

**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**

03.01.09 – математическая биология, биоинформатика.

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата биологических наук

Москва 2016

Работа выполнена в Федеральном государственном бюджетном учреждении Федеральном научно-клиническом центре физико-химической медицины Федерального медико-биологического агентства России.

Научный руководитель:

Кандидат биологических наук, доцент **Генерозов Эдуард Викторович**
Федеральное государственное бюджетное учреждение Федеральный научно-клинический центр физико-химической медицины федерального медико-биологического агентства России (ФГБУ ФНКЦ ФХМ ФМБА России).

Официальные оппоненты:

- Доктор биологических наук, **Лагунин Алексей Александрович**
Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский национальный исследовательский медицинский университет имени Н.И. Пирогова» Министерства здравоохранения Российской Федерации (ФГБОУ ВО РНИМУ Минздрава РФ).
- Кандидат биологических наук, **Кудрявцева Анна Викторовна**
Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук (ФГБУН ИМБ РАН).

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук (ФГБУН ИОГен РАН)

Защита состоится 19 декабря 2016 года на заседании Диссертационного совета Д.002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им.А.А.Харкевича Российской академии наук, расположенном по адресу: 127051, г.Москва, Большой Каретный переулок, 19, стр.1

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им.А.А.Харкевича Российской академии наук (ИППИ РАН), а также на сайте ИППИ РАН по адресу www.iitp.ru.

Автореферат разослан «___» _____ 2016 г.

Ученый секретарь диссертационного совета
д.б.н., профессор

/Рожкова Г.И./

Общая характеристика работы

Актуальность исследования

Математические методы становятся все более востребованными в области современных биомедицинских исследований. Особенно востребованы методы биоинформатики для анализа больших массивов генетических данных, получаемых с помощью технологий высокоплотной чип-гибридизации и секвенирования ДНК второго поколения. Такие данные чрезвычайно интересны как в исследовательском аспекте, так и для практического применения. Вместе с этим возникает и ряд новых и актуальных задач (Willet et al., 2014).

В области медицинской генетики такой задачей является определение роли генетических факторов в этиологии заболеваний, фенотипических проявлений и развитии физиологических характеристик человека. Особый интерес представляет изучение стратифицированных и ограниченных групп лиц с уникальными физиологическими характеристиками, например, профессиональных спортсменов. Информация о генетической основе физиологических особенностей таких людей является ценной с фундаментальной точки зрения, так как позволяет выявить общие механизмы адаптации организма к интенсивным физическим нагрузкам. Результаты таких исследований имеют и практическую значимость: применяются к спортивному отбору, коррекции параметров тренировочного процесса, а также диагностике и лечению пациентов, не являющихся профессиональными спортсменами (Maffulli et al., 2013).

Современная методология ассоциативных генетических исследований базируется на использовании технологий высокопроизводительного генетического анализа. В то же время, методы обработки получаемых данных существенно отстают и являются основным предметом для биоинформатического анализа и дискуссий (Manolio et al., 2010). На текущий момент, одним из популярных технологических решений, позволяющих в короткие сроки обработать большой объем генетической информации, является метод ионного полупроводникового секвенирования (Meldrum et al., 2011). Однако поставляемое

вместе с оборудованием штатное программное обеспечение не позволяет использовать весь объем получаемых генетических данных и исключить ряд ошибок секвенирования, а также имеет ограниченное количество пользовательских настроек. В случае использования протокола эксперимента с различными источниками геномных данных, необходимо приведение результатов генотипирования к единому стандарту для последующего сравнения. Эти и другие проблемы обуславливают приоритетность задачи разработки альтернативных биоинформатических протоколов, позволяющих оптимизировать техническую первичную обработку получаемых с секвенатора данных, которая решается в рамках данной работы. Не менее актуальны и другие решаемые задачи: вычисления при проведении ассоциативных генетических исследований, связанных со спортивной деятельностью, в группах профессиональных спортсменов (Ahmetov et al., 2015) и разработка инструментов, позволяющих осуществлять индивидуальную аннотацию выявленных в ходе генетического исследования вариантов, имеющих клиническое значение.

