На правах рукописи

Скитченко Ростислав Константинович

# ВЛИЯНИЕ ЧАСТОТНОГО СПЕКТРА АЛЛЕЛЕЙ НА РИСКИ ЗАБОЛЕВАНИЙ В РАМКАХ КОГОРТНЫХ ИССЛЕДОВАНИЙ

Специальность 1.5.7 – генетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Санкт-Петербург, 2022

Работа выполнена в международной лаборатории «компьютерных технологий» факультета информационных технологий и программирования Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет ИТМО», г. Санкт-Петербург

**Научный руководитель:** **Артемов Никита Николаевич**

доцент-исследователь центра геномного разнообразия, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», г. Санкт-Петербург

**Официальные оппоненты:** **Чекунова Елена Михайловна**

доктор биологических наук, старший преподаватель кафедры генетики, Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет», г. Санкт-Петербург

**Черняева Екатерина Николаевна**

кандидат биологических наук, заведующий лабораторией биоинформатики Центра постгеномных технологий, Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью», Федерального медико-биологического агентства, , г. Москва

**Ведущее учреждение:** Федеральное государственное бюджетное научное учреждение «Томский национальный исследовательский медицинский центр Российской академии наук», г. Томск

---

Защита диссертации состоится «___» _____ 2024 г. в ___ часов на заседании диссертационного совета 24.1.088.01 (Д 002.214.01) на базе Федерального государственного бюджетного учреждения науки Институт общей генетики им. Н.И. Вавилова РАН в конференц-зале Института по адресу: г. Москва, улица Губкина, д. 3, 119991, тел: (499) 135-62-13, Факс: (499) 132-89-62, e-mail: dissovet@vigg.ru

С диссертацией можно ознакомиться в библиотеке и на сайте www.vigg.ru Института общей генетики им. Н.И. Вавилова РАН.

Автореферат разослан «___» _____ г.

Учёный секретарь диссертационного совета, доктор биологических наук — Горячева И. И.

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**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 (ДНК – дезоксирибонуклеиновая кислота) связаны с рисками заболевания и как влияют на выживаемость пациента, является ключевой для развития персонализированной медицины и направленного поиска новых лекарственных препаратов (Dawood et al., 2008; Eheman et al., 2012; Jemal et al., 2011).

Так, ДНК-вариации с большим размером эффекта (ОШ≥2, [ОШ – 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 (Barton et al., 2017; Karczewski and Martin, 2020).

В контексте генетики сложных заболеваний, таких как воспалительные заболевания кишечника (ВЗК) (Rivas et al., 2011), гипертония (Surendran et al., 2016), рак (Huyghe et al., 2019), диабет (Sarnowski et al., 2019), аутизм (Leblond et al., 2019), шизофрения (Singh et al., 2022), научные изыскания ведутся преимущественно в двух направлениях. Во-первых, до сих пор проблемным является вопрос о фактическом количестве генов и/или генетических вариаций, вовлеченных в формирование того или иного признака. Во-вторых, генетическая основа большинства распространенных хронических заболеваний предполагает наличие нескольких генетических факторов, совместно определяющих индивидуальную предрасположенность (Schork et al., 2009). Последние результаты показывают, что риски таких полигенных заболеваний, как ВЗК, гипертония и шизофрения, связаны не только с носительством комбинации частых ДНК-вариантов, но и с предельно редкими RV, приводящими к “менделевской” форме патологии (Lifton, 1996; Reich and Lander, 2001; Rivas et al., 2011; Swales, 1985). Для простоты далее по тексту будут введены аббревиатуры CV (CV – common variants [частые варианты]) и RV (RV – rare variants [редкие варианты]) для ДНК-вариантов с аллельными частотами больше/равно 0.01 и менее 0.01, соответственно.

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. Данный эффект называется плейотропией. Классическими примерами плейотропии

являются пары фенотипов: серповидно-клеточная анемия и резистентность к малярии (Ashley-Koch et al., 2000), болезнь Хантингтона и снижение риска некоторых видов рака (Sørensen et al., 1999), 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 – один из немногих комплексных фенотипов, для которого не было проведено полногеномного ассоциативного исследования (GWAS – genome wide association studies). Это объясняется существующим предубеждением о том, что CV не несут значимого риска возникновения ФСГС. Однако последние исследования показывают, что CV 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.

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.

**Состояние научной разработанности.** Первые попытки сбора и систематизации информации об аллельных частотах в популяциях появились сразу после окончания проекта “Геном человека” (International Human Genome Sequencing Consortium et al., 2001). Среди достойных упоминания инициатив стоит выделить японский национальный проект JSNP (JSNP – The Japanese Society of Neuropsychopharmacology), который насчитывал информацию об 150 000 одиночных полиморфизмов (SNP – Single-nucleotide polymorphism) (Hirakawa et al., 2002). Однако первичные исследования фенотипических рисков, связанных с носительством аллелей на когортном уровне требовали большей информации о различиях в геномной последовательности ДНК между популяциями. Такие исследования стали возможны с завершением проекта HapMap и широким внедрением технологии генотипирования (International HapMap Consortium, 2003). Как следствие, это привело к внедрению полногеномных ассоциативных исследований (GWAS) (Ikegawa, 2012). В последствии, исследования CV приобрели большой масштаб в рамках национальных и международных биобанков – UK biobank и FinnGen (Kurki et al., 2022; Sudlow et al., 2015).

В свою очередь исследования RV получили массовость с удешевлением технологии экзомного секвенирования. Проекты Exome Sequencing Project (ESP), ExAC, gnomAD, стали последовательным развитием базы данных RV и оценки их возможной роли в моногенных заболеваниях (Karczewski et al., 2017; Karczewski et al., 2020; Kurki et al., 2022). Итогом

эволюционного развития баз данных секвенирования стала недавняя работа Karczewski и соавт. по анализу фенотипических ассоциаций RV в британском биобанке (Karczewski et al., 2022).

В России подобная инициатива получила свое развитие относительно недавно. На настоящий момент, крупнейшим хранилищем геномной информации является RuSeq (Barbitoff et al., 2021). Эта база данных основана в том числе на результатах текущего исследования и ссылается на соответствующую статью про полноэкзомное секвенирование жителей Северо-Западного региона России (Barbitoff et al., 2019).

Таким образом, данная работа построена на последних достижениях анализа частотного спектра в крупнейших когортах и представляет собой систематизацию существующих подходов к ассоциативным исследованиям и изучению плейотропных вариантов эффектов, а также их применению к новым, ранее не изученным заболеваниям.

**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.** 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 популяции.

Новым научным достижением является систематический анализ плейотропии 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 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 ДНК-варианты, 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 метаисследовании при создании российского экзомного браузера RuSeq (Barbitoff et al., 2021).

Выявленные генетические CV, 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. Найденная ДНК-вариация и соответствующий ей 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 (Zondervan and Cardon, 2007).

