ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ НАУЧНО-КЛИНИЧЕСКИЙ ЦЕНТР ФИЗИКО-ХИМИЧЕСКОЙ МЕДИЦИНЫ ФЕДЕРАЛЬНОГО МЕДИКО-БИОЛОГИЧЕСКОГО АГЕНТСТВА»

на правах рукописи

Кулемин Николай Александрович

Биоинформатический анализ данных высокопроизводительного генотипирования в применении к поиску маркеров спортивной успешности

Диссертация на соискание ученой степени
кандидата биологических наук

Специальность 03.01.09 – математическая биология, биоинформатика.

Научный руководитель
к.б.н., доцент Генерозов Э.В.

Москва 2016

Оглавление

Оглавление 2
Список используемых сокращений 5
Введение 6
Актуальность исследования 6
Цель исследования 8
Задачи работы 8
Научная новизна и практическая ценность работы 8
Положения диссертации, выносимые на защиту 10
Апробация работы 12
Глава 1. Обзор литературы 15
1.1 Биоинформатические методы в современных генетических исследованиях 15
1.2 Полногеномные ассоциативные исследования в медицинской генетике 17
1.2.1 Генетические исследования в спортивной медицине 21
1.3 Биоинформатический анализ геномных данных 24
1.3.1.1 Современные технологии секвенирования ДНК 25
1.3.1.2 Задачи биоинформатической обработки первичных результатов секвенирования 32
1.3.2.1 Современные технологии чип-гибридизации 42
1.3.2.2 Биоинформатический анализ первичных результатов высокоплотной чип-гибридизации 44
1.4 Биоинформатическая основа ассоциативных исследований 51
1.5 Биоинформатические подходы к оценке индивидуального генетического риска 55
1.5.1 Каталогизация клинически значимых ассоциаций 55
1.5.2 Оценка рискового статуса потенциально патогенных генетических вариаций 57
1.5.3 Биоинформатические подходы к оценке совокупного риска 58

1.6 Обоснование поставленных задач 62
Глава 2. Материалы и методы 63
2.1 Характеристика исследуемых групп 63
2.1.1 Группа профессиональных спортсменов из РФ 64
2.1.2 Группа спортсменов из других стран 65
2.2 Методы генетического анализа 66
2.2.1 Пробоподготовка 66
2.2.2 Ионное полупроводниковое секвенирование 66
2.2.3 Генотипирование с использованием ДНК-чипов 67
2.2.4 Масспектрометрическое минисеквенирование 67
2.3 Биомедицинские исследования 68
2.3.1 Определение максимального потребления кислорода 68
2.3.2 Анализ состава мышечных волокон 69
2.3.3 Биохимический и клинический анализ крови 70
2.4 Распределение испытуемых по группам сравнения 70
2.4.1 Распределение испытуемых при проведении полногеномного ассоциативного исследования выносливости 70
2.4.2 Распределение образцов при ассоциативном исследовании генетических маркеров и мышечной структуры 71
2.4.3 Распределение образцов при ассоциативном исследовании генетических маркеров и состояний оксидативного стресса 72
2.4.4 Распределение образцов при поиске общего генетического профиля спортсмена-стайера 73
2.5 Использованное программное обеспечение 73
2.5.1 Обработка данных NGS-секвенирования и генотипирования на ДНК чипах 73
2.5.2 Статистический анализ и визуализация 74

2.5.3 Другое программное обеспечение 75
2.6 Аппаратные средства 75
Глава 3. Результаты и обсуждение 76
3.1 Обработка результатов ионного полупроводникового секвенирования 76
3.1.1 Оптимизация протокола на стандартизированных прочтениях 79
3.1.2 Оптимизация протокола на экспериментально полученных прочтениях тестовых образцов ДНК 85
3.2 Адаптация результатов чип-гибридизации 88
3.3 Применение разработанных протоколов 92
3.3.1 Анализ генетических основ выносливости 94
3.3.2 Анализ генетических факторов, ассоциированных со структурой мышечных волокон 97
3.3.3 Генетический анализ факторов, ассоциированных с оксидативным стрессом 100
3.3.4 Полногеномный ассоциативный анализ общего генетического профиля спортсменов-стайеров 103
3.4 Медицинская аннотация индивидуальных особенностей на основании данных генетических исследований 106
3.4.1 Разработка аннотационного инструмента по известным клинически-значимым полиморфизмам ДНК 106
3.4.2 Разработка релевантной модели прогнозирования состава мышечных волокон 114
Заключение 118
Выводы 119
Список литературы 119

Список используемых сокращений

БД - база данных.

Гб (Gb) - гигабайт.

ДНК - дезоксирибонуклеиновая кислота.

Мб (Mb) - мегабайт.

млн. - миллионы.

МПК - максимальное потребление кислорода.

ОЗУ - оперативное запоминающее устройство.

п.н. (пн) - пара нуклеотидов.

ПЦР - полимеразная цепная реакция.

ПСА - Простатический специфический антиген.

ПО - программное обеспечение.

РНК - рибонуклеиновая кислота.

СУБД - система управления базами данных.

Тб (Tb) - терабайт.

тыс. - тысячи.

AUC - Area Under Curve, площадь под графиком.

BLAST (n) - Basic Local Alignment Search Tool (for nucleotides), простое средство поиска локальных соответствий (для нуклеотидных последовательностей).

ClinVar - Clinical Variation database, база данных клинических состояний.

GIAB – «Genome In A Bottle», «геном в бутылке» - международный проект по стандартизации результатов обработки данных секвенирования.

GWAS - Genome Wide Association Study, полногеномное ассоциативное исследование.

ICD (рус.МКБ) - International Classification of Diseases, Международный Классификатор Болезней.

LOVD - Leiden Open Variation Database, открытая база данных вариаций Ляйдена.

OMIM - Online Mendelian Inheritance in Man, онлайн справочник заболеваний человека, наследуемых по Менделю.

PharmGKB - The PharmacoGenomics Knowledge Base, база знаний по фармакогеномике.

ROC - Receiver Operating Characteristic, рабочая характеристика приёмника.

SNP - single nucleotide polymorphism, однонуклеотидный полиморфизм.

SNPedia - Single Nucleotide Polymorphism encyclopedia, энциклопедия однонуклеотидных полиморфизмов.

UMLS - Unified Medical Language System - унифицированный язык медицинских систем.

Введение

Актуальность исследования

Современные исследования в биологии и медицине все сложнее представить без использования математических методов. Постоянно увеличивающийся объем биологических данных, и развитие новых вычислительных технологий привели к формированию отдельного научного направления – биоинформатики. На сегодняшний день, методы биоинформатики особенно востребованы в современных генетических исследованиях, использующих новейшие технологии высокоплотной чип-гибридизации или секвенирования ДНК нового поколения [1].

Особую и актуальную нишу среди них составляют исследования медицинской направленности, посвященные оценке вклада генетических факторов в патогенез социально значимых заболеваний, развитии тех или иных физиологических характеристик индивидуума [1]. В этом отношении, отдельную ценность представляют генетические исследования стратифицированных и ограниченных групп лиц с уникальными физиологическими характеристиками, например, групп профессиональных спортсменов. Статус элитного спортсмена, в большинстве своем, подразумевает фенотипически экстремальное состояние человеческого организма. Информация о генетической основе таких фенотипических проявлений является ценной как с фундаментальной точки зрения, так как позволяет прояснить общие механизмы адаптации организма к интенсивным физическим нагрузкам, так и имеет выраженную практическую составляющую. Полученные в научных исследованиях результаты допустимо применять, впоследствии, для спортивного отбора, коррекции параметров тренировочного процесса, для диагностики и лечения пациентов, которые не являются профессиональными спортсменами [2].

Современная методология ассоциативных генетических исследований базируется на использовании технологий высокопроизводительного генетического анализа. Как и в большинстве современных развивающихся

технологий, при активном развитии приборной базы, технологии информационной обработки сформированного объема данных существенно отстают и являются основным предметом для биоинформатического анализа и дискуссий [3]. На текущий момент, одним из популярных и перспективных технологических решений, позволяющих в короткие сроки проанализировать большой объем генетической информации является метод ионного полупроводникового секвенирования [4]. К сожалению, штатное программное обеспечение, предоставляемое вместе с секвенаторами, использующими указанный метод, не предоставляет возможности использовать весь объем получаемых генетических данных. Кроме того, оно не позволяет исключить ряд ошибок секвенирования, связанных с технологическими особенностями приборов, имеет ограниченное количество пользовательских настроек. В случае если протокол эксперимента предполагает различные источники геномных данных, то всегда требуется приведение результатов генотипирования к единому стандарту для последующего сравнения. В связи с этим, разработка альтернативных биоинформатических протоколов, позволяющих оптимизировать техническую первичную обработку получаемых с секвенатора данных, включая их сборку, картирование на референсный геном, идентификацию уникальных вариантов является актуальной методической задачей, как в рамках настоящей работы, так и для других ассоциативных генетических исследований в целом.

Другой аспект актуальности биоинформатических исследований в настоящей работе связан с впервые полученными данными по новым генетическим показателями в группах профессиональных спортсменов, ассоциированных с различными особенностями организма, информативными в спортивной деятельности [5]. В дополнение к полученным результатам по групповым ассоциациям в настоящей работе разработаны инструменты, позволяющие осуществлять и индивидуальную аннотацию выявленных в ходе генетического исследования генетических вариантов, ассоциированных с медицински-значимыми характеристиками. Такая аннотация позволяет сгруппировать по нозологическому принципу известные генетические варианты и

является основой для последующей разработки алгоритмов, позволяющих оценивать совокупный индивидуальный риск, учитывающий как генетические, так и негенетические факторы. Предполагается, что именно они в будущем и будут замыкать полный цикл генетического исследования пациентов с последующим переходом к персонифицированной медицине.

Таким образом, решения, которые были получены в данной работе, являются актуальными и уникальными. В части разработки протоколов первичного анализа результатов ионного полупроводникового секвенирования они представляют собой готовый программный продукт, подходящий для использования при решении разнообразных научных задач, использующих этот метод секвенирования. В части дальнейшего анализа полученных данных впервые получены результаты по ассоциации у спортсменов: полиморфизма в гене *SOD2* (rs4880) и оксидативного стресса, полиморфизма в гене *AGTR2* (rs11091046) и структуры мышц, полиморфизма в гене *NFIA-AS2* (rs1572312) и эритропоэза, а также разработан программный продукт для индивидуальной аннотации выявленных в ходе генетического исследования медицински значимых генетических вариаций.

Цель исследования

Целью настоящей работы является разработка комплексного биоинформатического протокола, предназначенного для первичной обработки и анализа данных высокопроизводительного генотипирования, проведения ассоциативных генетических исследований и аннотации выявленных генетических вариаций в группах профессиональных спортсменов.

Задачи работы

1) Разработать оптимизированный алгоритм обработки геномных данных, получаемых с использованием технологий ионного полупроводникового секвенирования (ThermoFisher) и высокоплотной чип-гибридизации (Illumina).

2) Используя полученные данные высокопроизводительного генотипирования групп профессиональных спортсменов, осуществить анализ

генетических маркеров, ассоциированных с показателями выносливости, оксидативного стресса, эритропоэза и структуры мышечных волокон.

3) Разработать программный инструмент для комплексной аннотации медицински значимых генетических вариаций, включающий анализ известных нуклеотидных полиморфизмов и мутаций, размещенных в доступных базах данных, анализ потенциальной патогенности редких вариантов и впервые выявленных мутаций.

Научная новизна и практическая ценность работы

Проведенное исследование осуществлено на генетических данных, полученных с использованием современных технологий: ионное полупроводниковое секвенирование нового поколения, высокоплотная чип-гибридизация.

Для решения поставленных задач были использованы современные биоинформатические алгоритмы, позволившие существенно оптимизировать первичную обработку генетических данных, реализовать возможность пользовательской настройки параметров анализа, исключить технические ошибки, обусловленные особенностями использованной технологии полупроводникового секвенирования. Так как технология ионного полупроводникового секвенирования является на текущий момент самой дешевой в расчете на 1 нуклеотид прочтения, то разработанные программные протоколы увеличения точности обработки данных позволят уменьшить требуемый объем и снизить итоговую стоимость секвенирования, что позволит использовать ионное полупроводниковое секвенирование для решения более широкого круга задач.

Также был модифицирован алгоритм обработки данных чип-гибридизации для приведения их в соответствие с результатами секвенирования, а затем для последующего совместного использования информации, полученной с использованием разных технологий в одном эксперименте.

С использованием разработанного оптимизированного алгоритма был проанализирован массив данных высокоплотной чип-гибридизации более 1000

профессиональных спортсменов Российской Федерации. В ходе последующего генетического ассоциативного исследования были получены новые данные по генетическим маркерам, ассоциированными с показателями оксидативного стресса, особенностями эритропоэза и структуры мышечных волокон.

Анализ более чем 1,5 млн. полиморфизмов среди спортсменов показал, что полиморфизм rs1572312 в гене *NFIA-AS2* достоверно различает между собой стайеров, лиц, не занимающихся спортом, и спринтеров.

Впервые была показана ассоциация полиморфизма rs11091046 гена *AGTR2* со структурой мышц и потреблением кислорода.

Также, впервые было показано, что полиморфизм rs4880 в гене *SOD2* связан с изменением уровней креатинкиназы, креатинина и прочих маркеров мышечного повреждения, что однозначно влияет на устойчивость к внезапному оксидативному стрессу.

Отдельную практическую ценность представляет разработанный программный продукт, позволяющий в полногеномном масштабе осуществлять индивидуальную аннотацию медицински значимых генетических вариантов. Разработанный для этого алгоритм может использовать как данные, полученные на высокопроизводительных ДНК-секвенаторах, так и в ходе чип-гибридизации. Использование таких или аналогичных алгоритмов в будущем позволит перейти на новый уровень индивидуализации данных пациента в составе технологий персонифицированной медицины.

Впервые была проведена разработка и апробация генетической модели для не инвазивной оценки состава мышечных волокон на основании совокупного анализа вклада генетических факторов.

Положения диссертации, выносимые на защиту

1) Разработан альтернативный алгоритм обработки и анализа генетических данных ионного полупроводникового секвенирования (ThermoFisher), основными характеристиками которого являются: высокая точность идентификации геномных состояний (99,66%), гибкие настройки параметров качества прочтений,

исключение ошибок прочтения инсерций и делеций, ассоциированных с технологическими особенностями метода.

2) Разработан алгоритм обработки данных высокоплотной чип-гибридизации (Illumina), позволяющий исключать ошибки, связанные с неправильным позиционированием генетических вариаций и предоставляющий возможность пользовательской настройки выходных параметров качества идентификации вариантов.

3) По результатам проведенного полногеномного ассоциативного исследования полиморфизм rs1572312 гена *NFIA-AS2* достоверно ассоциирован с уровнем гемоглобина в крови и, как следствие, с аэробными возможностями спортсменов.

4) Полиморфизм rs11091046 гена *AGTR2* достоверно ассоциирован со структурой мышечных волокон и показателями потребления кислорода.

5) По результатам ассоциативного исследования группы профессиональных спортсменов показано, что полиморфизм rs4880 гена *SOD2* влияет на устойчивость организма спортсмена к оксидативному стрессу, а также ассоциирован с изменением уровня креатинкиназы и креатинина.

6) Несмотря на наличие множества генетических вариаций, ассоциированных с показателями физической выносливости, в настоящее время не выявлено единого для большинства популяций генетического профиля спортсмена-стайера.

7) Разработан программный продукт, позволяющий осуществлять индивидуальную аннотацию медицински информативных генетических вариантов, полученных различными методами высокопроизводительного генотипирования.

8) Разработан и апробирован алгоритм индивидуальной комплексной оценки структуры мышечных волокон на основании анализа данных генетического анализа.

Апробация работы

По теме диссертации опубликовано пять работ в рецензируемых научных журналах:

1) Mustafina LJ, Naumov VA, Cieszczyk P, Popov DV, Lyubaeva EV, Kostryukova ES, Fedotovskaya ON, Druzhevskaya AM, Astratenkova IV, Glotov AS, Alexeev DG, Mustafina MM, Egorova ES, Maciejewska-Karłowska A, Larin AK, Generozov EV, Nurullin RE, Jastrzębski Z, Kulemin NA, Ospanova EA, Pavlenko AV, Sawczuk M, Akimov EB, Danilushkina AA, Zmijewski P, Vinogradova OL, Govorun VM, Ahmetov II. *AGTR2 gene polymorphism is associated with muscle fibre composition, athletic status and aerobic performance* // Exp Physiol. 2014; 99(8):1042-52.

2) Ahmetov II, Naumov VA, Donnikov AE, Maciejewska-Karłowska A, Kostryukova ES, Larin AK, Maykova EV, Alexeev DG, Fedotovskaya ON, Generozov EV, Jastrzębski Z, Zmijewski P, Kravtsova OA, Kulemin NA, Leonska-Duniec A, Martykanova DS, Ospanova EA, Pavlenko AV, Podol'skaya AA, Sawczuk M, Alimova FK, Trofimov DY, Govorun VM, Cieszczyk P. *SOD2 gene polymorphism and muscle damage markers in elite athletes* // Free Radic Res. 2014; 48(8):948-55.

3) Морозова Е.Б., Тараскина А.Е., Кострюкова Е.С., Генерозов Э.В., Кулемин Н.А., Ахметов И.И., Говорун В.М. *Изучение индивидуальных особенностей генетического статуса высококвалифицированных спортсменов //* Спортивная медицина: наука и практика. Москва, 2014; 3:19-29

4) Ahmetov I.I., Kulemin N.A., Popov D.V., Naumov V.A., Akimov E.B., Kostryukova E.S., Bravy Y.R., Egorova E.S., Galeeva A.A., Mustafina L.J., Larin A.K., Generozov E.V., Ospanova E.A., Pavlenko A.V., Żmijewski P., Alexeev D.G., Vinogradova O.L., Govorun V.M. *Genome–wide association study identifies three novel genetic markers associated with elite endurance performance* // Biol. Sport 2015; 32:3-9.

5) Tuomo Rankinen, Noriyuki Fuku, Bernd Wolfarth, Guan Wang, Mark A. Sarzynski, Dmitry G. Alexeev, Ildus I. Ahmetov, Marcel R. Boulay, Pawel Cieszczyk, Nir Eynon, Maxim L. Filipenko, Fleur C. Garton, Edward V. Generozov, Vadim M.

Govorun, Peter J. Houweling, Takashi Kawahara, Elena S. Kostryukova, Nickolay A. Kulemin, Andrey K. Larin, Agnieszka Maciejewska-Karłowska, Motohiko Miyachi, Carlos A. Muniesa, Haruka Murakami, Elena A. Ospanova, Sandosh Padmanabhan, Alexander V. Pavlenko, Olga N. Pyankova, Catalina Santiago, Marek Sawczuk, Robert A. Scott, Vladimir V. Uyba, Thomas Yvert, Louis Perusse, Sujoy Ghosh, Rainer Rauramaa, Kathryn N. North, Alejandro Lucia, Yannis Pitsiladis, Claude Bouchard. No *Evidence of a Common DNA Variant Profile Specific to World Class Endurance Athletes* // PLoS One, January 29, 2016. Online publication, DOI:10.1371/journal.pone.0147330.

Основные положения и результаты диссертации были доложены и обсуждены на следующих конференциях:

1) Конференция " Постгеномные методы анализа в биологии, лабораторной и клинической медицине". Н.А.Кулемин, В.А.Наумов, А.В.Павленко, Д.Г.Алексеев. *Методы анализа однонуклеотидных полиморфизмов генома человека* // Казань, 22.11.2012.
2) Конференция ФГАОУ ВПО «МФТИ ГУ». Н.А.Кулемин, В.А.Наумов, А.В.Павленко, Д.Г.Алексеев. *Методы анализа однонуклеотидных полиморфизмов генома человека* // Москва, НИИ ФХМ ФМБА РФ, 19.11.2012.
3) Конференция ECCB’12. I.A.Altukhov, D.S.Ischenko, N.A.Kulemin, D.G.Alekseev et al. *SRAGEN – Web-service for comparative structural analysis of procariotes* // Basel, 12.09.2012.
4) Конференция НИИ ФХМ ФМБА РФ. Н.А.Кулемин, В.А.Наумов, А.В.Павленко, Д.Г.Алексеев. *Методы анализа однонуклеотидных полиморфизмов генома человека* // Москва, НИИ ФХМ ФМБА РФ, 25.05.2012.
5) Конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». Н.А.Кулемин, В.А.Наумов, А.В.Павленко, Э.В.Генерозов. *Алгоритм медицинской аннотации геномных данных* //

Казань, 21.11.2014. (Докладу присуждено 1 место конкурса Nucleic Acid Research.)

6) Конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». E.S.Egorova, A.A.Galeeva, L.J.Mustafina, N.A.Kulemin et al. *Genome-Wide Association Study in Russian Strength Athletes* // Казань, 20.11.2014.

7) Конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». L.J.Mustafina, V.A.Naumov, E.S.Egorova, A.A.Galeeva, N.A.Kulemin et al. *Genome-Wide Association Analysis Identifies Loci on DMD and UGT2B4 Genes for Power Athlete Status in Russians* // Казань, 20.11.2014.

8) Конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». I.I.Ahmetov, N.A.Kulemin et al. *Markers Associated with Elite Endurance Perfomance in Russians* // Казань, 20.11.2014.

9) Конференция «NGS в медицинской генетике». Н.А.Кулемин, К.А.Бабалян, Э.В.Генерозов. *Адаптация алгоритмов поиска полиморфизмов к результатам ионного полупроводникового секвенирования* // Суздаль, 24.04.2016.

10) Конференция «NGS-2016». Н.А.Кулемин, К.А.Бабалян, А.В.Гришин, Э.В.Генерозов. *Алгоритм поиска полиморфизмов в результатах ионного полупроводникового секвенирования* // Москва, 19.05.2016.

Глава 1. Обзор литературы

1.1 Биоинформатические методы в современных генетических исследованиях

С развитием физических методов исследований в середине 20 века появилось множество возможностей по изучению различных биологических процессов. Стало возможным получать все больший объем первичных измерений, что привело к активному применению статистики в биологии. В наше время, в связи с активным развитием вычислительных мощностей, решение большинства задач обработки биологической информации возможно ускорить с использованием информационных технологий.

Биоинформатика - отрасль информатики (теории информации), занимающаяся теоретическими вопросами хранения и передачи информации в биологических системах.

Основной целью биоинформатики является уточнение понимания и формализация биологических процессов.

Биоинформатика как самостоятельная наука включает в себя:

1) математические методы компьютерного анализа в сравнительной геномике, протеомике, метаболомике и прочих разделах молекулярной биологии;

2) вычислительная биология эволюции;

3) сравнительная геномика;

4) генетика заболеваний и ассоциативные исследования, а также онкогеномика;

5) анализ экспрессии генов, протеомика;

6) анализ структуры и свойств молекул в биологии;

7) системная и сетевая биология, анализ биологических взаимодействий;

8) анализ публикаций, а также анализ изображений, полученных в результате работы каких-либо приборов [6].

Из всех разделов, описанных выше, геномика (раздел молекулярной генетики, посвящённый изучению генома и генов живых организмов) является

самой крупной ветвью биоинформатики. По некоторым оценкам, более 70% публикаций, отнесенных к биоинформатике, содержат в себе понятие генома в том или ином контексте [7].

Чаще всего к методам информатики прибегают при анализе непосредственных сигналов какого-либо физического метода измерения с целью получить осмысленный биологический результат. Решение таких задач, как определение последовательностей ДНК, РНК или белков в настоящее время невозможно представить без методов биоинформатики.

Другим важнейшим направлением биоинформатики является статистическое сравнение и поиск биологических различий между организмами. При наличии большого числа измерений в задачах статистического сравнения информатика играет важнейшую роль потому, что информационная сложность большинства задач сравнения составляет величину, пропорциональную произведению числа образцов и числа учитываемых параметров.

В генетических исследованиях кроме анализа непосредственных данных секвенирования ДНК или чип-гибридизации биоинформатические методы также применяют для аннотирования функциональных последовательностей организмов, геномная последовательность которых получена впервые, как на основе сравнительного анализа с филогенетически близкородственными организмами, так и с использованием информации о фундаментальных биологических взаимодействиях.

Кроме того, множество приложений биоинформатики присутствуют и в системной биологии: моделирование экспрессии, генетических эффектов, молекулярного взаимодействия и биологических процессов. В генетике человека применение биоинформатики также широко: ассоциативные исследования, моделирование фенотипических проявлений при наличии определенных генотипов, оценка уровня экспрессии генов, аннотация неизученных участков генома.

На 2015 год количество последовательностей полных человеческих геномов, размещенных в открытом доступе, приблизилось вплотную к 1

миллиону образцов [8]. Так как большинство исследований требуют больших выборок и множества сопутствующих данных, то принято размещать геномы в открытом доступе для того, чтобы каждый имел возможность их использовать обезличено в собственных исследованиях. Объем информации, которую возможно получить при мета-анализе как уже отсеквенированных, так и новых геномов, значителен, а результаты исследований могут иметь множество приложений: от сельского хозяйства и животноводства до персонифицированной медицины.

Хотя уже более 200 лет известно о генетической природе многих редких наследственных заболеваний, огромный объем уже проведенных исследований показал, что развитие большинства распространенных, социально значимых заболеваний (сердечно-сосудистые синдромы, рак, диабеты и прочие) также происходит под влиянием генетических факторов. Поэтому, значительное число биоинформатических исследований ведется в медицинской генетике [9].

В будущем информация о геноме пациента позволит корректировать назначение лекарственных препаратов и применение специализированных медицинских средств, а также предсказывать наступление многих симптомов различных заболеваний.

1.2 Полногеномные ассоциативные исследования в медицинской генетике

Полногеномное ассоциативное исследование (англ. GWAS, Genome-Wide Association Studies) - направление биомедицинских исследований, связанных с исследованием ассоциаций между геномными вариантами и фенотипическими признаками [3].

Основным фенотипическим признаком для такого исследования может быть как какой-либо элемент внешнего вида пациентов (цвет глаз, форма ушных раковин и т.п.), так и наличие какого-либо охарактеризованного заболевания или симптома.

Результатами полногеномных ассоциативных исследований может быть информация об обнаруженных ассоциациях с различными генетическими аберрациями: хромосомными мутацями (целые хромосомы или их фрагменты), крупными вставками или делециями (100 – 100 000 нуклеотидов), небольшими вставками или делециями (от 1 до 100), единичными заменами нуклеотидов (однонуклеотидными полиморфизмами, SNP). Частота встречаемости более крупных мутаций обратно пропорциональна их протяженности. Указанный факт следует из того, что более крупные вставки или делеции влияют на большее число систем в организме, что зачастую приводит к смерти носителя мутации в раннем возрасте или до рождения. Именно поэтому большинство ассоциаций, обнаруженных в процессе полногеномных исследований, являются однонуклеотидными полиморфизмами.

В среднем однонуклеотидные различия между геномами двух людей обнаруживаются с частотой 1 на 1000 оснований, при этом во всем человеческом геноме содержится 3,2 миллиарда пар нуклеотидов [10]. Каждый полиморфизм всегда имеет как минимум 2 состояния: минорное (более редкое) и мажорное (более 50% встречаемости в популяции). Частота встречаемости того или иного генотипа в конкретном полиморфизме зачастую различна у представителей различных рас и этносов, что приводит в том числе и к медицинским особенностям у представителей таких народов.

Было показано, что нуклеотидные полиморфизмы влияют не только на внешние различия между людьми, но и на устойчивость организма к различным заболеваниям и внешним воздействиям [11]. Сейчас известно более 165 миллионов полиморфизмов генома человека (согласно базе данных нуклеотидных полиморфизмов Национального Центра Биотехнологической Информации Соединенных Штатов Америки (dbSNP NCBI USA)) [12]. Те генетические вариации, каждая из аллелей которых встречается более чем у 0,01% человечества, были пронумерованы и им был присвоен индекс rs [3] [13].

При проведении GWAS в области медицинской генетики существует сложность, как выбора общего признака, так и последующей интерпретации

результата, связанная с мультифакториальностью большинства часто встречающихся заболеваний. Такие заболевания чаще всего вызываются большим количеством различных полиморфизмов, вносящих вклад как по-отдельности, так и в комбинации. Одни и те же проявления, например, диабета 1 типа могут быть вызваны как крупной наследственной мутацией, так и множеством полиморфизмов в различных генах, которые влияют на риск возникновения и развития заболевания в совокупности. В рамках ассоциативного исследования формируются как минимум 2 группы испытуемых: те, кто обладают указанным признаком; а также те, кто им не обладают. В результате исследований составляется список маркеров, которые достоверно отделяют одну группу от другой. Согласно законам статистики, чем большее число пациентов вовлечено в исследование, тем достовернее результат ассоциации. С другой стороны, увеличение числа образцов увеличивает и математическую сложность обработки результатов. Биоинформатические методы позволяют обходить проблему вычислений для большого числа образцов с помощью применения ЭВМ и специализированных алгоритмов.

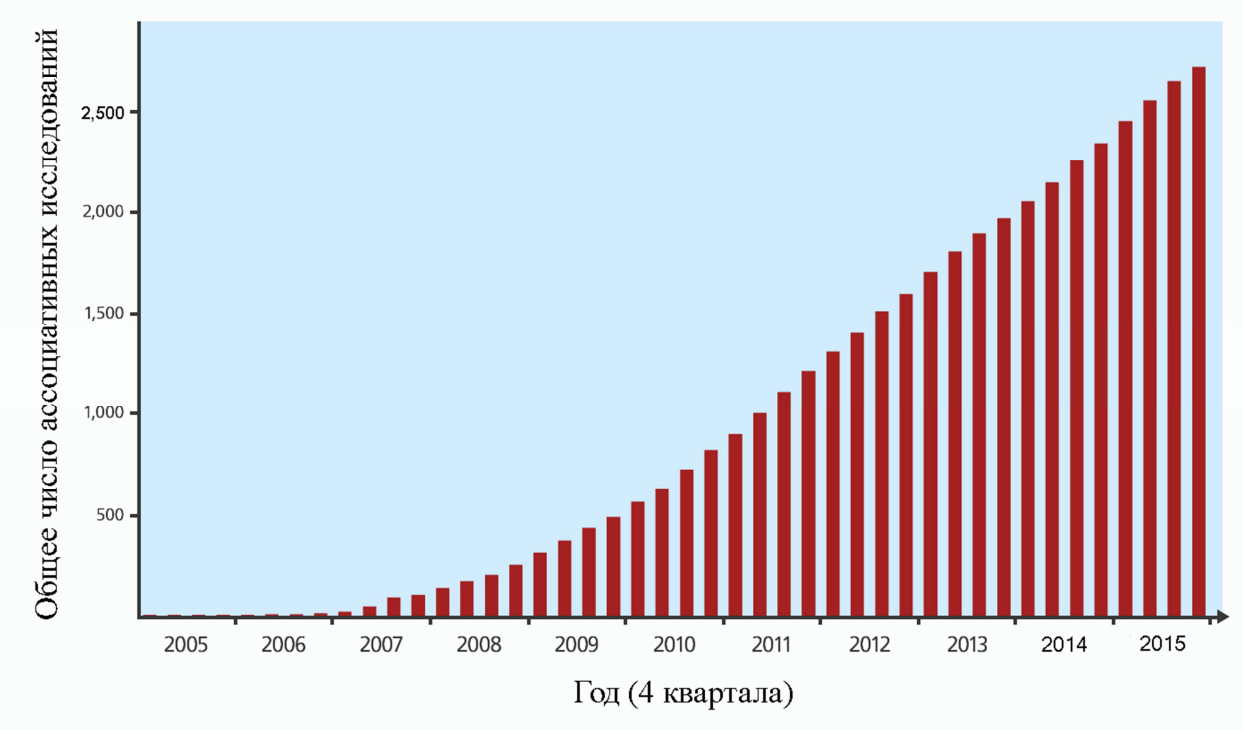

Рисунок 1. Количество опубликованных ассоциативных исследований к 2015 году [14] [15].

