пакета AutoGMM [488]. В ходе анализа было получено 8 кластеров (рис. 14), которые были соотнесены с данными проекта 1000 геномов (рис. 15). Четыре минорных кластера соответствующих популяциям AMR, SAS, и EAS, были исключены из дальнейшего рассмотрения в силу недостатка случаев в них. Это приводит к низкой статистической мощности набора данных (рис. 16).

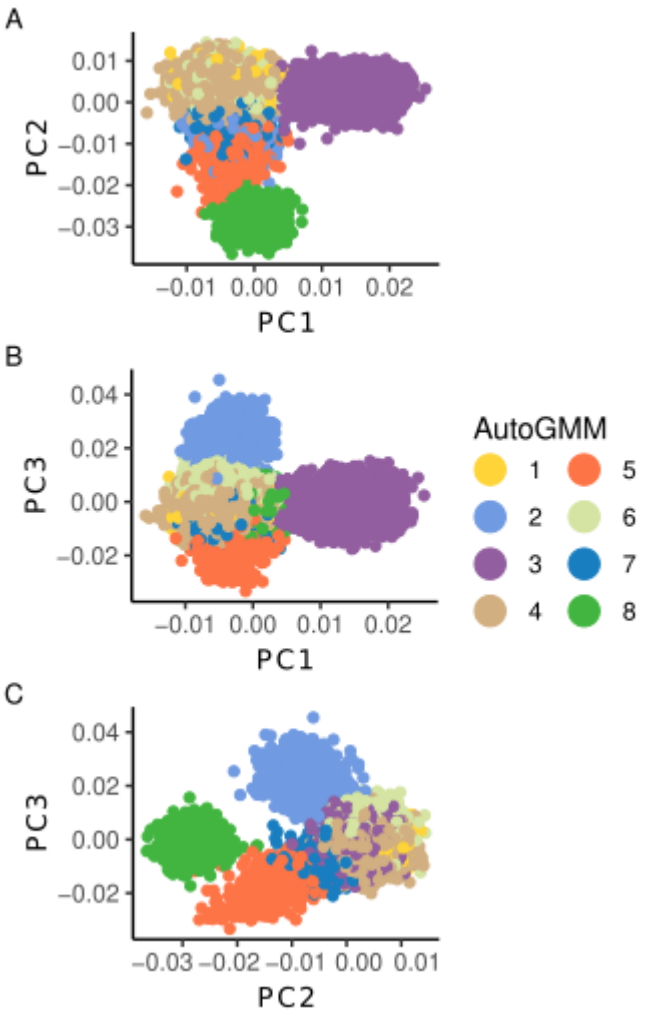

Рисунок 14. Кластеризация популяционных данных. Цвет точек соответствует принадлежности кластерам, назначенным пакетом AutoGMM: A) проекция компонент PC1 и PC2; B) проекция компонент PC1 и PC3; C) проекция компонент PC2 и PC3.

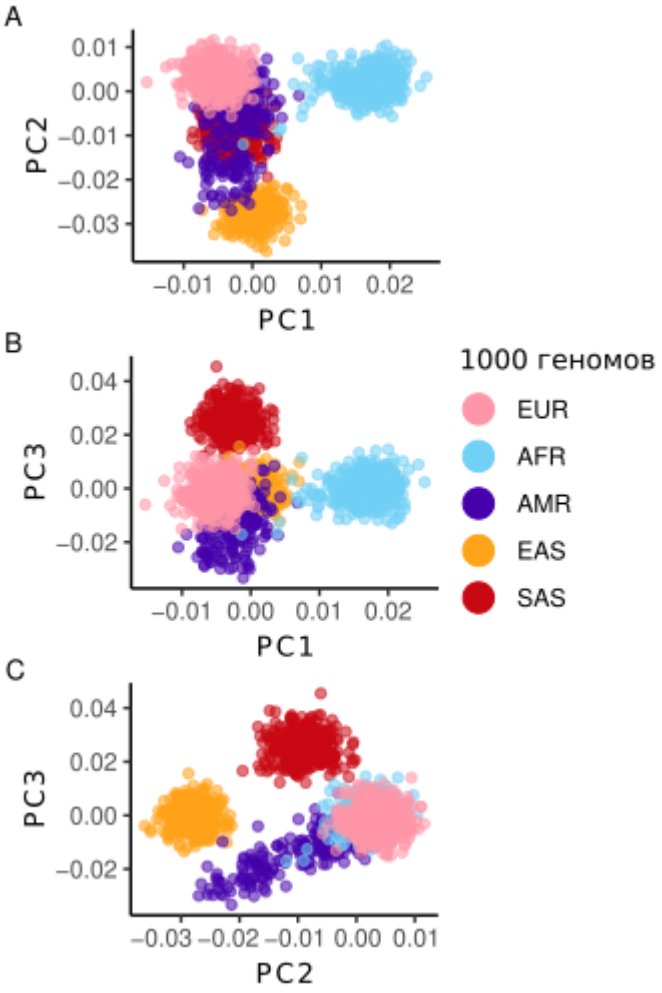

Рисунок 15. Кластеризация популяционных данных. Цвет точек соответствует принадлежности кластерам из данных 1000 геномов: A) проекция компонент PC1 и PC2; B) проекция компонент PC1 и PC3; C) проекция компонент PC2 и PC3.

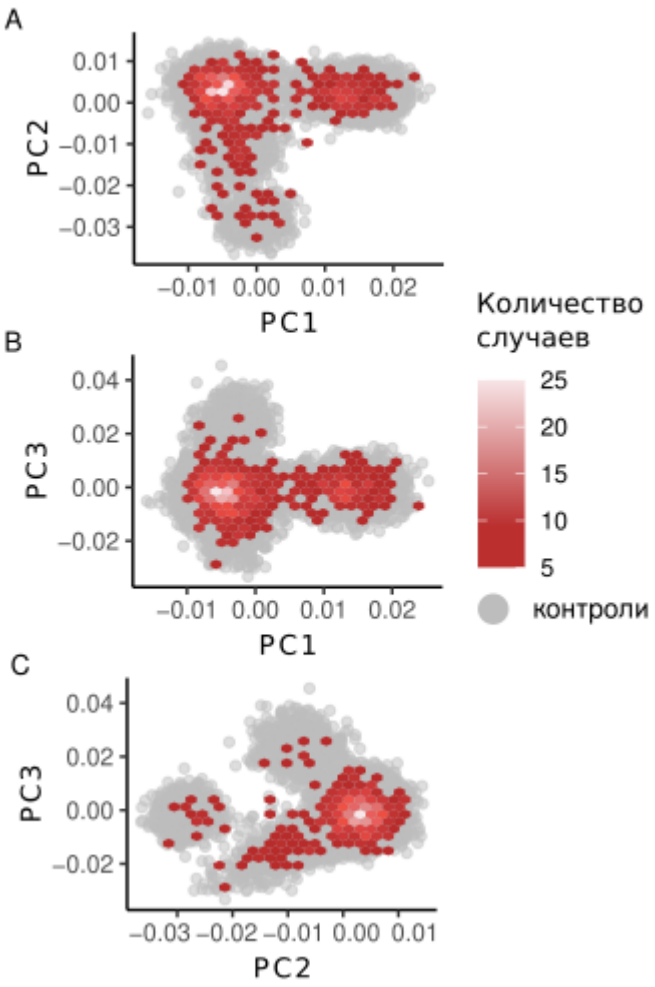

Рисунок 16. Кластеризация популяционных данных. Цвет отражает наличие случаев для каждого квантиля данных: A) проекция компонент PC1 и PC2; B) проекция компонент PC1 и PC3; C) проекция компонент PC2 и PC3.

В дальнейшем анализе участвовали четыре мажорных кластера, три из которых принадлежали европейской популяции и были объединены в один кластер, а четвертый представлял лиц африканского происхождения (рис. 17, 18).

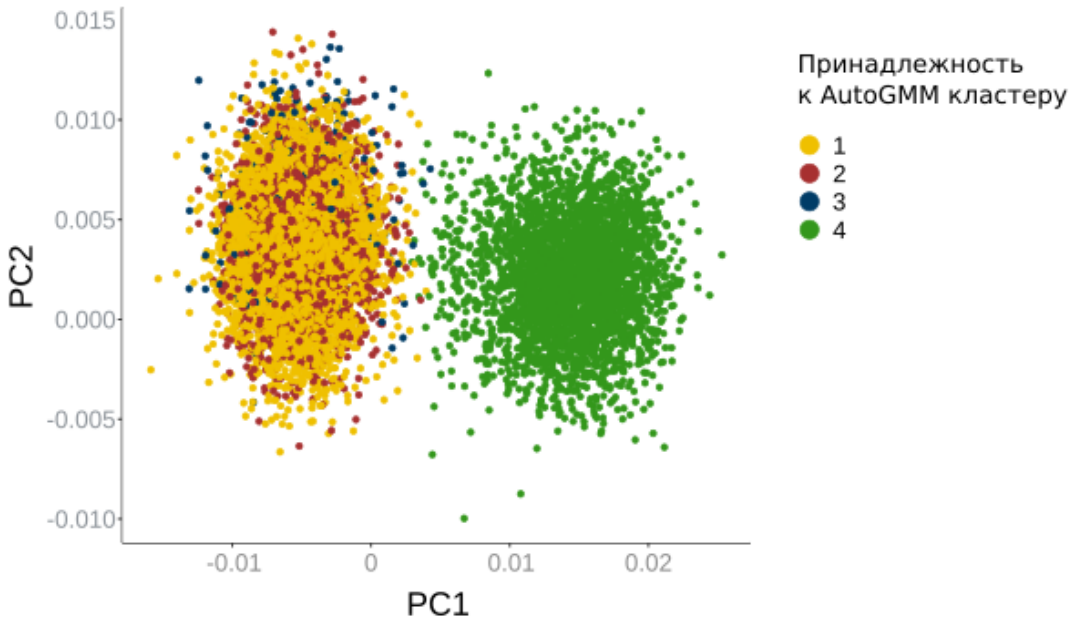

Рисунок 17. Кластеризация. Цвет точек соответствует принадлежности кластерам, назначенным пакетом AutoGMM.

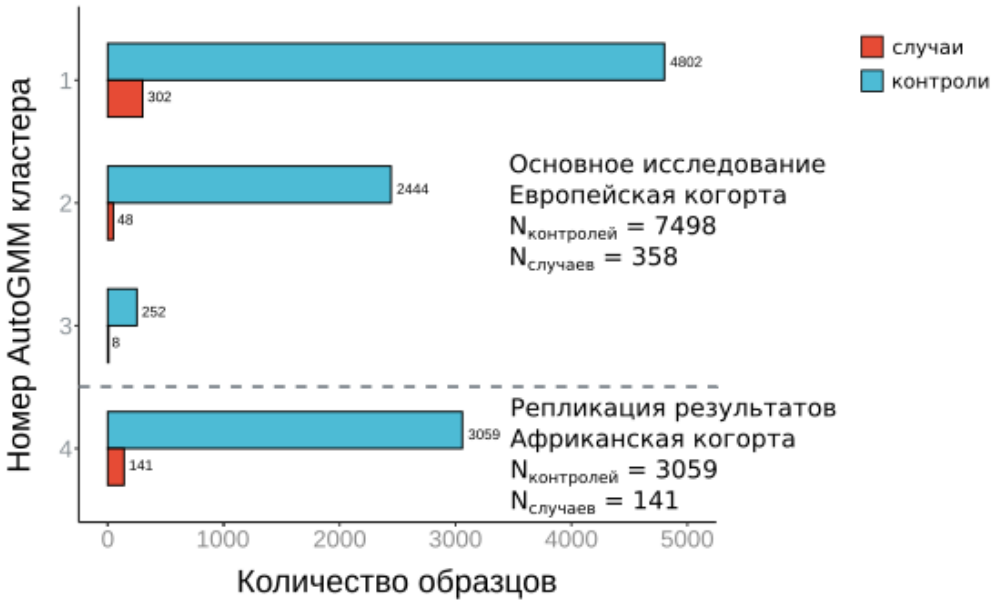

Рисунок 18. Кластеризация. Состав случаев и контролей для каждого кластера.

Дальнейшее сопоставление «случай-контроль» было проведено с помощью пакета Matchit [545] (рис. 19 – 21). Окончательный набор данных составил 358 случаев и 1 466 контролей для европейского кластера и 141 случай и 595 контролей для африканского кластера.

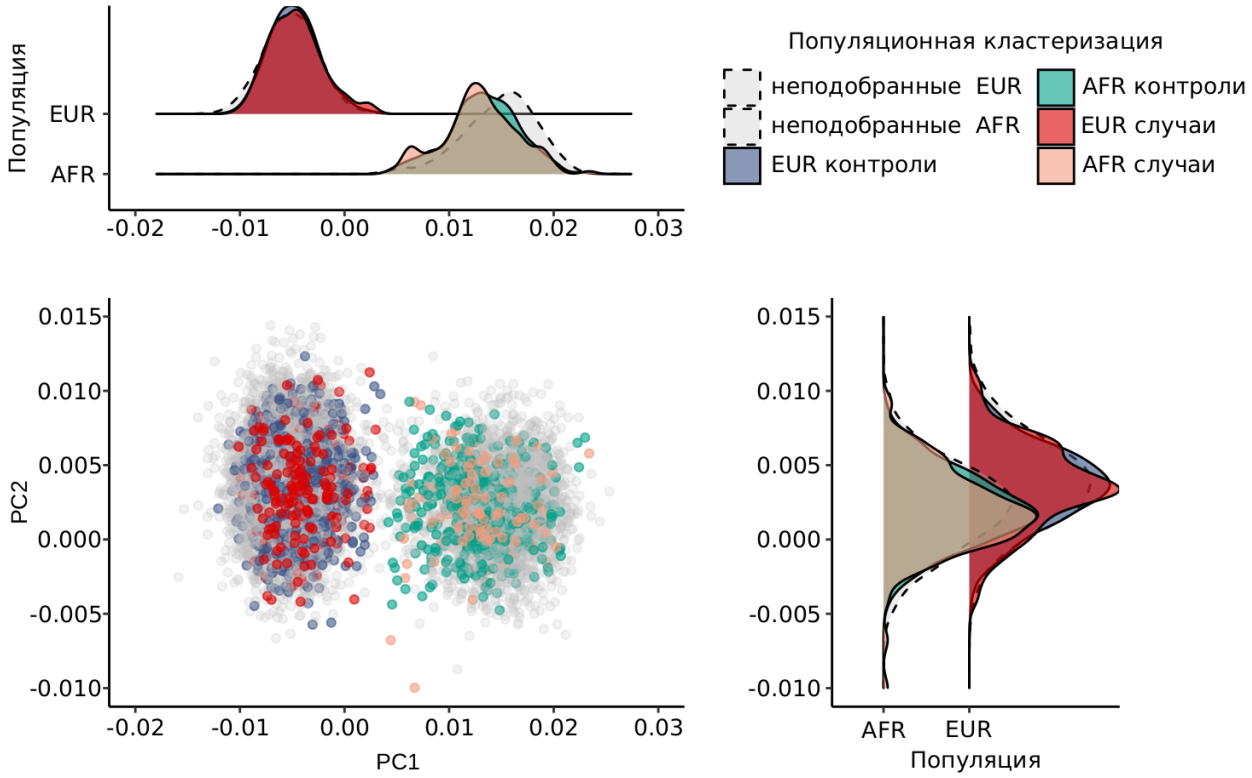

Рисунок 19. PCA, иллюстрирующий уровень соответствия случай-контроль в европейской и африканской популяциях

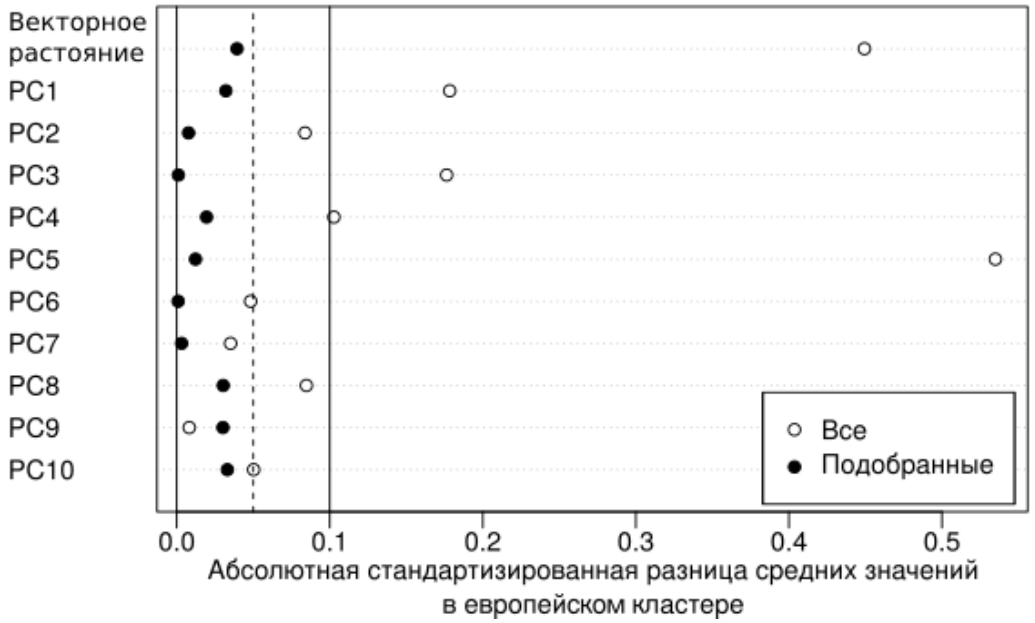

Рисунок 20. Вклад главных компонент в дисперсию между случаями и контролями для европейского кластера

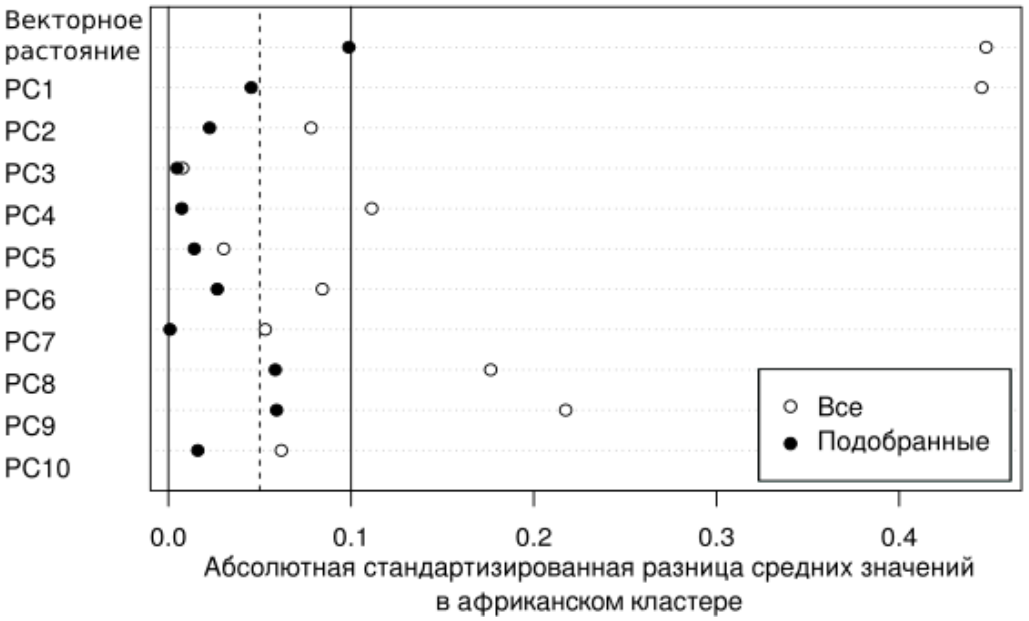

Рисунок 21. Вклад главных компонент в дисперсию между случаями и контролями для африканского кластера

Анализ мощности европейского набора данных показал многократное превосходство над другими когортными исследованиями ФСГС, а также то, что при существующем количестве случаев значимое увеличение мощности не может быть достигнуто при большем количестве контролей (рис. 22, 23).