В работе используется методология и методы обработки данных генотипирования и секвенирования (Van der Auwera et al., 2013). В рамках этого исследования были рассмотрены одиночные точечные мутации и небольшие инсерции и делеции. Основной платформой при анализе популяционных данных служит программный продукт Hail 0.2 (Maes, 2021) для языка программирования Python, а также ряд биоинформатических библиотек для языка программирования R (Aulchenko et al., 2007), с помощью которых осуществлялся поиск генетических ассоциаций и статистическая обработка найденных результатов. На этапе оценки плейотропности генетических локусов использовалась кластеризация ДНК-вариантов по сходным фенотипам (Rousseeuw, 1987).

**Основные положения, выносимые на защиту:**

1) российская популяция, проживающая 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 частых ДНК-вариантов. Редкие ДНК-варианты, испытывающие избыточное эволюционное давление, как правило, селективно связаны с одним фенотипом;
3) риски заболевания в разных популяционных группах могут быть вызваны одними и теми же генетическими ДНК-вариантами, при этом их частота определяется факторами эволюционного давления, существующими в конкретной популяции.

**Личный вклад автора.** 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 – 2021г.) и ФГБУ «НМИЦ им. В. А. Алмазова» (2022г.), а также на конференциях: 1) XLVIII научная и учебно-методическая конференция Университета ИТМО (29 января – 1 февраля 2019 г., Санкт-Петербург, Россия); 2) 2020 European Society of Human

Genetics (6-9 июня 2020 г., Берлин, Германия); 3) BGRS/SB-2020: 11th International Multiconference “Bioinformatics of Genome Regulation and Structure/Systems Biology" (6-10 июля 2020 г., Новосибирск, Россия).

**Публикации.** Результаты исследования представлены в 4 научных публикациях, все из которых индексируются системами цитирования Scopus и Web of Science, а также входят в списки ВАК.

**Структура и объем диссертации.**

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.

## 1. ОБЗОР ЛИТЕРАТУРЫ

Этот раздел начинается с краткого ретроспективного обзора ключевых этапов развития генетики человека, охватывающего период с начала XX века до настоящего времени в области генетики. Детально рассматриваются временные периоды и значимые события, связанные с становлением технологий секвенирования и их развитием, связанные с проектом генома человека, а также с масштабными исследованиями постгеномной эры, такими как HapMap и GNOMAD.

Следующий этап подробно освещает предмет и область исследования в популяционной генетике, а также описывает применяемые в настоящее время методы для оценки рисков полигенных эффектов. Рассматриваются методы менделевской рандомизации и генетических корреляций. Далее особое внимание уделяется практическим аспектам выбора платформы для секвенирования и особенностям работы с экзомными данными.

В заключительной части литературного обзора рассматривается перспектива будущего развития генетических технологий, в частности, обсуждается интеграция искусственного интеллекта в сферу медицинской генетики. Также подробно описывается использование новых видов данных в клинической практике, таких как технологии секвенирования длинных прочтений и пангеном человека в качестве референсной последовательности ДНК.

## 2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

### 2.1. Сведения об анализируемых когортах.

**2.1.1. Получение референсной информации по частотам аллелей для изучения редких ДНК-вариантов.** Для проведения исследования экзомов жителей северо-западного региона России было использовано 694 образца, секвенированых с помощью Illumina HiSeq 2500 и HiSeq 4000. У участников исследования были собраны образцы ДНК и клинические данные, в сотрудничестве с ФГБНУ "НИИ АГиР им. Д. О. Отта" и СПБ ГБУЗ Городской Больницей №40. Данные частично описаны в ранних работах: Glotov O.S. 2019 (Glotov et al., 2019) и Barbitoff Y.A. 2018 (Barbitoff et al., 2018). Большинство людей по самоотчету были русской национальности (~80%) и европеоидной расы; участники исследования преимущественно проживали в Северо-Западном регионе России.

Был проведен анализ присутствия найденных генетических ДНК-вариантов в базе известных ДНК-вариантов - dbSNP (версия 151) (The Single Nucleotide Polymorphism Database). Оценка распространенности аллелей риска аутосомно-рецессивных заболеваний, выбранных на основе известных патогенных ДНК-вариантов из ClinVar, проводилась в сравнении с базой данных gnomAD r. 2.1 (The Genome Aggregation Database) (Karczewski et al., 2020). Основными инструментами для работы с геномными данными стали языки программирования python и R и их библиотеки: numpy, scipy.stats, ggplot2, reshape2, dplyr.

**2.1.2. Получение результатов ассоциативных исследований для исследования генетических локусов ответственных за систематическое проявление множественных ассоциаций сразу с несколькими сложными признаками.** Результаты полногеномных ассоциативных исследований были получены с сайта лаборатории Бенджамина Нила (UK Biobank GWAS results imputed v2). Набор данных включал как генотипированные, присутствовавшие на микрочипе, так и “импутированные” ДНК-варианты, в общей сложности 10 894 597 ДНК-вариантов. Из полного списка наследуемых фенотипов было отобрано 543 сложных признака, которые имеют ненулевые оценки наследуемости ($h^2$, $p<0.01$). Для дальнейшего анализа использовались ДНК-варианты (одиночные точечные мутации и небольшие инсерции и делеции), значимо ассоциированные с каждым из фенотипов (p-значение $< 5\times10^{-9}$). Суммарная статистика ассоциаций для каждого варианта по каждому признаку была объединена в сводную матрицу для дальнейшего анализа.

**2.1.3. Исследование плейотропии для объяснения дисбаланса частоты аллелей на примере когортного исследования фокального сегментарного гломерулосклероза.** Когорта “случаев” формировалась из пациентов, участвовавших в многоцентровом исследовании, проводимом национальным институтом здоровья США (NIH – National Institutes of Health) и Вашингтонским университетом (Ст. Луис, США). ДНК образцы

“случаев” были собраны у пациентов с диагнозом ФСГС, подтвержденным методом биопсии. Файлы секвенирования контрольных образцов полноэкзомного секвенирования были получены из нескольких когорт базы данных dbGAP, которые состоят из пациентов без нефропатологий в истории болезни.

Данные были получены с помощью секвенирования генетической панели, состоящей из 2,482 генов ("подоцитный экзом"). В данную панель были включены:

- 5 генов, ранее идентифицированные как гены риска для семейной формы ФСГС;
- 200 генов, функционально связанных с предыдущей группой генов;
- 677 генов с высокой экспрессией в микропрепарированных гломерулах человека;
- 1 600 генов, которые высоко экспрессируются в мышиных подоцитах и имеют ортологи у человека.

Для анализа ДНК-вариантов (одиночные точечные мутации и небольшие инсерции и делеции) был использован набор инструментов для геномного анализа GATK (Genome Analysis Toolkit). Помимо основных программных продуктов, применяемых в анализе вариаций человеческого генома, GATK предоставляет пользователю набор подробно документированных руководств по последовательному практическому использованию входящих в него инструментов (best practice), в соответствии с которыми проводился анализ данных секвенирования и анализ качества данных ДНК-вариантов (Van der Auwera et al., 2013).