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

Разработанные программные протоколы, предназначенные для увеличения точности обработки данных, снижают итоговую стоимость секвенирования, что позволит использовать ионное полупроводниковое секвенирование для решения большего числа задач.

Модифицированный алгоритм обработки данных чип-гибридизации, стандартизирующий получаемые данные чип-гибридизации и секвенирования, даст возможность использовать в одном эксперименте данные, полученных на разных технологических платформах.

С использованием разработанных оптимизированных алгоритмов был проанализирован массив данных высокоплотной чип-гибридизации, более 1000 профессиональных спортсменов Российской Федерации, и были получены новые ассоциации различных фенотипических показателей с генетическими маркерами.

Отдельную практическую ценность представляет разработанный алгоритм, позволяющий на основе комплексного анализа полученных генетических данных осуществлять оценку индивидуальных медицинских рисков. Использование алгоритмов такого рода приближает внедрение персонифицированной медицины в реальную клиническую практику. Впервые была разработана и апробирована модель для не инвазивной оценки структуры мышечных волокон на основании генетического анализа.

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 секвенирования.

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

Личный вклад автора

Результаты биоинформатического анализа, изложенные в диссертации, получены автором лично. Диссертант лично участвовал в подготовке всех публикаций по теме диссертации в разделах работ, связанных с биоинформатической подготовкой и обработкой геномных данных. Совместные исследования по генетическим ассоциациям у профессиональных спортсменов, координация работ с зарубежными соисполнителями, работы по сбору и характеристике биологического материала осуществлялась д.м.н. Ахметовым И.И. и соавторами. Во всех опубликованных работах биоинформатическая обработка исходных данных и расчеты при выполнении ассоциативных исследований выполнены лично диссертантом.

Структура и объем работы

Диссертация состоит из введения, трех глав, выводов и библиографии. Общий объем диссертации, включая 14 рисунков и 19 таблиц, составляет 130 страниц, в том числе библиографический список, состоящий из 118 наименований на 10 страницах.

Содержание работы

Материалы и методы

Характеристика образцов

Для проведения исследования были использованы генетические данные 17668 человек, из которых 3543 являются профессиональными спортсменами различных стран мира. Генетические данные представителей Российской Федерации были проанализированы в ходе выполнения работы. Генетические данные представителей других стран были предоставлены в виде результатов генотипирования.

В исследовании принимали участие спортсмены в возрасте 20 ± 3 лет. К группе контроля привлекались представители европеоидной расы, не являющиеся профессиональными спортсменами, не имеющие какого-либо спортивного разряда или звания, не обладающие существенными медицинскими отклонениями

(группой потери трудоспособности), входящие в аналогичную спортсменам возрастную группу.

Для 80 ведущих атлетов сборной России было проведено измерение объема максимального потребления кислорода (МПК, $VO_2\max$). В анализе мышечной структуры приняли участие 40 атлетов, относящихся к группе выносливости. Для 61 спортсмена-стайера был проведен классический тест состава венозной крови (гемоглобин, моноциты, эритроциты и прочие показатели). У 1444 спортсменов сборной России были протестированы уровни различных показателей метаболизма.

Для дополнительного контроля анализировались данные генотипирования 11521 образцов, размещенных в банке данных Illumina iControlDB (<http://www.illumina.com/documents/icontribdb>) без учета возраста, но также европеоидного этноса. Для тестирования программного протокола использовались результаты проекта «Genome In A Bottle» (GIAB) (Zook et al., 2013), образец GCat59, являющийся общепринятым стандартом для сравнения различных методов секвенирования геномной ДНК человека.