Количество опубликованных ассоциативных исследований увеличивается с каждым годом все быстрее из-за активного развития методов генотипирования (Рис.1). На начало 2016 года известно более 3000 опубликованных ассоциативных исследований с различным количеством участников [15].

Так как самым распространенным результатом, полученным в результате ассоциативного исследования, является однонуклеотидный полиморфизм, то для большинства изучаемых заболеваний существуют обширные списки из обнаруженных ассоциированных полиморфизмов, которые требуется обосновать с точки зрения биологических процессов. Например, на текущий момент известно более 2000 полиморфизмов, ассоциированных с поликистозом почек, но только менее чем для 1300 полиморфизмов показан биологический механизм влияния на развитие заболевания [16] [17].

В противоположность поликистозу почек, существует множество мультифакторных фенотипов, при анализе которых путем ассоциативного исследования удавалось достичь высокой предиктивной точности полученных полиморфизмов. К таким заболеваниям можно отнести возрастную макулодистрофию или наследственные формы рака молочной железы [18]. Макулодистрофия – заболевание, которое присутствует в различных формах более чем у 5% жителей старше 60 лет, но, с другой стороны, ее развитие возможно задержать путем изменения образа жизни и принятии определенных лекарственных препаратов. Крупному научно-клиническому консорциуму, ведущему деятельность на трех континентах, удалось путем наблюдения за пациентами в клиниках в течение более чем 10 лет построить релевантную модель генетического теста для оценки вероятности развития макулодистрофии с предиктивной точностью более 85% (для различных популяций точность различается) [18].

Как и в любых других научных экспериментах, существует значительное число трудностей, которые приходится преодолевать при проведении ассоциативных исследований. Существует множество технических препятствий, самым значимым из которых является огромное количество различных платформ

и подходов к генотипированию, в результате работы каждой из которых формируется свой собственный список маркеров со своим профилем физических ошибок. Кроме научной стороны, существуют и трудности отбора пациентов для проведения исследования. В первую очередь это необходимость проведения юридических согласований и получения информированного согласия. Во-вторых, обязательна доступность максимального объема медицинской информации: если у двух сформированных групп есть какой-либо другой признак (например, сопутствующее заболевание), который их различает, то в результате ассоциативного исследования возможно обнаружение маркеров, ассоциированных с ним, а не с основным признаком.

Часто для изучения генетических факторов развития некоторых заболеваний обосновано использование групп профессиональных спортсменов как лиц с уникальными физиологическими характеристиками. Такие ассоциативные исследования обладают высокой информативностью, потому что помогают изучить границы физиологической адаптации, что ценно как для развития медицины, так и для дальнейших спортивных достижений.

1.2.1 Генетические исследования в спортивной медицине

Большинство физиологических показателей, актуальных для спортсменов, во многом зависят от генетических характеристик их обладателя. Еще в 20 веке было замечено, что во многих видах спорта дети могут быть настолько же успешными, как и их родители - особенно, если они оба спортсмены. Было показано, что такие факторы, как максимальная скорость бега, время реакции рукой на зрительное воздействие, быстрота и согласованность движений руками, статическая сила, взрывная сила (разгибание) наследуются на 60-70%, тогда как время коленного рефлекса, максимальная частота нажатий ладонью на кнопку или динамическая сила мышц (при отжимании) - на 95 - 99% [19].

Генетические ассоциации с указанными показателями могут быть полезны и в персонифицированной медицине. Так, информация об объеме вдыхаемого кислорода может быть полезна при лечении проявлений астмы [20], а количество

гемоглобина в крови может быть связано с различными заболеваниями печени [21].С другой стороны, исследования индивидуальных особенностей спортсменов могут подсказать тренерам оптимальную стратегию тренировок и питания, а также для детей может быть проведена оценка потенциальной успешности в том или ином виде спорта [19].

Наличие функционально значимых ДНК-полиморфизмов в генах, участвующих в функционировании сердечно-сосудистой системы и опорно-двигательного аппарата, предполагает выявление их взаимосвязи с физическими качествами человека, развивающимися в онтогенезе под значительным влиянием среды. Установление ассоциаций полиморфизмов данных генов с предрасположенностью к выполнению физических упражнений различной длительности и интенсивности, а также с фенотипами, значимыми в условиях спортивной деятельности, позволит разработать систему критериев прогностической оценки физических способностей человека. В одном из крупнейших полногеномных ассоциативных исследований (320485 участников исследования, более 20 клинических центров) проводилось сравнение возраста, пола и спортивной активности с формой тела у взрослых и поиск генетических маркеров, которые регулируют указанные процессы [22]. В качестве технологии для определения геномной информации использовалась специализированная ДНК-панель с нанесенными олигонуклеотидными зондами, тестирующими геномные состояния, которые находятся в генах, для которых была ранее найдена ассоциация с индексом массы тела и скоростью метаболизма различных липидов. Исследование показало, что более 60% совокупной дисперсии индекса массы тела, как у спортсменов, так и у тех, кто не занимается спортом регулярно, объясняются генетическими факторами. С другой стороны, некоторые гены, влияющие на обмен липидов у женщин, не являются значимыми у мужчин и наоборот. Полученные результаты допустимо применять как для лечения ожирения (путем корректировки баланса веществ в организме), так и для достижения спортивной формы при занятиях любительским и профессиональным спортом [22].

Все циклические виды спорта допустимо разделить на две большие, противостоящие друг другу группы: виды спорта на выносливость (категория спортсменов: стайеры) и виды спорта на быстроту (категория спортсменов: спринтеры). Типичный спортсмен-спринтер должен обладать возможностями быстро производить и использовать энергию во время мышечной деятельности, тогда как спортсмен-стайер должен обладать способностью производить больше энергии, но с меньшей скоростью для выполнения продолжительных по времени физических нагрузок. Все описанные особенности, так или иначе, ассоциированы с генетическими факторами, но исследований, ассоциирующих конкретные физические показатели со спортивной успешностью в каждой из групп, было выполнено немного.

Одним из важных критериев в видах спорта, требовательных к повышенному уровню выносливости, который во многом детерминирован на генетическом уровне, является максимальная скорость потребления кислорода (МПК или VO_2max; измеряется в абсолютных (л/мин) либо относительных (мл/мин/кг) единицах). МПК зависит от двух основных факторов: доставки кислорода в мышцы (легочная система и транспортная система крови) и его утилизации мышечными волокнами. МПК определяют с помощью газоанализатора на различных тренажерах (беговая дорожка, велоэргометр, гребной эргометр и т.д.) в тесте со ступенчато возрастающей до предельно возможной нагрузкой. Во множестве исследований было показано, что потребляемый кислород во время нагрузок используется для аэробной генерации энергии [23].

Процесс усвоения кислорода состоит из множества физических и химических взаимодействий, в которых принимает участие более 100 различных белков и других факторов, участвующих в метаболизме жиров и углеводов. Баланс и обмен многих из них разительно влияет на усвоение кислорода организмом.

Другим важным для спортсменов-атлетов свойством является состав мышечных волокон, который у взрослых практически не изменяется под

влиянием физических упражнений. В зависимости от их свойств волокна подразделяют на 3 типа: волокна I типа (медленные, окислительные), волокна II типа (быстрые окислительные) и волокна IIX типа (быстрые, гликолитические). Способом определения структуры мышц является биопсия небольшого объема мышечного материала рабочей мышцы (например, четырехглавой мышцы бедра) с дальнейшим иммуногистохимическим анализом под микроскопом. Состав мышечных волокон выражают в виде процентного соотношения каждого из типов волокон [24].

К настоящему моменту не проводились масштабные полногеномные исследования аэробных возможностей и структуры мышц у ведущих спортсменов России, с целью обнаружения генетических ассоциаций, присущих именно жителям нашей страны. На технологии каждого из шагов анализа как геномной, так и биофизической информации стоит остановиться подробнее.

1.3 Биоинформатический анализ геномных данных

Современная методология ассоциативных генетических исследований базируется на использовании различных протоколов масштабного генетического анализа. К таким технологиям относят современные варианты секвенирования ДНК, а также методы высокоплотного анализа нуклеотидных полиморфизмов на ДНК-чипах. Как и в большинстве современных развивающихся технологий, при активном развитии приборной базы, технологии информационной обработки сформированного объема данных существенно отстают и являются основным предметом для биоинформатического анализа и дискуссий. Так, штатное программное обеспечение, предоставляемое вместе с секвенаторами, зачастую не предоставляет возможности использовать весь объем получаемой генетической информации. Кроме того, оно не позволяет в достаточной мере снизить частоту ошибок секвенирования, связанных с технологическими особенностями приборов и имеет ограниченное количество пользовательских настроек. Обработки данных ДНК-чипов является технически более простой, но и в этом случае есть ряд

сложностей, связанных с необходимостью приведения результатов генотипирования к единому формату для последующего сравнения.

В связи с этим, разработка альтернативных биоинформатических протоколов, позволяющих оптимизировать техническую первичную обработку получаемых с секвенаторов данных, включая их сборку, картирование на референсный геном, идентификацию уникальных вариантов является актуальной методической задачей.

Другой точкой приложения биоинформатических методов является непосредственно выполнение необходимых расчетов при проведении полногеномных ассоциативных исследований.

И, наконец, сами результаты, полученные в ходе таких исследований, являются предметом дальнейшей биоинформатической обработки. Так, типичным результатом групповых ассоциативных генетических исследований является информация по достоверной ассоциации генетических маркеров с одним определенным фенотипическим признаком. При этом существенный интерес представляет задача разработки алгоритмов комплексной оценки генетического риска, аккумулирующих общедоступную информацию по множеству других информативных генетических маркеров и, в зависимости от их комбинации у исследуемого индивидуума, оценивающих его персональный совокупный генетический риск по различным медицинским и физиологическим показателям. Такие алгоритмы моделирования персонального риска могут учитывать как генетические, так и негенетические факторы (возраст, пол, курение и подобные). Предполагается, что именно они в будущем и будут завершать полный цикл генетического исследования пациентов с последующим переходом к персонифицированной медицине.

1.3.1.1 Современные технологии секвенирования ДНК

Собственно, до разработки технологий секвенирования второго поколения потребности в биоинформатическом анализе первичных данных в генетических исследованиях были незначительными. Это связано, в первую очередь, как с

низкой производительностью используемых технологий, так и с относительной простотой обработкой приборных данных.

Геномика человека в различных форматах и с применением различных технологий активно развивалась еще задолго до появления высокопроизводительных секвенаторов (Табл.1). Каждый из используемых подходов имел свои достоинства и недостатки, но только возможность относительно дешево и относительно быстро прочитать полный геном человека позволила всерьез заговорить о наступлении эры развития персонифицированной медицины.

Среди всех методов «золотым стандартом» секвенирования остается технология терминирующих меченых нуклеотидов, разработанная в 1980х годах Фредериком Сэнгером. Реализованная в формате детекции результатов секвенирования посредством капиллярного электрофореза в геле с флуоресцентными красителями, данная технология обеспечила высокое качество и точность определяемых последовательностей ДНК. На сегодняшний день, в большинстве стран мира данный метод до сих пор остается единственным клинически-подтвержденным методом диагностики наследственных генетических заболеваний. Это означает, что любая находка в геноме пациента с помощью других методов обязательно должна быть подтверждена результатами секвенирования по Сэнгеру. Однако чтение полных геномов, экзомов и таргетных панелей возможно исключительно в формате технологий секвенирования второго поколения. Данные технологии используют в своей основе различные базовые принципы определения нуклеотидной последовательности ДНК – сиквенс посредством синтеза, сиквенс посредством лигирования, пиросеквенирование, полупроводниковое секвенирование. Однако их роднит крайне высокая производительность и сложность интерпретации получаемой первичной информации, делает задачу получения непосредственной структуры последовательности совсем не тривиальной и не однозначной.

Таблица 1. Основные подходы в арсенале геномики человека до появления платформ NGS [25].

Метод	**Технология**	**Область применения**	**Ограничения в применении**
Метод Сэнгера (метод терминирующих нуклеотидов)	Нуклеотиды, меченные флуоресцентными красителями Фрагменты ДНК разделяются капиллярным электрофорезом	“Золотой стандарт”, секвенирование первого поколения	Требует много времени, дорог в использовании, относительно низкая производительность. Не может обнаружить делеции, транслокации и изменение числа копий генов
Минисеквени-рование с детекцией результатов в режиме MALDI-TOF	Удлинение праймера-зонда с участием терминирующих нуклеотидов. Детекция различия масс в зависимости от концевого нуклеотида	Высокая чувствительность, высокое разрешение, обнаружение мутаций со сдвигом рамки считывания, обнаружение SNP	Разрешение масс-спектрометрии вкупе с требованиями к дизайну праймер-зондов для ПЦР ограничивают масштабируемость
Аллель-специфичная ПЦР в реальном времени	Праймеры ограничивают участки интереса на ДНК Зонды обнаруживают конкретные известные мутации	Крайне высокая чувствительность, широко используется в клинической диагностике онкогенных мутаций	Применение ограничено "горячими точками" генома
ПЦР в реальном времени с анализом кривой плавления	Температура плавления гетерогенных ПЦР-продуктов отличается от таковой у гомогенных	Высокая чувствительность, даёт информацию о соотношениях мутантных и нормальных генов	Часто трудно уловить различия между кривыми плавления, трудно стандартизовать, возможности мультиплексирования ограничены.

На сегодняшний день в распоряжении молекулярных биологов находятся несколько платформ, каждая из которых может найти свое применение в решении задач секвенирования генома [25]. Эти приборы отличаются друг от друга по таким параметрам, как максимальная длина прочтения (в парах нуклеотидов), производительность за один запуск (в миллиардах пар нуклеотидов), длительность одного запуска (в сутках либо часах), а так же стоимость одного запуска (варьирует в зависимости от типа библиотеки, типа прочтения и количества образцов). Соответственно, эффективность применения каждого конкретного прибора для решения той или иной задачи должна зависеть в первую очередь от параметров требуемого результата.

Выбор той или иной платформы определяется поставленной перед исследователем задачей. При этом оценка эффективности применения каждого из приборов часто затрудняется тем, что в число основных характеристик производитель чаще всего выносит максимально возможные показатели для каждого из приборов (Табл.2).

Секвенирование генома человека с применением технологий NGS может осуществляться в нескольких форматах. В первую очередь следует отметить, что все работы в данной области, следующие за первым прочитанным полным геномом человека [26] [27], осуществляются в режиме ресиквенса. Данная процедура не подразумевает сборки геномной последовательности *de novo* и не требует предварительного разделения образца ДНК на отдельные хромосомы. Анализ массива первичных данных осуществляется посредством выравнивания ридов на референс, причем в данном случае рекомендуется использовать не менее чем 30-кратную глубину покрытия [28]. В целом, чем выше покрытие при секвенировании, тем выше точность определения каждого из нуклеотидов генома [29]. Рост точности в данной ситуации не является линейным, и достигает насыщения приблизительно к величине покрытия равной 100 [30]. Разнообразие результатов, которые могут быть получены в процессе анализа, напрямую зависят от типа используемой библиотеки и длины индивидуального прочтения.

Таблица 2. Основные характеристики платформ секвенирования нового поколения (NGS) [31].

Платформа (технология**)**	**Область применения**	**Особенности**
Roche 454 GS FLX+ Эмульсионная ПЦР, сиквенс посредством синтеза, детекция пирофосфата	De novo секвенирование небольших геномов, направленное секвенирование, глубокое секвенирование ампликонов	**Длительность запуска:** ~34 часа **Длина прочтения:** ~700 пн. **Производительность:** 700 Mb **Преимущества:** Длинные прочтения, средняя производительность. **Недостатки:** Высокая стоимость, ошибки при прочтении гомополимеров, отсутствие системы автоматизации ПЦР.
Roche 454 GS Junior Эмульсионная ПЦР, сиквенс посредством синтеза, детекция пирофосфата	De novo секвенирование небольших геномов, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** 10 часов **Длина прочтения:** ~400 пн. **Производительность:** 35 Mb **Преимущества:** Компактный формат прибора, длинные прочтения, небольшая длительность запуска. **Недостатки:** Низкая производительность, ошибки при прочтении гомополимеров, отсутствие автоматизации пробоподготовки.
Illumina HiSeq 1500/2500 Мостиковая ПЦР, сиквенс посредством синтеза с удаляемыми терминаторами, флуоресцентно-меченные нуклеотиды	De novo секвенирование геномов любого размера, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** ~7 дней, быстрый прогон ~27 часов **Длина прочтения:** 1x50 пн., 2x100 пн., 2x 125 пн., 2x250 пн. **Производительность:** 200 Gb - 1 Tb **Преимущества:** Наличие двух модулей, позволяющих работать в форматах максимальной и минимальной загрузок, высокая производительность. **Недостатки:** Высокая стоимость, длительный запуск при максимальной загрузке, короткие риды, ошибки определения однонуклеотидных замен.
Illumina MiSeq Мостиковая ПЦР, сиквенс посредством синтеза с удаляемыми терминаторами, флуоресцентно-меченные нуклеотиды	De novo секвенирование небольших геномов, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** 4 - 39 часов **Длина прочтения:** 1x50 пн., 2x75 пн., 2x 150 пн., 2x250 пн., 2x300 пн. **Производительность:** 540 Mb - 11 Gb **Преимущества:** Компактный формат, высокая производительность, небольшая длительность запуска. **Недостатки:** Короткие риды, ошибки определения однонуклеотидных замен.

Life Technologies SOLiD Эмульсионная ПЦР, сиквенс посредством лигирования, флуоресцентно-меченные зонды	De novo секвенирование геномов любого размера, направленное ресеквенирование	**Длительность запуска:** 5-14 дней **Длина прочтения:** 50-75 пн. **Производительность:** 77-155 Gb **Преимущества:** Высокая точность, высокая производительность, гибкий формат. **Недостатки:** Длительный запуск, короткие риды.
Life Technologies Ion Torrent Эмульсионная ПЦР, сиквенс посредством синтеза, pH-метрия	De novo секвенирование небольших геномов, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** 2-4 часа **Длина прочтения:** 200-400 пн. **Производительность:** 50 Mb - 1 Gb **Преимущества:** Компактный формат, средние прочтения, небольшая длительность запуска. **Недостатки:** Ошибки при прочтении гомополимеров.
Complete Genomics Амплификация «клубков» ДНК, сиквенс посредством лигирования	Полногеномное секвенирование	**Длительность запуска:** 12 дней **Длина прочтения:** 70 пн. **Производительность:** 20-60 Gb **Преимущества:** Возможность прочтения полного генома. **Недостатки:** Отсутствие других областей применения.
Helicos Единичная молекула, сиквенс посредством синтеза с удаляемыми терминаторами, флуоресцентно-меченные нуклеотиды	Полногеномное секвенирование, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** 8 дней **Длина прочтения:** 35 пн. **Производительность:** 21-35 Gb **Преимущества:** Отсутствие этапа клональной амплификации. **Недостатки:** Низкая точность.
Pacific Biosicence PacBioRS Единичная молекула, нанопоры	Полногеномное секвенирование, направленное ресеквенирование, глубокое секвенирование ампликонов	**Длительность запуска:** ~24 часа **Длина прочтения:** 3-10 пн **Производительность:** 1.2 Gb **Преимущества:** Отсутствие этапа клональной амплификации, наибольшая длина прочтения, небольшая длительность запуска. **Недостатки:** Наиболее низкая точность.

Теоретически, фрагментные библиотеки любого секвенатора второго поколения позволяют эффективно обнаруживать единичные нуклеотидные полиморфизмы в составе генома [32]. При этом длина прочтения существенно не

влияет на итоговый результат, и на первый план выступает производительность прибора [33].

Несмотря на тенденцию к значительному снижению себестоимости сиквенса генома человека, все равно цена за образец остается слишком высокой, чтобы научное сообщество могло себе позволить по-настоящему масштабные исследования. В данном случае альтернативным является подход, позволяющий направленно секвенировать экзомы (кодирующие участки генома) или любые другие фрагменты ДНК человека. Существует два конкурирующих подхода к формированию библиотек с целевыми фрагментами и обогащению ДНК:

1) Гибридизационный – когда с использованием образца геномной ДНК человека исходно готовится стандартная фрагментная библиотека, после чего в результате процедуры гибридизации со специфическими зондами осуществляется ее направленное обогащение кодирующими последовательностями генома человека. Особенностью метода является нормальное распределение достоверности определения нуклеотида по отношению к расстоянию от зонда, которое определяется случайностью фрагментации ДНК при приготовлении исходной библиотеки. Данная технология реализуется в формате коммерчески доступных наборов, которые предназначены для последующего секвенирования экзома с использованием различных секвенаторов линейки Illumina либо ThermoFisher.

2) Амплификационный - данная технология построена на реакции амплификации ДНК. Предварительно кодирующая часть генома человека (экзом) разбивается на перекрывающиеся фрагменты ДНК, и подбираются пары специфических праймеров, которые позволяют осуществить направленную амплификацию данных участков ДНК. В формате, пригодном для экзомного секвенирования данный подход реализован исключительно в формате Ion Ampliseq Exome.

В последние годы наметилась тенденция унификации структуры библиотек по отношению к последующей технологии секвенирования, что благоприятно

сказалось на увеличении числа и качества реагентов для приготовления библиотеки [34].

Самой популярной технологией, применяемой для секвенирования экзомов и целевых фрагментных панелей генома человека, на текущий момент является секвенирование посредством синтеза, реализованным в формате линейки секвенаторов производства компании Illumina, и с ее помощью получено более 80% всей генетической информации в 2015 году. Самой распространенной ошибкой данной технологии является неточность в определении флуоресцентного сигнала [35], что приводит к ошибочному распознаванию нуклеотида в конкретной позиции, и в свою очередь, при дальнейшем анализе данных – к появлению ложных нуклеотидных замен в том или ином гене.

Второй по популярности технологией является полупроводниковое секвенирование, реализацией которой осуществляет компания ThermoFisher. Самой распространенной ошибкой данной технологии является невозможность точного разрешения гомополимерных последовательностей, что при дальнейшем анализе данных приводит к появлению ложных вставок/делеций, и как следствие – ложных обрывов рамки считывания в том или ином гене.

На текущий момент стоимость прочтения одного нуклеотида с использованием полупроводникового секвенирования ниже, чем при технологии секвенирования синтезом. Именно различием в цене и объясняется такой высокий интерес научного сообщества к продуктам линейки Ion Proton [36].

Таким образом, в настоящее время направленное ресеквенирование генома человека прочно вошло в набор методов анализа ДНК, и в зависимости от области интереса для осуществления процедуры могут подходить различные платформы NGS.

1.3.1.2 Задачи биоинформатической обработки первичных результатов секвенирования

Обработка результатов высокопроизводительного секвенирования является сложной и ресурсоемкой задачей. В качестве входных данных протокола и, одновременно, выходных данных со всех современных секвенаторов

используются файлы формата FASTQ (Рис.2). Данный формат файла является текстовым, поэтому его возможно открыть любым текстовым редактором в любой операционной системе.

```
@nimblegen_GEN:846_2155_1795
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
X[[[]]]]]]]]]]]]]]]XX]]]][]]]]]]]]]]]]]]]]ZMY]
@nimblegen_GEN:462_3939_166
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
]]]]]]]]]]]]]]]]]]]]]]]]]]]YW]]]]Q!#LT]X!.]V
```

Рисунок 2. Структура FASTQ-файла на примере двух прочтений.

FASTQ-файл состоит из блоков по 4 строки, каждый блок относится к единичному прочтению. Первая строка внутри блока - это техническое название прочтения, начинающееся с "@" и содержащее любые символы. Во второй строке идет нуклеотидная последовательность прочтения, а в третьей всегда расположен символ "+", являющийся знаком-разделителем. Четвертая строка в точности соответствует по длине второй строке, а каждый символ четвертой строки является указателем на техническую достоверность определения каждого нуклеотида соответственно. [37]

После получения FASTQ-файлов требуется выровнять прочтения на известную последовательность генома человека, а затем произвести поиск тех нуклеотидов, которые отличаются от стандартной референсной последовательности.

В зависимости от типа используемой химии для обогащения генома требуется предварительная обработка полученных прочтений с прибора.

В случае если технология является гибридизационной и предполагает случайное разделение нити ДНК на первом шаге, то, с точки зрения теории вероятностей, впоследствии не должно получаться абсолютно одинаковых прочтений. Предполагается, что они являются побочными продуктами амплификации, поэтому они могут содержать ошибки полимеразы а,

следовательно, вносить существенные ошибки в результат определения геномных состояний. Для предварительной обработки таких геномных данных используется программный пакет trimmomatic [38] [39].

В случае же ампликонной технологии допускается возможность сохранения части биологических адаптеров при формировании прочтений, которые используются секвенатором для определения начала и конца каждого нового ампликона. Для увеличения точности требуется подсчитать количество прочтений, начинающихся с одинаковых 3 - 5 символов. Если данное число выходит за рамки нормального распределения, то данный фрагмент требуется удалить - он является частью неотрезанного адаптера. Кроме того, для данных, полученных по технологии ионного полупроводникового секвенирования также стоит удалять прочтения, содержащие в своей последовательности один повторяющейся нуклеотид более чем в 40% длины всего прочтения - скорее всего такая ситуация является следствием ошибки определения силы ионного тока [40]. Для решения обеих задач обычно используют пакет cutadapt.

На следующем этапе требуется сформировать файл, содержащий информацию о выравнивании полученных прочтений формата BAM. Данный тип файла формируется специальными программами-выравнивателями, которые используя различные математические алгоритмы, размещают полученные прочтения на известной последовательности генома человека [41]. На данный момент актуальной версией человеческого генома является GRCh38, но количество баз данных, содержащих записи в системе координат этой версии мало, поэтому до сих пор для большинства задач используется выравнивание прочтений на сборку hg19 (GRCh37) в ее финальном варианте от 15.10.2013. В случае если используемый набор для обогащения был изготовлен в системе координат hg19 или старше, потери данных не происходит. В настоящее время начали появляться панели целевого обогащения, которые разработаны позднее конца 2013 года и используют уже систему координат GRCh38 при изготовлении праймеров или зондов, например Roche MedExome или Agilent SureSelect V6. В

случае выравнивания таких наборов на более старые сборки генома возможно искусственное формирование ошибок.

Выравнивание прочтений является одной из основных проблем биоинформатики и является в большинстве своем задачей именно алгоритмической, нежели биологической. Понятно, что при выравнивании требуется предусмотреть возможность несоответствия нескольких нуклеотидов или даже целых участков исходному геному (мутации, полиморфизмы тестируемого образца). Также, существуют наборы для обогащения, которые допускают получение в итоге прочтений, состоящих из двух объединенных фрагментов, полученных с разных отрезков генома. Тогда, такие прочтения должны быть разделены на два отдельных и выровнены отдельно. Кроме всего прочего, наборы, в которых секвенирование ведется с обоих концов некоторого отрезка ДНК, ограниченного праймерами, позволяют уточнить результаты выравнивания информацией о парной связи таких отрезков (Paired End у Illumina, Mate Pair у Life Technologies/ThermoFisher).

Известно большое число различных алгоритмов, которые отличаются скоростью работы или адаптацией к различному соотношению параметров: длина референса, длина прочтения, число прочтений. Для большого числа прочтений и длинных референсных последовательностей, которой и является геном человека, предпочтительным является алгоритм индексирования последовательностей Барроуза-Виллера (Burrows-Wheeler transform) и алгоритм поиска места выравнивания Смита-Ватермана (Smith–Waterman algorithm) [42].

Алгоритм Смита-Ватермана является алгоритмом динамического программирования, при котором вычисляется матрица соответствия от последовательности нуклеотидов в зависимости от возможного места начала выравнивания. Технологию работы алгоритма возможно изучить на следующем примере. Предположим, что последовательность 1 состоит из нуклеотидов ACACACTA, а последовательность 2 - AGCACACA. Тогда, $f(a_n, b_n) = f(a_{n-1}, b_{n-1}) + 2$, если нуклеотиды одинаковые и $f(a_n, b_n) = f(a_{n-1}, b_{n-1}) - 1$ если различные. На основании такого утверждения составляется матрица соответствия H. После ее

заполнения производится последовательный переход по ячейкам матрицы от самого большого значения (12) в одну из трех соседних ячеек, содержащую максимальное числовое значение до тех пор, пока не достигается ячейка с нулевым значением. Таким образом, формируется матрица перехода (Рис.3), на основании которой имеется возможность восстановить выравненные последовательности: A-CACACTA и AGCACAC-A. Каждый переход вверх на матрице переходов обозначает делецию в выравниваемом отрезке, а каждый переход влево - вставку. [43]

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 1 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 1 & 3 & 2 & 3 & 2 & 3 & 2 & 2 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 4 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 6 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 8 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 10 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix} \quad T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \end{pmatrix}$$

Рисунок 3. Матрица состояний и матрица переходов.

Алгоритм Смита-Ватермана используется для выравнивания одного прочтения на отрезок генома, тогда как его дополнение преобразованием Барроуза-Виллера позволяет одновременно вычислять значение матрицы Смита-Ватермана для всех прочтений, причем без загрузки всех последовательностей в память. Основой преобразования является формирование специального FM-индекса всех прочтений. Для каждого прочтения он формируется в следующем порядке. Пусть исходное прочтение представляет собой отрезок ATGC., где точка - конец строки. Далее циклически производится перестановка символов влево по порядку и формируется матрица перестановок (Рис.4), которую затем сортируется по алфавиту, начиная с первого символа.