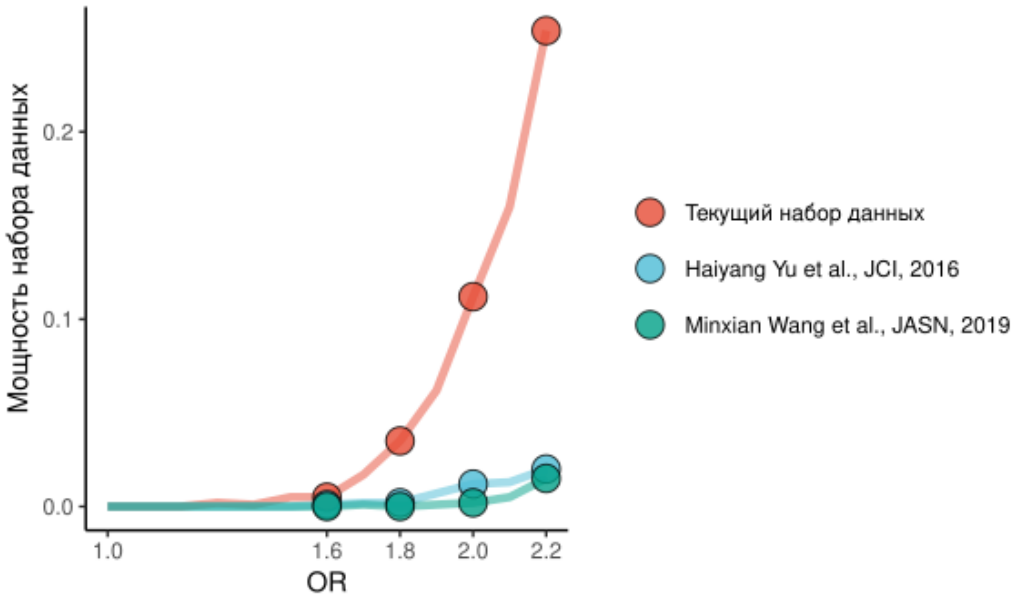

Рисунок 22. Анализ мощности для европейского кластера с использованием сопоставленных случаев и контролей. Были использованы следующие параметры для сравнения (текущее исследование: 356 случаев, 1 466 контролей, 2 482 гена; Haiyang Yu et al, JCI, 2016: 179 случаев, 378 контролей, 2 482 гена; Minxian Wang et al, JASN, 2019: 363 случая, 363 контроля, 19 000 генов). Для исследования JASN не было возможности получить точное количество генов, которые было непосредственно включены в анализ, поэтому использовано стандартное количество генов для полноэкзомного секвенирования.

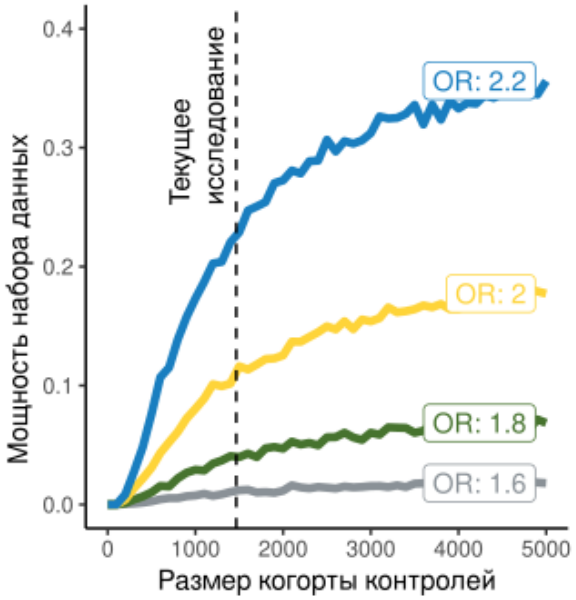

Рисунок 23. Анализ мощности при фиксированном числе случаев и различном числе добавляемых контролей

Далее было проведено ассоциативное исследование с использованием синонимичных CV (gnomAD популяционный специфический AF≥0.01). Предполагается, что синонимичные варианты не смогут повлиять на образование фенотипа, таким образом любое отклонение между когортами случаев и контролей будет свидетельствовать о наличие технических неполадок в анализе, например, о присутствие родственных образцов в данных. Так, для европейских и африканских наборов данных после сопоставления было подтверждено отсутствие систематического смещения между группами случаев и контролей (рис. 24, 25).

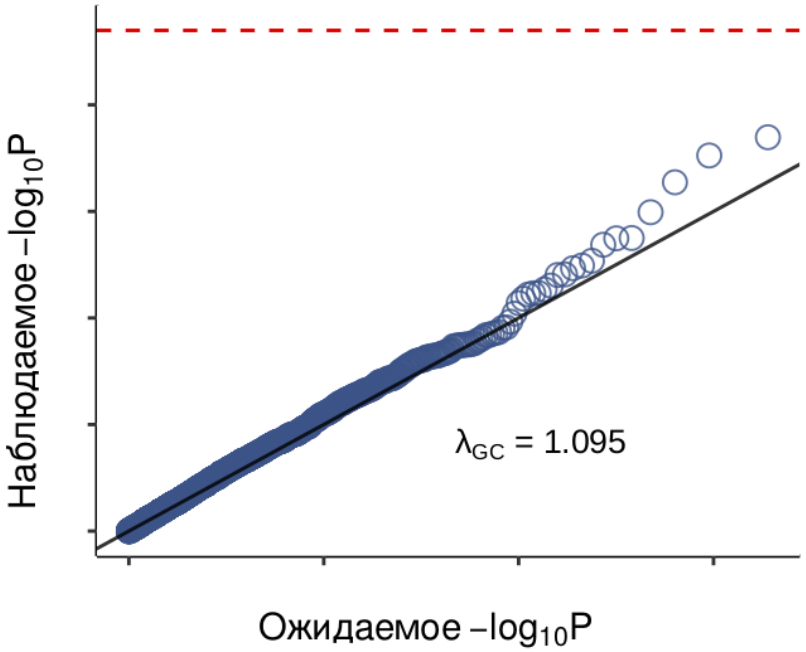

Рисунок 24. QQ–графики для исследования ассоциации распространенных (gnomAD population specific AF≥0.01) синонимичных вариантов, иллюстрирующие степень соответствия случай–контроль для европейской популяции.

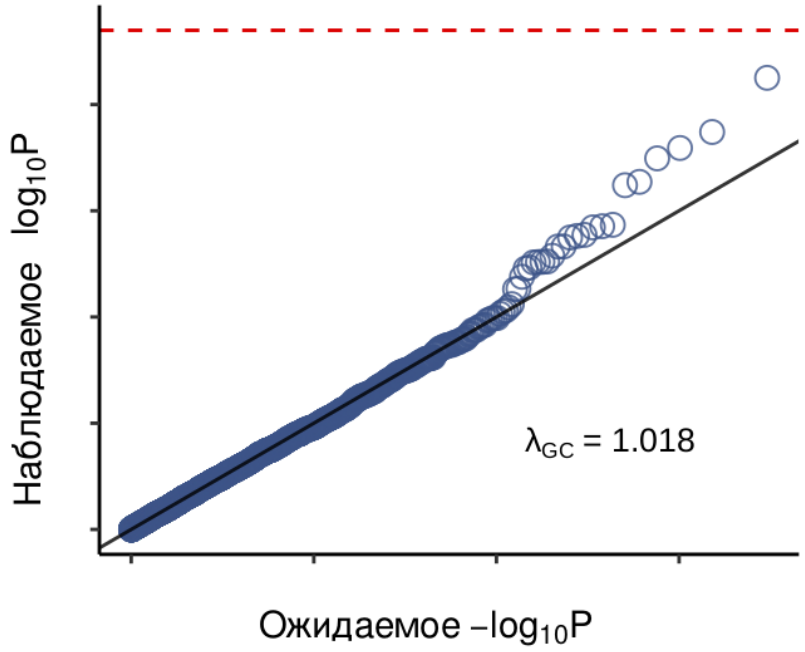

Рисунок 25. QQ–графики для исследования ассоциации распространенных (gnomAD population specific AF≥0.01) синонимичных вариантов, иллюстрирующие степень соответствия случай–контроль для африканской популяции.

3.3.3. Анализ вариантов на наличие ассоциации с риском возникновения фокального сегментарного гломерулосклероза

Далее была проведена серия из нескольких ассоциативных исследований с использованием миссенс- и PTV-вариантов. Во-первых, было проведено классическое повариантное ассоциативное исследование с применением для каждого ДНК-варианта линейной регрессионной модели. В регрессионной модели противопоставляется число случаев с 0, 1 и 2 альтернативными аллелями аналогичным показателям контрольной группы. В европейском кластере два варианта оказались значимыми (rs601314 – $p=8.1\times10^{-9}$; соотношение шансов для минорной аллели OR=13.24; миссенс эффект; ген – *EFEMP2*; и rs117071588 – $p=4.0\times10^{-6}$; соотношение шансов для минорной аллели OR=11.66; миссенс; *CCDC82*; рис. 26, табл. 3).

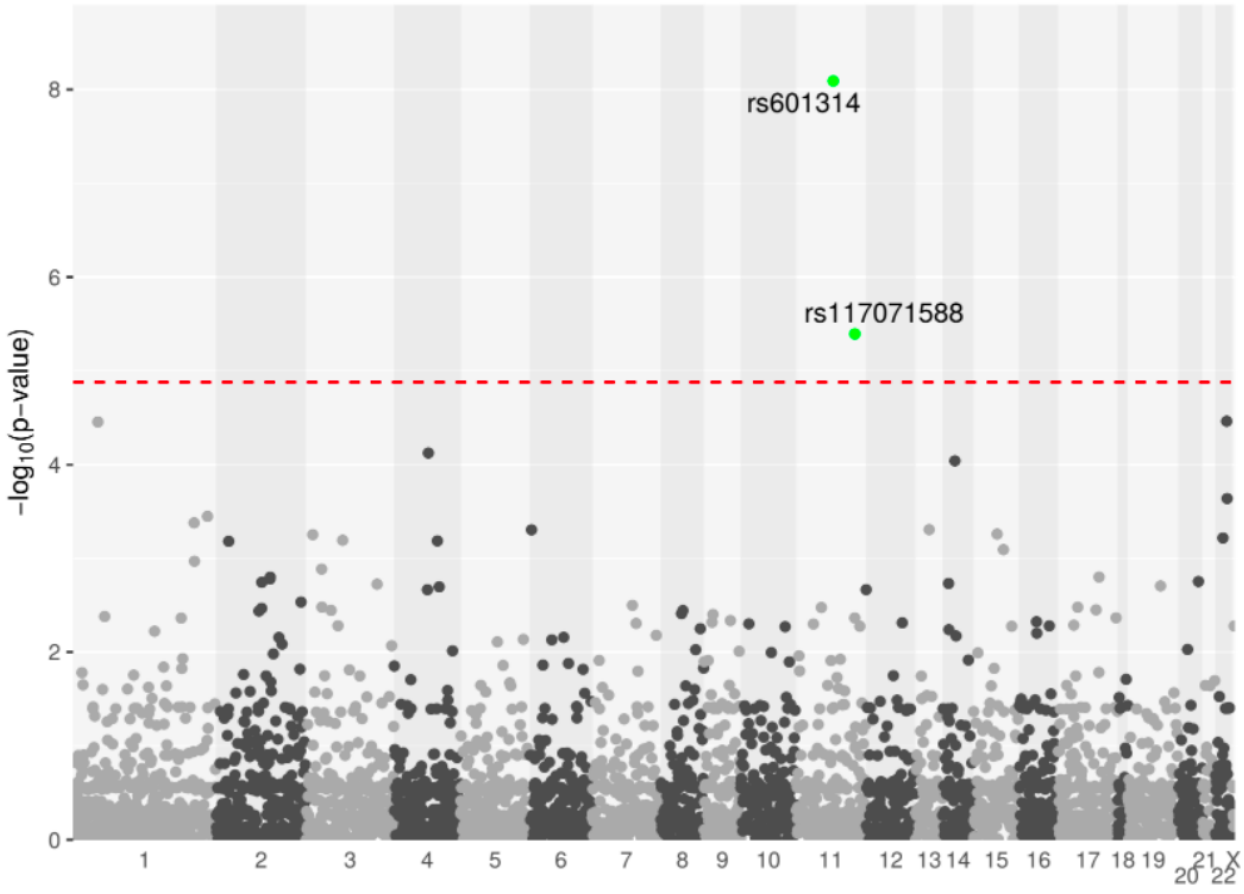

Рисунок 26. Повариантный ассоциативный анализ в европейском кластере

Таблица 3. Топ-20 результатов повариантного ассоциативного анализа в европейской популяции

			Популяционные частоты в gnomAD		N генотипов в случаях			N генотипов в контролях			Результат линейной регрессии	
Локус	ID в dbSNP	Ген	NFE	AFR	Реф. гом.	Гетер.	Альт. гом.	Реф. гом.	Гетер.	Альт. гом.	beta	pvalue
1:65636053	rs601314	EFEMP2	0.998	0.6702	0	12	331	0	4	1462	-0.5653931958	0.00000000810060
11:96117537	rs117071588	CCDC82	0.002398	0.0006788	325	8	0	1439	3	0	0.5430323645	0.00000403510515
2:36661906	rs73885319	APOL1	0.00011	0.23	347	8	0	1460	4	0	0.4746356761	0.0000344611979
1:23415518	rs12066671	LUZP1	0.0008999	0.2251	348	8	0	1462	4	0	0.4744014733	0.00003504848976
4:111397664	rs72890294	ENPEP	0.0003102	0.09177	350	5	0	1465	1	0	0.6404958678	0.00007510445923
14:55609418	rs10148371	LGALS3	0.0002472	0.09487	348	5	0	1426	1	0	0.6371664788	0.00009103490109
22:36662034	rs60910145	APOL1	0.0000708	0.2316	343	7	0	1425	4	0	0.4423591115	0.0002299533263
1:230840034	rs1805090	AGT	0.002049	0.0006768	351	6	0	1462	3	0	0.4730649016	0.0003563533784
1:207782856	rs17047660	CR1	0.001037	0.2445	348	6	0	1423	3	0	0.4701675136	0.0004179665537
13:45147305	rs144252895	TSC22D1	0.00254	0.0002462	349	5	0	1464	2	0	0.5217870932	0.0004931414383
6:10709581	rs34857240	PAK1IP1	0.0001055	0.05538	340	4	0	1464	1	0	0.6115299335	0.0004955719624
15:54008845	rs690346	WDR72	0.9997	0.869	0	4	345	0	1	1464	-0.6092868988	0.0005506053588
3:12861608	rs3817121	CAND2	0.1725	0.03601	255	72	7	946	430	49	-0.05977747606	0.0005584768315
22:36556768	rs11089781	APOL3	0.0002902	0.2224	346	5	1	1461	4	0	0.3764680851	0.0006060539975
3:89521664	rs17801309	EPHA3	0.08851	0.01834	306	37	1	1188	262	12	-0.07830302434	0.0006410183188
4:126373653	rs17009684	FAT4	0.00008814	0.07663	353	4	0	1463	1	0	0.6056167401	0.0006508991499
2:46574131	rs149898744	EPAS1	0.0006075	0.0000615	354	4	0	1465	1	0	0.6053875756	0.0006575620282
15:72638892	rs1800431	HEXA	0.9982	0.6041	1	9	340	0	12	1453	-0.2656870178	0.0008070888848

1:207782889	rs17047661	CR1	0.003016	0.6293	340	12	0	1411	12	1	0.2479001876	0.001072576324
3:46007823	rs13079478	FYCO1	0.1113	0.01655	297	40	3	1154	280	19	-0.06891267478	0.001301400058

Попытка реплицировать эти варианты в африканской когорте не привели к значимому результату (rs601314 – p=0.056, OR=1.35; rs117071588 – отсутствует в африканском наборе данных).

Анализ RV проводился с использованием миссенс- и PTV-вариантов с популяционной частотой менее 0.01. Отсечку выбирали по наличию сигнала в каждом из квартилей распределения частоты аллелей в интервале [0; 0.01] (рис. 27).

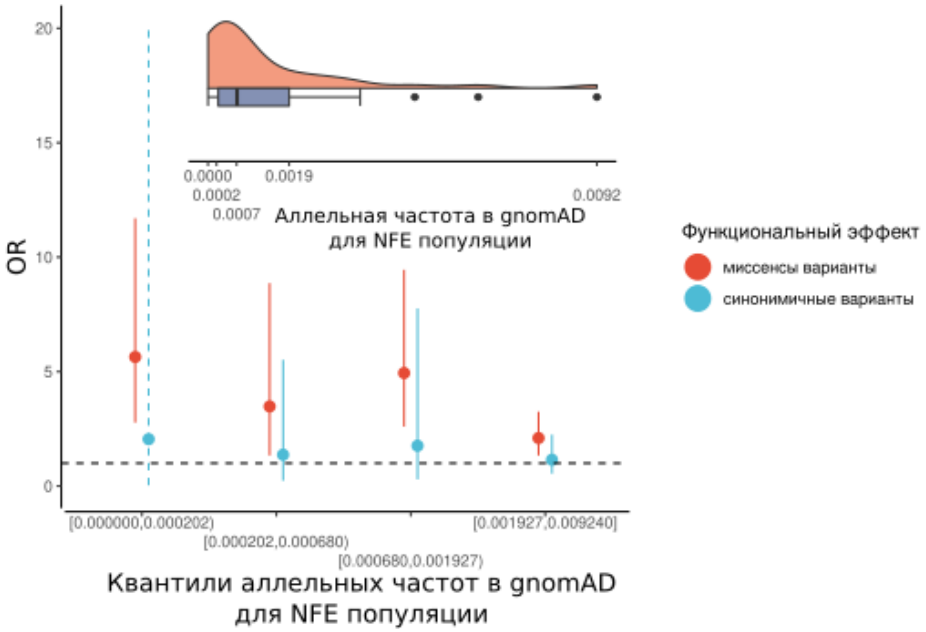

Рисунок 27. Распределение силы сигнала ассоциации между известными ассоциированными генами ФСГС - *KANK1*, *COL4A4*, *WNK4*, *APOL1*, *IL36G* по квантилям аллельных частот в gnomAD для NFE популяции.

