

На правах рукописи

Борисов Олег Витальевич

**Исследование вклада генетических факторов в варибельность
морфофункциональных характеристик мышечных волокон человека с
помощью комбинации биоинформатических методов полногеномного
ассоциативного анализа**

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2020

Работа прошла апробацию на кафедре молекулярной и трансляционной медицины Федерального государственного автономного образовательного учреждения высшего образования "Московский физико-технический институт (национальный исследовательский университет)"

Научный руководитель: Эдуард Викторович Генерозов, кандидат биологических наук, доцент, заведующий лабораторией молекулярной генетики человека Федерального государственного бюджетного учреждения «Федеральный научно-клинический центр физико-химической медицины Федерального Медико-биологического Агентства».

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук

Защита состоится «30» ноября 2020 года в 14.00 на заседании диссертационного совета ФБМФ03.01.09.003, по адресу: 141701, г. Долгопрудный Московской обл., Институтский переулок, д. 9.

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национальный исследовательский университет) <https://mipt.ru/education/post-graduate/soiskateli-biologicheskie-nauki.php>.

Работа представлена «24» августа 2020 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук в соответствии с п.3.1 ст. 4 Федерального закона «О науке и государственной научно - технической политике»

Общая характеристика работы

Актуальность исследования. Среди современных методов генетического анализа важное место занимает полногеномное ассоциативное исследование. Оно применяется для поиска ассоциации между генетическими факторами и изучаемыми наследственными признаками. Благодаря стремительному развитию технологий генотипирования такие исследования позволяют изучать ассоциации с сотнями тысяч генетических вариантов. За последние годы этот метод позволил установить десятки тысяч генетических ассоциаций с тысячами различных признаков человека. Следующий этап после установления ассоциации между генетическими факторами и наследственными признаками заключается в понимании стоящих за ней механизмов. Полногеномное ассоциативное исследование позволяет выявлять ассоциации отдельно для каждого из исследуемых генетических вариантов. Информативным дополнением к этому подходу является метод полнотранскриптомного ассоциативного исследования, который позволяет устанавливать связь наследственных признаков с изменением предсказанных уровней экспрессии генов. Большинство наследственных признаков человека ассоциированы с несколькими или значительным числом генетических вариантов. Для поиска взаимосвязи между изучаемым признаком и совокупным эффектом этих вариантов используется метод полногеномных баллов, который позволяет объединить эффекты отдельных генетических вариантов в единую меру. Все три метода (полногеномное ассоциативное исследование, полнотранскриптомное ассоциативное исследование и моделирование полногеномных баллов) могут проводиться с использованием единых входных генетических данных, полученных в ходе высокоплотного генотипирования. Это делает возможным их совместное применение при наличии таких данных, что позволяет более детально исследовать генетическую архитектуру изучаемых наследственных признаков.

Особенную актуальность представляет совместное применение этих методов к сложным в изучении признакам, для анализа которых проблематично сформировать репрезентативные выборки достаточного объема. Примером могут служить исследования редких наследственных патологий или анализ таких признаков, для изучения которых требуется проведение дорогостоящих и инвазивных исследований. К такой категории относятся исследования морфофункциональных характеристик волокон скелетных мышц, для прямой оценки которых требуется проведение инвазивного биопсийного исследования. В скелетных мышцах человека выделяют волокна нескольких типов, их соотношение определяется как средовыми, так и генетическими факторами. Исследование мышечных волокон у монозиготных и дизиготных близнецов показало, что генетический вклад в соотношение мышечных волокон составляет 40-50%. Соотношение мышечных волокон различного типа

определяет потенциал к развитию выносливости, силовым характеристикам и часто определяет успешность в различных видах спортивной деятельности. Также установлена связь состава мышечных волокон и предрасположенности к ряду хронических заболеваний, в первую очередь – компонентам метаболического синдрома.

Таким образом, изучение генетических факторов морфофункциональных характеристик мышечных волокон представляет актуальность для медицинских исследований и возможности оценки функционального потенциала в спортивном отборе. Совместное применение методов полногеномного ассоциативного исследования, полнотранскриптомного ассоциативного исследования и моделирования полногеномных баллов делает возможным систематическое изучение генетических факторов сразу на трех уровнях: ассоциации с отдельными генетическими вариантами, с уровнями экспрессии генов, а также с объединенным эффектом генетических вариантов.

Цели и задачи исследования. Цель настоящего исследования – применение биоинформатического подхода, основанного на комбинации методов полногеномного и полнотранскриптомного ассоциативного анализа и моделирования полногеномных баллов, для анализа генетических факторов, детерминирующих морфофункциональные характеристики мышечных волокон человека.

Для достижения цели были поставлены следующие задачи:

- 1) формирование генетически и фенотипически охарактеризованных групп сравнения для проведения полногеномного ассоциативного анализа;
- 2) биоинформатическая обработка исходных генетических данных, осуществление контроля качества, фазирования и импутации генотипов;
- 3) выполнение полногеномного ассоциативного исследования, полнотранскриптомного ассоциативного исследования с использованием предсказанных уровней экспрессии генов, а также анализ полногеномных баллов в пределах сформированных групп сравнения;
- 4) создание программного конвейера для автоматизации импутации генотипов и полногеномных ассоциативных исследований.

Научная новизна и практическая ценность работы. В рамках исследования сформирована экспериментальная группа из 171 человека. Для участников группы проведено высокоплотное генотипирование, получена биопсия мышечных волокон, исследованы морфофункциональные характеристики мышечных волокон с помощью иммуногистохимического анализа, а также проведен ряд антропометрических измерений и биоимпедансный анализ. Сформированная группа является в этом отношении уникальной по своим охарактеризованным генетическим и фенотипическим характеристикам и самой большой в России. Применение к данным сформированной группы комбинации биоинформатических методов ассоциативного анализа, включающей полногеномный и полнотранскриптомный ассоциативный анализ и моделирование полногеномных баллов, позволило охарактеризовать новые

генетические факторы, ассоциированные с морфофункциональными характеристиками мышечных волокон человека: соотношением волокон разного типа и площадью поперечного сечения. Настоящая работа имеет как теоретическую, так и практическую значимость. Теоретическая значимость состоит в описании новых генетических факторов, ассоциированных с морфофункциональными характеристиками мышечных волокон, которые дополнили уже известные и описанные ранее факторы. Генетические ассоциации также были установлены и для дополнительно доступных фенотипов (антропометрические изменения и биоимпедансный анализ). В ходе настоящей работы для изучения морфофункциональных характеристик мышечных волокон использовался инвазивный метод мышечной биопсии. Объединение результатов настоящей работы с данными генетических ассоциаций, полученных в рамках аналогичных исследований, может позволить в перспективе использовать генетические данные для неинвазивной оценки морфофункциональных характеристик мышечных волокон. Рассматриваемая комбинация биоинформатических методов может применяться для более детального анализа фенотипа в тех случаях, когда сформировать репрезентативные выборки для исследования сложно из-за редкой встречаемости анализируемого признака или сложности его измерения. В ходе биоинформатического анализа применялся метод импутации генотипов для увеличения числа доступных для изучения вариантов, полученных с помощью метода высокоплотного генотипирования. Используемый в ходе работы программный код был оформлен в виде программного конвейера с помощью системы Snakemake, что значительно упрощает его запуск и настройку работы. Программный конвейер позволяет автоматизировать процесс импутации генотипов и может использоваться в дальнейших аналогичных исследованиях.