Затем сохраняется номер строки, в которой находится исходная последовательность, а также последний столбец полученной матрицы перестановок. Полученная строка обладает некоторыми интересными свойствами: 1) Если в исходной последовательности много одинаковых символов, то в

результате FM преобразования они будут сгруппированы последовательно, что сильно облегчает архивирование (подход используется в архиваторе bzip). 2) По FM-индексу гораздо проще восстанавливать матрицу переходов в алгоритме Смита-Ватермана [44] [45].

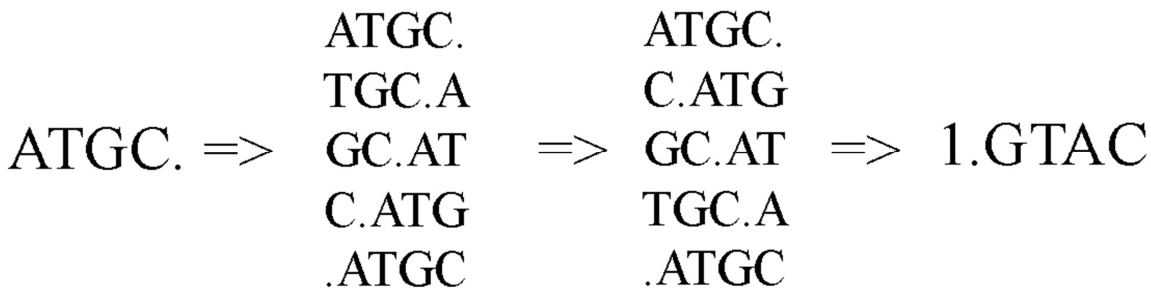

Рисунок 4. Схема преобразования матрицы перестановок.

На основе пары описанных алгоритмов существует значительное множество программных пакетов, различие между которыми заключается в предварительной адаптации к биологическим особенностям того или иного секвенатора. Самыми популярными программными пакетами для выравнивания являются BWA, Bowtie, Bowtie 2.

Для обработки результатов секвенирования ДНК человека, полученных почти со всех современных секвенаторов чаще всего используется пакет BWA, тогда как для меньших объемов прочтений (порядка 1 ГБ) или для меньших геномов (например, бактериальных) используется пакет Bowtie 2 (для длинных прочтений, например полученных с секвенатора Roche 454) или пакет Bowtie (для более коротких или прочтений с секвенатора SOLiD (в colourspace индексе)) [42].

У технологии ионного полупроводникового секвенирования, с точки зрения формирования прочтений, нет каких либо специфических особенностей, таких как парность прочтений или соединение нескольких прочтений в одно, которое должно быть разделено и помещено в различные отрезки генома. Поэтому для получения итогового результата, в целом, подходит любой из перечисленных пакетов. Прочтения, которые содержат гомополимеры, при использовании любого из алгоритмов никак не влияют на итоговое качество сборки - ошибочные

нуклеотиды (прочитанные или пропущенные) будут зафиксированы как вставки или удаления при выравнивании. Пакет BWA выравнивает 15 ГБ прочтений на ЭВМ (24 ядра, 64 ГБ ОЗУ, полностью свободный узел) приблизительно за 30 минут.

В результате работы любого из программных пакетов формируется файл формата BAM, который затем индексируется, например, пакетом samtools index для ускорения работы приложений на дальнейших этапах и формируется файл формата BAI [41].

Важнейшим моментом обработки результатов секвенирования является проверка соответствия полученных результатов исходным параметрам: соответствует ли биоматериал человеческой ткани; насколько качественно были произведены подготовительные операции к секвенированию; соответствует ли результат исходно выбранным регионам; насколько равномерно прочтения покрывают ожидаемые регионы генома. Для ответа на поставленные вопросы используется пакет NGSRich. Этот пакет представляет собой небольшую программу, написанную на языке java, которая на вход принимает индексированный BAM файл и список целевых регионов в формате BED (хромосома, начало региона, конец региона), а в результате работы формирует отчет, который содержит ответы на все поставленные вопросы. В случае низкого качества (менее 90% прочтений попадают в требуемые регионы) или прочих биологических и химических ошибок на этом этапе обычно анализ прекращается, и данный образец в последующую обработку и интерпретацию не передается [46].

В случае если технический контроль качества пройден успешно, то после формирования BAM-файла для медицинской аннотации и группового сравнения геномов требуется формирование файлов формата VCF (Variant Call Format, формат вызова вариантов), который является стандартом при предоставлении списков полиморфизмов и мутаций данного конкретного человека [47]. Существует достаточно много программных пакетов, позволяющих формировать VCF файл из BAM. Основными факторами при выборе пакета являются: возможность формировать как список определившихся отличий от основной

версии генома человека, так и выделять полиморфизмы, которые находятся в заранее сформированном списке; корректность оценки качества прочтения нуклеотида в точке, с учетом особенностей конкретного прибора; вычислительное быстродействие.

Самыми популярными пакетами сейчас являются GATK, SAMtools, FreeBayes. Именно этот этап отвечает за специфичность всего набора программ для альтернативного анализа результатов секвенирования, тогда как этап формирования BAM-файл влияет в основном на чувствительность (определился нуклеотид или нет). Согласно публикациям и проведенной проверки на тестовом образце, наиболее точными (93% нуклеотидов определено верно) при определении полиморфизмов с заменой нуклеотидов без изменения общего числа являются GATK и SAMtools, тогда как при определении вставок и удалений наиболее точным является FreeBayes. При сравнении GATK и SAMtools следует учитывать также и скорость работы алгоритма. При использовании одинаковых входных данных SAMtools производит вычислительную обработку быстрее на 32%, чем GATK [48].

Кроме того, технической особенностью технологии ионного полупроводникового секвенирования являются ошибки, связанные с определением ионного тока, что проявляется в виде ошибок в числе идущих подряд одинаковых нуклеотидов и детектируется как ложная делеция. Точность определения делеций в итоге не превышает 50%, что, фактически, является случайностью [40].

При работе с различными программными пакетами допустимо снижать требования встроенных дополнительных параметров при определении полиморфизмов. Сравнение различных алгоритмов на реальных и виртуальных данных было проведено в рамках данной работы.

После проведения выравнивания и поиска полиморфизмов имеется возможность получить результат секвенирования по аналогии с результатом чип-гибридизации - в виде текстового файла, который содержит в каждой строке хромосому, позицию и обнаруженное геномное состояние. Чаще всего такой файл

для всех полученных позиций не формируют ввиду его большой величины даже для простых вариантов секвенирования. Например, секвенирование только экзонов всех известных генов человека (известное как экзонное секвенирование) получает информацию о 60 млн. 500 тыс. нуклеотидах, а файл, который содержит в себе записи обо всех прочтенных нуклеотидах, имеет объем порядка 5 гигабайт, что делает его очень неудобным для дальнейшей обработки.

В рамках проекта bioplanet.com существует консорциум «Genome In A Bottle» (GIAB) [49], который предоставляет неограниченному кругу пользователей возможность скачать результаты секвенирования и чип-гибридизации генома одного из участников проекта 1000 геномов (GCat59) на всех доступных технологиях. Проект существует в качестве контроля для любого из разработчиков программного обеспечения, которые оптимизируют свои алгоритмы выравнивания и поиска полиморфизмов. Кроме того, участники консорциума агрегируют полученные результаты в единый VCF-файл со списком всех обнаруженных геномных состояний, достоверность которого превышает 99,99999%. В указанный файл вносятся записи о найденном полиморфизме только в том случае, если он был обнаружен на не менее чем 2х технологиях в не менее чем 3х технологических повторах для каждой. К VCF-файлу также прилагается BED-файл, указывающий на те геномные отрезки, которые были проверены при формировании VCF-файла по указанным выше правилам. Необходимость создания двух отдельных файлов обусловлена тем, что общепринятым стандартом передачи информации о мутациях и полиморфизмах является VCF-файл, в котором содержатся отличия от референсного генома (например, hg19 или hg38), тогда как информация о тех отрезках, на которых находятся нуклеотиды, совпадающие с референсом, теряется. Именно для предотвращения ошибок такие отрезки хранятся в отдельном текстовом файле [50]. Описанный подход является «золотым стандартом» сравнения для различных протоколов обработки данных секвенирования. Все материалы проекта являются открытыми, а декларация о свободном распространении и использовании размещена на сайте консорциума [51].

Для формирования VCF-файла после ионного полупроводникового секвенирования существует проприетарное решение - Torrent Suite. Выбор полиморфизмов в данном программном пакете происходит с использованием собственного алгоритма и собственного дополнительного пакета VariantCaller. Существенным минусом данного пакета, как и многих других аналогичных, является отсутствие возможности формирования VCF-списка по заданным геномным состояниям вне зависимости от соответствия референсу. Такая информация часто требуется для медицинского генотипирования - в типовом VCF-листе присутствуют только те состояния, которые были определены качественно и отличаются от референса, остальные же могут быть либо качественные, но совпадающие с референсом (гомозиготы), либо некачественно прочитанными. Также большим неудобством данного пакета является отсутствие возможности изменять параметры фильтрации по качеству и по покрытию. Согласно информации от сотрудников ThermoFisher (Life Technologies), данные возможности реализованы в платном облачном продукте Ion Reporter.

В современных версиях Torrent Suite в качестве первичных данных наравне с FASTQ предоставляется и BAM-файл, который формируется в процессе секвенирования, путем выравнивания прочитанных последовательностей на геном человека сборки hg19 (GRCh_37) с использованием алгоритма выравнивания Барроуза — Уилера (Burrows-Wheeler Aligner). Преимуществом выравнивания прочтений в Torrent Suite является учет технических особенностей конкретного запуска: схему отрезков используемого набора для обогащения ДНК и, как следствие, верное выравнивание особо коротких ридов (порядка 50 нуклеотидов). Но указанное преимущество незначительно по сравнению с отсутствием возможности изменения параметров фильтрации по качеству в результате работы плагина, что ведет к сильному уменьшению числа обнаруженных вариантов в итоговом VCF-файле.

Получив, в результате секвенирования, известные генотипы для каждого образца, возможно разделить дальнейшее исследование на 2 ветви: групповое ассоциативное сравнение и индивидуальное исследование генома пациента.

Групповое сравнение предполагает поиск одинаковых либо различающихся полиморфизмов внутри одной или нескольких групп, тогда как индивидуальное - поиск известных или вновь обнаруженных индивидуальных геномных особенностей в открытых базах данных или по известным биохимическим алгоритмам.

1.3.2.1 Современные технологии чип-гибридизации

Для работы с данными, получаемыми с использованием ДНК-чипов, по сравнению с NGS-секвенированием требуется меньше вычислительных ресурсов. Обусловлено это высокой стандартизацией метода, что позволило разработать достаточно удобные пакеты программных решений, которые широко применяются во многих лабораториях мира для решения стандартных задач. В то же время, различные форматы ДНК чипов, особенности биологических задач, которые ставятся для решения – диктуют необходимость внимательного первичного анализа генетических данных, исключающих досадные ошибки.

ДНК-микрочип (DNA microarray) - технология, используемая в молекулярной биологии и медицине. Современный ДНК-микрочип состоит из тысяч ДНК-зондов, сгруппированных в виде микроскопических точек и закреплённых на твёрдой подложке. В основе всех технологий генотипирования на биочипах находится метод гибридизации фрагментов однонитевой ДНК с зондами.

На сегодняшний день можно выделить следующие основные разновидности ДНК-чипов по технологии изготовления: печатные (printed), фотолитографические (*in situ* synthesized) и высокоплотного кварцевого нанесения (high-density bead arrays).

В печатной технологии зонды наносятся путем струйной печати с использованием высокоточных пикомиллиметровых сопел. Данная технология имеет минимальные ограничения на структуру зондов и позволяет изготавливать ДНК-чипы практически для любого организма и в любом требуемом количестве,

но она не подходит для полногеномных исследований в связи с неудовлетворительной максимальной плотностью нанесения зондов на чип.

Фотолитографическая технология заключается в активации лазером через литографическую матрицу специальных молекул-линкеров, нанесенных на подложку чипа, а затем промывку чипа раствором с присоединяемым зондом. Плотность таких чипов значительно выше, но на каждый тестируемый полиморфизм требуется минимум по 4 отдельных зонда (по одному на каждый из нуклеотидов), а значит, что на весь чип требуется изготовить более 1 млн. матриц, что составляет значительный вклад в стоимость и делает чипы экономически невыгодными для проведения исследований с малым числом тестируемых образцов.

Чипы, изготовленные по технологии высокоплотного кварцевого нанесения, являются наиболее подходящими для ассоциативных исследований, поэтому стоит остановиться на технологии их изготовления подробнее. В процессе изготовления чипа формируются кремниевые бусины с закрепленными на них олигонуклеотидами. Нанесение производится электрохимическим методом в случайном порядке, поэтому структура каждого отдельного чипа различна при одинаковом наборе зондов. Для повышения точности определения нуклеотидов, на каждый чип наносится до 30 бусин, несущих одинаковые олигонуклеотидные зонды, которые попадают в различные места стекла-подложки. При проведении анализа, на первом шаге производится ПЦР с использованием биотинилированного праймера, а затем промывка продуктов реакции для освобождения от остатков неиспользуемой ДНК. Полученный раствор наносится на чип и производится гибридизация фрагментов ДНК к комплементарным им зондам, а затем присоединение флуоресцентно-меченых дидезоксирибонуклеотидов в следующей позиции, после окончания зонда (Рис.5).

На последнем шаге чип сканируется и по яркости и цвету свечения определяют статус полиморфизма. На одном таком чипе возможно измерить всего порядка 20 миллионов полиморфизмов, которые могут быть разбиты на секторы, предназначенные для нанесения различных образцов [52]. Программное

обеспечение приборного комплекса устроено так, что в результате работы сканера формируется текстовый файл, содержащий для каждой прочитанной позиции нуклеотид и достоверность его определения в виде величины «геномного качества» (GC-score, Gene Call Score).

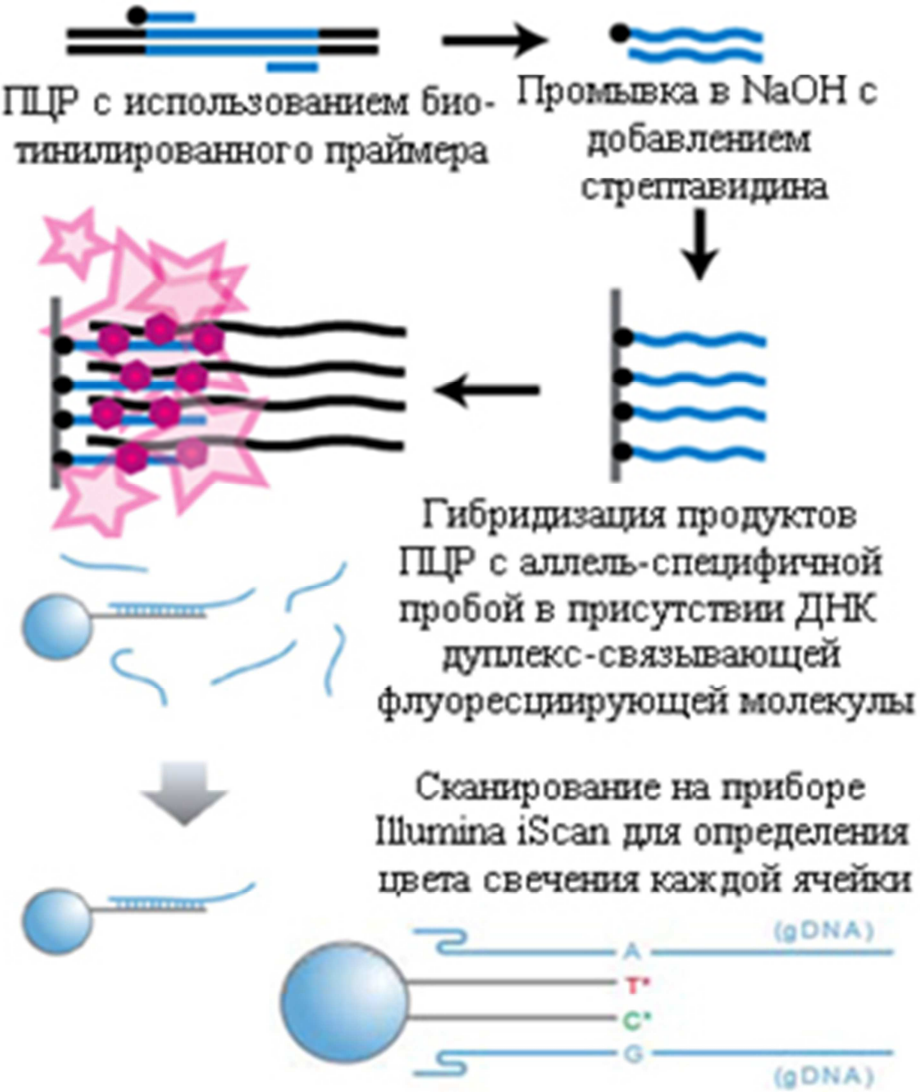

Рисунок 5. Протокол использования ДНК-чипа высокоплотного кварцевого нанесения [52].

1.3.2.2 Биоинформатический анализ первичных результатов высокоплотной чип-гибридизации

Технология чип-гибридизации применяется в биотехнологии уже более 20 лет, и вычислительная сложность обработки результатов достаточно низкая (по сравнению с секвенированием), поэтому крупнейшим корпорациям уже удалось

разработать программное обеспечение, которое могло бы достаточно точно определять однонуклеотидные полиморфизмы (SNP) по результатам свечения ячеек чипа. Немного хуже, но тоже достаточно успешно реализуется определение копийности фрагментов генов (CNV, copy number variation). У каждой из компаний-производителей сканеров для чипов существует свое программное обеспечение, которое, в целом, удовлетворяет всем поставленным задачам на необходимом уровне. У компании Illumina для определения генотипов и копийности используется пакет Genome Studio.

При работе с чипами высокоплотного кварцевого нанесения (технология BeadChips от Illumina), как было упомянуто выше, статус полиморфизма определяется по интенсивности свечения флуорофора, ассоциированного с аллельным вариантом. При этом следует учитывать, что уровни интенсивности нормализуются как по всем бусинам чипа (порядка 30 повторов), так и по всем образцам чипа или запуска (не менее 12 образцов).

Обычно полученные интенсивности на графике группируются в 3 кластера: 2 гомозиготы и гетерозигота (Рис.6). Для каждого полиморфизма вычисляют вероятность ошибки при отклонении нулевой гипотезы (P-значение, P-value) классификации в ожидаемый кластер (альтернативная гипотеза) против попадания в какой-либо другой (нулевая гипотеза). Если полученное значение P-value > 0,05, то считается, что точка не попадает в указанный кластер и данный полиморфизм у данного образца считается определившимся некорректно. Для каждой прошедшего первичный контроль полиморфизма вычисляется внутренний параметр GC-score. Эта величина является непосредственной характеристикой яркости флуоресценции зонда и измеряется как относительная интенсивность в сравнении со специальной технической бусиной, яркость которой всегда максимальная, так как все зонды этой бусины флуоресцируют. Поэтому, GC-score может принимать значения от 0 (свечение отсутствует) до 1 (все зонды всех бусин одного полиморфизма светятся максимально). Производителем установлен минимальный уровень флуоресценции в 0,15 единиц GC-score для того, чтобы учитывать наличие соответствующего сигнала.

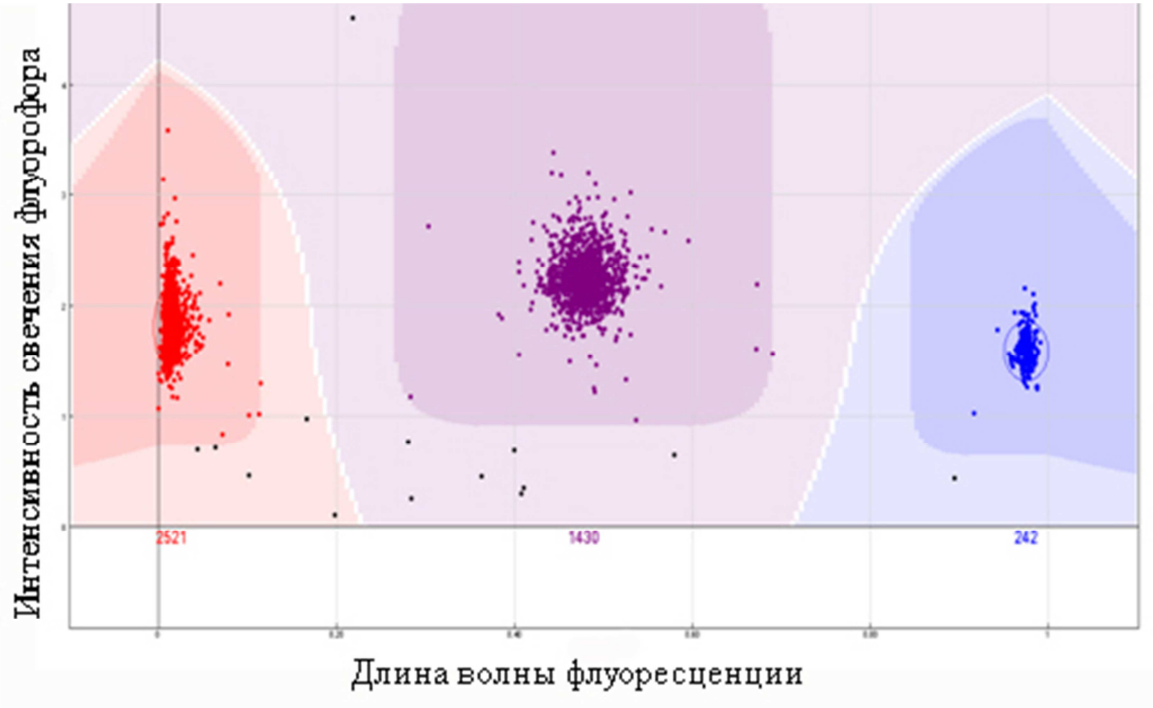

Рисунок 6. Пример распределения по длинам волн средних интенсивностей флуоресценции зондов для одного полиморфизма и 48 образцов. Каждая точка на графике – единичный образец, каждый кластер сформирован из образцов, обладающих одинаковым генотипом (слева направо: гомозигота, гетерозигота, гомозигота).

Исходя из этого, могут существовать точки с низким P-value и высоким GC-score в случае каких-либо нестандартных соотношений по числу копий хромосом (например, трисомия, 2 + 1 гетерозигота), а также с высоким P-value и низким GC-score в случае каких-либо технических проблем с гибридизацией или неполного соответствия ДНК образца последовательности зонда [53].

Если производить фильтрацию качества только по P-value без учета GC-score, то для стандартных чипов (Illumina 1M_Omni_Quad, например) примерно 1% точек будут промаркированы как ошибочные для каждого образца, а точность определения остальных нуклеотидов составит более 99% [54]. Для задач, связанных со множественным сравнением, например, при полногеномных ассоциативных исследованиях, настройки по умолчанию (учитывающие только P-

value), являются подходящими, так как ошибки такого рода случайные, поэтому они не влияют на общее распределение при достаточно большом наборе образцов.

Но, если необходимо максимально точно охарактеризовать генотип одного конкретного образца, например, при клиническом генетическом анализе, то такие параметры являются неприемлемыми. Для этого допустимо в ущерб количеству идентифицированных полиморфизмов, изменить параметры биоинформатической обработки для получения максимально точных данных. Согласно проведенным исследованиям, порог GC-score > 0,5 удаляет порядка 7% значений, при том, что точность определения повышается до 99,75% (1 ошибка на 400), а критерий GC-score > 0,7 фильтрует 30% точек с повышением точности до 99,88% [55]. Указанные значения GC-score и точности существенно зависят от условий эксперимента и концентрации образцов ДНК, но общая корреляция остается неизменной.

Другой проблемой, более актуальной при обработке данных генотипирования на ДНК-чипах, чем при секвенировании, является ошибочная запись генотипа полиморфизма по комплементарной нити ДНК. Эта проблема более типична для ДНК-чипов, так как для разработчиков в зависимости от нуклеотидного контекста вокруг полиморфизма иногда удобнее подбирать зонд, комплементарный обратной цепи ДНК.

Научным сообществом принято, что та цепь, которая со стороны начала нумерации нуклеотидов в хромосомах имеет 5’ конец называется TOP (верхняя), а комплементарная ей – BOTTOM (нижняя) (Рис.7). Но, к сожалению, не всегда удобно указывать значения нуклеотидов по схеме TOP/BOTTOM, поэтому были разработаны альтернативные схемы обозначения. Схема FORWARD подразумевает, что нуклеотид будет записан согласно расположению гена, последовательность которого была отсеквенирована. Таким образом, при схеме FORWARD требуется указать ген и интересующее нас положение, а эту информацию, зная на какой из цепей TOP или BOTTOM расположен ген, можно легко перевести в более простые форматы. Кроме FORWARD, существует еще и DESIGNER схема, которой пользуются различные производители оборудования для генотипирования, чтобы

хранить необработанные геномные данные. При этой схеме информация о соответствии каждой позиции TOP/BOTTOM цепям хранится в специальном файле, который является отдельным для каждого из производителей и именно из-за этого факта схема DESIGNER в открытых базах данных не используется [56].

Рисунок 7. TOP, BOTTOM, FORWARD схемы на примере двух различных полиморфизмов, находящихся внутри экзонов различных генов.

При всех удобствах каждой из схем, в большинстве баз данных используется для хранения информации схемы TOP и FORWARD, а в публикациях авторы стараются указывать, какой схемой они пользовались.

К сожалению, до появления 38-ой версии сборки человеческого генома (Human Genome 38, hg38) не существовало единого стандарта на запись нуклеотидных состояний по всему миру, поэтому различные компании-производители для каждого из форматов (TOP, FORWARD) использовали геномную последовательность, которая имела координатную структуру генома hg19 (или hg18), но информация об "основной" цепи различалась. Поэтому очень частой является ситуация не соответствия результатов чип-гибридизации и секвенирования именно в части направления цепей [56].

Несмотря на то, что в аннотационном файле к ДНК-чипу указана правильная ориентация для всех зондов, зачастую ошибочное позиционирование не определяется в ходе предварительной подготовки данных и вносит существенную ошибку в последующих этапах анализа. Особенно актуален этот вопрос при

проведении групповых исследований данных чип-гибридизации, полученных на различных технологических платформах с использованием ДНК-чипов разных производителей [56].

Кроме подбора критериев правильной идентификации положения полиморфизма на цепи ДНК, существуют и технические погрешности, связанные с физическими особенностями самого метода анализа. Так, в ходе реакции достраивания зонда на нуклеотидное звено комплементарное точке полиморфизма существует определенная вероятность присоединения гуанина (G) в пару к другому гуанину (G) на комплементарной цепи. Энергия такого объединения всего лишь на 10 - 15% выше энергии объединения нуклеотидов в пару GC [57]. Разработка зондов для детекции и последующий анализ таких полиморфизмов является значительно более трудоемким и, при наличии альтернативы, производители ДНК-чипов выбирают для разработки полиморфизмы другого типа (Табл.3). Но, согласно базе данных UCSC (hg19_snp138) в геноме человека такие полиморфизмы составляют порядка 2% от всех известных полиморфизмов (порядка 3000000 записей из 150482731 в таблице hg38_snp146), из которых около 1000 носят индекс клинически-значимого патогенного варианта (индекс pathogenic в БД ClinVar, dbSNP). Этот факт также несет большое значение как при работе с чипами, изготовленными на заказ, так и при работе с низкой концентрацией ДНК (ниже концентрация – слабее различие между интенсивностями флуоресценции).

При изготовлении чипов под заказ, особенно в целях диагностики клинически значимых полиморфизмов, избежать включения полиморфизмов C/G в чип не удается, потому как среди всех известных ассоциированных с заболеваниями однонуклеотидных полиморфизмов таких не менее 10% [58]. Но указанная статистика не означает, что по таким полиморфизмам, включенным на чип, нельзя будет получить достоверный результат. Достаточно при анализе таких полиморфизмов (чипов) придерживаться тщательного контроля качества.

Таблица 3. Количество зондов различных типов, размещенных на чипе Illumina 1M Omni Quad.

Чип	1M Omni Quad v1.2	CVD 55 v1.0	Omni Express 12 v1.1	Omni Express 24 v1.1	Core Exome Plus v1.2
Общее число зондов	1134515	49094	719665	713014	546929
Число детектируемых трансверсий	1006409	42498	716888	710343	501401
Число детектируемых транзиций	33155	6596	2671	2570	30792
Число детектируемых вставок/делеций	449	0	0	0	12222
Число зондов оценки числа копий	49	0	0	0	0

Но, к сожалению, это не все возможные трудности, которые возникают при чип-гибридизации. Для каждой партии Bead-чипов изготавливается набор ДНК-зондов на бусинах, с помощью которых возможно изготовить более 100000 одинаковых чипов. Поэтому, при печати стандартных чипов, изготавливаются тестовые образцы, а затем производится проверка их качества на производстве. В случае несоответствия заявленным параметрам зонд-бусина бракуется и изготавливается другой [59]. Стоимость изготовления набора бусин высока, но при значительных масштабах производства она незначительно влияет на итоговую цену [60]. Именно поэтому, при работе с не каталожными версиями чипов следует проводить дополнительный контроль качества, в особенности для бусин, изготовленных под заказ.

Исходя из всего вышесказанного, при работе с чипами допустимо использовать стандартное программное обеспечение, но всегда следует проводить предварительную обработку исходных данных по протоколам, соответствующим целям исследования, а также осуществлять дополнительный контроль качества данных, полученных с чипов, изготовленных на заказ.

1.4 Биоинформатическая основа ассоциативных исследований

Полногеномный поиск ассоциаций (GWAS, Genome-Wide Association Studies) – направление биомедицинских исследований, связанных с изучением ассоциаций между геномными вариантами и фенотипическими признаками. Несколько сотен лет врачами наблюдалось наследование различных заболеваний внутри отдельных семей, и, именно для поиска генетических маркеров такого наследования, был разработан указанный метод.

