НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ АКУШЕРСТВА,
ГИНЕКОЛОГИИ И РЕПРОДУКТОЛОГИИ ИМЕНИ Д.О. ОТТА

На правах рукописи

Глотов Олег Сергеевич

СЕКВЕНИРОВАНИЕ ЭКЗОМА ЧЕЛОВЕКА И ПЕРСПЕКТИВЫ ПРЕДИКТИВНОЙ МЕДИЦИНЫ

Научная специальность 1.5.7. ГЕНЕТИКА

ДИССЕРТАЦИЯ
на соискание ученой степени доктора биологических наук

Санкт-Петербург, 2023

Данная диссертация посвящается светлой памяти моего учителя, наставника, научного руководителя и консультанта доктора медицинских наук, профессора, член-корр. РАН Владислава Сергеевича Баранова.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ. Изучение генетических основ здоровья человека: от функции к патологии, от мутации к вариантам. Концепция предиктивной медицины и генетического клинического паспорта здоровья в эпоху секвенирования нового поколения

Научно-технический прогресс в биологии и медицине в конце XX и начале XXI вв. привел к появлению новых высокотехнологичных методов ранней диагностики, выявлению многих белковых и генетических маркеров-мишеней,внедрению в клиническую практику новых стратегий скрининга и протоколов таргетной терапии.Это способствовало выявлению причин редких моногенных заболеваний, улучшению профилактики и повышению эффективности лечения многофакторных социально-значимых заболеваний, что в итоге, улучшило качество здоровья и продолжительность жизни людей в экономически развитых странах [Баранов *и др.,* 2021]. Все эти достижения способствовали изменению парадигмы всего здравоохранения, связанной с переходом от групповой к предиктивной, превентивной, персонализированной медицине (ПМ) и терапии, учитывающей не только клинический диагноз заболевания, его стадию, пол, возраст пациентов, но и индивидуальные молекулярно-генетические профили биомаркеров, ассоциированных с риском развития патологии, ее прогнозом, исходом и эффективностью лечения. Успехи в развитии генетики и информационных технологийпривели к переосмыслению термина «мутация», появлению новых дисциплин – геномики (протеомики, метаболомики, транскриптомики, фармакогеномики), разработке стандартных критериев обработки больших массивов данных с применением биоинформатики с внедрением высокопроизводительных методов исследования структуры генов, прежде всего секвенирования ДНК нового поколения (NGS) [Баранов *и др.,* 2021].

Создание технологии секвенирования ДНК первого поколения в 80-х гг. открыло возможность расшифровать последовательность генома человека. С

этой целью в 1990 г. Национальным Институтом здоровья США был запущен проект «Геном человека» (HGP, TheHumanGenomeProject), в котором, помимо США, участвовали Великобритания, Япония, Франция, Германия, Испания и Китай. Проект завершился в 2003г., когда Национальным центром биотехнологической информации США (NCBI) была опубликована первая версия сборки полного генома человека (hg17) [IHGSC. Finishingtheeuchromatic, 2004]. Однако, данная версия сборки содержала много пропусков, которые были впоследствии досеквенированы. Последняя версия генома человека GRCh38.p14 была опубликована Консорциумом исследований генома (GRC) 09 мая 2022г. [Genome Reference Consortium, 2021].

Следующим этапом после опубликования генома человека было создание карты генетических вариантов или карты гаплотипов (проект HaplotypeMap, HapMap) у 270 человек, относящихся к четырем расам. В 2007 г. была опубликована вторая сборка HapMap, включающая 3,1 млн. вариантов [The International Hapmap, 2007]. Продолжением проекта HapMap была инициатива по секвенированию 1000 геномов человека (проект 1000 Genomes) [Genomes Project Consortium, 2012]. Этот проект позволил идентифицировать 38 млн. однонуклеотидных вариантов (SNV, SNP), а также 1,4 млн. биаллельных инсерций или делеций (инделов) и 14000 больших делеций у 1092 человек, относящихся к 14 этническим популяциям. Проекты HapMap и 1000 Genomes способствовали выявлению редких SNP с высокой пенетрантностью, которые, как считается, являются причинами развития моногенных заболеваний человека [Freund *et al.,* 2018]. Определение SNV поставило следующую задачу: аннотирование их патологических вариантов и поиск ассоциаций с заболеваниями. Данные обоих проектов HapMap и 1000 Genomes были применены для разработки методологии полногеномных исследований этих ассоциаций (GWAS, genome-wideassociationstudies). С помощью GWAS были охарактеризованы популяционные частоты многих SNV/SNP, ассоциированных с многофакторными заболеваниями: сахарным диабетом 1-го типа [Stankov *et*

al., 2013], сахарным диабетом 2-го типа [Sladek *et al.,* 2007], раком молочной железы [Fanale *et al.,* 2012] и другими, а также старением организма и долголетием человека [Deelen *et al.,* 2019]. Сегодня, имея представление о типе наследования, патогенезе заболевания и популяционных частотах SNV/SNP, можно аннотироватьи клинически интерпретировать патогенные варианты генов, найденные с помощью NGS [Rabbani *et al.,* 2012]. Важной особенностью идентификации генов и их вариантов, ассоциированных с заболеванием (состоянием), является возможность изучения ихпатогенетической роли. Первое свидетельство применения NGS для обнаружения генетических аберраций было предоставлено группой Шендуре в 2009 г, когда с помощью полноэкзомного секвенирования был открыт синдром Миллера-Фишера – редкая рецессивная воспалительная (аутоиммунная) демиелинизирующая полирадикулоневропатия [Ng *et al.,* 2009; Ng *et al.,* 2010].

В последние годы возрастает число работ, посвященных изучению особенностей аллельных частот функционально значимого полиморфизма в различных популяциях. Обязательным начальным этапом таких исследований является изучение популяционных частот функционально значимых аллелей генов. Важность популяционных исследований заключается в том, что различия в частотах функциональных полиморфных генных вариантов в разных популяциях могут зависеть от географических условий, региона проживания, особенностей пищевого рациона, расовой и этнической принадлежности и многих других факторов [Хуснутдинова *и др.,* 1997; Nebert and Carvan, 1997; Степанов и Пузырев, 2000; Степанов, 2002]. Поэтому не вызывает сомнения, что относительный вклад одного и того же генетического полиморфизма в этиологию и патогенез конкретного заболевания может различаться в разных популяциях и этнических группах [Баранов *и др.,* 2000; Степанов, 2002; Глотов О. *и др.,* 2004].

На основании вышеизложенного становится понятным, что еще необходимы популяционно-ориентированные исследования фундаментальных

основ здоровья человека с помощью новых высокопроизводительных методов, в том числе и NGS. Поэтому одним из ключевых направлений повышения эффективности внедрения NGS в практику для задач предиктивной медицины является развитие популяционных баз (в том числе - отечественных) для оценки частот генных вариантов, играющих роль в патогенезе наследственных и мультифакториальных заболеваний, совершенствование биоинформатических и статистических протоколов обработки и анализа данных секвенирования ДНК.

История геномных исследований и технологий секвенирования нуклеиновых кислот представлена на рисунке 1. В основе всех технологических платформ NGS лежит способность одновременного прочтения сразу многих участков генома, что является главным отличием от более ранних методов секвенирования. В ходе NGS за один рабочий цикл могут генерироваться более нескольких миллионов нуклеотидных последовательностей для последующего анализа.

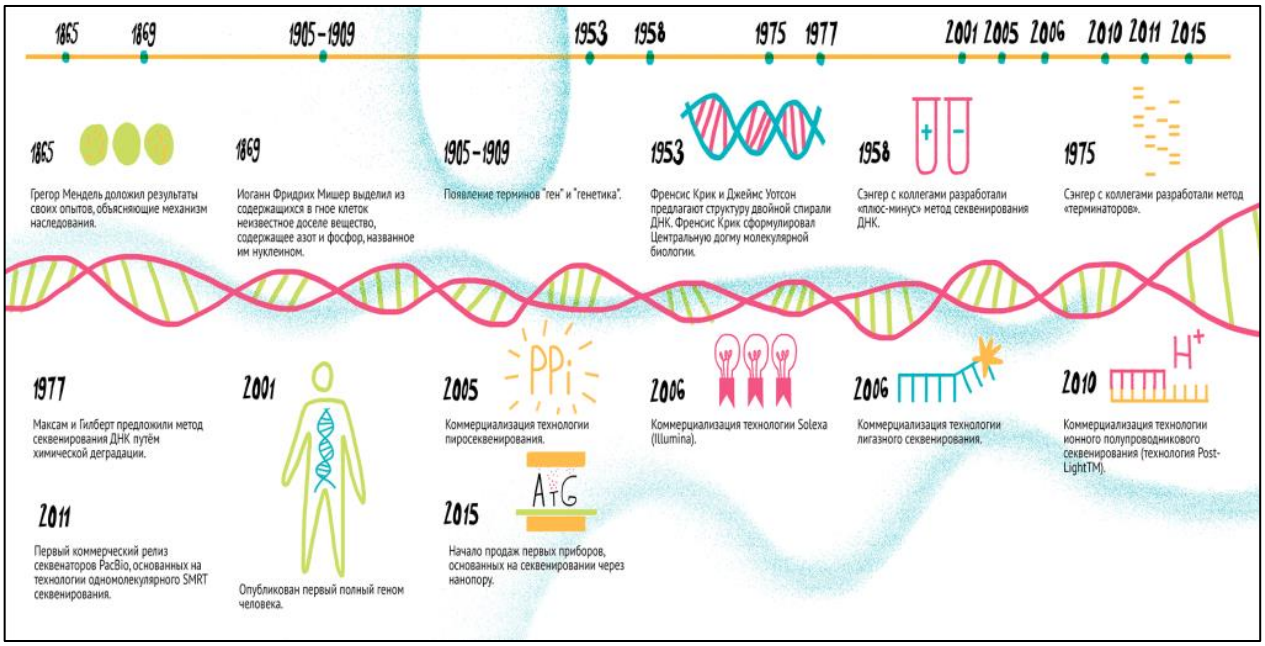


Рисунок 1. Исторические этапы геномных исследований и секвенирования нуклеиновых кислот (biomolecula.ru).

Платформы для NGS: преимущества и недостатки

Технологию NGS можно разделить на секвенирование с короткими (от 100 до 600 п.н.) и длинными (до 900 тыс. п.н.) прочтениями. В настоящее время наиболее часто применяются методы с короткими прочтениями последовательностей, поскольку они дешевле и имеет более высокую точность, чем методы с длинными прочтениями. При этом технологии с короткими прочтениями не могут использоваться для секвенирования повторяющихся или гетерозиготных последовательностей, для которых применяют технологии секвенирования с длинными прочтениями [Morganti *et al.,* 2020]. На этом основании технологии секвенирования формально делят на NGS второго и третьего поколения. Ко второму поколению относятся секвенаторы, позволяющие получить большое количество коротких прочтений (25–800 п.н.), среди них приборы 454 Life Sciences (уже снятый с производства), Illumina, Ion Torrent, MGI. К третьему поколению относятся секвенаторы Pacific Biosciences и Oxford Nanopore, позволяющие прочитывать более длинные участки генов [Бархатов *и др.,* 2016]. Сейчас наиболее широко используют секвенаторы Illumina, в ряде клинических приложений используются приборы IonTorrent (для неинвазивного пренатального тестирования (НИПТ) и пренатального генетического тестирования (ПГТ)), активно внедряются машины компании MGI. Технология Oxford Nanopore тоже распространена в России, но ее применяют для решения сугубо фундаментальных задач (данные приборы активнее используют биоинформатики, чем лабораторные генетики) [Шиков *и др.,* 2019]. Считается, что технология Oxford Nanopore является менее точной, чем методики секвенирования второго поколения. Однако, это не всегда так. В частности, до наших исследований не было показано, что при большом покрытии (около 4000x) применимость данных секвенирования нанопор для определения вариантов митохондриального генома согласуется с данными Illumina [Shikov *et al.,* 2021].

Область применения NGS

Варианты применения NGS в медицинских исследованиях многообразны. По целевому назначению технологии высокопроизводительного секвенирования могут быть разделены на следующие группы:

1) определение последовательности всей ДНК (полногеномное секвенирование —whole-genome sequencing, WGS);
2) определение последовательности кодирующих белок участков генома (полноэкзомное секвенирование — whole-exome sequencing, WES);
3) определение последовательности интересующих генов (от «клинических экзомов», включающих около 4-5000 значимых в медицинском отношении генов, до малых таргетных панелей на 1–3 гена или локуса);
4) секвенирование транскриптома (секвенирование РНК, RNA-seq);
5) оценкабиологического разнообразиябактериального микробиома [Goloshchapov *et al.*, 2020; Баранов *и др.,* 2021].

Особенности различных применяемых в медицинской практике технологий NGS представлены в таблице 1.

Таблица 1. Преимущества и недостатки различных технологий NGS.

Технология	**Особенность**	**Плюсы**	**Минусы**	**Возможность импортозамещения**
Таргетное NGS секвенирование (специализированные «панели вариантов генов»)	Включает только выбранные варианты генов, выявляемые данным методом (SNP)	Быстрый и дешевый метод 100% эффективность	В панелях отсутствуют распространённые варианты генов, характерные для той или иной популяции Не позволяет выявлять CNV, мини и микросателлитные последовательности	Существует
Секвенирование экзома человека методом NGS («экзом»)	Включает более 20 тыс. вариантов генов (практически	Идеально подходит для близкородственных пар Не чувствителен к популяционным особенностям	Дорог, 50% диагностическая эффективность. Не позволяет выявлять CNV, мини и микросателлиты	Решения разрабатываются

	все моногенные заболевания)		Требует наличия биоинформатической обработки данных	
Секвенирование генома человека методом NGS («геном»)	Весь геном человека	Позволяет устанавливать молекулярную природу практически всех заболеваний Не чувствителен к популяционным особенностям	Дорог, неизвестная эффективность. Требует биоинформатической обработки больших баз данных	Существует
Комплексный подход (NGS, MLPA, ПЦР-РВ)	Включает выбранные варианты генов	Дешевый метод 100% эффективность	Чувствителен к популяционным особенностям	Существует

Таргетное секвенирование конкретных генов (TS) стало одним из первых практических приложений NGS. Благодаря этой технологии стало возможным прочитывать отдельные гены в 1000 раз дешевле, чем ранее, с использованием ПЦР для длинных фрагментов (long-range ПЦР), позволяющей амплифицировать участки до 50–100 тыс. п. н. Во-первых, данный подход особенно эффективен для небольших генов [Glotov A. *et al.,* 2018]. Другим приложением таргетного секвенирования является подход, когда на основании Ampliseq-технологии можно амплифицировать до 27 тысяч ПЦР-продуктов, охватывающих целевые фрагменты, для последующего массового параллельного секвенирования. Данный подход используется при исследовании сравнительно небольшого числа генов [Glotov *et al.,* 2015]. Третьим приложениемтаргетного секвенирования может быть анализ целевых фрагментов гена (или генов), при котором на стадии подготовки проб используют технологию обогащения с помощью целевых зондов, благодаря которым «вытягиваются» из генома нужные фрагменты гена(ов), которые в дальнейшем секвенируются. Такая технология нашла свое применение для диагностики всех мутаций в гене муковисцидоза, а также в генах других моногенных заболеваний, например, болезни Вильсона –Коновалова [Balashova *et al.,* 2020]. Несомненным достоинством TS перед WGS, WES является низкая

стоимость (более чем в 50 раз), более короткое время выполнения секвенирования и биоинформационного анализа данных [Gonzalez-Garay, 2014].

Полногеномное секвенирование позволяет провести полный анализ всех генов человека, включающего как структурные, так и регуляторные гены и около 3 млрд. п.н. в гаплоидном наборе. Однако, по данным последней опубликованной версии генома человека GRCh38, структурные гены, кодирующие белки (экзом) составляют всего 3,09% (90 млн. п.н.) от общего количества генов, но содержат около 85% функциональных вариантов, ассоциированных с клиническими проявлениями заболевания [Guo *et al.,* 2017; Majewski *et al.,* 2011]. Поэтому, если целью исследования является изучение только белок-кодирующих нуклеотидных последовательностей, то использование полноэкзомного секвенирования (WES) дает возможность обеспечить широкое покрытие 22000-25000 генов, в более экономичном плане, чем WGS, и обнаруживать более редкие патологические генетические варианты (Single Nucleotide Polymorphism, SNP, инсерции, делеции), которые могут служить причиной заболевания [Suwinski *et al.*, 2019].

Несмотря на кажущееся преимущество полногеномного подхода над экзомным, анализ диагностической ценности двух этих методов для поиска клинически значимых мутаций не показал такой закономерности. Полногеномное тестирование позволяет выявлять, по разным оценкам, от 44 до 50 % всех искомых мутаций, тогда как экзомное уступает ему всего на 2 %. Во многом это связано с тем, что при анализе экзома используют более длинные фрагменты и более совершенные конструкции зондов [Barbitoff *et al.,* 2020]. Поэтому в настоящее время, в совокупности с относительно низкой ценой исследования, экзомное секвенирование выглядит более привлекательным для клинического применения.

С технической точки зрения решающее значение при проведении WES является выбор набора зондов для целевой гибридизации (захвата) белок-кодирующих генов (экзома), а не выбор платформы. На рынке секвенирования

доступны различные наборы: Agilent SureSelect XT, Agilent SureSelect QXT, NimbleGen SeqCap EZ и Illumina Nextera Rapid Capture Exome, которые включают биотинилированную ДНК, гибридизирующуюся с фрагментами ДНК библиотеки, но отличающихся методом фрагментации ДНК, длиной, область охвата генов-мишеней [Suwinski *et al.,* 2019]. Применение WES также значительно уменьшает объем анализируемой базы данных до 5–6 ГБ по сравнению с WGS (90 ГБ) [AllSeq, 2018].

Что касается научных исследований, то здесь задача несколько иная: получить больше информации о геноме в целом. Поэтому, в масштабных научных геномных проектах более активно используют полногеномное секвенирование, например, в проектах «100 000 геномов», «Российские геномы» и др. [Zhernakova *et al.,* 2020]. В ближайшее время, вероятно, в рутинной практике начнут применять комбинацию методов массового параллельного секвенирования, что позволит идентифицировать до 99% всех известных и ранее не выявленных мутаций.

Благодаря внедрению NGS, сегодня достигнут существенный прогресс в диагностике орфанных заболеваний. С помощью новых технологий стало возможным уточнить частоты распространенности многих моногенных заболеваний. В частности, было обнаружено, что в Северо-Западном регионе наиболее распространенной патологией является не муковисцидоз, а болезнь Штаргардта [Barbitoff *et al.,* 2019]. Повысилась диагностическая эффективность некоторых заболеваний олигогенной природы. Например, при использовании экзомного секвенирования более чем в 10 раз возросла эффективность диагностики мутаций, связанных с МОДИ-диабетом [Glotov *et al.,* 2019].

Экзомное секвенирование становится методом поиска новых генетических маркеров также и мультифакторных заболеваний. Так, в нашем исследовании благодаря данному методу и уникальному биоинформатическому подходу удалось выявить более 10 новых маркеров сахарного диабета 2-го типа [Barbitoff *et al.,* 2018].

Технологические особенности секвенирования

Как и любая другая технология, NGSимеет ряд недостатков, связанных именно с техническими особенностями. Так, получение большого объема данных при WGS и WES связано с проблемой их анализа и обработки: часто бывает трудно отличить некоторыегенетические варианты от случайных ошибок, возникающих в процессе секвенирования [Hofmann *et al.,* 2017]. Также основным ограничением WES является неравномерность покрытия зондами последовательностей генов-мишеней генома, что приводит к появлению областей с низким покрытием, которые в результате последующего анализа препятствуют точному аннотированию или пропуску локальных генных вариантов [Wang Q. *et al.,* 2017]. Данные WES могут характеризоваться изменчивостью, т.е. содержать аномалии и выбросыили иметь непостоянную скорость загрузки в репозиторий. Могут встречаться и наследственные ограничения: GC ошибка, проблемы с аннотированием паралогичных последовательностей и фазированием аллелей, ассоциацией вариантов с биологическими свойствами и фенотипом. Также трудности возникают с трансляцией данных секвенирования в понятные медицинские шаблоны, подобно формам представления клинико-диагностических тестов [Suwinski *et al.,* 2019].

Биоинформатическая обработка является также частью технологии NGS. Поэтому установление генетической природы заболевания во многом зависит от качественного биоинформатического протокола анализа данных секвенирования [Barbitoff *et al.,* 2017; 2020]. Необходимо учитывать, что в референсной последовательности встречаются ошибки, связанные с так называемыми референсными минорными вариантами —RMA (позициями референсного генома, в которые инкорпорирован редкий или даже патогенный вариант). Такие ошибки необходимо корректировать при проведении биоинформатического анализа [Barbitoff *et al.,* 2018].

Важно отметить, что на протяжении более чем 20 лет при секвенировании генома человека, животных, бактерий и вирусов преобладала автоматизированная технологияСэнгера. Однако необходимость более быстрогоскрининга генома стимулировала развитие новыхтехнологий NGS - мультиплексного секвенирования ДНК.В нашем исследовании было проведено сравнение двух методов секвенирования (Сенгер и NGS) и оценка их эффективности. На образцах ДНК потенциальных доноровгемопоэтических клеток, был проведен сравнительныйанализ технологий. Было установлено, что метод NGS позволяет выявлять редкие или новые варианты аллелей. Этот подход подтвержден в качестве болеечувствительного и более экономичного, особенно вбольших лабораториях по HLA-типированию [Glotov O. *et al.*, 2018].

Тем не менее несмотря на то, что секвенирование по Сенгеру менее производительно и не так экономически выгодно, оно сохраняет актуальность для ряда практических задач и чаще всего используется для поиска новых или известных мутаций в сравнительно небольшой области молекулы ДНК (от 100 до 1200 п.н.) и в качестве «золотого стандарта» для подтверждения мутаций, найденных у пациента методом NGS (рис.2).

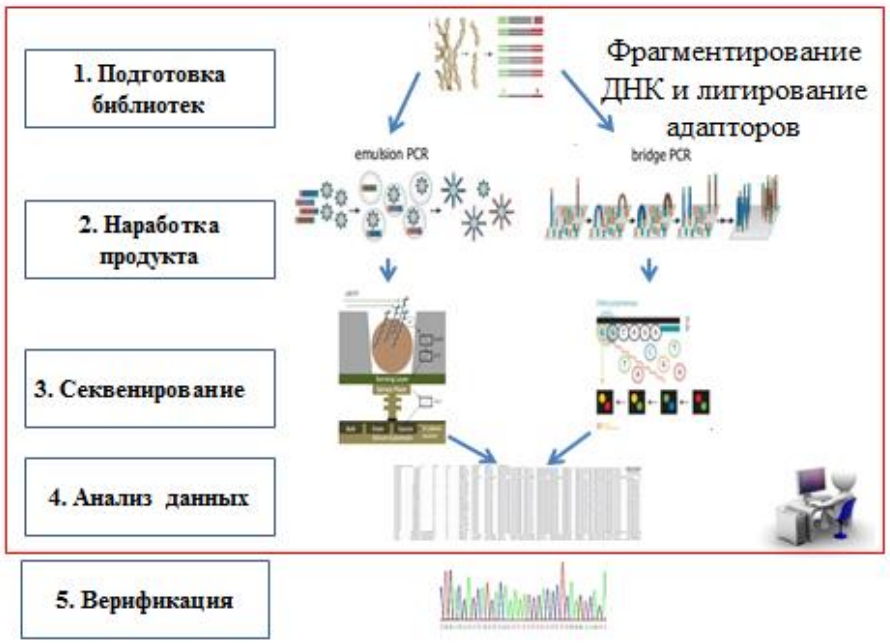


Рисунок 2. Этапы секвенирования следующего поколения.

С помощью секвенирования по Сенгеру удобно осуществлять поиск клинически актуальных мутаций в «горячих точках» хорошо изученных генов, определять нуклеотидную последовательность коротких (содержащих мало экзонов) и безынтронных генов, изучать короткие повторы [Баранов *и др.,* 2021], а также при поиске «второго» патогенного варианта после или вместе с NGS [Федяков *и др.,* 2021].

Генетическая терминология

Необходимо отметить существенно важное изменение в генетической терминологии, связанное с использованием новых методов. Вместо широко распространенных терминов «мутация» и «полиморфизм» с 2015 г. в США Американский колледж медицинской генетики и геномики (англ. American College of Medical Genetics and Genomics — ACMG), с 2017 Российское общество медицинской генетики рекомендуют использовать термин «вариант нуклеотидной последовательности» со следующими пятью характеристиками: патогенный (pathogenic); вероятно патогенный (likely pathogenic); неопределенного значения (uncertain significance); вероятно доброкачественный (likely benign); доброкачественный (benign) [Richards *et al.*, 2015; Рыжкова *и др.,* 2017, 2019]. Инструменты для правильного описания вариантов нуклеотидной последовательности в соответствии с номенклатурой HGVS представлены на специальном сайте [https://mutalyzer.nl.]. Данную терминологию активно стали применять не только в диагностическом плане [Рыжкова *и др.,* 2019], но и при проведении различных исследований. Благодаря использованию термина вариант с пятью характеристиками удается более четко характеризовать функцию того или иного варианта в геноме.

Сегодня есть четкое понимание, что генетический вариант является основным носителем предикции патогенности заболеваний с двумя основными характеристиками:

- пенетрантностью (процент носителей соответствующего генотипа, у которых проявляется признак;
- экспрессивностью (варьирующее проявление признака у особей с одинаковым генотипом [Инге-Вечтомов, 2010].

Таким образом, термины, предложенные еще в 1925 году Тимофеевым–Ресовским, оказались настолько важными и опережающими свое время [Инге-Вечтомов, 2010], что сейчас их сущность позволят объяснить, почему такой термин как «мутация» уходит в прошлое, и теперь остается только термин «вариант» в пяти состояниях [Рыжкова *и др.,* 2019].

Перспективы геномной медицины

Знание о структуре гена, особенностях генетического полиморфизма и функциях различных вариантов в геноме с учетом популяционной специфичности дает понимание о наследственной природе того или иного моногенного или мультифакториального заболевания (МФЗ), и способствует диагностике, профилактике и лечению этих болезней, а также меняет классификацию заболеваний человека (рис.3), приводит к смене парадигм, способствует появлению и быстрому развитию новой науки — молекулярной медицины [Collins, McKusick, 2001; Peltonen, McKusick, 2001].

В настоящее время ситуация в сфере здравоохранения такова, что традиционная медицинская помощь адресована, как правило, уже заболевшему человеку. На протяжении всего исторического развития информация о здоровом человеке и так называемых донозологических состояниях, предшествующих развитию болезни, часто оставалась в тени. Сегодня здравоохранение стоит на пороге больших перемен. Основными звеньями работы с пациентом должны стать предикция (предсказание) риска заболевания, ранняя (возможно – доклиническая) диагностика с определением его стадии и своевременное адекватное вмешательство (фармакология, питание и др.) с целью профилактики заболевания или его перехода в тяжелую стадию. Эти принципы легли в основу

принципиально новой стратегии «трех П» - предиктивной, превентивной и персонализированной медицины- ПМ [Баранов *и др.,* 2000; Тайц, 2019]. Следует отметить, что становление терминов ПМ и генетического паспорта (ГП) имеет место в Санкт-Петербурге еще с 2000 г. и сами термины постепенно внедряются в общественное сознание [Баранов *и др.,* 2021].