Методология и методы исследования. В основе настоящей работы лежат три биоинформатических метода ассоциативных исследований: полногеномное ассоциативное исследование, полнотранскриптомное ассоциативное исследование и моделирование полногеномных баллов. Основная часть биоинформатического анализа была реализована в среде shell с использованием ряда биоинформатических программ и скриптов, а также инструментов unix для обработки текстовых файлов. Для большинства этапов контроля качества использовалась программа PLINK 1.9. Для анализа главных компонент и ассоциативного анализа с помощью линейной регрессии использовалась программа PLINK 2.0. Для фазирования и импутации генотипов применялся метод Монте Карло по схеме марковской цепи, реализованный в программах SHAPEIT2 и IMPUTE2. Дополнительно были проведены фазирование и импутация с использованием программ eagle и Minimac3. Полнотранскриптомное ассоциативное исследование было основано на методе эластичной сети, реализованном в программе UTMOST. Для моделирования полногеномных баллов использовалась программа PRSice-2. Для обработки и

визуализации результатов применялись скрипты в среде R и дополнительные библиотеки: data.table, ggplot2, qqman. Для расчетов использовался вычислительный кластер Федерального научно-клинического центра физико-химической медицины (18 вычислительных узлов по 64 - 256 Гб ОЗУ).

Положения диссертации, выносимые на защиту.

1. Применение комбинации методов полногеномного, полнотранскриптомного ассоциативного анализа и метода моделирования полногеномных баллов позволило охарактеризовать новые генетические варианты, ассоциированные с морфологическими показателями мышечных волокон человека и антропометрическими характеристиками исследуемых лиц.
2. Полногеномное ассоциативное исследование выявило ряд ассоциаций на уровне тенденции ($p < 1e-5$) с показателями соотношения волокон I и II типа и площади поперечного сечения волокон в четырехглавой мышце бедра, а также ассоциации с полногеномной значимостью ($p < 5e-8$) для основного обмена.
3. Полнотранскриптомное исследование ассоциации установило значимую ассоциацию между относительной площадью мышечных волокон и генетическим компонентом регуляции экспрессии гена KDR ($q < 0.1$) и ассоциацию на уровне тенденции для гена PAX3 ($q = 0.18$).
4. Применение метода моделирования полногеномных баллов показало значимую ассоциацию между совокупным эффектом генетических вариантов и процентной долей мышечных волокон ($p = 0.00012$), доля объясняемой дисперсии модели составила $R^2 = 8.7\%$.
5. С использованием дополнительных независимых данных генотипирования групп профессиональных спортсменов был валидирован ряд ассоциаций с морфофункциональными характеристиками мышечных волокон для вариантов в генах UBR5, FTO, SEMA4A, MLN.
6. Программный код для импутации генотипов и проведения полногеномного ассоциативного исследования организован в программный конвейер для возможности использования в дальнейших исследованиях.

Апробация работы. По теме диссертации опубликовано девять работ в рецензируемых научных журналах. Результаты работы были представлены и обсуждались на следующих конференциях: ежегодной конференции немецкого общества генетики человека (Jahrestagung der Deutschen Gesellschaft für Humangenetik, Мюнстер, Германия, 16.03.2018), конференции Федерации европейских биохимических обществ 2018 (the FEBS Congress Прага, Чешская Республика, 09.07.2018), научной конференции молодых ученых Федерального научно-клинического центра физико-химической медицины Федерального Медико-биологического Агентства (Москва, Россия, 24.04.2018). Исследование

было награждено стипендией молодым ученым Федерации европейских биохимических обществ.

Личный вклад автора. Все этапы биоинформатического анализа в рамках настоящего исследования проведены соискателем лично: обработка и контроль качества данных генотипирования, фазирование и импутация генотипов, полногеномное ассоциативное исследование, полнотранскриптомное ассоциативное исследование, моделирование полногеномных баллов.

Структура и объем работы. Диссертация изложена на 103 страницах машинописного текста и состоит из следующих разделов: введение, обзор литературы, материалы и методы, результаты и обсуждение, заключение и выводы. Диссертация включает 27 рисунков и 9 таблиц. Список литературы включает 151 ссылку.

Содержание работы

Во введении рассматривается актуальность и степень разработанности темы исследования, приводятся цели и задачи исследования, использованные в ходе работы методы, формулируется научная новизна, теоретическая и практическая ценность работы, а также положения, выносимые на защиту.

В первой главе приводится обзор литературы по теме исследования. В первых двух разделах обзора литературы дается общая характеристика строения генома человека, вариабельных элементов в геноме и их классификация. Далее рассматриваются существующие технологии исследования генетических данных и методы их биоинформатического анализа. Обсуждается метод высокоплотного генотипирования на ДНК микрочипах, который применялся в рамках настоящей работы. В разделе биоинформатического анализа рассматриваются существующие форматы для хранения и работы с генетическими данными. Подробно описывается процесс биоинформатической обработки и контроля качества данных высокоплотного генотипирования для генетических вариантов и участников исследования. Далее рассматриваются методы фазирования гаплотипов и импутации генотипов, которые позволяют увеличить число изучаемых вариантов на порядок за счет заполнения негенотипированных вариантов, находящихся в неравновесном сцеплении с исследованными на чипе вариантами. В следующем разделе описывается методология полногеномных ассоциативных исследований, рассматривается применение регрессионного анализа для поиска генетических ассоциаций, расчет отношения шансов, проблема поправки на множественное сравнение, формирование сводной статистики и визуализация данных. Большинство генетических вариантов, для которых были установлены ассоциации, расположено в некодирующих участках ДНК. Один из механизмов их влияния на развитие фенотипа – регуляция экспрессии генов. Лocus в геноме, который ассоциирован с изменением уровня экспрессии генов, называется локусом количественных признаков экспрессии. На результатах исследования этих локусов основано полнотранскриптомное исследование ассоциаций, его методология рассматривается в следующем разделе. Описываются основные этапы для формирования моделей, позволяющих оценивать изменение уровней экспрессии генов с использованием данных генотипирования, а также существующие статистические методы и биоинформатические программы для проведения полнотранскриптомного исследования ассоциаций. Для анализа совокупного влияния генетических факторов на фенотип используется метод моделирования полногеномных баллов, рассматриваемый в следующем разделе. С помощью полногеномных баллов эффекты генетических вариантов объединяются в аддитивную сумму, и далее оценивается ассоциация совокупного влияния генетических вариантов с фенотипом. Приводится описание основных этапов моделирования и анализа полногеномных баллов,

различных подходов для учета неравновесного сцепления между включаемыми в модель вариантами, существующие биоинформатические программы для анализа полногеномных баллов, а также возможности практического применения полногеномных баллов. В рамках настоящей работы осуществлялся систематический биоинформатический анализ генетических факторов, определяющих состав мышечных волокон человека. В следующем разделе обзора литературы описываются морфофункциональные характеристики мышечных волокон человека. Рассматривается их связь с рядом заболеваний и физиологическими способностями. Также обсуждаются средовые и генетические факторы, влияющие на соотношение волокон I и II типа. Описываются установленные к настоящему времени генетические факторы, ассоциированные с составом мышечных волокон.