RVAS проводился с использованием пяти тестов, представляющих различные статистические классы методов для каждого гена (точный тест Фишера, C-альфа, ASUM, KBAC), чтобы охватить все потенциальные модели риска (рис. 28).

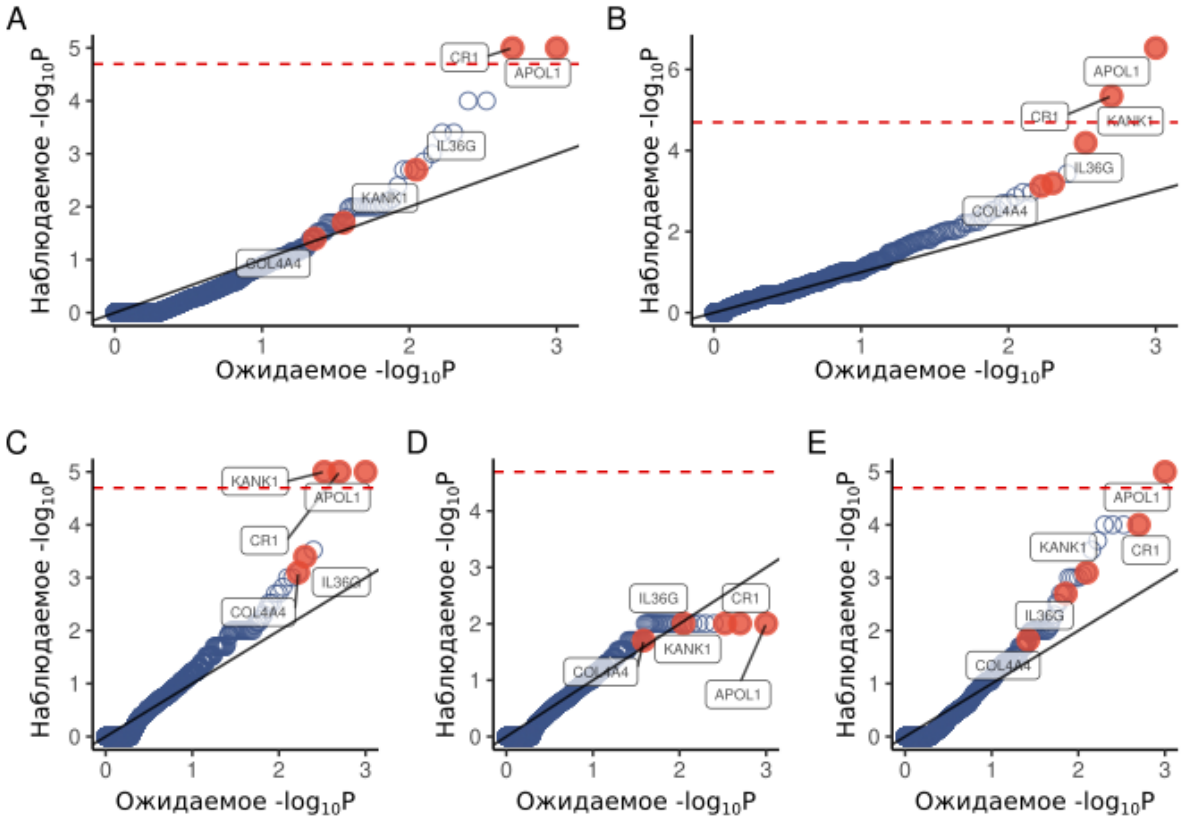

Рисунок 28. QQ-график для каждого из тестов на редкие варианты, включенных в агрегирующий метод Саймса. (A) C-alpha тест; (B) точный тест Фишера; (C) WSS тест; (D) KBAC тест; (E) ASUM тест.

Полученные p-значения были объединены с помощью метода Саймса, подходящего для объединения статистики зависимых тестов (табл. 4).

Таблица 4. Топ-20 результатов анализа редких вариантов в NFE популяции

Ген	число минорных аллелей		частота минорной аллели		OR	p.value статистических тестов, входящих в метод Саймса							Simes (метод Саймса)
	Случаи	Контроли	Случаи	Контроли		Лин. регресс.	тест Фишера по носителями	тест Фишера по аллелям	C-alpha	KBAC	ASUM	WSS	
APOL1	22	17	0.030726	0.0057980	5.432	0.0000000062	0.0000002941	0.000000391	0.00001	0.00990	0.00001	0.00001	0.000001470
CR1	22	23	0.030726	0.0078444	4.073	0.0000011988	0.0000045391	0.000009377	0.00001	0.00990	0.0001	0.00001	0.000016666
KANK1	50	105	0.069832	0.0358117	2.021	0.0000916226	0.0000644077	0.000223805	0.02	0.00990	0.0008	0.00001	0.00005
ENPEP	6	2	0.008379	0.0006821	12.36	0.0000751101	0.0010771691	0.001081530	0.0001	0.00990	0.0001	0.01	0.00025
LGALS3	6	2	0.008379	0.0006821	12.36	0.0000790770	0.0010771691	0.001107781	0.0001	0.00990	0.0001	0.0015	0.00025
CCDC82	15	27	0.020949	0.0092087	2.302	0.0051814834	0.0161830403	0.009198897	0.0004	0.00990	0.0001	0.01	0.0005
PLA2R1	30	55	0.041899	0.0187585	2.287	0.0002141506	0.0003717308	0.000487961	0.004	0.00990	0.001	0.0003	0.000929327
EPAS1	12	26	0.016759	0.0088676	1.905	0.0683492987	0.0585002411	0.066988654	0.11	0.02970	0.0002	0.06	0.001
LUZP1	11	14	0.015363	0.0047748	3.251	0.00155504963	0.0045006347	0.003891977	0.0004	0.00990	0.0003	0.005	0.001
IL36G	9	6	0.012569	0.0020463	6.204	0.00002891317	0.0006410395	0.000393873	0.002	0.00990	0.002	0.0004	0.001602598
COL4A4	26	46	0.036312	0.0156889	2.363	0.00033107807	0.0007589016	0.000858131	0.04	0.01980	0.015	0.0008	0.002
XYLT1	23	43	0.032122	0.0146657	2.230	0.00215405231	0.0019477991	0.002914233	0.01	0.00990	0.002	0.001	0.003333333
KIAA0226	28	53	0.039106	0.0180763	2.210	0.00081540161	0.0013794614	0.001834892	0.008	0.00990	0.01	0.001	0.003448653
SLC6A13	6	4	0.008379	0.0013642	6.182	0.00127005733	0.0056410854	0.005714385	0.0014	0.01980	0.001	0.03	0.0035
AGT	13	28	0.018156	0.0095497	1.917	0.04935119379	0.0701915829	0.072542577	0.01	0.02970	0.0009	0.01	0.0045
PTPN1	14	19	0.019553	0.0064802	3.056	0.00090175969	0.0027851746	0.002955042	0.001	0.00990	0.006	0.002	0.004641957
CLU	11	15	0.015363	0.0051159	3.033	0.00333673015	0.0097216668	0.010024518	0.04	0.00990	0.001	0.004	0.005

MPRIP	8	7	0.011173	0.0023874	4.719	0.00148411034	0.0036045594	0.004732187	0.01	0.00990	0.002	0.002	0.005
PAX4	8	6	0.011173	0.0020463	5.507	0.00052173294	0.0020240154	0.002435736	0.007	0.00990	0.001	0.02	0.005
CRYAB	5	3	0.006983	0.0010231	6.861	0.00225325376	0.0094218712	0.009583358	0.002	0.00990	0.003	0.013	0.0075

В топ-10 ассоциированных генов вошли *APOL1*, *KANK1*, *COL4A4* и *IL36G*, ранее отмеченные в исследованиях ассоциаций с ФСГС. Два гена достигли значимости после коррекции Бонферрони ($P=0.05/2482=2.015\times10^{-5}$) – *APOL1* ($P=1.47\times10^{-6}$), известный ген предрасположенности к ФСГС, и *CR1* ($P=1.67\times10^{-5}$), новый ген кандидат (рис. 29).

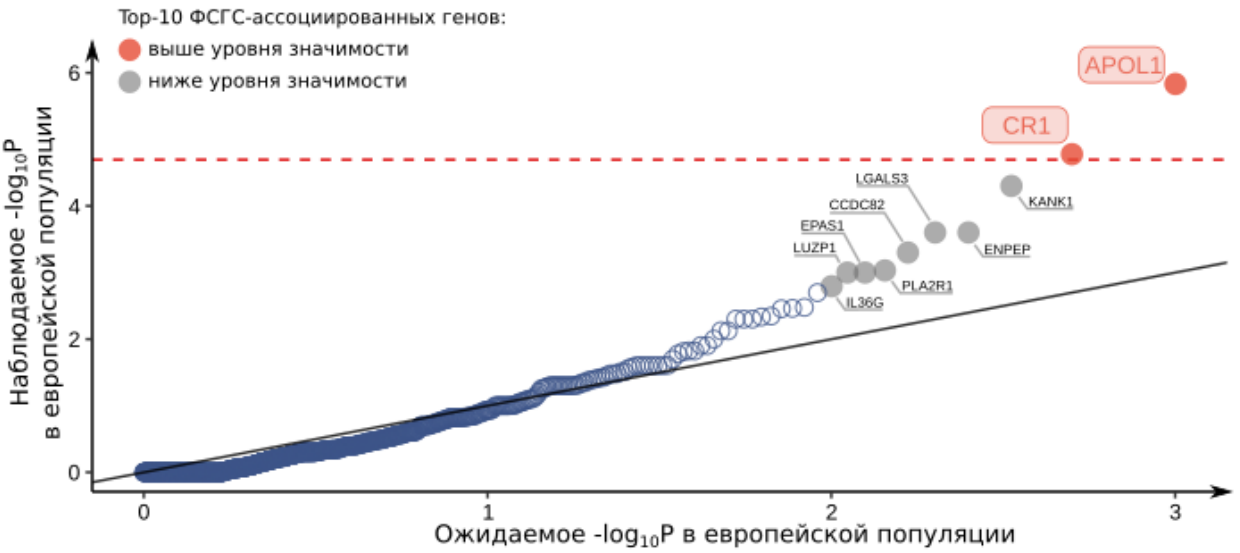

Рисунок 29. Исследование ассоциации редких вариантов в европейском кластере (gnomAD_EUR_AF < 0,01; миссенс- и PTV-варианты; метод Саймса - точный тест Фишера, C-альфа, ASUM, WSS, KBAC);

Ассоциированные гены в европейской когорте повторно проверялись на репликацию в когорте африканского происхождения (табл. 5). Ни *APOL1*, ни *CR1* не были воспроизведены с помощью анализа RV. Ранее наблюдавшийся положительный отбор, действующий на варианты *APOL1* [546] в африканской популяции, позволяет предположить, что варианты риска ФСГС могут быть слишком распространены, чтобы попасть в RVAS в африканской популяции. Поэтому далее был использован повариантный анализ для воспроизведения сигнала ассоциации в *APOL1* и *CR1*.

Таблица 5. Попытка репликации *APOL1* и *CR1* в африканской популяции в анализе редких вариантов

Ген	Число минорных аллелей		Частота минорного аллеля		OR	p.value статистических тестов, входящих в метод Саймса							Simes (метод Саймса)
	Случаи	Контроли	Случаи	Контроли		Лин. регресс.	Тест Фишера по носителями	Тест Фишера по аллелям	C-alpha	KBAC	ASUM	WSS	
APOL1	93	296	0.129888	0.1009549	1.485741	0.019588784	0.014785973	0.014917479	0.008	0.03960	0.017	0.02	0.025
CR1	2	11	0.002793	0.0037517	0.765714	0.695803170	1	1	0.27	0.83168	0.82	0.64	1

Сначала были определены RV в европейской когорте, определяющие сигнал ассоциации в RVAS для генов *APOL1* и *CR1* (рис. 30). Было определено четыре варианта: 1) парный локус G1 в *APOL1* – rs60910145 и rs7388531918; 2) два близких варианта в *CR1* – rs17047661 и rs17047660. Все четыре варианта прошли порог значимости при введении поправки на множественную проверку гипотезы (p=0.05/10 вариантов=0.005). Далее эти четыре варианта приняли участие в репликации на образцах из африканской когорты. Оба варианта из пары G1 *APOL1* и rs17047660 в *CR1* успешно прошли порог значимости репликации (p=0.05/4 = 0.0125) (табл. 6).

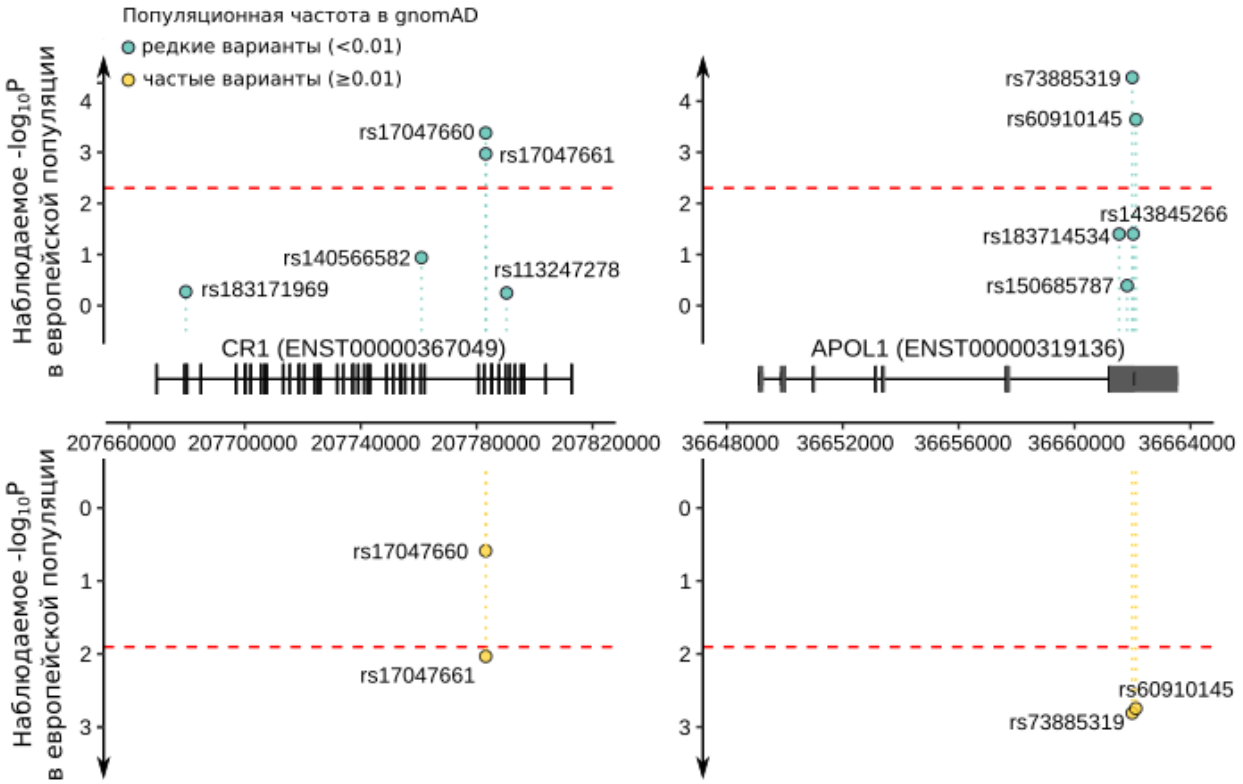

Рисунок 30. Результаты репликации между европейской и африканской популяцией.

Таблица 6. Репликация *APOL1* и *CR1* в повариантном тесте

Гены и варианты, вошедшие в репликацию		Результаты ассоциативного исследования в NFE поп.		Репликация результатов ассоциативного исследования в AFR поп.		Популяционные аллельные частоты в gnomAD	
Ген	ID в dbSNP	p.value	OR	p.value	OR	NFE	AFR
APOL1	rs73885319	0.0004759	0.1201409	0.001779	0.6275252	0.00008794	0.2315
APOL1	rs60910145	0.001744	0.1388619	0.002271	0.632818	0.00007081	0.2316
CR1	rs17047661	0.002079	0.2850085	0.008714	1.434805	0.003016	0.6293
CR1	rs17047660	0.002936	0.1233033	0.2755	1.204588	0.001037	0.2385

Частоты аллелей для реплицированных вариантов значительно отличаются между популяциями, что свидетельствует о положительном отборе (rs60910145: gnomAD_NFE_AF=8.6×10^{-5}, gnomAD_AFR_AF=0.23; rs73885319: gnomAD_NFE_AF=1.1×10^{-4}, gnomAD_AFR_AF=0.23; rs17047661: gnomAD_EUR_AF=3.0×10^{-3}, gnomAD_AFR_AF=0.62).

3.3.4. Плейотропия как объяснение высокой частоты аллелеи в популяции

Ген *CR1* является важным участником системы комплемента. В соответствии с гипотезой о возможной роли иммунной системы в возникновении ФСГС у некоторой группы пациентов, было решено выяснить, связан ли статус носительства rs17047661 с каким-либо типом HLA. Для этого был проведен анализ с использованием данных 1000 геномов, в которых была доступна информация об HLA-типировании. Образцы были разделены на те, которые являются носителями определенного HLA-генотипа, и те, которые не являются. Между группами производилась оценка наличия аллелей rs17047661. Пять номинально значимых ассоциаций в европейской популяции были воспроизведены у африканцев. Наиболее обогащенным HLA-генотипом как в европейской, так и в африканской популяциях был HLA-DRB*15:01 (P=0.0264; P=0.0942; соответственно) (табл. 7, полная таблица доступна в Приложении).