В США на базе NCBI (National Center for Biotechnology Information) ведется каталог опубликованных ассоциативных исследований GWAS-catalog. Исследования, опубликованные в нем, основаны на сравнении геномных данных группы людей, имеющих общий признак (например, болезнь Альцгеймера) с геномными данными испытуемых, у которых этот признак отсутствует (контрольная группа) [3]. Ограничением минимального числа пациентов в группе является итоговая ожидаемая величина погрешности. На основании различающихся состояний одного и того же полиморфизма, одна из аллелей которого в исследуемых образцах встречается существенно чаще, чем в контроле, утверждается, что минорная аллель такого SNP связана с проявлением склонности к тому же признаку, который проявляется у исследуемых лиц. Впервые такое исследование было применено в генетике в 2005 году группой европейских ученых для поиска зависимости генома и проявлений возрастного помутнения сетчатки глаза (возрастная макулодистрофия) [61].

В количественном представлении выглядит это следующим образом: берется группа из **n** исследуемых лиц с общим признаком и группа из **m** лиц контроля. Конечно, значения чисел **m** и **n** должны быть одного порядка, а также такими, чтобы **1/m** и **1/n** были меньшими, чем частота минорного аллеля самого редкого из исследуемых полиморфизмов. Затем тем или иным методом генотипирования определяется значительное число состояний различных полиморфизмов для образцов ДНК из обеих групп, и вычисляется частота каждого генотипа в каждом из полиморфизмов для обеих групп. Число тестируемых полиморфизмов может варьироваться от нескольких сотен до

полного генома, в зависимости от цели исследования. Далее, в случае предположения того, что все исследуемые полиморфизмы имеют только два состояния (таких полиморфизмов более 95% от общего числа известных в базе данных dbSNP), тогда для дальнейшего анализа целесообразно использовать величину, называющуюся «отношение шансов» (англ. odds ratio, OR). Ее вычисляют по формуле:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 \times q_2}{p_2 \times q_1}$$

где p_1 – частота редкой аллели у исследуемых образцов, а p_2 – у контроля [62].

В случае существенного различия между частотами у исследуемых образцов и у контроля OR будет существенно отличаться от 1. Дополнительно к OR всегда вычисляют P-value методом критерия согласия Пирсона (критерий согласия хи-квадрат, критерий χ^2). В качестве стандартного начального уровня критерия значимости чаще всего предполагается P-value < 0,05 (аналог уровня погрешности в 5% от значения), но допустимо изменять уровень значимости в зависимости от биологического смысла анализируемого процесса [63].

При проведении масштабных ассоциативных исследований с множеством потенциальных результатов всегда встает вопрос о поправке к уровню достоверности, связанной с резким повышением вероятности ошибочно допустить редкое предположение при увеличении числа статистических тестов. Множество дискуссий ведется про возможности применения того или иного метода коррекции, но самыми популярными на сегодняшний день можно считать два из них: поправка Бонферрони-Холма (современная версия поправки Бонферрони) и критерий Бенджамини-Хохберга (более популярное название – FDR).

При использовании поправки Бонферрони для m статистических предположений следует считать, что результат является ложным в том случае,

если Pval(i) > Pval.norm. / m, где Pval(i) – оценка значимости i-ой гипотезы (при условии выстраивания гипотез в порядке возрастания уровней значимости), а Pval.norm. - некоторая стартовая базовая оценка достоверности. Холм привлек метод последовательных итераций для решения проблем падения мощности теста при увеличении числа гипотез. Так, на каждом шаге фильтрации гипотез предполагается отсев не прошедших, а затем уменьшение общего числа m на количество отсеянных гипотез n. Тем самым критерий становится менее жестким, и большее число гипотез начинают удовлетворять ему. Количество итераций в методе Бонферрони-Холма может быть любым, и они производятся до тех пор, пока количество прошедших фильтр гипотез не будет изменяться [64].

В качестве альтернативного подхода допустимо применение метода Бенджамини-Хохберга, отличающаяся от метода Бонферрони-Холма только критерием значимости. Предполагается, что результат является ложным в том случае, если Pval(i) > i * Pval.norm. / m. В зависимости от количества гипотез и распределения уровней значимости, допустимо применение любого из описанных методов [65].

Сейчас существует множество программ для ЭВМ, которые позволяют вычислять эти величины достаточно быстро. Согласно данным Broad Institute, являющимся одним из ведущих научно-исследовательских институтов в сфере биоинформатики, большинство научных групп используют для проведения статистических вычислений программу PLINK [66], GenABEL [67], KGG [68], web-сервис SNAP [69] или различные библиотеки в среде программирования вычислений R [70]. Каждый из пакетов реализует фактически один и тот же математический подход, но оптимизирован для различных входных форматов и данных различного объема. Для значительных объемов данных и при отсутствии нестандартных вычислительных задача чаще всего применяется пакет PLINK [71].

Классический подход метода отношения шансов был разработан для двух групп и бинарного состояния анализируемого признака, тогда как для тройных и непрерывных состояний существуют различные адаптации описанного метода.

При проведении GWAS-исследований требуется следить за сопутствующими общими признаками, как в группе исследуемых пациентов, так и к группе контроля. Дело в том, что существенное количество полиморфизмов связано с этническим происхождением каждого из испытуемых. Поэтому, при сравнении 100 европейцев, больных диабетом, и 100 здоровых африканцев с большей вероятностью будут обнаружены различия, связанные с национальным происхождением, нежели чем с самим заболеванием. В случае если удастся выделить все такие SNP, то для болезней, не связанных с происхождением (одинаково встречающихся во всех популяциях), будет сильно упрощена процедура анализа [63] [72].

В случае поиска корреляций с непрерывными биохимическими показателями полный набор тестируемых образцов разбивают на небольшие группы, внутри которых образцы имеют близкое значение исследуемого биохимического показателя, а затем проводят ассоциативное сравнение двух групп, значение показателя в которых максимально и минимально. На следующем шаге производят проверку корреляции обнаруженного набора полиморфизмов с биохимическим показателем с использованием промежуточных групп. Данный прием является стандартным методом машинного обучения: на части данных производится обучение, а на другой части - валидация.

Результаты GWAS-исследований, конечно, являются одним из первичных источников информации об ассоциации, но, так как большинство социально значимых заболеваний вызываются совокупностью различных генетических и негенетических факторов, значимость единичного полиморфизма в них мала. Кроме того, биологически полученные полиморфизмы могут быть не значимы, а находиться в группе сцепления с другими полиморфизмами, которые не были приняты в исследование, но являются значимыми. Поэтому, для проверки биологической значимости полученных вариантов необходимы исследования на клеточных линиях или модельных организмах. Но, с другой стороны, результаты ассоциативных исследований возможно использовать в качестве маркеров риска для дальнейшего изучения организма пациента в обнаруженных направлениях.

1.5 Биоинформатические подходы к оценке индивидуального генетического риска

При проведении групповых исследований результатом является утверждение об ассоциации генотипа какому-либо фенотипу. Как уже было сказано, такие выводы имеет смысл применять в медицинских целях при определении индивидуальных особенностей конкретного пациента. Для множества моногенных заболеваний существует значительное число биологически доказанных ассоциаций, которые являются основаниями для постановки формального диагноза. С другой стороны, подавляющее большинство социально значимых заболеваний являются мультифакторными, и для них уже обнаружено множество статистически достоверных маркеров, для которых нет ни прямого, ни комплексного биологического обоснования описанной ассоциации. Для решения обозначенной проблемы, возможно выделить несколько основных биоинформатических направлений, на которых стоит остановиться подробнее.

1.5.1 Каталогизация клинически значимых ассоциаций

Существует множество проектов, предоставляющих возможность исследователям размещать полученную информацию в базе данных. К таким относятся, например, участники проекта Leiden Open Variation Database [73]. LOVD - это программный модуль и название консорциума, который занимается анализом биологических связей между генотипом и фенотипом на малом числе образцов. В большинстве своем это моногенные заболевания и синдромы, а также моногенные варианты мультифакторных заболеваний. Кроме того, существует также проект SNPedia [74], являющейся открытой энциклопедией геномных состояний и их проявлений. Основой проекта является интерфейс WikiPedia - электронная энциклопедия, в которой каждую страницу может написать и отредактировать любой из пользователей.

Кроме баз данных первого уровня (заполняемых непосредственно исследователями), существуют и активно разрабатываются также базы данных второго уровня, которые создаются биоинформатическими научными группами

путем поиска интересной им информации в публикациях и последующего размещения ее на страницах своих проектов. К таким проектам относится упомянутый выше GWAS-catalog [15], а также проекты Online Mendelian Inheritance in Man (OMIM) – публично доступная, курируемая база данных по заболеваниям, наследуемым согласно законам Менделя [75], Human Genome Mutation Database (HGMD) - платная курируемая база данных по мутациям генома человека и их клиническому обоснованию [76], PharmGKB - открытая база данных по ассоциациям генотипов и успешности лечения различными препаратами [77].

Проект Human Genome Navigator [78] содержит много различных данных о геноме человека, а также информацию о рисках заболеваний, взятых из GWAS-catalog [15]. Проект отличает более полная техническая информация о геноме. По сути, он является интернет-реализацией базы данных genome калифорнийского университета [79], но доступ пользователей, не обладающих навыками работы с базами данных MySQL, в эту базу затруднен.

При индивидуальном анализе геномных состояний вручную обычно обращаются к GWAS-catalog [15], LOVD [73], OMIM [75], SNPedia [74], PharmGKB [77], COSMIC (Catalogue of Somatic Mutations in Cancer) [80], ClinVar [81], UCSC [77].

К сожалению, на данный момент не существует какого-либо простого единого интерфейса, который по генотипу человека формировал бы отчет по заболеваниям и нозологиям, которые могут проявиться у данного индивида, хотя бы на уровне ссылок по опубликованным ассоциативным и биологическим проектам.

Одной из самых крупных попыток реализовать общий интерфейс для удобной аннотации персональных геномных данных с приемлемым результатом для пациентов - не являющихся специалистами, является программа Promethease проекта SNPedia.com. Она позволяет по файлу с генотипами в различных позициях SNP в обозначениях стандарта rs отображать все описания из проекта

SNPedia, которые верны для исследуемого человека. Сейчас в базе данных проекта описано всего 35605 полиморфизмов [74].

Также существует специализированный биоинформатический проект GEMINI [82], основной идеей которого является создание удобного интерфейса обращения к известным открытым базам данных, а также помощь при сборке и аннотации геномов. Данный программный пакет является консольным и созданным для среды Linux, что автоматически ограничивает его использование кругом разработчиков и биоинформатиков.

1.5.2 Оценка рискового статуса потенциально патогенных генетических вариаций

В отдельную группу задач, которые необходимо решить при оценке индивидуального риска, относится оценка рискового статуса обнаруженных новых мутаций. Такая оценка, в большинстве своем, проводится с использованием информации о фундаментальных биологических взаимодействиях (центральная догма молекулярной биологии, метаболические пути и прочее) и реализуется в существующих различных алгоритмах: AVSIFT [83], LJB_PhyloP, LJB_PolyPhen2 [84]. Все указанные алгоритмы основаны на различных схемах оценки патогенности различных аминокислотных замен и стоп-кодонов. Данные алгоритмы реализованы в различных программных продуктах: VEP [85], AnnoVar [86], GEMINI [82], SnpEFF [87]. Все программы на вход принимают VCF-файл, содержащий только записи о нуклеотидах, отличающихся от референса, а на выходе формируют также VCF-файл, в который добавляется информация о значимости полиморфизма по большинству описанных алгоритмов, а также указывается информация о наличии у него rs индекса в дополнительных столбцах. Существенные различия между пакетами заключаются в алгоритмах оценки биологической значимости полиморфизма и влиянии его на экспрессию белков, однако для большинства мутаций в генах, однозначно связанных с заболеваниями, все программы предсказывают рисковые состояния примерно одинаково [88].

1.5.3 Биоинформатические подходы к оценке совокупного риска

Самой малоизученной и сложной задачей является разработка алгоритмов по учету совокупного риска развития заболевания при выявлении комбинации из множества генетических полиморфизмов, ассоциированных в той или иной степени с данным заболеванием. Зачастую, при обнаружении у пациента двух различных полиморфизмов, каждый из которых описан как патогенный по отношению к некоторому общему заболеванию, определить итоговый рисковый статус по этому фенотипу не представляется возможным. На первый взгляд кажется, что риск должен возрастать, но в реальности такая ситуация реализуется очень редко. Дело в том, что очень часто такие полиморфизмы находятся на генах, наследование которых сцепленное и наличие двух, трех, четырех и более полиморфизмов с рисковым статусом никак не отражается на общем риске развития исследуемого заболевания. Кроме сцепленного наследования возможно также взаимное участие двух описанных генов в метаболической цепи последовательно, тогда риск будет возрастать, но незначительно по отношению к рисковому статусу каждого из полиморфизмов. Иногда встречаются ситуации, в которых совокупный риск оказывается ниже потому, что оба гена находятся в независимых цепях метаболического обмена.

Самым простым алгоритмом учета совокупного риска до сих пор является построение дерева решений. Каждый из элементов дерева является критерием для перехода к следующему элементу и так до тех пор, пока не будут пройдены все полиморфизмы. Деревья решений аналогичны по своей сути условиям алгебры Буля (и, или, не), а их структура напрямую вытекает из результатов эксперимента, в котором изучались все полиморфизмы или мутации модели одновременно. [89]

Другими методами, позволяющими оценить рисковый статус пациента для мультифакторных заболеваний в общем случае, являются различные методы кластеризации (метод главных компонент, метод карт Кохонена и прочее). Чаще всего их применяют с использованием результатов ассоциативного исследования в качестве обучающей выборки, а затем предполагают, к какой группе относится исследуемый пациент. Основная проблема этого метода заключается как в

отсутствии количественной оценки риска, так и в невозможности его применения в случае отсутствии информации о том или ином полиморфизме. [90] [91]

Основываясь на всем вышеописанном, ведущими коллективами биоинформатиков были разработаны две основных простых математических алгоритма для построения релевантной модели оценки совокупного риска.

Первая из них - мультипликативная была разработана компанией Navigenics и опубликована как научно-исследовательская работа. [92]Модель хорошо работает в следующих приближениях:

1) Рисковый статус каждого из полиморфизмов не превышает 5 (в единицах OR, отношения шансов).

2) Полиморфизмы не являются зависимо наследуемыми. Находятся в разных генах, на разных хромосомах, достаточно далеко друг от друга, нет исследований показывающих взаимосвязь в наследовании.

Пусть рисковый статус аллели для каждого из полиморфизмов задан отношением шансов - r (OR), а также известны частоты аллелей в популяции, к которой относится индивид, для каждого из генотипов Caa, Cab, Cbb (aa – рисковый генотип), тогда средний риск заболевания равен $M = r^2xCaa+rxCab+1xCbb$, а относительные риски при соответствующих генотипах $Raa = r^2/M$, $Rab = r/M$, $Rbb = 1/M$. В результате преобразования и при условии соответствия полиморфизма требованиям, указанным выше, допустимо вычислять совокупный относительный риск Rобщ как произведение относительных рисков по каждому из генотипов. Полученный совокупный риск при умножении на частоту заболеваемости в популяции показывает вероятность заболеть данным заболеванием конкретному человеку. [92]

У указанной модели есть и известные трудности. Во-первых, она никак не учитывает совокупное наследование, которое встречается очень часто (чаще чем в 50% случаев). Во-вторых, модель резко увеличивает максимальный относительный риск при увеличении числа низкорисковых полиморфизмов. Эта особенность связана с тем, что при числе полиморфизмов более 30 обязательно присутствует пара, которая имеет неполную сцепленность при наследовании,

тогда реальный риск внутри такой пары должен быть ниже, чем произведение относительных рисков.

В противоположность мультипликативной модели компанией 23andMe была разработана аддитивная или регрессионная модель оценки риска. Такая модель работает хорошо только при следующих условиях:

1) Число полиморфизмов в модели строго фиксировано.

2) Рисковый статус всех полиморфизмов приблизительно одинаковый.

Для каждого полиморфизма и каждого генотипа присваивается некоторый численный рисковый статус. Чаще всего в качестве рискового статуса используется балловая система с шагом 1 за каждую рисковую аллель. На следующем шаге вычисляется минимальная, максимальная и средняя по популяции сумма для всех аллелей, входящих в модель. В данном случае, среднее по популяции является тем же самым, что и величина M в мультипликативной модели L_{norm} = 2xCaa+1xCab+0xCbb, в случае если аллель **A** является рисковой с баллом 1. На графике показано, как соответствует относительный риск и количество рисковых баллов в случае, если в модели учитывается 63 полиморфизма. Соответственно, если у пациента суммарный рисковый статус ниже L_{norm}, то он обладает протективным генотипом, а если выше - то рисковым [93].

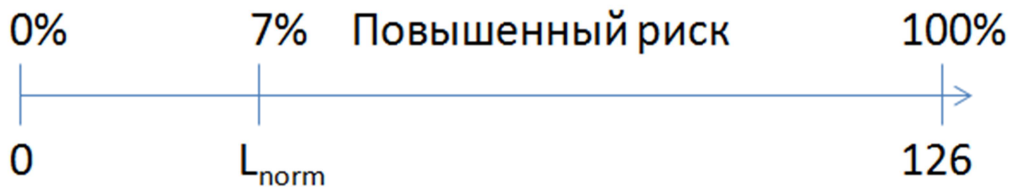

Рисунок 8. Схема распределения состояний рискового статуса в зависимости от общего суммарного рискового балла.

Основные трудности при использовании аддитивной модели возникают в случае, если не известен набор полиморфизмов, который будет получен у пациента. Дело в том, что каждый отсутствующий полиморфизм приводит к пересчету модели, а также к оценке неучтенных рисков. Кроме того, не очевидно,

как учитывать в такой модели полиморфизмы с резко различающимся рисковым статусом, например, $OR_1 = 5$ и $OR_2 = 1,2$.

Большинство исследователей сходятся во мнении, что для каждого заболевания требуется комбинировать данные модели для разных групп полиморфизмов, а затем учитывать результаты оценки рисков по каждой из модели с использованием какой-либо специально разработанной математической формулы [94] [95].

Отдельного обсуждения требуют технологии оценки качества диагностического теста. Чаще всего для расчета используют характеристическую кривую диагностического теста (ROC, Receiver Operating Characteristic), а также площадь под ней (AUC, Area Under Curve) [96]. Варьируя число полиморфизмов, используемых в исследовании, вычисляется риск для пациентов с известным статусом, а затем оценивается точность соответствия вычисленного риска реальному значению [97]. В случае отсутствия возможности наблюдать и генотипировать существенное число пациентов допустимо проводить моделирование искусственных групп людей, распределение которых в точности бы соответствовало описанному в статьях с ассоциативными исследованиями, а затем проводить оценку точности на полученных выборках [98].

По всему миру существует множество коммерческих компаний, которые предлагают собственные услуги по исследованию персонального генома, а также осуществляют попытки разработки моделей совокупного учета риска для тех или иных заболеваний: 23andMe, Navigenics, AriaDNA, Pathway Genomics и другие (93). Результаты предсказанных рисков у различных компаний могут значительно отличаться, но для большинства хорошо изученных заболеваний генетический тест, в целом, может быть хорошей основой для направления пациента на дальнейшее обследование [99].

К сожалению, спорные результаты оценки рисков у различных компаний и при использовании различных методов, показывают огромный объем возможной работы, но регулярно воспроизводящиеся удачные модели для некоторых заболеваний означают успешность приложенных трудовых затрат.

1.6 Обоснование поставленных задач

Исходя из всего вышеописанного, можно сформулировать следующий перечень задач, решение которых в рамках настоящей работы требует применения биоинформатических методов.

Во-первых, это разработка оптимизированного алгоритма обработки и анализа генетических данных, получаемых с использованием технологий ионного полупроводникового секвенирования (ThermoFisher, inc) и чип-гибридизации (Illumina). Это обусловлено как недостатками существующего программного обеспечения, используемого по умолчанию, так и особенностью используемых методов, как таковых.

Во-вторых – это адаптация данных генотипирования, полученных с использованием разных технологических платформ для проведения полногеномных ассоциативных исследований в рамках текущей работы и непосредственно биоинформатическая обработка и анализ, полученных в них данных.

В-третьих - разработка вспомогательных информационных ресурсов и алгоритмов для оценки индивидуального генетического риска на основе полученных данных высокоплотной чип-гибридизации или секвенирования.

Решение перечисленных биоинформатических задач будет актуальным и необходимым условием для проведения полногеномных ассоциативных исследований в группах профессиональных спортсменов по поиску генетических маркеров, ассоциированных с показателями оксидативного стресса, эритропоэза и структуры мышечных волокон.

Проведение генетических исследований в профессиональном спорте позволит как обнаружить новые маркеры спортивной успешности, провести индивидуальное исследование состояния здоровья каждого спортсмена отдельно, а также использовать полученные результаты в персонифицированной медицине.

Глава 2. Материалы и методы

Проведенное исследование подтверждено этическими комитетами ФГБУ «Федеральный научный центр физической культуры и спорта» (Москва, Россия), Pomeranian Medical University (Szczecin, Польша), Российским Национальным комитетом по биоэтике РАН, ФГБОУ ВПО «Санкт-Петербургский государственный университет».

2.1 Характеристика исследуемых групп

В исследовании приняло участие 17668 человек из различных стран мира, из которых 3543 являлись спортсменами и 14125 человек относились к контрольной группе (Табл.4).

Таблица 4. Этнический состав испытуемых, принимавших участия в исследованиях.

Происхождение	Спортсмены	Контроль
Россия	**1444**	734
Польша	1220	441
Кения	276	83
Австралия	215	258
Испания	170	198
Япония	143	692
Эфиопия	75	198
Illumina iContol DB	0	11521
Всего	**3543**	**14125**

Для выявления ассоциаций генетических маркеров с различными фенотипами исследуемых выборок было применено два различных методических подхода. При первом подходе проводилось сравнение частоты генетических маркеров между двумя выборками: спортсменами различных специализаций и

контрольной группой. Предполагалось, что если частота определенного генетического маркера встречается значимо выше, например, в группе атлетов-спринтеров по сравнению с контрольной группой, тогда данный маркер можно считать благоприятствующим к занятиям бегом на короткие дистанции и проявлению взрывной силы.

Во втором подходе применение данных групп контроля не требовалось, поскольку проводился корреляционный анализ внутри исследуемой выборки спортсменов по двум-трем генотипам конкретного полиморфизма. В данном случае определялись генетические маркеры, ассоциированные с различными биохимическими или спортивно значимыми показателями, например, с уровнем креатинкиназы.

Для выявления ассоциаций генетических маркеров с фенотипами при подборе выборок учитывались следующие критерии: 1) пол (частоту генетического маркера между спортсменами и контрольной группой сравнивали как в совокупности, так и раздельно по половой принадлежности); 2) расовая принадлежность (предполагалось разбиение по трем основным этническим группам); 3) возраст (сравниваемые выборки должны были быть примерно одного возраста, возраст испытуемых составил 20±3 лет); 4) спортивная специализация (спортсмены подразделялись как по отдельным специализациям, так и объединялись в одну подгруппу по типу энергообеспечения соревновательной нагрузки); 5) спортивная квалификация (имеет значение для выявления зависимости частоты генетического маркера от квалификации, которая характеризует спортивную успешность).

2.1.1 Группа профессиональных спортсменов из РФ

В исследовании приняли участие 1444 спортсмена сборной РФ различной специализации и квалификации (393 женщины, 951 мужчина 20±3 лет). Образцы биологического материала собирались у спортсменов на основании добровольного информированного согласия в поликлиниках №1 ФГБУ Федерального Научно-Клинического Центра Физико-Химической Медицины

ФМБА России, ФГБУ Федерального Научно-Клинического Центра Спортивной Медицины ФМБА России и ФГБОУ ВО "Поволжская государственная академия физической культуры, спорта и туризма". Таким образом, в выборке спортсменов присутствуют воспитанники школ олимпийского резерва, представляющие многие регионы России.

На момент забора биологического материала для исследований ДНК, 23 спортсмена являлись заслуженными мастерами спорта (ЗМС), 52 – мастерами спорта международного класса (МСМК), 311 - мастерами спорта (МС), 472 – кандидатами в мастера спорта (КМС) и 586 спортсменов имели взрослый разряд.

Для 80 ведущих атлетов сборной России (стайеры и средневики) было проведено измерение максимального потребления кислорода (VO_2max). 40 физически активных мужчин приняли участие в анализе состава мышечных волокон.

61 спортсмену-стайеру был проведен клинический анализ крови (гемоглобин, лейкоциты, эритроциты и прочие элементы). 1444 спортсмена сборной России прошли биохимический анализ крови (креатинкиназа, АЛТ, АСТ, креатинин и другие).

2.1.2 Группа спортсменов из других стран

В исследовании приняло участие 2099 спортсменов из различных стран мира. Образцы биологического материала забирались и анализировались в национальных научно-исследовательских лабораториях указанных стран, а затем передавались в виде генетических данных и сопутствующей информации внутри коллаборации.

2.1.3 Группа контроля

Для проведения ассоциативных исследований использовалось 734 геномных образца жителей России, 1870 – жителей других стран мира. Указанные представители группы контроля не являются профессиональными спортсменами и не имеют какого-либо взрослого спортивного разряда или звания. Кроме того, все они относятся к возрастной категории 20±3 лет, европеоидной расы и не

обладают существенными медицинскими отклонениями (не присвоена группа утраты трудоспособности).

В качестве дополнительных контрольных данных без учета возраста использовались также геномы европейцев, размещенные в базе данных проекта Illumina iControl DB [100] и доступных, согласно принятому соглашению для научно-исследовательских целей.

Для разработки программного протокола обработки данных ионного полупроводникового секвенирования использовался тестовый образец GCat59 [49], данные о генотипировании которого являются «золотым стандартом» и размещены в открытом доступе [51], а описание и декларация о свободном распространении и использовании размещена на сайте консорциума.

Для тестирования разработанного протокола анализа результатов ионного полупроводникового секвенирования использовались 7 геномных образцов жителей города Москва.

2.2 Методы генетического анализа

2.2.1 Пробоподготовка

Для последующего генетического анализа производился забор 4 мл венозной крови в пробирки, содержащие EDTA (Vacuette EDTA tubes, Greiner Bio-One, Austria), с использованием специального оборудования от компаний Qiagen и Eppendorf квалифицированным медицинским персоналом.

Выделение ДНК проводилось с использованием специального биохимического набора от компании Qiagen.

2.2.2 Ионное полупроводниковое секвенирование

Экзомное обогащение образцов геномной ДНК человека и подготовку библиотек для секвенирования осуществляли с использованием наборов ION Ampliseq Exome RDY-IC и Ion Xpress Barcode Adapters 1-16 (ThermoFisher) согласно рекомендациям производителя. Нанесение готовых библиотек на чип-стекло проводилось в парах или тройках с использованием оборудования Ion Chef

(ThermoFisher), а затем осуществлялось полупроводниковое секвенирование образцов ДНК с использованием секвенатора Ion Proton (ThermoFisher) и наборов расходных материалов Ion PI Chef Kit или Ion PI Hi-Q Chef Kit (ThermoFisher). Средняя выходная длина прочтений в результате секвенирования составила порядка 250 пар нуклеотидов на библиотеку.

Экспериментальные процедуры осуществлялись в лаборатории постгеномных исследований в биологии Федерального Научно-Клинического Центра Физико-Химической Медицины ФМБА России.

2.2.3 Генотипирование с использованием ДНК-чипов

Чип-гибридизация проводилась с использованием различных чипов, изготовленных по технологии Illumina Infinium:1M Omni Quad v1.2 (1 млн. 240 тыс. полиморфизмов, 4 образца на чип), Omni Express 12 v1.1 (775 тыс. полиморфизмов, 12 образцов на чип), Omni Express 24 v1.1 (773 тыс. полиморфизмов, 24 образца на чип), Core Exome Plus v1.2 (568 тыс. полиморфизмов, 12 образцов на чип).

Нанесение на чип производилось автоматически согласно рекомендациям производителя с использованием оборудования от компании Illumina, сканирование чипов производилось по технологии Illumina iScan в лаборатории постгеномных исследований в биологии Федерального Научно-Клинического Центра Физико-Химической Медицины ФМБА России.

2.2.4 Масспектрометрическое минисеквенирование

Определение искомых геномных состояний проводилось, в том числе, и с использованием реакции минисеквенирования. Очищенные ионнообменной смолой (Sequenom, США) образцы наносили на кристаллы матрицы (AnchorChip 400/384, Bruker Daltonics; Fluka, Германия) и проводили масс-спектрометрический анализ. Спектры получали с использованием технологии MALDI анализа (Matrix Assisted Laser Desorbtion/Ionization) и времяпролетного масс-спектрометра Autoflex (Bruker Daltonics, Германия). Для получения каждого масс-спектра использовали 50 импульсов лазера с мощностью излучения, установленной на

уровне минимального порогового значения, достаточного для десорбции-ионизации образца.

2.3 Биомедицинские исследования

2.3.1 Определение максимального потребления кислорода

Определение аэробных возможностей гребцов в тесте с нарастающей нагрузкой производили на механическом гребном эргометре PM 3 (Concept II, США) на базе федерального государственного бюджетного учреждения науки «Государственный научный центр Российской Федерации - Институт медико-биологических проблем РАН». Начальная нагрузка составила 150 Вт для мужчин и 100 Вт – для женщин, длительность ступени 3 мин, время отдыха между ступенями 30 с, прирост нагрузки 50 Вт на каждую последующую ступень. Работа выполнялась до отказа (снижение мощности гребка > 30 Вт от заданной мощности, дыхательный коэффициент > 1.1). Во время теста постоянно (каждый дыхательный цикл) регистрировали показатели газообмена и частоту сердечных сокращений (ЧСС, уд/мин) (газоанализатор MetaMax 3B, Cortex, Германия и Vmax 229, SensorMedics, США). Максимальное потребление кислорода (МПК, мл/мин/кг) определяли по значениям усредненных за последние 30 с каждой ступени теста показателей газообмена.

МПК у байдарочников определяли с помощью гребного тренажера Ефремова (Москва). Начальная нагрузка составила 8 кг для мужчин и 5 кг для женщин, длительность ступени 2 мин, время отдыха между ступенями 30 с, прирост нагрузки 1 кг на каждую последующую ступень. Во время теста постоянно (каждый дыхательный цикл) регистрировали показатели газообмена и частоту сердечных сокращений (ЧСС, уд/мин) (газоанализатор MetaLyzer II, Cortex, Германия). МПК (мл/мин/кг) определяли по значениям усредненных за последние 30 с каждой ступени теста показателей газообмена.