Во второй главе описаны материалы и методы настоящей работы, вторая глава состоит из 5 разделов.

Общая характеристика сформированной группы. Численность сформированной группы – 171 человек (115 мужчин, 56 женщин). Образцы геномной ДНК участников получали из лейкоцитарной фракции венозной крови. Для анализа генетических вариантов использовался ДНК-чип Illumina Omni Express Exome 8 v1.4 (Illumina, США), исследующий около 960 тысяч вариантов для каждого образца. Генотипирование образцов проводилось сотрудниками Федерального научно-клинического центра физико-химической медицины (Москва). Для исследования морфофункциональных характеристик мышечных волокон применялся метод биопсии и иммуногистохимический анализ, которые выполнялись сотрудниками Института медико-биологических проблем РАН (Москва). В ходе исследования анализировались следующие 3 характеристики мышечных волокон каждого типа: процентная доля волокон, площадь поперечного сечения волокон, а также относительная площадь, занимаемая мышечными волокнами на срезе. Взятие образцов мышечной ткани из срединной части наружной головки четырехглавой мышцы бедра было проведено с помощью методики игольчатой биопсии по Бергстрому с аспирацией. Мышечные волокна I и II типов выявляли с применением антител против медленных и быстрых изоформ тяжелых цепей миозина. Для определения процентной доли волокон I типа подсчитывалось количество окрашенных волокон после обработки срезов антителами против тяжелых цепей миозинов I типа, аналогично определялась процентная доля волокон II типа. Площадь поперечного сечения мышечных волокон рассчитывалась в условных единицах и переводилась в площадь, выраженную в квадратных микрометрах, с использованием снимка микрометровой шкалы. Относительная площадь, занимаемая мышечными волокнами I типа на срезе (S_{отн}) определялась как $S_{отн} = SI \times FI / ((SI \times FI) + (SII \times FII))$, где SI и SII – площадь поперечного сечения волокон I и II типа, соответственно; FI и FII – процентная доля волокон I и II типа, соответственно. Аналогично была вычислена

относительная площадь волокон II типа. Распределение относительной площади волокон I и II типа показано на рисунке 1.

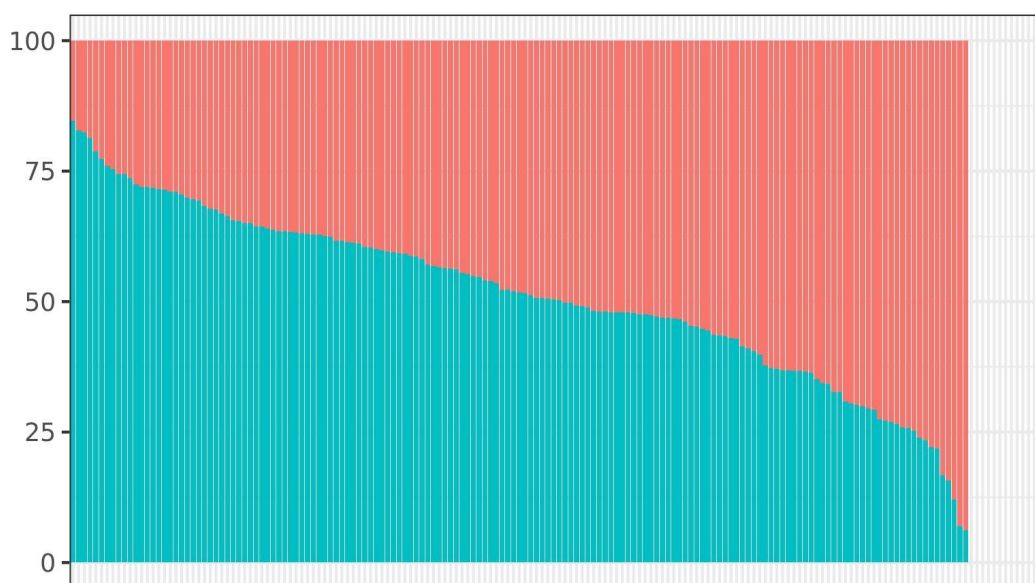


Рисунок 1 - Столбиковая диаграмма распределения относительной площади (ось ординат) волокон I типа (выделены красным цветом) и волокон II типа (выделены голубым цветом). Каждый столбик соответствует распределению относительной площади волокон у одного из участников выборки.

Для части из участников группы были также исследованы данные дополнительных фенотипов, включивших антропометрические измерения и биоимпедансный анализ, которые были проведены сотрудниками НИИ и Музея антропологии Московского государственного университета имени М.В. Ломоносова (Москва).

Таким образом, наличие полногеномных генетических данных, иммуногистохимических данных состава мышечных волокон и дополнительных фенотипических признаков делают возможным детальное и систематическое изучение генетических факторов морфофункциональных характеристик мышечных волокон.

Контроль качества генетических данных, фазирование, импутация. Для контроля качества осуществлялась фильтрация вариантов по частоте минорного аллеля (0.1%), по частоте пропусков (2%), по отклонению от равновесия Харди-Вайнберга ($p < 1e-6$), также удалялись варианты дубликаты. Контроль качества для образцов включил проверку соответствия пола образцов по генетическим данным, удаление образцов с высоким уровнем пропусков вариантов (более 2%), поиск родственных образцов до третьей степени родства. Одна пара образцов первой степени родства была обнаружена, один образец был удален. Для контроля популяционных различий использовался анализ главных компонент, в ходе которого были извлечены 10 собственных векторов (главных компонент). Для определения отклоняющихся образцов использовалось расстояние в 6 стандартных отклонений от среднего значения как минимум

одного из 10 собственных векторов. В итоге 7 образцов были удалены из дальнейшего анализа (рисунок 2).