Таблица 7. Топ-10 результатов анализа на присутствие rs17047661 в конкретном HLA-типе. Полная версия в Приложении

HLA тип	генотип	p.value в AFR	pvalue в NFE
HLA-DQB1	06:02	0.4079437114	0.005091652191
HLA-A	29:02	0.4967179612	0.02691250052
HLA-DRB1	15:01	0.09424731899	0.02700897226
HLA-A	02:01/30	1	0.0353949089
HLA-C	16:01	0.2366706976	0.04319061145
HLA-DQB1	03:01	0.5391910071	0.05579866535
HLA-DRB1	04:05	0.129983287	0.07706973155
HLA-B	44:03	0.3429098007	0.09234732559
HLA-DRB1	11:04	0.1123856918	0.09981746544
HLA-A	03:01	0.5149344467	0.1219476206

По совпадению, этот HLA-тип особенно обогащен в регионах с высоким уровнем заболеваемости малярией, что согласуется с известным защитным эффектом FSGS против этого заболевания (рис. 31).

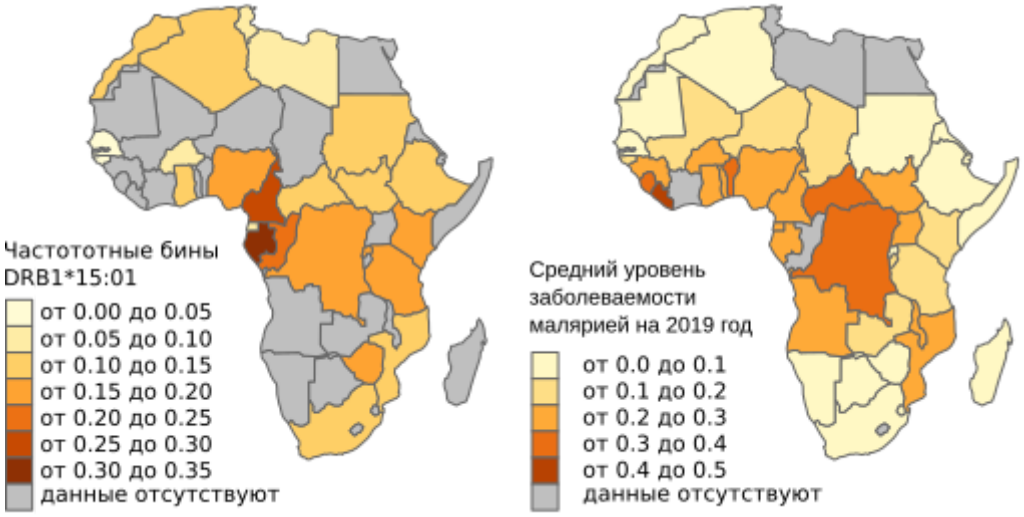

Рисунок 31. Распределение HLA-типа DRB1*15:01(слева) и малярии (справа) на африканском континенте. Данные по DRB1*15:01 и малярии были получены с сайтов http://www.allelefrequencies.net/ и https://malariaatlas.org/ соответственно.

Ранее[обз] высказанная гипотеза об иммунном компоненте в ФСГС находит новые подтверждения благодаря обнаруженному ассоциированному гену *CR1*. Степень достоверности ассоциации увеличивается, так как результаты реплицированы в двух независимых глобальных популяциях. Недавние исследования уже описали влияние системы комплемента [547] при различных гломерулопатиях. Аутоантитела, реагирующие на экспрессируемые почками аутоантигены или комплексы антитело/антиген, оседающие в почках, считаются возбудителями различных заболеваний почек человека. Известны случаи С1-опосредованного воспаления и отложения С3 [548]. Также было показано, что ингибирование С3 снижает протеинурию в животных моделях [549]. Что касается конкретно ФСГС, то в пораженных гломерулах часто наблюдаются отложения IgG и С3, но патогенез до сих пор неясен, а терапия против системы комплемента не изучена [550].

CR1 является негативным регулятором, влияющим на активацию С3 и уменьшающим его отложение. С3 - центральный элемент системы комплемента, который может активироваться либо по классическому иммунному пути с инициацией IgM и IgG, либо по альтернативному пути неспецифического связывания с антигенами на мембранах или с остатками маннозы по лектиновому пути. При классическом и лектиновом путях *CR1* обладает так называемой активностью активирующей распад комплекса, то есть, связывая С4b, он предотвращает образование С3-конвертазы, которая, в свою очередь, должна активировать С3 путем расщепления на фрагменты С3а и С3b. Альтернативный

путь предполагает, что *CR1* затем действует как кофактор для расщепления активных фрагментов C3b (на фрагменты C3c и C3dg), значительно уменьшая осаждение фрагментов C3b, которые могли бы реактивировать C3. Фрагмент C3b, являясь опсонином как IgG, открывает группировку, способную образовывать эфирную связь с OH-группами клеточной мембраны, благодаря чему C3b фиксируется на клеточной мембране бактерии, что повышает вероятность поглощения нейтрофилами и макрофагами.

CR1 значительно снижает отложение C3b, примерно на 80 % по сравнению с классическим путем, но наилучший эффект достигается при активации альтернативного пути (более 95 % снижения отложения C3b) [548].

Мы предположили, что эволюционное давление на *CR1*, связанное с защитой от малярийной инфекции, приводит к гораздо большей распространенности идентифицированных вариантов *CR1* у лиц африканского происхождения по сравнению с европейцами. Выдвинутая гипотеза, что *CR1* является рецептором эритроцитов, используемым *Plasmodium falciparum* для инвазии в клетку, нашла подтверждение в ранних работах [551,552]. Усиленное действие *CR1* в альтернативном пути активации системы комплемента, включающем активацию врожденного иммунного ответа, также согласуется с этой гипотезой. Более того, было показано, что вариант rs17047661 приводит к аминокислотной замене R1601G, тем самым защищая от церебральной малярии в популяциях Африки к югу от Сахары [552]. Такой эффект приводит к резкому различию в частоте аллелей этого варианта между европейскими и африканскими популяциями и, как следствие к возникновению плеотропности.

Известно, что гены, вовлеченные в регуляцию иммунитета, обладают высокой степенью плейотропии. В связи с этим, в завершение данного исследования, был проведен анализ плейотропии *CR1* с помощью ресурса PheWeb. Данными послужил тот же самый набор фенотипов из лаборатории Бенджамина Нила, который использовался в предыдущем анализе плейотропии [494]. Имеющиеся результаты позволяют заключить, что CR1 является характерным примером плейотропного гена, что вместе с влиянием естественного отбора, объясняет дисбаланс в частотах аллели между европейской и африканской популяцией (табл. 8).

Таблица 8. Топ-10 фенотипов, ассоциированных с вариантами в *CR1* в UK Biobank

p.value	Фенотип, указанный в базе PheWeb (Набор данных – UKBiobank Neale v1)
1.4×10^{-56}	Underlying (primary) cause of death: ICD10: K80.2 Calculus of gallbladder without cholecystitis

9.8×10^{-43}	Pulse rate, automated reading
$1.3x10^{-35}$	Surgery/amputation of toe or leg: Yes, leg below the knee
1.9×10^{-34}	Underlying (primary) cause of death: ICD10: K26.4 Chronic or unspecified with haemorrhage
9.5×10^{-33}	Underlying (primary) cause of death: ICD10: C14.0 Pharynx, unspecified
7.6×10^{-28}	Underlying (primary) cause of death: ICD10: C73 Malignant neoplasm of thyroid gland
3.1×10^{-22}	Underlying (primary) cause of death: ICD10: C55 Malignant neoplasm of uterus, part unspecified
5.2×10^{-22}	Underlying (primary) cause of death: ICD10: I71.1 Thoracic aortic aneurysm, ruptured
9.5×10^{-22}	Underlying (primary) cause of death: ICD10: C18.1 Appendix
1.1×10^{-21}	Underlying (primary) cause of death: ICD10: J69.0 Pneumonitis due to food and vomit

ЗАКЛЮЧЕНИЕ

Подводя итог, следует отметить, что многие заболевания имеют наследственный компонент, состоящий из нескольких частотных групп вариантов ДНК. Обычно это связано с характеристиками заболевания у данного пациента. RV обычно приводят к тяжелым семейным проявлениям, а сочетания частых вариантов приводят к спорадическим формам заболевания с более поздней манифестацией.

В первой части исследования был показан ожидаемый частотный профиль для аллелей с сильным дезадаптирующим действием на примере российской когорты. Однако, далее было выдвинуто предположение о том, что плейотропия может способствовать закреплению аллелей с большим размером эффекта в популяции на более высоких частотах. На втором этапе исследования было выяснено, что плейотропные варианты имеют склонность быть частыми из-за чего вносят значительный вклад в риск развития полигенных заболеваний.

Тезис о том, что плейотропия может способствовать закреплению сильно дезадаптирующих аллелей в популяции был продемонстрирован в заключительной части исследования на примере ФСГС. Особенность этиологии заболевания, приводит к тому, что частота аллели риска значимо меняется в зависимости от популяции и условий среды проживания. Таким образом, плейотропный эффект для частых вариантов у африканской популяции можно наблюдать в европейской популяции, где те же варианты имеют редкую популяционную частоту.

С учетом этих выводов необходимо принимать во внимание особенности популяционной структуры и использовать это при подборе контрольных групп. Такая практика позволяет улучшить корректность ассоциативных исследований и расширить возможности для поиска новых ассоциаций в независимых когортах.

ВЫВОДЫ

1. В результате оценки частотного спектра аллелей, связанных с носительством аутосомно-доминантных наследственных заболеваний у жителей Северо-Западного региона Российской Федерации было выяснено, что данная когорта обладает собственными рисками носительства патогенных аллелей по сравнению с европейской популяцией;

2. На примере когорты UK Biobank было показано, что для локусов с множественной фенотипической ассоциацией характерно повышенное число частых аллелей, что может быть объяснено давлением очищающего отбора;

3. На основе когортного исследования, посвященного фокальному сегментарному гломерулосклерозу, проанализированы межпопуляционные риски наследственных заболеваний, связанные с носительством конкретных аллелей. Также выявлены потенциальные различия в распространенности и воздействии этих аллелей в европейских и африканских популяциях. Полученные данные указывают на то, что плейотропия может представлять собой важный фактор, предсказывающий наличие дисбаланса аллельной частоты при определении межпопуляционных рисков.

СПИСОК ЛИТЕРАТУРЫ

1. Wray N.R., Goddard M.E., Visscher P.M. Prediction of individual genetic risk to disease from genome-wide association studies. // Genome Res. 2007. Vol. 17, № 10. P. 1520–1528.
2. Mathur S., Sutton J. Personalized medicine could transform healthcare. // Biomed. Rep. 2017. Vol. 7, № 1. P. 3–5.
3. van Schaik R.H.N. CYP450 pharmacogenetics for personalizing cancer therapy. // Drug Resist. Updat. 2008. Vol. 11, № 3. P. 77–98.
4. Klomp S.D. et al. Phenoconversion of cytochrome P450 metabolism: A systematic review. // J. Clin. Med. 2020. Vol. 9, № 9.
5. Lee M.T.M., Klein T.E. Pharmacogenetics of warfarin: challenges and opportunities. // J. Hum. Genet. 2013. Vol. 58, № 6. P. 334–338.
6. Barton N.H., Etheridge A.M., Véber A. The infinitesimal model: Definition, derivation, and implications. // Theor. Popul. Biol. 2017. Vol. 118. P. 50–73.
7. Rivas M.A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. // Nat. Genet. 2011. Vol. 43, № 11. P. 1066–1073.
8. Surendran P. et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. // Nat. Genet. 2016. Vol. 48, № 10. P. 1151–1161.
9. Huyghe J.R. et al. Discovery of common and rare genetic risk variants for colorectal cancer. // Nat. Genet. 2019. Vol. 51, № 1. P. 76–87.
10. Sarnowski C. et al. Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. // Am. J. Hum. Genet. 2019. Vol. 105, № 4. P. 706–718.
11. Leblond C.S. et al. Both rare and common genetic variants contribute to autism in the Faroe Islands. // NPJ Genom. Med. 2019. Vol. 4. P. 1.
12. Singh T. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. // Nature. 2022. Vol. 604, № 7906. P. 509–516.
13. Reich D.E., Lander E.S. On the allelic spectrum of human disease. // Trends Genet. 2001. Vol. 17, № 9. P. 502–510.
14. Rosenberg N.A. et al. Genome-wide association studies in diverse populations. // Nat. Rev. Genet. 2010. Vol. 11, № 5. P. 356–366.
15. Wiberg R.A.W. et al. Identifying consistent allele frequency differences in studies of stratified populations. // Methods Ecol. Evol. 2017. Vol. 8, № 12. P. 1899–1909.
16. Cirulli E.T., Goldstein D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. // Nat. Rev. Genet. 2010. Vol. 11, № 6. P. 415–425.
17. Feldman M.W., Lewontin R.C. The heritability hang-up. // Science. 1975. Vol. 190, №

4220. P. 1163–1168.
18. Sudlow C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. // PLoS Med. 2015. Vol. 12, № 3. P. e1001779.
19. Hillenmeyer M.E. et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. // Science. 2008. Vol. 320, № 5874. P. 362–365.
20. Khera A.V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. // Nat. Genet. 2018. Vol. 50, № 9. P. 1219–1224.
21. Ashley-Koch A., Yang Q., Olney R.S. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. // Am. J. Epidemiol. 2000. Vol. 151, № 9. P. 839–845.
22. Sørensen S.A., Fenger K., Olsen J.H. Significantly lower incidence of cancer among patients with Huntington disease: An apoptotic effect of an expanded polyglutamine tract? // Cancer. 1999. Vol. 86, № 7. P. 1342–1346.
23. Yu H. et al. A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. // J. Clin. Invest. 2016. Vol. 126, № 3. P. 1067–1078.
24. Barbitoff Y.A. et al. Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples // medRxiv. 2021.
25. Carson A.R. et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. // BMC Bioinformatics. 2014. Vol. 15. P. 125.
26. Tanjo T. et al. Practical guide for managing large-scale human genome data in research. // J. Hum. Genet. 2021. Vol. 66, № 1. P. 39–52.
27. Hail Team. Hail 0.2. [Electronic resource]. URL: https://github.com/hail-is/hail (accessed: 25.08.2022).
28. Sepulveda J.L. Using R and bioconductor in clinical genomics and transcriptomics. // J. Mol. Diagn. 2020. Vol. 22, № 1. P. 3–20.
29. Zlotina A. et al. A 300-kb microduplication of 7q36.3 in a patient with triphalangeal thumb-polysyndactyly syndrome combined with congenital heart disease and optic disc coloboma: a case report. // BMC Med. Genomics. 2020. Vol. 13, № 1. P. 175.
30. Glotov O.S. et al. Whole-exome sequencing in Russian children with non-type 1 diabetes mellitus reveals a wide spectrum of genetic variants in MODY-related and unrelated genes. // Mol. Med. Report. 2019. Vol. 20, № 6. P. 4905–4914.
31. Barbitoff Y.A. et al. Whole-exome sequencing provides insights into monogenic disease prevalence in Northwest Russia. // Mol. Genet. Genomic Med. 2019. Vol. 7, № 11. P. e964.
32. Shikov A.E. et al. Phenome-wide functional dissection of pleiotropic effects highlights key molecular pathways for human complex traits. // Sci. Rep. 2020. Vol. 10, № 1. P. 1037.
33. Satzinger H. Theodor and Marcella Boveri : chromosomes and cytoplasm in heredity and development // Humboldt-Universität zu Berlin. 2008.
34. Roll-Hansen N. Sources of Wilhelm Johannsen's genotype theory. // J. Hist. Biol. 2009. Vol. 42, № 3. P. 457–493.
35. Davenport C.B. Color inheritance in mice. // Science. 1904. Vol. 19, № 472. P. 110–114.
36. McClung C.E. Notes on the accessory chromosome // Anat Anz. 1901. P. 220–226.
37. Brush S.G. Nettie M. Stevens and the discovery of sex determination by chromosomes. // Isis. 1978. Vol. 69, № 247. P. 163–172.
38. Allen G.E. Thomas Hunt Morgan and the Problem of Sex Determination, 1903-1910 //

Proceedings of the American Philosophical Society. 1966. Vol. 110, № 1. P. 48.
39. Fraser R.J. An introduction to medical genetics. // Oxford: Oxford University Press. 1940.
40. Følling A. Über Ausscheidung von Phenylbrenztraubensäure in den Harn als Stoffwechselanomalie in Verbindung mit Imbezillitat // Hoppe-Seylers 2 Physiol Chem. 1934. Vol. 227. P. 169–176.
41. Bearn A.G., Miller E.D. Archibald Garrod and the development of the concept of inborn errors of metabolism. // Bull. Hist. Med. 1979. Vol. 53, № 3. P. 315–328.
42. Neel J.V. The inheritance of sickle cell anemia. // Science. 1949. Vol. 110, № 2846. P. 64–66.
43. Pauling L., Itano H.A. Sickle cell anemia a molecular disease. // Science. 1949. Vol. 110, № 2865. P. 543–548.
44. Haldane J.B.S. The rate of mutation of human genes // Hereditas. 1949. Vol. 35, № S1. P. 267–273.
45. Ford C.E. et al. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). // Lancet. 1959. Vol. 1, № 7075. P. 711–713.
46. Jacobs P.A., Strong J.A. A case of human intersexuality having a possible XXY sex-determining mechanism. // Nature. 1959. Vol. 183, № 4657. P. 302–303.
47. Polani P.E. Human and clinical cytogenetics: origins, evolution and impact // Eur. J. Hum. Genet. 1997. Vol. 5, № 3. P. 117–128.
48. Caskey C.T., McKusick V.A. Medical genetics. // JAMA. 1990. Vol. 263, № 19. P. 2654–2656.
49. Leder A. et al. Comparison of cloned mouse alpha- and beta-globin genes: conservation of intervening sequence locations and extragenic homology. // Proc Natl Acad Sci USA. 1978. Vol. 75, № 12. P. 6187–6191.
50. Orkin S.H. The duplicated human alpha globin genes lie close together in cellular DNA. // Proc Natl Acad Sci USA. 1978. Vol. 75, № 12. P. 5950–5954.
51. Fritsch E.F., Lawn R.M., Maniatis T. Characterisation of deletions which affect the expression of fetal globin genes in man. // Nature. 1979. Vol. 279, № 5714. P. 598–603.
52. Ingram V.M. Abnormal human haemoglobins. III. The chemical difference between normal and sickle cell haemoglobins. // Biochim. Biophys. Acta. 1959. Vol. 36. P. 402–411.
53. Nathans D., Smith H.O. Restriction endonucleases in the analysis and restructuring of dna molecules. // Annu. Rev. Biochem. 1975. Vol. 44. P. 273–293.
54. Maniatis T. et al. The isolation of structural genes from libraries of eucaryotic DNA. // Cell. 1978. Vol. 15, № 2. P. 687–701.
55. Maxam A.M., Gilbert W. A new method for sequencing DNA. // Proc Natl Acad Sci USA. 1977. Vol. 74, № 2. P. 560–564.
56. Sanger F. et al. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. // Proc Natl Acad Sci USA. 1973. Vol. 70, № 4. P. 1209–1213.
57. Orkin S.H. et al. Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster. // Nature. 1982. Vol. 296, № 5858. P. 627–631.
58. Antonarakis S.E., Kazazian H.H., Orkin S.H. DNA polymorphism and molecular pathology of the human globin gene clusters. // Hum. Genet. 1985. Vol. 69, № 1. P. 1–14.
59. Antonarakis S.E. et al. Nonrandom association of polymorphic restriction sites in the