МПК у конькобежцев определяли в тесте с нарастающей нагрузкой с помощью велоэргометра Ergoselect 200K (Ergoline, Германия). МПК у

биатлонистов и лыжников определяли в тесте с нарастающей нагрузкой на беговой дорожке HP Cosmos (Германия).

2.3.2 Анализ состава мышечных волокон

Биопсия скелетных мышц у физически активных молодых мужчин проводилась на базе федерального государственного бюджетного учреждения науки «Государственный научный центр Российской Федерации - Институт медико-биологических проблем РАН». Для определения состава мышечных волокон, предварительно из *m. vastus lateralis* (под местной анестезией 1% раствором лидокаина) методом игольчатой биопсии по Бергстрему брали пробы мышечной ткани и замораживали в жидком азоте. Серийные поперечные срезы толщиной 10 μm готовили в криостате при -20ºС и монтировали на предметные стекла.

Для иммуногистохимического выявления изоформ ТЦМ использовали иммунопероксидазную технику. Применяли антитела против медленных (MHCs) и быстрых (MHCf) цепей миозина (клоны NCL-MHCs и NCL-MHCf (a+в) (Novocastra Laboratories)).

Срезы, которые инкубировали без первичных антител, использовали как контроль для выявления неспецифической окраски. Для усиления метки антиген-антитело был использован Vectrastain ABC kit (Vector Labs, CA), который визуализировался диаминобензидин пероксидазной реакцией.

Распределение волокон выражали как соотношение между числом волокон каждого типа на срезе к общему количеству волокон. Измеряли все волокна (200-300 волокон) на каждом срезе. Площадь поперечного сечения (ППС) измеряли не менее чем у 100 волокон каждого типа с помощью системы анализа изображений QUANTIMET-500 (Leica) с цветной цифровой видеокамерой JVC TK-1280E. Все сравниваемые срезы готовили и окрашивали одновременно реагентами фирмы Sigma (USA).

2.3.3 Биохимический и клинический анализ крови

Анализ крови проводился в двух различных вариантах: простом и расширенном. В простом случае проводился клинический анализ крови (гемоглобин, лейкоциты, эритроциты и прочие элементы). Расширенный вариант анализа включал кроме того биохимическое обследование показателей метаболизма и маркеров мышечного повреждения (как результата окислительного стресса): креатинкиназы, креатинина, аланинтрансаминазы, аспартаттрансаминазы, щелочной фосфатазы.

Для проведения анализа у спортсменов была забрана венозная кровь объемом 10 мл в утреннее время после продолжительного сна.

2.4 Распределение испытуемых по группам сравнения

В зависимости от поставленной задачи, а также от доступных сопутствующих данных спортсмены были разделены на различные группы для последующих сравнений.

2.4.1 Распределение испытуемых при проведении полногеномного ассоциативного исследования выносливости

Для исследования выносливости все исследуемые атлеты (449 человек) были разделены на 3 группы (табл.5).

В некоторых случаях группы 1 и 2 объединялись в общую группу выносливости, численностью 219 человек. Кроме того, внутри групп сохранялось разделение по полу.

Кроме разделения по видам спорта, также проводилось деления по уровню максимальных достижений. Из всех групп 234 атлета были классифицированы как «элитные» (заслуженные мастера спорта и мастера спорта международного класса), 168 атлетов как «субэлита» (мастера спорта и кандидаты в мастера спорта), 47 атлетов как «не элитные» (I взрослый разряд).

Таблица 5. Структура групп спортсменов, участвовавших в полногеномном ассоциативном исследовании выносливости.

#	Характеристика (время соревнования)	Общее кол-во	Количество спортсменов по видам спорта
1	Длинные и средние дистанции (более 5 минут) «выносливость»	158	гребцы (33 человека), бег 3 - 10 км (15 человек), биатлонисты (28 человек), конькобежцы на 5 - 10 км (3 человека), лыжники кросс-кантри (38 человек), марафонцы (2 человека), пловцы на 5 - 25 км (14 человек), пловцы на 800 - 1500 м (7 человек), спортивная ходьба (8 человек), триатлонисты (10 человек)
2	Короткие дистанции (от 45 секунд до 5 минут) «смешанная»	61	пловцы на 200 - 400 метров (10 человек), бег 800 - 1500 м (15 человек), конькобежцы на 1500 - 3000 м и шорт-трек конькобежцы (15 человек), гребцы на байдарках и каноэ 1000 м (14 человек), 1 - 4 км велогонщики на треке (7 человек)
3	Силовые спортсмены «силовая»	230	арм-рестлеры (3 человека), пловцы на 50 - 100 м (26 человек), гребцы на байдарках и каноэ на 200 м (19 человек), десятиборцы (3 человека), спринт-велосипедисты (22 человека), бег на 100 - 400 м (29 человек), 500 - 100 м конькобежцы и шорт-трек конькобежцы (20 человек), прыжки с шестом (23 человека), семиборцы (2 человека), пауэрлифтеры (16 человек), метатели копья (10 человек), тяжелоатлеты (57 человек)

2.4.2 Распределение образцов при ассоциативном исследовании генетических маркеров и мышечной структуры

Для ассоциативного исследования генетических маркеров, влияющих на структуру мышц 2178 спортсменов из России и Польши были разделены на 4 группы: "выносливость" (группа 1, n = 487, атлеты длинных и средних дистанций), "смешанная" (группа 2, n = 907, виды спорта, использующие и аэробный, и анаэробный методы получения энергии, игровые виды спорта, например), "спринтеры" (группа 3, n = 615, атлеты коротких дистанций), "силовая" (группа 4, n = 169, тяжелая атлетика). Кроме того, как и в одном из предыдущих исследований, 679 атлетов были отмечены как «элитные», 1050 – «субэлитные» и 449 – «не элитные».

2.4.3 Распределение образцов при ассоциативном исследовании генетических маркеров и состояний оксидативного стресса

В ассоциативном исследовании маркеров оксидативного стресса использовалась генетическая информация о 2664 спортсменах из России и Польши, которые были разделены на 7 групп по видам спорта (Табл.6).

Таблица 6. Разделение спортсменов по группам для ассоциативного исследования маркеров оксидативного стресса.

#	Характе-ристика	Общее кол-во	Количество спортсменов по видам спорта
1	Низкая интен-сивность дыхания	180	керлинг - 16 человек, гребцы - 50, стрелки - 69, художественные гимнасты - 7, прыгуны в воду - 17, шахматисты - 11, конный спорт - 10.
2	Длинные дистанции	235	пловцы более чем на 5 км - 30 человек, лыжники - 79, биатлонисты - 61, марафонцы - 9, спортивная ходьба - 8, шоссейные велосипедисты - 28, триатлонисты - 20.
3	Средние дистанции	187	бег на 3 - 10 км - 20 человек, лыжные гонки на 5 - 10 км - 3, пловцы на 800 - 1500 м - 24, гребцы - 140.
4	Короткие дистанции	86	пловцы на 200 - 400 метров - 22 человека, бег на 800 - 1500 м - 25, лыжные гонки на 1500 - 3000 м - 17, гребцы на байдарках 500 - 1000 м - 10, 1500 - 3000 м шорт-трек - 7, 500 - 1000 м каноэ - 5.
5	Смешанная группа	1452	бадминтонисты - 24 человека, баскетболисты - 109, бейсболисты - 38, боксеры - 143, гандболисты - 92, хоккеисты - 111, каратисты - 22, тэквондисты - 18, хоккеисты на траве - 19, синхронное плавание - 27, фехтовальщики - 64, фристайлисты - 11, фигурное катание - 76, стрелки из лука - 24, зимние триатлонисты - 10, сноубордисты - 33, футболисты - 36, пятиборцы - 23, софтболисты - 31, рэгбисты - 48, настольный теннис - 11, волейболисты - 115, минифутболисты - 9, водное поло - 59, водные слаломисты - 5, борцы - 294.
6	Скоростно-силовая группа	321	50 - 100 м плавание - 53 человека, 100 - 400 м бег - 57, 500 - 1000 м конькобежцы - 15, горнолыжники - 26, тройной прыжок - 37, прыжки с трамплина - 7, метатели копья - 18, гребцы на байдарках 200 м - 26, 500 - 1000 м шорт-трековцы - 16, 200 м каноэ - 24, велосипедисты-спринтеры - 31, десятиборцы - 8, семиборцы - 3.
7	Силовая группа	203	спортивная гимнастика - 63 человека, пауэрлифтинг - 35, тяжелоатлеты - 105.

2.4.4 Распределение образцов при поиске общего генетического профиля спортсмена-стайера

Для поиска общего генетического профиля спортсменов-стайеров использовались геномные данные проектов GENATHLETE [101] (Германия, Финляндия, Канада, США) и ассоциативного исследования атлетов Японии [102]. Для валидации результатов этого исследования использовалось различное количество представителей различных стран мира (Табл.7). Генотипирование проводилось в национальных университетах указанных стран.

Таблица 7. Количество генетических данных, использованных при поиске общего генетического профиля выносливости.

	Атлеты	"Элитные атлеты"	Контроль
Фаза поиска ассоциаций			
GENATHLETE	315	168	320
Япония	60	60	116
Фаза репликации			
Австралия	215	167	258
Эфиопия	75	75	198
Япония	143	na	692
Кения	276	66	83
Польша	113	61	161
Россия	153	68	734
Испания	170	170	198
Всего	**1520**	**835**	**2760**

2.5 Использованное программное обеспечение

2.5.1 Обработка данных NGS-секвенирования и генотипирования на ДНК чипах

Для подготовки результатов генотипирования на ДНК чипах использовался стандартный программный пакет Illumina Genome Studio (версии 2011.1), а затем программный алгоритм собственной разработки на языке PHP в среде Xubuntu Linux Server x64 14.04.3 LTS.

Для подготовки результатов секвенирования использовались следующий программный протокол:

1) cutadapt (версии 1.8.1) с параметрами -f fastq -e 0.1 -q 14 -O 13;

2) bwa (версии 0.7.12-r1039) с параметрами mem -t 23;

3) samtools (версии 0.1.19) index

4) samtools mpileup с ключами -I -u –f

5) vcf-tools (vcf-annotate, версии 0.1.14) с ключом<+/d=22/q=30/Q=30/-a>,

6) NGSrich (версии 0.7.8)

Кроме того, использовался программный пакет Torrent Suite (версии 5.0) в стандартной комплектации от ThermoFisher, inc. для сравнения результатов его работы с полученными данными в результате использования свободного программного обеспечения.

Torrent Suite и Genome Studio являются проприетарными пакетами, лицензия на использование которых входит в оферту при приобретении прибора и является доступной для любого образца, отсеквенированного или отсканированного с применением указанных приборов.

Все остальное программное ПО является свободным и распространяется по различным версиям лицензии GNU/GPL.

2.5.2 Статистический анализ и визуализация

Для статистического сравнения полиморфизмов различных групп спортсменов использовался программный пакет PLINK 1.9, который автоматизирует проведение ассоциативных исследований. В качестве алгоритмов поиска ассоциаций использовался непарный Т-тест, а также Тест Фишера (ANOVA). Результаты фильтровались, в зависимости от постановки задачи, с использованием поправки Бонферрони-Холма или Бенджамини-Хохберга (FDR). Для сравнения качества алгоритмов применялся критерий F1-мера, позволяющий производить тестирование сложных статистических гипотез.

2.5.3 Другое программное обеспечение

Для приведения форматов данных в соответствии с требованиями готовых программных пакетов использовались программы собственной разработки на языках программирования bash, Python, Perl, PHP и SQL (процедуры базы данных). Графические интерфейсы были реализованы на языках PHP и HTML (в версиях 4.0 и 5.0) с использованием WEB-протоколов.

В качестве среды для разработки баз данных использовалась СУБД MySQL (версия mysql-server-5.7).

2.6 Аппаратные средства

Анализ данных секвенирования и чип-гибридизации производился на вычислительном кластере ФГБУ ФНКЦ ФХМ ФМБА России на десяти 24 - ядерных узлах с 64 Гб ОЗУ и двух 48 - ядерных узлах с 256 Гб ОЗУ (общая мощность 3 терафлопс); операционная система CentOS 7.0, планировщик Sun Grid Engine (SGE).

Для хранения и обработки промежуточных и конечных результатов обсчета использовалось 2 дисковых полки по 10 Тб каждая с резервированием стандарта RAID 5.

Глава 3. Результаты и обсуждение

3.1 Обработка результатов ионного полупроводникового секвенирования

Как уже было сказано, существует значительное количество технических трудностей при обработке результатов ионного полупроводникового секвенирования. Для их решения на рынке присутствует готовое программное решение – Torrent Suite, а также достаточно много свободных программных продуктов для выравнивания прочтений и для определения нуклеотидов по заданным позициям из сформированных файлов.

Для оценки точности определения геномных состояний с помощью штатного программного обеспечения на реальных образцах было проведено сравнение результатов анализа секвенирования двух спортсменов сборной России (AmpliSEQ Exome Kit RDY) с помощью пакета Torrent Suite (преобразованного к списку полиморфизмов) и чип-гибридизации (Core Exome Plus v1.2, с фильтрацией по GC-score > 0,9). Предполагалось, что при таком сравнении ошибка чип-гибридизации не является существенной по отношению к ошибкам секвенирования. Требуемое число определенных геномных состояний с использованием секвенирования составило порядка 24 тысяч, в зависимости от GC-score чип-гибридизации (Табл.8). В случае, если результаты секвенирования и чип-гибридизации совпали, то считалось, что генотип определен верно. Кроме того, состояниям, которые определялись путем секвенирования, но не совпадали с результатами чип-гибридизации, присваивался флаг «ошибочные», а тем, которые не определялись совсем – «не определившиеся». Программный пакет Torrent Suite запускался с набором стандартных параметров, которые выставлялись автоматически при указании всех используемых реактивов при секвенировании.

Было выдвинуто предположение, что точность результатов (Табл.8) возможно повысить путем использования открытого программного обеспечения со значительным числом дополнительных настроек. За основу настраиваемого протокола были взяты пакеты BWA для выравнивания прочтений на референс и

samtools для определения геномных состояний. На первом шаге было проведено составление программного конвейера и определение начальных параметров запуска каждой из программ (Рис.9).

Таблица 8. Результаты сравнения определения геномных состояний пакетом Torrent Suite с оптимальными параметрами, согласно рекомендациям компании-производителя, с результатами чип-гибридизации по технологии Illumina Infinium.

Образец	Tri04	Tri05
Пол	M	M
Число ожидаемых геномных состояний после фильтрации чип-гибридизации	24493	24133
Определено верно	19957	19124
Определено неверно	212	215
Не определено при секвенировании	4324	4794
Точность протокола	**98.94%**	**98.88%**

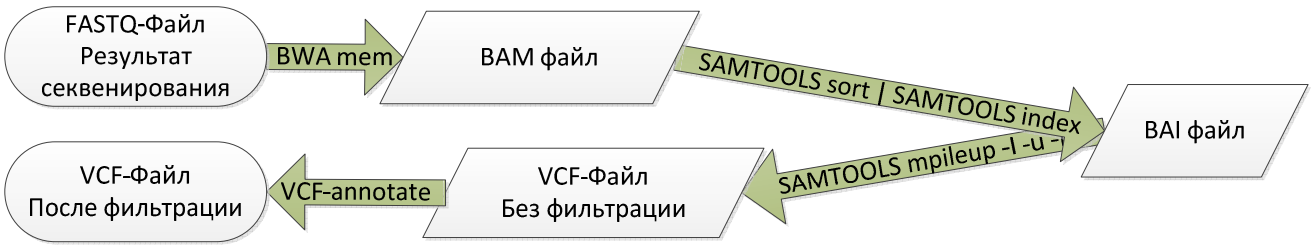

Рисунок 9. Блок-схема работы настраиваемого алгоритма выравнивания прочтений и идентификации генетических вариантов после ионного полупроводникового секвенирования.

Так как технология формирования библиотеки является амплификационной, то на этапе предварительной обработки было проведено удаление адаптеров секвенирования с помощью программного пакета cutadapt. Выравнивание прочтений производилось пакетом BWA с алгоритмом MEM, а затем пакетом

samtools (view, sort, index) производилось формирование отсортированного по хромосомам, индексированного файла формата BAM. У указанных пакетов отсутствуют какие-либо настроечные параметры, за исключением флага –t у пакета BWA, который указывает число вычислительных узлов используемой ЭВМ и не влияет на результаты работы конвейера в целом.

На следующем шаге производится поиск достоверных геномных состояний. Выбор программных продуктов производился на основании сравнений, проведенных другими авторами, поэтому, был выбран пакет samtools mpileup, как наиболее подходящий для поиска полиморфизмов генома человека. Так как самой распространенной ошибкой прочтения ДНК у приборов с технологией Ion Proton являются гомополимеры, то на данном этапе поиск вставок и делеций не рассматривался и при запуске пакета samtools mpileup использовался дополнительный параметр –I, означающий удаление вставок и делеций из результатов работы протокола. На последнем шаге производится проверка качества прочтения и фильтрация качественных вариантов с использованием пакета VCF-tools (опция vcf-annotate), на которой стоит остановиться подробнее.

Основными вариабельными параметрами (используемыми в VCF-tools), по которым возможно определить ошибочность определения геномного состояния являются: покрытие в точке (dp, d); среднее качество прочтения нуклеотида (qual, q); минимальное качество картирования прочтения (minMQ, Q); ошибка чтения нуклеотидов в различных направлениях (StrandBias, ключ 1); ошибка оценки качества средней приборной ошибки (BaseQualBias, ключ 2); ошибка точности определения позиции прочтения в геноме (MapQualBias, ключ 3); ошибка, связанная с краевыми эффектами амплификации (EndDistBias, ключ 4). Остальные же характеристики никак не влияют на определение ошибочности геномного состояния.

Для того чтобы определить оптимальные настройки фильтрации результатов секвенирования с использованием VCF-tools, предполагалось произвести максимизацию итоговой точности путем простого программного подбора вариантов. Для оптимизации разрабатываемого алгоритма были

использованы как собственные экспериментально полученные данные, так и доступные референсные данные по секвенированию тестовых образцов ДНК. Для того чтобы исключить ошибки, связанные с экспериментальной работой, на первом шаге в качестве исходных данных использовались открытые результаты анализа 59 образца проекта Genome In A Bottle (GIAB) [51]: FASTQ-файл с прочтениями, полученными с использованием прибора Ion Proton (AmpliSEQ Exome Kit RDY), с одной стороны и искусственный VCF-файл, содержащий корректный список мутаций с другой. На втором шаге проводилось аналогичное сравнение с настройкой параметров с использованием собственных результатов секвенирования и чип-гибридизации.

3.1.1 Оптимизация протокола на стандартизированных прочтениях

Чтобы произвести настройку исключительно биоинформатических протоколов и исключить ошибки реального эксперимента поиск оптимальных параметров проводился следующим образом. В качестве достоверных результатов секвенирования был использован VCF-файл образца 59 проекта GIAB (Ion Proton, AmpliSEQ Exome Kit RDY, AmpliSEQ HiQ, 1 образец на чип-стекло), который был отфильтрован по тем фрагментам ДНК, которые должны секвенироваться при использовании указанного набора согласно информации производителя ThermoFisher. Затем производился программный перебор всех возможных дополнительных параметров и, предполагалось, что те записи, которые в VCF-файле имели индекс PASS, являлись верными при установленных параметрах. В качестве стартовых состояний всех исходных параметров использовались значения «по умолчанию», которые, согласно инструкции к пакету VCF-tools разработаны для конкурирующей платформы.

При оценке качества работы установленных фильтров требовалось определить метрику и алгоритм, по которому она будет производиться. Существует два различных подхода к решению указанного вопроса: первый подход делит все возможные правильно определенные состояния на 2 класса (positive и negative) и не идентифицированные варианты, а второй содержит всего

один класс и не идентифицированные варианты. При первом подходе (далее «метод 1») критерием оценки мог служить статистический критерий F1-мера, а при втором (далее «метод 2») - простое отношение числа ошибок к общему числу определенных полиморфизмов. В результате минимизации критерия в метрике первого подхода ожидался такой набор параметров, при котором определялось максимальное число отличий от референса при меньшем общем числе определенных генотипов, а во втором подходе - максимальное число верно определенных любых генотипов.

При сравнении, с использованием меры F1, требовалось разделить все возможные результаты на 2 класса и, соответственно, 4 состояния. Традиционно, за класс 1 (“positive”) принимается обнаружение отличия от референсного генома (hg19), а за класс 2 (”negative”) – определенное состояние, являющееся референсной гомозиготой. Понятно, что в случае отсутствия существенных сбоев при секвенировании, число результатов “negative” на несколько порядков больше, чем состояний “positive”. Поэтому, для оценки качества протокола используют термины точности (precision) и воспроизводимости (recall).

$$\text{Точность} = \frac{\text{верные класса 1}}{(\text{верные класса 1} + \text{ошибочные класса 1})}$$

$$\text{Воспроизводимость} = \frac{\text{верные класса 1}}{(\text{верные класса 1} + \text{ошибочные класса 2})}$$

$$\text{F1} - \text{мера} = \frac{2 \times \text{точность} \times \text{воспроизводимость}}{(\text{точность} + \text{воспроизводимость})}$$

В данном случае сравнение проводилось на уровне аллелей, и для каждого из ожидаемых полиморфизмов было выделено по 2 независимых варианта, каждый из которых относился к одному из 4 ожидаемых состояний:

к верно определенным класса 1 образцам относили нереференсные аллели, совпавшие с базовой VCF;

к верно определенным класса 2 образцам относили референсные аллели, совпавшие с базовой VCF;

к неверно определенным класса 1 образцам относили нереференсные аллели, не совпавшие с базовой VCF;

к неверно определенным класса 2 образцам относили референсные аллели, не совпавшие с базовой VCF;

Для гетерозиготных состояний сравнение проводилось для каждой аллели отдельно, но с условием максимизации верных состояний, то есть для пары нереференсной гомозиготы против гетерозиготы предполагалось, что одна аллель является верно определённой класса 1, а другая аллель неверно определена класса 1.

В результате эксперимента было проведено значительное число повторов вычислений с различными состояниями фильтров, использованные параметры и результаты некоторых из вычислений приведены в таблице 9.

Таблица 9. Параметры и результаты некоторых технических повторов при подборе параметров с помощью критерия F1.

		Параметры запусков			
Параметры	**Обозначение**	**1 (по умолчанию)**	**2**	**3**	**4 (оптимальный набор)**
Минимальное покрытие	DP	10	12	16	14
Минимальное качество картирования	MinMQ	10	10	20	30
Минимальное качество прочтения	Qual	10	10	20	10
Ошибка направления	1, StrandBias	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$
Ошибка оценки качества	2, BaseQualBias	0	0	0	0
Ошибка картирования	3, MapQualBias	0	0	0	0
Ошибка края	4, EndBias	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$
Результаты					
Верно определено отличие от референса	TruePositives	36436	33525	32744	33437
Неверно определено отличие от референса	FalsePositives	3998	2886	2768	2545
Неверно определен референс	FalseNegatives	4544	3015	2993	2763
F1-score	**из 100%**	**88,95%**	**91,12%**	**91,99%**	**92,98%**

Такой алгоритм определения оптимальных параметров является хорошим со стороны математической статистики, но если проанализировать биологический смысл получившегося набора параметров, то возможно заметить, что такой протокол будет определять значительное число отличий от референса ценой существенно меньшей точности определения референсных состояний. С другой стороны, при работе с мультифакторными заболеваниями, каждый изученный полиморфизм в любом из своих состояний осуществляет вклад в изменение рискового статуса его носителя, поэтому определение максимального числа качественных нуклеотидов вне зависимости от согласия с референсом кажется наиболее логичным.

Для решения такой задачи возможно сформировать всего 1 класс верно определенных нуклеотидов:

к верно определенным образцам относили генотипы, совпавшие с базовой VCF проекта GIAB;

к неверно определенным образцам относили генотипы, не совпавшие с базовой VCF;

Гетерозиготные состояния рассматривались как целый генотип, и ожидалось строгое соответствие именно гетерозиготного генотипа. Кроме того, состояния, которые должны были присутствовать в итоговом VCF-файле, но были помечены как ошибочные, были отнесены к не определенным.

Регулируя дополнительные параметры, допустимо достичь различных предельных точек, интересных с биологической точки зрения. В первую очередь, это такой набор фильтров, при котором минимизируется число ошибок. Понятно, что в такой ситуации число верно определенных генотипов будет невелико, но достоверность их будет максимальная. Другой предельной точкой будет являться такой набор параметров, при котором минимизируется отношение ошибочных точек к верно определенным. При усилении фильтров до достижения этой точки резко уменьшается число ошибочных вариантов, при том, что число верных вариантов остается неизменным, а после преодоления данного предела, все происходит наоборот: число ошибок практически постоянно, а количество верных

генотипов снижается. И, наконец, третьей предельной точкой будет максимальное число вариантов при некотором высоком, но допустимом, относительном количестве ошибок. Таким критерием может выступать, например, типовая точность чип-гибридизации Illumina в 99,4%.

В данном эксперименте производилась оптимизация общей точности определения полиморфизмов при условии, что допустимое количество отфильтрованных состояний не должно превышать 10%. Для данных проекта GIAB ожидаемое количество определенных геномных состояний пересечения набора для обогащения и итогового VCF-файла составило 36595, поэтому все наборы параметров, формирующих в результате работы программного конвейера число состояний менее 32936, были исключены из анализа. В качестве стартовых состояний всех исходных параметров пакета VCF-tools использовались значения «по умолчанию».

В результате эксперимента было проведено значительное число технических повторов с различными состояниями фильтров, использованные параметры и результаты 4 тестов, являющиеся самыми показательными, отображены на таблице 10.

Было замечено, что покрытие, качество картирования и качество определения прочтения зависимы и ситуация, при которой покрытие высоко, а качество картирования, например, низко, не реализуется, как и обратная. Кроме того, определено, что безусловный предел по качеству достигается при сочетании набора параметров <DP; minMQ; Qual> = <50; 50; 50> (точность 99,99%, потеря данных 50%), оптимум при <22; 30; 30> (точность 96,20%, потеря данных 8%).

В качестве дополнительной оптимизации было принято утверждение, что при определении гомозиготы допустимо меньшее общее покрытие, чем при определении гетерозиготы. Поэтому, был написан модифицирующий программный пакет, который в VCF-файлах, получаемых с критерием <22;30;30> добавлял дополнительный собственный фильтр «LOW_GETERO» и устанавливал его статус равным «не пройден» для точек, покрытие в которых было ниже 28.

Таблица 10. Параметры и результаты некоторых технических повторов при подборе параметров с помощью критерия одногрупповой точности.

		Параметры запусков			
Параметры	**Обозначение**	**1 (по умол-чанию)**	**2**	**3**	**4 (опти-мальный набор)**
Минимальное покрытие	DP	10	22	28	22/28
Минимальное качество картирования	MinMQ	10	30	30	30
Минимальное качество прочтения	Qual	10	30	30	30
Ошибка направления	1, StrandBias	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$
Ошибка оценки качества	2, BaseQualBias	0	0	0	0
Ошибка картирования	3, MapQualBias	0	0	0	0
Ошибка края	4, EndBias	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$
Результаты					
Верно определено состояний	TruePositives	36734	33800	32937	33782
Не определено состояний	FalseNegatives	187	3115	3987	3133
Определено ошибочных состояний	FalsePositives	14736	1275	1045	1030
Точность	**из 100%**	**59,9%**	**96,2%**	**96,8%**	**97,0%**

Результаты сравнения показали, что действительно, в обсуждаемой метрике, алгоритм, состоящий из samtools, vcf-tools с параметрами <d=22;q=30;Q=30> и с удалением гетерозигот, покрытие которых ниже, чем 28 является оптимальным (Табл.10).

В результате проведенного анализа было получено несколько различных наборов дополнительных параметров программ разработанного конвейера, каждый из которых оптимизирует какой-то один из показателей. «Метод 1» максимально точно определяет состояния, отличные от референса, поэтому его стоит использовать для подготовки результатов секвенирования к ассоциативным исследованиям, а «метод 2» формирует VCF-файл с минимальным числом ошибок определения геномных состояний в целом, поэтому имеет смысл использовать его для индивидуального описания генома пациента. Стоит заметить, что различие между двумя алгоритмами при качественном

секвенировании незначительно и составляет не более нескольких тысяч определенных нуклеотидов на экзом, качество которых флуктуирует на грани пороговых значений.

В связи с тем, что экзомное секвенирование в дальнейшем использовалось исключительно для определения индивидуальных особенностей спортсменов, то дополнительная оптимизация на экспериментальных образцах проводилась с использованием «метода 2».

3.1.2 Оптимизация протокола на экспериментально полученных прочтениях тестовых образцов ДНК

Использование результатов проекта GIAB при настройке программного конвейера позволило исключить из общей ошибки факторы, связанные с особенностями конкретных образцов и конкретного оборудования. Конечно, для использования разработанного протокола на реальных образцах в будущем требуется его валидация с использованием результатов, полученных с использованием того же прибора и того же протокола подготовки библиотеки.

Для проверки протокола на экспериментальных данных было проанализировано 6 образцов ДНК: 3 мужских и 3 женских. Для них был отсеквенирован экзом с использованием технологии Ion Proton и набора для обогащения AmpliSEQ Exome Kit RDY.

В качестве основных референсов использовались результаты чип-гибридизации на различных чипах (1M Omni Quad v1.2, Omni Express 12 v1.1, Omni Express 24 v1.1, Core Exome Plus v1.2) от компании Illumina, поэтому число геномных состояний, присутствующих и в результатах чип-гибридизации и секвенирования одновременно, для каждой пары является различным. Результаты чип-гибридизации были отфильтрованы с GC-score более 0,90.

При кроссплатформенном парном сравнении образцов (Табл.11) точность определения генотипов, в среднем по образцам, превышает 99%, что означает высокую точность разработанного протокола.

В качестве дополнительного контроля было проведено сравнение результатов работы разработанного программного конвейера с VCF-файлом, полученным в результате работы пакета Torrent Suite с оптимальными параметрами, согласно рекомендациям компании-производителя. Было показано, что точность разработанного протокола существенно выше (Табл.12).

Таблица 11. Сравнение результатов работы разработанного программного конвейера с соответствующими результатами чип-гибридизации.

	Ph3078	Ph233	Ph3158	Tri04	Tri05	**CON_2X**
Сравниваемое число геномных состояний	30833	30755	16709	24493	24133	7280
Определено верно	27889	29041	15258	22018	21302	6439
Определено неверно	127	99	787	137	128	54
Не определено при секвенировании	2817	1615	664	2338	2703	787
Пол	Жен	Жен	Жен	Муж	Муж	Муж
Точность	**99,54%**	**99,66%**	**94,84%**	**99,38%**	**99,40%**	**99,16%**
Доля определенных геномных состояний	**90,45%**	**94,43%**	**91,32%**	**89,90%**	**88,27%**	**88,45%**

Таблица 12. Сравнение результатов работы разработанного программного конвейера с соответствующими VCF-файлами, полученными с помощью пакета Torrent Suite при оптимальных параметрах, согласно рекомендациям компании-производителя.