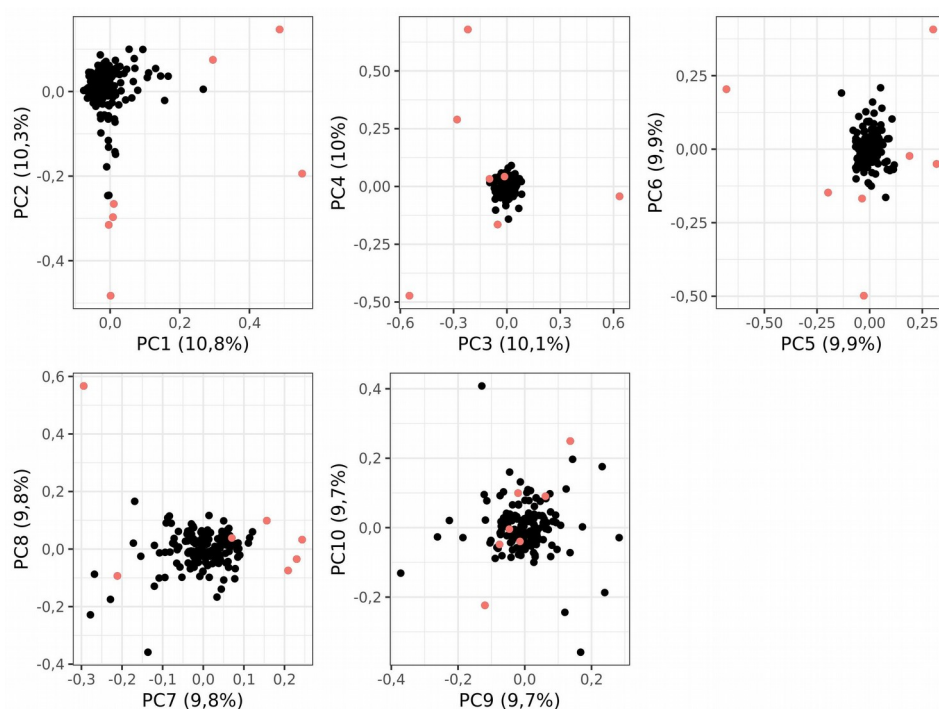


Рисунок 2 – Диаграмма рассеяния результатов анализа главных компонент для исследуемой выборки. Цветом выделены отклоняющиеся от общей выборки образцы

Входной контроль качества прошли 660218 вариантов для 163 образцов, для них выполнялось фазирование гаплотипов и импутация генотипов. В качестве референсной панели использовались данные проекта «1000 геномов» 3-й фазы. Этап фазирования был реализован с помощью программы SHAPEIT2, этап импутации – с помощью программы IMPUTE2. Импутация производилась для отдельных регионов каждой хромосомы размером 5 мегабаз для повышения точности и возможности параллелизации процессов. Общее число импутированных вариантов составило 15 миллионов 622 тысячи 415. Большинство вариантов было импутировано с высоким качеством, медиана info (критерий качества импутации) составила 0.962. После фильтрации по критерию info (0.6) с целью удаления вариантов, импутированных с низкой вероятностью, число вариантов составило 11 миллионов 887 тысяч 662. Кроме фазирования и импутации данных с помощью комбинации SHAPEIT2 и IMPUTE2, также были осуществлены фазирование и импутация с помощью комбинации программ eagle и Minimac3 для снижения вероятности установления ложно-положительных ассоциаций как артефактов импутации.

Полногеномное ассоциативное исследование. Все фенотипы были количественными, и для анализа использовалась линейная регрессия в программе PLINK 2.0. В качестве ковариат были включены пол образцов,

возраст образцов, а также 10 собственных векторов из анализа главных компонент. Для поправки на множественное сравнение использовался порог полногеномной значимости $p = 5e-8$ (genome-wide threshold). Для выявления ассоциаций на уровне тенденций использовался порог предположительной значимости $p = 1e-5$ (suggestive threshold). В результате ассоциативного анализа для каждого фенотипа была сформирована сводная статистика. Для визуализации использовался Манхэттенский график – диаграмма рассеяния, отображающая уровень значимости вариантов ассоциаций по оси ординат, и координаты вариантов на хромосомах по оси абсцисс. Обработка и визуализация результатов ассоциативного анализа проводилась в среде R с использованием библиотеки qqman.

Полнотранскриптомное ассоциативное исследование. Для поиска ассоциаций с фенотипами на уровне регуляции экспрессии генов применяются полнотранскриптомные ассоциативные исследования. Они используют результаты полногеномного ассоциативного исследования (сводную статистику) и локусы количественных признаков экспрессии (eQTL) для предсказания генетического компонента экспрессии генов. Для осуществления полнотранскриптомного ассоциативного исследования существует несколько алгоритмов, в настоящем исследовании использовалась программа UTMOST (unified test for molecular signatures). Для предсказания уровней экспрессии генов использовались модели, основанные на данных базы GTEx. Для настоящего исследования наибольший интерес представляла модель для ткани скелетных мышц (Muscle Skeletal). Для поправки на множественное сравнение использовался критерий Бенджамини-Хохберга. Результаты анализа в виде сводной статистики далее визуализировались в среде R с использованием библиотеки ggplot2.

Моделирование полногеномных баллов. Для расчета полногеномных баллов применялся метод валидации последовательным случайным сэмплированием, необходимый для снижения риска переобучения модели (overfitting). Выборка была разделена на обучающую подвыборку (2/3 от общего объема выборки) и тестовую подвыборку (1/3 от общего объема выборки). Ассоциативная статистика рассчитывалась для обучающей подвыборки, полногеномные баллы рассчитывались для тестовой подвыборки как сумма эффектов генетических вариантов, разделенная на общее число включенных вариантов. Эффект каждого варианта в аддитивной сумме рассчитывался как произведение оценки эффекта варианта, полученного из ассоциативной статистики обучающей подвыборки, и числа аллелей у индивида тестовой подвыборки (0, 1 или 2). Перед расчетом полногеномных баллов производилось прореживание (clumping) для исключения сильно коррелирующих друг с другом вариантов. В ходе валидации разделение выборки на обучающую и тестовую подвыборки было произведено 1000 раз. В ходе каждой итерации для расчета полногеномных баллов использовались 100 порогов значимости для включения вариантов в модель: от $p = 5e-5$ до $p = 1$ с логарифмическим шагом. В итоге

каждой из 1000 итераций были рассчитаны полногеномные баллы для образцов тестовой подгруппы для каждого из порогов значимости. Итоговые значения полногеномных баллов были получены путем усреднения полногеномных баллов из 1000 итераций для каждого образца на каждом из порогов значимости. Для полученных 100 моделей полногеномных баллов производилась оценка ассоциаций с фенотипом с помощью линейной регрессии. Модель с наивысшей предсказательной ценностью выбиралась с помощью максимизации коэффициента детерминации R^2 . Для расчета полногеномных баллов использовалась программа PRSice-2. Для анализа ассоциаций с рассчитанными полногеномными баллами и визуализации результатов использовалась среда R и библиотека ggplot2.

В третьей главе приводится описание результатов полногеномного ассоциативного исследования, полнотранскриптомного ассоциативного исследования и анализа полногеномных баллов для сформированной выборки, а также обсуждение полученных результатов.

Результаты полногеномного ассоциативного исследования. Для нескольких групп генетических вариантов была установлена ассоциация на уровне тенденции ($p < 1e-5$) с процентной долей мышечных волокон, одна из таких групп располагалась на 10-й хромосоме (рисунок 3, таблица 1). Для большинства из этих вариантов известна ассоциация с антропометрическими измерениями человека по данным биобанка Великобритании: предсказанная масса туловища, безжировая масса всего тела, масса воды всего тела.