Распределение образцов по группам

Для проведения ассоциативных исследований спортсмены разбивались на группы по трем основным признакам: виду упражнений, уровню максимальных достижений и полу. При делении по видам упражнений формировались следующие группы: длительной выносливости (дистанция более 5 км), средней выносливости (длительностью более 5 минут, но дистанцией менее 5 км), короткой выносливости (длительностью от 45 секунд до 5 минут), низкой интенсивности дыхания, игровых видов спорта, силовая, скоростно-силовая. В различных разделах при наличии общего признака группы объединялись.

Разделение по уровню максимальных достижений производилось на группы «элитные», «субэлитные», «не элитные». К «элитным» спортсменам были отнесены победители международных чемпионатов и кубков, «субэлитным» - участники международных соревнований, «не элитным» - все остальные спортсмены олимпийского резерва.

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.

Ионное полупроводниковое секвенирование

Экзомное обогащение образцов геномной ДНК человека и подготовку библиотек для секвенирования осуществляли с использованием набора ION Ampliseq Exome RDY-IC (ThermoFisher) согласно рекомендациям производителя. Полупроводниковое секвенирование образцов ДНК осуществлялось с использованием секвенатора Ion Proton (ThermoFisher) и набором расходных материалов Ion PI Hi-Q Chef Kit (ThermoFisher). Экспериментальные процедуры осуществлялись в лаборатории постгеномных исследований в биологии ФНКЦ ФХМ ФМБА России.

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 ФНКЦ ФХМ ФМБА России.

Определение максимального уровня потребляемого кислорода

Измерение объема поглощаемого кислорода (VO_{2max}) производилось при выполнении силовых упражнений на различных эргометрах (по видам спорта),

определение количества вдыхаемого и выдыхаемого кислорода с помощью системы газоанализации MetaMax 3B (Cortex, Leipzig, Германия). Начальная силовая нагрузка составляла 150 Ватт, шаг повышения нагрузки – 50 Ватт. Упражнение при каждом уровне нагрузки длилось 3 минуты с последующим перерывом в 30 секунд. В качестве критерия максимальности использовалось снижение силы более чем на 30 Ватт от целевой нагрузки упражнения и повышение обменного дыхательного соотношения более чем на 10% перед окончанием упражнения. Значение VO_2max фиксировалось как максимальное в течение последних 30 секунд теста.

Анализ состава мышечных волокон

Анализ состава мышечной структуры, выполнялся в Институте Медико-Биологических Проблем (ФГБУН ГНЦ ИМБП РАН) путем игольчатой биопсии мышц по Бергстрему, выполненной на латеральной четырехглавой мышце бедра мышечной ткани. Состав мышечных волокон определялся иммуногистохимическим способом, а в качестве результата анализа были получены численные значения составных элементов мышечной структуры в объеме в процентах.

Биохимический анализ крови

Биохимический анализ крови проводился в лабораториях ФГБУН ГНЦ ИМБП РАН в двух различных вариантах: простом и расширенном. В случае простого анализа, проводился классический тест состава венозной крови (гемоглобин, моноциты, эритроциты и прочие элементы). Расширенный вариант анализа дополнительно включал уровни следующих показателей метаболизма: креатинкиназы, креатинина, аланинтрансаминазы, аспартаттрансаминазы, алкалинфосфатазы. Для проведения анализа у спортсменов была забрана венозная кровь объемом 10 мл в утреннее время после продолжительного сна.

Программное обеспечение

Для обработки 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 обработки результатов секвенирования использовались программный протокол, состоящий из пакетов: cutadapt, bwa, samtools, vcftools, NGSrich. Кроме того, использовался программный пакет Torrent Suite (версии 5.0) в стандартной комплектации от ThermoFisher, inc.

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).

Для статистического сравнения полиморфизмов различных групп спортсменов использовался программный пакет PLINK 1.9. В качестве алгоритмов поиска ассоциаций использовался непарный Т-тест, а также тест Фишера (ANOVA). Поправка на множественность сравнений выполнялась по методу Бонферрони-Холма или Бенджамини-Хохберга (FDR). Для сравнения качества работы алгоритмов обработки данных секвенирования применялся статистический критерий F1-мера.