Образец	**Tri04**	**Tri04**	**Tri05**	**Tri05**
Пол	M	M	M	M
Использованный протокол	Наш	Torrent Suite	Наш	Torrent Suite
Среднее покрытие экзома	81,42	81,42	82,51	82,51
Среднее отклонение покрытия	51,86	51,86	55,4	55,4
Число ожидаемых геномных состояний после фильтрации результатов чип-гибридизации	24493	24493	24133	24133
Определено верно	22018	19957	21302	19124
Определено неверно	137	212	128	215
Не определено при секвенировании	2338	4324	2703	4794
Предполагаемое количество ошибочных геномных состояний на экзом (шт.)	**370 тыс.**	**631 тыс.**	**360 тыс.**	**668 тыс.**
Точность протокола	**99,38%**	**98,94%**	**99,40%**	**98,88%**

Дополнительно была проведена оптимизация и определение зависимости точности программного конвейера от плотности нанесения на чип-стекло и от типа используемого химического набора для секвенирования. Для решения указанного вопроса один и тот же образец был отсеквенирован с использованием химического набора для секвенирования Standart и различных нанесений на чип-стекло: одиночное, парное и тройное. Кроме того, был проведен сиквенс парного образца и с использованием AmpliSEQ HiQ химического набора, который, согласно рекомендации производителя призван увеличить техническую точность секвенирования. В качестве референса для проведенного сравнения были использованы результаты чип-гибридизации указанного образца на чипе Illumina 1M Omni Quad v1.2 с дополнительной фильтрацией GC-score более 0.9, поэтому количество геномных состояний, ожидаемых в результатах такого сравнения, составило 7280.

Таблица 13. Результаты анализа данных секвенирования одного образца с применением различных условий нанесения и различных химических наборов для секвенирования. Точность вычислена с помощью разработанного программного конвейера («метод 2»).

Образец	**CON_1X**	**CON_2X**	**CON_3X**	**CON_HI_3X**
Нанесение	**1 образец / чип**	**2 образца / чип**	**3 образца / чип**	**3 образца / чип**
Тип химического набора	Standart	Standart	Standart	HiQ
Среднее покрытие экзома	178,24	110,3	67,18	83,78
Среднее отклонение покрытия	139,9	86,5	54,42	49,96
Число ожидаемых геномных состояний после фильтрации чип-гибридизации	7280	7280	7280	7280
Определено верно	6583	6439	5835	6398
Определено неверно	59	54	60	57
Не определено при секвенировании	638	787	1385	882
Точность протокола	**99,12%**	**99,16%**	**98,97%**	**99,11%**
Доля определенных геномных состояний	**90,42%**	**88,45%**	**80,15%**	**87,88%**

На основании проведенного анализа (Табл.13) возможно сделать вывод, что тип нанесения на чип-стекло на точность работы программного конвейера не влияет при использовании AmpliSEQ HiQ химического набора, но доля определенных геномных состояний ожидаемо растет при уменьшении плотности нанесения и достигает максимума при нанесении 1 образец на чип-стекло.

Согласно полученным результатам, разработанный протокол является оптимальным для предварительной обработки результатов секвенирования по ионной полупроводниковой технологии для последующего анализа индивидуальных особенностей пациента.

3.2 Адаптация результатов чип-гибридизации

Наряду с экзомным секвенированием для получения генетической информации использовалась технология чип-гибридизации Illumina Infinium с различными вариантами ДНК-чипов (1M Omni Quad v1.2, CVD 55 v1.0, Omni Express 12 v1.1, Omni Express 24 v1.1, Core Exome Plus v1.2).

Как было замечено выше, несмотря на хорошее качество генотипирования, существует множество ошибок, связанных с отличием референсного генома, который использовался при дизайне зондов производителем и общепринятыми сборками (NCBI, Ensembl): несоответствие цепи позиционирования полиморфизма, несоответствие координат местоположения полиморфизма в геноме и несоответствие указателей rs текущей версии каталога dbSNP. Поэтому задачей, поставленной в данной работе, была разработка технологии предварительной фильтрации результатов чип-гибридизации для дальнейшего использования их как в полногеномных ассоциативных исследованиях, так и для анализа индивидуальных особенностей пациента.

В качестве исходных данных на биоинформатическую обработку поступали текстовые файлы, содержащие в каждой из строк наименование зонда, наименование образца, определившийся генотип, также показатели GC-score, оценивающие эффективность работы зонда.

Известно, что в технологии чип-гибридизации Illumina зонды могут быть комплементарны любой из цепей ДНК. Для стандартизации получаемых результатов с данными секвенирования в качестве референсной последовательности была использована сборка генома человека Human Genome 19 (hg19, GRCh37), размещенная в базе данных NCBI. Компания Illumina для каждого из чипов поставляет текстовый файл, который содержит последовательности зондов и их координаты в той версии генома, в которой они были разработаны.

Для исключения всех описанных ошибок был разработан алгоритм и написан программный продукт. На первом шаге производится поиск последовательностей зондов, полученных от производителя, по их координатам в референсном геноме (hg19) как в прямом, так и в комплементарно-обратном вариантах. В качестве критерия соответствия было установлено, что допустимое максимальное количество несоответствующих нуклеотидов при выравнивании не может превышать 2 на зонд. После первого этапа формировался список зондов, которые были обнаружены, вместе с указанием цепи их нахождения, а также список тех зондов, координаты которых не соответствуют заявленным.

На втором этапе зонды, координаты которых не соответствовали сборке hg19, с помощью программного пакета blastn [103] выравнивались на референсный геном. Считалось, что поиск был произведен успешно в том случае, если при таком выравнивании с учетом не более 2х несоответствий находился только один уникальный фрагмент ДНК.

Простым поиском, даже с учетом использования blastn, по полному геному не получалось обнаружить порядка сотен зондов на каждый чип, поэтому такие зонды просматривались вручную или помечались как ошибочные и отбраковывались.

Все указанные шаги были проделаны для всех чипов, которые были использованы в данной работе. Исходя из того, что результаты генотипирования на ДНК-чипах планировалось использовать для проведения ассоциативного исследования, то порог отсечения низкокачественных результатов был

установлен величиной в 0,5 GC-score, что следует из рекомендаций других разработчиков [55]. Результаты такой проверки по всем типам использованных ДНК-чипов представлены в таблице 14.

Таблица 14. Результаты анализа структуры зондов чипов компании Illumina, использованных в данной работе.

Чип	1M Omni Quad v1.2	CVD 55 v1.0	Omni Express 12 v1.1	Omni Express 24 v1.1	Core Exome Plus v1.2
Общее число полиморфизмов	1134515	49094	719665	713014	546929
Число корректно определяемых полиморфизмов	769926	38851	528969	522482	442889
Число тестируемых геномных состояний, координаты для которых указаны верно, а цепь - ошибочно	267684	9878	189467	189311	101294
Число тестируемых геномных состояний, координаты которых определены путем выравнивания blastn	2452	365	1123	1120	232
Число геномных состояний, исключенных из анализа по различным причинам	94453	0	106	101	2514
Сборки референсного генома, на основе которой производился дизайн зондов	hg18 / hg19	hg18	hg19	hg19	hg19 / hg38

Количество зондов с неверными координатами в файле-схеме чипа от производителя составляло порядка 1000 на каждый вариант чипа. Такое несоответствие возможно связано с тем, что дизайн чипа разрабатывался с использованием более старого референса (например, hg18) или в координатном пространстве контигов, которые не были встроены в общую последовательность той или иной хромосомы (в геноме человека сборки hg19 таких контигов более 50).

Существенное количество исключенных геномных состояний (особенно в чипе 1M Omni Quad v1.2) связано с различными факторами. В первую очередь это

состояния, которые располагались в контигах, которые были перенесены или исключены при изменении версии сборки референсного генома с более старых на hg19 или добавлены в результате перехода от hg19 к hg38 (в основном фрагменты 7 хромосомы и митохондриальной ДНК). Кроме того, в список исключенных попали те зонды, которые не имеют близкой однозначной последовательности либо не являются уникальными для генома при использовании в качестве референса сборку hg19.Особенность чипа CVD 55 v1.0 по отношению ко всем остальным заключается в том, что при его разработке были использованы те геномные состояния, ассоциация которых с сердечно-сосудистыми заболеваниями была уже показана и зонды для их детектирования были проверены неоднократно в составе других чипов компании Illumina.

В итоге, если обработка результатов генотипирования с использованием чипа 1M Omni Quad была бы проведена по стандартным протоколам Illumina, и в качестве источника данных был бы выбран файл-отчет с указанием генотипов в виде TOP (как подразумевается компанией Illumina), то в 16604 геномных состояниях из 267684 был бы ошибочно определен генотип без возможности последующего исправления методами биоинформатики. Указанные полиморфизмы являются транзициями (A/T, G/C), которые, согласно документации производителя, находятся на TOP цепи их сборки генома hg19, тогда как они расположены на BOTTOM цепи основной версии генома hg19 NCBI. Кроме того, у 253532 из 267684 полиморфизмов, являющихся трансверсиями, возникла бы ситуация несоответствия нуклеотидов результатам секвенирования из-за того, что они также находятся на BOTTOM цепи, но совпадают с направлением гена (т.е. FORWARD). Нуклеотиды таких полиморфизмов находятся в согласии с базами данных (например, dbSNP), но являются комплементарными тем парам, которые могут быть обнаружены NGS-методами. Для таких состояний потребовалось бы написать программу-обработчик, которая бы определяла такие состояния и автоматически заменяла значения генотипов на комплементарные пары у всех образцов.

Кроме того, со времени разработки чипа 1M Omni Quad (2012 год) некоторые полиморфизмы изменили свой индекс rs и на следующих чипах находятся с новой его версией. Именно такие зонды не находились по координатам, но легко обнаруживались сплошным поиском (blastn), и, в случае, если бы интеграция данных производилась по индексам rs, такие точки были бы утеряны. В процессе анализа было обнаружено 2452 таких полиморфизма.

Таким образом, следует заметить, что при использовании штатного программного обеспечения (Genome Studio версии 2011.1), процент неправильно идентифицированных полиморфизмов может достигать 2%. При использовании высокоплотных ДНК-чипов, таких как 1M Omni Quad v1.2 в эту категорию попадут более 20000 полиморфизмов. Поэтому, применение разработанного протокола обработки данных генотипирования на ДНК-чипах является необходимым этапом, особенно при проведении исследований, комбинирующих различные версии ДНК-чипов.

3.3 Применение разработанных протоколов

Разработанные алгоритмы первичной обработки и анализа генетических данных, получаемых в ходе ионного полупроводникового секвенирования и генотипирования на ДНК-чипах, были использованы для проведения серии ассоциативных генетических исследований в группах профессиональных спортсменов из Российской Федерации и других европейских стран.

Основными задачами этих исследований были поиск генетических ассоциаций с показателями скорости кислородного обмена, структуры мышечных волокон и общей выносливости.

В зависимости от исследуемого параметра, количество и состав исследуемых спортсменов (специализация, национальная принадлежность, спортивная успешность) в разных вариантах исследования существенно варьировали. Также были различными и методы, при помощи которых осуществлялось генотипирование: различные версии ДНК-чипов, NGS-секвенирование, минисеквенирование MALDI-TOF.

Разработанные биоинформатические протоколы были использованы для адаптации данных ко всем последующим ассоциативным исследованиям. Также осуществлялся независимый контроль этнического однообразия исследуемых групп.

Для контроля этнического однообразия была использована база данных Illumina iControl DB [100], из которой были отобраны образцы, для которых был указан этнический фенотип CEU (европеоид), CHB (монголоид) или YRI (негроид). Добавочным условием было то, что эти же образцы были отгенотипированы на тех же версиях ДНК-чипов, которые были использованы в наших исследованиях. Далее в пространстве всех полиморфизмов (1 млн. 200 тыс.) и 4х независимых групп (европеоиды, монголоиды, негроид и спортсмены) был применен метод главных компонент (в среде R), и его результаты были представлены в качестве графика по 2 главным осям, которые объясняют порядка 30% суммарного отличия групп (Рис.10).

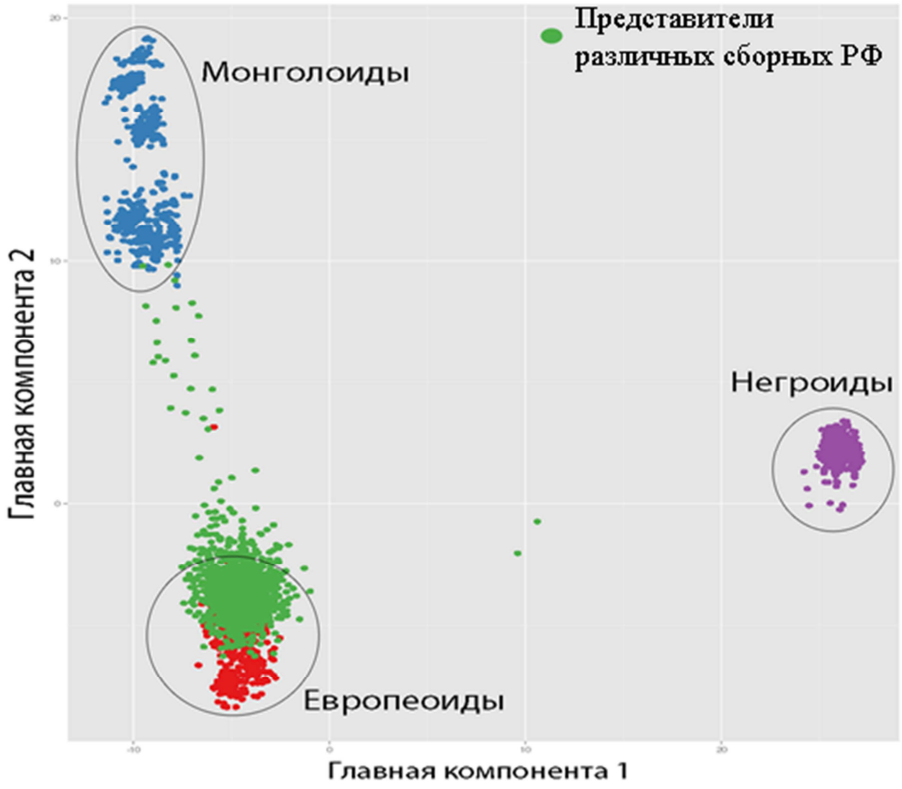

Рисунок 10. Кластеризация представителей основных этнических групп в пространстве двух главных компонент с наложением группы спортсменов различных сборных РФ.

В ходе сравнения собственных экспериментальных данных было показано, что собранные для исследования группы этнически гомогенны, за исключением 2 негроидов и 21 монголоида, которые были исключены из дальнейшего исследования.

Таким образом, общая группа из 1423 спортсменов сборной России является этнически однородной и ее допустимо использовать в дальнейших исследованиях.

3.3.1 Анализ генетических основ выносливости

Известно, что показатели, характеризующие выносливость организма во многом детерминированы на генетическом уровне. Одним из таких классических показателей является аэробная выносливость. Для поиска генетических маркеров, ассоциированных с показателями аэробной выносливости, было выполнено двухэтапное исследование, состоящее из поискового полногеномного ассоциативного исследования, выполненного на группе российских спортсменов и серии валидаций результатов в формате случай-контроль на расширенной выборке, включающей, в том числе, и группы контроля.

Результаты чип-гибридизации были обработаны согласно собственному протоколу, описанному в предыдущих разделах. Затем, были проведены требуемые групповые сравнения с использованием программного пакета PLINK. Различия в фенотипах между группами были проанализированы с использованием ANOVA-критерия или непарных t-критериев (сравнение изменения вариаций двух генотипов), а затем все данные были представлены как среднее и стандартное отклонение от среднего. Показатели сравнивались по алгоритму χ^2, использовалась непараметрическая корреляция Спирмана с применением ограничения: значения с P-value величиной менее 0,05 в единичном тесте предполагались статистически достоверными. В качестве уровня полногеномной значимости был установлен критерий $P < 10^{-5}$.

Поисковая часть исследования была выполнена на группе из 80 российских спортсменов – стайеров и средневиков, для которых имелись результаты теста VO_2max, при сравнении в пределах 5 категорий, учитывающих комбинации пола

и длины дистанции. Для генотипирования был использован ДНК-чип (Illumina 1M Omni Quad), позволяющий проанализировать 1140419 полиморфизмов. В качестве показателя аэробной выносливости анализировали данные максимального потребления кислорода (МПК, тест VO_2max).

По результатам поисковой фазы исследования было установлено изначально 15 нуклеотидных полиморфизмов, ассоциированных с МПК с уровнем достоверности, достигшим полногеномной значимости. Впоследствии, ассоциации 6 из них были подтверждены, по крайней мере, в одной из групп сравнения (Таб.15).

Таблица 15. Список обнаруженных и подтвержденных ассоциаций генетического статуса и МПК.

Группа обнаружения	Группы подтверждения	Поли-морфизм	Замена	Ген	Аллель вынос-ливости	P-value
Все атлеты	Мужчины-стайеры, мужчины атлеты, женщины атлеты	rs7144481	T/C	*TSHR*	C	$8.92*10^{-8}$
Все атлеты	Женщины-стайеры, мужчины средневики, женщины средневики, мужчины атлеты, женщины атлеты	rs1572312	C/A	*NFIA-AS2*	C	$1.66*10^{-6}$
Мужчины атлеты	Мужчины стайеры, мужчины средневики, женщины средневики	rs12893597	C/T	*ESRRB*	T	$7.44*10^{-6}$
Средневики	Мужчины средневики, женщины средневики	rs9922134	C/T	*CDH13*	C	$8.85*10^{-6}$
Средневики	Мужчины средневики, женщины средневики, женщины атлеты	rs296000	G/T	*SLIT3*	G	$1.66*10^{-5}$
Средневики	Мужчины средневики, женщины средневики, женщины атлеты	rs7191721	G/A	*RBFOX1*	G	$1.73*10^{-5}$

На следующем этапе для валидации полученных 6 генетических вариантов было выполнено ассоциативное исследование в формате эксперимент - контроль на расширенной выборке, включающей 218 российских спортсменов стайеров/средневиков и комбинированной группы сравнения, состоящей из 230

российских спортсменов скоростно-силовых дисциплин, а также 192 и 1367 человек, не занимающихся профессионально спортом из России и Европы, соответственно.

Изначально предполагалось, что аллели полиморфизмов, ассоциированные с выносливостью, будут существенно менее представлены в группах сравнения (спортсмены скоростно-силовых дисциплин и контроль). По результатам проведенного исследования ассоциация с показателями выносливости была подтверждена для 3 из 6 полиморфизмов: *NFIA-AS2* rs1572312 C, *TSHR* rs7144481 C, *RBFOX1* rs7191721 G. Наибольшая степень ассоциации была показана для полиморфизма в гене *NFIA-AS2* - rs1572312.

Так, для аллели С была показана ассоциация с наиболее высоким значением МПК: среди женщин-стайеров: среди женщин - спортсменов длительной выносливости (генотип АС: 54,4 ± 2,5 мл/(мин*кг), генотип СС: 59,9 ± 3,3 мл/(мин*кг); P = 0,0164), среди мужчин – средневиков (АС: 55,3 ± 7,5 мл/(мин*кг), СС: 61,0 ± 4,5 мл/(мин*кг); P = 0,0279), среди женщин - средневиков (АА/АС: 50,1 ± 4,8 мл/(мин*кг), СС: 56,9 ± 5,8 мл/(мин*кг); P = 0,0278), среди всех мужчин-спортсменов (АС: 55,3 ± 7,5 мл/(мин*кг), СС: 64,4 ± 6,4 мл/(мин*кг); P = 0,0051) и среди всех женщин-спортсменов (АА/АС: 51,5 ± 4,5 мл/(мин*кг), СС: 58,5 ± 4,7 мл/(мин*кг); P = 0,0005).

Полигенный анализ трех наиболее значимых полиморфизмов показал, что их совокупная доля объясняемой вариабельности МПК составляет 24,6% у мужчин и 48,8% у женщин. В случае вычисления общего суммарного рейтинга (аддитивная модель) по 3 полиморфизмам среднее значение его было достоверно выше у атлетов-стайеров (58,9 ± 18,2) чем у российской контрольной группы (54,2 ± 15,7, P = 0,0059) или спортсменов скоростно-силовой направленности (53,6 ± 18,4, P = 0,0022). У элитной группы стайеров указанный индекс был еще выше и составил 59,4 ± 17,2 с P = 0,01 при сравнении с российским контролем. Стоит обратить внимание на то, что никто из контрольной группы не обладал всеми 6 (3 полиморфизма) "аллелями выносливости", тогда как 4 из 20 элитных спортсмена - стайера такой комбинацией обладали.

Ген, в котором был выявлен наиболее значимый полиморфизм – *NFIA-AS2*, кодирует антисмысловую РНК, которая, как предполагается, регулирует экспрессию гена *NFIA*, который в свою очередь вовлечен в регуляцию эритропоэза и гранулопоэза, и, как следствие, связан с качественным составом крови у атлетов [104].

В ходе дополнительного анализа было показано, что мужчины, являющиеся носителями генотипа *NFIA-AS2* rs1572312 CC имели гораздо более высокий уровень гемоглобина в крови (161,7 ± 7,8 против 153,2 ± 7,6 г/л, P = 0,0467), уровень эритроцитов в 1 литре (5,17 ± 0,26 · 10^{12} против 4,86 ± 0,38 · 10^{12}, P = 0,0403) и число рекулоцитов в 1 литре (62,7 ± 22,4 · 10^{9} против 30,0 ± 14,1 · 10^{9}, P = 0,0072) чем среди носителей более редкой A аллели. С другой стороны, носители CC гомозиготы у мужчин имели более низкий уровень нейтрофилов на 1 литр (3,17 ± 1,1 · 10^{9} против 5,1 ± 1,8 · 10^{9}, P = 0,0029) и меньший размер лейкоцита и эритроцита (P = 0,0332) нежели носители аллели A. Таким образом, можно предположить, что аллель C полиморфизма rs1572312 гена *NFIA-AS2* ассоциирована с активацией эритропоэза, тогда как аллель A активирует гранулопоэз.

3.3.2 Анализ генетических факторов, ассоциированных со структурой мышечных волокон

Много исследований утверждают, что структура мышц влияет на физические показатели и соревновательные результаты [105] [106]. Спортсмены имеют различные возможности достижения высоких результатов в аэробных и скоростно-силовых видах спорта, что может быть объяснено вариабельностью в структуре мышечных волокон каждого из них. В целом, структура мышечной ткани сама по себе является хорошим маркером функциональных возможностей, но высокая доля медленных волокон ассоциирована с большим объемом митохондрий, более высоким кислородным обменом и высокой устойчивостью к усталости.

Известные кандидатные гены, вероятно детерминирующие структуру волокон кодируют сигнальные пути кальциневрина, альфа- и гамма-рецепторов, активируемых пероксисомными пролифераторами, кальций/кальмодулин-зависимые киназы, деацетилазы гистонов, тироидных гормоны и множество других сопутствующих факторов и метаболитов. В ходе проведенных ранее исследований публикаций были сформированы списки кандидатных полиморфизмов, являющихся наиболее перспективными для поиска ассоциаций. В связи с этим, в рамках нашего исследования был проведен ассоциативный генетический анализ 15 из таких кандидатных полиморфизмов: *ADRB2* G16R (rs1042713) и *Q27E* (rs1042714), *AGT* M235T (rs699), *AGTR1* A1166C (rs5186), *AGTR2* C3123A (rs11091046), *BDKRB2* T58C (rs1799722), *REN* G83A (rs2368564), *AMPD1* Q12X (rs17602729), *CALCR* L447P (rs1801197), *IGF1* CA-repeat (rs10665874), *MYF6* C964T (rs3121), *NOS3* 27-bp repeat, *TSHR* D727E (rs1991517), *VDR* TaqI (rs731236), *VEGFA* -2578 A/C (rs699947).

В рамках исследования было проведено общее генотипирование 3497 образцов с использованием ДНК-чипа Illumina 1M Omni Quad v1.2 или минисеквенирование с использованием технологии MALDI-TOF. Результаты чип-гибридизации (1165 образцов) обрабатывались по разработанному в данной работе протоколу.

Исследование проводилось в два этапа: на первом шаге производилась проверка ассоциаций указанных полиморфизмов со структурой мышц и МПК, тогда как на втором шаге производилась оценка частот встречаемости генотипов в различных группах.

В поисковой части исследования было проведено ассоциативное исследование на группе в 40 человек спортсменов (аэробные возможности) и 55 человек, не являющихся спортсменами, для которых были известны результаты биопсии четырехглавой мышцы бедра. Из 15 исследованных кандидатных полиморфизмов только *AGTR2* rs11091046 A/C показал значимую ассоциацию с составом мышечных волокон.

У носителей С аллеля средний процент волокон I типа составил 54,2%, в то время как у носителей А аллеля – 45,2% (P = 0,003). После коррекции уровня P по методу Бонферрони-Холма P = 0,045, что может считаться достоверным.

Таким образом, носители *AGTR2* С аллеля имеют большую долю медленных мышечных волокон. Суммарно исследованный генотип полиморфизма *AGTR2* может объяснить 15,2% вариабельности типа мышечных волокон в латеральной широкой мышце бедра у мужчин.

Дальнейшая часть исследования проводилась на расширенной группе спортсменов из России и Польши в составе 2178 человек, которые были разделены на группы по различным типам спортивной нагрузки («выносливость», «смешанная», «скоростно-силовая») и различной спортивной успешности («элита», «субэлита», «не элита»).

Частоты исследуемого полиморфизма rs11091046 не отличались у российских и польских спортсменов, поэтому все исследуемые образцы не делились на этнические группы. При сравнении группы выносливости и группы силы аллель С показала значимое различие (мужчины: 62,7% и 51,7%, P=0,0038; женщины: 56,6% и 48,1%, P=0,0169), также и между группой выносливости и контролем (мужчины: 62,7% и 51,0%, P=0,0006). Кроме того, частота аллеля А достоверно различалась между женщинами - спортсменами скоростно-силовой направленности в сравнении с контрольной группой (51,9% и 44,8%, P=0,0069). Также, аллель С является более частой среди мужчин группы "спринтеры" по отношению к мужскому контролю (65,8% и 51,0%, P=0,0037), но для женщин аналогичное утверждение не преодолело поправку Бонферрони. При сравнении носителей аллели А с носителями аллели С (мужчины) отношение шансов быть спортсменом - стайером составило 1,612 (95% доверительный интервал (CI): 1,227-2,119, P = 0,0006). С другой стороны, отношение шансов для женщины, занимающейся скоростно-силовыми видами спорта иметь аллель А (AC + AA) в сравнении с женщинами - контролем было 1,774 (95% CI: 1,267-2,483, P = 0,0007). Кроме того, у женщин было показано, что частота *AGTR2* С аллели была значимо

выше у «элитных» спортсменов к контрольной группе (65,1% к 55,2%, P = 0,0488) или к «неэлитным» спортсменам (65,1% к 52,4%, P = 0,0391).

В еще одной, независимой части исследования, был произведен поиск возможной ассоциации генотипа *AGTR2* с эргоспирометрическими параметрами (абсолютное и относительное потребление кислорода на основе теста VO_2max). Это исследование было выполнено на группе из 40 элитных спортсменов, тренирующих выносливость: гребцы-академисты, гребцы-байдарочники, конькобежцы-многоборцы.

При сравнении ассоциации полиморфизма гена *AGTR2* и МПК у мужчин - стайеров было достоверно показано, что МПК у носителей аллели С выше, чем у носителей аллели А (n=28; 62,3 (4,4) против 57,4 (6,0) мл/мин/кг; P = 0,0197), причем указанный полиморфизм объяснял 19,2% вариабельности МПК у спортсменов.

В ходе проведенной работы впервые были получены данные, демонстрирующие ассоциацию полиморфизма rs11091046 в гене *AGTR2* с составом мышечных волокон, предрасположенностью к занятиям видами спорта на выносливость и аэробными возможностями.

Ангиотензиновый рецептор 2 тип 2 является компонентом ренин-ангиотензиновой системы (PAC, RAS), которая является одним из регуляторов роста и дифференциации скелетных мышц [107] [108]. RAS связана со множеством других метаболических цепей, которые могут также влиять на скелетно-мышечные характеристики [109] [110] [111]. *AGTR2* регулирует эффект ангиотензина 2 при клеточной дифференциации и имеет метаболическую роль в сравнении с сосудистой ролью ангиотензинового рецептора 2 тип 1 (*AGTR1*) [112].

3.3.3 Генетический анализ факторов, ассоциированных с оксидативным стрессом

Оксидативный стресс – это процесс, протекающий преимущественно в клетках организма профессиональных спортсменов при сверхинтенсивных

нагрузках, при котором продукты кислородного обмена нарушают клеточный «редокс-статус», что приводит к повышенному уровню токсичных соединений с кислородом (например, окислов липидов) в клетке и к дальнейшему ускоренному ее разрушению.

Одним из кандидатных генов, вариабельность которого может быть ассоциирована с повреждением мышц является супероксид дисмутаза 2 типа (*SOD2*). Этот ген кодирует фермент – марганцевую супероксиддисмутазу, который катализирует дисмутацию супероксида в кислород и пероксид водорода и защищает мышцы от образующихся высокотоксичных кислородных радикалов. Предполагается, что полиморфизмы этого гена могут влиять на антиоксидантную активность. Так для, полиморфизма rs4880 (C/T) в гене *SOD2* было показано, что генотип TT, скорее всего уменьшает количество активного фермента [113], а в отношении профессиональных спортсменов то, что указанный генотип ассоциирован с повышением уровня креатинкиназы в крови после соревнований, что косвенно свидетельствует о большем повреждении мышц.

В ходе ранее проведенных нами исследований с использованием различных технологических платформ была получена информация по статусу полиморфизма rs4880 гена *SOD2* в больших выборках, представленных российскими и польскими спортсменами. В частности у 914 российских спортсменов, проанализированных с использованием ДНК-чипа Illumina 1M Omni Quad v1.2. После обработки данных с использованием оптимизированного протокола и совместным анализом данных генотипирования по другим российским и зарубежным спортсменам общая исследованная выборка составила 2664 человека, представленная профессиональными спортсменами из России и Польши и 917 человек группы контроля из тех же стран, представленная лицами, не занимающимися спортом.