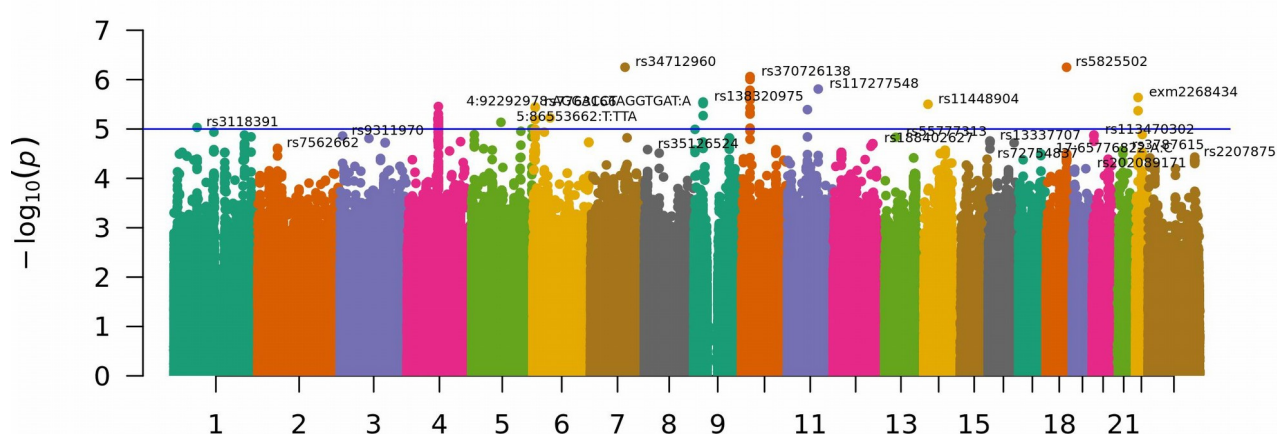


Рисунок 3 – Манхэттенский график (точечная диаграмма) ассоциаций с процентной долей мышечных волокон II типа. По оси абсцисс – варианты на каждой из хромосом, по оси ординат – отрицательный десятичный логарифм уровня значимости p

Ближайший ген к этой группе вариантов – ARHGAP21. По данным портала GTEx большинство из этих вариантов выступают значимыми регуляторами сплайсинга гена ARHGAP21 в ткани скелетных мышц, наибольшая значимость регуляции была показана для варианта rs7089436 ($p = 3.0e-18$). Ген ARHGAP21

относится к группе ГТФ-активирующих белков (Rho-GTPase-activating proteins), для него известно участие в энергетическом обмене.

Таблица 1 - Сводная статистика для части ассоциаций с процентной долей мышечных волокон II типа. CHR – хромосома, BP – позиция, SNP – идентификатор варианта, A1 – аллель, для которого рассчитана статистика, AX – другой аллель, BETA – коэффициент регрессии, SE – стандартная ошибка, T_STAT – T-статистика, p – уровень значимости

CHR	BP	SNP	A1	AX	BETA	SE	T_STAT	p
10	24879784	rs370726138	A	ATAC	10.8195	2.10171	5.14798	8.7e-7
10	24926724	rs72784284	G	T	10.5656	2.05741	5.1354	9.2e-7
10	24904420	rs7089436	C	A	10.8038	2.11085	5.11823	9.9e-7
10	24902921	rs72784277	A	G	10.6842	2.13473	5.00497	1.6e-6
10	24862664	rs72784264	T	A	10.3018	2.0739	4.96736	1.9e-6

При анализе площади поперечного сечения мышечных волокон I типа для нескольких групп вариантов была также выявлена значимость $p < 1e-5$. Интерес представляет группа вариантов на 14-й хромосоме. Лидирующий вариант – rs8014574, для него известны ассоциации с рядом антропометрических и биоимпедансных измерений (показатели жировой массы конечностей и всего тела). Полногеномное ассоциативное исследование относительной площади, занимаемой мышечными волокнами на срезе, выявило ассоциации с уровнем значимости $p < 1e-5$ для группы вариантов на 10-й хромосоме. Для варианта rs605089 этой группы ранее была найдена номинальная ассоциация с антропометрическими показателями и силой кисти в данных биобанка Великобритании.

Среди дополнительных доступных фенотипов была установлена ассоциация с полногеномной значимостью ($p < 5e-8$) с показателями основного обмена (рисунок 4).

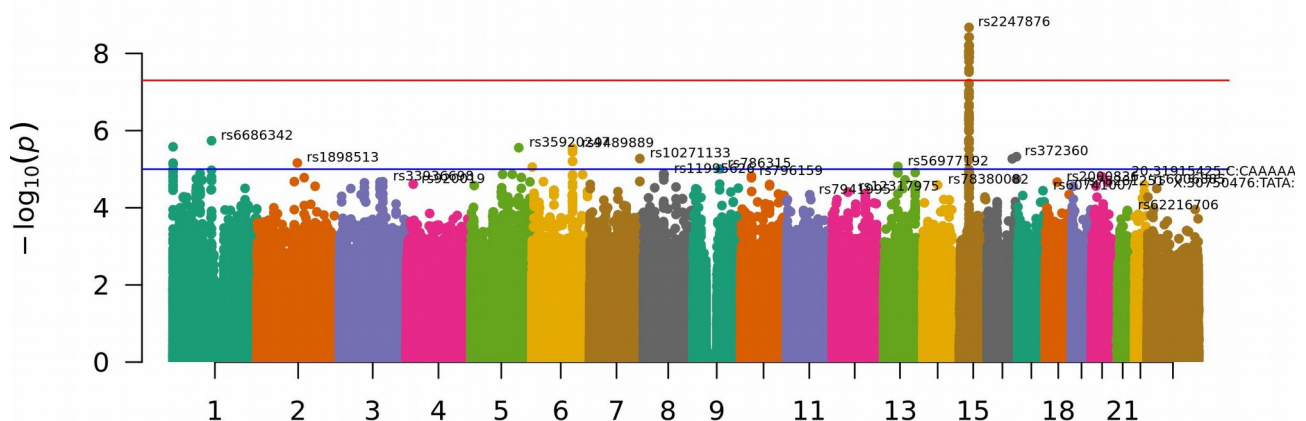


Рисунок 4 - Манхэттенский график (точечная диаграмма) ассоциаций с основным обменом.

По оси абсцисс – варианты на каждой из хромосом, по оси ординат – отрицательный десятичный логарифм уровня значимости p

Основной обмен оценивался с помощью биоимпедансометрии и был нормирован на площадь поверхности тела. Ассоциация для варианта rs611483

достигла уровня значимости $p=3.8e-9$. Для этого же варианта была выявлена номинальная ассоциация с площадью поперечного сечения мышечных волокон I типа ($p = 0.047$). С аллелем «А» варианта rs611483 были ассоциированы как более высокие показатели основного обмена, так и более высокая площадь поперечного сечения волокон I типа (рисунок 5). Биологически эта ассоциация может быть обусловлена высоким содержанием митохондрий в мышечных волокнах I типа, определяющими энергетический обмен в миоцитах.