Результаты

Обработка результатов ионного полупроводникового секвенирования

Была произведена разработка альтернативного протокола анализа данных ионного полупроводникового секвенирования для увеличения объема и качества получаемых результатов при сравнении со штатным программным обеспечением (Torrent Suite). Предварительный анализ эффективности работы штатного программного обеспечения показал уровень точности не превосходящий 99% при сравнении результатов с данными генотипирования на чипах. За основу альтернативного протокола был взят программный конвейер BWA – SAMtools – VCF-tools.

Для подбора оптимальных дополнительных параметров при применении конвейера к результатам ионного полупроводникового секвенирования использовались результаты проекта GIAB (образец GCat59), являющиеся

мировым стандартом тестирования аннотационных протоколов. В качестве входных данных использовался FASTQ-файл, полученный при ионном полупроводниковом секвенировании указанного образца, а для сравнения применялся агрегированный VCF-файл. Было выбрано 2 различных критерия оценки качества работы конвейера: F1-мера и точность. В качестве начальных параметров для последующего программного перебора использовались значения «по умолчанию» всех программных пакетов, входящих в конвейер. Так как на непосредственную фильтрацию результатов работы влияют исключительно дополнительные параметры пакета VCF-tools, в дальнейшем рассматриваться будут только они.

При оценке качества работы конвейера с помощью критерия F1-мера все возможные результаты были разбиты на два класса: прошедшие установленные дополнительные параметры и отфильтрованные. Класс прошедших фильтры был разделен на два подкласса: референсные состояния и альтернативные. Для каждого запуска вычислялись:

$$\begin{aligned} \text{Точность} &= \frac{\text{верные класса 1}}{(\text{верные класса 1} + \text{ошибочные класса 1})} \\ \text{Воспроизводимость} &= \frac{\text{верные класса 1}}{(\text{верные класса 1} + \text{ошибочные класса 2})} \\ F1 - \text{мера} &= \frac{2 \times \text{точность} \times \text{воспроизводимость}}{(\text{точность} + \text{воспроизводимость})} \end{aligned}$$

В результате оптимизации максимальное значение F1-меры составило 92,98%, и было достигнуто при следующих дополнительных параметрах, отличных от значений «по умолчанию»: минимальное покрытие (DP) 14, минимальное качество картирования (MinMQ) 30, минимальное качество прочтения (Qual) 10.

Такой алгоритм определения оптимальных параметров является хорошим со стороны математической статистики, но будет определять значительное число отличий от референса ценой существенно меньшей точности определения референсных состояний.

При использовании критерия одногрупповой точности формировалось два класса: прошедшие через установленные дополнительные параметры и

отфильтрованные. Затем производилось значительное число программных повторов, и для каждого вычислялась программная точность:

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

Считалось, что результаты являлись удовлетворительными лишь в том случае, если количество полученных полиморфизмов в итоге насчитывало не менее 90% от общего ожидаемого числа. В результате оптимизации максимальное значение точности составило 97%, и было достигнуто при следующих дополнительных параметрах, отличных от значений «по умолчанию»: минимальное покрытие (DP) 22 для гомозигот и 28 для гетерозигот, минимальное качество картирования (MinMQ) 30, минимальное качество прочтения (Qual) 30.

При анализе количественных показателей каждого из подходов было выдвинуто предположение, что при проведении ассоциативных исследований следует проводить оптимизацию с использованием критерия F1, а при индивидуальном анализе – с использованием критерия одnogрупповой точности.

Была проведена проверка работы полученного протокола на реальных биологических образцах: 3 мужских и 3 женских. В качестве данных для сравнения использовались результаты генотипирования на различных ДНК-чипах по технологии Illumina Infinium с завышенным требованием по качеству определенных геномных состояний (GC-score > 0,9), отфильтрованные по заявленным производителем фрагментам экзомного набора. Максимальная достигнутая точность составила 99,66%, средняя точность по шести образцам – 99,52%.