Для дальнейшего анализа были использованы следующие программные продукты: 1) библиотека AutoGMM (Athey et al., 2019) для языка программирования python (кластеризация для учёта популяционной стратификации “случай-контроль”); 2) библиотека MatchIt (Ho et al., 2011) для языка программирования R (“случай-контроль” подбор образцов); 3) библиотека Hail 0.2 для языка программирования python (построение логики ассоциативных исследований); 4) библиотеки Genabel (Aulchenko et al., 2007) и AssotesteR для языка программирования R (ассоциативные тесты и контроль “случай-контроль” подбора образцов).

**2.2. Обзор методов использованных в исследовании.** В данном подразделе описаны методы, использованные для получения результатов изложенных в разделе 3. Среди этих методов присутствуют: 1) анализ силуэта, 2) метод главных компонент, 3) смешанные гауссовские модели, 4) одно-вариантные тесты ассоциации, 5) анализ RV в ассоциативных исследованиях.

Особое внимание уделяется проблеме выбора и описания различных видов агрегирующих статистических тестов для анализа RV. Среди них включаются следующие категории: 1) тесты мутационной нагрузки, 2) адаптивные тесты мутационной нагрузки, 3) тесты на основе дисперсионных компонент, и 4) комбинированные тесты мутационной нагрузки. В последней части подраздела подробно рассматриваются методы проведения мета-анализа на основе одновременного применения нескольких статистических тестов. Например, в частности, рассматривается используемый в диссертации метод Фишера.

## 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. 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.

Эволюционный процесс не является постоянным и равномерным для разных популяций. Негомогенная окружающая среда, глобальные события, такие как эпидемии, войны, голод, вносят свой вклад в неравномерность распределения частот аллелей между популяциями. Из-за этого аллель-частотный спектр может значительно различаться в малоизученных народностях, недостаточно представленных в крупномасштабных проектах.

Недостаток референсной информации о частотном спектре аллелей в российской популяции затрудняет оценку рисков заболеваний, прогнозирование ответа пациентов на терапию и постановку точных клинических диагнозов. Наиболее интерпретируемыми с точки зрения рисков являются кодирующие RV, которые чаще всего ассоциированы с менделевскими (моногенными) формами заболеваний. Таким образом, создание крупной когорты экзомных сиквенсов в РФ – актуальная задача с высокой прикладной ценностью.

**3.1.1. Составление русской этнической когорты.** В рамках проводимого метаисследования была составлена когорта из 694 образцов, собранных из различных независимых клинических исследований (Barbitoff et al., 2018a; Barbitoff et al., 2018b; Barbitoff et al., 2018c). Набор данных содержал секвенированные образцы пациентов, агрегированные из различных исследовательских и клинических проектов (как контрольную когорту, так и когорту пациентов с диагностированными заболеваниями (основные фенотипы: диабет зрелого типа у молодых (MODY – maturity-onset diabetes of the young), диабет 2 типа (T2D – type 2 diabetes), ожирение, расстройства аутистического спектра (ASD –

autism spectrum disorder), заболевания соединительной ткани (CTD – connective tissue disease), нейрофиброматоз). По отчёту около 80% пациентов были русской национальности 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.** На первом этапе был охарактеризован спектр генетических ДНК-вариантов, обнаруженных в исследуемой выборке. Для полного набора из 694 участников исследования было определено в общей сложности 463 100 генетических ДНК-вариантов внутри целевых областей экзома. Из них для 420 187 ДНК-вариантов (90.7%) есть информация в dbSNP (версии 151), остальные 42 913 ДНК-вариантов ранее не были учтены в базе данных (не было найдено идентификатора базы данных rsID).

**3.1.3. Корреляция аллельных частот между исследуемой когортой и открытыми базами данными.** Далее была проведена оценка соответствия между частотами альтернативных аллелей в исследуемой выборке и в экзомах из базы данных gnomAD r. 2.1. С этой целью была произведена оценка коэффициентов линейной регрессии для аллельных ДНК-вариантов, которые имеют среднее покрытие не менее 15X в gnomAD. В итоге была обнаружена сильная корреляция между аллельными частотами в gnomAD и исследуемым набором данных ($R^2$=0.96; рис. 1).

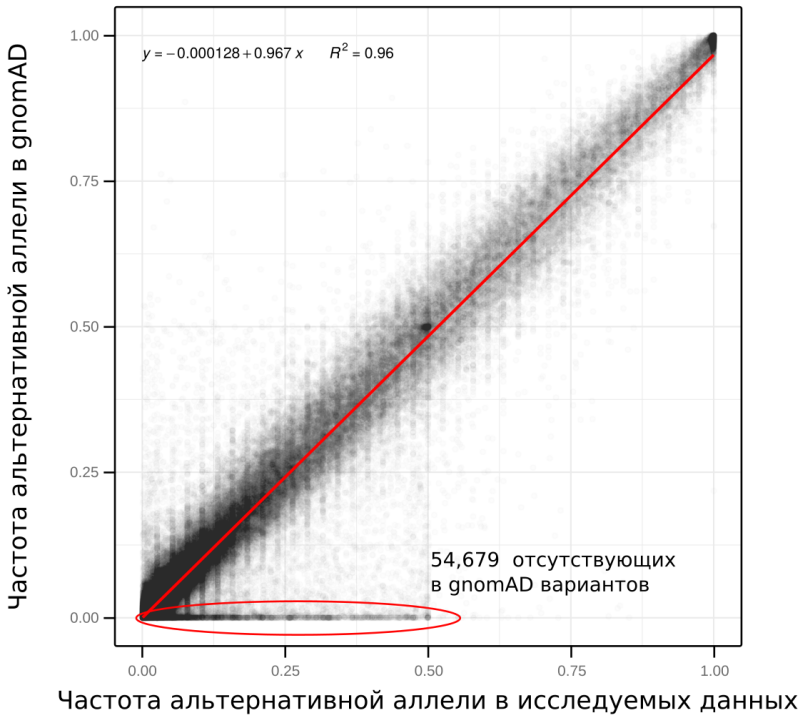


Рисунок 1. Диаграмма разброса частот альтернативных аллелей в наборе данных в сравнении с частотами на основе gnomAD; исключены сайты, использующие только gnomAD, а также мульти-аллельные записи и плохо покрытые регионы

Также стоит отметить, что в gnomAD отсутствовало 54 679 ДНК-вариантов. Большинства из них не было также и в dbSNP build v. 151 (37 338; 68.2%), что может свидетельствовать о том, что эти полиморфизмы составляют специфический компонент генетической структуры населения Северо-Запада России.

Следующим этапом исследования было изучение взаимосвязи между когортой жителей Северо-Запада России (NWR – Northwest Russia) и основными популяциями, присутствующими в gnomAD. Сначала была изучена корреляция между частотами альтернативных аллелей в NWR и в пяти основных глобальных популяциях: африканской (AFR – Africans), смешанной американской (AMR – Admixed Americans), восточноазиатской (EAS – East Asians), нефинской европейской (NFE – Non-Finnish European) и южноазиатской (SAS – South Asians). Как и ожидалось, частоты аллелей NWR более всего сопоставимы с частотами аллелей, полученными из популяции NFE (рис. 2).