Задачей текущего этапа исследования была проверка гипотезы о том, что частота генотипа TT полиморфизма rs4880 будет существенно ниже в группе профессиональных спортсменов, (с учетом и без учета их специализации) чем в группе контроля.

В процессе эксперимента требовалось найти ассоциацию указанного генетического маркера с биохимическими показателями (креатинкиназа, креатинин, аланинтрансаминаза, аспартаттрансаминаза, щелочная фосфатаза), а также со склонность к занятиям спортом. Соответственно, после типовой подготовки чиповых данных, производился стандартный поиск корреляций между генотипами rs4880 в различных группах, выделенных по видам спорта, и биохимическими показателями попарно. Показатели сравнивались по алгоритму χ^2, использовалась непараметрическая корреляция Спирмана. Были вычислены P-value и те ассоциации, P-value которых не превышал 0,05 и соответствовал критерию Бонферрони-Холма, были признаны статистически достоверными.

Было обнаружено, что действительно, распределение частот генотипов полиморфизма rs4880 (Ala16Val) гена *SOD2* различно в различных группах спортсменов, а также и по отношению к контрольной группе жителей, не являющихся спортсменами. Распределения генотипов у всех спортсменов и у контрольной группы находились в равновесии Харди-Вайнберга (у спортсменов $\chi^2 = 2,96$, $p = 0,228$; в контрольной группе $\chi^2 = 1,66$, $p = 0,436$).

Результаты анализа показали, что, как и ожидалось, частоты аллели T в группе спорта низкой интенсивности и силовых видов спорта отличались от частоты аллели в контрольной группе, но после коррекции Бонферрони эти ассоциации перестали быть значимыми.

С другой стороны, при сравнении частот генотипов TT, а также аллели T отдельно, между атлетами, занимающимися спортом высокой интенсивности по отношению к силовым атлетам и контрольной группе было обнаружено, что частота генотипа TT у силовых и скоростно-силовых спортсменов ниже: 18,6%, 524 человека по отношению к 25,0% 2140 человек ($p = 0,0076$). Кроме того, частота указанного генотипа ниже и по отношению к спортсменам с низкой интенсивностью нагрузок: 180 человек, 33,9%, $p< 0,0001$.

Была также исследована зависимость биохимических параметров и данного генотипа, где было обнаружено, что у тех, кто является носителем TT генотипа, увеличивается активность креатинкиназы и уровень креатинина, что негативно

сказывается на устойчивости к кислородному стрессу. Было показано, что указанный полиморфизм ассоциирован с активностью креатинкиназы (женщины: CC – 214 (224) ед./л, CT – 288(308) ед./л, TT – 231 (243) ед./л, p = 0,022; мужчины: CC – 340 (392) ед./л, CT – 392 (381) ед./л, TT – 402 (455) ед./л, p = 0,284) в группах III-VII (группы I и II были исключены потому, что имели самый низкий уровень креатинкиназы), а также и с уровнем креатинина (женщины: CC – 83,2 (12,9) мкМ/Л, CT – 84,7 (20,2) мкМ/Л, TT – 87,3 (20,4) мкМ/Л, p = 0,0276; мужчины: CC – 97,2 (14,4) мкМ/Л, CT – 100,8 (17,3) мкМ/Л, TT – 101,2 (15,9) мкМ/Л, p = 0,0135).

Стоит также обратить внимание, что уровни креатинкиназы и креатинина среди элитных спортсменов в любой из спортивных групп ниже, чем у тех, кто не является элитными. Поэтому, исходное утверждение о негативном влиянии полиморфизма на достижения в скоростных видах спорта является статистически достоверным.

3.3.4 Полногеномный ассоциативный анализ общего генетического профиля спортсменов-стайеров

Унификация и стандартизация биоинформатических протоколов обработки данных генотипирования позволяют использовать для проведения полногеномных ассоциативных исследований и валидаций данных, полученных в таких исследованиях, обширные выборки, проанализированные в разных странах на разных технологических платформах. Такая возможность была продемонстрирована нами в исследованиях по анализу генетики выносливости, структуры мышечных волокон, оксидативного стресса. В отличии от гипотез-ориентированного подхода, использованного в предыдущих исследованиях, также представляют интерес и попытки оценить отличия, присущие профессиональным спортсменам на глобальном уровне. С использованием проанализированных данных по генотипированию спортсменов России, такая работа была выполнена в

ходе анализа комбинированной выборки представленной спортсменами других стран, общим числом 4280 человек.

Было выдвинуто предположение, что, возможно, существует какой-либо единый генетический профиль олимпийского элитного спортсмена - стайера. Для поиска общего генетического профиля спортсменов-стайеров использовались геномные данные проектов GENATHLETE [101] (Германия, Финляндия, Канада, США) и ассоциативного исследования атлетов Японии [102]. Для валидации результатов этого исследования использовалось: 1165 спортсменов сборной России, 375 геномов атлетов из США, Германии и Японии, а также 436 жителей из этих же стран, не являющихся спортсменами.

Работа проводилась в два этапа. Сначала был проведен ассоциативный поиск маркеров в результатах проектов GENATHLETE и ассоциативного исследования атлетов Японии. Те полиморфизмы, чье распределение частот не соответствовало равновесию Харди-Вайнберга (условие - $p > 0{,}00001$) были исключены из анализа.

Приблизительно 143000 полиморфизмов прошли все предварительные фильтры. Наибольшую возможность сформировать ассоциацию показал локус 15q23. Пять полиморфизмов показали ассоциированность с p-value от $2{,}61 \times 10^{-5}$ до $7{,}37 \times 10^{-5}$. Пятьдесят полиморфизмов, находящихся в некодируемых областях в окрестности генов *ADRA1A*, *BCL2*, *BCL2L14*, *GLIS3*, *MYL2*, *MYO5B*, показали очень высокую достоверность (P-value $< 5{,}0 \times 10^{-4}$) . Для японской выборки было проведено аналогичное исследование, которое обнаружило 19 полиморфизмов (Табл.16) с P-value $< 5{,}0 \times 10^{-5}$.

В конечном итоге был сформирован список из 45 полиморфизмов (Табл.16), которые были проверены с использованием результатов генотипирования во всех странах консорциума. Некоторые полиморфизмы были заведомо исключены из итогового списка в связи со сцепленным наследованием или региональными этническими особенностями.

Таблица 16. Результаты предварительного поиска полиморфизмов в проекте GENEATHLETE и в ассоциативном исследовании японских атлетов.

Результаты GENEATHLETE					
Полиморфизм	Хромосома	Ген	Полиморфизм	Хромосома	Ген
rs17459307	1	*NOS1AP*	rs12821816	12	*NDUFA12*
rs6726044	2	*BMP10*	rs61940911	12	*ANKRD13A*
rs2118908	2	*ACOXL*	rs73195844	12	*CCDC63*
rs6799372	3	*CNTN3*	rs9543114	13	*DIS3*
rs10938202	4	*ATP8A1*	rs9301108	13	*DAOA*
rs4699824	4	*PPP3CA*	rs214003	14	*NRXN3*
rs10499127	6	*NKAIN2*	rs4776471	15	*RPLP1*
rs17055965	8	*ADRA1A*	rs11856981	15	*RPLP1*
rs7861665	9	*LOC101929330*	rs4288991	16	*TOX3*
rs17054974	9	*SEMA4D*	rs8065364	17	*CARD14*
rs12573965	11	*KCNQ1*	rs578211	18	*MYO5B*
rs7947391	11	*NPAS4*	rs4808571	19	*MYO9B*
rs11613185	12	*BCL2L14*	rs62135557	19	*MAU2*

Результаты Японского исследования					
Полиморфизм	Хромосома	Ген	Полиморфизм	Хромосома	Ген
rs10874242	1	*LPHN2*	rs9355947	6	*PARK2*
rs12047209	1	*AKT3*	rs6959675	7	*CFAP69*
rs921665	2	*TSSC1*	rs11975386	7	*BET1*
rs2694093	2	*TSSC1*	rs16906888	8	*FAM135B*
rs6548153	2	*TSSC1*	rs3780169	9	*PAX5*
rs2361506	2	*MROH2A*	rs17690338	10	*ZNF503-AS1*
rs7650685	3	*VGLL4*	rs2761291	10	*MYOF*
rs10007111	4	*MEPE*	rs9580890	13	*SPATA13*
rs558129	4	*GALNTL6*	rs4541108	17	*RBFOX3*
rs2910756	5	*GDNF-AS1*			

Тестирование указанного списка проводилось в 7 странах мира (не входящих в GENEATHLETE), в том числе и в России в рамках данной работы на существующей выборке российских элитных спортсменов – стайеров (234 человека). Ни один из указанных полиморфизмов не достиг отметки P-value < $5{,}0 \times 10^{-4}$ более чем в 2х странах консорциума, что могло бы послужить указателем на единый генетический профиль выносливости.

Возможно, что количество элитных спортсменов было недостаточно для преодоления поставленного порога P-value и при вовлечении в исследование большего числа спортсменов достоверные ассоциации были бы обнаружены.

На основании проведенного анализа допустимо сделать вывод о том, что, хотя выносливость у элитных спортсменов однозначно связана с сердечно-сосудистой системой и аэробным дыханием, на данный момент не обнаружено какого-либо набора полиморфизмов среди всех сборных мира, который мог бы достоверно соответствовать статусу элитного или даже не элитного стайера.

3.4 Медицинская аннотация индивидуальных особенностей на основании данных генетических исследований

Информация об ассоциациях, получаемая в результате групповых исследований является ценным источником данных, позволяющих ближе подступить к проблеме персонификации генетического риска. Так как активно развивающиеся технологии генетического анализа, применяемые ранее только для ассоциативных исследований, выходят на рынок персональных генетических услуг, то остро встает вопрос разработки аннотационных инструментов и алгоритмов, позволяющих осуществлять медицинскую аннотацию генетических данных, информативных для конкретного индивидуума.

В рамках настоящей работы был разработан информационный ресурс, позволяющий осуществлять аннотацию результатов генетического анализа, основываясь на информации, полученной из общедоступных баз данных. Кроме этого, были апробированы алгоритмы оценки генетического риска, основанные на совокупном анализе множества маркеров.

3.4.1 Разработка аннотационного инструмента по известным клинически-значимым полиморфизмам ДНК

Для оптимизации поиска клинических статусов полученных полиморфизмов в описанных ранее открытых источниках данных была создана собственная база данных и программа, позволяющая осуществлять медицинскую аннотацию генетических данных. Наполнение базы данных осуществлялось из

основных открытых геномных проектов (GWAS-catalog [15], LOVD [73], OMIM [75], SNPedia [74], PharmGKB [77], ClinVar [81]). Для наполнения были разработаны специальные программные инструменты, которые позволяли конвертировать информацию из всех указанных источников, используя web-интерфейс или API (при его наличии), а затем размещать ее в собственной базе данных. Кроме того, для баз данных, которые имеют API, SQL или текстовый доступ, при написании программных модулей была предусмотрена возможность автоматического обновления.

Для упрощения дальнейшей классификации полиморфизмов по значимости был разработан протокол назначения классификационных меток.

Все загружаемые записи об ассоциациях были условно отнесены к одному из следующих типов:

Тип 1 - Полиморфизмы, значительно ассоциированные с заболеванием. К этому типу относятся моногенные мутации, пенетрантность которых составляет 100% и явно доказана в нескольких публикациях. В случае, если в научном сообществе ведется дискуссия и присутствуют сомнения в полной пенетрантности, то ассоциации переносятся в другие типы. Основной источник таких данных – базы OMIM и LOVD.

Тип 2 - Полиморфизмы, сцепленные с количественными изменениями каких-либо биохимических показателей. Данные этого типа находились исключительно в результате ассоциативных исследований, поэтому они присутствовали только в базе GWAS-catalog.

Тип 3 - Полиморфизмы, с отношением шансов (OR) превышающим 5. Вероятность возникновения заболевания при их наличии заведомо высокая, но в небольшом числе случаев она может не развиваться.

Тип 4 - Полиморфизмы, с отношением шансов (OR) ниже 5. Вероятность возникновения заболевания при их наличии более низкая, также к ним могут относиться и протективные состояния. Основной источник данных типа 3 и типа 4 – GWAS-catalog, но они также встречаются в ClinVar и SNPedia.

Тип 5 – Редкие мутации, для которых не изучена пенетрантность. Такие полиморфизмы являются результатом исследований одного пациента или малочисленной группы пациентов, поэтому они были обнаружены в базах данных LOVD, OMIM, PharmGKB.

Тип 6 – Редкие полиморфизмы или мутации, которые были получены в результате малочисленных ассоциативных исследований (GWAS) или для которых использовалась редкая этническая выборка.

Тип 7 - Полиморфизмы, о которых известно лишь качественное влияние на признак без каких-либо количественных значений. Обычно об этом влиянии упоминается в публикациях как "выше" или "ниже". Такие полиморфизмы часто являются результатом изучения метаболических цепей у одного или небольшой группы пациентов с одинаковым и очень редким фенотипом. Такие ассоциации возможно обнаружить в любой из загруженных баз данных.

При загрузке ассоциаций типа 2 и типа 3 (имеющих численные значения) автоматически вычислялись относительные риски для каждого из возможных генотипов согласно классической мультипликативной модели для единичного полиморфизма (см.п.1.5.3. литературного обзора).

Для оценки информационной достоверности был применен следующий критерий. Для каждого утверждения об ассоциации полиморфизма и фенотипа производился поиск информации о суммарном общем числе пациентов и контроля, участвовавших при создании всех публикаций, а затем указанное число перемножалось с импакт-фактором журнала, делилось на 1000 и округлялось до целого. Результат был принят в качестве рейтинга доверия к ассоциации.

Одной из целей нашей работы была возможность предоставления результатов работы программы медицинским специалистам, которые, зачастую недостаточно понимают результаты биоинформатической обработки данных, поэтому для адаптации конечных результатов требовалось сгруппировать различные ассоциации по названиям заболеваний или фенотипов, а также найти их русскоязычные аналоги.

Для автоматического поиска синонимов одного и того же заболевания использовались данные из проекта UMLS [114], который содержит информацию о заболеваниях, описания, коды ICD-10, а также всевозможные мировые синонимы названий на различных языках. С помощью разработанного программного пакета для каждой ассоциации было проведено присвоение русского синонима и кода по международному классификатору ICD-10 (International Classification of Diseases 10^{th} Revision) [115]. В России прямым аналогом ICD-10 является база МКБ10 (Международный Классификатор Болезней), которая является дополненным переводом ICD-10. В результате всех операций были сформированы группы синонимичных наименований заболеваний, частично снабженные кодами ICD-10, описаниями и переводами на русский язык.

Условно максимальный набор полиморфизмов, входящих в базу, можно разделить на 25 нозологических групп, в пределах которых можно провести еще более детализированную классификацию. Среди основных нозологических групп можно выделить следующие: злокачественные опухоли, доброкачественные опухоли, психические и неврологические расстройства, нейродегенеративные заболевания, расстройства мозгового кровообращения, сердечно-сосудистые заболевания, бронхолегочные заболевания, болезни костно-мышечной системы и соединительной ткани, заболевания крови, болезни органов зрения, болезни органов слуха, болезни обмена веществ, заболевания желудочно-кишечного тракта, заболевания печени, заболевания почек, заболевания эндокринной системы, дефициты ферментов, аллергические и иммунологические заболевания, аутоиммунные заболевания, расстройства репродуктивной системы, кожные болезни, инфекционные заболевания, болезни зависимостей, особенности физического развития и фармакогенетически информативные маркеры.

Для ускорения загрузки и обработки списков полиморфизмов и мутаций, а также литературных источников и названий заболеваний была создана специальная структура в реляционной базе данных (в качестве среды использовался MySQL-server 5.7).

Таблицы полиморфизмов, генов, публикаций и блок таблиц по заболеваниям являются отдельными единицами, каждая из которых имеет свой уникальный первичный ключ (Рис.11). Основной таблицей в созданной базе данных являлась информация об ассоциациях в виде связок ключей из других таблиц (полиморфизм – источник – заболевание – база данных), а также сопутствующая информация, связанная исключительно с ассоциацией (уровень риска, количество пациентов в исследованиях, индекс информационной достоверности). Статистические данные по структуре разработанной базы данных, актуальные на 01.09.2016, указаны в табл.17.

Таблица 17. Статистические данные по структуре разработанной базы данных.

Всего ассоциаций	221543
Уникальных публикаций	49558
Уникальных полиморфизмов	162428
Уникальных заболеваний	3121
Синонимичных заболеваний	10511

Рисунок 11. Схема структуры реляционной базы данных для оценки индивидуальных рисков развития заболеваний.

Для автоматизации обработки данных, полученных в ходе генетического исследования, был написан программный модуль, принимающий в качестве входных данных как VCF-файлы результатов анализа данных секвенирования, полученные на любых платформах, так и файлы формата 23andMe, которые являются стандартными для результатов чип-гибридизации.

Формат выдачи результатов поиска записей в базе данных предусматривает только выдачу данных об ассоциациях без какой-либо последующей групповой обработки. Исходя из того, что вывод должен быть максимально достоверным, было принято решение о замене численных значений на текстовые записи согласно таблице 18.

Таблица 18. Соответствия уровней относительного риска и текстовых обозначений.

Риск < 1	Низкий
Риск = 1	Нормальный
1 < Риск < 2	Повышенный
2 < Риск < 4	Высокий
4 < Риск	Очень высокий

Все описанные программные модули были собраны в единый протокол, способный из данных секвенирования (fastq) или чип-гибридизации (23andMe) получать как список полиморфизмов, отличных от референсного генома с описанием их биологической значимости программным пакетом AnnoVar, так и список всех полиморфизмов, когда-либо принимавших участие в тех или иных исследованиях с их рисковыми статусами. Общая схема работы итогового протокола представлена на рис.12.

Для того, чтобы результаты работы программного протокола были удобными для применения врачами в медицинской практике, была разработана удобная HTML-форма выдачи (Рис.13), которая позволяет в табличной форме и с

использованием цветовых схем обратить внимание пользователя на те или иные полиморфизмы или состояния.

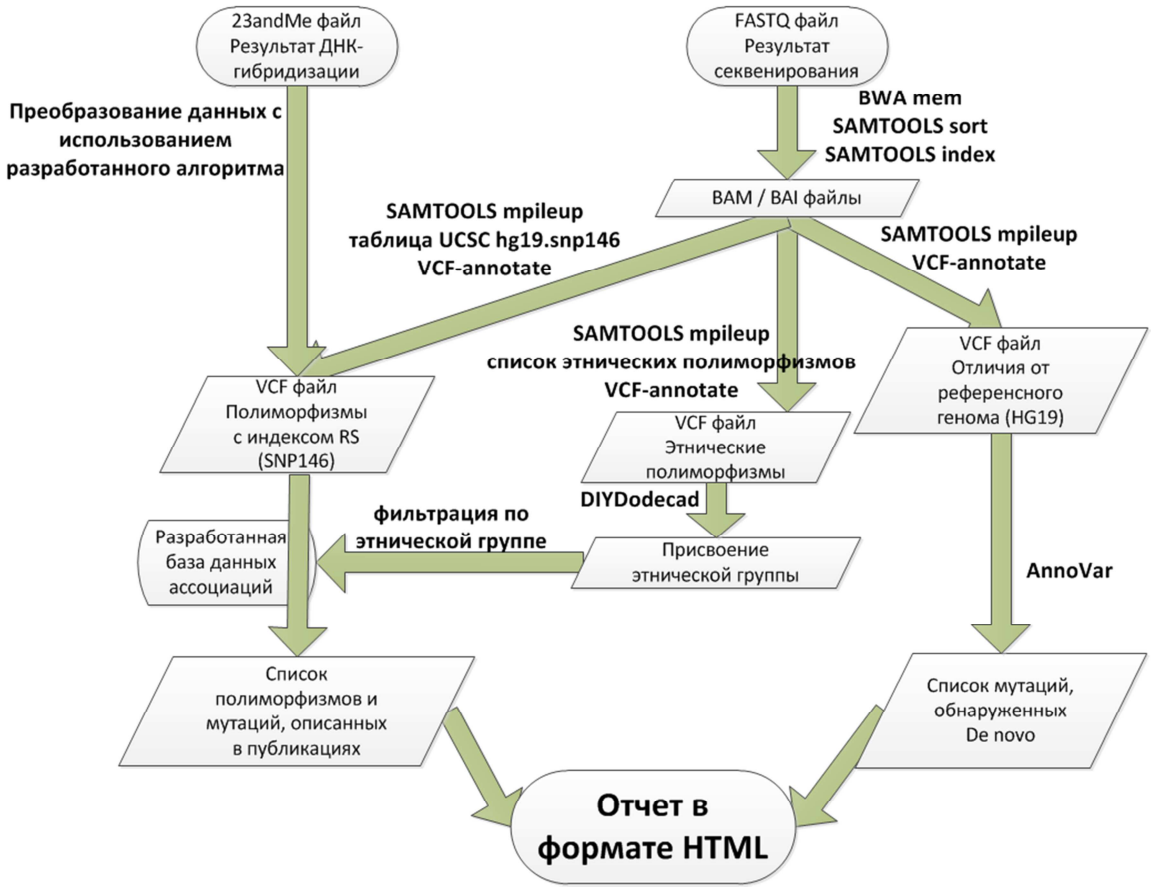

Рисунок 12. Схема организации разработанного программного продукта.

Таким образом, разработанный аннотационный инструмент, основанный на использовании собственной локальной базы данных по нуклеотидным полиморфизмам, ассоциированным с различными медицинскими заболеваниями и реализующий простой язык запросов позволяет выполнять простую и логичную аннотацию данных генотипирования. При этом осуществляется группировка выявленных вариантов по нозологическому принципу с сортировкой по величине, ассоциированного риска. Это позволяет быстро и удобно оценить характер и диапазон вариабельности в пределах нозологических групп, представляющих наибольший интерес для анализа. Следует отдельно отметить то, что разработанный аннотационный инструмент позволяет работать на вход как с данными высокопроизводительного секвенирования (в формате fastq), так и полученных при использовании ДНК-чипов (в формате 23andMe).

Список генетических полиморфизмов, ассоциированных с медицинскими показателями.

Хроническая обструктивная болезнь легких

Заболеваемость составляет 50 на 1000 у мужчин.

Хроническая обструктивная болезнь лёгких — самостоятельное заболевание, для которого характерно частично необратимое ограничение воздушного потока в дыхательных путях. Ограничение воздушного потока, как правило, имеет неуклонно прогрессирующий характер и спровоцировано аномальной воспалительной реакцией ткани лёгких на раздражение различными патогенными частицами и газами.

Полиморфизм	Позиция hg19	Ген	Генотип	Риск	Частота	Публикации
rs41511151	chr7:73482987	ELN	GG	Нейтральный	99.2%	PUBMED: 16081882

Аутизм

Полиморфизм	Позиция hg19	Ген	Генотип	Риск	Частота	Публикации
rs849563	chr2:206610502	NRP2	GG	Высокий	6%	PUBMED: 17427189
rs4150167	chr16:84213684	TAF1C	TC	Повышенный	1.7%	PUBMED: 20663923
rs149653283	chrX:148048494	AFF2	AA	Нейтральный	100%	PUBMED: 22773736
rs140927355	chrX:148044334	AFF2	GG	Нейтральный		PUBMED: 22773736
rs121917893	chrX:70387358	NLGN3	CC	Нейтральный		PUBMED: 12669065
rs147324438	chrX:23397873	PTCHD1	AA	Нейтральный		PUBMED: 20844286
rs201108123	chr22:19799963	GNB1L	GG	Нейтральный		PUBMED: 22095694
rs138483899	chr22:19776348	GNB1L	CC	Нейтральный		PUBMED: 22095694
rs201867922	chr12:66838406	GRIP1	TT	Нейтральный		PUBMED: 21383172
rs202029416	chr12:66800174	GRIP1	CC	Нейтральный		PUBMED: 21383172
rs144494437	chr12:66786171	GRIP1	AA	Нейтральный	99.5%	PUBMED: 21383172
rs201292141	chr11:115088681	CADM1	TT	Нейтральный		PUBMED: 18957284
rs121909239	chr10:89717730	PTEN	AA	Нейтральный		PUBMED: 15805158
rs121909240	chr10:89717697	PTEN	TT	Нейтральный		PUBMED: 15805158
rs121909238	chr10:89692794	PTEN	AA	Нейтральный		PUBMED: 15805158
rs121908445	chr7:147844634	CNTNAP2	TT	Нейтральный		PUBMED: 18179895
rs200604893	chr2:50149369	nrxn1b	CC	Нейтральный		PUBMED: 22504536
rs115466046	chr1:160012270	KCNJ10	CC	Нейтральный		PUBMED: 21458570
rs140646329	chr1:160012073	KCNJ10	CC	Нейтральный		PUBMED: 21458570
rs2297909	chr1:200960307	NR	GG	Нейтральный	56.2%	PUBMED: 23453885
rs6537825	chr1:114948281	TRIM33, BCAS2, DENND2C	AG	Нейтральный	27%	PUBMED: 24189344

Аксональная нейропатия

Полиморфизм	Позиция hg19	Ген	Генотип	Риск	Частота	Публикации
rs119103267	chr1:12069698	MFN2	CC	Нейтральный		PUBMED: 18458227

Синдром дефицита внимания и гиперактивности

Полиморфизм	Позиция hg19	Ген	Генотип	Риск	Частота	Публикации
rs11700812	chr21:44480590	CBS	CC	Высокий		PUBMED: 22267502
rs1057868	chr7:75615006	POR	TC	Повышенный	41%	PUBMED: 16467261
rs4148301	chr4:70462042	UGT2A1	TC	Повышенный	14.1%	PUBMED: 21164388

Рисунок 13. Фрагмент списка обнаруженных у пациента ассоциаций в формате HTML.

3.4.2 Разработка релевантной модели прогнозирования состава мышечных волокон

Результаты работы разработанного программного продукта оказывают значительную пользу при исследовании индивидуальных особенностей пациента. Тем не менее, такой аннотационный продукт формирует лишь промежуточный результат, с медицинской точки зрения. Если обратить внимание на блок с любой из мультифакторных нозологий, то можно заметить, что для каждого пациента в списке будет значительное число известных ассоциаций, некоторые из которых будут рисковыми, а другие – протективными. На основании таких результатов утверждать о рисковом статусе данного пациента затруднительно. Попытки разработать мультилокусные модели оценки генетического риска предпринимаются с того момента, как высокоплотные генетические исследования стали более доступными дли персонального применения. Несмотря на всю неоднозначность предпринимаемых подходов и различную успешность, многие из таких моделей показывают приемлемые показатели точности [98] [116].

Для применения результатов проведенных ассоциативных исследований по структуре мышц была произведена попытка создания релевантной модели, прогнозирующей состав мышечных волокон спортсмена на основании комплексного анализа известных полиморфизмов. Максимальное количество известных полиморфизмов, ассоциированных со структурой мышц, было описано в одной из публикаций, изданной И.И.Ахметовым в 2012 году (Табл.19) [108].

Таблица 19. Список полиморфизмов, ассоциированных с мышечной структурой.

Полиморфизм	Ген	Полиморфизм	Ген
rs1815739	*ACTN3*	rs1937	TFAM
rs11549465	*HIF1A*	rs660339	UCP2
rs2229309	*NFATC4*	rs1800849	UCP3
rs4253778	*PPARA*	rs2010963	VEGFA
rs2016520	*PPARD*	rs1870377	KDR
rs8192678	*PPARGC1A*	rs4646994	ACE
rs7732671	*PPARGC1B*	rs11091046	AGTR2

К сожалению, до сих пор не известно ни количественных соотношений, ни степени совместного зависимого наследования полиморфизмов из списка. Авторами указанной публикации было выдвинуто предположение, что степень влияния всех полиморфизмов из списка примерно одинаковая, а общий вклад генетики в мышечную структуру составляет порядка 50%. На основании этого утверждения допустимо предположить, что наилучший результат при оценке структуры мышц в процентном соотношении может дать использование закрытой аддитивной модели [93].

Каждый из описанных полиморфизмов является результатом ассоциативного исследования и имеет 3 состояния: с преобладанием 1-ого типа мышц, промежуточное и с преобладанием 2-ого типа мышц. Пусть каждой из аллелей, приводящей к повышению числа мышечных волокон 1-ого типа, будет присвоен весовой индекс равный 1. Общий рисковый индекс полиморфизма R будет равен сумме весовых индексов по обеим аллелям (0, 1, 2). На основании этого, максимальный вклад генетики в развитие мышц 1ого типа наступит при наборе 28 суммарных индексов (на 14 полиморфизмов), а 0 результат суммирования будет означать развитие мышц скорее 2ого типа, нежели 1ого. Суммарный рисковый индекс $R_{общ}$ будет равен простой сумме рисковых индексов по всем полиморфизмам модели.

$R_i = n$ – рисковый индекс равен числу ассоциированных аллелей с 1 типом.

$R_{общ} = \sum_i R_i$ – общий рисковый индекс равен сумме рисковых индексов.

Так как результат вычисления рискового индекса является лишь оценочным, то был произведен переход от числовой модели к статусам, определяющим общее соотношение или преобладание волокон определенного типа.

Проверка полученной модели проводилась на 55 испытуемых (30 спортсменов, 25 контроль), для которых имелись результаты биопсии мышцы бедра. Для всех образцов данной группы проводилось сравнение результата

работы модели и реального статуса на основании мышечной биопсии. В случае, если результат совпадал, то считалось, что тест пройден верно, иначе – ошибочно. Для построения ROC-кривой производилось варьирование критерия, по которому определялся преобладающий тип мышечной структуры (Рис.14).