Кроме сравнения с результатами чип-гибридизации также проводилось и сравнение с результатами работы штатного программного обеспечения (Torrent Suite). Разработанный протокол показывал в среднем на 0,5% более высокую точность, что соответствует 270-ти тысячам дополнительных достоверных геномных состояний на экзом человека, которые приведены в таблице.

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).

Для стандартизации результатов чип-генотипирования, полученных на различных версиях чипов, и снижения числа технически ошибочных результатов при использовании данных требовалось произвести проверку корректности интерпретируемых вариантов полиморфизмов на предмет исключения ошибок, связанных с несоответствием цепи позиционирования полиморфизмов и несоответствием координат местоположения полиморфизмов в стандартной версии генома человека (NCBI), а также несоответствие указателей rs текущей версии каталога dbSNP. Так как в данной работе производились ассоциативные исследования, то использовалась последняя стабильная сборка генома версии hg19. В качестве источника информации о структуре чипа использовалась техническая документация, представленная производителем и содержащая последовательности зондов и их координаты в одном из стандартизованных обновлений различных версий генома человека.

На первом шаге производился поиск указанных координат, и проверялось соответствие нуклеотидных последовательностей и цепей. Последовательности считались идентичными, если количество несовпадающих нуклеотидов не превышало двух на зонд. В случае если последовательности не совпадали, производился полногеномный поиск с использованием программного пакета blastn. В случае если обнаруживался подходящий уникальный фрагмент генома, то информация о его местоположении сохранялась и заменяла заводскую. Те зонды, которые не имели аналогичных фрагментов в последовательности генома hg19, а также те, которые имели несколько схожих фрагментов, маркировались как исключенные.

В результате такого анализа каждого из чипов был сформирован файл-схема с указаниями о замене генотипов, формируемых штатным программным обеспечением прибора, на комплементарные и с заметками об исключении отдельных зондов из анализа. Алгоритм позволил существенно снизить риск возникновения ошибок при ассоциативных исследованиях. В рамках данной

работы был разработан программный продукт для предварительной обработки результатов чип-гибридизации, и все образцы были обработаны с его использованием.

Разработанные протоколы обработки данных полупроводникового секвенирования и чип-гибридизации были использованы для единого препроцессинга генетических данных, полученных разными методами в независимых экспериментах для серии ассоциативных исследований.

Ассоциативные генетические исследования профессиональных спортсменов.

В рамках работы было произведено 4 различных ассоциативных исследования: анализ генетических факторов, связанных с выносливостью; факторов, ассоциированных со структурой мышечных волокон; факторов, ассоциированных с оксидативным стрессом; а также полногеномный ассоциативный анализ общего генетического профиля спортсменов-стайеров. Для предварительной фильтрации геномных образцов по этнической принадлежности было проведено сравнение выборки с этническими группами, размещенными в проекте Illumina iControlDB. 23 образца, не относящиеся к европейскому этносу, были исключены из дальнейшего исследования.

Анализ генетических основ выносливости

Полногеномное ассоциативное исследование маркеров выносливости было проведено с использованием геномных данных российских спортсменов, полученных методом генотипирования с использованием ДНК-чипа Illumina 1M Omni Quad (1140419 тестируемых геномных состояний). В качестве показателя аэробной выносливости были анализированы данные максимального потребления кислорода (МПК, VO_{2max}). Поиск производился в два этапа. На первом этапе производилось формирование списка кандидатных полиморфизмов на основании ассоциативного исследования 80 образцов с известным уровнем максимального потребления кислорода. На втором этапе производилась валидация достоверных ассоциаций с использованием всей доступной выборки спортсменов. 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 значимыми. В качестве уровня полногеномной значимости был установлен критерий $P < 10^{-5}$. По результатам поисковой фазы исследования изначально было установлено 15 нуклеотидных полиморфизмов, значимость 6 из которых была подтверждена (rs7144481, rs1572312, rs12893597, rs9922134, rs296000, rs7191721).