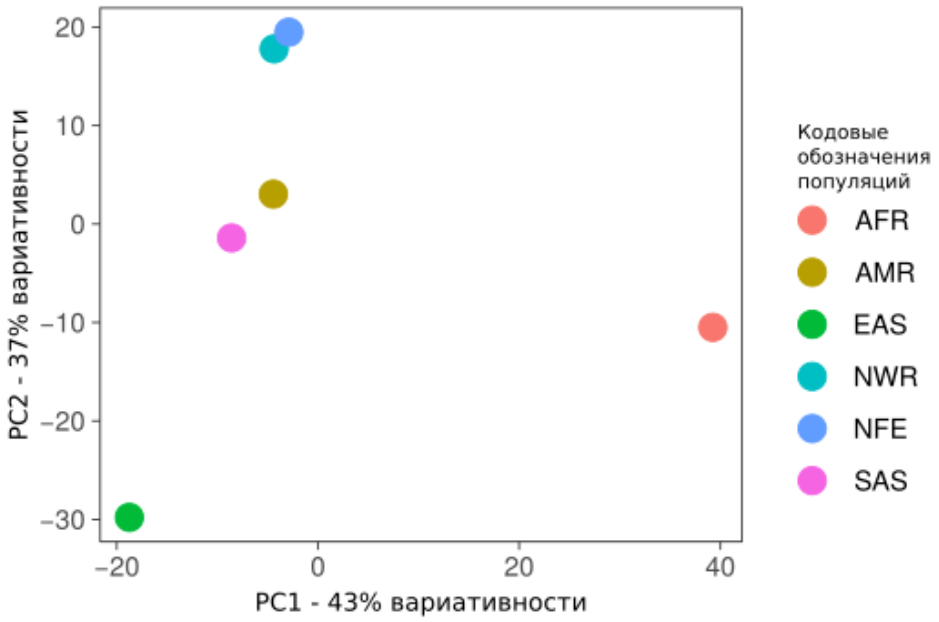


Рисунок 2. Анализ частот аллелей 121 171 ДНК-вариантов, представленных в Северо-Западной России и во всех популяциях gnomAD, на основе анализа главных компонент. AFR - африканская; AMR - смешанная американская; EAS - восточноазиатская; NFE - нефинская европейская; NWR - северо-запад России; SAS - южноазиатская.

**3.1.4. Оценка частот патогенных аллелей в исследуемой когорте.** Далее анализ был сосредоточен на ДНК-вариантах, которые являются патогенными согласно базе клинических полиморфизмов ClinVar и при этом имеют относительно низкую (<0.5%) частоту в глобальной популяции gnomAD и являются строго гетерозиготными в исследуемом наборе данных. Далее выборка пациентов была сужена до лиц со строго установленным фенотипом без признаков тяжелых заболеваний. В результате фильтрации была получена когорта из 372 не родственных между собой лиц, для которых было известно 314 902 из

463 100 (68.0%) ДНК-вариантов. В исследуемой выборке генотипической информации было обнаружено несколько примеров широко распространенных известных патогенных ДНК-вариантов аутосомно-рецессивных заболеваний. Наиболее часто встречались ДНК-вариант rs5030858 в гене *PAH* (MIM#612349; gnomAD_NFE_AF=0.0015; $p=7.9\times10^{-4}$), ДНК-вариант rs36209567 в гене фактора свертываемости крови VII (*F7*; MIM#613878) (gnomAD_NFE_AF=0.001, p=0.001), ДНК-вариант rs61754365 в гене *TYR*, связанный с тирозиназо-отрицательным альбинизмом (gnomAD_NFE_AF=$3.2\times10^{-4}$; $p=1.1\times10^{-4}$) (Takeda et al., 1990).

Интересно, что для генов, связанных с риском развития аутосомно-рецессивных заболеваний, не было выявлено высоко распространенных патогенных или вероятно патогенных ДНК-вариантов, отсутствующих в базах данных ClinVar или dbSNP, как и в генах, связанных с аутосомно-рецессивными заболеваниями. Это может указывать на то, что, по крайней мере, для рецессивных патологий, большая часть генетических детерминант является общей для российской и других популяций.

При этом риски, связанные с RV, не обязательно ограничиваются менделевскими заболеваниями. Так, в частности, известны примеры RV, вызывающих моногенные формы полигенных заболеваний (Barton et al, 2017). На следующих этапах работы были рассмотрены эффекты взаимодействия между RV и CV и 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.** CV обычно несут низкие риски заболеваний. В связи с этим для достижения достаточной статистической мощности необходимы когорты большого размера. Это оказывается чрезвычайно затратным для методов NGS, но доступным для микрочипового генотипирования. В качестве примера успешной агрегации информации о генетической и фенотипической изменчивости популяции можно привести проект UK Biobank (UKB), который собрал результаты ассоциативных исследований более чем для 10 000 признаков на когорте в более чем 500 000 жителей Британии (Bycroft et al., 2018).

**3.2.1. Оценка соотношения фенотипической и генетической информации в популяционных данных UK Biobank.** Для исследования человеческого генома на предмет количественной оценки частоты встречаемости плейотропных локусов 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 генотипированные, так и импутированные ДНК-варианты, в общей сложности 10 894 597 ДНК-вариантов. Анализ был сосредоточен только на 543 сложных признаках, которые были отфильтрованы по оценке наследуемости ($h^2$<0.01). Оказалось, что в 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 ДНК-вариантом. Интересно, что в результате анализа наблюдались многочисленные множественные ассоциации во всем наборе данных. Около половины всех SNP (230 296; 49.21%) имели более одного ассоциированного фенотипа, а для 57 856 (12.34%) SNP соответствовало более чем 10 ассоциированных фенотипов.

**3.2.2. Кластеризация фенотипической информации.** Затем была предпринята кластеризация тех признаков, которые имеют значительную долю общей генетической архитектуры. В качестве меры расстояния для кластеризации была использована восстановленная фенотипическая корреляция. Методология проведения кластеризации описана в полном тексте диссертации.

**3.2.3. Закономерность степени плейотропности относительно функционального эффекта варианта и аллельной частоты.** Важным выводом из результатов кластеризации стало то, что в данных присутствовало относительно небольшое количество RV с более чем 5 ассоциированными кластерами, несмотря на гораздо более высокую распространенность RV в исследуемом наборе данных. Другими словами, RV имеют тенденцию быть менее плейотропными по сравнению с распространенными (рис 3).

Склонность к низкой степени плейотропии RV может быть вызвана сильным очищающим естественным отбором, действующим против высокоплейотропных полиморфизмов с серьезными эффектами, в результате чего все плейотропные ДНК-варианты имеют более низкие размеры эффектов и более высокую частоту.