beta-globin gene cluster. // Proc Natl Acad Sci USA. 1982. Vol. 79, № 1. P. 137–141.
60. Chakravarti A. et al. Nonuniform recombination within the human beta-globin gene cluster. // Am. J. Hum. Genet. 1984. Vol. 36, № 6. P. 1239–1258.
61. Yamamoto T. et al. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. // Cell. 1984. Vol. 39, № 1. P. 27–38.
62. Myerowitz R., Proia R.L. cDNA clone for the alpha-chain of human beta-hexosaminidase: deficiency of alpha-chain mRNA in Ashkenazi Tay-Sachs fibroblasts. // Proc Natl Acad Sci USA. 1984. Vol. 81, № 17. P. 5394–5398.
63. Sorge J. et al. Molecular cloning and nucleotide sequence of human glucocerebrosidase cDNA. // Proc Natl Acad Sci USA. 1985. Vol. 82, № 21. P. 7289–7293.
64. Gitschier J. et al. Characterization of the human factor VIII gene. // Nature. 1984. Vol. 312, № 5992. P. 326–330.
65. Kwok S.C. et al. Nucleotide sequence of a full-length complementary DNA clone and amino acid sequence of human phenylalanine hydroxylase. // Biochemistry. 1985. Vol. 24, № 3. P. 556–561.
66. Kan Y.W., Dozy A.M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. // Proc Natl Acad Sci USA. 1978. Vol. 75, № 11. P. 5631–5635.
67. Botstein D. et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. // Am. J. Hum. Genet. 1980. Vol. 32, № 3. P. 314–331.
68. Gusella J.F. et al. A polymorphic DNA marker genetically linked to Huntington's disease. // Nature. 1983. Vol. 306, № 5940. P. 234–238.
69. Ott J. Linkage analysis and family classification under heterogeneity. // Ann. Hum. Genet. 1983. Vol. 47, № 4. P. 311–320.
70. A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. // Science. 1992. Vol. 258, № 5079. P. 67–86.
71. Dausset J. et al. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. // Genomics. 1990. Vol. 6, № 3. P. 575–577.
72. Donis-Keller H. et al. A genetic linkage map of the human genome. // Cell. 1987. Vol. 51, № 2. P. 319–337.
73. Warren A.C. et al. A genetic linkage map of 17 markers on human chromosome 21 // Genomics. 1989. Vol. 4, № 4. P. 579–591.
74. Gabriel S.B. et al. The structure of haplotype blocks in the human genome. // Science. 2002. Vol. 296, № 5576. P. 2225–2229.
75. Saiki R.K. et al. Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. // Nature. 1986. Vol. 324, № 6093. P. 163–166.
76. Burke D.T., Carle G.F., Olson M.V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. // Science. 1987. Vol. 236, № 4803. P. 806–812.
77. Botstein D., Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. // Nat. Genet. 2003. Vol. 33 Suppl. P. 228–237.
78. Antonarakis S.E., McKusick V.A. OMIM passes the 1,000-disease-gene mark // Nat. Genet. 2000. Vol. 25, № 1. P. 11–11.
79. Royer-Pokora B. et al. Cloning the gene for an inherited human disorder--chronic

granulomatous disease--on the basis of its chromosomal location. // Nature. 1986. Vol. 322, № 6074. P. 32–38.
80. Koenig M. et al. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. // Cell. 1987. Vol. 50, № 3. P. 509–517.
81. Monaco A.P. et al. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. // Nature. 1986. Vol. 323, № 6089. P. 646–650.
82. Fung Y.K. et al. Structural evidence for the authenticity of the human retinoblastoma gene. // Science. 1987. Vol. 236, № 4809. P. 1657–1661.
83. Riordan J.R. et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. // Science. 1989. Vol. 245, № 4922. P. 1066–1073.
84. Rommens J.M. et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. // Science. 1989. Vol. 245, № 4922. P. 1059–1065.
85. Malkin D. et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. // Science. 1990. Vol. 250, № 4985. P. 1233–1238.
86. Pelletier J. et al. WT1 mutations contribute to abnormal genital system development and hereditary Wilms' tumour. // Nature. 1991. Vol. 353, № 6343. P. 431–434.
87. Marchuk D.A. et al. cDNA cloning of the type 1 neurofibromatosis gene: complete sequence of the NF1 gene product. // Genomics. 1991. Vol. 11, № 4. P. 931–940.
88. Viskochil D. et al. Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. // Cell. 1990. Vol. 62, № 1. P. 187–192.
89. Wallace M.R. et al. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. // Science. 1990. Vol. 249, № 4965. P. 181–186.
90. Kinzler K.W. et al. Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers. // Science. 1991. Vol. 251, № 4999. P. 1366–1370.
91. Dietz H.C. et al. Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene. // Nature. 1991. Vol. 352, № 6333. P. 337–339.
92. Goate A. et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. // Nature. 1991. Vol. 349, № 6311. P. 704–706.
93. Verkerk A.J. et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. // Cell. 1991. Vol. 65, № 5. P. 905–914.
94. Lupski J.R. et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. // Cell. 1991. Vol. 66, № 2. P. 219–232.
95. Amir R.E. et al. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. // Nat. Genet. 1999. Vol. 23, № 2. P. 185–188.
96. Leach F.S. et al. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. // Cell. 1993. Vol. 75, № 6. P. 1215–1225.
97. Papadopoulos N. et al. Mutation of a mutL homolog in hereditary colon cancer. // Science. 1994. Vol. 263, № 5153. P. 1625–1629.
98. Sherrington R. et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. // Nature. 1995. Vol. 375, № 6534. P. 754–760.
99. Miki Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. // Science. 1994. Vol. 266, № 5182. P. 66–71.
100. Wooster R. et al. Identification of the breast cancer susceptibility gene BRCA2. // Nature. 1995. Vol. 378, № 6559. P. 789–792.

101. Savitsky K. et al. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. // Science. 1995. Vol. 268, № 5218. P. 1749–1753.
102. Rousseau F. et al. Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. // Nature. 1994. Vol. 371, № 6494. P. 252–254.
103. Shiang R. et al. Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. // Cell. 1994. Vol. 78, № 2. P. 335–342.
104. Lefebvre S. et al. Identification and characterization of a spinal muscular atrophy-determining gene. // Cell. 1995. Vol. 80, № 1. P. 155–165.
105. van Slegtenhorst M. et al. Identification of the tuberous sclerosis gene TSC1 on chromosome 9q34. // Science. 1997. Vol. 277, № 5327. P. 805–808.
106. Tartaglia M. et al. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. // Nat. Genet. 2001. Vol. 29, № 4. P. 465–468.
107. Krantz I.D. et al. Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. // Nat. Genet. 2004. Vol. 36, № 6. P. 631–635.
108. Tonkin E.T. et al. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. // Nat. Genet. 2004. Vol. 36, № 6. P. 636–641.
109. Vissers L.E.L.M. et al. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. // Nat. Genet. 2004. Vol. 36, № 9. P. 955–957.
110. Holley R.W., Madison J.T., Zamir A. A new method for sequence determination of large oligonucleotides // Biochemical and Biophysical Research Communications. 1964. Vol. 17, № 4. P. 389–394.
111. Holley R.W. et al. Structure of a Ribonucleic Acid // Science. 1965. Vol. 147, № 3664. P. 1462–1465.
112. Sanger F., Brownlee G.G., Barrell B.G. A two-dimensional fractionation procedure for radioactive nucleotides. // J. Mol. Biol. 1965. Vol. 13, № 2. P. 373–398.
113. Brownlee G.G., Sanger F. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli // J. Mol. Biol. 1967. Vol. 23, № 3. P. 337-IN9.
114. Cory S. et al. Primary structure of a methionine transfer RNA from Escherichia coli. // Nature. 1968. Vol. 220, № 5171. P. 1039–1040.
115. Dube S.K. et al. Nucleotide sequence of N-formyl-methionyl-transfer RNA. // Nature. 1968. Vol. 218, № 5138. P. 232–233.
116. Goodman H.M. et al. Amber suppression: a nucleotide change in the anticodon of a tyrosine transfer RNA. // Nature. 1968. Vol. 217, № 5133. P. 1019–1024.
117. Adams J.M. et al. Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA. // Nature. 1969. Vol. 223, № 5210. P. 1009–1014.
118. Min Jou W. et al. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. // Nature. 1972. Vol. 237, № 5350. P. 82–88.
119. Fiers W. et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. // Nature. 1976. Vol. 260, № 5551. P. 500–507.
120. McKusick V.A., Ruddle F.H. A new discipline, a new name, a new journal // Genomics. 1987. Vol. 1, № 1. P. 1–2.
121. Wu R., Kaiser A.D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. // J. Mol. Biol. 1968. Vol. 35, № 3. P. 523–537.

122. Wu R. Nucleotide sequence analysis of DNA // J. Mol. Biol. 1970. Vol. 51, № 3. P. 501–521.
123. Sanger F., Coulson A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. // J. Mol. Biol. 1975. Vol. 94, № 3. P. 441–448.
124. Sanger F. et al. Nucleotide sequence of bacteriophage phi X174 DNA. // Nature. 1977. Vol. 265, № 5596. P. 687–695.
125. Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. // Proc Natl Acad Sci USA. 1977. Vol. 74, № 12. P. 5463–5467.
126. Chidgeavadze Z.G. et al. 2’,3’-Dideoxy-3’ aminonucleoside 5’-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. // Nucleic Acids Res. 1984. Vol. 12, № 3. P. 1671–1686.
127. Smith L.M. et al. The synthesis of oligonucleotides containing an aliphatic amino group at the 5’ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. // Nucleic Acids Res. 1985. Vol. 13, № 7. P. 2399–2412.
128. Ansorge W. et al. A non-radioactive automated method for DNA sequence determination. // J. Biochem. Biophys. Methods. 1986. Vol. 13, № 6. P. 315–323.
129. Ansorge W. et al. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. // Nucleic Acids Res. 1987. Vol. 15, № 11. P. 4593–4602.
130. Prober J.M. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. // Science. 1987. Vol. 238, № 4825. P. 336–341.
131. Swerdlow H., Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. // Nucleic Acids Res. 1990. Vol. 18, № 6. P. 1415–1419.
132. Luckey J.A. et al. High speed DNA sequencing by capillary electrophoresis. // Nucleic Acids Res. 1990. Vol. 18, № 15. P. 4417–4421.
133. Hunkapiller T. et al. Large-scale and automated DNA sequence determination. // Science. 1991. Vol. 254, № 5028. P. 59–67.
134. Staden R. A strategy of DNA sequencing employing computer programs. // Nucleic Acids Res. 1979. Vol. 6, № 7. P. 2601–2610.
135. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. // Nucleic Acids Res. 1981. Vol. 9, № 13. P. 3015–3027.
136. Saiki R.K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. // Science. 1985. Vol. 230, № 4732. P. 1350–1354.
137. Saiki R.K. et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. // Science. 1988. Vol. 239, № 4839. P. 487–491.
138. Cohen S.N. et al. Construction of biologically functional bacterial plasmids in vitro. // Proc Natl Acad Sci USA. 1973. Vol. 70, № 11. P. 3240–3244.
139. Klenow H., Henningsen I. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. // Proc Natl Acad Sci USA. 1970. Vol. 65, № 1. P. 168–175.
140. Chen C.-Y. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. // Front. Microbiol. 2014. Vol. 5. P. 305.
141. Smith L.M. et al. Fluorescence detection in automated DNA sequence analysis. // Nature. 1986. Vol. 321, № 6071. P. 674–679.
142. Ansorge W.J. Next-generation DNA sequencing techniques. // N. Biotechnol. 2009. Vol. 25, № 4. P. 195–203.

143. Venter J.C. et al. The sequence of the human genome. // Science. 2001. Vol. 291, № 5507. P. 1304–1351.
144. Nyrén P., Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. // Anal. Biochem. 1985. Vol. 151, № 2. P. 504–509.
145. Hyman E.D. A new method of sequencing DNA. // Anal. Biochem. 1988. Vol. 174, № 2. P. 423–436.
146. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity // Anal. Biochem. 1987. Vol. 167, № 2. P. 235–238.
147. Ronaghi M. et al. Real-time DNA sequencing using detection of pyrophosphate release. // Anal. Biochem. 1996. Vol. 242, № 1. P. 84–89.
148. Ronaghi M., Uhlén M., Nyrén P. A sequencing method based on real-time pyrophosphate. // Science. 1998. Vol. 281, № 5375. P. 363, 365.
149. Margulies M. et al. Genome sequencing in microfabricated high-density picolitre reactors. // Nature. 2005. Vol. 437, № 7057. P. 376–380.
150. Tawfik D.S., Griffiths A.D. Man-made cell-like compartments for molecular evolution. // Nat. Biotechnol. 1998. Vol. 16, № 7. P. 652–656.
151. Voelkerding K.V., Dames S.A., Durtschi J.D. Next-generation sequencing: from basic research to diagnostics. // Clin. Chem. 2009. Vol. 55, № 4. P. 641–658.
152. Shendure J., Ji H. Next-generation DNA sequencing. // Nat. Biotechnol. 2008. Vol. 26, № 10. P. 1135–1145.
153. Fedurco M. et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. // Nucleic Acids Res. 2006. Vol. 34, № 3. P. e22.
154. Bentley D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. // Nature. 2008. Vol. 456, № 7218. P. 53–59.
155. Turcatti G. et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. // Nucleic Acids Res. 2008. Vol. 36, № 4. P. e25.
156. Balasubramanian S. Sequencing nucleic acids: from chemistry to medicine. // Chem. Commun. 2011. Vol. 47, № 26. P. 7281–7286.
157. Quail M.A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. // BMC Genomics. 2012. Vol. 13. P. 341.
158. McKernan K.J. et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. // Genome Res. 2009. Vol. 19, № 9. P. 1527–1541.
159. Shendure J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. // Science. 2005. Vol. 309, № 5741. P. 1728–1732.
160. Buermans H.P.J., den Dunnen J.T. Next generation sequencing technology: Advances and applications. // Biochim. Biophys. Acta. 2014. Vol. 1842, № 10. P. 1932–1941.
161. Glenn T.C. Field guide to next-generation DNA sequencers. // Mol. Ecol. Resour. 2011. Vol. 11, № 5. P. 759–769.
162. Drmanac R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. // Science. 2010. Vol. 327, № 5961. P. 78–81.
163. Rothberg J.M. et al. An integrated semiconductor device enabling non-optical genome sequencing. // Nature. 2011. Vol. 475, № 7356. P. 348–352.

164. Loman N.J. et al. Performance comparison of benchtop high-throughput sequencing platforms. // Nat. Biotechnol. 2012. Vol. 30, № 5. P. 434–439.
165. Schadt E.E., Turner S., Kasarskis A. A window into third-generation sequencing. // Hum. Mol. Genet. 2010. Vol. 19, № R2. P. R227-40.
166. Niedringhaus T.P. et al. Landscape of next-generation sequencing technologies. // Anal. Chem. 2011. Vol. 83, № 12. P. 4327–4341.
167. Pareek C.S., Smoczynski R., Tretyn A. Sequencing technologies and genome sequencing. // J. Appl. Genet. 2011. Vol. 52, № 4. P. 413–435.
168. Gut I.G. New sequencing technologies. // Clin. Transl. Oncol. 2013. Vol. 15, № 11. P. 879–881.
169. Braslavsky I. et al. Sequence information can be obtained from single DNA molecules. // Proc Natl Acad Sci USA. 2003. Vol. 100, № 7. P. 3960–3964.
170. Harris T.D. et al. Single-molecule DNA sequencing of a viral genome. // Science. 2008. Vol. 320, № 5872. P. 106–109.
171. Bowers J. et al. Virtual terminator nucleotides for next-generation DNA sequencing. // Nat. Methods. 2009. Vol. 6, № 8. P. 593–595.
172. GenomeWeb . 2012. Helicos BioSciences Files for Chapter 11 Bankruptcy Protection [Electronic resource]. URL: https://www.genomeweb.com/sequencing/helicos-biosciences-files-chapter-11-bankruptcy-protection (accessed: 02.05.2022).
173. van Dijk E.L. et al. Ten years of next-generation sequencing technology. // Trends Genet. 2014. Vol. 30, № 9. P. 418–426.
174. Eid J. et al. Real-time DNA sequencing from single polymerase molecules. // Science. 2009. Vol. 323, № 5910. P. 133–138.
175. Haque F. et al. Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. // Nano Today. 2013. Vol. 8, № 1. P. 56–74.
176. Kasianowicz J.J. et al. Characterization of individual polynucleotide molecules using a membrane channel. // Proc Natl Acad Sci USA. 1996. Vol. 93, № 24. P. 13770–13773.
177. Li J. et al. Ion-beam sculpting at nanometre length scales. // Nature. 2001. Vol. 412, № 6843. P. 166–169.
178. Dekker C. Solid-state nanopores. // Nat. Nanotechnol. 2007. Vol. 2, № 4. P. 209–215.
179. Clarke J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. // Nat. Nanotechnol. 2009. Vol. 4, № 4. P. 265–270.
180. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. // Nat. Biotechnol. 2012. Vol. 30, № 4. P. 295–296.
181. Loman N.J., Quinlan A.R. Poretools: a toolkit for analyzing nanopore sequence data. // Bioinformatics. 2014. Vol. 30, № 23. P. 3399–3401.
182. Branton D. et al. The potential and challenges of nanopore sequencing. // Nat. Biotechnol. 2008. Vol. 26, № 10. P. 1146–1153.
183. Loman N.J., Quick J., Simpson J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. // Nat. Methods. 2015. Vol. 12, № 8. P. 733–735.
184. Kilianski A. et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. // Gigascience. 2015. Vol. 4. P. 12.
185. Karlsson E. et al. Scaffolding of a bacterial genome using MinION nanopore sequencing. // Sci. Rep. 2015. Vol. 5. P. 11996.
186. Ashton P.M. et al. MinION nanopore sequencing identifies the position and structure of a

bacterial antibiotic resistance island. // Nat. Biotechnol. 2015. Vol. 33, № 3. P. 296–300.
187. Madoui M.-A. et al. Genome assembly using Nanopore-guided long and error-free DNA reads. // BMC Genomics. 2015. Vol. 16. P. 327.
188. Miga K.H. et al. Telomere-to-telomere assembly of a complete human X chromosome. // Nature. 2020. Vol. 585, № 7823. P. 79–84.
189. Logsdon G.A. et al. The structure, function and evolution of a complete human chromosome 8. // Nature. 2021. Vol. 593, № 7857. P. 101–107.
190. Dulbecco R. A turning point in cancer research: sequencing the human genome. // Science. 1986. Vol. 231, № 4742. P. 1055–1056.
191. Sinsheimer R.L. The santa cruz workshop--May 1985. // Genomics. 1989. Vol. 5, № 4. P. 954–956.
192. Report of the committee on mapping and sequencing the human genome. Washington, D.C.: National Academies Press, 1988.
193. Author N.G. Understanding our genetic inheritance: The US Human Genome Project, The first five years FY 1991--1995. U.S. Department of Energy: Technical Information Center, 1990.
194. Church G.M., Kieffer-Higgins S. Multiplex DNA sequencing. // Science. 1988. Vol. 240, № 4849. P. 185–188.
195. Strezoska Z. et al. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. // Proc Natl Acad Sci USA. 1991. Vol. 88, № 22. P. 10089–10093.
196. Venter J.C. et al. Shotgun sequencing of the human genome. // Science. 1998. Vol. 280, № 5369. P. 1540–1542.
197. International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome // Nature. 2001. Vol. 409, № 6822. P. 860–921.
198. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. // Nature. 2004. Vol. 431, № 7011. P. 931–945.
199. Hood L. A personal journey of discovery: developing technology and changing biology. // Annu Rev Anal Chem (Palo Alto Calif). 2008. Vol. 1. P. 1–43.
200. National Research Council (US) Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution. A New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution. Washington (DC): National Academies Press (US), 2009.
201. Ideker T., Galitski T., Hood L. A new approach to decoding life: systems biology. // Annu. Rev. Genomics Hum. Genet. 2001. Vol. 2. P. 343–372.
202. Benson D.A. et al. GenBank. // Nucleic Acids Res. 2002. Vol. 30, № 1. P. 17–20.
203. Kent W.J. et al. The human genome browser at UCSC. // Genome Res. 2002. Vol. 12, № 6. P. 996–1006.
204. Nurk S. et al. The complete sequence of a human genome // BioRxiv. 2021.
205. Encyclopedia of DNA Elements. [Electronic resource]. URL: http://encodeproject.org/ENCODE/ (accessed: 21.04.2022).
206. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. // Nature. 2012. Vol. 489, № 7414. P. 57–74.
207. Hood L., Rowen L. The Human Genome Project: big science transforms biology and medicine. // Genome Med. 2013. Vol. 5, № 9. P. 79.
208. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). // PLoS Biol. 2011. Vol. 9, № 4. P. e1001046.