По результатам проведенного исследования была подтверждена ассоциация с показателями выносливости трех из шести полиморфизмов: NFIA-AS2 rs1572312 C, TSHR rs7144481 C, RBFOX1 rs7191721 G. Среди них наиболее значимый полиморфизм был выявлен в гене – NFIA-AS2, который кодирует антисмысловую РНК2, регулирующую, как предполагается, экспрессию гена NFIA. Данный ген в свою очередь вовлечен в регуляцию эритропоэза и гранулопоэза, и влияет на структурный состав крови у атлетов (Starnes et al., 2009).

Анализ генетических факторов, ассоциированных со структурой мышечных волокон

Структура мышечных волокон, а именно, соотношение быстрых и медленных волокон также существенно влияет на физические показатели. Показано, что во многом этот параметр детерминируется генетическими факторами (Ahmetov et al., 2011). В результате поиска по публикациям был сформирован список из 15 кандидатных полиморфизмов для проведения ассоциативного исследования. Анализ также производился в два этапа. На первом этапе было проведено ассоциативное исследование 95 образцов (40 спортсменов и 55 контролей) для которых были известны результаты биопсии четырехглавой мышцы бедра. Только полиморфизм rs11091046 гена AGTR2 показал значимую ассоциацию с составом мышечных волокон. На втором этапе при изучении генотипов 2178 спортсменов и 1319 контролей данная гипотеза подтвердилась. Дополнительно был проведен поиск ассоциации полиморфизма с МПК для 44 элитных спортсменов, изученных в предыдущем исследовании. Результаты этого поиска впервые продемонстрировали ассоциацию полиморфизма rs11091046 в гене AGTR2 с составом мышечных волокон, показателями выносливости у

атлетов с параметрами кислородного обмена. Ген AGTR2 кодирует ангиотензиновый рецептор II типа, который является компонентом ренин-ангиотензиновой системы (РАС) - регулятора роста и дифференцировки скелетных мышц (Wei et al., 2000).

Генетический анализ факторов, ассоциированных с оксидативным стрессом

Оксидативный стресс достаточно часто проявляется у профессиональных спортсменов при интенсивных нагрузках, что объясняется повышенным накоплением токсичных продуктов кислородного обмена (например, окислов липидов), нарушающих клеточный метаболизм и приводящих к ускоренному разрушению клетки. В работе была выдвинута гипотеза, что вариант rs4880 (C/T) в гене SOD2 (генотип TT) уменьшает количество активного фермента (Ben-Zaken et al., 2013), защищающего мышцы от повреждений. Предполагалось, что у профессиональных спортсменов указанный генотип ассоциирован с повышением уровня креатинкиназы в крови после соревнований, что косвенно свидетельствует о большем повреждении мышц. Поиск ассоциаций проводился на выборке из 2664 спортсменов из России и Польши, а также 917 человек группы контроля. Гипотеза подтвердилась. Кроме того было замечено, что средние уровни креатинкиназы и креатинина среди элитных спортсменов в любой из спортивных групп значительно ниже, чем у не элитных. Показано, что негативное влияние генотипа TT в полиморфизме rs4880 полиморфизма на достижения его носителей в скоростных видах спорта является статистически значимым.