**3.3. Вклад плейотропии в объяснение дисбаланса частоты аллелей на примере когортного исследования фокального сегментарного гломерулосклероза.** Предположение о наличии очищающего естественного отбора согласуется с наблюдениями, что естественный отбор против дезадаптирующих ДНК-вариантов действует на некоторые формы сложных признаков (Gazal et al., 2018; Zeng et al., 2018). Таким образом, если плейотропные ДНК-варианты, влияющие на заболевания человека, имеют тенденцию быть пагубными, то можно ожидать, что высокоплейотропные ДНК-варианты будут удаляться из популяции или сохраняться при низких аллельных частотах (Paaby and Rockman, 2013). Чтобы добавить убедительности данной теории, было проведено кросс-популяционное когортное исследование на примере фокального сегментарного гломерулосклероза.

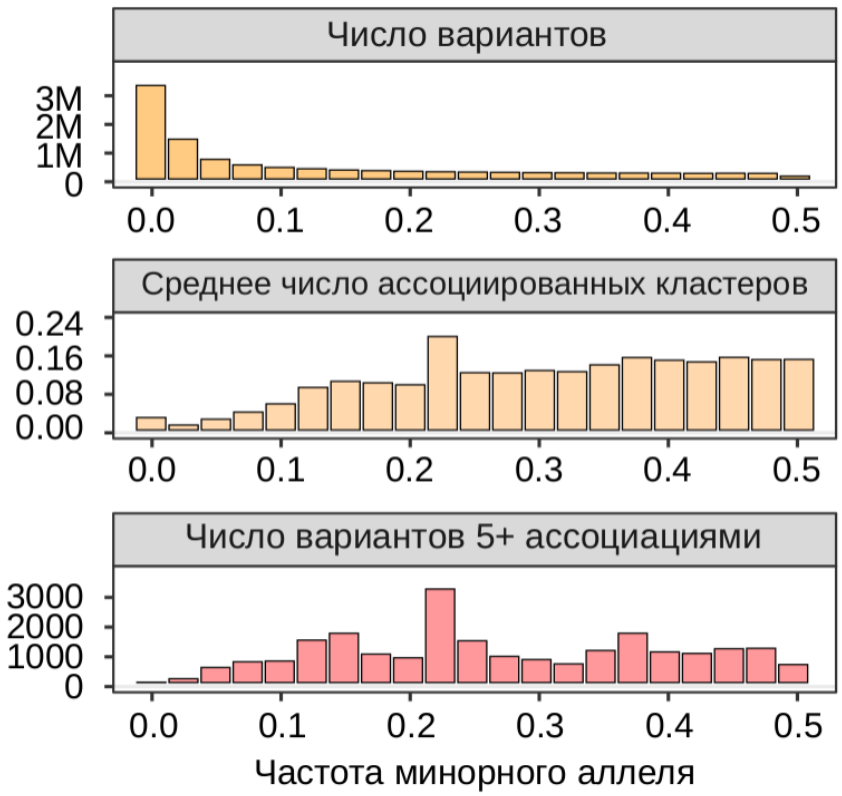


Рисунок 3. Сводная статистика ассоциаций для ДНК-вариантов с различной частотой минорных аллелей. Значения агрегированы по бинам размером 0.025.

Фокальный сегментарный гломерулосклероз (ФСГС) является основной причиной нефротического синдрома (Gipson et al., 2016) и встречается в 20-30% случаев хронической почечной недостаточности, а патогенез напрямую связан с подоцитами (Freedman et al., 2009). В настоящее время частота ФСГС оценивается как 1.9 случаев на миллион среди европейцев и 6.8 – среди представителей афроамериканской популяции.

Первые генетические исследования первичной формы ФСГС были основаны на наблюдении более высокой распространенности ФСГС в африканских и афроамериканских популяциях и гипотезе о наличии положительного отбора в отношении генов предрасположенности к ФСГС. Открытие ДНК-вариантов G1/G2 в гене *APOL1*, приводящих к риску развития ФСГС, дало одно из возможных объяснений дисбаланса распространенности заболевания среди африканской и европейской популяций. Так, ДНК-варианты G1/G2 также связаны со способностью клетки противостоять трипаносомозу, болезни, преимущественно локализованной на африканском континенте (Genovese et al., 2010).

**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. Когорта “случаев” формировалась из пациентов, участвовавших в многоцентровом исследовании, проводившимся национальным институтом здоровья США (NIH) и Вашингтонским университетом (Ст. Луис, США). ДНК образцы “случаев” были собраны у пациентов с диагнозом ФСГС, подтвержденным методом биопсии. Контрольные образцы полноэкзомного секвенирования получены из нескольких когорт базы данных dbGAP, которые состоят из пациентов без нефропатологий в истории болезни.

Данные были получены с помощью секвенирования генетической панели, состоящей из 2 482 генов (“подоцитный экзом”). В данную панель были включены:

- 5 генов, ранее идентифицированные как гены риска для семейной формы ФСГС;
- 200 генов, функционально связанных с предыдущей группой генов;
- 677 генов с высокой экспрессией в микропрепарированных гломерулах человека;
- 1 600 генов, которые высоко экспрессируются в мышиных подоцитах и имеют ортологи у человека.

Результаты полноэкзомного секвенирования прошли совместный анализ поиска ДНК-вариантов для создания набора данных “случай-контроль”. Получившаяся генотипическая информация была подвергнута процессу обработки, в результате которого финальная таблица включала в себя 499 образцов “случаев”, 10 557 контрольных образцов, 131 179 ДНК-вариантов.

**3.3.2. Контроль качества и популяционная стратификация генотипических данных.** Для учета неоднородности происхождения образцов из-за их популяционной стратификации был проведен совместный PCA-анализ генотипов когорт “случаев” и “контролей” с их последующей кластеризацией. Дальнейшее сопоставление “случай-контроль” было проведено для каждого кластера с помощью пакета Matchit (Ho et al., 2011). Окончательный набор данных составил 358 “случаев” и 1 466 “контролей” для европейского кластера и 141 “случай” и 595 “контролей” для африканского кластера (рис. 4).

“Анализ мощности” (eng. “Power analysis”) европейского набора данных показал многократное превосходство настоящего исследования по статистической мощности над другими когортными исследованиями ФСГС, а также то, что при существующем количестве случаев значимое увеличение мощности не может быть достигнуто при большем количестве контролей.