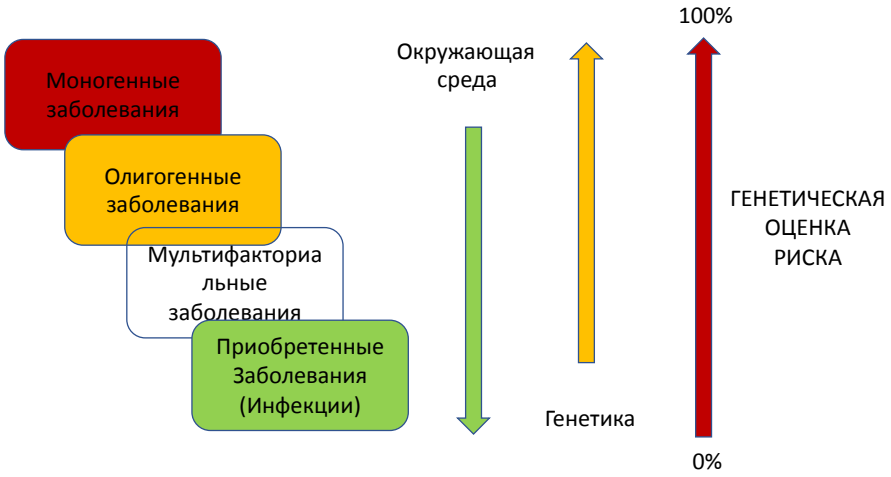


Рисунок 3. Классификация заболеваний человека с генетической точки зрения.

Основные этапы эволюции ПМ от медицины 3П (предиктивная, превентивная, персонализированная) к медицине 4П (participatory) и далее к трансляционной, интегративной медицине, завершающиеся «точной» (precision) медициной, показаны рисунке 4.

Рисунок 4. От генетического паспорта к NGS генетическому клиническому паспорту здоровья (по Баранов *и др.,* 2021 с изменениями).

Начиная с 2008 г. методом физического картирования было идентифицировано более 1500 генов, ассоциированных с МФЗ человека. Многие из генов риска-кандидатов МФЗ были изучены в различных отечественных лабораториях и центрах, дополнительные сведения об этом можно найти в монографиях и обзорах ряда авторов [Баранов *и др.,* 2000; Баранов, Хавинсон, 2001; Пузырев *и др.,* 2007]. Информацию о генах-кандидатах, а также о вариантах, ассоциированных с МФЗ, можно почерпнуть в различных международных базах и каталогах: OMIM [http://omim.org], HUGENAVIGATOR [https://phgkb.cdc.gov/PHGKB/hNHome.action], HumanGeneMutationDatabase [http://hgmd.cf.ac.uk/ac/index.php], GWAS [https://ebi.ac.uk/gwas], ClinVar [http://etal.,.ncbi.nlm.nih.gov/clinvar], dbSNP [http://ncbi.nlm.nih.gov/snp], HumanGenomeVariationSociety [http://hgvs.org/dblist/dblist.html], LeidenOpenVariationDatabase [http://lovd.nl], DECIPHER [https://decipher.sanger.ac.uk], база проекта «1000 геномов» [http://browser.1000genomes.org/index.html], база данных экзомов ExomeAggregationConsortium [http://exac.broadinstitute.org], объединенная база данных геномов GenomeAggregationDatabase [http://gnomad.broadinstitute.org] и др.

По мнению известных авторитетов в области генома человекам - Френсиса Колинса, Виктора МакКьюсика, Лиины Пелтонени других - золотой век предиктивной медицины наступит через 5-10 лет и медицина 4П постепенно эволюционирует в медицину 10П и будет активно привлекать самих пациентов [Тайц, 2019]. Медицина 10П будет включать следующие разделы [Тайц, 2019]:

1. Предсказательная (Predictive);
2. Превентивная (Preventive);
3. Персонализированная (Personalized);
4. Партисипативная (с участием пациента) (Participatory);
5. Практическая (Practical);
6. Непрерывная, постоянная (Permanent);

7. Проактивная, упреждающая (Proactive);
8. Позитивная (Positive);
9. Точная (Precision);
10. Пропагандистская (Promotional).

Таким образом, молекулярная медицина и ее дочерние направления — предиктивная медицина (генетический паспорт, фармакогенетика, генная терапия) — представляют результат широкого внедрения генетических знаний в медицинскую науку. Характерной особенностью молекулярной медицины, основанной на молекулярной структуре генома человека, является ее индивидуальный характер. В силу уникальности индивидуального генома молекулярная медицина направлена на коррекцию патологического процесса у конкретного больного с учетом его уникальных генетических особенностей [Баранов, 2000; Баранов, Баранова 2018]. Другая важнейшая особенность молекулярной медицины — профилактическая направленность. Полные сведения о геноме могут быть получены задолго до начала заболевания, а при необходимости — еще до рождения, поэтому правильно организованная профилактика может способствовать полной ликвидации или в значительной мере предупреждению развития тяжелого заболевания.

Вклад геномики в медицину трудно переоценить. Ее наиболее значимые достижения включают:

- точные, эффективные и универсальные методы диагностики наследственных болезней на любой стадии онтогенеза, в том числе и до рождения (пренатальная диагностика);
- новые научно-практические направления геномной медицины — онкогеномику, кардиогеномику, иммуногеномику, геномику репродуктивного здоровья, геномику старения, нутригеномику, спортивную геномику, метаболомику, протеомику, микробиомику и пр.; молекулярные тесты для идентификации личности — так называемую геномную дактилоскопию; изучение микробиома;

- экспериментальные и клинические основы генной терапии наследственных и ненаследственных болезней;
- клеточные модели наследственных болезней и разработку методов эффективного редактирования генома;
- диагностику и персонализированное лечение (фармакогенетика и фармакогеномика) пациентов с моногенными и частыми МФЗ;
- неинвазивную диагностику генных и хромосомных болезней на всех стадиях развития плода и диагностику редких (орфанных) заболеваний методом NGS;
- преконцепционное генетическое тестирование и молекулярные основы профилактической (предиктивной) медицины.

Поэтому сейчас мы уже можем говорить о технологиях NGS на «страже» здоровья человека:

- экзом/клинический экзом для NGS секвенирования образца ДНК больного ребенка и обследование семьи (носительство);
- генетический паспорт;
- ПГД;
- НИПТ.

Основными ограничениями применения технологий целевого секвенирования (WGS, WES, RNA-seq, ChIP-seq) в клинической практике является их высокая стоимость, в которую, помимо затрат на реактивы и оборудование, включаются расходы на хранение, передачу, обработку и биоинформационный анализ данных [Schmidt and Hildebrandt, 2017]. Например, во Франции стоимость секвенирования генных панелей в зависимости от количества анализируемых генов составляет 376-968 евро [Marino *et al.*, 2018]. Tan с коллегами на основании проанализируемых 10 исследований в США рассчитали среднюю стоимость секвенирования генной панели одного образца в размере 1609 долларов США (488-3443 долларов США) [Tan O. *et al.*, 2018]. Другая проблема состоит в клинической интерпретации данных секвенирования.

Знание данных секвенирования недостаточно для объяснения этиологии, патогенеза и симптомов многофакторных заболеваний, на развитие которых кроме генетических факторов сильно влияют внешние факторы и образ жизни [Lightbody *et al.,* 2019; Suwinski *et al.,* 2019].

Сегодня для реализации концепции предиктивной медицины, по инициативе руководителя программы «Геном человека», а ныне директора Национального института здоровья США Френсиса Коллинза, был создан специальный институт (Patient-Centered Outcomes Research Institute). Для улучшения качества диагностики, профилактики и лечения частых заболеваний сотрудники института должны составлять медицинские карты больных, содержащие результаты лабораторных анализов в сочетании с данными индивидуального генома. В 2015 г. при участии и с одобрения президента США Барака Обамы в США была запущена программа «Точная медицина» (Precision Medicine) с включением 1 млн человек и проведением масштабных клинических и геномных исследований. По мнению Френсиса Коллинза, результаты исследования на такой большой группе позволят получить доказательства, подтверждающие концепцию точной медицины. Идеологическим продолжением проекта «Точная медицина» является американский проект — «Мы все» (Allofus) [https://allofus.nih.gov/news-events], направленный на решение задач персонализированной медицины. Его цель — поиск и совершенствование путей интерпретации геномных и медицинских данных. В проекте участвуют крупнейшие американские университеты, предусмотрен сбор данных о здоровье более 200 000 добровольцев, а на основании полученных образцов ДНК, будут протестированы 59 генов тяжелых наследственных заболеваний. Наряду с проектом «Геномы англичан» проект «Мы все» будет представлять одну из самых больших в мире баз данных геномной и клинической информации. На март 2022 года уже получена информация о более чем 100000 полных геномах и 165000 чиповых датасетов [https://allofus.nih.gov/news-

events/announcements/research-roundup-genomic-data-release-opens-new-paths-discovery].

Уместно отметить, что двумя годами раньше (2013 г.) было создано Европейское общество предиктивной, превентивной и персонализированной медицины (EPMA), которое в 2015 г. опубликовало программу «Персонализированная медицина для европейцев: на пути к более точной медицине с целью диагностики, лечения и предотвращения болезни» [Golubnitschaja *et al.,* 2016; https://epmanet.eu].

К сожалению, ни в американской программе ПМ, ни в задачах Европейского общества не упоминаются пионерские исследования по ПМ в России, результаты которых неоднократно докладывались и были опубликованы не только в России, но и за рубежом [Баранов *и др.,* 2021].

Сегодня во всех программах подчеркивается, что особенно важную роль в прогрессе геномной медицины сыграл метод NGS, который значительно повысил клиническую значимость тестирования, хотя и не решил всех проблем ПМ. В 2016 г. Комитет FDA (США) снял с американской фирмы «23 andme» запрет на использование прогностического генетического тестирования (ГТ) десяти частых МФЗ «по запросу потребителя», в том числе болезни Крона, Альцгеймера, Паркинсона, рака простаты, молочной железы и др. [Greens, 2015]. В настоящее время многие крупные западные фирмы (например, Systemas Genomicos) предлагают генетические тесты для оценки наследственной предрасположенности к 100 МФЗ. Точность прогнозирования некоторых заболеваний (болезнь Крона, рак простаты) при таком ГТ возросла до 20 % (рак простаты) и даже до 80 % (болезнь Крона). Успех объясняется созданием сложных панелей с высокой плотностью маркеров, позволяющих одновременно исследовать несколько сотен генов предрасположенности или специальных локусов, включающих кластеры таких генов. Поиск патогенных вариантов осуществляется методом секвенирования ДНК с дальнейшим биоинформатическим анализом [Manolio *et al.,* 2010; Carere *et al.,* 2016].

Таким образом, прогресс современной ПМ, ее практическая ценность зависят от качества секвенирования генома и анализа его функций с позиции системной генетики. Согласно EPMA программе дорожная карта ПМ в качестве основного компонента включает массовое секвенирование индивидуальных геномов с целью выяснения их популяционных, этнических, социальных и даже межтканевых особенностей. Посредством интегративного анализа экспрессии генов, белок-белковых взаимодействий формируют индивидуальные омиксные профили, которые сопоставляют с результатами клинических и лабораторных данных того же пациента. С учетом этих сведений создают интегрированные генные сети участвующих в патологическом процессе органов и систем пациента и анализируют варианты развития патологического процесса. Следовательно, сам пациент является не только источником информации, но и пользователем данных ПМ [Баранов *и др.,* 2021].

Сегодня эволюцию медицинской генетики и ПМ можно представить в виде хронологической шкалы развития научных направлений и технологических решений (рис.7), где важным для перехода на следующий этап являются работы отечественных и зарубежных научных коллективов по изучению генетических основ формирования моногенных, мультифакториальных и инфекционных заболеваний человека.

Важно подчеркнуть, что разработка научных основ точной медицины, для изучения, диагностики и лечения как моногенных болезней, так и олигогенных, мультифакториальных и инфекционных заболеваний будет определяться эффективностью использования NGS технологий с учетом современных алгоритмов анализа и классических генетических понятий экспрессивности и пенетрантности. Дальнейшие этапы практического внедрения генетического тестирования включают: досимптоматическое (упреждающее) генетическое тестирование (ГТ) в семьях высокого риска; проспективное ГТ с обязательным последующим мониторингом состояния лиц групп высокого риска по

результатам тестирования; рандомизированное предиктивное тестирование [Barnoy, 2007, Баранов *и др.,* 2021].

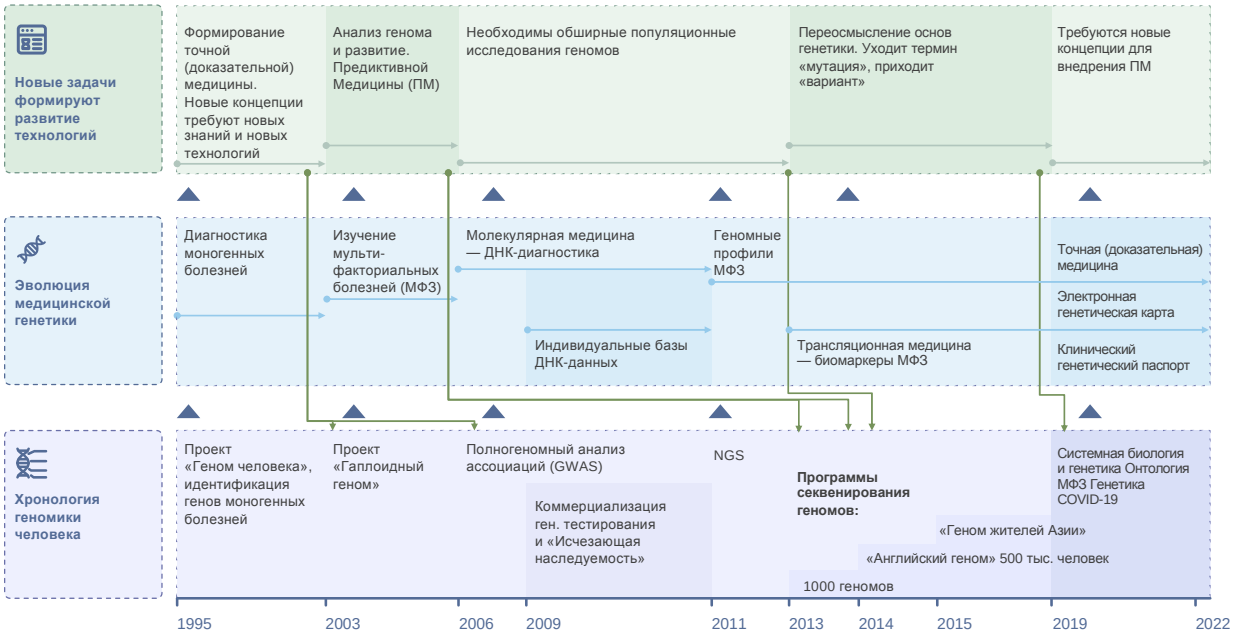


Рисунок 5. Хронологическая шкала развития научных направлений и технологических решений в медицинской генетике.

Цель настоящего исследования – на основании результатов экзомного секвенирования определить факторы риска социально-значимых заболеваний и разработать методологические подходы для выявления клинически значимых генных вариантов с целью оценки риска развития моногенной, олигогенной и мультифакториальной патологии, тяжести протекания некоторых вирусных инфекционных заболеваний у человека на примере Северо-Западного региона России.

Задачи:

1. На основании результатов экзомного секвенирования охарактеризовать структуру моногенных заболеваний и генных вариантов, приводящих к ним, в Северо-Западном регионе России.

2. Оценить эффективность использования NGS-технологии в детекции ранее не описанных патогенных вариантов генов, ассоциированных с моно- и олигогенными заболеваниями, в том числе - при сочетанных патологиях.

3. Сравнить эффективность использования технологии NGS с другими молекулярно-генетическими методами на основе ПЦР в детекции патогенных вариантов у пациентов с моногенным сахарным диабетом и болезнью Вильсона-Коновалова.

4. На основании результатов экзомного секвенирования оценить вариативность проявления патологий олигогенной и мультифакториальной природы в небольших когортах пациентов Северо-Западного региона России.

5. Оценить эффективность прогностического моделирования фенотипа пациента с применением регрессионного метода на основе диагностики генных вариантов и их сочетаний.

6. Описать спектр генетических вариантов, выявленных с помощью NGS, ассоциированных с различной степенью тяжести и исходом новой коронавирусной инфекции COVID-19.

7. Разработать комплекс генетических обследований, включающих экзомное секвенирование, для предсказания вероятности развития олигогенной и мультифакториальной патологии, тяжести протекания некоторых инфекционных заболеваний и объяснения патогенетического характера клинических проявлений ряда генных болезней у человека.

Научная новизна исследования

В результате применения NGS для секвенирования ДНК пациентов с моногенными заболеваниями с использованием оригинальных протоколов биоинформатической обработки по данным международных баз, и собственным результатам когортных исследований получены новые данные о распространенности моногенных заболеваний в Северо-Западном регионе РФ.

Показано, что 24,3% вариантов в российской популяции ранее не были описаны в мировой литературе. Показано, что наиболее часто встречающиеся наследственные моногенные болезни с рецессивным типом наследования являются: фенилкетонурия, недостаточность фактора VII4, синдром Элерса-Данлоса, кифосколиотический тип, 2; тирозиназонегативный кожно-глазной альбинизм; болезнь Вильсона-Коновалова. Впервые показано, что в Северо-Западном регионе России наиболее часто встречаются патогенные варианты в генах: *ABCA4* (дистрофия сетчатки, болезнь Штаргардта) и *CFTR* (муковисцидоз).

Впервые описана частота таких вариантов, как rs554847663 в гене *OTOG,* связанный с аутосомно-рецессивной глухотой, и вариант rs119473033 в гене *SMARCAL1*, вызывающий иммуно-костную дисплазию Шимке. Выявлено более 100 предполагаемых вариантов потери функции (pLoF), присутствующих у здоровых пациентов на момент обследования. Оценена распространенность варианта c.894G>A (1/262) и частота гетерозиготного носительства всех патогенных вариантов (1/130) в гене *LIPA* в российской популяции.

Установлено, что в референсном геноме есть ошибки и их необходимо исправлять для проведения корректного биоинформатического анализа. Впервые с использованием NGS описаны новые патогенные варианты в генах *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT* и проанализирована клиническая картина у пациентов с более чем одним генетическим вариантом в одном или разных генах-мишенях для моногенного сахарного диабета (МОДИ) и болезни Вильсона-Коновалова. Впервые в РФ описаны клинические случаи нескольких наследственных заболеваний у одного человека: совместного наследования Х-сцепленной и аутосомно-доминантной форм ихтиоза, болезни Вильсона-Коновалова и гемохроматоза. Показано, что эффективность NGS увеличивает диагностическую выявляемость с 15 до 50 % (для МОДИ), с 75 до 96% для болезни Вильсона-Коновалова.

Впервые в РФ проведена оценка клинического полиморфизма патологий, позволяющие понять природу заболеваний и предположить олигогенный характер наследования для гипертрофической кардиомиопатии, моногенного диабета, семейной гиперхолестеринемии. Впервые предложены 4 модели наследования для наследственных кардиомиопатий, из которых информативными оказались только 2 доминантные модели. Показана неполная пенетрантность для следующих вариантов: *MYBPC3* (c.977G>A и c.2678G>T), *CASQ2* (c. 1014+12delG). Впервые показано, что варианты в гене *TNNT2* (c.97+151delC, c.223+92G>C и c.223+93C>G) являются «защитными» от кардиомиопатии.

Впервые в РФ обнаружены генетические варианты в различных генах-мишенях у одного пациента: *GCK и HNF1A; GCK и BLK; GCK, BLK и WFS1* и показано, что клиническая картина у этих пациентов более типичной для МОДИ2. Впервые показана эффективность идентификации генов-кандидатов сахарного диабета 2 типа и ожирения при помощи экзомного секвенирования и оригинальных биоинформационных подходов для небольших когорт больных.

Впервые выявлены ассоциации вариантных локусов rs328 в гене *LPL*, rs11863726 в гене *HBQ1*, rs112984085 в гене *VAV3* для СД2 и ожирения, rs6271 в гене *DBH*, rs62618693 в гене *QSER1*, rs61758785 в гене *RAD51B*, rs34042554 в гене *PCDHA1* и rs144183813 в гене *PLEKHA5* для ожирения; и rs9379084 в гене *RREB1*, rs2233984 в гене *C6orf15*, rs61737764 в гене *ITGB6*, rs17801742 в гене *COL2A1* и rs685523 в гене *ADAMTS13* для СД2.

Впервые в РФ обнаружено, что rs328 в гене *LPL*, связан с СД2 и ожирением одновременно, rs6271 в гене *DBH* и rs62618693 в гене *QSER1* являются специфическими маркерами ожирения, rs2233984 в гене *C6orf15* был идентифицирован как специфический маркер СД2. Впервые обнаружены новые специфические варианты для СД2: rs9379084 в гене *RREB1*, rs61737764 в гене *ITGB6* и rs17801742 в гене *COL2A1*, rs139972217 в гене *TMC8*, rs61758785 в гене *RAD51B*, rs34042554 в гене *PCDHA1* и rs144183813 в гене *PLEKHA5*.

Впервые показана эффективность применения регрессионной модели для оценки прогноза фенотипа на основе данных о генотипах в генах *EFEMP1, ZBTB38, HHIP, LCORL, ADAMTSL3, CDH13, JAZF1, IGF1R, GHSR, CABLES1, IFNG, VDR3, IGFBP3* для роста человека, 36 генетических маркеров для предсказания количественных признаков (ИМТ, ОБ, ХС, ХС ЛПНП, ХС ЛПОНП), генах *AGTR2, NOS3, CNB1, ADRB2* для жизненной емкости легких (ЖЕЛ). С применением этой модели впервые показана большая генетическая детерминированность ЛПОНП по сравнению с ИМТ, ОБ, ХС, ХС ЛПНП, ХС.

Впервые в РФ приводится биоинформатическое обоснование, что идентификация факторов риска моногенных заболеваний путем секвенирования экзома поможет в снижении риска не только моногенных, но и МФЗ, подчеркивается отсутствие четкой грани между моногенными, олигогенными и МФЗ.

Впервые в РФ разработана прогностическая модель развития цитокинового шторма (ЦШ) у пациентов с диагностированной новой коронавирусной инфекцией COVID-19 и изучены генетические факторы риска развития тяжелых форм новой коронавирусной инфекции COVID-19. Показано, что количество мутаций в S-белке выше у пациентов 3 степенью тяжести по сравнению с более легкими. Выявлено, что мутации в линиях SARS-CoV-2 нероссийского генеза ассоциированы с повышенным риском летального исхода в группе российских пациентов.

Впервые установлено, что частоты пяти вариантов гена *ACE2* (rs35803318, rs41303171, rs113691336, rs971249, rs2285666 не различаются) в русской и европейских популяциях. Показано, что редкие варианты в гене *ACE2* могут косвенно играть роль в патологии новой коронавирусной инфекции COVID-19, влияя на важные нормальные функции белка и тяжесть болезни.

Впервые в РФ проведено секвенирование 840 экзомов у пациентов с COVID-19. Установлено 11 генетических вариантов в генах *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX,*

которые ассоциированы с количественными признаками, связанными с тяжестью и исходом новой коронавирусной инфекции COVID-19. Выдвинуто предположение, что варианты в генах *ATXN1, PKHD1, SETX, PIEZO1* и *CDH23* оказывают прямое влияние на фенотип новой коронавирусной инфекции COVID-19 путем изменения функции (в случае миссенсных вариантов в генах *ATXN1* и *CDH23*) или уровней экспрессии своего соответствующего гена; для трех из этих вариантов найдено подтверждение их роли в патогенезе новой коронавирусной инфекции COVID-19.

Впервые обоснована необходимость отказа от термина «мутация» и перехода к термину «вариант» как при анализе моногенных, так олигогенных, МФЗ и новой коронавирусной инфекции COVID-19. Впервые в РФ обоснована необходимость создания новой генетической классификации болезней с опорой не только на клинико-лабораторные данные, но и на результаты молекулярно-генетического анализа с учетом пенетрантности и экспрессивности генетических вариантов и предложена концепция генетического клинического паспорта (ГКП) как методического инструмента для последующей оценки риска всех типов заболеваний для использования в Предиктивной медицине (ПМ).

Теоретическая и практическая значимость результатов исследования

Полученные данные дополняют имеющиеся представления относительно молекулярно-генетических механизмах моногенных (ДЛКЛ, тугоухость, болезнь Вильсона-Коновалова, ихтиоза, синдром Ноя-Лаксовой, синдром Floating Harbor, анаукзетическая дисплазия), олигогенных (аритмогенная кардиомиопатия/дисплазия правого желудочка, семейная гиперхолестеринемия, моногенный и неонатальный диабет) и мультифакториальных заболеваний (СД2, ожирение), устойчивости к новой коронавирусной инфекции COVID-19.

В результате работы установлены механизмы возникновения изученных заболеваний, основанные на присутствии в генотипе специфических вариантов в одном или разных генах-мишенях. Полученная информация о популяционной

частоте аллелей известных и новых патогенных вариантов в генах *ABCA4, ABCC8, ACE2, ADAMTSL3, ADAMTS13, ADRB2, AGTR2, ALDOB, ALMS1, ALOXE3, APOB,ATP7B,ATXN1, BCKDHB, BLK, BLM, C6orf15, CABLES1, CASQ2, CCNO, CDH13, CDH23, CFTR, COL2A1, COL7A1, CPLANE1, DBH, DNAJB2, EFEMP1, EIF2AK3, EOGT, F7, F8, FGG, FKBP14, FLG, GAA, GABBR2, GALT, GATA6, GCK, GDAP1, GHSR, GJB2, HBQ1, HHIP, HNF1A, HFE, HTT, JAZF1, IFNG, IGF1R, IGFBP3, ITGB6, KCNJ11, LCORL, LDLR, LIPA, LIPC, LPL, LZTR1, MSH2, MTO1, MYBPC3, MYH14, NEB, NOS3, NPC1, NPHS2, OTOG, PAH, PAX4, PCDHA1, PHGDH, PIEZO1, PKHD1, PKP2, PLEKHA5, QSER1, RAD51B, RMRP, RREB1, SBF1, SCN11A, SETX, SLC19A2, SLC26A2, SMARCAL1, SRCAP6, STAC3 STS, SURF1,TGM5, TNNT2, TYR, VAV3, VDR3, WFS1, ZBTB38* в РФ может быть использована для интерпретации результатов NGS в клинической практике и дополнить отечественные и международные базы знаний.

Предложены подходы в биоинформатике, которые учитывают ошибки референсного генома, популяционные частоты и другие особенности для корректной интерпретации данных секвенирования.