209. Aebersold R., Mann M. Mass spectrometry-based proteomics. // Nature. 2003. Vol. 422, № 6928. P. 198–207.
210. Genomes Online Database: complete genome projects. [Electronic resource]. URL: http://www.genomesonline.org/cgi-bin/GOLD/index.cgi?page_requested=Complete+Genome+Projects (accessed: 21.04.2022).
211. Theobald D.L. A formal test of the theory of universal common ancestry. // Nature. 2010. Vol. 465, № 7295. P. 219–222.
212. Wolfe K.H., Li W.-H. Molecular evolution meets the genomics revolution. // Nat. Genet. 2003. Vol. 33 Suppl. P. 255–265.
213. Marques-Bonet T., Ryder O.A., Eichler E.E. Sequencing primate genomes: what have we learned? // Annu. Rev. Genomics Hum. Genet. 2009. Vol. 10. P. 355–386.
214. Noonan J.P. Neanderthal genomics and the evolution of modern humans. // Genome Res. 2010. Vol. 20, № 5. P. 547–553.
215. Stoneking M., Krause J. Learning about human population history from ancient and modern genomes. // Nat. Rev. Genet. 2011. Vol. 12, № 9. P. 603–614.
216. Sankararaman S. et al. The date of interbreeding between Neandertals and modern humans. // PLoS Genet. 2012. Vol. 8, № 10. P. e1002947.
217. Schatz M.C. Computational thinking in the era of big data biology. // Genome Biol. 2012. Vol. 13, № 11. P. 177.
218. SourceForge. [Electronic resource]. URL: http://sourceforge.net/ (accessed: 21.04.2022).
219. Bioconductor: open source software for bioinformatics. [Electronic resource]. URL: http://www.bioconductor.org/ (accessed: 21.04.2022).
220. Field D. et al. Megascience. 'Omics data sharing. // Science. 2009. Vol. 326, № 5950. P. 234–236.
221. Knoppers B.M. et al. Towards a data sharing Code of Conduct for international genomic research. // Genome Med. 2011. Vol. 3, № 7. P. 46.
222. International HapMap Consortium. A haplotype map of the human genome. // Nature. 2005. Vol. 437, № 7063. P. 1299–1320.
223. Lupski J.R. et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. // N. Engl. J. Med. 2010. Vol. 362, № 13. P. 1181–1191.
224. Marian A.J. Clinical interpretation and management of genetic variants. // JACC Basic Transl. Sci. 2020. Vol. 5, № 10. P. 1029–1042.
225. Kong A. et al. Rate of de novo mutations and the importance of father's age to disease risk. // Nature. 2012. Vol. 488, № 7412. P. 471–475.
226. Besenbacher S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. // Nat. Commun. 2015. Vol. 6. P. 5969.
227. Campbell C.D. et al. Estimating the human mutation rate using autozygosity in a founder population. // Nat. Genet. 2012. Vol. 44, № 11. P. 1277–1281.
228. Francioli L.C. et al. Genome-wide patterns and properties of de novo mutations in humans. // Nat. Genet. 2015. Vol. 47, № 7. P. 822–826.
229. Kloosterman W.P. et al. Characteristics of de novo structural changes in the human genome. // Genome Res. 2015. Vol. 25, № 6. P. 792–801.
230. Bergström A. et al. Insights into human genetic variation and population history from 929 diverse genomes // BioRxiv. 2019.
231. Fairley S. et al. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. // Nucleic Acids Res. 2020. Vol. 48, № D1. P.

D941–D947.
232. Jónsson H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. // Nature. 2017. Vol. 549, № 7673. P. 519–522.
233. Collins R.L. et al. An open resource of structural variation for medical and population genetics // BioRxiv. 2019.
234. Principles of Population Genetics. 4th ed. / ed. Hartl D.L. Sunderland, MA:Sinauer Assoc., 2006.
235. Ramachandran S. et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. // Proc Natl Acad Sci USA. 2005. Vol. 102, № 44. P. 15942–15947.
236. Martin A.R. et al. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. // Am. J. Hum. Genet. 2018. Vol. 102, № 5. P. 760–775.
237. Lim E.T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. // PLoS Genet. 2014. Vol. 10, № 7. P. e1004494.
238. Kimura M. An Introduction to Population Genetics Theory. 1st ed / ed. Crow J.F. Minneapolis, MN: Burgess., 1970.
239. Ohta T., Gillespie J.H. Development of Neutral and Nearly Neutral Theories // Theor. Popul. Biol. 1996. Vol. 49, № 2. P. 128–142.
240. Kimura M., Maruyama T., Crow J.F. The mutation load in small populations. // Genetics. 1963. Vol. 48. P. 1303–1312.
241. Henn B.M. et al. Estimating the mutation load in human genomes. // Nat. Rev. Genet. 2015. Vol. 16, № 6. P. 333–343.
242. Menozzi P., Piazza A., Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. // Science. 1978. Vol. 201, № 4358. P. 786–792.
243. Price A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. // Nat. Genet. 2006. Vol. 38, № 8. P. 904–909.
244. Visscher P.M., Hill W.G., Wray N.R. Heritability in the genomics era--concepts and misconceptions. // Nat. Rev. Genet. 2008. Vol. 9, № 4. P. 255–266.
245. Klein R.J. et al. Complement factor H polymorphism in age-related macular degeneration. // Science. 2005. Vol. 308, № 5720. P. 385–389.
246. Voutyritsa E. et al. PCSK9 Antibody-based Treatment Strategies for Patients With Statin Intolerance. // In Vivo. 2021. Vol. 35, № 1. P. 61–68.
247. Schoech A.P. et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. // Nat. Commun. 2019. Vol. 10, № 1. P. 790.
248. DeForest N., Majithia A.R. Genetics of Type 2 Diabetes: Implications from Large-Scale Studies. // Curr. Diab. Rep. 2022. Vol. 22, № 5. P. 227–235.
249. Yengo L. et al. A saturated map of common genetic variants associated with human height. // Nature. 2022. Vol. 610, № 7933. P. 704–712.
250. Abifadel M. et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. // Nat. Genet. 2003. Vol. 34, № 2. P. 154–156.
251. Manolio T.A. et al. Finding the missing heritability of complex diseases. // Nature. 2009. Vol. 461, № 7265. P. 747–753.
252. Yang J. et al. Common SNPs explain a large proportion of the heritability for human height. // Nat. Genet. 2010. Vol. 42, № 7. P. 565–569.
253. International Schizophrenia Consortium et al. Common polygenic variation contributes to

risk of schizophrenia and bipolar disorder. // Nature. 2009. Vol. 460, № 7256. P. 748–752.
254. Liu J.Z. et al. A versatile gene-based test for genome-wide association studies. // Am. J. Hum. Genet. 2010. Vol. 87, № 1. P. 139–145.
255. Wood A.R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. // Nat. Genet. 2014. Vol. 46, № 11. P. 1173–1186.
256. Marouli E. et al. Rare and low-frequency coding variants alter human adult height. // Nature. 2017. Vol. 542, № 7640. P. 186–190.
257. De Rubeis S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. // Nature. 2014. Vol. 515, № 7526. P. 209–215.
258. Fromer M. et al. De novo mutations in schizophrenia implicate synaptic networks. // Nature. 2014. Vol. 506, № 7487. P. 179–184.
259. Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. // Nucleic Acids Res. 2010. Vol. 38, № 16. P. e164.
260. Ng P.C., Henikoff S. SIFT: Predicting amino acid changes that affect protein function. // Nucleic Acids Res. 2003. Vol. 31, № 13. P. 3812–3814.
261. Adzhubei I.A. et al. A method and server for predicting damaging missense mutations. // Nat. Methods. 2010. Vol. 7, № 4. P. 248–249.
262. Yandell M. et al. A probabilistic disease-gene finder for personal genomes. // Genome Res. 2011. Vol. 21, № 9. P. 1529–1542.
263. Sunyaev S.R. Inferring causality and functional significance of human coding DNA variants. // Hum. Mol. Genet. 2012. Vol. 21, № R1. P. R10-7.
264. VEP Plugins [Electronic resource]. URL: http://www.ensembl.org/info/docs/tools/vep/script/vep_plugins.html (accessed: 28.04.2022).
265. de Leeuw C.A. et al. MAGMA: generalized gene-set analysis of GWAS data. // PLoS Comput. Biol. 2015. Vol. 11, № 4. P. e1004219.
266. Bycroft C. et al. The UK Biobank resource with deep phenotyping and genomic data. // Nature. 2018. Vol. 562, № 7726. P. 203–209.
267. Yengo L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. // Hum. Mol. Genet. 2018. Vol. 27, № 20. P. 3641–3649.
268. Astle W.J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. // Cell. 2016. Vol. 167, № 5. P. 1415-1429.e19.
269. Sinnott-Armstrong N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. // Nat. Genet. 2021. Vol. 53, № 2. P. 185–194.
270. Hill W.D. et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. // Mol. Psychiatry. 2019. Vol. 24, № 2. P. 169–181.
271. Thorp J.G. et al. Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. // Nat. Hum. Behav. 2021. Vol. 5, № 10. P. 1432–1442.
272. Christophersen I.E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. // Nat. Genet. 2017. Vol. 49, № 6. P. 946–952.
273. Ferreira M.A.R. et al. Age-of-onset information helps identify 76 genetic variants associated with allergic disease. // PLoS Genet. 2020. Vol. 16, № 6. P. e1008725.
274. Purves K.L. et al. A major role for common genetic variation in anxiety disorders. // Mol. Psychiatry. 2020. Vol. 25, № 12. P. 3292–3303.

275. Peterson R.E. et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. // Cell. 2019. Vol. 179, № 3. P. 589–603.
276. Van Hout C.V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. // Nature. 2020. Vol. 586, № 7831. P. 749–756.
277. Chong J.X. et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. // Am. J. Hum. Genet. 2015. Vol. 97, № 2. P. 199–215.
278. Lek M. et al. Analysis of protein-coding genetic variation in 60,706 humans. // Nature. 2016. Vol. 536, № 7616. P. 285–291.
279. MacArthur D.G. et al. Guidelines for investigating causality of sequence variants in human disease. // Nature. 2014. Vol. 508, № 7497. P. 469–476.
280. Richards S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. // Genet. Med. 2015. Vol. 17, № 5. P. 405–424.
281. Risch N. et al. A genomic screen of autism: evidence for a multilocus etiology. // Am. J. Hum. Genet. 1999. Vol. 65, № 2. P. 493–507.
282. Sella G., Barton N.H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. // Annu. Rev. Genomics Hum. Genet. 2019. Vol. 20. P. 461–493.
283. Pickrell J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. // Am. J. Hum. Genet. 2014. Vol. 94, № 4. P. 559–573.
284. Welter D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. // Nucleic Acids Res. 2014. Vol. 42, № Database issue. P. D1001-6.
285. Li Y. et al. A functional genomics approach to understand variation in cytokine production in humans. // Cell. 2016. Vol. 167, № 4. P. 1099-1110.e14.
286. Maurano M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. // Science. 2012. Vol. 337, № 6099. P. 1190–1195.
287. Farh K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. // Nature. 2015. Vol. 518, № 7539. P. 337–343.
288. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. // Nature. 2015. Vol. 518, № 7539. P. 317–330.
289. Furlong M. et al. Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3 to 12 years. // Cochrane Database Syst. Rev. 2012. № 2. P. CD008225.
290. Harpak A., Przeworski M. The evolution of group differences in changing environments. // PLoS Biol. 2021. Vol. 19, № 1. P. e3001072.
291. Gazal S. et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. // Nat. Genet. 2018. Vol. 50, № 11. P. 1600–1607.
292. O'Connor L.J. et al. Extreme polygenicity of complex traits is explained by negative selection. // Am. J. Hum. Genet. 2019. Vol. 105, № 3. P. 456–476.
293. Speed D., Holmes J., Balding D.J. Evaluating and improving heritability models using summary statistics. // Nat. Genet. 2020. Vol. 52, № 4. P. 458–462.
294. Zeng J. et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. // Nat. Commun. 2021. Vol. 12, № 1. P. 1164.

295. Tanaka Y. Apparent directional selection by biased pleiotropic mutation. // Genetica. 2010. Vol. 138, № 7. P. 717–723.
296. Zeng J. et al. Signatures of negative selection in the genetic architecture of human complex traits. // Nat. Genet. 2018. Vol. 50, № 5. P. 746–753.
297. Rands C.M. et al. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. // PLoS Genet. 2014. Vol. 10, № 7. P. e1004525.
298. Gulko B. et al. A method for calculating probabilities of fitness consequences for point mutations across the human genome. // Nat. Genet. 2015. Vol. 47, № 3. P. 276–283.
299. Jordan D.M., Verbanck M., Do R. HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. // Genome Biol. 2019. Vol. 20, № 1. P. 222.
300. Solovieff N. et al. Pleiotropy in complex traits: challenges and strategies. // Nat. Rev. Genet. 2013. Vol. 14, № 7. P. 483–495.
301. Hackinger S., Zeggini E. Statistical methods to detect pleiotropy in human complex traits. // Open Biol. 2017. Vol. 7, № 11.
302. Schmidt A.F. et al. Phenome-wide association analysis of LDL-cholesterol lowering genetic variants in PCSK9. // BMC Cardiovasc. Disord. 2019. Vol. 19, № 1. P. 240.
303. Watanabe K. et al. A global overview of pleiotropy and genetic architecture in complex traits. // Nat. Genet. 2019. Vol. 51, № 9. P. 1339–1348.
304. Devlin B., Roeder K. Genomic control for association studies. // Biometrics. 1999. Vol. 55, № 4. P. 997–1004.
305. Lee S., Wright F.A., Zou F. Control of population stratification by correlation-selected principal components. // Biometrics. 2011. Vol. 67, № 3. P. 967–974.
306. Ning C. et al. A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. // Bioinformatics. 2018. Vol. 34, № 11. P. 1817–1825.
307. Kang H.M. et al. Variance component model to account for sample structure in genome-wide association studies. // Nat. Genet. 2010. Vol. 42, № 4. P. 348–354.
308. Loh P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. // Nat. Genet. 2015. Vol. 47, № 3. P. 284–290.
309. Bulik-Sullivan B. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. // Nat. Genet. 2015. Vol. 47, № 3. P. 291–295.
310. Berg J.J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. // eLife. 2019. Vol. 8.
311. Sohail M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. // eLife. 2019. Vol. 8.
312. Leslie S. et al. The fine-scale genetic structure of the British population. // Nature. 2015. Vol. 519, № 7543. P. 309–314.
313. Karakachoff M. et al. Fine-scale human genetic structure in Western France. // Eur. J. Hum. Genet. 2015. Vol. 23, № 6. P. 831–836.
314. Kerminen S. et al. Fine-Scale Genetic Structure in Finland. // G3 (Bethesda). 2017. Vol. 7, № 10. P. 3459–3468.
315. Haworth S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. // Nat. Commun. 2019. Vol. 10, № 1. P. 333.
316. Raveane A. et al. Population structure of modern-day Italians reveals patterns of ancient

and archaic ancestries in Southern Europe. // Sci. Adv. 2019. Vol. 5, № 9. P. eaaw3492.
317. Torkamani A., Wineinger N.E., Topol E.J. The personal and clinical utility of polygenic risk scores. // Nat. Rev. Genet. 2018. Vol. 19, № 9. P. 581–590.
318. Knowles J.W., Ashley E.A. Cardiovascular disease: The rise of the genetic risk score. // PLoS Med. 2018. Vol. 15, № 3. P. e1002546.
319. Abdellaoui A. et al. Genetic correlates of social stratification in Great Britain. // Nat. Hum. Behav. 2019. Vol. 3, № 12. P. 1332–1342.
320. Price A.L. et al. New approaches to population stratification in genome-wide association studies. // Nat. Rev. Genet. 2010. Vol. 11, № 7. P. 459–463.
321. Lippert C. et al. FaST linear mixed models for genome-wide association studies. // Nat. Methods. 2011. Vol. 8, № 10. P. 833–835.
322. Listgarten J. et al. Improved linear mixed models for genome-wide association studies. // Nat. Methods. 2012. Vol. 9, № 6. P. 525–526.
323. Zhou X., Stephens M. Genome-wide efficient mixed-model analysis for association studies. // Nat. Genet. 2012. Vol. 44, № 7. P. 821–824.
324. Mathieson I., McVean G. Differential confounding of rare and common variants in spatially structured populations. // Nat. Genet. 2012. Vol. 44, № 3. P. 243–246.
325. Listgarten J., Lippert C., Heckerman D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. // Nat. Genet. 2013. Vol. 45, № 5. P. 470–471.
326. Mathieson I., McVean G. Reply to: “FaST-LMM-Select for addressing confounding from spatial structure and rare variants”. // Nat. Genet. 2013. Vol. 45, № 5. P. 471.
327. Zhang Y., Shen X., Pan W. Adjusting for population stratification in a fine scale with principal components and sequencing data. // Genet. Epidemiol. 2013. Vol. 37, № 8. P. 787–801.
328. Babron M.-C. et al. Rare and low frequency variant stratification in the UK population: description and impact on association tests. // PLoS ONE. 2012. Vol. 7, № 10. P. e46519.
329. Liu Q., Nicolae D.L., Chen L.S. Marbled inflation from population structure in gene-based association studies with rare variants. // Genet. Epidemiol. 2013. Vol. 37, № 3. P. 286–292.
330. Wang C. et al. Ancestry estimation and control of population stratification for sequence-based association studies. // Nat. Genet. 2014. Vol. 46, № 4. P. 409–415.
331. Pirinen M., Donnelly P., Spencer C.C.A. Including known covariates can reduce power to detect genetic effects in case-control studies. // Nat. Genet. 2012. Vol. 44, № 8. P. 848–851.
332. Zaitlen N. et al. Informed conditioning on clinical covariates increases power in case-control association studies. // PLoS Genet. 2012. Vol. 8, № 11. P. e1003032.
333. Moskvina V. et al. Design of case-controls studies with unscreened controls. // Ann. Hum. Genet. 2005. Vol. 69, № Pt 5. P. 566–576.
334. Fry A. et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. // Am. J. Epidemiol. 2017. Vol. 186, № 9. P. 1026–1034.
335. Pirastu N. et al. Genetic analyses identify widespread sex-differential participation bias. // Nat. Genet. 2021. Vol. 53, № 5. P. 663–671.
336. Choi S.W., Mak T.S.-H., O’Reilly P.F. Tutorial: a guide to performing polygenic risk score analyses. // Nat. Protoc. 2020. Vol. 15, № 9. P. 2759–2772.
337. Wang Y. et al. Challenges and opportunities for developing more generalizable polygenic risk scores. // Annu. Rev. Biomed. Data Sci. 2022. Vol. 5. P. 293–320.
338. Yang J. et al. GCTA: a tool for genome-wide complex trait analysis. // Am. J. Hum. Genet.