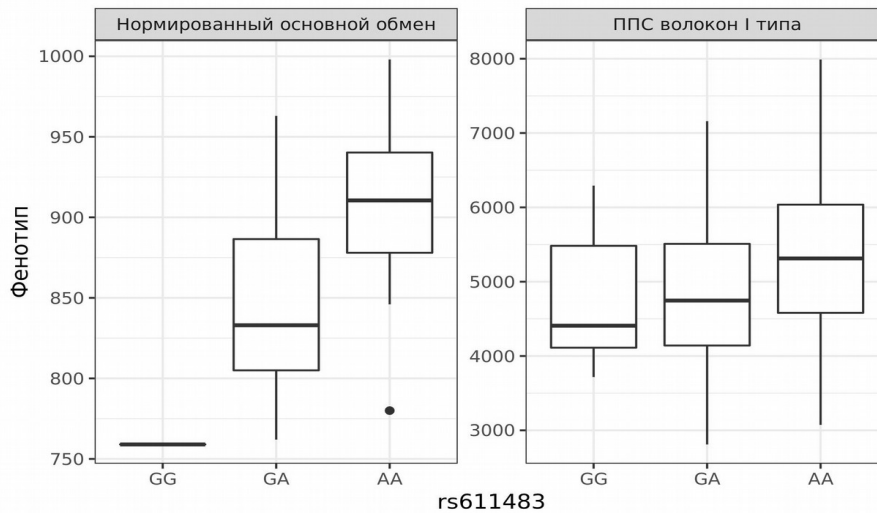


Рисунок 5 - Диаграмма ассоциаций с основным обменом (ккал/м²) и площадью поперечного сечения (кв. микрометры) волокон I типа. По осям абсцисс – генотипы варианта rs611483, по осям ординат – значения фенотипов

Таким образом, полногеномное ассоциативное исследование установило ряд генетических ассоциаций на уровне тенденции с морфофункциональными характеристиками мышечных волокон ($p < 1e-5$) и ассоциацию с полногеномной значимостью ($p < 5e-8$) для основного обмена. Ассоциации для части вариантов косвенно подтверждаются их функциональной ролью в регуляции экспрессии или спайсинга ближайших генов, а также ассоциацией с дополнительными фенотипами из открытых баз данных. Относительно небольшой объем изучаемой выборки ограничивает мощность ассоциативного исследования, и ассоциации с уровнем значимости $p < 1e-5$ могут рассматриваться как предположительные и требуют дополнительной валидации.

Результаты полнотранскриптомного ассоциативного исследования. В ходе полнотранскриптомного исследования ассоциаций были установлены два гена, которые представляют наибольший интерес для изучаемых фенотипов: KDR и PAX3. При исследовании процентной доли мышечных волокон была получена ассоциация с уровнями значимости $p = 1.05e-4$ для гена KDR и $p = 9.51e-5$ для гена PAX3. При исследовании относительной площади, занимаемой мышечными волокнами на срезе, наиболее значимым был ген KDR ($p = 4.3e-6$), уровень значимости для PAX3 составил $p = 9.99e-05$ (рисунок 6).

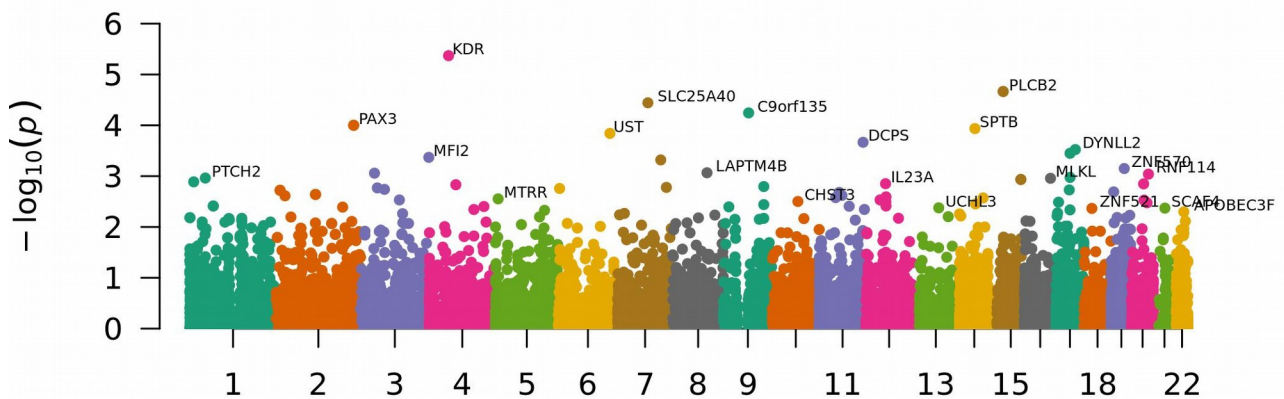


Рисунок 6 - Манхэттенский график (точечная диаграмма) полнотранскриптомного ассоциативного анализа относительной площади мышечных волокон на срезе. По оси абсцисс – гены на каждой из хромосом, по оси ординат – отрицательный десятичный логарифм уровня значимости p

При поправке на множественное сравнение с помощью критерия Бенджамини-Хохберга рассчитывался скорректированный уровень значимости q (q -value). После поправки уровень значимости для гена KDR составил $q = 0.053$, для гена PAX3 $q = 0.18$. Ассоциация с KDR может рассматриваться как статистически значимая при использовании порога значимости $q < 0.1$, в то время как ассоциация с PAX3 может рассматриваться только как предположительная. Дополнительные данные о возможной ассоциации генов PAX3 и KDR с изучаемыми фенотипами были найдены в литературе. Для замены His472Gln в гене KDR (синоним – VEGF, эндотелиальный фактор роста сосудов) была ранее описана ассоциация с процентной долей мышечных волокон, а также участие в регенерации скелетных мышц *in Vivo*. Для гена PAX3 ранее в ряде публикаций было описано участие в развитии клеток-предшественников миоцитов и в функционировании скелетных мышц.

Результаты моделирования полногеномных баллов. Полногеномные баллы для процентной доли мышечных волокон были рассчитаны с использованием 100 различных моделей. Для визуализации параметров полученных моделей использовалась столбиковая диаграмма, отражавшая уровни значимости (p) и предсказательную ценность (R^2) моделей на разных порогах уровня p для включения вариантов в модель (рисунок 7). Параметры полученных моделей показали характерную тенденцию для генетической архитектуры полигенных признаков. Наблюдалось увеличение объясняемой дисперсии при включении в модель большего числа вариантов до модели с наивысшей предсказательной ценностью, а затем ее снижение. Увеличение дисперсии может объясняться ассоциацией фенотипа с частью вариантов, которые не достигли уровня предположительной значимости ($p < 1e-5$). Для выявления ассоциаций этих вариантов требуется увеличение объема выборки и мощности исследования. Вместе с этим, предсказательная ценность модели начала снижаться после определенного порога ($p = 0.0742$), что связано с включением в последующие модели большой доли ложно-положительных вариантов.