На основании проведенного исследования предложены алгоритмы генетической диагностики для МОДИ, болезни Вильсона-Коновалова, аритмогенной кардиомиопатии/дисплазии правого желудочка, семейной гиперхолестеринемии, и других заболеваний. Выявленные новые данные о наличии протективных вариантов, в частности, в гене *TNNT2* (c.97+151delC, c.223+92G>C и c.223+93C>G) расширяют наши представления о патогенезе заболеваний.

Обнаруженные генетические варианты в различных генах-мишенях у одного пациента позволяют скорректировать тактику его консультирования, наблюдения и лечения. На примере гена *GCK* показано, что патогенные варианты в нем могут быть ассоциированы с разными нозологическими формами, что позволяет предположить, что нозологические формы — это всего

лишь различные варианты изменений в работе одних и тех же основных генов с учетом существенного вклада факторов внешней среды.

Разработанные оригинальные алгоритмы и биоинформационные подходы для небольших когорт больных показали эффективность технологий секвенирования всего экзома (WES) для поиска новых маркеров многофакторных заболеваний в когортах ограниченного размера в плохо изученных популяциях.

Найденные специфические и неспецифические генные маркеры для СД и ожирения позволяют проводить дифференциальную диагностику и последующую индивидуальную терапию влияя на специфические метаболические нарушения, характерные для определенной нозологии. Предложены и апробированы новые GLM модели для предсказания роста, ИМТ, ОБ, ХС, ХС ЛПНП, ХС ЛПОНП и ЖЕЛ. Показана эффективность применения регрессионной модели для оценки прогноза фенотипа на основе генетических, анамнестических и клинических данных, которая может быть полезна для создания полигенных предикторов риска МФЗ.

Приведено обоснование того, что внедрение в геномную медицину секвенирования экзома полезно, так как может помочь выявить редкие состояния, которые скрыты в рамках сложной диагностики МФЗ. Предложены алгоритмы использования NGS для НИПТ и ПГД. Обоснована необходимость комплексного подхода к планированию беременности с использованием всего арсенала молекулярно-генетических, цитогенетических, эмбриологических методов. Показано, что современные представления о генетическом варианте концептуально меняют наши знания о том, какие болезни являются моногенными, олигогенными и мультифакториальными.

Предложенная прогностическая модель развития цитокинового шторма у пациентов с диагностированной новой коронавирусной инфекцией COVID-19 может быть использована в клинической практике для выявления на ранней стадии заболевания пациентов с худшим прогнозом. Представленные клинико-

генетические корреляции течения цитокинового шторма при новой коронавирусной инфекции COVID-19 могут быть использованы для определения прогноза заболевания и показаний для проведения медико-генетического консультирования.

Редкие варианты в гене *ACE2* могут косвенно играть роль в патологии новой коронавирусной инфекции COVID-19, поэтому о необходимо более тщательное наблюдение пациентов с редкими патогенными вариантами в случае заражения новой коронавирусной инфекцией COVID-19. Выявленные новые данные о корреляции тяжести и летальности новой коронавирусной инфекции COVID-19 с количеством мутаций в S-белке, происхождением вируса SARS-COV-2 (российский или нероссийский) можно использовать для прогноза заболевания. Найденная схожая частота пяти вариантов (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) в гене *ACE2* в русской и европейских популяциях предполагает схожий уровень инфицирования и тяжести заболевания. Выявленные новые 11 вариантов в генах *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX*, ассоциированные с количественными признаками и связанные с тяжестью и исходом заболевания можно использовать для прогноза исхода новой коронавирусной инфекции COVID-19 с показателем ROC/AUC = 0,59. Невысокий уровень репликации результатов подчеркивает особую роль дизайна исследования и структуры популяции для достоверного выявления генетических факторов риска инфекционных заболеваний.

Разработана концепция генетического клинического паспорта здоровья и новая методология определения генетических детерминант с последующей интерпретацией найденных вариантов, которая подтверждает клиническое значение генетических предикторов при формировании групп риска моногенных, мультифакториальных заболеваний и новой коронавирусной инфекции COVID-19.

Основные положения выносимые на защиту

1. Сопоставление результатов NGS ДНК пациентов с моногенными заболеваниями с использованием оригинальных биоинформатических протоколов, как с данными международных баз, так и с популяцией Северо-Западного региона России, позволяет определять новые патогенные варианты в генах *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT,* и выявлять патологию, обусловленную их сочетанием (совместное наследование Х-сцепленной и аутосомно-доминантной форм ихтиоза, болезни Вильсона-Коновалова и гемохроматоза).

2. Применение NGS по сравнению с классическими молекулярно-генетическими методами (ПРЦ, ПЦР-ПДФР, секвенирование по Сенгеру) у пациентов с клинической картиной моногенного сахарного диабета и болезни Вильсона-Коновалова эффективно в связи с увеличением выявляемости патогенных вариантов с 15 до 50 % и с 75 до 96%, соответственно.

3. В Северо-Западном регионе России вариативность проявления патологий как олигогенной, так и мультифакториальной природы обусловлена не только экспрессивностью, пенетрантностью, но и комплексными гаплотипами, ассоциированных с заболеваниями генов, выявленных с помощью экзомного секвенирования и оригинального биоинформатического анализа, адаптированного для небольших когорт.

4. Варианты в генах *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX,* в том числе редкие в гене *ACE2:* rs35803318, rs41303171, rs113691336, rs971249, rs2285666, выявленные с помощью NGS, ассоциированы с тяжестью, оцененной в соответствии с разработанной шкалой, и исходом новой коронавирусной инфекции COVID-19.

5. Комплекс генетических обследований - «генетический клинический паспорт здоровья человека», включающий экзомное секвенирование и позволяющий предсказать возможное развитие олигогенной и

мультифакториальной патологии, тяжесть протекания некоторых инфекционных заболеваний и объяснить патогенетический характер клинических проявлений болезней разной природы, с учетом не только экспрессивности, пенетрантности, но и сочетания патогенных вариантов в генах, может быть новой основой Предиктивной медицины.

Степень достоверности и апробации результатов исследования

Достоверность обеспечивается многообразием применяемых методов, соответствующих цели и задачам исследования, статистической значимостью результатов, согласованностью полученных данных и имеющихся результатов клинических и экспериментальных исследований по теме диссертации, а также репрезентативным объемом выборок (суммарно образцов биологического материала от более 4670 пациентов).

Основные результаты по теме диссертационной работы представлены и обсуждены в виде устных докладов на российских и международных конференциях и конгрессах, в том числе: Biologie Prospective – Santorini Conference, 2004, Santorini Island; II International conference «Medicine of longevity and quality of life», 2006, Moscow, Russia; VI European Congress Internation Assosiation of Gerontology and Geriatrics, 2007, Saint-Petersburg, Russia; The six the international conference on bioinformatics of genome regulation and structure – BGRS 2008, Novosibirsk, Russia; 2nd International conference Genetic of aging and longevity, 2012, Moscow, Russia; 3nd International Conference «Genetics of Aging and Longevity», 2014, Sochi, Russia; съезды ВОГиС, 2004, 2009, Москва, 2014, Ростов-на-Дону, 2019 Санкт-Петербург, Россия; Первая международная научно-практическая конференция МЕДБИОТЕК–2005, Москва, Россия; Всероссийская конференция «Перспективы фундаментальной геронтологии, 2006, Санкт-Петербург Россия; III Международная anti-ageing конференция «Медицина долголетия и качества жизни», 2007, Москва, Россия; Международная школа-конференция посвященная 100-летию со дня рождения М.Е. Лобашева

«Системный контроль генетических и цитогенетических процессов, 2007, Санкт-Петербург, Россия; III Международная научная конференция Донозология 2007, Донозология 2009, Санкт-Петербург, Россия; Всероссийский семинар «Генетика продолжительности жизни и старения», 2008, 2009, Сыктывкар, Россия; 12-я Международная Пущинская школа-конференция молодых ученых (Биология – наука XXI века), 2008, Пущино, Россия; V Всероссийская конференция с международным участием «Пренатальная диагностика и генетический паспорт – основа профилактической медицины в век нанотехнологий», 2012, Санкт-Петербург, Россия; Российский конгресс с международным участием «Молекулярные основы клинической медицины – возможное и реальное», 2012, 2020, 2021, 2022 Санкт-Петербург, Россия; II Международный конгресс «Медицина долголетия и качества жизни», 2013, Москва, Россия; Всероссийская научно-практическая конференция с международным участием «Молекулярная диагностика», 2014, 2017, 2021, Москва, Россия; II Всероссийская научно-практическая конференция с международным участием «Превентивная медицина 2014. Инновационные методы диагностики, лечения и реабилитации социально значимых заболеваний», 2014, Москва, Россия; Международный Научный Конгресс «Спорт, Человек, Здоровье», 2015, Санкт-Петербург, Россия; конференция Института Трансляционной Биомедицины СПбГУ (ИТБМ СПбГУ) «Актуальные проблемы трансляционной биомедицины», 2017, 2019, Санкт-Петербург, Россия; XI научная конференция «генетика человека и патология», посвященная 35-летию НИИ медицинской генетики, 2017, Томск, Россия; Международная научно-практическая конференция «NGS в медицинской генетике», 2018, 2019, 2021, Суздаль, Россия; XI Всероссийский съезд неврологов, 2019, Санкт-Петербург, Россия; XXI Зимняя молодежная школа по биофизике и молекулярной биологии, 2020, Санкт-Петербург, Россия; Всероссийская научно-практическая конференция «Фундаментальные и прикладные проблемы здоровьесбережения человека на Севере», 2020, Сургут, Россия; III национальный конгресс с международным участием ЛАБРиН 2021,

Москва, Россия; Онлайн-конгресс с международным участием «Молекулярная диагностика и биобезопасность, 2021, Москва, Россия; VI Всероссийская научно-практическая конференция с международным участием «Генетика опухолей кроветворной системы – от диагностики к терапии»,2021, Санкт-Петербург, Россия; 1-й Международный Форум Геномных и биомедицинских технологий «От рождения до активного долголетия», 2021, Сургут, Россия.

Кроме того, результаты работы были представлены в виде постерных докладов на Европейских конгрессах по генетике человека (European Congress of Human Genetics Conference (ESHG), 2002, Strasbourg, France; 2003, Birmingham, England; 2004, Munich, Germany; 2006, Amsterdam, The Netherlands; 2007, Nice, France; 2008, Barcelona, Spain; 2010, Gothenburg, Sweden; 2013, Paris, France; 2014, Milan, Italy; 2015, Glasgow, Scotland, United Kingdom; 2018, Milan, Italy; 2019, Gothenburg, Sweden; 2020, Virtual conference); на конференции по перспективным исследованиям в биологии (Biologie Prospective - Santorini Conference), 2008, 2018, Santorini Island, Greece; Anti-Aging Medicine World Congress 2006, Paris, France; 12th International Conference on Advanced Technologies & Treatments for Diabetes (ATTD 2019) Berlin, Germany.

Результаты работы представлены и обсуждены в 35 научных работах, из них: 24 основных статьи (среди которых 24 в журналах, индексируемых в международных базах данных WoS и Scopus, 24 статей в журналах по перечню ВАК Минобрнауки РФ), 1 монография и 1 методические рекомендации для врачей.

Результаты диссертационного исследования внедрены в практику лечебной и учебной работы ФГБНУ «НИИ АГиР им.Д.О.Отта», ФГБУ «ДНК ЦИБ» ФМБА РФ, СПбГУ, ИМБ РАН, СПб ГБУЗ «Городская больница № 40», ФГБНУ «МГНЦ», СПб ГКУЗ МГЦ, ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России, СПб ГБУЗ «КДЦД», СЗГМУ им. И.И. Мечникова Минздрава России, ФГБОУ ВО СПбГПМУ Минздрава России, ФГАОУ ВО

Первый МГМУ им. И.М. Сеченова Минздрава России, ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России, ФГБНУ «ИЭМ».

В диссертационной работе использованы экспериментальные и аналитические материалы, полученные автором лично или под его непосредственным руководством. Диссертант координировал и участвовал в сборе биологического материала и составлении клинической базы, осуществлял пробоподготовку и секвенирование образцов для молекулярно-генетического анализа с помощью NGS, секвенирования по Сенгеру и ПЦР-ПДРФ методов, а также руководил процессом биоинформатической обработки и статистического анализа полученных данных. Автор лично разрабатывал дизайн исследования и определял методологию исследования, выбирал показатели для анализа из баз данных, определял регионы генома для изучения, проводил анализ литературы, обобщал полученные клинические, лабораторные и молекулярно-генетические результаты, представлял результаты исследований в виде статей и тезисов докладов, выступал с докладами на научных конференциях.

Результаты, представленные в диссертации, получены при поддержке грантов Российского Фонда Фундаментальных Исследований № 09-04-13849-офи_ц, Российского научного фонда № 14-50-00069, Президента Российской Федерации МК-4113.2009.7, субсидий в виде грантов Комитета по науке и высшей школе Правительства Санкт-Петербурга, 2008, 2009, 2011, гранта ИАС СПбГУ №1.38.79.2012.

Диссертация включает введение, три главы и заключение; изложена на 242 страницах текста, содержит 35 рисунков, 26 таблиц и 341 ссылок на использованные источники литературы.

ГЛАВА I. Секвенирование нового поколения и моногенные болезни человека.

В рамках выполнения программы «Геном человека» и аналогичных проектов, уже к 2003 г. были идентифицированы и исследованы на наличие мутаций гены 1485 наследственных болезней человека. В дальнейшем были определены мутации, ассоциированные с еще примерно 3000 наследственных синдромов и болезней с выраженным генетическим компонентом. В то же время на сегодняшний день описывают более 1500 фенотипов с пока невыясненной молекулярно-генетической основой [Hamosh *et al.*, 2021].

В базе OMIM (на 02.08.2022) имеется информация о 7221 известных наследственных болезнях и синдромах, имеющих молекулярную основу [https://www.omim.org/statistics/geneMap]. Из них для 6152 фенотипов показана связь с единственным геном, т.е. моногенная природа генетического признака и или синдрома. Предполагают, что дальнейшая расшифровка генома и идентификация новых генов не сильно отразятся на этих цифрах, хотя и приведут к уточнению генов-кандидатов, вызывающих конкретные наследственные заболевания, а также генов-модификаторов, наличие которых в том или ином варианте может существенно сказываться на фенотипических особенностях болезни.

Совершенствование медицинской генетики во многом связано с развитием технологий полногеномного секвенирования ДНК, то есть определения точной последовательности составляющих ее нуклеотидов как во всем геноме, так и отдельно в его белок- кодирующей части – экзоме. Создание в 80-х гг. технологии секвенирования ДНК открыло возможность расшифровки последовательности генов человека. В последующем особенно важную роль в развитии медицинской генетики сыграла технология параллельного анализа последовательностей ДНК - секвенирования нового поколения (Next Generation Sequencing – NGS).

Благодаря применению NGS в сравнительно короткий срок (1,5 года) удалось провести секвенирование геномов почти у 2 000 жителей Европы (2013г.), начать программы по секвенированию 500 000 геномов коренных жителей Великобритании (2015 г.) и 200 000 жителей Северной Америки (2015 г.). Геномные проекты инициированы в Азии (Геномы Азии (Япония, Южная Корея, Китай и другие), Германии, Франции, США, Голландии). Геномный проект уже реализован в Эстонии, где секвенировано более 2000 геномов и проведено секвенирование более 10000 экзомов человека [https://genomics.ut.ee/en].

Генетическое тестирование и развитие национальных биобанков привело к принципиальному изменению здравоохранения в европейских странах.В Великобритании полагают, что в результате осуществления программы по секвенированию 500000 геномов будет осуществлена разработка новой генетической карты для каждого гражданина, что приведет к изменению принципов не только организации медицины, но индивидуальной терапии каждого пациента, внедрению технологии редактирования геномов для исправления «дефектных генов» и организации новых бизнес-процессов в области страховой медицины [https://www.genomicsengland.co.uk].

Информация о проектах по расшифровке генома(ов) человека HaplotypeMap, 1000 Genomes и др. приведены в предыдущем разделе (Введение) и в таблице 2. Имеются уже завершенные проекты с опубликованными результатами: 1000 Genomes Project [http://www.internationalgenome.org/], NHLBI Exome Sequencing Project [https://evs.gs.washington.edu/EVS/], The Genome Aggregation Database [https://gnomad.broadinstitute.org/], The 100,000 Genomes Project [https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/].

Помимо вышеуказанных проектов, аннотированию вариантов, ассоциированных с многофакторными и редкими моногенными заболеваниями, также посвящены проекты по секвенированию экзома (The NHLBI (MD, USA)

Exome Sequencing, CHARGE Consortia), персонализированного генома (The Personal Genome), генотипированию вариантов у добровольцев из Исландии (Next Code Health) (Табл. 2).

Таблица 2. Список наиболее известных геномных и экзомных проектов [по Gonzalez-Garay, 2014].

Проект	Доступ	Содержание	Источник
HapMap	Открытый доступ	Проект HapMap содержит SNP с частотой минорных аллелей ≥5%.	The International Hapmap, 2007
1000 Genomes	Открытый доступ	Проект 1000 Genome охватывает 98% SNP с частотой минорных аллелей ≥1% от 1092 человек из 14 популяций.	Genomes Project Consortium, 2012
The NHLBI (MD, USA) Exome Sequencing	Открытый доступ	Проект, направленный на обнаружение белок-кодирующих генов, ассоциированных с заболеваниями сердца, легких и крови и содержит частоты аллелей каждого SNP.	www.evs.gs.washington.edu/EVS/
The Personal Genome	Открытый доступ	Проект «Персональный геном» включает геномы 174 человек и экзомы более 400 добровольцев.	www.personalgenomes.org
NextCode Health	Коммерческий	Проект содержит 40 млн. анатированных вариантов, полученных из генотипов 140000 добровольцев из Исландии.	www.nextcode.com
CHARGE consortia	Открытый доступ	1000 наборов данных полного экзома хорошо фенотипированных индивидуумов из консорциума CHARGE.	www. dnanexus.com/usecases-charge

Благодаря технологиям NGS достигнут существенный прогресс в диагностике орфанных заболеваний [Biesecker, Green, 2014]. Важно отметить, что с помощью новых технологий стало возможным уточнять частоты распространенности многих моногенных заболеваний [Ng *et al.*, 2009]. Разработка программного обеспечения и биоинформатических подходов для эффективного и точного анализа данных секвенирования и интерпретации генных вариантов, т.е. их классификация по предполагаемой патогенности, стала наиболее сложным и важным шагом на пути от необработанных данных

геномных последовательностей к окончательному молекулярному диагнозу заболевания [Nykamp *et al.*, 2017; Richards *et al.*, 2015].

1.1. Популяционные генетические проекты

Как было отмечено выше, одним из наиболее значительных достижений в оценке клинических эффектов, ассоциированных с генными вариантами, стали глобальные проекты по секвенированию ДНК человека, такие как проект «1000 геномов» [Auton *et al.,* 2015], Genome Aggregation Database (gnomAD) или программа Национального института сердца, крови и легких (NHLBI) TopMed [Bick *et al.,* 2020]. Простой аргумент о том, что многие варианты, ранее перечисленные как патогенные, встречаются у здоровых людей слишком часто, чтобы вызывать заболевание, наследуемое по менделевскому типу, стал наиболее мощным инструментом уменьшения ложноположительных ассоциаций между генными вариантами и фенотипическими проявлениями. С этой целью информация о частоте популяционных аллелей (AF) в настоящее время широко используется для интерпретации вариантов в клинической практике [Lek *et al.,* 2016].

Генетическая структура человеческих популяций широко изучается во всем мире. Частота нереферентной аллели (т.е. не аллели «дикого типа») - в популяции является одним из наиболее важных факторов, влияющих на клиническую интерпретацию генетического варианта. Несмотря на чрезвычайно большое количество образцов в данных Gnomad (125 748 для версии v. 2.1), генетическая изменчивость во многих регионах мира все еще плохо изучена. Множество текущих крупномасштабных проектов по геному направлены на характеризацию вариантов, которые сохраняются в конкретной стране или регионе, например, GenomicsEngland [Walter *et al.,* 2015]. Однако на генетической карте мира все еще остается много белых пятен, которые вряд ли будут заполнены в ближайшие годы. Сегодня уже многие страны предпринимают усилия, чтобы восполнить этот пробел, осуществляя

национальные проекты по секвенированию генома (например, два недавних исследования населения Катара [Fakhro *et al.*, 2016; Rodriguez Flores *et al.*, 2014]). Одна из таких инициатив - проект «Геномы России», запущенный в 2015 году, - направлена на характеристику спектра генетических вариаций в различных этнических группах по всей России [Oleksyk *et al.*, 2015; Zhernakova *et al.*, 2018]. Однако проект "Геномы России" еще далек от завершения, и поэтому Россия остается одним из регионов, мало изученных в этом плане [Oleksyk *et al.*, 2015].

Поэтому имеющихся данных пока недостаточно для понимания всего спектра частоты патогенных генетических вариантов в РФ. В связи с этим мы впервые в РФ предприняли попытку охарактеризовать спектр носительства наследственных заболеваний на примере Северо-Западного региона РФ [Barbitoff *et al.*, 2019]. Мы использовали набор данных из 694 образцов ДНК, секвенированных с помощью различных наборов реагентов для секвенирования всего экзома (Agilent SureSelect V6, Illumina Nextera Rapid Capture, Roche SeqCap EZ MedExome и Illumina TruSeq Exome) и клинического экзома (CES; Набор для секвенирования Illumina TruSight One) для целей клинической молекулярной диагностики и/или научных проектов в Санкт-Петербурге.

Биоинформационный анализ данных секвенирования экзома был выполнен с использованием специального конвейера на основе программного продукта для выравнивания bwa [Li&Durbin, 2009], GenomeAnalysisToolkitv. 3.5., и Picardtoolsv. 2.2.2. Схема приема и обработки информации (пайплайн) была построена в соответствии с рабочим протоколом GATK Best Practices [De Pristo *et al.*, 2011]. Все образцы были обработаны с идентичными настройками конвейера и совместно генотипированы с использованием метода когортного генотипирования GATK. Варианты были аннотированы с помощью SnpEff/SnpSift [Cingolani *et al.*, 2012] с использованием следующих ресурсов: частоты аллелей проекта 1000 геномов [Auton *et al.*, 2015], частоты аллелей gnomADr.2.1 [Karczewski *et al.*, 2019; Lek *et al.*, 2016], частоты аллелей ESP6500

[Fu *et al.*, 2013], предикция патогенности оценивалась с использованием dbNSFP [Liu *et al.*, 2016], ClinVar databas ev. 2019.

Весь протокол анализа данных NGS, используемый во всех наших публикациях, включая анализ качества прочтений, выравнивание последовательностей, обнаружение вариантов, аннотацию вариантов, визуализацию вариантов, интерпретацию данных и использование геномных инструментов для анализа данных, детально приведен в нашем обзоре [Шиков *и др.*, 2019].

Для всей нашей выборки, состоящий из образцов ДНК от 694 участников исследования, мы определили в общей сложности 463 100 вариантов внутри целевых регионов экзома. Мы показали, что значительная часть генетических вариаций в Северо-Западном регионе России специфична для изучаемого региона. Например, информация о 9,3% (42 913) вариантах, идентифицированных в нашем исследовании, отсутствует в последней сборке dbSNP 151.

Как и ожидалось, европейцы были наиболее близки к нашей популяции по спектру генетических вариантов [Barbitoff *et al.*, 2019]. В целом, анализируя данные из базы gnomAD, мы наблюдали высокое соответствие по частотам аллелей для вариантов экзома между Северо-Западом России и Европейскими популяциями (рис. 6).

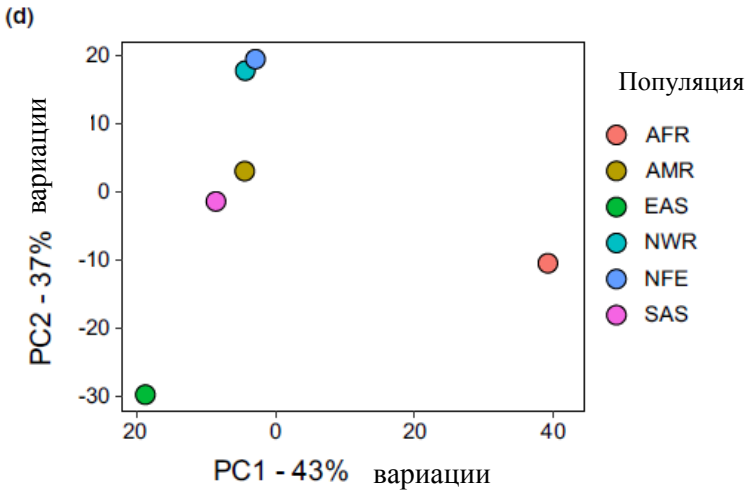


Рисунок 6. Анализ главных компонент частот 121171 варианта в популяции СЗ.

РФ и gnomAD (Genome Aggregation Database) популяциях (AFR, Африканцы; AMR, Смешанные американцы; EAS, Восточная Азия; NFE, нефинские европейцы; NWR, Северо-запад РФ; SAS, Южная Азия Asian [Barbitoff *et al.*, 2019].

Полученные результаты подтверждали необходимость создания крупномасштабной национальной базы данных генетических вариаций в России, которая поддерживала бы как локальную, так и глобальную информацию о проводимых клинико-генетических исследованиях.

Интересно отметить, что многие из ранее описанных патогенных аллелей, которые широко распространены в европейской популяции, преобладают также и у жителей Северо-Запада России (Табл. 3). Причем частоты многих этих аллелей в России выше, чем у европейцев, за исключением финнов. Среди них можно отметить доминантную аллель rs76151636 в гене *ATP7B* [de Bie *et al.*, 2007] и распространенную мутацию R408W (rs5030858) в гене *PAH* [Tighe *et al.*, 2003].

Таблица 3. Наиболее часто встречающиеся наследственные моногенные болезни с рецессивным типом наследования у жителей Северо-Запада России [Barbitoff *et al.*, 2019].