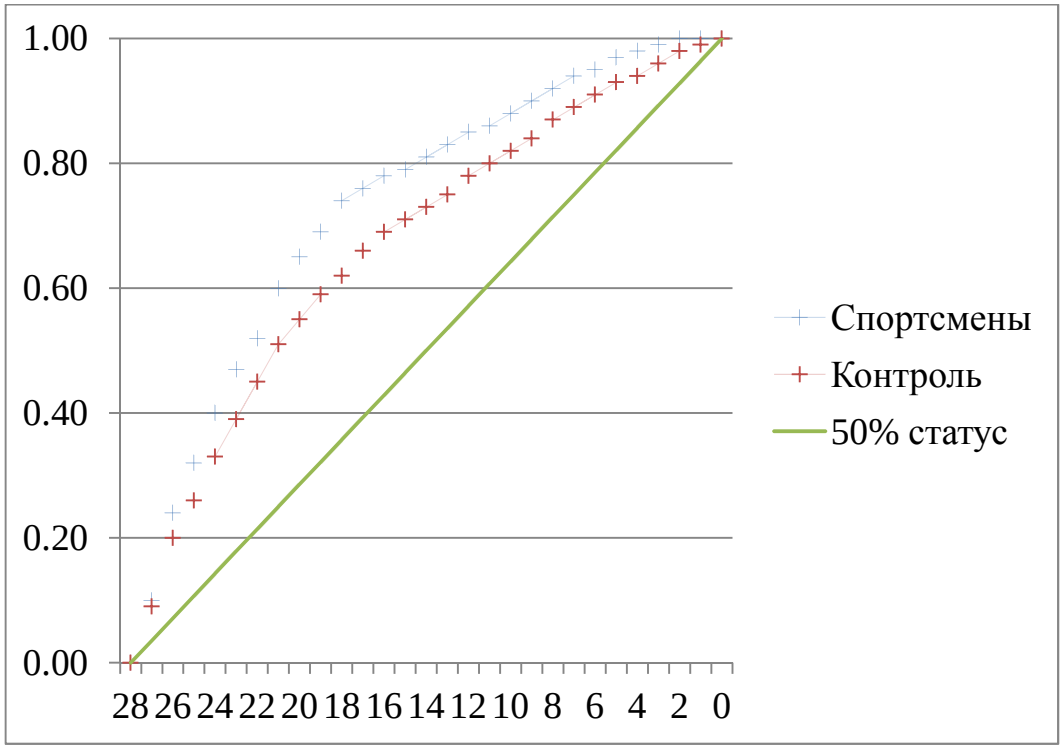

Рисунок 14. Зависимость числа образцов, отнесенных тестом к группе обладателей мышечной структурой 1 типа, от численного значения порога определения мышечного типа (ROC-кривые) для модели прогнозирования мышечной структуры у спортсменов и в контрольной группе.

На основании теста точность работы модели (AUC) составила 81% для профессиональных спортсменов и 73% для лиц, не являющихся спортсменами. Наибольшая точность достигалась при следующих критериях отнесения к одной из групп: 0-4 балла означали значительное преобладание волокон 2-ого типа, 5 – 10 баллов - преобладание волокон 2-ого типа, 11 – 17 баллов – равномерную структуру мышц, 18 – 24 балла - преобладание волокон 1-ого типа, более 25 баллов – значительное преобладание волокон 1-ого типа.

Можно предположить, что небольшая разница между группами возникает вследствие существенного влияния на развитие мышечной структуры регулярных занятий спортом, что напрямую подтверждает утверждение о 50% величине вклада генетических факторов в структуру мышц. Полученные итоговые результаты показывают, что даже простейшие аддитивные генетические модели такого рода допустимо применять для формирования индивидуальных рекомендаций как спортсменам (по выбору оптимальной дистанции на соревнованиях), так и обычным пациентам (по оптимизации занятий спортом).

Данный подход, как и многие исследования последних лет, все больше говорит о перспективности моделирования персонального риска мультифакториальных состояний на основании индивидуальных генетических данных. Так, хорошие результаты были получены по данным трехлетнего пилотного проекта по анализу генетической предрасположенности к раку молочной железы и раку предстательной железы: CGEMS (Cancer Genetics Markers of Susceptibility) [117]; на данных исследований, финансируемых консорциумом Wellcome Trust Case Control Consortium по болезни коронарных сосудов, болезни Крона, ревматоидному артриту, диабету 2 типа [118].

Полученные значения площади под характеристической кривой диагностического теста (AUC) – ROC (Receiver Operating Characteristic) при проверке на вышеупомянутых генетических данных составляли в диапазоне 0.69-0.77. Причем эти характеристики были получены без учета данных семейного анамнеза, учета воздействия внешних факторов и других клинически значимых характеристик. С учетом того, что большинство используемых на сегодняшний день клинических негенетических тестов, таких как анализ ПСА, тредмил-тест, маммография имеют показатели AUC в диапазоне (0,62-0,85), можно констатировать перспективность предлагаемого способа расчета генетического риска для практического использования.

Заключение

В настоящее время очень остро стоит вопрос обработки данных, которые возможно получать с помощью современных геномных технологий и использовать, впоследствии, для изучения способностей групп людей или даже для постановки диагнозов и лечения конкретного пациента. К сожалению, физика и химия значительно обгоняют в прогрессе биоинформатические технологии.

В данной работе был разработан протокол получения геномной информации из результатов ионного полупроводникового секвенирования и чип-гибридизации. Для данных полупроводникового секвенирования впервые была достигнута точность в 99,66%, что равно точности определения полиморфизмов с помощью чип-гибридизации. Для чип-гибридизации впервые разработан протокол автоматической адаптации выдачи результатов по положительной или отрицательной цепям в зависимости от используемой сборки генома. Указанные разработки позволят использовать обе описанные технологии наиболее полно при различных исследованиях. Кроме того, была создана база данных полиморфизмов и мутаций, являющаяся собирательной из 6 открытых баз данных с адаптацией к публикационному описанию генетической информации одного исследуемого пациента.

Все указанные протоколы были применены на практике к более, чем 1200 спортсменам различных сборных России, и впервые был сделан вывод о корреляции некоторых полиморфизмов с мышечной структурой и с максимальным потреблением кислорода в клетках спортсменов. Кроме того, было показано, что не существует единого генетического профиля спортсмена - стайера. Полученные результаты могут быть использованы для адаптации тренировок к конкретным спортсменам с целью улучшения их достижений.

Выводы

1) Точность разработанного альтернативного протокола обработки данных ионного полупроводникового секвенирования выше штатного на 0,5%, что позволяет при полноэкзомном формате исследования дополнительно корректно идентифицировать около 270 тысяч нуклеотидных позиций.

2) Разработанный алгоритм обработки данных высокоплотной чип-гибридизации позволяет исключать ошибки, связанные с неправильным позиционированием генетических вариаций и предоставляет возможность пользовательской настройки выходных параметров качества идентификации вариантов.

3) Полиморфизм rs1572312 гена *NFIA-AS2* для представителей европейского этноса достоверно ассоциирован с уровнем гемоглобина в крови и, как следствие, с аэробными возможностями спортсменов.

4) Полиморфизм rs11091046 гена *AGTR2* для представителей европейского этноса достоверно ассоциирован со структурой мышечных волокон и показателями потребления кислорода.

5) Полиморфизм rs4880 гена *SOD2* для представителей европейского этноса влияет на устойчивость организма спортсмена к оксидативному стрессу, а также ассоциирован с изменением уровня креатинкиназы и креатинина.

6) В настоящее время не выявлено единого для большинства популяций генетического профиля элитного спортсмена-стайера.

7) Разработанный программный продукт позволяет осуществлять индивидуальную аннотацию медицински информативных генетических вариантов, полученных различными методами высокопроизводительного генотипирования.

8) Точность разработанного прототипа модели индивидуальной комплексной оценки структуры мышечных волокон составила 81% в группах профессиональных спортсменов.

Список литературы

1. Willet CE,Wade CM. *From the phenotype to the genotype via bioinformatics* // Methods Mol Biol. 2014. V. 1168. P. 1-16 (DOI: 10.1007/978-1-4939-0847-9_1).

2. Maffulli N, Margiotti K, Longo UG et al. *The genetics of sports injuries and athletic performance* // Muscles Ligaments Tendons J. 2013. V. 3. P. 173-189 (PMID: 24367777).

3. Manolio TA,Guttmacher AE. *Genomewide association studies and assessment of the risk of disease* // N. Engl. J. Med. 2010. V. 363. P. 166–176 (DOI: 10.1056/NEJMra0905980).

4. Meldrum C, Doyle MA, Tothill RW. *Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective* // Clin Biochem Rev. 2011. V. 32. P. 177-195 (PMID: 22147957).

5. Ahmetov II,Fedotovskaya ON. *Current Progress in Sports Genomics* // Adv Clin Chem. 2015. V. 70. P. 247-314 (DOI: 10.1016/bs.acc.2015.03.003).

6. Несговорова Г.П.. *Биоинформатика: Пути Развития и Перспективы* // Сборник научных трудов института систем информатики им. А.П. Ершова. СО РАН. 2012. P. 71-89.

7. Saviotti PP, deLooze MA, Michelland S et al. *The changing marketplace of bioinformatics* // Nature Biotechnology. 2000. V. 18. P. 1247 - 1249 (DOI: 10.1038/82351).

8. Bannon J. Heading for $100: The Declining Costs of Genome Sequencing & The Consequences // [ARK invest] URL: https://ark-invest.com/research/genome-sequencing.

9. Smerecnik CM, Mesters I, de Vries NK et al. *Educating the general public about multifactorial genetic disease: applying a theory-based framework to understand current public knowledge.* // Genet Med. 2008. V. 10. P. 251-258 (DOI: 10.1097/GIM.0b013e31816b4ffd).

10. Luger K, Mäder AW, Richmond RK et al. *Crystal structure of the nucleosome core particle at 2.8 A resolution* // Nature. 1997. V. 429. P. 251-260 (DOI: 10.1038/38444.).

11. Sachidanandam R, Weissman D, Schmidt SC et al. *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms* // Nature. 2001. V. 409. P. 928–933 (DOI:10.1038/35057149).
12. National Institute for Biotechnological Information. Human Genome Statistics (NCBI) // [National Institute for Biotechnological Information] URL: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606.
13. McEntyre J,Ostell J *The NCBI Handbook* — Bethesda (MD): National Center for Biotechnology Information (US), 2015 (Chapter 5).
14. Manolio TA. *Bringing genome-wide association findings into clinical use* // Nature Reviews Genetics. 2013. V. 14. P. 549-558 (DOI: 10.1038/nrg3523).
15. GWAS Catalog: The NHGRI-EBI Catalog of published genome-wide association studies // [сайт] URL: https://www.ebi.ac.uk/gwas.
16. Boucher C,Sandford R. *Autosomal dominant polycystic kidney disease (ADPKD, MIM 173900, PKD1 and PKD2 genes, protein products known as polycystin-1 and polycystin-2)* // Eur J Hum Genet. 2004. V. 12. P. 347-354 (DOI: 10.1038/sj.ejhg.5201162).
17. Igarashi P,Somlo S. *Genetics and pathogenesis of polycystic kidney disease* // J Am Soc Nephrol. 2002. V. 13. P. 2384-2398 (PMID: 12191984).
18. Buitendijk GH, Rochtchina E, Myers C et al. *Prediction of age-related macular degeneration in the general population: the Three Continent AMD Consortium.* // Ophthalmology. 2013. V. 120. P. 2544-2655 (DOI:10.1016/j.ophtha.2013.07.053).
19. Солодков А.С.,Сологуб Е.Б. *Физиология спорта* Россия — Санкт-Петербург: СПбГАФК им. П. Ф. Лесгафта, 2003. — 373 с.
20. Freeman W, Williams C, Nute MG. *Endurance running performance in athletes with asthma* // J Sports Sci. 1990. V. 8. P. 103-117 (DOI: 10.1080/02640419008732138).
21. Christman AL, Lazo M, Clark JM et al. *Low Glycated Hemoglobin and Liver Disease in the U.S. Population* // Diabetes Care. 2011. V. 34. P. 2548-2550 (DOI: 10.2337/dc11-0944).

22. Winkler TW, Justice AE, Graff M et al. *The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study.* // PLoS Genet. 2015. V. 11. (DOI: 10.1371/journal.pgen.1005378).
23. Беловол А.Н., Князькова И.И., Гридасова Л.Н.. *Диагностика хронической сердечной недостаточностиу пациентов с хронической обструктивной болезнью легких* // Научные ведомости Белгородского государственного университета. Серия: Медицина. Фармация. 2014. V. 24. №28, P. 17.
24. Wilkie DR *Muscle* The UK — London: Edward Arnold, 1976. — 15-18 c.
25. Ross JS,Cronin M. *Whole cancer genome sequencing by next-generation methods* // Am J Clin Pathol. 2011. V. 136. P. 527-539 (DOI: 10.1309/AJCPR1SVT1VHUGXW).
26. Venter JC, Adams MD, Myers EW et al. *The sequence of the human genome* // Science. 2001. V. 291. P. 1304-1351 (PMID: 11181995).
27. Ledford H. *All about Craig: the first 'full' genome sequence* // Proc. Natl. Acad. Sci. USA. V. 74. P. 560-564 (DOI: 10.1073/pnas.74.2.560).
28. Koboldt DC, Ding L, Mardis ER et al. *Challenges of sequencing human genomes* // Brief Bioinform. 2010. №11, P. 484-498 (DOI: 10.1093/bib/bbq016).
29. Staden R. *A strategy of DNA sequencing employing computer programs* // Nucleic Acids Research. 1979. V. 6. P. 2601-2610 (DOI: 10.1093/nar/6.7.2601).
30. Ajay SS, Parker SC, Abaan HO et al. *Accurate and comprehensive sequencing of personal genomes* // Genome Res. 2011. V. 21. P. 1498–1505 (DOI: 10.1101/gr.123638.111).
31. Guan YF, Li GR, Wang RJ et al. *Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer* // Clin J Cancer. 2012. V. 31. P. 463-470 (DOI: 10.5732/cjc.012.10216).
32. Pleasance ED, Cheetham RK, Stephens PJ et al. *A comprehensive catalogue of somatic mutations from a human cancer genome* // Nature. 2009. V. 463. P. 191-196 (DOI: 10.1038/nature08658).
33. Vrijenhoek T, Kraaijeveld K, Elferink M et al. *Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects* // European Journal of Human Genetics. 2015. V. 23. P. 1142-1150 (DOI: 10.1038/ejhg.2014.279).

34. Samorodnitsky E, Jewel BM, Hagopian R et al. *Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing* // Hum Mutat. 2015. V. 36. P. 903-914 (DOI: 10.1002/humu.22825).
35. Smith LM, Sanders JZ, Kaiser RJ et al. *Fluorescence detection in automated DNA sequence analysis* // Nature. 1986. V. 321. P. 674-679 (DOI: 10.1038/321674a0).
36. Boland JF, Chung CC, Roberson D et al. *The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing* // Hum Genet. 2013. V. 132. P. 1153-1163 (DOI: 10.1007/s00439-013-1321-4).
37. Cock PJ, Fields CJ, Goto N et al. *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants* // Nucleic Acids Res. 2010. V. 38. P. 1767–1771 (DOI: 10.1093/nar/gkp1137).
38. Del Fabbro C, Scalabrin S, Morgante M et al. *An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis* // PLoS One. 2013. V. 8. (DOI: 10.1371/journal.pone.0085024).
39. Bolger AM, Lohse M, Usadel B. *Trimmomatic: a flexible trimmer for Illumina sequence data* // Bioinformatics. 2014. V. 30. P. 2114–2120 (DOI: 10.1093/bioinformatics/btu170).
40. Bragg LM, Stone G, Butler MK et al. *Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data* // PLoS Comput Biol. 2013. V. 9. (DOI: 10.1371/journal.pcbi.1003031).
41. Li H, Handsaker B, Wysoker A et al. *The Sequence Alignment/Map format and SAMtools* // Bioinformatics. 2009. V. 25. P. 2078-2079 (DOI: 10.1093/bioinformatics/btp352).
42. Yu X, Guda K, Willis J et al. *How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?* // BioData Min. 2012. V. 5. (DOI: 10.1186/1756-0381-5-6).
43. Arslan AN, Eğecioğlu O, Pevzner PA. *A new approach to sequence comparison: normalized sequence alignment.* // Bioinformatics. 2001. V. 17. P. 327-337 (PMID: 11301301).

44. Burrows M,Wheeler D *A block sorting lossless data compression algorithm.* — s.l.: Digital Equipment Corporation, Palo Alto, 1994.

45. Simpson JT,Durbin R. *Efficient construction of an assembly string graph using the FM-index* // Bioinformatics. 2010. V. 26. P. 367-375 (DOI: 10.1093/bioinformatics/btq217).

46. Frommolt P, Abdallah AT, Altmüller J et al. *Assessing the enrichment performance in targeted resequencing experiments.* // Hum Mutat. 2012. V. 33. P. 635-641 (DOI: 10.1002/humu.22036).

47. Danecek P, Auton A, Abecasis G et al. *The variant call format and VCFtools* // Bioinformatics. 2011. V. 27. P. 2156–2158 (DOI: 10.1093/bioinformatics/btr330).

48. Cornish A,Guda C. *A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference* // Biomed Res Int. 2015. (DOI: 10.1155/2015/456479).

49. Zook JM, Chapman B, Wang J et al. *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls* // Nat. Biotech. 2013. V. 32. P. 246–251 (DOI: 10.1038/nbt.2835).

50. Zook JM,Salit M. *Genomes in a bottle: creating standard reference materials for genomic variation - why, what and how?* // Genome Biol. 2011. V. 12. (DOI: 10.1186/gb-2011-12-s1-p31).

51. GCat59 data download FTP // [Bioplanet.com] URL: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/ion_exome/.

52. Illumina, inc. *Infinum LCG Assay Protocol Guide* — s.l.: s.n., 2011.

53. Illumina, inc. *Infinium Genotyping Data Analysis* USA — s.l.: Illumina, inc, 2014 (http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf).

54. Teo YY, Inouye M, Small KS et al. *A genotype calling algorithm for the Illumina BeadArray platform* // Bioinformatics. 2007. V. 23. P. 2741-2746 (DOI: 10.1093/bioinformatics/btm443).

55. Ritchie ME, Liu R, Carvalho BS et al. *Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips* // BMC Bioinformatics. 2011. V. 12. P. 68-77 (DOI: 10.1186/1471-2105-12-68).

56. Zuvich RL, Armstrong LL, Bielinski SJ et al. *Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality* // Genet Epidemiol. 2011. V. 35. P. 887–898 (DOI: 10.1002/gepi.20639).

57. Freier SM, Sinclair A, Neilson T et al. *Improved free energies for G.C base-pairs.* // J Mol Biol. 1985. V. 185. P. 645-652 (PMID: 4057259).

58. Anderson CA, Pettersson FH, Clarke GM et al. *Data quality control in genetic case-control association studies* // Nat Protoc. 2010. V. 5. P. 1564-1573 (DOI: 10.1038/nprot.2010.116).

59. Mille MB,Tang MW. *Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology* // Clin Microbiol Rev. 2009. V. 22. P. 611-633 (DOI: 10.1128/CMR.00019-09).

60. Tindall EA, Petersen DC, Nikolaysen S et al. *Interpretation of custom designed Illumina genotype cluster plots for targeted association studies and next-generation sequence validation* // BMC Res Notes. 2010. V. 3. P. 39-44 (DOI: 10.1186/1756-0500-3-39).

61. Haines JL, Hauser MA, Schmidt S et al. *Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration* // Science. 2005. V. 5720. P. 419-421 (DOI: 10.1126/science.1110359).

62. Mosteller F. *Association and Estimation in Contingency Tables* // J.Amer.Stat.Ass. 1968. V. 321. P. 1-28 (DOI: 10.2307/2283825).

63. Clarke GM, Anderson CA, Pettersson FH et al. *Basic statistical analysis in genetic case-control studies* // Nat Protoc. 2011. V. 6. P. 121-133 (DOI:10.1038/nprot.2010.182).

64. Phipson B,Smyth GK. *Permutation P-values Should Never Be Zero: Calculating Exact P-values when Permutations are Randomly Drawn* // Statistical Applications in Genetics and Molecular Biology. 2010. V. 9. (DOI: 10.2202/1544-6155.1585).

65. Westfall P, Tobias R, Wolfinger R *Multiple comparisons and multiple testing using SAS, 2nd edn* — s.l.: SAS Institute.

66. Purcell S, Neale B, Todd-Brown K et al. *PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses* // Am. J. Hum. Genet. 2007. V. 81. P. 559-579 (DOI: 10.1086/519795).
67. Karssen LC, van Duijn CM, Aulchenko YS. *The GenABEL Project for statistical genomics* //. 2016. №5, (DOI: 10.12688/f1000research.8733.1).
68. Li MX, Kwan JS, Sham PC. *HYST: A HYbrid Set-based Test for genome-wide association studies, with application to protein-protein interaction-based association analysis* // Am J Hum Genet. 2012. V. 7. P. 478-488
(DOI: 10.1016/j.ajhg.2012.08.004).
69. Broad Institute. SNAP: SNP Annotation and Proxy Search // [SNAP: SNP Annotation and Proxy Search] URL: http://www.broadinstitute.org/mpg/snap/.
70. R-project Group. The R Project for Statistical Computing // [The R Project for Statistical Computing] URL: https://www.r-project.org/.
71. Gui H, Li M, Sham PC et al. *Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's Disease dataset* // BMC Research Notes. 2011. V. 4.
(DOI: 10.1186/1756-0500-4-386).
72. Novembre J, Johnson T, Bryc K et al. *Genes mirror geography within Europe* // Nature. 2008. V. 7218. P. 98-101 (DOI: 10.1038/nature07331).
73. Fokkema IF, Taschner PE, Schaafsma GC et al. *LOVD v.2.0: The Next Generation in Gene Variant Databases* // J. Hum. Mutat. 2010. V. online publication.
(DOI: 10.1002/humu.21438).
74. Cariaso M,Lennon G. *SNPedia: a wiki supporting personal genome annotation, interpretation and analysis* // Nucleic Acids Res. 2012. V. 40. P. 1308-1312
(DOI: 10.1093/nar/gkr798).
75. Hamosh A, Scott AF, Amberger JS et al. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders* // Nucleic Acids Res. 2005. V. 33. (DOI: 10.1093/nar/gki033).
76. Stenson PD, Mort M, Ball EV et al. *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine* // Hum Genet. 2014. V. 133. P. 1-9
(DOI: 10.1007/s00439-013-1358-4).

77. Gong L, Owen RP, Gor W et al. *PharmGKB, an Integrated Resource of Pharmacogenomic Data and Knowledge* // Curr Protoc Bioinformatics. 2014. V. 14. (DOI: 10.1002/0471250953.bi1407s23).
78. Human Genome Navigator Association. Human Genome Navigator // [сайт] URL: http://hugenavigator.net.
79. Rosenbloom KR, Armstrong J, Barber GP et al. *The UCSC Genome Browser database: 2015 update* // Nucleic Acids Res. 2015. V. 43. (DOI: 10.1093/nar/gku1177).
80. Bamford S, Dawson E, Forbes S et al. *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website* // Br J Cancer. 2004. V. 91. P. 355-358 (DOI: 10.1038/sj.bjc.6601894).
81. Landrum MJ, Lee JM, Riley GR et al. *ClinVar: public archive of relationships among sequence variation and human phenotype* // Nucleic Acids Res. 2014. V. 42. (DOI: 10.1093/nar/gkt1113).
82. Paila U, Chapman BA, Kirchner R et al. *GEMINI: integrative exploration of genetic variation and genome annotations* // PLoS Comput Biol. 2013. V. 9. (DOI: 10.1371/journal.pcbi.1003153).
83. Ng PC,Henikoff S. *SIFT: predicting amino acid changes that affect protein function* // Nucleic Acids Res. 2003. №31, P. 3812–3814 (PMID: 12824425).
84. Adzhubei IA, Schmidt S, Peshkin L et al. *A method and server for predicting damaging missense mutations* // Nat Methods. 2010. V. 7. P. 248-249 (DOI: 10.1038/nmeth0410-248).
85. McLaren W, Gil L, Hunt SE et al. *The Ensembl Variant Effect Predictor* // Genome Biol. 2016. V. 17. P. 122-123 (DOI: 10.1186/s13059-016-0974-4).
86. Wang K, Li M, Hakonarson H. *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data* // Nucleic Acids Res. 2010. V. 38. (DOI: 10.1093/nar/gkq603).
87. Cingolani P, Platts A, Wang L et al. *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* // Fly (Austin). 2012. V. 6. P. 80-92 (DOI: 10.4161/fly.19695).

88. Andrew Jesaitis. The State of Variant Annotation: A Comparison of AnnoVar, snpEff and VEP // [Best practices in genetic analysis • Technology review] URL: http://blog.goldenhelix.com/ajesaitis/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/.

89. Özgüç M. *Genetic testing: predictive value of genotyping for diagnosis and management of disease* // EPMA J. 2011. V. 2. P. 173-179 (DOI: 10.1007/s13167-011-0077-y).

90. Brugger D, Bogdan M, Rosenstiel W. *Automatic cluster detection in Kohonen's SOM* // IEEE Trans Neural Netw. 2008. V. 19. P. 442-459 (DOI: 10.1109/TNN.2007.909556).

91. Yeung KY,Ruzzo WL. *Principal component analysis for clustering gene expression data* // Bioinformatics. 2001. V. 17. P. 763-774 (PMID: 11590094).

92. Navigenics, "The science behind the Navigenics service," 2012.

93. 23andMe, "Estimating Genotype-Specific Incidence for One or Several Loci," 2007.

94. Chatterjee N, Shi J, García-Closas M. *Developing and evaluating polygenic risk prediction models for stratified disease prevention* // Nat Rev Genet. 2016. V. 17. P. 392-406 (DOI: 10.1038/nrg.2016.27).

95. Kalf RR, Mihaescu R, Kundu S et al. *Variations in predicted risks in personal genome testing for common complex diseases* // Genet Med. 2014. №16, (DOI: 10.1038/gim.2013.80).

96. Wray NR, Yang J, Goddard ME et al. *The genetic interpretation of area under the ROC curve in genomic profiling* // PLoS Genet. 2010. V. 6. (DOI: 10.1371/journal.pgen.1000864).

97. Hajian-Tilaki K. *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation* // Caspian J Intern Med. 2013. V. 4. P. 627-635 (PMID: 24009950).

98. Kundu S, Mihaescu R, Meijer CM et al. *Estimating the predictive ability of genetic risk models in simulated data based on published results from genome-wide association studies* // Front Genet. 2014. V. 5. P. 179 (DOI: 10.3389/fgene.2014.00179).

99. Adams SD, Evans JP, Aylsworth AS. *Direct-to-consumer genomic testing offers little clinical utility but appears to cause minimal harm.* // N C Med J. 2013. V. 74. P. 494-498 (PMID: 24316774).
100. Illumina iControlDB // [Illumina iControlDB] URL: http://www.illumina.com/documents/icontroldb/agreement_download.pdf.
101. Wolfarth B, Rankinen T, Mühlbauer S et al. *Endothelial nitric oxide synthase gene polymorphism and elite endurance athlete status: the Genathlete study.* // Scand J Med Sci Sports. 2008. V. 18. P. 485-490 (DOI: 10.1111/j.1600-0838.2007.00717.x).
102. Mikami E, Fuku N, Takahashi H et al. *Polymorphisms in the control region of mitochondrial DNA associated with elite Japanese athlete status.* // Scand J Med Sci Sports. 2013. V. 23. P. 593-599 (DOI: 10.1111/j.1600-0838.2011.01424.x).
103. Madden T *The BLAST Sequence Analysis Tool* — s.l.: s.n., 2013 (http://www.ncbi.nlm.nih.gov/books/NBK153387/).
104. Starnes LM, Sorrentino A, Pelosi E et al. *NFI-A directs the fate of hematopoietic progenitors to the erythroid or granulocytic lineage and controls beta-globin and G-CSF receptor expression.* // Blood. 2009. V. 114. P. 1753-1763.
105. Zawadowska B, Majerczak J, Semik D et al. *Characteristics of myosin profile in human vastus lateralis muscle in relation to training background.* // Folia Histochem Cytobiol. 2004. V. 42. P. 181-190 (PMID: 15493580).
106. Ahmetov II, Druzhevskaya AM, Lyubaeva EV et al. *The dependence of preferred competitive racing distance on muscle fibre type composition and ACTN3 genotype in speed skaters.* // Exp Physiol. 2011. V. 96. P. 1302-1310 (DOI: 10.1113/expphysiol.2011.060293).
107. Zhang B, Tanaka H, Shono N et al. *The I allele of the angiotensin-converting enzyme gene is associated with an increased percentage of slow-twitch type I fibers in human skeletal muscle.* // Clin Genet. 2003. V. 63. P. 139-144 (PMID: 12630962).
108. Ahmetov II, Vinogradova OL, Williams AG. *Gene polymorphisms and fiber-type composition of human skeletal muscle.* // Int J Sport Nutr Exerc Metab. 2012. V. 22. P. 292-303 (PMID: 22645169).

109. Wei C, Cardarelli MG, Downing SW et al. *The effect of angiotensin II on mitogen-activated protein kinase in human cardiomyocytes.* // Renin Angiotensin Aldosterone Syst. 2000. V. 1. P. 379-384 (DOI: 10.3317/jraas.2000.070).
110. Yang D, Ma S, Tan Y et al. *Increased expression of calpain and elevated activity of calcineurin in the myocardium of patients with congestive heart failure.* // Int J Mol Med. 2010. V. 26. P. 159-164 (PMID: 20514436).
111. Carrillo-Sepúlveda MA, Ceravolo GS, Furstenau CR et al. *Emerging role of angiotensin type 2 receptor (AT2R)/Akt/NO pathway in vascular smooth muscle cell in the hyperthyroidism.* // PLoS One. 2013. V. 8. (DOI: 10.1371/journal.pone.0061982).
112. Chai W, Wang W, Dong Z et al. *Angiotensin II receptors modulate muscle microvascular and metabolic responses to insulin in vivo.* // Diabetes. 2011. V. 60. P. 2939-2946 (DOI: 10.2337/db10-1691).
113. Ben-Zaken S, Eliakim A, Nemet D et al. *Increased prevalence of MnSOD genetic polymorphism* // Free Radic Res. 2013. V. 47. P. 1002-1008 (DOI: 10.3109/10715762.2013.838627).
114. Unified Medical Language System. UMLS Terminology Services // [Unified Medical Language System] URL: https://uts.nlm.nih.gov/.
115. World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision // [World Health Organization: ICD-10] URL: http://apps.who.int/classifications/icd10/browse/2010/en.
116. James KM, Cowl CT, Tilburt JC et al. *Impact of Direct-to-Consumer Predictive Genomic Testing on Risk Perception and Worry Among Patients Receiving Routine Care in a Preventive Health Clinic* // Mayo Clin Proc. 2011. V. 86. P. 933-940 (DOI: 10.4065/mcp.2011.0190).
117. Yeager M, Orr N, Hayes RB et al. *Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.* // Nat Genet. 2007. V. 39. P. 645-649 (DOI: 10.1038/ng2022).
118. Wellcome Trust Case Control Consortium. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* // Nature. 2007. V. 447. P. 661-678 (DOI: 10.1038/nature05911).