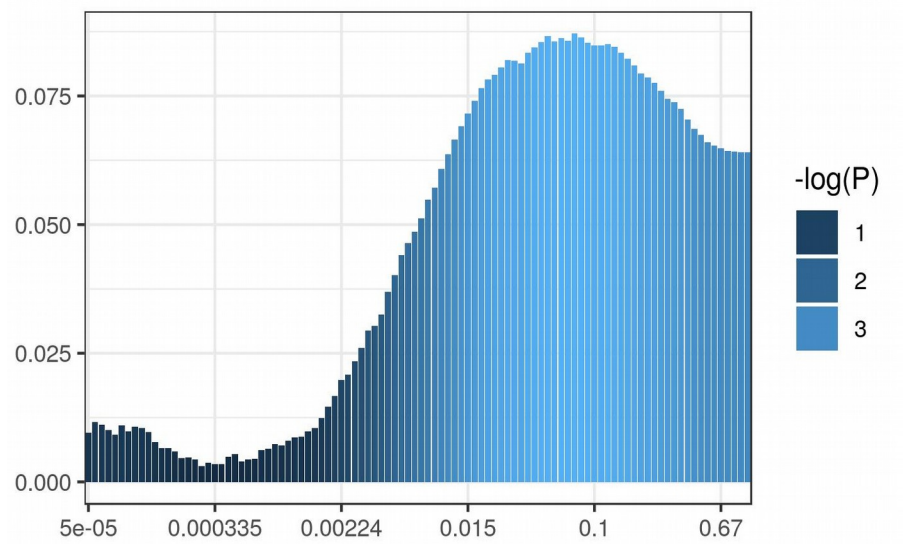


Рисунок 7 - Столбиковая диаграмма параметров моделей полигенных баллов. По оси абсцисс отображены уровни значимости p для включения вариантов в модель, по оси ординат – значения R^2 моделей, цвет заполнения отражает уровни значимости моделей

Среди 100 моделей наибольшую предсказательную ценность имела модель с порогом $p = 0.0742$. Регрессионный анализ полногеномных баллов этой модели показал значимую ассоциацию ($p = 0.00012$), рисунок 8.

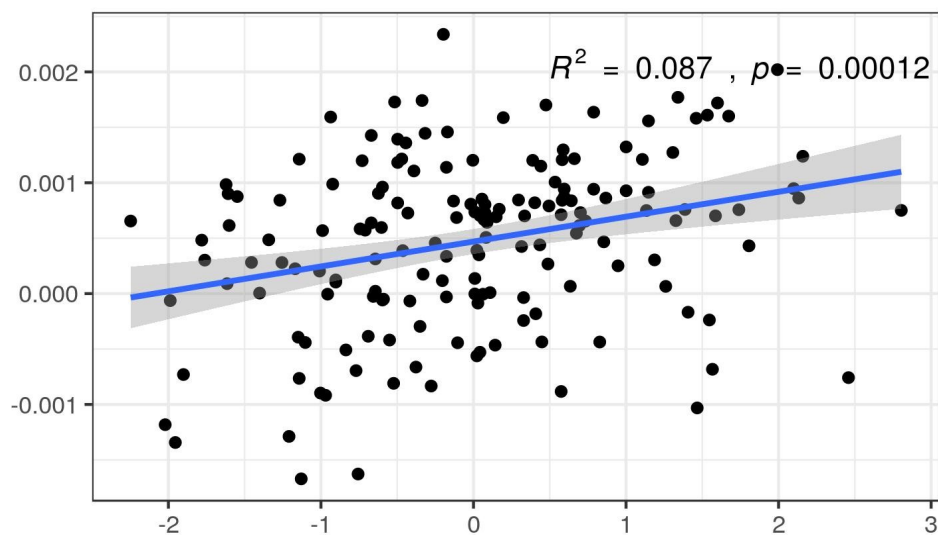


Рисунок 8 - Диаграмма рассеяния и регрессионный анализ. По оси ординат – значения полигенных баллов, по оси абсцисс – значения процентной доли мышечных волокон (значения стандартизованы).

Таким образом, моделирование полногеномных баллов позволило установить значимую ассоциацию между совокупным эффектом генетических вариантов и изучаемым фенотипом. Доля объясняемой вариации модели (коэффициент детерминации) составила $R^2 = 0.087$.

Валидация результатов с использованием дополнительных данных. В ходе работы использовались дополнительные данные высокоплотного генотипирования спортсменов различной специализации. Соотношение мышечных волокон связано с успешностью в определенных видах физической деятельности, более высокое содержание волокон I типа дает преимущество при тренировке выносливости, более высокое содержание волокон II типа – при силовых и спринтерских тренировках. Так, группа профессиональных спортсменов использовалась для косвенного подтверждения ассоциаций, полученных при анализе основных фенотипов. Дополнительные данные для ассоциаций с номинальной значимостью были получены для вариантов в генах UBR5, FTO, SEMA4A, MLN, HFE. Для аллеля A варианта rs4734621 была установлена ассоциация с высокими показателями площади поперечного сечения мышечных волокон II типа в четырехглавой мышце бедра и преобладание в группах спортсменов силовых видов спорта по сравнению со стайерами. Этот вариант ассоциирован с уровнем экспрессии гена UBR5 в ткани скелетных мышц по данным GTEx ($p = 0.000013$). При моделировании состояния гипертрофии и восстановлении после атрофии мышечного волокна было показано увеличение уровня экспрессии UBR5 (работа выполнена в исследовательском институте в Великобритании – Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Liverpool, UK). Аллель A варианта rs9939609 (ген FTO) был ассоциирован с низкой долей волокон I типа в четырехглавой мышце бедра и более часто встречался в группе профессиональных спортсменов силовой специализации. Аллель T варианта rs12401573 в гене SEMA4A был ассоциирован с высокой долей мышечных волокон II типа и спринтерскими характеристиками в группе профессиональных спортсменов. Для аллеля G варианта rs12055409 (ближайший ген MLN) показана ассоциация с показателями гипертрофии волокон II типа и силовыми характеристиками в группе профессиональных спортсменов тяжелоатлетов. Также был проведен мета-анализ для варианта rs1799945 в гене HFE (homeostatic iron regulator), который участвует в регуляции обмена железа. В общей сложности в мета-анализ были включены 586 спортсменов и 1416 контролей. Генотипы CG и GG варианта rs1799945 чаще наблюдались в группе спортсменов по сравнению с контрольной группой ($p = 1.7e-9$).

Создание репозитория с использованным программным кодом. Биоинформатический анализ данных генотипирования в рамках настоящего исследования включил применение ресурсоемкого метода импутации генотипов. Для возможности использования в дальнейших исследованиях примененного программного кода был создан программный конвейер с помощью системы Snakemake, который доступен в репозитории <https://github.com/oborisov/imputation-tools>. Система Snakemake делает возможным запуск программного конвейера на разных машинах за счет

настройки параметров в файле с конфигурациями. Конвейер использует ряд пакетов и программ, которые устанавливаются с помощью среды conda. Конвейер принимает на вход данные в бинарном формате PLINK 1.9, разделяет данные по хромосомам, выполняется выравнивание генотипов для соответствия референсному геному. Далее выполняется фазирование и импутация генотипов с использованием референсной панели проекта «1000 геномов». Необходимые файлы для выравнивания генотипов, фазирования и импутации скачиваются автоматически в ходе работы конвейера. После импутации конвейер выполняет ассоциативный анализ и производит построение Манхэттенского графика.