RsID	Ген	gnomAD AF	gnomAD NFE AF	p-value	Заболевание/состояние
rs5030858	*PAH*	7.6 × 10−4	0.0015	7.9 × 10−4 (1.1 × 10−5)	Фенилкетонурия
rs36209567	*F7*	5.6 × 10−4	0.0010	0.0010(5.7 × 10−4)	Недостаточность фактора VII
rs542489955	*FKBP14*	5.5 × 10−4	0.0010	0.0061(9.6 × 10−5)	Синдром Элерса-Данлоса, кифосколиотический тип, 2
rs61754365	*TYR*	2.9 × 10−4	3.2 × 10−4	1.1 × 10−4 (2.4 × 10−8)	Тирозиназонегативный кожно-глазной альбинизм
rs76151636	*ATP7B*	9.2 × 10−4	0.0013	0.0159 (0.0095)	Болезнь Вильсона-Коновалова
rs200488568	*SBF1*	3.3 × 10−4	1.4 × 10−4	1.2 × 10−4 (5.0 × 10−5)	Болезнь Шарко-Мари-Тута, тип 4B3

rs549794342	*NEB*	2.7 × 10−4	4.7 × 10−4	0.0050(5.9 × 10−5)	Немалиновая миопатия
rs201544686	*MTO1*	1.7 × 10−4	2.0 × 10−4	4.9 × 10−4 (5.4 × 10−6)	Комбинированный кисл. фос. дефицит 10
rs386834233	*BCKDHB*	5.5 × 10−4	3.9 × 10−4	3.0 × 10−3(2.3 × 10−3)	Болезнь кленового сиропа
rs775051461	*CCNO*	9.8 × 10−5	4.7 × 10−5	4.6 × 10−4(0.0015)	Дискинезия ресничек
rs121434233	*ALOXE3*	1.5 × 10−4	2.8 × 10−4	0.018(5.2 × 10−5)	Аутосомно-рецессивный врожденный ихтиоз 3
rs543206298	*NPC1*	7.6 × 10−5	1.1 × 10−4	0.0033(5.2 × 10−4)	Болезнь Нимана-Пика
rs104894080	*GDAP1*	3.2 × 10−5	7 × 10−5	0.0013(0.0044)	Полиневропатия, болезнь Шарко-Мари-Тута тип А
rs1307458231	*ALMS1*	2.0 × 10−5	4.4 × 10−5	5.0 × 10−4(1.8 × 10−3)	Синдром Альстрома

Мы не выявили никаких широко распространенных патогенных или вероятных патогенных вариантов, отсутствующих в ClinVar или dbSNP, для генов, связанных с аутосомно-рецессивным наследованием. Это указывает на то, что по крайней мере для рецессивных патологических состояний, большая часть генетических детерминант является общей для России и других групп населения. Полученные данные позволили нам сделать первые оценки распространенности моногенных заболеваний на основе изучения экзома в регионе (табл. 4).

Таблица 4. Наиболее часто встречающиеся патогенные варианты у жителей Северо-Запада России [Barbitoff *et al.*, 2019].

Заболевание/ состояние	**Ген**	**Количество аллелей**	**Частота носительства (нижний/верхний интервал)**	**Частота болезни (нижний/верхний интервал)**	**Известная частота**	**Комментарии/ ссылки**
Дистрофия сетчатки, болезнь Штаргардта	*ABCA4*	13 (23)	0.0350 (0.0206/0.0589)	3.1×10^{-4} (1.1×10^{-4}/8.8×10^{-4})	1 in 10,000	Zol'nikova, 2016; Sheremet *et al.*, 2017
Муковисцидоз	*CFTR*	11 (19)	0.0296 (0.0167/0.0522)	2.2×10^{-4} (6.9×10^{-5}/6.9×10^{-4})	1 in 10,000	Частота носительства 0.032 (Abramov*etal.*, 2015)
Фенилкетонурия	*PAH*	11 (18)	0.0296(0.0167/0.052	2.2×10^{-4}(6.9×10^{-5}/6.9	1 in 10,000	Частота носительства

			2)	$\times 10^{-4}$)		0.029 (Abramov*etal.*, 2015)
Врожденная Афибриногенемия	*FGG*	7 (10)	0.0190(0.0092/0.0387)	9.0×10^{-5}(2.1×10^{-5}/3.8×10^{-4})	n.a.	Частота 1на 1,000,000 (Mannucci*etal.*, 2004)
Дефицит печеночной липазы	*LIPC*	6 (14)	0.0162(0.0075/0.0359)	6.6×10^{-5}(1.4×10^{-5}/3.1×10^{-4})	n.a.	—
Тирозиназонегативный кожно-глазной альбинизм	*TYR*	6 (12)	0.0162(0.0075/0.0359)	6.6×10^{-5}(1.4×10^{-5}/3.1×10^{-4})	1 in 39,000	—
Синдром шелушащейся кожи	*TGM5*	5 (8)	0.0135(0.0058/0.0311)	4.5×10^{-5}(8.3×10^{-6}/2.5×10^{-4})	n.a.	—
Недостаточность фактора VII	*F7*	5 (7)	0.0135(0.0058/0.0311)	4.6×10^{-5}(8.3×10^{-6}/2.5×10^{-4})	n.a.	Частота 1 на 500,000 (Wulff etal., 2000)
Болезнь Вильсона-Коновалова	*ATP7B*	4 (6)	0.0108(0.0042/0.0274)	2.9×10^{-5}(4.3×10^{-6}/1.9×10^{-4})	1 in 30,000.	Сообщалось о аналогичной глобальной заболеваемости (Ala, Walker, Ashkan, Dooley, &Schilsky, 2007)
Синдром Элерса-Данлоса, кифосколиотический тип, 2	*FKBP14*	4 (8)	0.0108(0.0042/0.0274)	2.9×10^{-5}(4.3×10^{-6}/1.9×10^{-4})	n.a.	—
Наследственная непереносимость фруктозы	*ALDOB*	4 (7)	0.0108(0.0042/0.0274)	2.9×10^{-5}(4.3×10^{-6}/1.9×10^{-4})	n.a.	—
Галактоземия	*GALT*	4 (5)	0.0108(0.0042/0.0274)	2.9×10^{-5}(4.3×10^{-6}/1.9×10^{-4})	1 in 20,000	Частота носительства 0.006 (Abramov *et al.*, 2015)

Несмотря на низкую точность оценок из-за ограниченного размера выборки, наши данные по двум наиболее распространенным патологиям - муковисцидозу и фенилкетонурии - согласуются с предыдущими оценками на уровне изучения отдельных генов [Barbitoff *et al.*, 2019]. Интересно отметить, что, по нашим данным, для Северо-Западного региона наиболее распространенной патологией является не муковисцидоз, а болезнь Штаргардта [Barbitoff *et al.*, 2019]. Обнаружив это, мы решили проанализировать частоту

каждого выявленного распространенного варианта в gnomAD, отличного от Северо-Запада России. Мы обнаружили, что один из вариантов rs38683423 в гене *BCKDHB* также «избыточно» представлен в финской популяции, что, возможно, указывает либо на поток генов между Северо-Западом России и финской популяцией, либо на специфичный для региона отбор против соответствующего состояния.

В ходе нашего исследования мы сфокусировали внимание на некоторых патогенных вариантах для ряда заболеваний человека. Результаты этого анализа представлены в таблице 4. Анализ показал, что наиболее распространенными были следующие нарушения: (а) Болезнь Штатгарда с *ABCA4* (MIM#601691) в качестве основного гена, как сообщалось ранее [Sheremet*etal.*, 2017], с частотой 1:3226; (б) муковисцидоз (ген *CFTR*, MIM#602421, F508del (7:117199644:ATCT>A) - мажорная мутация, с частотой 1:5263; (в) фенилкетонурия (ген *PAH*, с частотой 1:5556); (г) дефицит печеночной липазы (ген *LIPC*, MIM#151670, 1:10000, с одним патогенным вариантом rs113298164), и (д) тирозин-негативный кожно-глазной альбинизм (ген *TYR*, 1:13158). Наши результаты согласуются с предыдущим крупномасштабным анализом частот носителей патогенных аллелей при муковисцидозе у лиц европеоидной расы [Lazarin *et al.*, 2013]. Более того, предполагаемая заболеваемость фенилкетонурией в России также аналогична той, о которой сообщалось в различных популяциях, включая недавнее исследование в Китае [Zhao *et al.*, 2019]. Полученные нами данные по оценке частот муковисцидоза, галактоземии и фенилкетонурии также сходны с предыдущими оценками ряда исследований [Abramov *et al.*, 2015; Abramov *et al.*, 2017]. В то же время оценки частот других заболеваний (недостаточность фактора VII и врожденная афибриногенемия) в 20-100 раз выше, чем по мировым данным [Mannucci, Duga, &Peyvandi, 2004; Wulff *et al.*, 2000]. Возможно, это связано с небольшим объемом нашей выборки на момент исследования [Barbitoff *et al.*, 2019].

Хотя глобальные базы данных о частотах аллелей по-прежнему полезны, недавно была подчеркнута дополнительная ценность справочных баз данных по конкретным популяциям. Даже в оригинальной публикации ExAC, Лек с соавторами [Lek *et al.,* 2016] показали, что фильтрация вариантов-кандидатов по максимальной частоте аллелей в разных популяциях существенно уменьшает количество потенциально вызывающих заболевание вариантов, наблюдаемых в отдельной выборке экзомных данных.

В последнее время значительные усилия геномного сообщества были направлены на создание более разнообразных и эксклюзивных справочных данных о населении, а также ресурсов, охватывающих различные этнические и расовые группы [Wong *et al.,* 2020; Martin *et al.,* 2018]. Во многих странах были проведены общенациональные проекты по секвенированию, в том числе «Геном Нидерландов» (GoNL, [Boomsma *et al.,* 2014]) или База данных генома китайцев-ханьцев [Gao *et al.,* 2020]. Как было уже упомянуто выше, лишь несколько исследований были посвящены изучению геномной изменчивости населения России. К ним относятся пилотная фаза проекта «Геномы России» [Zhernakova *et al.,* 2020], а также наше исследование распространенности моногенных заболеваний у 694 пациентов [Barbitoff *et al.,* 2019] и таргетное секвенирование 242 известных генов заболеваний у 1658 здоровых людей из Ивановской области [Ramensky *et al.,* 2021]. В этих работах было выявлено несколько важных аспектов вариабельности генома у российских пациентов. Однако всем этим исследованиям не хватает полноты анализа, либо из-за малого размера выборки (например, в проекте «Геномы России»), либо из-за малого разнообразия анализируемых генов [Ramensky *et al.,* 2021]. С целью увеличения объема анализируемой выборки мы создали расширенный референтный набор генетических вариантов, проанализировав 6096 образцов, собранных в двух крупных городах России - Москве и Санкт-Петербурге [Barbitoff *et al.,* 2021].

Наиболее распространенный подход к получению достоверной информации о частоте аллелей по сей день основывается на стандартной и

единообразной обработке генетических данных, сгенерированных несколькими центрами секвенирования и геномными лабораториями. В то же время для выполнения такой интеграции централизованным способом требуется сбор огромного объема данных, что является очень трудоемким с точки зрения времени и вычислительных ресурсов, а также потенциально требует обмена конфиденциальной или защищенной информацией. Возможный способ обойти эти трудности состоит в том, чтобы выполнить идентичные начальные этапы анализа данных в различных лабораториях, а затем объединить результаты на уровне анализа вариантов для каждой выборки. Такой подход мы применили в нашей работе для объединения данных из двух лабораторий [Barbitoff *et al.*, 2021].

После тщательного контроля качества образцов и вариантов [Barbitoff *et al.*, 2021] был получен массив данных по 5268 биообразцов и 2 092 456 выявленных генных вариантов. Из них 349 811 вариантов совпадали с теми, о которых сообщалось в нашей предыдущей публикации [Barbitoff *et al.*, 2019]. Из всех вариантов 75,7% были известны (найдены в последней сборке dbSNP), а 24,3% (509 409) были новыми. В общей сложности 1 459 530 вариантов (69,8%) были либо некодирующими, либо «молчащими» вариантами. 579 974 (27,7%) были миссенс или другими вариантами с умеренным воздействием, а 52 952 (2,5%) варианта были вариантами предполагаемой потери функции (loss-of-function, pLoF). Как и в предыдущих исследованиях, варианты, встречавшиеся редко и повреждающие структуру белка, были в значительной мере представлены среди новых вариантов. Например, только 23,6% некодирующих и «молчащих» вариантов были новыми по сравнению с 41,3% всех pLoF-вариантов. Аналогичным образом, 89,6% (1 875 600) всех генных вариантов были редкими (MAF <1% в общей выборке) по сравнению с 99,2% (505 427) для новых вариантов. Наконец, частоты вариантов в выборках каждой из участвующих сторон (лабораторий) показали идеальную корреляцию (r2 =

0,999), что отражено на рис. 7. В связи с тем, что кластеры выглядели как части ступни ноги, мы их назвали «пятка», «пальцы ног», «лодыжка».

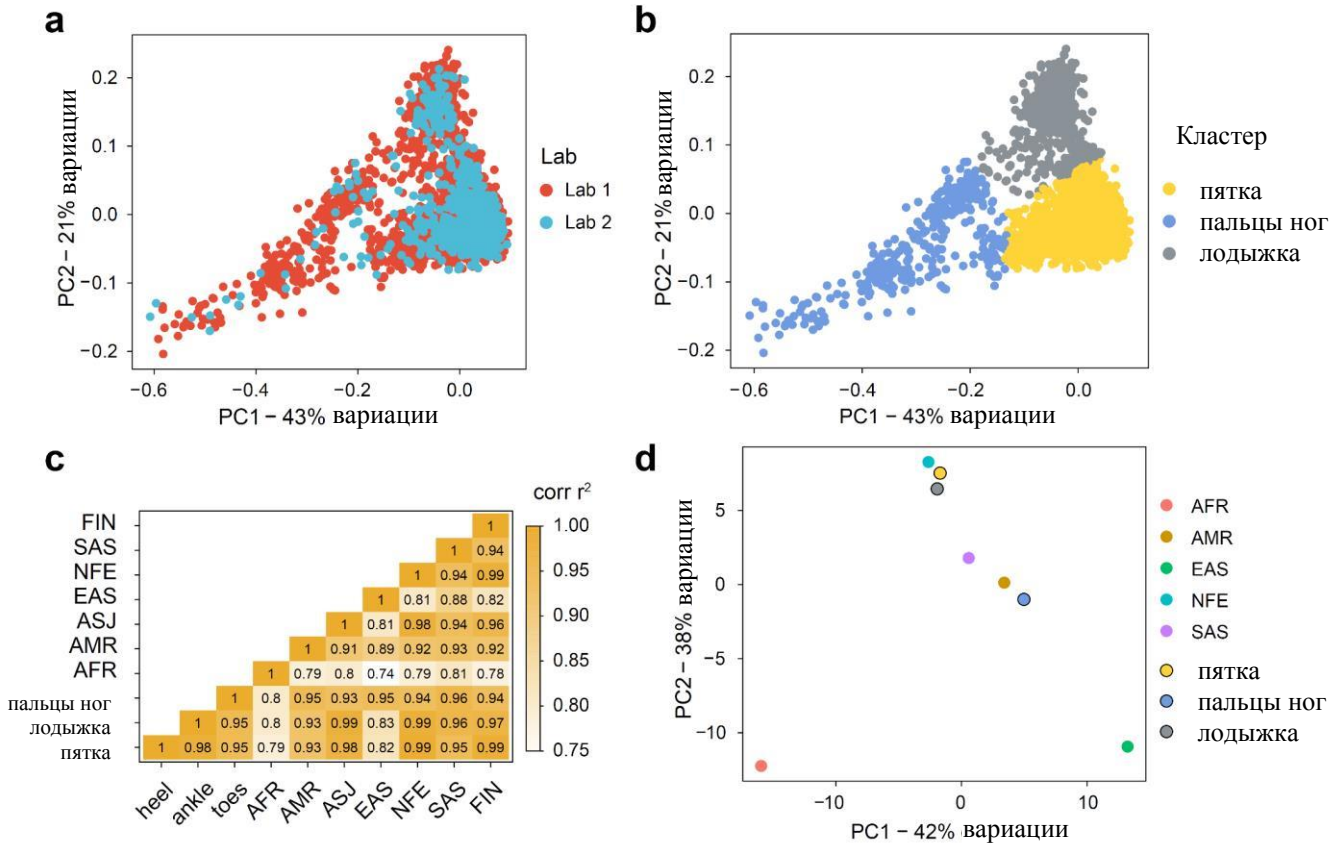


Рисунок 7. Анализ субструктуры выборки генеральной российской популяции [Barbitoff *et al.*, 2021]. Примечания: по данным анализа основных компонентов генотипов, отмечены разными цветами лаборатории **(a)**, или результаты кластеризации k-средних в пространстве первых 10 основных компонентов **(b)**. **(c)** Тепловая карта, показывающая корреляцию Пирсона между частотами общих вариантов аллелей в группах предков gnomAD и тремя кластерами («пятка, лодыжка, пальцы ног»), идентифицированных с помощью k-средних в **(b)**. **(d)** Анализ основных компонентов частот аллелей общих вариантов в gnomAD и трех кластерах («пятка, лодыжка, пальцы ног»), идентифицированных с помощью k-средних в **(b)**.

Определив три отдельные подгруппы образцов («пятка, лодыжка, пальцы ног»), мы решили проверить, какая из глобальных групп предков ближе всего к этим кластерам. Чтобы ответить на этот вопрос, мы сначала проанализировали

корреляцию между частотами аллелей по общим вариантам между каждым кластером и основными семью популяциями gnomAD (африканская (AFR), смешанная американская (AMR), евреи-ашкенази (ASJ), восточноазиатская (EAS), европейско-финская (FIN) и не финская (NFE) и южноазиатская (SAS)). Основной кластер русских людей («пятка») имел наибольшую корреляцию с группой gnomADNFE (r2=0,991), за которой следует выборка населения Финляндии (r2=0,988). В то же время второй («лодыжечный») кластер имел высокую степень корреляции AF с группой NFE (r2 =0,987), был ближе к евреям-ашкенази, чем к финским пробандам из gnomAD (r2=0,987 и r2=0,980 соответственно) и имел гораздо более высокую корреляцию AF как с EAS, так и с SAS. Наконец, третий кластер («пальцы ног») показал наибольшую корреляцию с субпопуляцией SAS из gnomAD (r2=0,975) и в равной степени отличался от EAS и NFE (r2=0,964 для обоих, рис. 7 (c)). В дополнение к наблюдениям, сделанным с помощью анализа корреляции AF, мы провели анализ основных компонентов частот общих вариантов аллелей, используя данные из трех кластеров и пяти основных групп предков gnomAD (AFR, AMR, EAS, NFE и SAS). Этот анализ показал, что, как и ожидалось, кластеры «пятка» и «лодыжка» были близки к группе NFE, а «лодыжка» оказалась ближе к SAS / EAS. В то же время гетерогенный кластер «пальцы ног» находился гораздо дальше и казался ближе к gnomADSAS (рис. 7 (d)).

Учитывая эти наблюдения, мы пришли к выводу, что первый кластер представляет лиц европейского происхождения, т.е. коренных жителей Центральной и Северо-Западной России; второй кластер представляет население Юга России и Северного Кавказа, в то время как третий кластер соответствует пациентам, происходящим из сибирских регионов и/или азиатских республик бывшего СССР. Эти предположения были дополнительно подтверждены имеющейся информацией о пациентах из обоих центров секвенирования [Barbitoff *et al.*, 2021].

В двух крупных исследованиях российской популяции, основанных на секвенировании, было зарегистрировано около двух десятков распространенных и чрезмерно представленных патогенных вариантов [Barbitoff *et al.,* 2019; Ramensky *et al.,* 2021]. Сначала мы решили подтвердить в нашем новом наборе данных частоты вариантов, наиболее представленных в более ранних исследованиях. В целом, мы взяли в обработку 22 варианта, о которых сообщалось в этих двух публикациях: 14 в нашем исследовании 2019 года и 10 в исследовании Раменского и соавт., с 2 перекрывающимися вариантами [Barbitoff *et al.,* 2019; Ramensky *et al.,* 2021, соответственно]. Из них избыточная представленность 10 вариантов была успешно подтверждена с использованием данных о здоровых донорах, а 15 - в полном наборе из 5268 образцов.

Затем мы перешли к выявлению всех вариантов, которые были избыточно представлены в нашем массиве данных. В результате 19 аллелей, ассоциированных с заболеваниями, были идентифицированы как чрезмерно представленные в подгруппе здоровых доноров (табл. 5). Они включали как известные варианты, встречающиеся с высокой частотой, такие как rs5030654 (ген *PAH*) и rs1555287300 (ген *ATP7B*), связанные, соответственно с фенилкетонурией и болезнью Вильсона-Коновалова, так и варианты, о которых ранее не сообщалось как чрезмерно представленных. Последняя категория включала такие варианты как rs554847663 в гене *OTOG*, связанный с аутосомно-рецессивной глухотой, и вариант rs119473033 в гене *SMARCAL1*, вызывающий иммуно-костную дисплазию Шимке. Высокая частота этих вариантов ранее не отмечалась [Barbitoff *et al.,* 2021]. Поэтому данная информация представляет значительный клинический интерес, особенно в российской популяции.

Охарактеризовав перечень избыточно представленных патогенных аллелей в нашем массиве данных, мы перешли к выявлению потенциально клинически значимых вариантов, которые присутствуют у здоровых пациентов, но не обнаружены в gnomAD. Мы начали с выявления известных патогенных вариантов, отсутствующих в данных gnomADv2.1.1, которые представлены в

базе данных ClinVar. В общей сложности мы обнаружили 72 таких варианта, причем 25 из них расположены в генах с аутосомно-доминантным наследованием заболеваний [Barbitoff *et al.,* 2021]. В дополнение к этим вариантам мы также искали потенциально клинически значимые варианты, которые отсутствуют в gnomAD, но присутствуют в подгруппе здоровых лиц. В общей сложности было выявлено более 100 предполагаемых вариантов потери функции (pLoF), присутствующих у здоровых пациентов. Из них 27 вариантов локализованы в генах с высокой степенью эволюционной сохранности в соответствии с метриками, полученными из gnomAD (PLI, LOEUF), и с известной связью с аутосомно-доминантными фенотипами [Barbitoff *et al.,* 2021].

Таблица 5. Известные патогенные варианты, присутствующие с высокой частотой в расширенной группе жителей России [Barbitoff *et al.,* 2021].

Вариант ID	Ген	gnomAD NFE AF	Количество аллелей	RUSeq AF*	p-value	Заболевание
rs200482683	*NPHS2*	0.02%	5	0.35%	2.07E-05	Нефротический синдром
rs549794342	*NEB*	0.05%	8	0.56%	5.86E-07	Muscular dystrophy
rs119473033	*SMARCAL1*	0.01%	4	0.28%	4.62E-05	Иммуно-костная дисплазия Шимке
rs775288140	*COL7A1*	0.00%	3	0.21%	8.84E-06	Рецессивный дистрофический буллезный эпидермолиз
rs777686211	*CPLANE1*	0.02%	4	0.29%	3.41E-04	Нефронофтизис
rs104893924	*SLC26A2*	0.02%	5	0.35%	2.91E-05	Остеохондродисплазия
rs386834233	*BCKDHB*	0.04%	5	0.35%	2.83E-04	Болезнь мочи с кленовым сиропом
rs782316919	*SURF1*	0.02%	4	0.28%	7.59E-05	Мозжечковая атаксия
rs554847663	*OTOG*	0.08%	6	0.46%	5.86E-04	Умственная отсталость и аутосомно-рецессивная глухота
rs371720347	*STAC3*	0.01%	3	0.21%	1.06E-04	Врожденная миопатия Бейли-Блоха
rs5030858	*PAH*	0.15%	10	0.70%	7.45E-05	Фенилкетонурия
rs76151636	*ATP7B*	0.13%	9	0.63%	1.26E-04	Болезнь Вильсона-Коновалова
rs36209567	*F7*	0.10%	7	0.49%	8.87E-04	Аномальное кровотечение
rs200389141	*BLM*	0.03%	5	0.35%	6.55E-05	Синдром Блума
rs375470378	*GAA*	0.03%	4	0.28%	7.21E-04	Болезнь накопления гликогена
rs387906455	*F8*	0.00%	3	0.21%	9.34E-07	Недостаточность фактора VIII

* - AFполученас учетом данных референса подгруппы здоровых доноров.

Среди известных и ожидаемых патогенных вариантов, присутствующих в нашей выборке, наиболее интересен вариант rs397516471 в гене *TNNT2*. Этот вариант считается патогенным для семейной рестриктивной кардиомиопатии и

патологии левого желудочка; однако в нашем массиве данных он присутствует *в* гетерозиготном состоянии у 1 здорового человека без симптомов заболевания. Другим интересным примером является вариант rs1064793825 в гене *MSH2*, ассоциированный с повышенным риском возможного развития наследственного колоректального рака. Однако этот гетерозиготный носитель также не имеет соответствующей патологии. Однако в случае rs397516471 и rs1064793825 в семейном анамнезе носителей патогенных вариантов указано несколько случаев сопутствующих заболеваний (кардиомиопатия или онкология, соответственно), что позволяет предположить, что заболевание, вероятно, проявится в ближайшем будущем.

В целом, наши результаты демонстрируют острую необходимость в генетических базах данных для конкретных популяций. Это крайне необходимо для интерпретации вариантов и выявления факторов риска заболеваний в малоизученных популяциях. Хотя нынешний размер выборки позволяет делать более объективные выводы относительно генетической структуры населения России, мы все еще можем ожидать большого количества редких генетических вариантов у остальной части населения, которые не были охвачены нашим анализом. Следовательно, дальнейшее объединение данных из центров секвенирования по всей России, секвенирование большего числа здоровых доноров и включение пациентов из разных регионов необходимы для полной характеристики спектра генетических вариаций современной России, чтобы сделать уверенные предположения о распространенности моногенных болезней и о популяционных частотах аллелей заболеваний [Barbitoff *et al.*, 2019; 2021]. Тем не менее, мы надеемся, что представленные данные помогут медицинским генетическим исследованиям и клиническим генетическим анализам как в России, так и за ее пределами.

1.2. Популяционные исследования для оценки частот вариантов

Как нами было описано выше, на основе популяционных генетических работ появляется возможность не только оценить частоту того или иного заболевания, но и проанализировать распространенность некоторых вариантов, например, частоты встречаемости дефицита лизосомной кислой липазы в российской популяции [Fedyakov *et al.,* 2018]. Дефицит лизосомной кислой липазы (ДЛКЛ) — редкая наследственная прогрессирующая болезнь обмена липидов, приводящая к развитию атеросклероза, гепатоспленомегалии, цирроза печени, мальабсорбции и других симптомов. При отсутствии специфического лечения прогноз для пациента неблагоприятный, поэтому крайне важна своевременная диагностика болезни. В зависимости от остаточной активности фермента лизосомной кислой липазы выделяют различные формы заболевания. Болезнь Вольмана — ранняя тяжелая форма, активность фермента — менее 1% [Abramov *et al.,* 1956; Aslanidis *et al.,* 1996]. Болезнь накопления эфиров холестерина — поздняя форма, более мягкое течение, активность фермента в пределах 1–12% [Fredrickson *et al.,* 1963]. Суммарная частота ДЛКЛ составляет 1/40 000–300 000 для различных популяций [Lohse *et al.,* 2000; Muntoni *et al.,* 2007]. Данные о частоте заболевания в России до нашей работы отсутствовали. Ожидаемая цифра составляла 1/100 000–150 000 [Baranov *et al.,* 2016; Strokova *et al.,* 2017]. Для оценки частоты ДЛКЛ в российской популяции были выгружены данные экзомного и геномного секвенирования 523 человек из Северо-Западного и Центрального регионов РФ. В данную выборку вошли образцы крови пациентов с моногенными (моногенные формы сахарного диабета, наследственные нарушения соединительной ткани, наследственные нарушения обмена веществ и другие орфанные заболевания) и мультифакторными заболеваниями (ожирение, сахарный диабет 2-го типа), а также образцы крови контрольной популяционной группы. Пациентов с клиническим диагнозом ДЛКЛ в представленной выборке не было. Проводился анализ генотипов по

патогенному варианту c.894G>A. В результате анализа данных в 2 образцах из 523 был выявлен вариант c.894G>A в гетерозиготном состоянии, что соответствует частоте 1/262 в исследуемой популяции. Таким образом, расчетная частота гетерозиготного носительства всех патогенных вариантов в гене *LIPA* может достигать 1/130. По закону Харди–Вайнберга, это соответствует оценочной частоте заболевания 1/67 600 [Fedyakov *et al.*, 2018].