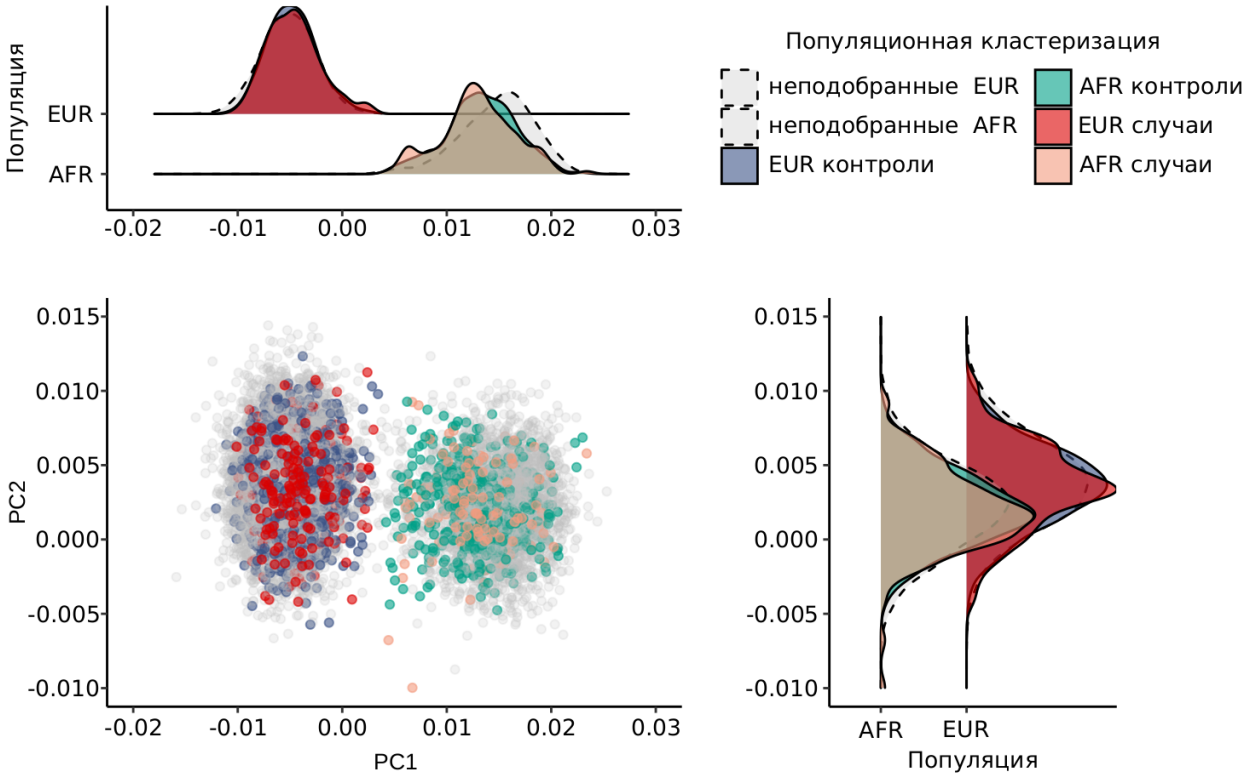


Рисунок 4. PCA, иллюстрирующий уровень соответствия “случай-контроль” в европейской и африканской популяциях

**3.3.3. Анализ вариантов на наличие ассоциации с риском возникновения фокального сегментарного гломерулосклероза.** В полном тексте диссертации дополнительно описывается анализ CV для европейской когорты (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*). Эти результаты не были воспроизведены в независимой африканской когорте.

Анализ RV (RVAS – Rare Variant Association Analysis) проводился для миссенс- и PTV-вариантов (PTV – protein truncating variants) с популяционной частотой менее 0.01, 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. 5).

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

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. 6).

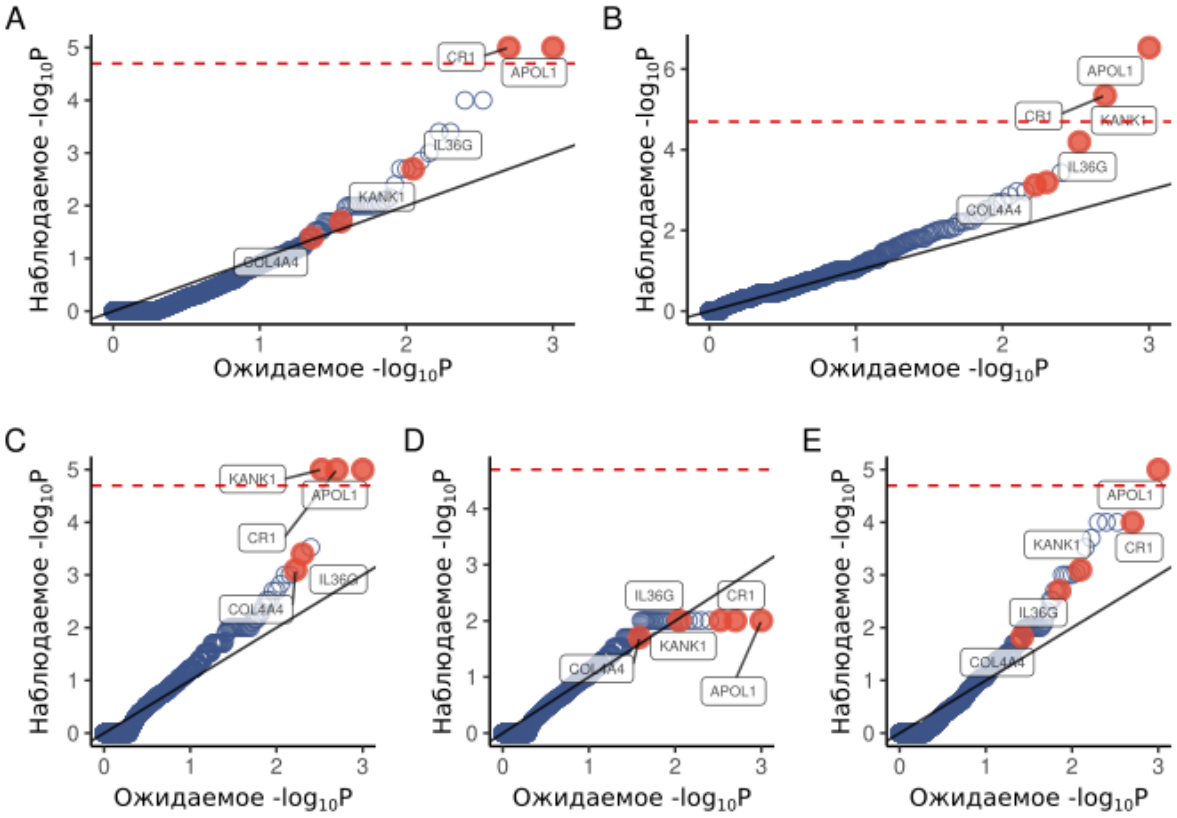


Рисунок 5. QQ-график для каждого из тестов на редкие ДНК-варианты, включенных в агрегирующий метод Саймса. (A) C-alpha тест; (B) точный тест Фишера; (C) WSS тест; (D) KBAC тест; (E) ASUM тест