Полногеномный ассоциативный анализ общего генетического профиля спортсменов-стайеров

Гипотеза о существовании единого генетического профиля элитного спортсмена-стайера была экспериментально проверена в ходе совместного исследования с консорциумом GAMES ДНК российских и зарубежных спортсменов. В рамках консорциума производилось подтверждение на выборках атлетов-стайеров из различных стран мира утверждения о включении в единый профиль 26 полиморфизмов, обнаруженных в рамках глобального ассоциативного проекта GENATHLETE (Wolfarth et al., 2008), а также 19 полиморфизмов,

полученных в ассоциативном исследовании атлетов-стайеров Японии (Mikami et al., 2013). Каждая страна, входящая в консорциум GAMES производила проверку ассоциаций самостоятельно, а затем производилось сравнение полученных результатов. Предполагалось, что в случае значимого воспроизведения какого-либо полиморфизма в двух странах или более, данный полиморфизм относится к общему генетическому профилю стайеров. В результате проверки гипотезы в различных странах были выявлены различные наборы полиморфизмов, прошедшие все требуемые критерии. Таким образом, рассматриваемая гипотеза не подтвердилась.

Медицинская аннотация индивидуальных генетических особенностей на основании данных генетических исследований

Помимо проведения ассоциативных генетических исследований, важной биоинформатической задачей является разработка алгоритмов, позволяющих аннотировать медицински значимые полиморфизмы в рамках высокоплотного индивидуального генетического теста. В третьем разделе работы производилась разработка таких инструментов аннотации индивидуальных генетических особенностей на основании информации о геномных состояниях пациента.

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, LOVD, OMIM, SNPeDia, PharmGKB, ClinVar). Для наполнения были разработаны специальные программные инструменты, которые конвертировали информацию из всех указанных источников с использованием web-интерфейса или API (при его наличии), а затем размещали ее в собственной базе данных. При включении записей в базу данных все ассоциации были сгруппированы в 7 подтипов по уровню медицинской значимости. Также была проведена оценка информационной достоверности. Для объединения ассоциаций в группы по нозологическим признакам в случае различных синонимичных наименований

заболеваний использовался список синонимов проекта UMLS, а также международный классификатор ICD-10. Структура базы данных приведена на рисунке 1.



Рисунок 1. Схема структуры реляционной базы данных для оценки индивидуальных рисков развития заболеваний.

Для удобства использования был написан программный пакет, принимающий в качестве входных данных как VCF-файлы результатов анализа данных секвенирования, полученные на любых платформах, так и файлы формата 23andMe. 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.

Все описанные программные пакеты были собраны в единый протокол, способный из данных секвенирования (fastq) или чип-гибридизации (23andMe) получать список полиморфизмов, отличных от референсного генома с описанием их биологической значимости программным пакетом AnnoVar, а также список всех вариантов, когда-либо включенных в различные исследования, с их рисковыми статусами.

Для применения в клинической практике была разработана удобная html-форма выдачи, в которой осуществляется сортировка всех выявленных полиморфных медицински ассоциированных вариантов на блоки, сформированные по нозологическому принципу. В составе каждого блока выявленные варианты отсортированы по рисковому статусу с использованием цветовых схем, что позволяет обратить внимание медицинского специалиста на определенные варианты и состояния.

Разработка релевантной модели прогнозирования состава мышечных волокон

В большинстве случаев список известных генетических ассоциаций с мультифакторными нозологиями для каждого индивидуума содержит как рисковые, так и протективные статусы. На основании таких результатов утверждать о рисковом статусе данного пациента затруднительно. В рамках настоящей работы была разработана и апробирована модель, позволяющая прогнозировать на основании генетических факторов состав мышечных волокон. Для ее разработки был произведен поиск опубликованных ассоциаций генетических состояний и мышечной структуры. В состав разработанной панели было включено 14 полиморфизмов: rs1815739 в гене 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.