Важно отметить, что на основе популяционных генетических работ появляется возможность не только оценить частоту того или иного заболевания, но и проанализировать распространенность некоторых вариантов, например, в гене *ACE2*, полиморфизм в котором играет роль в патологии COVID-19, влияя на важные нормальные функции белка. Мы провели сравнительный анализ частот пяти вариантов (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) в гене *ACE2* русской и европейских популяций. Было установлено, что русские похожи на другие европейские популяции, что предполагает сходный уровень встречаемости и тяжести заболевания. Это было важно для понимания эпидемиологической ситуации в начале эпидемии в марте-апреле 2020 года [Shikov *et al.*, 2020]. Более подробно данная часть исследования приведена в Главе III.

Популяционный анализ крайне необходим при оценке вклада патогенных вариантов в риск развития того или иного заболевания. Несмотря на наличие больших геномных баз, исследование различных групп российских пациентов имеет большое значение. Одним из примеров является наше исследование тухоухости [Bliznetz *et al.*, 2017]. Хотя мутации в последовательности гена *GJB2* составляют большинство вариантов, вызывающих аутосомно-рецессивную несиндромную потерю слуха, было показано, что несколько крупных делеций в локусе *DFNB1* способствуют глухоте. В настоящее время генетическое тестирование на потерю слуха включает секвенирование гена *GJB2* и анализ двух распространенных крупных делеций - del (GJB6-D13S1830) и del (GJB6-D13S1854). Нам удалось выявить новую делецию 101 кб, del (GJB2-D13S175). В

многонациональной когорте из 1104 неродственных пациентов с потерей слуха с биаллельными мутациями в локусе *DFNB1* была определена частота аллели с del (*GJB2*-D13S175) до 0,5% (11/2208) и было показано, что эта аллель преимущественно связана с глубокой нейросенсорной потерей слуха. Кроме того, были описаны восемь ранее неопубликованных мутаций в гене *GJB2* [Bliznetz *et al.,* 2017]. Все пациенты, носители del (*GJB2*-D13S175), имели ингушское происхождение. Среди людей с нормальным слухом del (*GJB2*-D13S175) наблюдалась в Республике Ингушетия с частотой носителей ~ 1% (2/241). Анализ гаплотипов, связанных с делецией, выявил общего основателя у ингушей, возраст делеции составляет ~ 3000 лет. Анализ на del (*GJB2*-D13S175) был добавлен в рутинную стратегию тестирования на потерю слуха [Bliznetz *et al.,* 2017]. Таким образом, поиск не только точечных мутаций, но и делеций является важным при постановке диагноза. В современное время это невозможно без качественного биоинформатического протокола анализа данных.

1.3. Биоинформатическая обработка данных NGS

Биоинформатическая обработка является частью технологии NGS. Поэтому установление генетической природы заболевания во многом зависит от качественного биоинформатического протокола анализа данных секвенирования [Barbitoff *et al.,* 2018; 2020]. Начинается она с картирования последовательности отсеквенированных участков. Для этого используют референсную последовательность генома, доступную на ресурсах RefSeq Национального центра биотехнологической информации [http://www.ncbi.nlm.nih.gov/RefSeq/] (с указанием номера версии) или Locus Reference Genomic [http://www.lrg-sequence.org/]. Необходимо учитывать, что в референсной последовательности встречаются ошибки, связанные с так называемыми референсными минорными вариантами —RMA (позициями референсного генома, в которые инкорпорирован редкий или даже патогенный вариант). Такие ошибки

необходимо корректировать при проведении биоинформатического анализа. Для коррекции данных ошибок была разработана специальная программа [Barbitoff *et al.*, 2018]. Первым шагом в биоинформатической обработке данных NGS является определение вариантов (variant calling) — стадия, на которой программа определяет варианты, отличающиеся от референсной последовательности. Затем проводят аннотацию вариантов (variant annotation) — стадия, на которой описывается патогенность хорошо охарактеризованных замен (на основании баз ClinVar, OMIM и др.) (см рис. 8 и табл. 6), а для малоизвестных замен предсказывается эффект с помощью одной из программ — предикторов эффекта (PROVEAN, SIFT, Polyphen, MutPred и др.) (табл. 7).

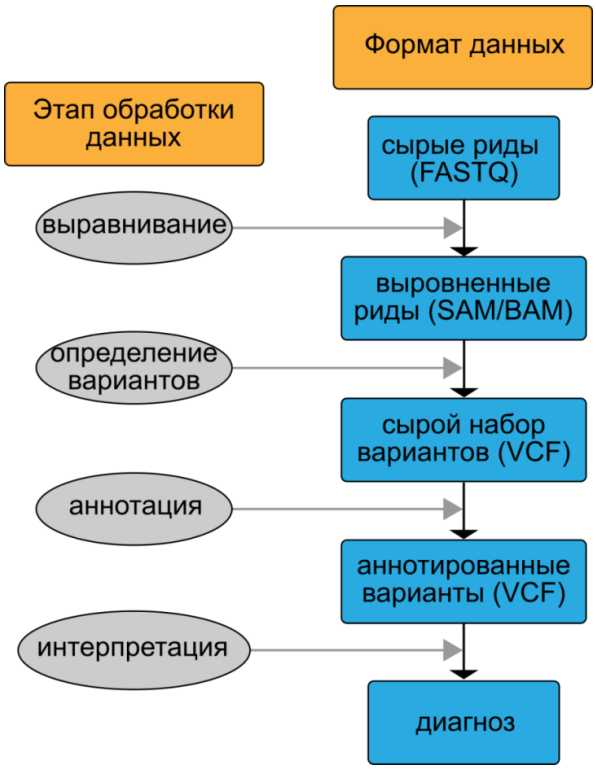


Рисунок 8. Алгоритм биоинформатического анализа данных секвенирования.

Обработка заканчивается ранжированием вариантов: найденные и аннотированные варианты сортируют по ряду критериев, таких как частота аллели в публичных базах данных, патогенность, гомо-/гетерозиготность замены и др. Ключевыми нюансами первой стадии являются международные базы геномов и детекция RMA, вторая стадия, опять-таки, во многом зависит от

наличия хорошо охарактеризованных баз данных, а вот третья в значительной степени определяется мощностью программ-предикторов, которые «обучаются» на основании множества факторов — «сильной» программы, наличия собственной базы данных и точности поставленного клинического диагноза. И если на первых двух стадиях главным «фигурантом» процесса является биоинформатик, то на последней стадии не менее, а может быть и более важны клинический опыт и знания врача-генетика или генетического консультанта.

1.4. Интерпретация данных NGS

Все современные стратегии и рекомендации по интерпретации вариантов, такие как рекомендации Американского колледжа медицинской генетики и геномики (ACMG) [Richards *et al.*, 2015], российские рекомендации по интерпретации вариантов [Рыжкова *и др.*, 2018] или рекомендации Шерлока [Nykamp *et al.*, 2017], используют AF в здоровых популяциях для классификации эффектов вариантов, что становится особенно критичным для аутосомно-доминантных (AD) заболеваний.

С внедрением NGS меняется и язык генетики. Вместо широко распространенных терминов «мутация» и «полиморфизм» рекомендуют использовать термин «вариант нуклеотидной последовательности» со следующими пятью характеристиками: патогенный (pathogenic); вероятно патогенный (likely pathogenic); неопределенного значения (uncertainsignificance); вероятно доброкачественный (likely benign); доброкачественный (benign) [Рыжкова *и др.*, 2019]. Инструменты для правильного описания вариантов нуклеотидной последовательности в соответствии с номенклатурой HGVS представлены на специальном сайте [https://mutalyzer.nl.]. Все обнаруженные варианты необходимо классифицировать по патогенности. Оценка патогенности выявленных вариантов включает изучение медицинской и научной литературы, а также баз данных. Для поиска описанных ранее вариантов рекомендуется использовать базы данных, указанные в таблице 6.

Варианты, описанные ранее в нескольких источниках (кроме ClinVar), являющиеся причиной развития интересующего фенотипа и подходящие под тип наследования, классифицируются как патогенные. Для интерпретации остальных вариантов предлагается использовать два набора критериев: первый для классификации вероятно патогенных вариантов, второй для классификации вероятно доброкачественных вариантов. Каждый патогенный критерий может оцениваться несколькими категориями: очень сильный (PVS1), сильный (PS1-4); средней тяжести (PM1-5) или вспомогательный (PP1-5). Каждый доброкачественный критерий - как очень сильный (независимый) (BA1), сильный (BS1-4) или вспомогательный (BP1-6). Нумерация признаков внутри категории не дает никаких усилений варианта, нумерация нужна для упрощения их использования. Для каждого варианта специалист выбирает подходящие признаки, которые затем объединяет в соответствии с указанными ниже правилами, классифицируя значимость варианта по пятиуровневой системе: патогенный, вероятно патогенный, неопределенного значения, вероятно доброкачественный, доброкачественный. Если вариант не отвечает критериям, используемым любым из этих наборов (патогенных или доброкачественных), или доказательства для доброкачественности и патогенности противоречивы, то вариант по умолчанию становится неопределенного значения [Рыжкова *и др.*, 2019].

Таблица 6. Рекомендуемые базы данных для анализа патогенности вариантов нуклеотидной последовательности [Рыжкова *и др.*, 2019].

Популяционные	
Exome Aggregation Consortium http://exac.broadinstitute.org/	База данных вариантов, найденных при проведении экзомного секвенирования у 61,486 неродственных индивидуумов, являющихся участниками различных болезнь-специфичных и популяционных генетических исследований. Лица, с наследственными заболеваниями, проявляющимися в детстве, были исключены из выборки.
genome Aggregation Database http://gnomad.broadinstitute.org/	Расширенная база данных геномных вариантов, основанная на базе ExomeAggregationConsortium, включающая данные по 123 136 экзомов и 15 496 геномов.

Exome Variant Server http://evs.gs.washington.edu/EVS/	База данных вариантов, найденных вовремя экзомного секвенирования нескольких крупных когорт лиц европейского и афроамериканского происхождения. Включает в себя данные о покрытии, для учета информации об отсутствии варианта.
1000 Genomes Project http://browser.1000genomes.org/index.html	База данных вариантов, найденных во время геномного и таргетного секвенирования с низким и высоким покрытием среди 26-и популяций. Обеспечивает большее разнообразия по сравнению ExomeVariantServer, но содержит данные более низкого качества, а некоторые когорты содержат данные о родственных индивидуумах.
dbSNP http://www.ncbi.nlm.nih.gov/snp	База данных коротких генетических вариантов (как правило, ≤50 п.о.), собранных из различных источников. Наряду с популяционными полиморфизмами содержит и множество патогенных вариантов.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	База данных структурных вариантов (как правило, >50 п.о.), составленная из многих источников.
Фенотипы	
OMIM http://www.omim.org/	База данных генов человека и генетических состояний, которая содержит репрезентативную выборку вариантов, ассоциированных с заболеваниями.
Human Gene Mutation Database http://www.hgmd.cf.ac.uk/ac/index.php	База данных аннотированных вариантов, опубликованных в литературе. Доступ к основной части контента требует оплаты. В базе встречаются полиморфизмы, необходимо уточнять клиническую значимость по литературным данным.
ClinVar http://www.ncbi.nlm.nih.gov/clinvar/	База данных утверждений о клинической значимости и фенотипической взаимосвязи вариантов. Содержит данные низкого качества, ***не рекомендуется***. Использование может быть ограничено только поиском ссылок на литературные источники.
Специфичные базы данных (локус/ болезнь/ этно/ другие)	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html Leiden Open Variation Database http://www.lovd.nl	Сайт общества по изучению вариаций генома человека создало список из тысяч баз данных, которые предоставляют варианты аннотации на конкретные разновидности человеческой вариации. Большая доля баз данных представлена системе LeidenOpenVariationDatabasesystem.
DECIPHER https://decipher.sanger.ac.uk/	Молекулярно-цитогенетическая база данных для врачей и исследователей, связывающая геномные данные, полученных с помощью микрочипов, с фенотипом, используя геномный браузер Ensembl.
Кодирующая последовательность	
NCBI Genome http://www.ncbi.nlm.nih.gov/genome	Ресурс полногеномных референсных последовательностей.

RefSeqGene http://www.ncbi.nlm.nih.gov/refseq/rsg/	Ресурс референсных последовательностей клинически релевантных генов.
MitoMap http://www.mitomap.org/MITOMAP/	Исправленная кембриджская референсная последовательность митохондриальной ДНК человека

При использовании баз данных лаборатории должны проверить следующую информацию:

1. частота обновлений, осуществление курирования базы данных (использовать новейшую версию и/или базу, куратором которой является институт с хорошей репутацией);
2. подтвердить использование HGVS номенклатуры и указать референсные последовательности для сборки генома и транскриптов, используемые при наименовании вариантов;
3. оценить все показатели качества, приводящиеся для оценки точности данных (может потребоваться чтение соответствующих публикаций);
4. определить источник и независимость исследований, в которых содержится информация о варианте.

Если вариант не был описан в литературе ранее и не представлен ни в одной из баз данных, или сведения о нем недостаточны, для решения о его значимости можно использовать результаты программ предсказаний патогенности. Ниже представлены адреса сайтов и краткое описание наиболее используемых на данный момент программ предсказания (табл. 7).

Таблица 7. Наиболее часто используемые компьютерные программы предсказания патогенности вариантов нуклеотидной последовательности [Рыжкова *и др.,* 2019].

Название - Вебсайт	**Основа**
Миссенс замены	
Align GVGD -http://agvgd.hci.utah.edu/agvgd_input.php	Структура/функция белка и эволюционная консервативность
MAPP – http://mendel.stanford.edu/SidowLab/ downloads/ MAPP/index.html	

MutationTaster - http://www.mutationtaster.org/	
MutPred - http://mutpred.mutdb.org/	
PolyPhen-2 - http://genetics.bwh.harvard.edu/pph2/	
PROVEAN -http://provean.jcvi.org/index.php	Выравнивание и измерение сходства между последовательностью варианта и последовательностью гомологичного белка
SIFT- http://provean.jcvi.org/index.php	
nsSNPAnalyzer - http://snpanalyzer.uthsc.edu/	Выравнивание множества последовательностей и анализ структуры белка
Condel - http://bg.upf.edu/fannsdb/	Объединяет SIFT, PolyPhen-2 и MutationAssessor
Изменения в сайтах сплайсинга	
GeneSplicer - http://ccb.jhu.edu/software/genesplicer/	Модели Маркова
Human Splicing Finder - http://www.umd.be/HSF/	Основанное на положении варианта
MaxEntScan – http://genes.mit.edu/burgelab/maxent/ Xmaxentscanscoreseq.html	Принцип максимальной энтропии
NetGene2 – http://www.cbs.dtu.dk/services/NetGene2/	Нейронные сети
NNSplice - http://www.fruitfly.org/seq_tools/splice.html	Нейронные сети
ASSP- http://wangcomputing.com/assp/	Нейронные сети
FSPLICE - http://www.softberry.com/berry.phtml? topic=fsplice& group=programs&subgroup=gfind	Видо-специфичный предиктор сайтов-сплайсинга, основанный на модели весовой матрицы

1.5. Поиск новых вариантов в геноме пациентов методом NGS

Новые подходы в интерпретации данных, наряду с использованием продвинутых биоинформатиких алгоритмов, позволяют нам описывать не только известные варианты, но и новые. Одним из примеров является описанный нами клинический случай обнаружения новой мутации сдвига рамки считывания в гене *PKP2* с помощью молекулярно-генетического тестирования с использованием метода NGS [Fedyakov *et al.*, 2019].

Аритмогенная кардиомиопатия/дисплазия правого желудочка (ARVC) - прогрессирующее заболевание миокарда, поражающее в первую очередь правый желудочек. Она развивается преимущественно в молодом возрасте, и первым симптомом часто является внезапная сердечная смерть (SCD), связанная со злокачественной желудочковой аритмией. Диагностика с использованием стандартной оценки состояния сердца может быть затруднена из-за незначительных и неспецифических клинических признаков на ранней стадии заболевания, особенно у родственников пациента. Молекулярно-генетическое тестирование может предоставить больше информации для принятия клинических решений. В нашем исследовании было показано, что пробанд и трое его детей были носителями патогенного варианта в гене *PKP2* [Fedyakov *et al.*, 2019].

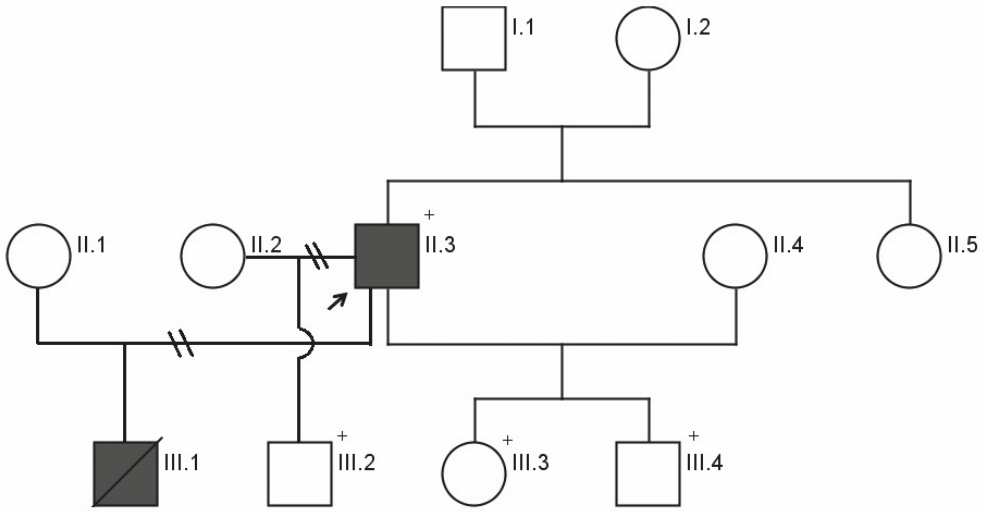


Рисунок 9. Родословная. Пробанд обозначен стрелкой; сплошные фигуры, заполненные черным цветом, представляют собою пациентов с ARVC+; + представляет индивидуумов с мутацией c.355delT.

Об этом варианте до настоящего времени (2018 год) не сообщалось в базах данных или научной литературе; он также отсутствовал в Exome Aggregation Consortium, сервере Exome Variant Server и проекте 1000 Genomes. Ген *PKP2* кодирует белок плакофилин-2, который участвует в образовании десмосом. Десмосомы являются основными компонентами межклеточных взаимодействий

и особенно распространены в эпидермисе и миокарде. Дефекты межклеточных соединений являются основным фактором, приводящим к ARVC [Huber, 2003]. Делеция тимина в позиции 355 кДНК приводит к смещению рамки, вызывая преждевременное обрывание цепи p.Y119fs*23 (и в конечном итоге приводит к образованию усеченного белка). Согласно критериям ACMG [Richards *et al.*, 2015], выявленная мутация была классифицирована как «вероятная патогенная». Результаты молекулярной диагностики позволили нам оценить риск развития ARVC и SCD у родственников пробанда, а также разработать индивидуальные протоколы оценки состояния сердца каждые 2-3 года до достижения 10-летнего возраста и ежегодно после 10 лет. Полученные результаты подчеркивают важность семейного скрининга при выявлении патогенной мутации у пробанда и демонстрируют эффективность генетического тестирования в кардиологической практике.

Другим примером успешности применения секвенирования является идентификация новых вариантов в гене *LDLR* у российских пациентов с семейной гиперхолестеринемией [Miroshnikova *et al.*, 2021]. Семейная гиперхолестеринемия (СГХ) ассоциирована с мутациями в различных генах, включая гены *LDLR, APOB* и *PCSK9* [Масленников, 1999; Мандельштам *и др.,* 2002]. В настоящем исследовании был проведен скрининг мутаций гена *LDLR* и других генов, ассоциированных с СГХ, у пациентов с диагностированной СГХ с использованием NGS [Miroshnikova *et al.*, 2021]. В общей сложности было исследовано 59 неродственных пациентов, разделенных на две отдельные группы в зависимости от их возраста: взрослые (n=31; средний возраст 49 лет; возрастной диапазон 23-70 лет) и дети/подростки (n=28; средний возраст 11 лет; возрастной диапазон 2-21 год). Варианты, ассоциированные с FH, были выявлены у 18 взрослых и 25 детей, что отражает частоту обнаружения мутаций 58 и 89% для взрослых и детей/подростков, соответственно. В группе взрослых у 13 пациентов были СГХ-ассоциированные мутации в гене *LDLR*, включая два новых варианта NM_000527.4: c.433_434dupGp.(val145glyfS*35) и c.1186G>C

(GLY396ARG); у 3 пациентов были выявлены мутации в гене *APOB* и у двух - в генах *ABCG5*/*G8*. В группе детей/подростков у 21 пациента были выявлены мутации в гене *LDLR*, включая пять новых вариантов NM_000527.4: c.325T>Gp.(Cys109Gly), c.401G>Cp.(Cys134Ser), c.616A>Cp.(Ser206Arg), c.1684_1691delTGGCCCAAp.(Pro563His) и c.940+1_c.940+4delgtga. В нашем исследовании описано семь новых вариантов в гене *LDLR*, которые считаются патогенными или вероятными патогенными. Среди них четыре миссенс-варианта были расположены в кодирующих областях, которые соответствовали функциональным доменам белка. Были также идентифицированы два варианта со сдвигом рамки считывания, продуцирующие дефектные белки. Эти варианты наблюдались только однократно у разных пациентов, тогда как вариант, приводящий к альтернативному сплайсингу в интроне 6 (c.940+1_c.940+4delGTGA) был обнаружен у четырех неродственных индивидуумов. Вариант p.Gly592Glu в гене *LDLR* был обнаружен у 6 пациентов, что составляет 10% от всех случаев СГХ в нашем исследовании. Таким образом, он может быть наиболее распространенным вариантом для СГХ среди населения России.

Несмотря на то, что СГХ является одним из наиболее распространенных генетических заболеваний, она по-прежнему часто остается невыявленной и не лечится во всем мире [Taranto *et al.,* 2020; Wiegman *et al.,* 2015]. Скрининг в детском возрасте может повысить выявляемость лиц с этим заболеванием до развития клинически значимой сердечно-сосудистой патологии [Vander Graaf *et al.,* 2011]. Поскольку на метаболизм липопротеинов у детей влияет меньше факторов окружающей среды, чем у взрослых, разница в уровнях холистерина липопротеинов низкой плотности (ХС ЛПНП) между детьми с СГХ и без нее более выражена [Santos *et al.,* 2016]. СГХ у детей диагностируется фенотипически по наличию повышенных уровней ХС ЛПНП, в дополнение к семейному анамнезу преждевременной ишемической болезни сердца (ИБС), высоким исходным уровням триглицеридов (ТС) у одного из родителей и/или

мутации, вызывающей СГХ [ander Graaf *et al.,* 2011]. Следует отметить, что не существует универсальных методов лечения для снижения уровня ХС ЛПНП в случае диагностики СГХ у детей. В связи с этим, раннее выявление СГХ является решающим для назначения профилактической терапии и снижения инвалидизации в будущем у пациентов, а подобного рода исследования, выявляющие семь новых вариантов в гене *LDLR* и расширяющие спектр мутаций в генах, связанных с СГХ, в Российской Федерации, следует считать значимыми в популяционно-генетическом аспекте.

Таблица 8. Анализ патогенности новых вариантов в гене *LDLR.*

Ген	**Пациент ID**	**Экзон/интрон**	**Вариант**	**Частота аллели вGnomAD**	**Частота аллели в [Barbitoff *etal.*, 2019]**	**Классификация патогенности вариантов по ACMG**
LDLR	G31	4	c.316_328delCCCAAGACG TGCT p.(Lis107Argfs*95)	Not found	Not found	Патогенный (PVS1 PS1 PM1 PM2 PP3)
LDLR	G29	4	c.325T>G p.(Cys109Gly)	Not found	Not found	Вероятно, патогенный (PS1 PM1 PM2 PM5 PP3)
LDLR	G36	4	c.401G>C (p.Cys134Ser)	Not found	Not found	Вероятно, патогенный (PS1 PM1 PM2 PM5 PP3)
LDLR	1	4	c.433_434insG p(Val145Glyfs*35)	Not found	Not found	Патогенный (PVS1 PM2 PP3)
LDLR	G18	4	c.616A>C (p.Ser206Arg)	Not found	Not found	Неопределённого значения (PM2 PP1 PP3)
LDLR	G21	IVS6	c.940+1_c.940+4 delGTGA (g.18154_18157delGTGA)	Not found	Not found	Патогенный (PVS1 PM1 PM2 PP3)
LDLR	32	8	c.1186G>C p.(Gly396Arg)	Not found	Not found	Патогенный (PVS1 PM1 PM2 PM5 PP3)
LDLR	G26	IVS8	c.1186+1G>T (g.22279G>T)	Not found	Not found	Патогенный (PVS1 PM2 PP3)
LDLR	G17	11	c.1684_1691delTGGCCCAA p.(Pro563Hisfs*14)	Not found	Not found	Патогенный (PVS1 PM1 PM2 PP3)

Третьим примером успешности применения NGS секвенирования является идентификация новых вариантов у российских детей с сахарным диабетом не 1 типа [Glotov O. *et al.,* 2019]. В нашей работе представлены данные о частоте и спектре генетических вариантов, вызывающих моногенный диабет у российских детей с сахарным диабетом не 1 типа. В исследовании приняли участие 60 неродственных российских детей с не диагностированным в возрасте до 18 лет сахарным диабетом 1 типа. Моногенный диабет составляет 1-6% детей, больных сахарным диабетом, и характеризуется наибольшей частотой среди пациентов,

страдающих сахарным диабетом не 1 типа в детском или подростковом возрасте [Hattersley *et al.,* 2018].