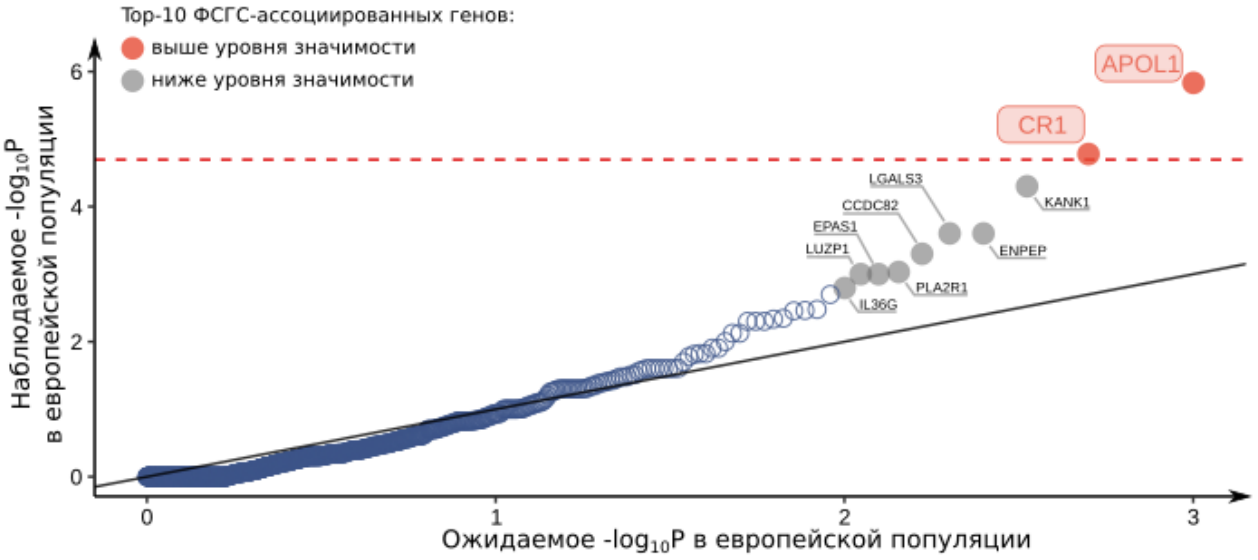


Рисунок 6. Исследование ассоциации редких ДНК-вариантов в европейском кластере (gnomAD_NFE_AF < 0.01; миссенс- и PTV-варианты; метод Саймса – точный тест Фишера, С-альфа, ASUM, WSS, KBAC);

Значимо ассоциированные с ФСГС гены, установленные для европейской когорты, повторно проверялись на реплицируемость в когорте африканского происхождения (табл. 5).

Ни *APOL1*, ни *CR1* не были воспроизведены с помощью анализа RV. Ранее наблюдавшийся положительный отбор, действующий на ДНК-варианты *APOL1* (Genovese et al., 2010) в африканской популяции, позволяет предположить, что ДНК-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*.

Сначала были определены RV в европейской когорте, определяющие сигнал ассоциации в RVAS для генов *APOL1* и *CR1* (рис. 7). Было выявлено четыре ДНК-варианта: парный локус G1 в *APOL1* (rs60910145 и rs7388531918), а также два близких ДНК-варианта в *CR1* – rs17047661 и rs17047660. Все четыре ДНК-варианта прошли порог значимости при введении поправки на множественную проверку гипотезы (p=0.05/10 ДНК-вариантов=0.005). Далее эти четыре ДНК-варианта приняли участие в репликации на образцах из африканской когорты. Оба ДНК-варианта из пары G1 *APOL1* и rs17047660 в *CR1* успешно прошли порог значимости репликации (p=0.05/4=0.0125) (рис. 7).

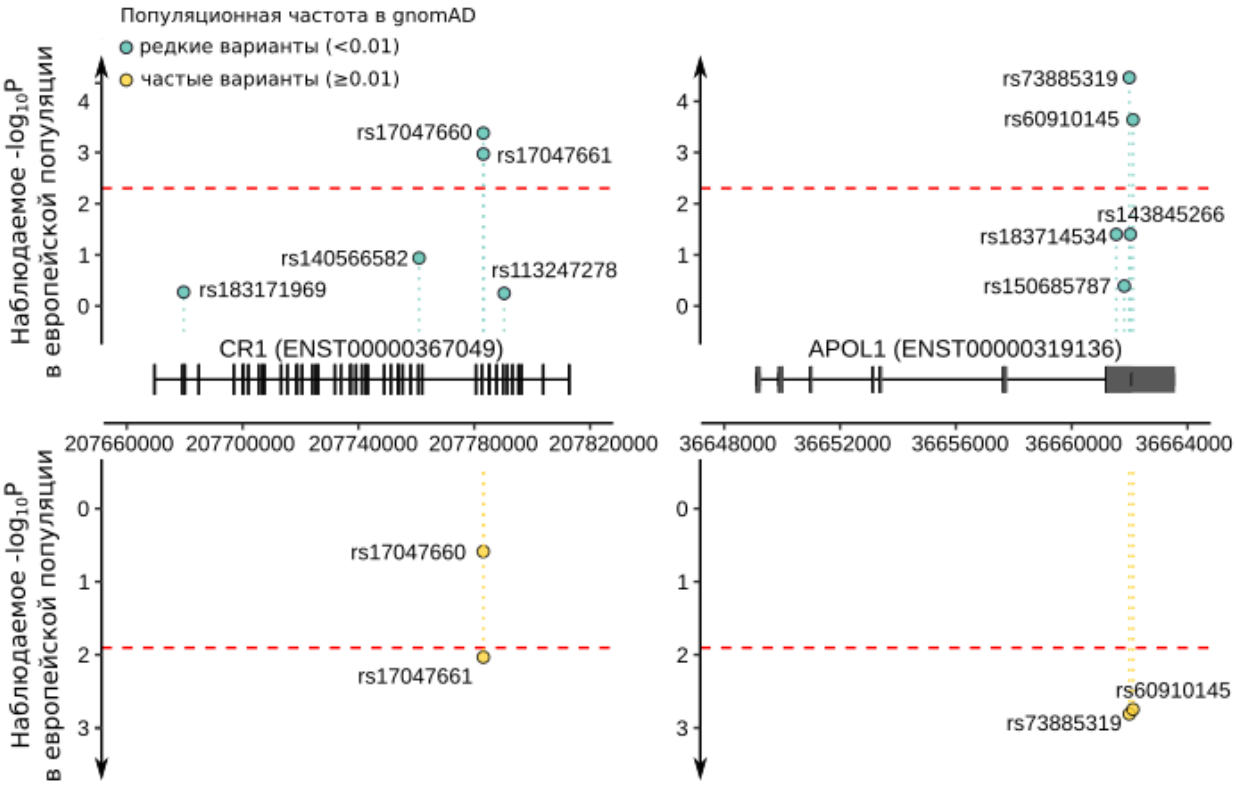


Рисунок 7. Результаты репликации между европейской и африканской популяцией.

Частоты аллелей для реплицированных ДНК-вариантов значительно отличаются между популяциями, что свидетельствует о положительном отборе (rs60910145: gnomAD_NFE_AF=$8.6\times10^{-5}$, gnomAD_AFR_AF=0.23;

rs73885319: gnomAD_NFE_AF=$1.1\times10^{-4}$, gnomAD_AFR_AF=0.23; rs17047661: gnomAD_NFE_AF=$3.0\times10^{-3}$, gnomAD_AFR_AF=0.62).