2011. Vol. 88, № 1. P. 76–82.
339. Lee J.J. et al. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. // Genet. Epidemiol. 2018. Vol. 42, № 8. P. 783–795.
340. Sanderson E. et al. Mendelian randomization. // Nat. Rev. Methods Primers. 2022. Vol. 2.
341. Richmond R.C., Davey Smith G. Mendelian randomization: concepts and scope. // Cold Spring Harb. Perspect. Med. 2022. Vol. 12, № 1.
342. Zhang Y. et al. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. // Brief. Bioinformatics. 2021. Vol. 22, № 5.
343. de Andrade M., Mazo Lopera M.A., Duarte N.E. Bivariate traits association analysis using generalized estimating equations in family data. // Stat. Appl. Genet. Mol. Biol. 2020. Vol. 19, № 2.
344. Bulik-Sullivan B. et al. An atlas of genetic correlations across human diseases and traits. // Nat. Genet. 2015. Vol. 47, № 11. P. 1236–1241.
345. Zhou X., Im H.K., Lee S.H. CORE GREML for estimating covariance between random effects in linear mixed models for complex trait analyses. // Nat. Commun. 2020. Vol. 11, № 1. P. 4208.
346. Sazonovs A., Barrett J.C. Rare-Variant Studies to Complement Genome-Wide Association Studies. // Annu. Rev. Genomics Hum. Genet. 2018. Vol. 19. P. 97–112.
347. Benner C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. // Bioinformatics. 2016. Vol. 32, № 10. P. 1493–1501.
348. Chen W. et al. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. // Nat. Commun. 2021. Vol. 12, № 1. P. 7117.
349. Quick C. et al. emeraLD: rapid linkage disequilibrium estimation with massive datasets. // Bioinformatics. 2019. Vol. 35, № 1. P. 164–166.
350. Benner C. et al. Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. // Am. J. Hum. Genet. 2017. Vol. 101, № 4. P. 539–551.
351. Boycott K.M. et al. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. // Nat. Rev. Genet. 2013. Vol. 14, № 10. P. 681–691.
352. Sawyer S.L. et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. // Clin. Genet. 2016. Vol. 89, № 3. P. 275–284.
353. Li Y. et al. Low-coverage sequencing: implications for design of complex trait association studies. // Genome Res. 2011. Vol. 21, № 6. P. 940–951.
354. Pasaniuc B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. // Nat. Genet. 2012. Vol. 44, № 6. P. 631–635.
355. 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. // Nature. 2012. Vol. 491, № 7422. P. 56–65.
356. Morrison A.C. et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. // Nat. Genet. 2013. Vol. 45, № 8. P. 899–901.
357. Macrogen Europe [Electronic resource]. URL: https://www.macrogen-europe.com/ (accessed: 21.04.2022).
358. Pujar S. et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. // Nucleic Acids Res. 2018. Vol. 46, № D1. P. D221–D228.

359. Do R., Kathiresan S., Abecasis G.R. Exome sequencing and complex disease: practical aspects of rare variant association studies. // Hum. Mol. Genet. 2012. Vol. 21, № R1. P. R1-9.
360. Zhan X. et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. // Nat. Genet. 2013. Vol. 45, № 11. P. 1375–1379.
361. Hu Y. et al. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. // Am. J. Hum. Genet. 2013. Vol. 93, № 5. P. 891–899.
362. Jun G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. // Am. J. Hum. Genet. 2012. Vol. 91, № 5. P. 839–848.
363. Gorski M.M. et al. Whole-exome sequencing to identify genetic risk variants underlying inhibitor development in severe hemophilia A patients. // Blood. 2016. Vol. 127, № 23. P. 2924–2933.
364. LaHaye S. et al. Utilization of whole exome sequencing to identify causative mutations in familial congenital heart disease. // Circ. Cardiovasc. Genet. 2016. Vol. 9, № 4. P. 320–329.
365. Gambin T. et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. // Nucleic Acids Res. 2017. Vol. 45, № 4. P. 1633–1648.
366. Gupta S. et al. Whole exome sequencing: Uncovering causal genetic variants for ocular diseases. // Exp. Eye Res. 2017. Vol. 164. P. 139–150.
367. Weigelt B. et al. The landscape of somatic genetic alterations in breast cancers from ATM germline mutation carriers. // J Natl Cancer Inst. 2018. Vol. 110, № 9. P. 1030–1034.
368. Haiman C.A. et al. Multiple regions within 8q24 independently affect risk for prostate cancer. // Nat. Genet. 2007. Vol. 39, № 5. P. 638–644.
369. Haiman C.A. et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. // Nat. Genet. 2011. Vol. 43, № 6. P. 570–573.
370. Yeager M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. // Nat. Genet. 2007. Vol. 39, № 5. P. 645–649.
371. Amundadottir L.T. et al. A common variant associated with prostate cancer in European and African populations. // Nat. Genet. 2006. Vol. 38, № 6. P. 652–658.
372. Hunt K.A. et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. // Nature. 2013. Vol. 498, № 7453. P. 232–235.
373. Tang H. et al. A large-scale screen for coding variants predisposing to psoriasis. // Nat. Genet. 2014. Vol. 46, № 1. P. 45–50.
374. Johansen C.T. et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. // Nat. Genet. 2010. Vol. 42, № 8. P. 684–687.
375. Voight B.F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. // PLoS Genet. 2012. Vol. 8, № 8. P. e1002793.
376. Cortes A., Brown M.A. Promise and pitfalls of the Immunochip. // Arthritis Res. Ther. 2011. Vol. 13, № 1. P. 101.
377. Verlouw J.A.M. et al. A comparison of genotyping arrays. // Eur. J. Hum. Genet. 2021. Vol. 29, № 11. P. 1611–1624.
378. Huyghe J.R. et al. Exome array analysis identifies new loci and low-frequency variants

influencing insulin processing and secretion. // Nat. Genet. 2013. Vol. 45, № 2. P. 197–201.
379. Kryukov G.V. et al. Power of deep, all-exon resequencing for discovery of human trait genes. // Proc Natl Acad Sci USA. 2009. Vol. 106, № 10. P. 3871–3876.
380. Guey L.T. et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. // Genet. Epidemiol. 2011. Vol. 35, № 4. P. 236–246.
381. Barnett I.J., Lee S., Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. // Genet. Epidemiol. 2013. Vol. 37, № 2. P. 142–151.
382. Li D. et al. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. // Genet. Epidemiol. 2011. Vol. 35, № 8. P. 790–799.
383. Emond M.J. et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. // Nat. Genet. 2012. Vol. 44, № 8. P. 886–889.
384. Ng S.B. et al. Exome sequencing identifies the cause of a mendelian disorder. // Nat. Genet. 2010. Vol. 42, № 1. P. 30–35.
385. Ng S.B. et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. // Nat. Genet. 2010. Vol. 42, № 9. P. 790–793.
386. O'Roak B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. // Nature. 2012. Vol. 485, № 7397. P. 246–250.
387. Jeste S.S., Geschwind D.H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. // Nat. Rev. Neurol. 2014. Vol. 10, № 2. P. 74–81.
388. Pierson T.M. et al. Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. // PLoS Genet. 2011. Vol. 7, № 10. P. e1002325.
389. Yang Y. et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. // N. Engl. J. Med. 2013. Vol. 369, № 16. P. 1502–1511.
390. Tennessen J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. // Science. 2012. Vol. 337, № 6090. P. 64–69.
391. Fu W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. // Nature. 2013. Vol. 493, № 7431. P. 216–220.
392. Lange L.A. et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. // Am. J. Hum. Genet. 2014. Vol. 94, № 2. P. 233–245.
393. Stephens Z.D. et al. Big data: astronomical or genomical? // PLoS Biol. 2015. Vol. 13, № 7. P. e1002195.
394. He K.Y., Ge D., He M.M. Big data analytics for genomic medicine. // Int. J. Mol. Sci. 2017. Vol. 18, № 2.
395. Karczewski K.J. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D840–D845.
396. Perkel J.M. LIFE SCIENCE TECHNOLOGIES: exome sequencing: toward an interpretable genome // Science. 2013. Vol. 342, № 6155. P. 262–264.
397. Illumina (2018). Scalability for Sequencing Like Never Before [Electronic resource]. URL: https://sapac.illumina.com/systems/sequencing-platforms/novaseq/specifications.html

(accessed: 22.04.2022).
398. Robasky K., Lewis N.E., Church G.M. The role of replicates for error mitigation in next-generation sequencing. // Nat. Rev. Genet. 2014. Vol. 15, № 1. P. 56–62.
399. Hofmann A.L. et al. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. // BMC Bioinformatics. 2017. Vol. 18, № 1. P. 8.
400. Wang Q. et al. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. // Sci. Rep. 2017. Vol. 7, № 1. P. 885.
401. Hoischen A., Krumm N., Eichler E.E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. // Nat. Neurosci. 2014. Vol. 17, № 6. P. 764–772.
402. Tabet A.-C. et al. Complex nature of apparently balanced chromosomal rearrangements in patients with autism spectrum disorder. // Mol. Autism. 2015. Vol. 6. P. 19.
403. PrecisionFDA [Electronic resource]. URL: https://precision.fda.gov/ (accessed: 22.04.2022).
404. Patel Z.H. et al. The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. // Front. Genet. 2014. Vol. 5. P. 16.
405. DeepVariant. DeepVariant is an Analysis Pipeline that Uses a Deep Neural Network to Call Genetic Variants From Next-Generation DNA Sequencing Data. [Electronic resource]. URL: https://github.com/google/deepvariant (accessed: 22.04.2022).
406. Homer N. et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. // PLoS Genet. 2008. Vol. 4, № 8. P. e1000167.
407. Gymrek M. et al. Identifying personal genomes by surname inference. // Science. 2013. Vol. 339, № 6117. P. 321–324.
408. Harmanci A., Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. // Nat. Methods. 2016. Vol. 13, № 3. P. 251–256.
409. Wright C.F. et al. Returning genome sequences to research participants: Policy and practice. [version 1; peer review: 2 approved] // Wellcome Open Res. 2017. Vol. 2. P. 15.
410. Kaye J. et al. Can I access my personal genome? The current legal position in the UK. // Med. Law Rev. 2014. Vol. 22, № 1. P. 64–86.
411. Belkadi A. et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. // Proc Natl Acad Sci USA. 2015. Vol. 112, № 17. P. 5473–5478.
412. Visel A., Rubin E.M., Pennacchio L.A. Genomic views of distant-acting enhancers. // Nature. 2009. Vol. 461, № 7261. P. 199–205.
413. Shigemizu D. et al. Performance comparison of four commercial human whole-exome capture platforms. // Sci. Rep. 2015. Vol. 5. P. 12742.
414. K Y., G H. Structural Variation Detection from Next Generation Sequencing // Next Generat. Sequenc. & Applic. 2015. Vol. 01, № S1.
415. Shang J. et al. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. // Biomed Res. Int. 2014. Vol. 2014. P. 309650.
416. Cho N. et al. De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries. // Nat. Commun. 2015. Vol. 6. P. 8351.
417. Deng X. et al. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. // Nucleic Acids Res. 2015. Vol. 43, № 7. P. e46.

418. Hung C.-M. et al. The de novo assembly of mitochondrial genomes of the extinct passenger pigeon (Ectopistes migratorius) with next generation sequencing. // PLoS ONE. 2013. Vol. 8, № 2. P. e56301.
419. Nagasaki H. et al. DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. // DNA Res. 2013. Vol. 20, № 4. P. 383–390.
420. Menon R. et al. VDAP-GUI: a user-friendly pipeline for variant discovery and annotation of raw next-generation sequencing data. // 3 Biotech. 2016. Vol. 6, № 1. P. 68.
421. Altschul S.F. et al. Basic local alignment search tool. // J. Mol. Biol. 1990. Vol. 215, № 3. P. 403–410.
422. Jones P. et al. InterProScan 5: genome-scale protein function classification. // Bioinformatics. 2014. Vol. 30, № 9. P. 1236–1240.
423. DePristo M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. // Nat. Genet. 2011. Vol. 43, № 5. P. 491–498.
424. Cibulskis K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. // Nat. Biotechnol. 2013. Vol. 31, № 3. P. 213–219.
425. Ahn D.H. et al. Whole-exome tumor sequencing study in biliary cancer patients with a response to MEK inhibitors. // Oncotarget. 2016. Vol. 7, № 5. P. 5306–5312.
426. Engelhardt K.R. et al. Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by Whole Exome Sequencing. // J. Clin. Immunol. 2017. Vol. 37, № 1. P. 42–50.
427. Kim B.-Y. et al. Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. // PLoS ONE. 2017. Vol. 12, № 8. P. e0182272.
428. Coudray A. et al. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. // PeerJ. 2018. Vol. 6. P. e5362.
429. Han Z. et al. The identification of growth, immune related genes and marker discovery through transcriptome in the yellow drum (Nibea albiflora) // Genes Genomics. 2018. Vol. 40, № 8. P. 881–891.
430. Zhu P. et al. OTG-snpcaller: an optimized pipeline based on TMAP and GATK for SNP calling from ion torrent data. // PLoS ONE. 2014. Vol. 9, № 5. P. e97507.
431. Malcolmson J. et al. SCN8A mutation in a child presenting with seizures and developmental delays. // Cold Spring Harb Mol Case Stud. 2016. Vol. 2, № 6. P. a001073.
432. Romanel A. et al. ASEQ: fast allele-specific studies from next-generation sequencing data. // BMC Med. Genomics. 2015. Vol. 8. P. 9.
433. Faltas B.M. et al. Clonal evolution of chemotherapy-resistant urothelial carcinoma. // Nat. Genet. 2016. Vol. 48, № 12. P. 1490–1499.
434. Decap D. et al. Halvade-RNA: Parallel variant calling from transcriptomic data using MapReduce. // PLoS ONE. 2017. Vol. 12, № 3. P. e0174575.
435. Wang J., Ling C., Gao J. Cnndel: calling structural variations on low coverage data based on convolutional neural networks. // Biomed Res. Int. 2017. Vol. 2017. P. 6375059.
436. Wang Y. et al. GT-WGS: an efficient and economic tool for large-scale WGS analyses based on the AWS cloud service. // BMC Genomics. 2018. Vol. 19, № Suppl 1. P. 959.
437. D'Aurizio R. et al. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. // Nucleic Acids Res. 2016. Vol. 44, № 20. P. e154.
438. Maxwell E.K. et al. KaryoScan: abnormal karyotype detection from whole-exome sequence // BioRxiv. 2017.
439. Gameiro D. N. AstraZeneca Partners up With Genomics Elite for new Biobank [Electronic

resource]. URL: https://labiotech.eu/medical/astrazeneca-partners-up-with-genomics-elite-for-new-biobank/ (accessed: 22.04.2022).
440. Delivering ourpipeline throughscientific leadership. 2016.
441. Johannessen C.M., Boehm J.S. Progress towards precision functional genomics // Current Opinion in Systems Biology. 2017.
442. Wang T. et al. The Human Pangenome Project: a global resource to map genomic diversity. // Nature. 2022. Vol. 604, № 7906. P. 437–446.
443. Halvorsen M. et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. // Nat. Commun. 2020. Vol. 11, № 1. P. 1842.
444. Huddleston J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. // Genome Res. 2017. Vol. 27, № 5. P. 677–685.
445. Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies. // Nat. Rev. Genet. 2016. Vol. 17, № 6. P. 333–351.
446. Ho S.S., Urban A.E., Mills R.E. Structural variation in the sequencing era. // Nat. Rev. Genet. 2020. Vol. 21, № 3. P. 171–189.
447. Mahmoud M. et al. Structural variant calling: the long and the short of it. // Genome Biol. 2019. Vol. 20, № 1. P. 246.
448. Weckselblatt B., Rudd M.K. Human structural variation: mechanisms of chromosome rearrangements. // Trends Genet. 2015. Vol. 31, № 10. P. 587–599.
449. Zook J.M. et al. A robust benchmark for detection of germline large deletions and insertions. // Nat. Biotechnol. 2020. Vol. 38, № 11. P. 1347–1355.
450. Chaisson M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. // Nat. Commun. 2019. Vol. 10, № 1. P. 1784.
451. Beyter D. et al. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease // BioRxiv. 2019.
452. Wu Z. et al. Structural variants in Chinese population and their impact on phenotypes, diseases and population adaptation // BioRxiv. 2021.
453. Majidian S., Sedlazeck F.J. PhaseME: Automatic rapid assessment of phasing quality and phasing improvement. // Gigascience. 2020. Vol. 9, № 7.
454. Hiatt S.M. et al. Long-read genome sequencing for the diagnosis of neurodevelopmental disorders // BioRxiv. 2020.
455. Sone J. et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. // Nat. Genet. 2019. Vol. 51, № 8. P. 1215–1221.
456. de la Morena-Barrio B. et al. Long-read sequencing resolves structural variants in *SERPINC1* causing antithrombin deficiency and identifies a complex rearrangement and a retrotransposon insertion not characterized by routine diagnostic methods // BioRxiv. 2020.
457. Smedley D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. // Nat. Protoc. 2015. Vol. 10, № 12. P. 2004–2015.
458. Deep Genomics [Electronic resource]. URL: https://www.deepgenomics.com/ (accessed: 02.05.2022).
459. Poplin R. et al. A universal SNP and small-indel variant caller using deep neural networks. // Nat. Biotechnol. 2018. Vol. 36, № 10. P. 983–987.