Большая, клинически гетерогенная группа доминирующих наследственных заболеваний, связанных с первичной дисфункцией βклеток, классифицируется как диабет с наступлением зрелости у молодых (МОДИ). На сегодняшний день известно 13 генов, вызывающих 13 типов МОДИ [Barbetti *et al.,* 2018]. МОДИ обычно диагностируется в возрасте до 25 лет; болезнь не зависит от инсулина, и ее симптомы обычно умеренные. Однако из-за разнообразия клинических форм, вызванных широким спектром мутаций в генах, связанных с МОДИ, используются различные стратегии лечения: от соответствующей диеты и физической активности до пероральной и/или инсулинотерапии. Моногенный диабет также включает ряд немодифицированных переходных или постоянных неонатальных форм, возникающих в возрасте до 6 месяцев. Известно, что более 20 генов связаны с врожденным неонатальным диабетом [Lemelman *et al.,* 2018]. В зависимости от задействованного гена, неонатальный диабет может следовать паттернам доминантного или рецессивного наследования, может быть изолированным или связанным с различными синдромными признаками [Greeley *et al.,* 2011]. Однако из-за очень раннего начала диабета гипергликемия часто диагностируется до появления других синдромных признаков. Стратегия лечения немодифицированного неонатального диабета зависит от конкретного генетического дефекта, вызывающего диабетический фенотип. Нами была предложена панель из 13 генов, вызывающих МОДИ (*HNF4A* (МОДИ1), *GCK* (МОДИ2), *HNF1A* (МОДИ3), *PDX1* (МОДИ4), *HNF1B* (МОДИ5), *NEUROD1* (МОДИ6), *KLF11* (МОДИ7), *CEL* (МОДИ8), *PAX4* (МОДИ9), *INS* (МОДИ10), *BLK* (МОДИ11), *ABCC8* (МОДИ12) и *KCNJ11* (МОДИ13)) и 22 генов, патогенные варианты в которых являются причинами транзиторного или постоянного неонатального диабета, в том числе связанного с конкретными синдромами (*EIF2AK3, RFX6, WFS1, ZFP57, FOXP3, AKT2, PPARG, APPL1,*

PTF1A, GATA4, GATA6, GLIS3, IER3IP1, LMNA, NEUROG3, PAX6, PLAGL1, SLC19A2, SLC2A2, SH2B1, SERPINB4 и *MADD*). Панель из 35 генов, вызывающих диабет раннего возраста (МОДИ) и транзиторный или постоянный неонатальный диабет, была исследована методом секвенирования всего экзома (WES). Проверка результатов WES была осуществлена секвенированием по Сенгеру. В целом, 33 из 60 пациентов (55%) имели генетические варианты в генах-мишенях (табл. 9; 21-40). У 12 пациентов родители были доступны для генетического тестирования, и благодаря этому было определено происхождение генетических вариантов. В 11 случаях генетические варианты были унаследованы от родителей, и в одном случае был подтвержден генетический вариант denovo. В общей сложности 38 генетических вариантов были выявлены у 33 из 60 пациентов (55%). У большинства пациентов (27/33, 81,8%) были варианты в генах, связанных с МОДИ: *GCK* (n=19), *HNF1A* (n=2), *PAX4* (n=1), *ABCC8* (n=1), *KCNJ11* (n=1), *GCK+HNF1A* (n=1), *GCK+BLK* (n=1) и *GCK+BLK+WFS1* (n=1). В общей сложности у 6 пациентов (6/33, 18,2%) были варианты в генах, не связанных с МОДИ: *GATA6* (n=1), *WFS1* (n=3), *EIF2AK3* (n=1) и *SLC19A2* (n=1). В общей сложности 15 из 38 вариантов были новыми, включая варианты в генах *GCK, HNF1A, BLK, WFS1, EIF2AK3* и *SLC19A2* [Glotov O. *et al.*, 2019]. Спектр генетических вариантов в гене *GCK* показан на рисунке 10 и представлен в таблице 9.

Миссенс-мутации в гене *HNF1A* (МОДИ3) были зарегистрированы у двух пациентов. Другие генетические варианты, связанные с МОДИ, включали три случая миссенс мутаций в генах *PAX4* (МОДИ 9), *ABCC8* (МОДИ12) и *KCNJ11* (МОДИ 13).

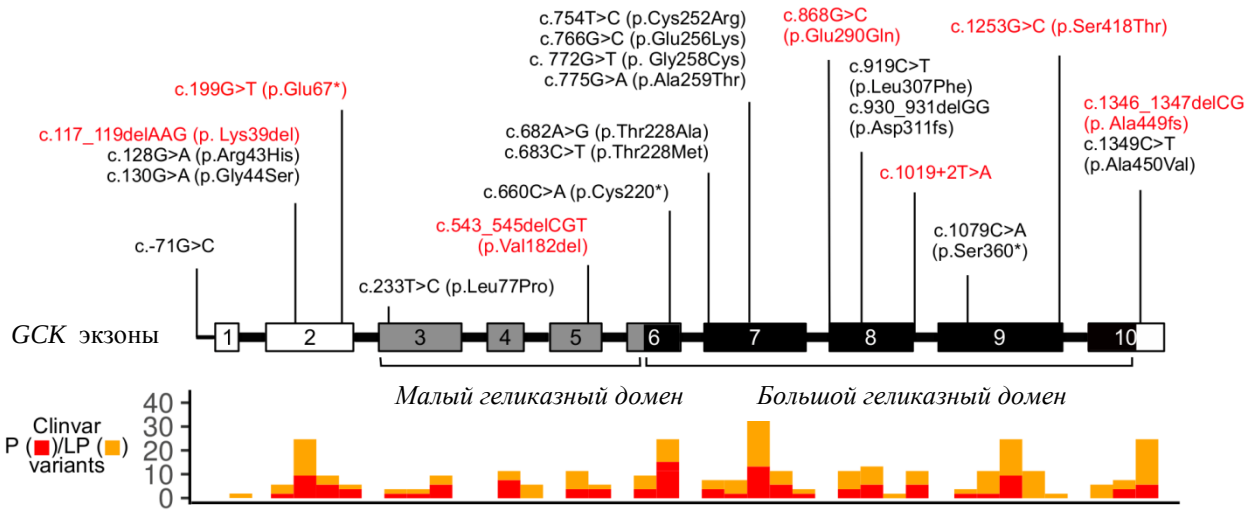


Рисунок 10. Спектр идентифицированных генетических вариантов в *GCK* гене [Glotov O. *et al.*, 2019]. Красным цветом отмечены впервые выявленные патогенные варианты у российских пациентов.

Этнические различия играют важную роль в определении эпидемиологии моногенного диабета, особенно МОДИ. Крупные популяционные исследования европейских европеоидов показали общую тенденцию к увеличению частоты *HNF1A*МОДИ в Северной Европе, в то время как *GCK*МОДИ преобладает в популяциях Южной Европы [Kleinberger and Pollin, 2015). Здесь мы сообщаем о *GCK*МОДИ у 19 и *HNF1A*МОДИ только у 2 из 27 МОДИ положительных российских пациентов. Наши показатели распространённости мутаций, по-видимому, ближе к таковым в популяциях Южной Европы, чем у жителей Северной Европы. Это может указывать на специфический для популяции тип частотных модификаций у российских пациентов [Glotov O. *et al.*, 2019].

Наличие генетических вариантов в различных генах-мишенях было обнаружено у трех пациентов. В одном из них делеция в гене *GCK* сопровождалась миссенс мутацией в гене *HNF1A* (пациент №226). В другом случае присутствовали две миссенс-мутации в генах *GCK* и *BLK* (пациент №529). У третьего пациента (№ 662) присутствовал дефект сплайсинга в генах*GCK* и миссенс-мутации в генах *BLK* и *WFS1*. Остальные найденные варианты присутствуют с низкой частотой в 9 различных генах, что в общей

сложности составляет ~ 50% случаев и подчеркивает эффективность использования именно WES в случаях, не связанных с геном *GCK* [Glotov O. *et al.,* 2019]. Эффективность WES для выявления патогенных вариантов оказалась значительно выше, чем секвенирования по Сэнгеру, которое обычно ограничивается анализом нескольких генов, связанных с МОДИ, и подтверждает примерно только 15% случаев МОДИ [Shields *et al.,* 2010]. Таким образом, более высокая частота выявления патогенных вариантов в нашем исследовании достигается за счет увеличения числа тестируемых генов и тщательного клинического отбора пациентов с возможным моногенным диабетом. В этой связи следует отметить еще одно преимущество WES: данные секвенирования ДНК могут быть легко сохранены для дальнейшего анализа вновь обнаруженных генов-кандидатов. Учитывая, что моногенный диабет может быть связан с делециями и дупликациями, мы проанализировали возможное присутствие CNV в генах-мишенях. Мы не обнаружили никаких признаков CNV в генах-мишенях ни в одном образце. Однако следует отметить, что у технологии WES есть ограничения, которые не позволяют уверенно обнаруживать маломасштабные CNV. Мы проанализировали взаимосвязь выявленных генетических вариантов с фенотипами пациентов с диабетом. Среди 38 обнаруженных генетических вариантов, о 23 вариантах ранее сообщалось как об ассоциированных с моногенным диабетом, а 15 вариантов были новыми (табл. 9).

Таблица 9. Генетические варианты, выявленные у российских детей с сахарным диабетом не 1 типа [Glotov O. *et al.,* 2019].

Пациент ID	**Ген**	**Нуклеотидная замена (замена в белке)**	**Тип мутации**	**Происхождение мутации**	**Известный /новый вариант**	**Классификация патогенности вариантов по ACMG**
59	*GCK*	c.772G>T (p.Gly258Cys)	миссенс	неизвестно	известный	вероятно, патогенный
62	*GCK*	c.930_931delGG (p.Asp311fs)	Сдвиг рамки считывания	неизвестно	известный	патогенный
83	*GCK*	c.930_931delGG (p.Asp311fs)	Сдвиг рамки считывания	неизвестно	известный	патогенный
95	*GCK*	c.130G>A	миссенс	отец	известный	вероятно,

		(p.Gly44Ser)				патогенный
167	*GCK*	c.128G>A (p.Arg43His)	миссенс	мать	известный	вероятно, патогенный
197	*GCK*	c.233T>C (p.Leu77Pro)	миссенс	отец	известный	вероятно, патогенный
426	*GCK*	c.683C>T (p.Thr228Met)	миссенс	неизвестно	известный	вероятно, патогенный
460	*GCK*	c.682A>G (p.Thr228Ala)	миссенс	мать	известный	вероятно, патогенный
580	*GCK*	c.775G>A (p.Ala259Thr)	миссенс	неизвестно	известный	вероятно, патогенный
663	*GCK*	c.1079C>A (p.Ser360*)	стопкодон	неизвестно	известный	патогенный
665	*GCK*	c.660C>A (p.Cys220*)	стопкодон	неизвестно	известный	патогенный
176	*GCK*	c.1349C>T (p.Ala450Val)	миссенс	неизвестно	известный	вероятно, патогенный
661	*GCK*	c.1349C>T (p.Ala450Val)	миссенс	неизвестно	известный	вероятно, патогенный
118	*GCK*	c.117_119delAAG (p.Lys39del)	делеция	неизвестно	новый	неизвестного клинического значения
119	*GCK*	c.1346_1347delCG (p.Ala449fs)	Сдвиг рамки считывания	неизвестно	новый	патогенный
434	*GCK*	c.868G>C (p.Glu290Gln)	миссенс	мать	новый	неизвестного клинического значения
578	*GCK*	c.1253G>C (p.Ser418Thr)	миссенс	неизвестно	новый	патогенный
27	*GCK*	c.754T>C (p.Cys252Arg)	миссенс	неизвестно	известный	вероятно, патогенный
		c.-71G>C	промотор	неизвестно	известный	вероятно, патогенный
78	*GCK*	c.199G>T (p.Glu67*)	стопкодон	мать	новый	патогенный
		c.766G>C (p.Glu256Lys)	миссенс	мать	известный	вероятно, патогенный
153	*HNF1A*	c.709A>G (p.Asn237Asp)	миссенс	неизвестно	известный	неизвестного клинического значения
422	*HNF1A*	c.485T>G (p.Leu162Arg)	миссенс	неизвестно	новый	неизвестного клинического значения
215	*PAX4*	c.574C>A (p.Arg192Ser)	миссенс	неизвестно	известный	неизвестного клинического значения
114	*ABCC8*	c.4139G>A (p.Arg1380His)	миссенс	неизвестно	известный	вероятно, патогенный
134	*KCNJ11*	c.406C>A (p.Arg136Ser)	миссенс	неизвестно	известный	неизвестного клинического значения
68	*GATA6*	c.1477C>T (p.Arg493*)	стопкодон	de novo	известный	патогенный
266	*WFS1*	c.2452C>T (p.Arg818Cys)	миссенс	мать	известный	вероятно, доброкачественный
408	*WFS1*	c.2327A>T (p.Glu776Val)	миссенс	мать	известный	вероятно, доброкачественный
133	*WFS1*	c.1124G>A (p.Arg375His)	миссенс	неизвестно	новый	неизвестного клинического

						значения
411	*EIF2AK3*	c.1912C>T (p.Arg638*)	стопкодон	от отца	новый	патогенный
	EIF2AK3	c.1912C>T (p.Arg638*)	стопкодон			
432	*SLC19A2*	c.164delC (p.Pro55fs)	Сдвиг рамки считывания	мать	новый	патогенный
	SLC19A2	c.161C>A (p.Thr54Asn)	миссенс	отец	новый	неизвестного клинического значения
226	*GCK*	c.543_545delCGT (p.Val182del)	делеция	неизвестно	новый	неизвестного клинического значения
	HNF1A	c.92G>A (p.Gly31Asp)	миссенс	неизвестно	известный	вероятно, патогенный
529	*BLK*	c.939G>C (p.Glu313Asp)	миссенс	неизвестно	новый	неизвестного клинического значения
	GCK	c.919C>T (p.Leu307Phe)	миссенс	неизвестно	новый	неизвестного клинического значения
662	*GCK*	c.1019+2T>A	нарушение сплайсинга	неизвестно	новый	патогенный
	BLK	c.1148G>A (p.Arg383Gln)	миссенс	неизвестно	новый	неизвестного клинического значения
	WFS1	c.1957C>T (p.Arg653Cys)	миссенс	неизвестно	известный	вероятно, патогенный

Согласно рекомендациям Американского колледжа медицинской генетики и геномики (ACMG) [Richards *et al.,* 2015], большинство обнаруженных генетических вариантов (18 ранее зарегистрированных и 6 новых) были классифицированы как патогенные или вероятные патогенные и, таким образом, они рассматривались как причина фенотипа диабета у исследуемых пациентов. Однако связь обнаруженной миссенс мутации в гене *KCNJ11* с фенотипом диабета не была очевидной, поскольку ранее было показано, что она связана с гиперинсулинизмом [Mohnike *et al.,* 2014], которого не было у пациента № 134. Три ранее зарегистрированных и 9 новых генетических вариантов были классифицированы как имеющие неопределенное значение, а два генетических варианта, вероятно, были доброкачественными (табл. 9). Эти варианты включали 12 миссенс мутаций; для них мы провели дополнительный анализ insilico с использованием I-мутанта 2.0 [Capriotti *et al.,* 2005] (табл. 10).

Таблица 10. In silico прогнозирование увеличения/уменьшения стабильности белка, вызванного миссенс-мутациями неопределенного клинического значения и вероятно доброкачественными [Glotov O. *et al.*, 2019].

Пациент ID	Ген	Нуклеотидная замена (замена в белке)	Классификация патогенности вариантов по ACMG	Стабильность белка, предсказанная с помощью I-Mutant
434	*GCK*	c.868G>C (p.Glu290Gln)	неизвестного клинического значения	снижение
153	*HNF1A*	c.709A>G (p.Asn237Asp)	неизвестного клинического значения	снижение
422	*HNF1A*	c.485T>G (p.Leu162Arg)	неизвестного клинического значения	снижение
215	*PAX4*	c.574C>A (p.Arg192Ser)	неизвестного клинического значения	снижение
134	*KCNJ11*	c.406C>A (p.Arg136Ser)	неизвестного клинического значения	снижение
266	*WFS1*	c.2452C>T (p.Arg818Cys)	вероятно, доброкачественный	снижение
408	*WFS1*	c.2327A>T (p.Glu776Val)	вероятно, доброкачественный	увеличение
133	*WFS1*	c.1124G>A (p.Arg375His)	неизвестного клинического значения	снижение
432	*SLC19A2*	c.161C>A (p.Thr54Asn)	неизвестного клинического значения	снижение
529	*BLK*	c.939G>C (p.Glu313Asp)	неизвестного клинического значения	снижение
	GCK	c.919C>T (p.Leu307Phe)	неизвестного клинического значения	снижение
662	*BLK*	c.1148G>A (p.Arg383Gln)	неизвестного клинического значения	снижение

Во всех случаях, кроме одного, моделирование insilico свидетельствовало о снижении стабильности белка, что свидетельствует о патогенном эффекте проверенных генетических вариантов. Особый интерес представляли два новых генетических варианта в гене *WFS1*, первоначально классифицированных как, вероятно, доброкачественные. Пациент № 266 унаследовал генетический вариант от матери, не страдающей диабетом, в то время как пациент № 408 унаследовал генетический вариант от матери, страдающей диабетом. Гомозиготные мутации в гене *WFS1* приводят к развитию синдрома Вольфрама, аутосомно-рецессивного расстройства, характеризующегося перечнем клинических признаков, включая двустороннюю прогрессирующую атрофию зрительного нерва, глухоту и сахарный диабет [FraserandGunn, 1977]. Сообщалось, что гетерозиготные носители мутаций в гене *WFS1* имеют риск раннего развития сахарного диабета [Bennett *et al.,* 2011]. Последнее нельзя

исключать у наших пациентов. Однако интригующим моментом является то, что генетический вариант в гене *WFS1* у пациента № 408, который унаследовал его от матери, страдающей диабетом, по-видимому, не снижал стабильность белка в соответствии с I-мутантом, что делает его патогенность сомнительной. Наконец, мы проанализировали клиническую картину у пациентов с более чем одним генетическим вариантом в одном или разных генах-мишенях (табл. 11).

Одновременное наличие двух генетических вариантов в гене *GCK* у пациента №27 подняло вопрос об их локализации в одном или обоих аллелях. Родители были недоступны для анализа.

Таблица 11. Клиническая характеристика пациентов с множественными генетическими вариантами в моногенных генах, связанных с диабетом [Glotov O. *et al.*, 2019].

Пациент ID	**Ген Нуклеотидная замена (замена в белке)**	**Возраст постановки диагноза, мес.**	**Диабетический кетоацидоз**	**С-белок нг/мл**	**HbA1C, %**	**SDS BMI**	**Лечение**
27	*GCK* c.754T>C (p.Cys252Arg)	3	нет	0.7	6	-0,63	диета
	GCK c.-71G>C						
78	*GCK* c.199G>T (p.Glu67*)	39	нет	0.63	6.4	+0,83	диета
	GCK c.766G>C (p.Glu256Lys)						
226	*GCK* c.543_545delCGT (p.Val182del)	36	нет	1.1	6	-1,69	диета
	HNF1A c.92G>A (p.Gly31Asp)						
411	*EIF2AK3* c.1912C>T (p.Arg638*)	3	кетонурия	0.2	9.2	-0,72	инсулин
	EIF2AK3 c.1912C>T (p.Arg638*)						
432	*SLC19A2* c.164delC (p.Pro55fs)	48	кетонурия	1.1	5.3	-1.0	инсулин/ диета
	SLC19A2 c.161C>A (p.Thr54Asn)						
529	*BLK* c.939G>C (p.Glu313Asp)	10	нет	0.43	6.7	-0,46	диета
	GCK c.919C>T (p.Leu307Phe)						
662	*GCK* c.1019+2T>A	22	нет	1.1	6.82	-1,32	диета
	BLK c.1148G>A (p.Arg383Gln)						
	WFS1 c.1957C>T (p.Arg653Cys)						

Клиническая картина была легкой и типичной для МОДИ2. Это контрастировало с тяжелым заболеванием, обычно отмечаемым у пациентов с поражением обоих аллелей в гене *GCK*, что позволяет предположить, что у пациента № 27 оба генетических варианта присутствовали в одной и той же аллели и, следовательно, не имели накопительного эффекта. У пациента №78, который также был носителем двух генетических вариантов в гене *GCK*, клиническая картина была типичной для МОДИ2. Поскольку оба генетических варианта были унаследованы от матери, мы пришли к выводу, что была затронута только одна аллель. Более того, из них только вариант c.199G>T оказался клинически значимым, поскольку полученный им стоп кодон завершает трансляцию перед сайтом c.766G>C. Клиническая картина у пациента № 226, у которого были генетические варианты в обоих генах *GCK* и *HNF1A*, была более типичной для МОДИ2, чем для МОДИ3: у него была умеренная гипергликемия натощак и после приема пищи, у него не было гликозурии, и он успешно лечился с помощью диеты. У пациента № 411 была гомозиготная мутация в гене *EIF2AK3*, унаследованная от кровных родителей и связанная с синдромом Уолкотта-Раллисона, который, в свою очередь, как сообщалось, является наиболее распространенной генетической причиной постоянного неонатального диабета в кровных семьях [Rubio-Cabezas *et al.,* 2009]. У пациента № 432 было два новых генетических варианта, влияющих на обе аллели в гене *SLC19A2*. Гомозиготные мутации в гене *SLC19A2* вызывают синдром Роджерса: мегалобластную анемию, чувствительную к тиамину, связанную с сахарным диабетом и глухотой [Labay *et al.,* 1999]. Среди других клинических признаков: врожденные пороки сердца, дегенерация сетчатки, кетонурия, карликовость и неврологическая симптоматика [ShawSmith *et al.*, 2012]. Следует отметить, что у пациента № 432 был только сахарный диабет, дегенерация сетчатки, кетонурия и неврологическая симптоматика, и поэтому у него не было типичной клинической картины. У обоих пациентов №529 и №662 были типичные клинические признаки *GCK*-МОДИ, а не *BLK*-МОДИ, что свидетельствует об

отсутствии сильного накопления патогенного эффекта выявленных генетических вариантов. Среди генетических вариантов, выявленных в нашем исследовании, 60,5% уже были зарегистрированы у больных сахарным диабетом, а 39,5% были новыми. С одной стороны, эти результаты указывают на значительную повторяющуюся вариабельность генов, связанных с моногенным диабетом. С другой стороны, они предполагают, что, несмотря на множество исследований моногенного диабета, многие варианты все еще остаются неопознанными. Выявление новых генетических вариантов, а также накопление данных о ранее известных причинах моногенного диабета имеет большое значение как для фундаментального понимания патогенеза заболевания, так и для клинической практики. Подводя итог, можно сказать, что настоящее исследование демонстрирует высокую частоту и широкий спектр генетических вариантов, вызывающих моногенный диабет у российских детей с сахарным диабетом не 1 типа. Спектр включает ранее известные и новые варианты в генах, связанных и не связанных с МОДИ, с несколькими вариантами у ряда пациентов. Распространенность вариантов в гене *GCK* указывает на то, что диагностика моногенного диабета у российских детей может начинаться с тестирования на МОДИ2.

Еще одним примером обнаружения новых патогенных вариантов с помощью современных технологий секвенирования является исследование спектра мутаций у пациентов с болезнью Вильсона-Коновалова - БВК [Balashova *et al.*, 2020]. Болезнь Вильсона-Коновалова, или гепатоцеребеллярная дегенерация, – аутосомно-рецессивное генетическое заболевание, возникающее вследствие мутаций в гене *ATP7B*. Этот ген кодирует белок-транспортер меди АТФ-азу 7В, который ответственен за распределение меди, поступающей в гепатоциты. АТФ-аза 7В включает медь в церулоплазмин либо выводит ее в желчь через апикальную мембрану гепатоцитов. При дисфункции этого фермента в первую очередь страдают гепатоциты, накопление меди в которых приводит к их разрушению. При своевременной диагностике и лечении

медьэлиминирующими препаратами (терапия хелаторами) прогноз для пациентов с БВК значительно улучшается. Без лечения БВК приводит к ранней инвалидизации и смерти, которая обычно наступает вследствие декомпенсации цирроза печени или его осложнений. Целью нашего исследования было установление частоты патогенных вариантов в гене *ATP7B* в Российской популяции среди пациентов с БВК. Используя разработанную нами таргетную панель удалось провести секвенирование и обнаружить патогенные варианты во всех экзонах гена *ATP7B*, кроме 1, 3, 5, 9, 10, 12 и 21 (рис.11). По своему эффекту варианты распределились следующим образом: 5 аллелей (16,7%) - фреймшифт, 5 (16,7%) - (стоп) нонсенс, 15 (50%) – миссенс, 4 (13,3%) – мутации сайтов сплайсинга, 1 (3,3%) - индел.

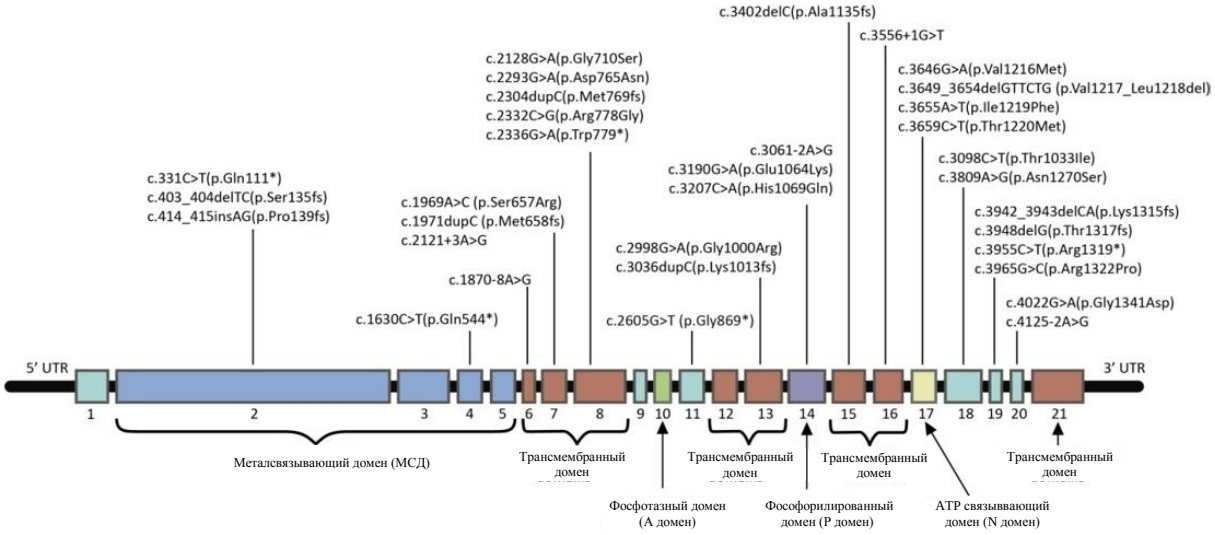


Рисунок 11. Распределение выявленных мутаций по гену *ATP7B* [Balashova *et al.*, 2020].

Мутация c.3207C>A (p.His1069Gln) обнаружилась у 72,9% обследованных. При этом только у 28,2% она присутствовала в гомозиготной форме, в то время как у остальных 44,7% обследованных данная замена была выявлена в компаунд-гетерозиготной форме. Всего с помощью NGS была обнаружена 30 уникальных мутаций. Перечень и характеристика обнаруженных мутаций (в том

числе и неописанных ранее), а также количество выявленных аллелей приведены в таблице 12.

Таблица 12. Генетические варианты, выявленные у российских пациентов с болезнью Вильсона-Коновалова [Balashova *et al.*, 2020].