На основании ранее полученных результатов (Ahmetov et al., 2012), было допущено предположение, что наилучший результат при оценке структуры мышечных волокон может быть достигнут при использовании закрытой аддитивной модели. Каждый из включённых в модель полиморфизмов может оцениваться в 3-х состояниях: ассоциации с преобладанием волокон 1-ого типа, промежуточном состоянии и состоянии ассоциации с преобладанием 2-ого типа волокон. Каждой из аллелей, приводящей к повышению числа мышечных волокон 1-ого типа, присваивается весовой индекс, равный 1. Для получения суммарного рискового индекса производится суммирование рисковых индексов

по всем 14 полиморфизмам для тестируемого пациента. Эффективность разработанной модели была оценена путем построения характеристической кривой (ROC) и варьирования критерия отнесения к каждой из групп. Точность работы модели (AUC) составила 81% для профессиональных спортсменов и 73% для лиц, не являющихся спортсменами. Применение такой панели на практике позволит не инвазивно, только на основании генетического теста предсказывать вероятное соотношение мышечных волокон в скелетных мышцах. Такой тест может быть востребован для профориентации молодых спортсменов и для оптимизации тренировочного процесса.

Заключение

Проблема биоинформатической обработки данных, получаемых с помощью современных геномных технологий, как для последующих задач ассоциативных исследований, так и для индивидуальных медицинских генетических тестов является актуальной и неоднозначной, в зависимости от возможных вариантов ее решений. На сегодняшний день технологические возможности получения геномных данных в значительной степени идут впереди биоинформатических технологий.

В данной работе был разработан оптимизированный протокол, позволяющий на более качественном уровне обрабатывать и стандартизовать исходные генетические данные, полученные с помощью ионного полупроводникового секвенирования, при использовании которого была достигнута точность, превышающая показатели штатного программного обеспечения (99,66%). Разработанный модифицированный протокол обработки результатов чип-гибридизации позволяет исключить ошибки неправильного позиционирования определяемых генетических вариантов при анализе данных, получаемых на различных версиях ДНК-чипов. Указанные разработки позволят использовать обе описанные технологии наиболее полно при различных исследованиях.

Разработанные протоколы были применены при проведении серии ассоциативных исследований с использованием генетических данных более, чем

1200 спортсменов РФ., По результатам проведённых исследований получены новые данные по генетическим детерминантам структуры мышечных волокон, максимальным уровнем потреблением кислорода, оксидативного стресса.

Дополнительно разработан аннотационный инструмент, основанный на собственной обновляемой базе данных по медицински значимым генетическим вариациям. Данный программный продукт позволяет осуществлять медицинскую аннотацию генетических данных, полученных с использованием современных технологий генотипирования в формате индивидуального генетического теста.

В совокупности, применение на практике разработанных протоколов и полученных с их использованием генетических данных может быть востребовано как в различных научных дисциплинах, так и на практике. В частности, протоколы обработки данных полупроводникового секвенирования являются универсальными и могут быть настроены и использованы в качестве базовых при решении любых задач, связанных с получением первичных генетических данных на этой платформе. Полученные результаты ассоциативных исследований могут быть использованы для адаптации тренировочного процесса, и его индивидуализации на основании генетических данных конкретного спортсмена. А возможности аннотационного инструмента не ограничиваются только границами спортивной медицины и позволяют его применять и при решении других задач, в которых необходима аннотация генетических вариантов, ассоциированных с медицинскими показателями.

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 публикаций по теме диссертации

По теме диссертации опубликовано пять работ в рецензируемых научных журналах:

1) Mustafina LJ, Naumov VA, Cieszczyk P, Kulemin NA et al. *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, Kulemin NA et al. *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. et al. *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, Nickolay A. Kulemin et al. *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 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.
- 3) 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.)
- 4) Конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». I.I.Ahmetov, N.A.Kulemin et al. *Markers Associated with Elite Endurance Performance in Russians* // Казань, 20.11.2014.
- 5) Конференция «NGS в медицинской генетике». Н.А.Кулемин, К.А.Бабалаян, Э.В.Генерозов. *Адаптация алгоритмов поиска полиморфизмов к результатам ионного полупроводникового секвенирования* // Суздаль, 24.04.2016.
- 6) Конференция «NGS-2016». Н.А.Кулемин, К.А.Бабалаян, А.В.Гришин, Э.В.Генерозов. *Алгоритм поиска полиморфизмов в результатах ионного полупроводникового секвенирования* // Москва, 19.05.2016.