460. Camacho D.M. et al. Next-Generation Machine Learning for Biological Networks. // Cell. 2018. Vol. 173, № 7. P. 1581–1592.
461. Rabbani B., Tekin M., Mahdieh N. The promise of whole-exome sequencing in medical genetics. // J. Hum. Genet. 2014. Vol. 59, № 1. P. 5–15.
462. Mikheyev A.S., Tin M.M.Y. A first look at the Oxford Nanopore MinION sequencer. // Mol. Ecol. Resour. 2014. Vol. 14, № 6. P. 1097–1102.
463. Shihab H.A. et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. // Hum. Mutat. 2013. Vol. 34, № 1. P. 57–65.
464. Liu F. et al. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. // Sci. Rep. 2016. Vol. 6. P. 28517.
465. McLaren W. et al. The ensembl variant effect predictor. // Genome Biol. 2016. Vol. 17, № 1. P. 122.
466. Alyass A., Turcotte M., Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. // BMC Med. Genomics. 2015. Vol. 8. P. 33.
467. Shorten C., Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning // J. Big Data. 2019. Vol. 6, № 1. P. 60.
468. Metcalfe J., Kornell N., Finn B. Delayed versus immediate feedback in children's and adults' vocabulary learning. // Mem. Cognit. 2009. Vol. 37, № 8. P. 1077–1087.
469. Vassy J.L. et al. The Impact of Whole-Genome Sequencing on the Primary Care and Outcomes of Healthy Adult Patients: A Pilot Randomized Trial. // Ann. Intern. Med. 2017. Vol. 167, № 3. P. 159–169.
470. QIAGEN [Electronic resource]. URL: https://corporate.qiagen.com/ (accessed: 02.05.2022).
471. Golden Helix [Electronic resource]. URL: https://www.goldenhelix.com/products/VarSeq/ (accessed: 02.05.2022).
472. ADVAITA [Electronic resource]. URL: https://advaitabio.com/ivariantguide/ (accessed: 02.05.2022).
473. LifeMap Sciences [Electronic resource]. URL: https://www.lifemapsc.com/ (accessed: 02.05.2022).
474. Krämer A. et al. Causal analysis approaches in Ingenuity Pathway Analysis. // Bioinformatics. 2014. Vol. 30, № 4. P. 523–530.
475. Ben-Ari Fuchs S. et al. Geneanalytics: an integrative gene set analysis tool for next generation sequencing, rnaseq and microarray data. // OMICS. 2016. Vol. 20, № 3. P. 139–151.
476. Stelzer G. et al. The genecards suite: from gene data mining to disease genome sequence analyses. // Curr. Protoc. Bioinformatics. 2016. Vol. 54. P. 1.30.1-1.30.33.
477. Barbitoff Y.A. et al. Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. // Genet. Med. 2018. Vol. 20, № 3. P. 360–364.
478. Barbitoff Y.A. et al. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage // BioRxiv. 2018.
479. Allelic frequencies in Russian population using WES [Electronic resource]. URL: https://github.com/bioinf/afpaper (accessed: 15.05.2022).
480. UK Biobank [Electronic resource]. URL: http://www.nealelab.is/uk-biobank (accessed: 15.05.2022).
481. PheWAS on UK Biobank [Electronic resource]. URL:

https://github.com/bioinf/ukb_phewas (accessed: 15.05.2022).
482. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65.
483. Zhao S. et al. Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. // Biol. Proced. Online. 2018. Vol. 20. P. 5.
484. Lovmar L. et al. Silhouette scores for assessment of SNP genotype clusters. // BMC Genomics. 2005. Vol. 6. P. 35.
485. Jolliffe I.T., Cadima J. Principal component analysis: a review and recent developments. // Philos. Transact. A Math. Phys. Eng. Sci. 2016. Vol. 374, № 2065. P. 20150202.
486. Reich D., Price A.L., Patterson N. Principal component analysis of genetic data. // Nat. Genet. 2008. Vol. 40, № 5. P. 491–492.
487. Karczewski K.J., Martin A.R. Analytic and translational genetics // Annu. Rev. Biomed. Data Sci. 2020. Vol. 3, № 1.
488. Athey T.L. et al. AutoGMM: Automatic and Hierarchical Gaussian Mixture Modeling in Python // arXiv. 2019.
489. Asimit J., Zeggini E. Rare variant association analysis methods for complex traits. // Annu. Rev. Genet. 2010. Vol. 44. P. 293–308.
490. Ma C. et al. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. // Genet. Epidemiol. 2013. Vol. 37, № 6. P. 539–550.
491. Asimit J.L. et al. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. // Hum. Hered. 2012. Vol. 73, № 2. P. 84–94.
492. Morgenthaler S., Thilly W.G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). // Mutat. Res. 2007. Vol. 615, № 1–2. P. 28–56.
493. Li B., Leal S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. // Am. J. Hum. Genet. 2008. Vol. 83, № 3. P. 311–321.
494. Morris A.P., Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. // Genet. Epidemiol. 2010. Vol. 34, № 2. P. 188–193.
495. Madsen B.E., Browning S.R. A groupwise association test for rare mutations using a weighted sum statistic. // PLoS Genet. 2009. Vol. 5, № 2. P. e1000384.
496. Zawistowski M. et al. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. // Am. J. Hum. Genet. 2010. Vol. 87, № 5. P. 604–617.
497. Han F., Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. // Hum. Hered. 2010. Vol. 70, № 1. P. 42–54.
498. Hoffmann T.J., Marini N.J., Witte J.S. Comprehensive approach to analyzing rare genetic variants. // PLoS ONE. 2010. Vol. 5, № 11. P. e13584.
499. Lin D.-Y., Tang Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. // Am. J. Hum. Genet. 2011. Vol. 89, № 3. P. 354–367.
500. Neale B.M. et al. Testing for an unusual distribution of rare variants. // PLoS Genet. 2011. Vol. 7, № 3. P. e1001322.
501. Wu M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. // Am. J. Hum. Genet. 2011. Vol. 89, № 1. P. 82–93.
502. Wu M.C. et al. Powerful SNP-set analysis for case-control genome-wide association studies. // Am. J. Hum. Genet. 2010. Vol. 86, № 6. P. 929–942.
503. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. //

Genet. Epidemiol. 2009. Vol. 33, № 6. P. 497–507.
504. Derkach A., Lawless J.F., Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. // Genet. Epidemiol. 2013. Vol. 37, № 1. P. 110–121.
505. Rödel E. Fisher, R. A.: statistical methods for research workers, 14. aufl., oliver & boyd, edinburgh, london 1970. XIII, 362 S., 12 abb., 74 tab., 40 s // Biom. J. 1971. Vol. 13, № 6. P. 429–430.
506. Sun J., Zheng Y., Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. // Genet. Epidemiol. 2013. Vol. 37, № 4. P. 334–344.
507. Sarkar S.K., Chang C.-K. The simes method for multiple hypothesis testing with positively dependent test statistics // J. Am. Stat. Assoc. 1997. Vol. 92, № 440. P. 1601.
508. Chen L.S. et al. An exponential combination procedure for set-based association tests in sequencing studies. // Am. J. Hum. Genet. 2012. Vol. 91, № 6. P. 977–986.
509. Cruchaga C. et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. // Nature. 2014. Vol. 505, № 7484. P. 550–554.
510. Liu D.J. et al. Meta-analysis of gene-level tests for rare variant association. // Nat. Genet. 2014. Vol. 46, № 2. P. 200–204.
511. Zeggini E., Ioannidis J.P.A. Meta-analysis in genome-wide association studies. // Pharmacogenomics. 2009. Vol. 10, № 2. P. 191–201.
512. Lin D.Y., Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. // Genet. Epidemiol. 2010. Vol. 34, № 1. P. 60–66.
513. Evangelou E., Ioannidis J.P.A. Meta-analysis methods for genome-wide association studies and beyond. // Nat. Rev. Genet. 2013. Vol. 14, № 6. P. 379–389.
514. Liu L. et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. // PLoS Genet. 2013. Vol. 9, № 4. P. e1003443.
515. Lee A.M. STOUFFER, SAMUEL A., EDWARD A. SUCHMAN, LELAND C. DEVINNEY, SHIRLEY A. STAR, and ROBIN M. WILLIAMS, JR. The American Soldier: Adjustment During Army Life. Vol. I. Pp. xii, 599. STOUFFER, SAMUEL A., ARTHUR A. LUMSDAINE, MARION HARPER LUMSDAINE, ROBIN M. WILLIAMS, JR., M. BREWSTER SMITH, IRVING L. JANIS, SHIRLEY A. STAR, and LEONARD S. COTTRELL, JR. The American Soldier: Combat and Its Aftermath. Vol. II. Pp. 675. Princeton: Princeton University Press, Studies in Social Psychology in World War II, 1949. Vols. I and II together, $13.50; separately, $7.50 per vol // Ann. Am. Acad. Pol. Soc. Sci. 1949. Vol. 265, № 1. P. 173–175.
516. Lee S. et al. General framework for meta-analysis of rare variants in sequencing association studies. // Am. J. Hum. Genet. 2013. Vol. 93, № 1. P. 42–53.
517. Tang Z.-Z., Lin D.-Y. MASS: meta-analysis of score statistics for sequencing studies. // Bioinformatics. 2013. Vol. 29, № 14. P. 1803–1805.
518. Hu Y.-J. et al. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. // Am. J. Hum. Genet. 2013. Vol. 93, № 2. P. 236–248.
519. Morris A.P. Transethnic meta-analysis of genomewide association studies. // Genet. Epidemiol. 2011. Vol. 35, № 8. P. 809–822.
520. Han B., Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. // Am. J. Hum. Genet. 2011. Vol. 88, № 5. P. 586–598.
521. Ross M.G. et al. Characterizing and measuring bias in sequence data. // Genome Biol.

2013. Vol. 14, № 5. P. R51.
522. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. // Nature. 2015. Vol. 526, № 7571. P. 68–74.
523. Karczewski K.J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes // BioRxiv. 2019.
524. Exome Aggregation Consortium et al. Analysis of protein-coding genetic variation in 60,706 humans. // BioRxiv. 2015.
525. Fakhro K.A. et al. The Qatar genome: a population-specific tool for precision medicine in the Middle East. // Hum. Gen. Variation. 2016. Vol. 3. P. 16016.
526. Rodriguez-Flores J.L. et al. Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. // Genome Res. 2016. Vol. 26, № 2. P. 151–162.
527. Oleksyk T.K., Brukhin V., O'Brien S.J. The Genome Russia project: closing the largest remaining omission on the world Genome map. // Gigascience. 2015. Vol. 4. P. 53.
528. Zhernakova D.V. et al. Analytical "bake-off" of whole genome sequencing quality for the Genome Russia project using a small cohort for autoimmune hepatitis. // PLoS ONE. 2018. Vol. 13, № 7. P. e0200423.
529. Barbitoff Y.A. et al. Identification of Novel Candidate Markers of Type 2 Diabetes and Obesity in Russia by Exome Sequencing with a Limited Sample Size. // Genes (Basel). 2018. Vol. 9, № 8.
530. Cingolani P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. // Fly (Austin). 2012. Vol. 6, № 2. P. 80–92.
531. Tighe O. et al. Genetic diversity within the R408W phenylketonuria mutation lineages in Europe. // Hum. Mutat. 2003. Vol. 21, № 4. P. 387–393.
532. Wulff K., Herrmann F.H. Twenty two novel mutations of the factor VII gene in factor VII deficiency. // Hum. Mutat. 2000. Vol. 15, № 6. P. 489–496.
533. Takeda A. et al. Molecular basis of tyrosinase-negative oculocutaneous albinism. A single base mutation in the tyrosinase gene causing arginine to glutamine substitution at position 59. // J. Biol. Chem. 1990. Vol. 265, № 29. P. 17792–17797.
534. de Bie P. et al. Molecular pathogenesis of Wilson and Menkes disease: correlation of mutations with molecular defects and disease phenotypes. // J. Med. Genet. 2007. Vol. 44, № 11. P. 673–688.
535. Baumann M. et al. Mutations in FKBP14 cause a variant of Ehlers-Danlos syndrome with progressive kyphoscoliosis, myopathy, and hearing loss. // Am. J. Hum. Genet. 2012. Vol. 90, № 2. P. 201–216.
536. Pritchard J.K., Cox N.J. The allelic architecture of human disease genes: common disease-common variant...or not? // Hum. Mol. Genet. 2002. Vol. 11, № 20. P. 2417–2423.
537. Batty G.D. et al. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. // BMJ. 2020. Vol. 368. P. m131.
538. McInnes I.B. et al. Comparison of baricitinib, upadacitinib, and tofacitinib mediated regulation of cytokine signaling in human leukocyte subpopulations. // Arthritis Res. Ther. 2019. Vol. 21, № 1. P. 183.
539. Canela-Xandri O., Rawlik K., Tenesa A. An atlas of genetic associations in UK Biobank. //

Nat. Genet. 2018. Vol. 50, № 11. P. 1593–1599.
540. Glanville K.P. et al. Investigating pleiotropy between depression and autoimmune diseases using the UK biobank. // Biological Psychiatry Global Open Science. 2021. Vol. 1, № 1. P. 48–58.
541. van Rheenen W. et al. Genetic correlations of polygenic disease traits: from theory to practice. // Nat. Rev. Genet. 2019. Vol. 20, № 10. P. 567–581.
542. Paaby A.B., Rockman M.V. The many faces of pleiotropy. // Trends Genet. 2013. Vol. 29, № 2. P. 66–73.
543. Gipson D.S. et al. Complete remission in the nephrotic syndrome study network. // Clin. J. Am. Soc. Nephrol. 2016. Vol. 11, № 1. P. 81–89.
544. Freedman B.I. et al. Polymorphisms in the non-muscle myosin heavy chain 9 gene (MYH9) are strongly associated with end-stage renal disease historically attributed to hypertension in African Americans. // Kidney Int. 2009. Vol. 75, № 7. P. 736–745.
545. Ho D.E. et al. MatchIt : Nonparametric Preprocessing for Parametric Causal Inference // J. Stat. Softw. 2011. Vol. 42, № 8.
546. Genovese G. et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. // Science. 2010. Vol. 329, № 5993. P. 841–845.
547. Mathern D.R., Heeger P.S. Molecules great and small: the complement system. // Clin. J. Am. Soc. Nephrol. 2015. Vol. 10, № 9. P. 1636–1650.
548. Poppelaars F., Thurman J.M. Complement-mediated kidney diseases. // Mol. Immunol. 2020. Vol. 128. P. 175–187.
549. Salant D.J. et al. A new role for complement in experimental membranous nephropathy in rats. // J. Clin. Invest. 1980. Vol. 66, № 6. P. 1339–1350.
550. Strassheim D. et al. IgM contributes to glomerular injury in FSGS. // J. Am. Soc. Nephrol. 2013. Vol. 24, № 3. P. 393–406.
551. Tham W.-H. et al. Complement receptor 1 is the host erythrocyte receptor for Plasmodium falciparum PfRh4 invasion ligand. // Proc Natl Acad Sci USA. 2010. Vol. 107, № 40. P. 17327–17332.
552. Opi D.H. et al. Two complement receptor one alleles have opposing associations with cerebral malaria and interact with α+thalassaemia. // eLife. 2018. Vol. 7.

ПРИЛОЖЕНИЯ

Таблица 1. Топ-35 результатов анализа на присутствие rs17047661 в HLA-типе

HLA	GT	pvalue.afr	pvalue.eur
HLA-DQB1	06:02	0.4079437114	0.005091652191
HLA-A	29:02	0.4967179612	0.02691250052
HLA-DRB1	15:01	0.09424731899	0.02700897226
HLA-A	02:01/30	1	0.0353949089
HLA-C	16:01	0.2366706976	0.04319061145
HLA-DQB1	03:01	0.5391910071	0.05579866535
HLA-DRB1	04:05	0.129983287	0.07706973155
HLA-B	44:03	0.3429098007	0.09234732559
HLA-DRB1	11:04	0.1123856918	0.09981746544
HLA-A	03:01	0.5149344467	0.1219476206
HLA-C	07:01	0.5304281676	0.1349462547
HLA-DQB1	02:02	0.5223787831	0.1540105812
HLA-C	01:02	0.463423715	0.1564288925
HLA-B	27:05	0.629351584	0.1747812312
HLA-DQB1		0.6691557398	0.2171496429
HLA-B	14:02	0.2437127992	0.2342380209
HLA-C	06:02	0.3837185433	0.2686524299
HLA-DRB1	11:01	0.5178700056	0.2925069588
HLA-B	35:03	0.5575376828	0.335884148
HLA-A	11:01	0.5190161032	0.3543632728
HLA-A	23:01	0.5244479256	0.3619178277
HLA-B	08:01	0.5704299903	0.3699811914
HLA-B	51:01	0.2622559147	0.3735979591
HLA-A	31:01	0.06257607473	0.373939811
HLA-DQB1	03:02	0.1191560865	0.3837662304
HLA-C	08:02	0.5483190411	0.3868243105

HLA-A	02:01	0.3374718734	0.3923045945
HLA-C	15:02	0.0854397171	0.3995753568
HLA-B	13:02	0.542788398	0.4269431164
HLA-C	04:01	0.5026495179	0.5061149568
HLA-DRB1	07:01	0.4754458669	0.5141193376
HLA-B	57:01	1	0.5639125283
HLA-DQB1	03:03	0.557763495	0.5703277809
HLA-B	44:02	0.519746026	0.5879936814
HLA-C	07:02	0.4524248745	0.5895768036