Домен	Экзон	Вариант	Эффект	N выявленных аллелей	Процент	Описана ранее в литературе как патогенная при БВК
Металл-связывающий	2	c.331C>T (p.Gln111*)	стопкодон	2	1,34%	Да
	4	c.1630C>T (p.Gln544*)	стопкодон	1	0,67%	Да
Трансмембранный	6	c.1870-8A>G	изменение сплайсинга	1	0,67%	Нет
	7	c.1969A>C (p.Ser657Arg)	миссенс	1	0,67%	Да
		c.1971dupC (p.Met658fs)	Сдвиг рамки считывания	1	0,67%	Нет
		c.2121+3A>G	изменение сплайсинга	1	0,67%	Да
	8	c.2128G>A (p.Gly710Ser)	миссенс	1	0,67%	Да
		c.2293G>A (p.Asp765Asn)	миссенс	1	0,67%	Да
		c.2304dupC (p.Met769fs)	изменение сплайсинга	7	4,70%	Да
		c.2332C>G (p.Arg778Gly)	миссенс	3	2,01%	Да
		c.2336G>A (p.Trp779*)	стопкодон	1	0,67%	Да
Фосфатазный	11	c.2605G>T (p.Gly869*)	стопкодон	1	0,67%	Да
Трансмембранный	13	c.2998G>A (p.Gly1000Arg)	миссенс	1	0,67%	Да
		c.3036dupC (p.Lys1013fs)	Сдвиг рамки считывания	2	1,34%	Нет
Фосфорилирующий	14	c.3098C>T(p.Thr1033Ile)	миссенс	1	0,67%	Да
		c.3190G>A (p.Glu1064Lys)	миссенс	12	8,05%	Да
		c.3207C>A (p.His1069Gln)	миссенс	84	56,38%	Да
Трансмембранный	15	c.3402delC (p.Ala1135fs)	Сдвиг рамки считывания	10	6,71%	Да
	16	c.3556+1G>T	изменение сплайсинга	1	0,67%	Да
АТФ-связывающий	17	c.3646G>A (p.Val1216Met)	миссенс	1	0,67%	Да
		c.3649_3654delGTTCTG (p.Val1217_Leu1218del)	индел	6	4,03%	Да
		c.3655A>T (p.Ile1219Phe)	миссенс	1	0,67%	Нет
		c.3659C>T (p.Thr1220Met)	миссенс	1	0,67%	Да
Фосфорилирующий	18	c.3809A>G (p.Asn1270Ser)	миссенс	1	0,67%	Да
		c.3948delG (p.Thr1317fs)	Сдвиг рамки считывания	1	0,67%	Да
Трансмембранн	19	c.3955C>T (p.Arg1319*)	стопкодон	1	0,67%	Да

ый		c.3965G>C (p.Arg1322Pro)	миссенс	1	0,67%	Да
		c.4022G>A (p.Gly1341Asp)	миссенс	1	0,67%	Да
С-конец (стабилизирующая функция)	20	c.4125-2A>G	изменение сплайсинга	3	2,01%	Да

Мы сравнили спектр выявленных мутаций с мутациями, входящими в наиболее часто применяющуюся в РФ панель. Оказалось, что только 4 мутации из 12 встретились в нашей выборке пациентов (c.2304insC, c.3207C>A, c.3402delC,c.3649_3654del6). Кроме того, еще 3 мутации относительно часто встречались среди наших пациентов, но не входили в данную панель (c.2332C>G (p.Arg778Gly), c.4125-2A>G, c.3190G>A (p.Glu1064Lys). Расчетная информативность при анализе только на частую мутацию c.3207C>A составляет 75% пациентов и 82% аллелей, при анализе с помощью панели на 12 частых мутаций – 86% и 93% соответственно. Применение NGS позволило поднять уровень информативности до 96% [Balashova *et al.*, 2020]. Установление факта наличия мутаций в гене *ATP7B* играет важную роль в подтверждении диагноза БВК, так как на момент манифестации клиническая картина БВК может быть крайне неопределенной, и генетическое тестирование на данном этапе является единственным способом подтвердить наличие БВК до формирования выраженных признаков.

Еще одной важной особенностью использования результатов современных методов секвенирования является то, что мы можем выявлять сразу несколько наследственных заболеваний у одного человека. И как оказывается это не такая большая редкость. Одним из примеров является клинический случай совместного наследования Х-сцепленной и аутосомно-доминантной форм ихтиоза [Alaverdian *et al.*, 2019]. Согласно современной классификации, существует две формы наследственного ихтиоза: синдромная и несиндромная, и каждая из них состоит из более чем десяти различных нозологий. Наиболее распространенными типами ихтиоза являются Х-сцепленный рецессивный (распространенность 1/2000-6000 у мужчин) и аутосомно–доминантный, или

вульгарный ихтиоз с неполной пенетрантностью (1/250-1000). Х-сцепленная форма связана с мутациями в гене стероидной сульфатазы *STS*; примечательно, что в 90% случаев наблюдается полная делеция гена. Ихтиоз обыкновенный вызывается гетерозиготными мутациями в гене *FLG*, кодирующем филаггрин. Важно отметить, что клинические формы этих заболеваний неотличимы. Целью этого исследования был поиск патогенных или вероятных патогенных мутаций, которые связаны с различными формами наследственного ихтиоза, такими как другие наследственные заболевания со сходными фенотипическими признаками. Идентифицированная мутация p.Arg2037Ter в гетерозиготном состоянии в гене *FLG* ранее описывалась в базах данных как патогенная. Кроме того, у нашего пациента была выявлена полная делеция в гене *STS*; таким образом, наш пациент несет две патогенные мутации, которые связаны с различными формами наследованного ихтиоза. Для генетического консультирования такая информация может быть очень ценной из-за сходства клинических признаков. На основании проведенного нами исследования рекомендуется для постановки клинического диагноза у таких пациентов анализировать как ген *STS*, так и ген *FLG*, чтобы исключить комбинированные формы ихтиоза.

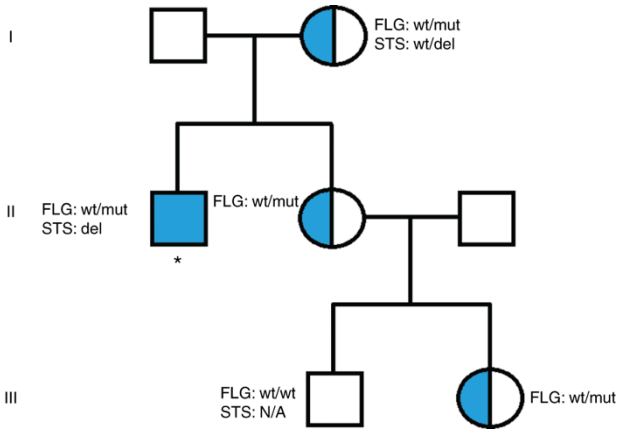


Рисунок 12. Родословная карта пробанда (*). Мать, сестра и дочь сестры имеют симптомы вульгарного ихтиоза, вызванного мутацией в гене *FLG* (p.Arg2037Ter: c.6109C>T) [Alaverdian *et al.*, 2019].

Другим примером сочетанного наследования является внутрисемейный клинический полиморфизм болезни Вильсона-Коновалова [Тулзуновская *и др.*,

2017; Balashova *et al.*, 2020]. В одной семье среди 3-х сибсов, больных БВК, было обнаружено 3 различных мутации в гене *ATP7B*. У всех трех братьев на одной из копий гена была выявлена мутация c.3649_3654delGTTCTG (p.Val1217_Leu1218del). У старшего брата на второй копии гена присутствовала мутация c.3036dupC (p.Lys1013fs), а у младших братьев, являющихся дизиготными близнецами (дизиготность подтверждена фенотипом и различными группами крови, а также данными NGS), мутация c.3207C>A (p.His1069Gln). Клиническая картина заболевания у каждого больного характеризовалась значительными различиями. Только у старшего брата был выявлен патогномоничный для БВК симптом – кольцо Кайзера–Флейшера и неврологические проявления в виде неустойчивости при ходьбе, нарушения речи, почерка. Проведённый скрининг двум младшим братьям, которые на момент обследования не предъявляли жалоб, выявил у одного первоначальные проявления хронического гепатита низкой степени активности (без цирроза печени), у второго – хронический гепатит с трансформацией в цирроз печени, что позволило клиницистам диагностировать БВК. К признакам, которые выявлены у всех 3 больных, относятся: биохимические показатели функции печени, снижение уровня церулоплазмина. Таким образом, у старшего брата установлен диагноз «БВК смешанная форма. Цирроз печени с синдромом портальной гипертензии (спленомегалия), дрожательно-ригидная форма поражения центральной неврной системы». У братьев диагноз был выставлен в результате скринингового обследования родственников пациента с БВК, диагностирована абдоминальная форма БВК с более мягким течением заболевания. Кроме патогенных вариантов в основном гене выявлены варианты в гене модификаторе – *HFE*, ассоциированном с наследственным гемохроматозом (рис.13).

Таким образом, наблюдаемое клиническое разнообразие БВК может быть связано с различиями молекулярных дефектов, приводящих к развитию заболевания. Однако модифицирующее действие различных мутаций и

полиморфизмов других генов, а также разнообразие экзогенных и эндогенных воздействий могут существенным образом изменять течение и исход данного наследственного заболевания [Тулзуновская *и др.,* 2017].

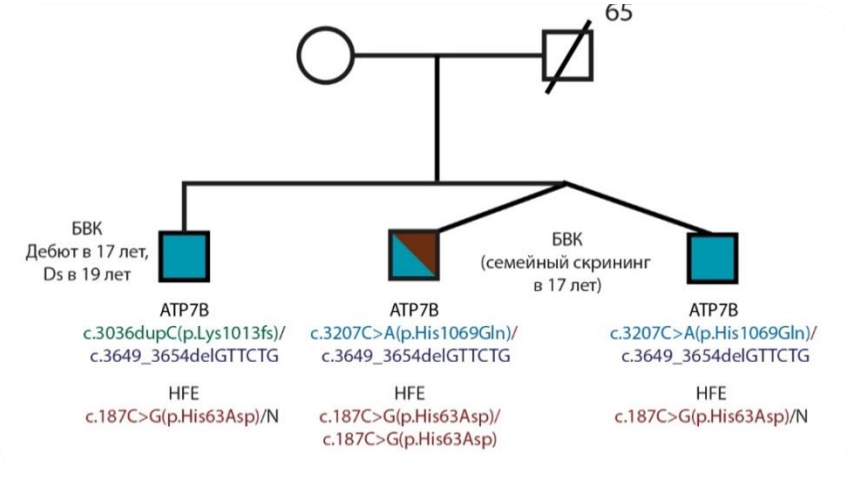


Рисунок 13. Родословная с болезнью Вильсона-Коновалова [Тулзуновская *и др.,* 2017].

Необходимо отметить, что NGS позволяет выявлять патогенные или вероятно патогенные варианты в генах, где ранее мы предполагали лишь один тип мутаций [Koshevaya *et al.,* 2022]. Например, синдром Лопеса-Масиэля-Родана (LOMARS; OMIM 617435) — это редкое аутосомно-рецессивное заболевание, обусловленное предполагаемыми LoF вариантами в гене гентингтина (*HTT*). Экспансия нестабильного CAG-повтора в гене *HTT* (OMIM 613004) приводит к развитию нейродегенеративного заболевания - хореи Гентингтона [MacDonald *et al.,* 1993], однако же о самой функции гентингтина у людей известно мало. До нашего исследования в научной литературе было описано несколько семей, в которых две предполагаемые LoF мутации в гене *HTT* в компаунд-гетерозиготном состоянии являются причиной развития редкого врожденного заболевания с Rett-подобной неврологической симптоматикой – синдрома Лопеса-Масиэля-Родана [Lopes *et al.,* 2016; Rodan *et al.,* 2021]. По описанным данным синдром LOMARS включает в себя

разнообразные симптомы: спастичность конечностей, снижение мышечного тонуса, стереотипные движения рук, дистония, атаксия, эпилепсия, миопия, бруксизм и др. По результатам секвенирования полного экзома были выявлены два неописанных варианта неизвестного клинического значения в гене *HTT*, которые могут отношение к фенотипу: c.2350C>T и c.8440C>A. Дообследование родителей показало, что вариант c.2350C>T был унаследован от отца, а c.8440C>A от матери. Учитывая схожесть клинических данные и результаты дообследования родителей, был заподозрен синдром Лопеса-Масиэля-Родана (LOMARS). Учитывая представленные клинические и молекулярно-генетические данные, мы рекомендуем рассмотреть вероятность LOMARS у детей с неврологическими симптомами, подобными Ретту, и провести молекулярно-генетическое тестирование для поиска предполагаемых мутаций LoF в последовательности гена *HTT* [Koshevaya *et al.,* 2022].

Тщательная диагностическая оценка и адекватная интерпретация результатов имеют решающее значение для быстрой и точной диагностики наследственных нервно-психических заболеваний. Настоящее исследование является попыткой повысить осведомленность врачей об этом редком заболевании и облегчить его диагностику и молекулярно-генетическое подтверждение в будущем [Koshevaya *et al.,* 2022].

1.6. Общая стратегия и алгоритм применения NGS для диагностики генной патологии у человека

Для многих заболеваний важнейшим результатом является корректная терапия. Синдром Floating Harbor (FHS) - чрезвычайно редкое заболевание; в мире зарегистрировано чуть более ста случаев. FHS вызвано гетерозиготными мутациями в гене *SRCAP*. Однако мало что известно о патогенезе FHS или эффективности его лечения. В нашем исследовании мы сообщаем о первом случае FHS в Российской Федерации [Turkunova *et al.,* 2022]. Мужчина-пробанд обладал большинством типичных фенотипических признаков FHS, включая

низкий рост, черты скелета и лица, задержку роста и развития костной ткани, высокий голос и интеллектуальные нарушения. У пробанда также был частичный дефицит гормона роста. Была выявлена патогенная мутация c.7466 C>G (p.Ser2489*) в последнем экзоне связанного с FHS гена *SRCAP*. Систематический обзор литературы и анализ доступных наборов данных о генетических вариациях выявили необычное распределение патогенных вариантов в гене *SRCAP* и подтвердили отсутствие патогенности для вариантов за пределами экзонов 33 и 34. Для понимания клинического эффекта патогенного варианта мы предложили новую модель патогенеза, которая обеспечивает возможную основу для доминантной негативной природы мутаций, вызывающих FHS, и объясняет ограниченные эффекты гормонального лечения при FHS. Наши результаты увеличивают число зарегистрированных случаев FHS и дают новое представление о генетике заболевания и эффективности гормональной терапии у пациентов с FHS [Turkunova *et al.*, 2022].

Исходя из проведенных исследований об оценке эффективности диагностики наследственных заболеваний (табл.13), нами предложена стратегия и алгоритмы экономически эффективной генетической диагностики для МОДИ, болезни Вильсона-Коновалова и других моногенных заболеваний, где есть мажорные мутации (рис.14).

Таблица 13. Эффективность диагностики наследственных заболеваний.

Нозология	**Эффективность диагностики до NGS, в %**	**Эффективность диагностики после NGS, в %**	**Эффективность диагностики с новыми вариантами, в %**	**Ссылка**
Муковисцидоз	45-55 (1 мутация) 58 (35 мутаций)	67-80	-	Неопубликованные данные
Вильсона-Коновалова	До 75 (4 мутации) До 86 (12 мутаций)	До 96	97	Balashova et al., 2020
МОДИ	15-35	40-50	55	Glotov O. et al., 2019

Следует отметить, что не всегда NGS достаточно для постановки диагноза и в ряде случаев при поиске «второго» патогенного варианта после или вместе с NGS необходимо делать секвенирование по Сенгеру.

В частности, у пациента с анаукзетической дисплазией – чрезвычайно редкой формой аутосомно-рецессивной скелетной хондродисплазии - методом NGS секвенирования был идентифицирован вариант rs387906533 (n.91_92delinsGC) нуклеотидной последовательности (chr9:35657924-35657925delCTinsGC) в первом экзоне гена РНК-процессируемой эндорибонуклеазы (*RMRP*) в гетерозиготном состоянии [Федяков *и др.*, 2021].

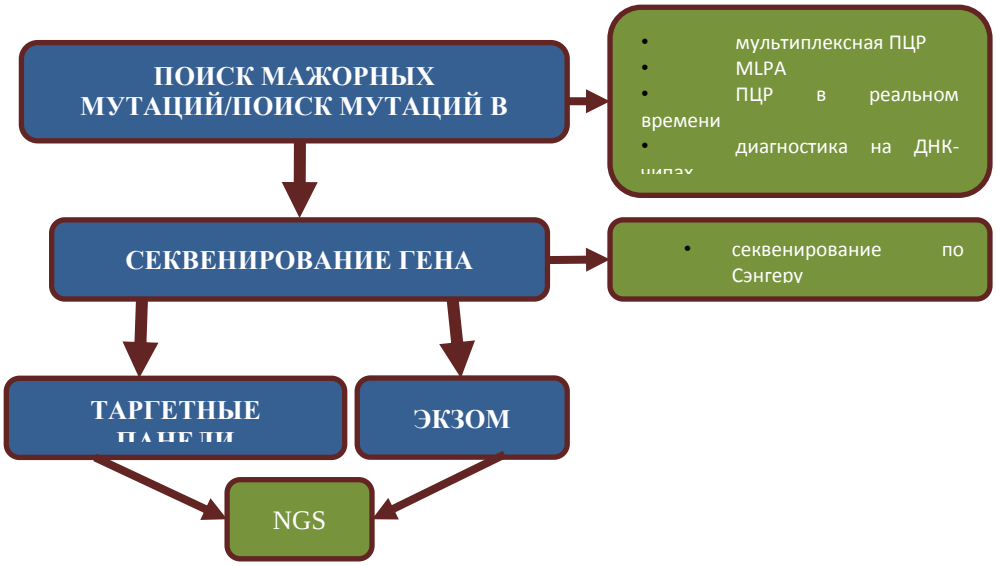


Рисунок 14. Стратегия выбора метода диагностики.

В связи с отсутствием второй мутации по данным секвенирования нового поколения был проанализировал критический регион в промоторной области гена *RMRP* и принято решение провести прямое автоматическое секвенирование по Сэнгеру гена *RMRP* у пробанда (рис.15).

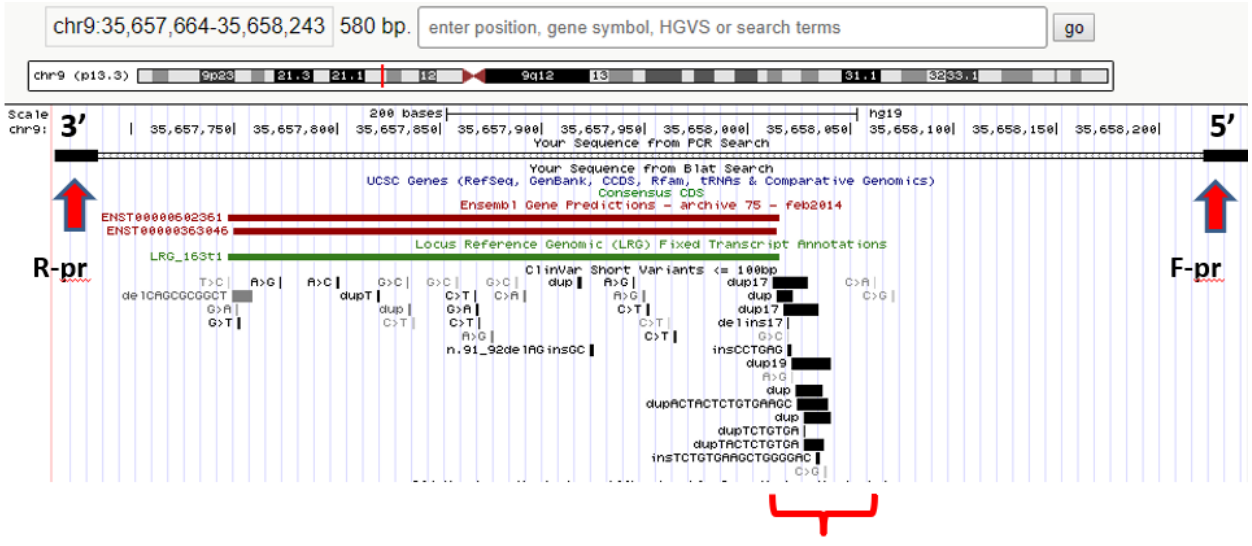


Рисунок 15. Критический регион в промоторной области гена *RMRP* у пробанда.

В ходе исследования было подтверждено наличие мутации n.91_92delinsGC в гетерозиготном состоянии. Кроме того, в промоторной области гена была выявлена ранее неописанная в литературе замена n.-6_-5insTCTCAGCTTCAC (chr9:g.35658020-35658021insTCTCAGCTTCAC) (рис.16).

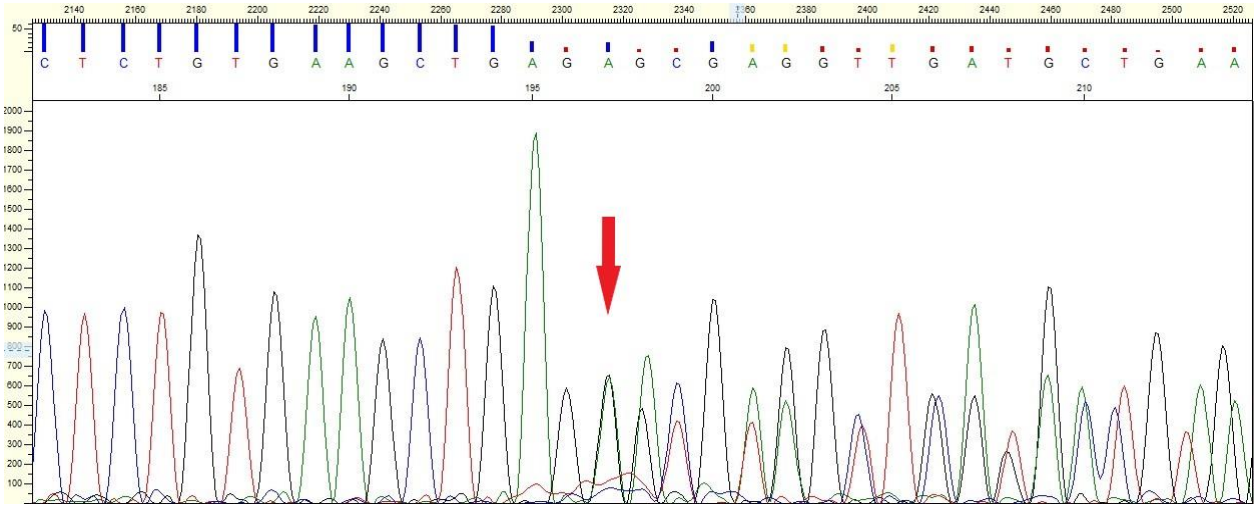


Рисунок 16. Электроферограмма. Патогенный вариант в гетерозиготном состоянии n.-6_-5insTCTCAGCTTCAC в гене *RMRP* у пробанда [Федяков *и др.,* 2021].

Данный вариант представляет собой вставку 12 нуклеотидов в область между ТАТА-боксом и стартом транскрипции. Методом прямого автоматического секвенирования по Сэнгеру был проведен анализ гена *RMRP* у родителей пробанда. Было выяснено, что мутация n.-6_-5insTCTCAGCTTCAC имеет отцовское происхождение, а мутация n.91_92delinsGC - материнское. Инсерция в промоторном регионе гена *RMRP* при АД выявлена впервые, экстраскелетных проявлений (характерных для носителей подобных мутаций) у пациентки на текущий момент не наблюдается [Федяков *и др.,* 2021].

1.7. NGS при планировании семьи для профилактики тяжелых наследственных заболеваний

Сегодня без применения современных методом секвенирования не обходится ни одна сфера медицины. Важнейшим вопросом медицины является

решение проблем репродукции. В нашей работе мы описываем фенотип и репродуктивную историю взрослой пациентки с несбалансированным кариотипом: терминальные делеции 8p23 и 18p11.3 и дубликация 8p22 [Pendina *et al.*, 2019]. Показанием для кариотипирования 28-летней пациентки была структурная перестройка в ее образце биоматериала выкидыша: 45,XX,der(8;18)t(8;18)(p23;p11.3). Неожиданно, что у пациента был тот же кариотип только с одной нормальной хромосомой 8, одной нормальной хромосомой 18 и производной хромосомой, которая была продуктом слияния хромосом 8 и 18 с потерей их концевых областей короткого плеча. У пациента была незначительная дисморфия лица и черепа при отсутствии выраженных физических или психических отклонений. Она была социально нормальной, имела высшее образование и состояла в браке с 26 лет. После четырех неудачных циклов ЭКО/ПГТ пациентка зачала третью естественную беременность. Методом NGS был проведен НИПТ образца крови пациента, который выявил дополнительный материал хромосомы 18. Для проверки результатов НИПТ с помощью цитогенетического анализа был проведен забор ворсинок хориона. Был обнаружен аномальный кариотип, так как имелась только одна нормальная хромосома 8, две нормальные хромосомы 18 и аберрантная хромосома der(8;18). Скорее всего, плод унаследовал аберрантную хромосому der(8;18) вместе с одной нормальной хромосомой 18 от пациента, еще одна нормальная хромосома 18 поступила от супруга пациента, таким образом, сформировав кариотип с фактически тремя копиями хромосомы 18. Беременность была прервана по медицинским показаниям. Следует подчеркнуть, что наследование производной хромосомы пациентки ее потомством вместе с нормальными хромосомами 8 и 18 от супруга пациента нежелательно. Даже если в этом случае у плода будет тот же кариотип, что и у пациента, фенотипический эффект аберрации может быть непредсказуемым. Данный случай демонстрирует необходимость комплексного подхода с

использованием всего арсенала молекулярно-генетических, цитогенетических, эмбриологических методов при планировании беременности.

Другим примером использования NGS может быть применение данной технологии для ПГТ, в частности для поиска хромосомного мозаицизма [Saifitdinova *et al.,* 2020]. Постзиготические митотические ошибки могут привести к появлению клеточных клонов с неодинаковыми наборами хромосом внутри одного и того же эмбриона. Это явление было описано как эмбриональный мозаицизм. Мозаицизм следует тщательно учитывать при обследовании предимплантационных эмбрионов на стадии бластоцисты. Многочисленные данные свидетельствуют об относительно высокой частоте мозаицизма в бластоцисте [Weissman *et al.,* 2017]. Между тем мозаичность клеток трофэктодермы (TE) не обязательно соответствует мозаичности клеток внутренней клеточной массы (ICM) [Munne *et al.,* 2017]. Этот факт может значительно усложнить интерпретацию результатов предимплантационного генетического тестирования (ПГТ). Наблюдение за тем, что мозаицизм в предимплантационных эмбрионах человека выявляется почти в каждом цикле ЭКО с помощью предимплантационного генетического тестирования на анеуплоидию (ПГТ-А), неизбежно поднимает вопрос о том, как можно достоверно оценить уровень мозаичности. Быстрое развитие секвенирования следующего поколения (NGS) в сочетании с методами амплификации всего генома одной клетки стимулировало внедрение этого подхода для ПГТ-А. Высокая чувствительность метода NGS позволяет с высокой степенью достоверности идентифицировать мозаичность в образцах ДНК TE с единичными аномальными клетками (20% для образцов с 5 клетками). С помощью подхода NGS было показано, что частота мозаицизма в предимплантационных бластоцистах варьируется от 17% до 47% в девяти различных центрах вспомогательных репродуктивных технологий (ВРТ) [Sachdev *etal.,* 2016]. Тем не менее, имеющиеся сравнительные данные о значениях хромосомного мозаицизма в клетках TE и ICM остаются

ограниченными и противоречивыми [Munne *et al.,* 2017]. Шесть человеческих бластоцист с мозаицизмом в их клетках TE были разделены на три части, две из которых содержали клетки TE, а одна – преимущественно с клетками ICM, и проанализированы отдельно. Наши данные показывают, что доля анеуплоидных клеток в биопсии, взятой для анализа ПГТ-А, не обязательно отражает истинный хромосомный статус всего эмбриона и не может быть экстраполирована на таковой в клетках ICM. Примечательно, что качественные и количественные характеристики мозаичного состояния могут не совпадать между различными частями одних и тех же эмбрионов, и в образцах, содержащих TE, прилегающих к ICM, мозаичность имеет тенденцию к увеличению, что может иметь физиологическое значение для имплантации. Результаты нашего исследования, очевидно, подтверждают вывод о том, что мозаицизм, выявленный в бластоцисте, снижает вероятность обнаружения эуплоидного набора хромосом в других частях эмбриона [Saifitdinova *et al.,* 2020]. Аномалии у мозаичных эмбрионов непредсказуемо разнообразны. Это может привести не только к потере зачатия, но и к развитию генетического заболевания. Это значительно усложняет интерпретацию результатов предимплантационного генетического тестирования и требует дополнительных исследований для улучшения клинических рекомендаций по переносу эмбрионов.