**3.3.4. Плейотропия как объяснение высокой частоты аллелеи в популяции.** Ген *CR1* является важным участником системы комплемента. Исходя из гипотезы 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 (HLA – Human Leukocyte Antigens). 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 обладают этим HLA-генотипом. Между группами производилась оценка наличия аллельного ДНК-варианта rs17047661. Пять номинально значимых ассоциаций в европейской популяции были воспроизведены в африканской популяции. Наиболее обогащенным HLA-генотипом как в европейской, так и в африканской популяциях был HLA-DRB*15:01 (P=0.026; P=0.094; соответственно) (табл. 1).

Таблица 1. Топ-5 результатов анализа на присутствие rs17047661 в конкретном HLA-типе.

| HLA тип | генотип | p.value в AFR | pvalue в NFE |
|---|---|---|---|
| HLA-DQB1 | 06:02 | 0.41 | 0.0051 |
| HLA-A | 29:02 | 0.50 | 0.027 |
| HLA-DRB1 | 15:01 | 0.094 | 0.027 |
| HLA-A | 02:01/30 | 1 | 0.035 |
| HLA-C | 16:01 | 0.24 | 0.043 |

Как выяснилось, HLA-DRB1*15:01 широко распространён в регионах с высоким уровнем заболеваемости малярией, что согласуется с известным протективным эффектом ФСГС против этого заболевания (рис. 8).

Недавние исследования уже описали влияние системы комплемента при различных гломерулопатиях (Mathern and Heeger, 2015). Ранее высказанная гипотеза об иммунном компоненте в ФСГС находит новые подтверждения благодаря обнаруженному ассоциированному гену *CR1*. Аутоантитела, реагирующие на экспрессируемые почками аутоантигены или комплексы антитело-антиген, оседающие в почках, считаются возбудителями различных заболеваний почек человека. Известны случаи C1-опосредованного воспаления и отложения C3 (Poppelaars and Thurman, 2020). Также было показано, что ингибирование C3 снижает протеинурию в животных моделях (Salant et al., 1980). Что касается конкретно ФСГС, то в пораженных гломерулах часто наблюдаются отложения IgG и C3, но патогенез до сих пор неясен, а терапия против системы комплемента не изучена (Strassheim et al., 2013).

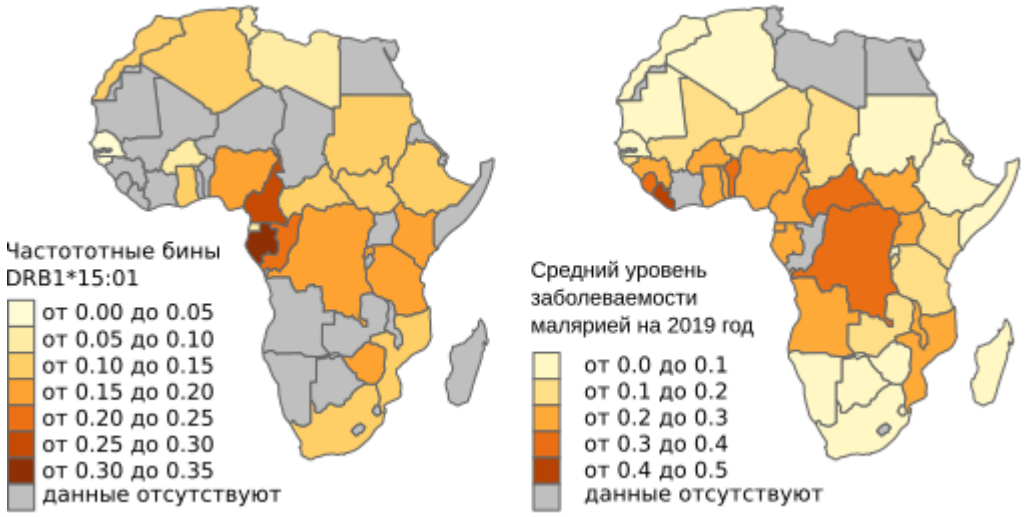


Рисунок 8. Распределение HLA-типа DRB1*15:01(слева) и малярии (справа) на африканском континенте. Данные по DRB1*15:01 и малярии были получены с сайтов http://www.allelefrequencies.net/ и https://malariaatlas.org/ соответственно.

*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) (Poppelaars and Thurman, 2020).

Усиление действия альтернативного пути, включающего активацию врожденного иммунного ответа, согласуется с нашими наблюдениями об эволюционном давлении на *CR1*, связанном с защитой от малярии. Была выдвинута гипотеза, что *CR1* является рецептором эритроцитов, используемым P. falciparum для инвазии независимо от сиаловой кислоты (Tham et al., 2010).

## 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 РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

**Публикации в журналах, соответствующих Перечню ВАК, и индексируемые системами цитирования Web of Science и Scopus:**

1. Zlotina A, 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 / A. Zlotina, O. Melnik, Y. Fomicheva, **R. Skitchenko**, A. Sergushichev, E. Shagimardanova, O. Gusev, G. Gazizova, T. Loevets, T. Vershinina, I.

Kozyrev, M. Gordeev, E. Vasichkina, T. Pervunina, A. Kostareva // BMC Med. Genomics. 2020. Vol. 13, № 1. P. 175;

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 / O.S. Glotov, E.A. Serebryakova, M.E. Turkunova, O.A. Efimova, A.S. Glotov, Y.A. Barbitoff, Y.A. Nasykhova, A.V. Predeus, D.E. Polev, M.A. Fedyakov, I.V. Polyakova, T.E. Ivashchenko, N.Y. Shved, E.S. Shabanova, A.V. Tiselko, O.V. Romanova, A.M. Sarana, A.A. Pendina, S.G. Scherbak, E.V. Musina, A.V. Petrovskaia-Kaminskaia, L.R. Lonishin, L.V. Ditkovskaya, L.A. Zhelenina, L.V. Tyrtova, O.S. Berseneva, **R.K. Skitchenko**, E.N. Suspitsin, E.B. Bashnina, V.S. Baranov // Mol. Med. Report. 2019. Vol. 20, № 6. P. 4905–4914;
3. **Skitchenko R.K.*** & Barbitoff Y.A.* et al. Whole-exome sequencing provides insights into monogenic disease prevalence in Northwest Russia / **R.K. Skitchenko**, Y.A. Barbitoff, O.I. Poleshchuk, A.E. Shikov, E.A. Serebryakova, Y.A. Nasykhova, D.E. Polev, A.R. Shuvalova, I.V. Shcherbakova, M.A. Fedyakov, O.S. Glotov, A.S. Glotov, A.V. Predeus // Mol. Genet. Genomic Med. 2019. Vol. 7, № 11. P. e964. (* – совместное первое авторство);
4. Shikov A.E. et al. Phenome-wide functional dissection of pleiotropic effects highlights key molecular pathways for human complex traits / Shikov AE, **R.K. Skitchenko**, Predeus AV, Barbitoff YA // Sci. Rep. 2020. Vol. 10, № 1. P. 1037.