Заключение

Благодаря современным технологиям генотипирования стали возможными широкомасштабные исследования наследственных признаков человека. В ходе настоящей работы выполнялось изучение генетических факторов, ассоциированных с морфофункциональными характеристиками мышечных волокон: процентной долей и площадью поперечного сечения волокон I и II типа. Для этого была сформирована экспериментальная группа, для которой были получены данные высокоплотного генотипирования, исследованы морфофункциональные характеристики мышечных волокон четырехглавой мышцы бедра, а также проведены дополнительные антропометрические измерения и биоимпедансный анализ. Для анализа изучаемых фенотипов были осуществлены три биоинформатических метода: полногеномное исследование ассоциаций, полнотранскриптомное исследование ассоциаций и моделирование полногеномных баллов. С помощью комбинации этих методов был охарактеризован ряд генетических факторов состава мышечных волокон: установлены ассоциации на уровне тенденции с отдельными генетическими вариантами, значимая ассоциация с предсказанным уровнем экспрессии гена KDR, а также значимая ассоциация с совокупным эффектом генетических вариантов. С использованием дополнительных данных генотипирования групп профессиональных спортсменов был валидирован ряд ассоциаций. Наконец, использованный в работе программный код для импутации генотипов был оформлен в виде программного конвейера с помощью системы Snakemake для возможности использования в дальнейших исследованиях.

Выводы

1. Сформирована группа испытуемых численностью 171 человек, для которых получены данные высокоплотного генотипирования, морфологических характеристик мышечных волокон, антропометрических измерений и биоимпедансного анализа.
2. Итоговое количество генетических вариантов, доступных для полногеномных ассоциативных исследований, составило более 10

миллионов после этапов контроля качества, фазирования и импутации генотипов.

3. Полногеномное ассоциативное исследование выявило ряд ассоциаций на уровне тенденции ($p < 1e-5$) с показателями соотношения волокон I и II типа и гипертрофии волокон в четырехглавой мышце бедра, а также ассоциации с полногеномной значимостью ($p < 5e-8$) для основного обмена.
4. Полнотранскриптомное исследование ассоциации установило ассоциацию между относительной площадью мышечных волокон и генетическим компонентом регуляции экспрессии гена KDR ($q < 0.1$) и ассоциацию на уровне тенденции для гена PAX3 ($q = 0.18$).
5. Применение метода моделирования полногеномных баллов показало значимую ассоциацию между совокупным эффектом генетических вариантов и процентной долей мышечных волокон ($p = 0.00012$), доля объясняемой дисперсии модели составила $R^2 = 8.7\%$.
6. С использованием дополнительных независимых данных генотипирования групп профессиональных спортсменов был валидирован ряд ассоциаций с морфофункциональными характеристиками мышечных волокон для вариантов в генах UBR5, FTO, SEMA4A, MLN.
7. Программный код для импутации генотипов и проведения полногеномного ассоциативного исследования организован в программный конвейер для возможности использования в дальнейших исследованиях.

Список публикаций по теме диссертации

По теме диссертации опубликовано девять работ в рецензируемых научных журналах:

- 1) **Borisov O.**, Kulemin N., Ahmetov I., Generozov E. A Novel Multilocus Genetic Model Can Predict Muscle Fibers Composition // International Symposium on Computer Science in Sport. Springer, 2017. P. 164–168. https://doi.org/10.1007/978-3-319-67846-7_16
- 2) Семенова Е.А., Хабибова С.А., **Борисов О.В.**, Генерозов Э.В., Ахметов И.И. Вариабельность структуры ДНК и состав мышечных волокон человека // Физиология человека, 2019, том 45, ном. 2, стр.128–136. <https://doi.org/10.1134/S0131164619010132>
- 3) Guilherme J.P.L., Egorova E., Semenova E., Kostyukova E., Kulemin N., **Borisov O.**, Khabibova S., Larin A., Ospanova E., Pavlenko A., Lyubaeva E., Popov D., Lysenko E., Vepkhvadze T., Lednev E., Govorun V., Generozov E., Ahmetov I., Lancha Junior A. The A-allele of the FTO Gene rs9939609 Polymorphism Is Associated With Decreased Proportion of Slow Oxidative Muscle Fibers and Over-represented in Heavier Athletes // J. Strength Cond.

- Res. 2019. Vol. 33, № 3. P. 691–700.
<https://doi.org/10.1519/JSC.0000000000003032>
- 4) Seaborne, R. A.; Hughes, D. C.; Turner, D. C.; Owens, D. J.; Baehr, L. M.; Gorski, P.; Semenova, E. A.; **Borisov, O. V.**; Larin, A. K.; Popov, D. V.; et al. UBR5 Is a Novel E3 Ubiquitin Ligase Involved in Skeletal Muscle Hypertrophy and Recovery from Atrophy. *J Physiol* 2019, JP278073. <https://doi.org/10.1113/JP278073>
 - 5) Pickering, C.; Suraci, B.; Semenova, E. A.; Boulygina, E. A.; Kostryukova, E. S.; Kulemin, N. A.; **Borisov, O. V.**; Khabibova, S. A.; Larin, A. K.; Pavlenko, A. V.; et al. A Genome-Wide Association Study of Sprint Performance in Elite Youth Football Players: *Journal of Strength and Conditioning Research* 2019, 1. <https://doi.org/10.1519/JSC.0000000000003259>
 - 6) Grishina, E. E.; Zmijewski, P.; Semenova, E. A.; Ciężczyk, P.; Humińska-Lisowska, K.; Michałowska-Sawczyn, M.; Maculewicz, E.; Crewther, B.; Orysiak, J.; Kostryukova, E. S.; Kulemin, N. A.; **Borisov, O. V.**; et al. Three DNA Polymorphisms Previously Identified as Markers for Handgrip Strength Are Associated With Strength in Weightlifters and Muscle Fiber Hypertrophy: *Journal of Strength and Conditioning Research* 2019, 33 (10), 2602–2607. <https://doi.org/10.1519/JSC.0000000000003304>
 - 7) **Borisov O.**, Maj C., Kulemin N., Semenova E., Krawitz P., Ahmetov I., Generozov E. Polygenic Modeling of Muscle Fibers Composition // *Proceedings of the 12th International Symposium on Computer Science in Sport, Advances in Intelligent Systems and Computing*, vol 1028. Springer, Cham. https://doi.org/10.1007/978-3-030-35048-2_18
 - 8) Denis Kusić, Joanne Connolly, Heikki Kainulainen, Ekaterina A Semenova, **Oleg V Borisov**, Andrey K Larin, Daniil V. Popov, Edward V Generozov, Ildus I Ahmetov, Steve L. Britton, Lauren G. Koch, and Jatin G Burniston. Striated muscle-specific serine/threonine-protein kinase beta (SPEGβ) segregates with high- versus low-responsiveness to endurance exercise training // *Physiological Genomics* 2019. <https://doi.org/10.1152/physiolgenomics.00103.2019>
 - 9) Ekaterina A Semenova, Eri Miyamoto-Mikami, Egor B Akimov, Fatima Al-Khelaifi, Haruka Murakami, Hirofumi Zempo, Elena S Kostryukova, Nikolay A Kulemin, Andrey K Larin, **Oleg V Borisov** et al. The association of HFE gene H63D polymorphism with endurance athlete status and aerobic capacity: novel findings and a meta-analysis // *European Journal of Applied Physiology* 2020. <https://doi.org/10.1007/s00421-020-04306-8>