Успешностью использования различных молекулярно-генетических технологий продемонстрировано в нашем исследовании по возможности оказания медицинской помощи на примере семьи с наследственной патологией [Лязина *и др.,* 2017]. В данной работе представлен сложный и длительный путь диагностики наследственной патологии. У пробанда был выявлены варианты в компаунд-гетерозиготном состоянии: c.851T>G и c.242A>T в гене *PHGDH*. Мутации в данном гене ассоциированы с двумя заболеваниями: синдромом Ноя-Лаксовой (OMIM:#256520) и дефицитом фосфоглицератдегидрогеназы (OMIN:#601815). У ребенка не было классической формы синдрома Ноя-Лаксовой [Лязина *и др.,* 2017]. Семье по результатам диагностики, проведено

медико-генетическое консультирование в декабре 2016 года и дана информация о возможностях пренатальной диагностики. Была разработана тест-система для быстрой диагностики выявленных мутаций. Повторно семья обратилась в феврале 2017 года в связи с беременностью, наступившей естественным путем. При проведении УЗИ первого триместра диагностирована дихориальная диамниотическая двойня. Было показано, что один плод имеет такой же генотип, что и больной пробанд. Прогноз для ребенка крайне неблагоприятный. Методов лечения не разработано. Второй плод – здоровой [Лязина *и др.,* 2017]. Проведена редукция больного плода в сроке беременности 16/17 недель. Беременность здоровым плодом была завершена успешно. Родилась здоровая девочка.

Данный пример демонстрирует необходимость внедрения нового алгоритма преконцепционного обследования семей с использованием всего арсенала молекулярно-генетических методов, включая секвенирование нового поколения как метод первого звена при планировании беременности, а также метода ПГТ и НИПТ для последующего мониторинга беременности. Таким образом, развитие технологий меняет парадигму при планировании семьи и профилактики тяжелых наследственных заболеваний человека. Появляется необходимость в преконцепционном генетическом скрининге для «здоровой» первой беременности (рис.17).

Преконцепционный скрининг:

1. позволяет оптимизировать алгоритм ведения будущей беременности:
- *выбор диагностических процедур;*
- *рекомендации по медицинскому прерыванию;*
- *консультирование;*
- *междисциплинарный подход;*
2. может быть использован на этапе планирования беременности:
- *донорство;*
- *ПГТ;*

3. снижает количество перинатальных потерь;
4. важен при психологической поддержке будущих родителей (позволяет уменьшить количество самообвинений среди тех, кто столкнулся с самопроизвольным прерыванием беременности, в том числе из-за моногенных заболеваний);
5. безопасен и прост для пациента.

Рисунок 17. Изменение парадигмы при планировании семьи для профилактики тяжелых наследственных заболеваний человека.

Выделяют следующие группы моногенных заболеваний, связанные с репродуктивными потерями:

- *моногенные заболевания матери;*
- *моногенные заболевания плода;*
- *другие генетические факторы (гены-предрасположенности, РНК).*

В таблице 14 приведены некоторые моногенные заболевания матери, связанные с высоким риском репродуктивных потерь.

Таблица 14. Некоторые моногенные заболевания матери, связанные с высоким риском репродуктивных потерь.

Заболевание (OMIM)	**Ген (OMIM)**	**Тип наследования**	**Характеристика**
Миотоническая дистрофия, I тип (160900)	*DMPK* (605377)	АД	Повышенный риск прерывания и акушерских осложнений на любом сроке включая выкидыш, преждевременные роды, отёки, внутриутробную гибель плода.
ВДКН (201910)	*CYP21A2* (613815)	АР	Невынашивание беременности встречается у 25% женщин, частота его достоверно снижается после лечения заболевания.
Дефицит субъединицы А фактора 13 (613225)	*F13A1* (134570)	АР	Дефицит FXIIID приводит к кровотечениям, самопроизвольным абортам и другим осложнениям во время беременности. Высока вероятность потери беременности на ранних сроках.
Поликистозная болезнь почек, I тип (173900)	*PKD1* (601313)	АД	При заболевании развивается гипертония и преэклампсия, увеличивается вероятность прерывания беременности.
Синдром удлинённогоQT, I-III типы (192500, 613688, 603830)	*KCNQ1, KCNH2, SCN5A* (607542, 152427, 600163)	АД	Повышенная вероятность гибели плода, а также наблюдается задержка роста у выживших новорождённых.
Болезнь накопления гликогена, Ib тип (232220)	*SLC37A4* (602671)	АР	Повышенный риск самопроизвольных абортов и гибели плода.

Важной проблемой также являются летальные фенотипы у плода, обусловленные моногенными заболеваниями (табл. 15): аутосомно-рецессивными заболеваниями (α-талассемия; синдром множественных птеригиумов, летальный тип; галактосиалидоз; мукополисахаридоз, VII тип), аутосомно-доминантными заболеваниями (танатофорная дисплазия; несовершенный остеогенез, II тип; ахондроплазия; туберозный склероз, I тип), Х-сцепленными заболеваниями (синдром недержания пигмента (Блоха-

Сульцбергера); синдром Гольца (фокальная кожная гипоплазия); синдром Ретта; синдром иммунной дисрегуляции, полиэндокринопатии и энтеропатии).

Таблица 15. Летальные фенотипы у плода, связанные с новыми геномными вариантами.

Год	Гены	Патологии/ механизм	Ссылка
2011	*KIF7*	Гидролетальный синдром, акрокаллезный синдром, синдром Жубера	Putoux *et al.,* 2011
2013	*WDR60*	Перинатальный летальный синдром коротких ребер – полидактилия типа II	McInerney-Leo AM *et al.,* 2013
2014	*FGFR3, COL2A1, OFD1, PRKDC, DLC1, RERE, ACF1, FRAS1*	Летальная скелетная дисплазия, повышенной прозрачностью затылочной кости, трикуспидальной недостаточностью и аномалиями нижних конечностей, вентрикуломегалия и агенез мозолистого тела и др.	Carss *et al.,* 2014
2014	*KIF14*	Цилиопатия	Filges *et al.,* 2014
2015	*THSD1*	Нарушение проницаемости сосудов и поддержания целостности сосудов	Shamseldin *et al.,*2015
2015	*GLE1, RYR1*	Синдром акинезии плода с артрогрипозом	Ellard *et al.,* 2015
2016	*DYNC2H1, ALOX15*	Цилиопатия, плацентарная дисфункция	Qiao *et al.,* 2016
2016	*FOXP3*	Х-сцепленный синдром иммунной дисрегуляции, полиэндокринопатии и энтеропатии (IPEX)/нарушения развитие и поддержание CD3 + CD4 + CD25 + регуляторных Т-клеток	Rae *et al.,* 2016
2018	*ASPM, ATAD3A, ATRX, B3GLCT, BBS9, BBS10, CENPJ, DYNC2H1, ERCC5, ETFA, EXOSC3, FRAS1, GLE1, IFT122, ITGA8, LRP4, MKS1, MRPS22, NEK9, POMGNT1, RYR1, SASS6, TMEM67, TRIP11*	Скелетные дисплазии, акинезия плода, врожденная микроцефалия, синдром Барде-Бидля, синдром Фрейзера и др.	Stals *et al.,* 2018

Сегодня в России в случае, если оба супруга являются носителями генетических мутаций, но хотят иметь здорового ребенка, специалисты рекомендуют проводить процедуру экстракорпорального оплодотворения (ЭКО)

с последующим преимплантационным генетическим тестированием (ПГТ) эмбрионов, что гарантирует родителям возможность иметь здорового малыша.

На основе собственных исследований и данных литературы нами предложен следующий алгоритм генетического обследования пациента с нарушением репродуктивной функции (рис. 18).

Предлагаются следующие последовательные шаги по внедрению преконцепционного скрининга:

1. консенсус специалистов (формирование четких критериев отбора заболеваний для скрининга носительства и критериев для определения круга лиц высокого риска, для которых показано тестирование);
2. решение этико-правовых вопросов;
3. пилотные проекты;
4. разработка тарифа высоких медицинских технологий (ВМП) или программ ОМС профилактики генетических заболеваний для супружеских пар с высоким риском рождения ребёнка с тяжёлой моногенной патологией путем преконцепционного скрининга с последующим преимплатационным генетическим тестированием и пренатальной диагностикой;
5. всесторонняя организационно-консультационная поддержка семей детей, больных моногенным заболеванием или носителей, выявленных в рамках неонатального скрининга.

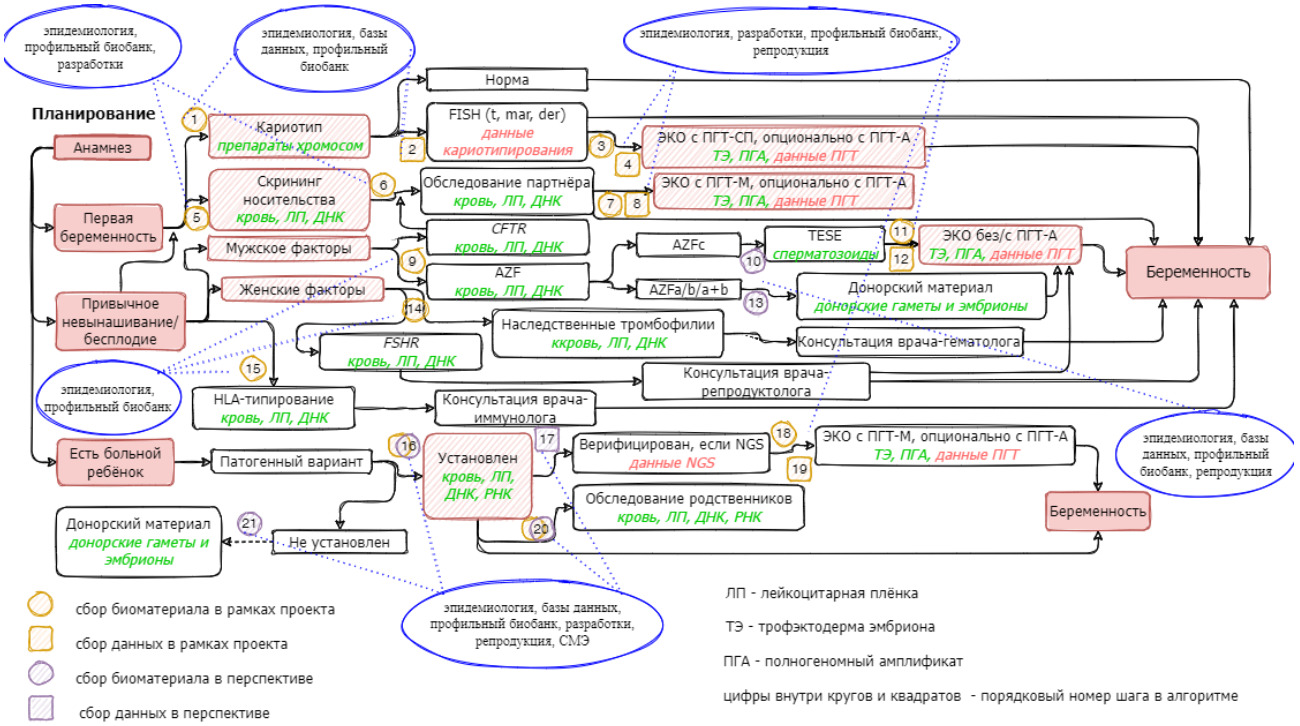


Рисунок 18. Алгоритм генетического обследования пациента с нарушением репродуктивной функции.

Таким образом, концепция Предиктивной медицины - генетического клинического паспорта здоровья для решения задач преконцепционного скрининга, ПГТ, рождения здорового потомства, постановки диагноза и так далее, в разрезе моногенных заболеваний должна опираться на секвенирование нового поколения как базовый метод с использованием собственных баз, алгоритмов и биоинформатики, а также всего арсенала «вспомогательных методов».

ГЛАВА II. Секвенирование нового поколения, анализ фенотипа, олигогенные и мультифакториальные болезни

В предыдущей главе нашего исследования было установлен значительный вклад методов NGS в изучение моногенных заболеваний человека, в понимание значения тех или иных вариантов генома в аспекте современной концепции генетического клинического паспорта здоровья человека. Однако значительно более распространенными являются наследственные болезни, в развитии которых задействованы патогенные варианты сразу нескольких разных генов. Это так называемые олигогенные наследственные болезни [Agarwal and Moorchung, 2005, Kousi and Katsanis, 2015]. Олигогенные болезни представляет собой промежуточное звено между моногенным заболеванием, при котором признак определяется одним геном-возбудителем, и полигенным заболеванием, при котором на признак влияют многие гены и часто факторы окружающей среды. Более того, с точки зрения известного ученого Л.-Ч. Тсуи, все моногенные болезни следует рассматривать как олигогенные [Баранов *и др.,* 2021].

2.1. Кардиомиопатии как олигогенные болезни

Поэтому важным этапом для оценки риска заболевания является понимание его природы. Моногенное, олигогенное или мультифакториальное? Иногда ответить на этот вопрос достаточно сложно. Мы попытались продемонстрировать это положение в наших работах по изучению наследственных кардиомиопатий [Glotov *et al.,* 2015; Komissarova *et al.,* 2016]. Основной целью нашего исследования было выявление генетических маркеров у пациентов с гипертрофической кардиомиопатией (ГКМ) и у студентов из группы риска, которые не вошли в контрольную группу студентов, и выявление маркеров, присутствующих только в контрольной группе. Были рассмотрены две аддитивные модели. Для этого мы предположили, что клинический эффект зависит от наличия альтернативных аллелей или их комбинаций (доминантные

модели), и альтернативных генотипов или их комбинаций между обнаруженными вариантами (рецессивные модели).

Эти предположения привели нас к формированию четырех различных моделей. Наши рецессивные модели не показали достоверного значения для определения группы риска [Glotov *et al.,* 2015]. Однако доминантные модели были более информативными для этих целей, что соответствовало ранее полученным данным об аутосомно-доминантном типе наследования кардиомиопатии [Millat *et al.,* 2014]. Для каждой модели была сгенерирована отдельная таблица. Результаты, полученные от доминантных моделей, получились разными [Glotov *et al.,* 2015]. Исходя из этого можно предположить, что для таких заболеваний необходимо обязательно выяснять тип наследования.

Нами были обнаружены следующие генные варианты только у пациентов с кардиомиопатией и в группе риска студентов: *MYBPC3* (c.977G>A и c.2678G>T) и *CASQ2* (c.1014+12delg), соответственно; *MYBPC3* (c.977G>A) был обнаружен у двух пациентов и одного студента из группы риска; *MYBPC3* (c.2678G>T) был выявлен у двух пациентов и четырех студентов из группы риска; *CASQ2* (c.1014+12delG) был выявлен у пяти пациентов и одного студента из группы риска. Эти варианты были признаны значимыми для тестирования на наличие кардиомиопатий в нашей группе риска, и мы предположили, что людям и членам их семей с этими вариантами следует предлагать предиктивное генетическое тестирование варианта c.977G>A в гене*MYBPC3* [Glotov *et al.,* 2015].

Однако высокая распространенность патогенных вариантов в генеральной популяции является трудностью в диагностике кардиомиопатии, возможно, из-за их неполной пенетрантности [Teekakirikul *et al.,* 2013], что значительно усложняет диагностический скрининг кардиомиопатий, особенно ГКМ.

Анализ SnpSiftпоказал патогенность вариантов гена *TNNT2* (Табл. 16). Некоторые из них (c.97+151delC, c.223+92G>C и c.223+93C>G) были выявлены только в контрольной группе студентов. Таким образом, эти варианты можно

рассматривать как протективные для кардиомиопатии [Glotov *et al.,* 2015]. Между тем, анализ сложных вариантов в других генах, включая *MYBPC3* (c.706A>G) - *MYH7* (c.3973-30A>G), *MYBPC3* (c.3288G>A) - *MYH7* (c.1095G>A), *MYBPC3* (c.3815-66C>T) - *MYH7* (c.1128C>T), *MYBPC3* (c.706A>G) - *MYH7* (c.3853+27T>A), *MYBPC3* (c.706A>G) - *CASQ2* (c.939+23C>T) и *MYBPC3* (c.1223+29G>A) - *MYH7* (c.1095G>A) [Glotov *et al.,* 2015], также может быть полезен, особенно потому, что количество патогенных вариантов у индивидуума может влиять на тяжесть заболевания [Zou *et al.,* 2013]. Мы предполагаем, что гипертрофическая кардиомиопатия, скорее всего, является не моногенным, а более сложным наследственным одногенным заболеванием. Поэтому для понимания его патогенеза необходимо выявлять как генетические, так и средовые причины заболевания [Glotov *et al.,* 2015].

Таблица 16. Основные генетические варианты для кардиомиопатий, выявленные у пациентов и в группе риска по сравнению с контрольной группой [Glotov *et al.,* 2015].

Ген	**Изменение нуклеотида**	**Пациенты/группа риска/контроль, в %**	**Риск**	**Риск2**	**p-оценка**	**Polyphen 2**	**SIFT**	**Клиническая верификация**
MYBPC3	**c.977G>A (NM_000256.3)**	5/4/0	19	-99	0.41	BENIGN	Damaging	Jääskeläinen *et al.,* 2014
MYBPC3	**c.2678G>T (NM_000256.3)**	5/17/0	16	-96	-	PROBABLY DAMAGING	Damaging	-
CASQ2	**c.1014+12delG (NM_001232.3)**	13/4/0	49	-249	8.62E-05	-	-	-
TNNT2	**c.97+151delC (NM_000364.3)**	0/0/10	-100	20	1.80E-05	-	-	-
TNNT2	**c.223+92G>C (NM_000364.3)**	0/0/29	-300	60	1.902E-07	-	-	-
TNNT2	**c.223+93C>G (NM_000364.3)**	0/0/33	-350	70	2.535E-04	-	-	-

2.2. Моногенный сахарный диабет

Еще одним примером «сложных» заболеваний является так называемый «моногенный» диабет (МОДИ), который составляет 1-6% среди - детей и подростков, страдающих тяжелым сахарным диабетом (СД) [Hattersley *et al.,* 2018]. В отличие от СД типа I, имеющего аутоиммунную природу, МОДИ ассоциирован с различными генными мутациями. На сегодняшний день известно 13 генов, вызывающих 13 типов МОДИ, ведущих к развитию умеренной или выраженной гипергликемии у пациентов МОДИ [Barbetti *et al.*, 2018].

Это заболевание обычно диагностируется в возрасте до 25 лет, оно не является инсулинозависимым, и его симптомы обычно умеренные. Однако из-за разнообразия клинических форм, вызванных широким спектром мутаций в генах, связанных с МОДИ, используются различные стратегии лечения: от соответствующей диеты и физической активности до пероральной и/или инсулинотерапии. Сегодня благодаря использованию полногеномного экзомного секвенирования (WES) в случаях, не связанных с геном глюкокиназы *GCK*, удается обнаружить патогенные варианты в других генах [Glotov O. *et al.*, 2019].

По нашим данным, особого внимания заслуживают обнаруженные генетические варианты в различных генах-мишенях у одного пациента. Такие находки выявлены у трех пациентов из обследуемой нами группы. У одного из них делеция в гене *GCK* сопровождалась миссенс-мутацией в гене *HNF1A* (пациент №226). В другом случае присутствовали две миссенс-мутации: в генах *GCK* и *BLK* (пациент №529). У третьего пациента (№ 662) присутствовал дефект сплайсинга в гене *GCK* и миссенс-мутации в генах *BLK* и *WFS1*. Клиническая картина у пациента №226, у которого были патогенные варианты в генах *GCK* и *HNF1A*, была более типичной для МОДИ2, чем для МОДИ3: у него была умеренная гипергликемия натощак и после приема пищи, у него не было

глюкозурии, и он успешно лечился с помощью диеты. У обоих пациентов №529 и №662 были типичные клинические признаки *GCK*-МОДИ, а не *BLK*-МОДИ. Клинические проявления у пациентов с более чем одним генетическим вариантом в одном или разных генах-мишенях подробно представлены в Главе I (раздел «Поиск новых вариантов в геноме пациентов - методом NGS»). Полученные данные свидетельствует об отсутствии существенного накопления патогенного эффекта выявленных генетических вариантов, не связанных с мутацией *GCK* [Glotov O. *et al.,* 2019]. Тем не менее, не исключается их эффект с возрастом, так как клинические признаки, например, МОДИ3, проявляются позднее, чем МОДИ2, что требует дальнейшего наблюдения за этими пациентами.

Выявление новых генетических вариантов, а также накопление данных о ранее известных причинах моногенного диабета имеет большое значение как для фундаментального понимания патогенеза заболевания, так и для клинической практики. Вопрос об отнесении данного заболевания к олигогенным болезням (участии нескольких генов) остается открытым. То же касается и семейной гиперхолестеринемии [Miroshnikova *et al.,* 2021], и наследственных кардиомиопатий [Glotov *et al.,* 2015].

Другим важным аспектом при интерпретации выявленных генетических вариантов является то, что патогенные варианты в одном и том же гене могут приводить к различным заболеваниям. На рисунке 19 представлен спектр патогенных вариантов, вовлеченных в развитие неонатального сахарного диабета (НСД) и МОДИ диабета. На примере гена *GCK* видно, что патогенные варианты в нем могут быть ассоциированы с разными клиническими формами заболевания [Glotov O. *et al.,* 2019].

Известно, что неонатальные формы СД возникают в возрасте до 6 месяцев. Сегодня выявлено более 20 генов, связанных с врожденным неонатальным диабетом [Lemelman *et al.,* 2018]. В зависимости от характера проявления мутаций гена, неонатальный диабет может иметь доминантное или рецессивное

наследование, проявляясь в виде изолированной клинической формы или в рамках различных клинических синдромов [Greeley *et al.*, 2011]. Однако из-за очень раннего начала диабета гипергликемия часто диагностируется до появления других синдромальных признаков. Поэтому неонатальный диабет выделяют в отдельную нозологическую группу. Стратегия лечения немодифицированного неонатального диабета зависит от конкретного генетического дефекта, вызывающего диабетический фенотип.

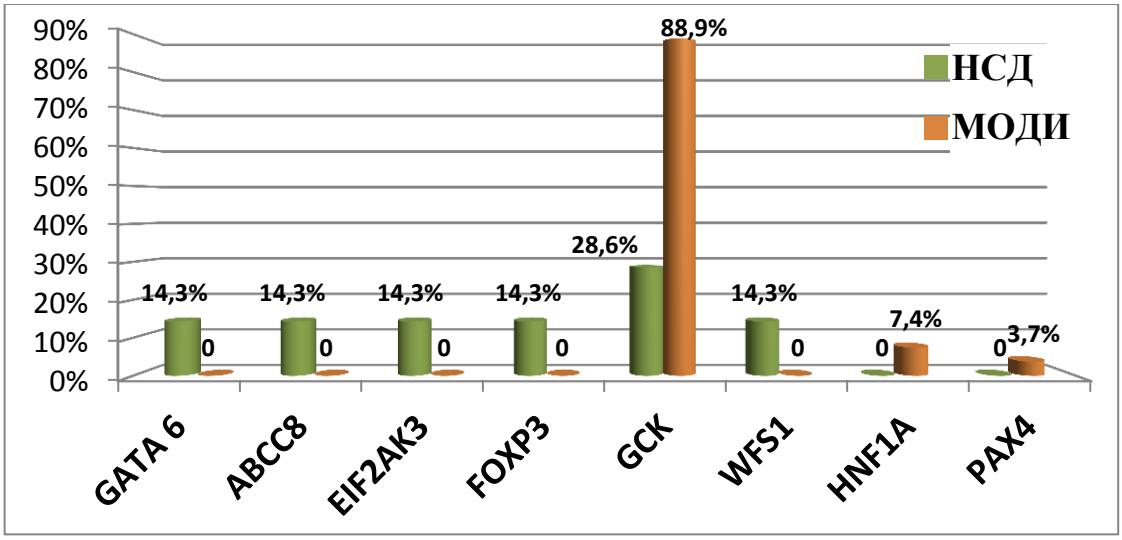


Рисунок 19. Различия в спектре частоте вариантов в генах, вовлеченных в развитие НСД и МОДИ-диабета.

Полученные нами данные о связи патогенных вариантов в одном гене с различными заболеваниями или проявлениями СД (МОДИ и НСД) подчеркивают, насколько важно знать и с осторожностью использовать в медицинской практике понятия «пенетрантность» и «экспрессивность» по отношению к таким генетическим синдромам. Возможно, нозологические формы — это всего лишь различные варианты изменений в работе одних и тех же основных генов с учетом некоторого вклада факторов внешней среды.

2.3. Основы предиктивной медицины

Сегодня мы уже говорим о возможности применения уточненных методов диагностики, профилактики, лечения и прогноза комплексных заболеваний

[Franks *et al.*, 2021], однако широко распространенные хронические заболевания имеют сложную, многофакторную этиологию, которая включает взаимодействие, как генетической восприимчивости, так и факторов риска окружающей среды, которые в широком смысле определяются образом жизни, поведением, профессиональным или экологическими воздействиями, что необходимо учитывать при прогнозе риска заболеваний [Chatterjee and García-Closas, 2016]. Исторически сложилось так, что исследования семейных случаев привели к выявлению редких вариантов с высокой пенетрантностью, лежащих в основе некоторых сложных заболеваний, например семейной гиперхолестеринемии. Благодаря этим открытиям генетическое тестирование стало частью клинического ведения лиц из семей высокого риска. Известно, что вклад средового фактора сводится к минимуму при моногенных и олигогенных заболеваниях, в то время как при болезнях с поздней манифестацией, и особенно в случае мультифакториальных заболеваний, его роль особенно существенна.

Мультифакториальные заболевания (МФЗ), к которым относятся и все наиболее частые хронические заболевания включают: атеросклероз, сахарный диабет, ожирение, бронхиальную астму, остеопороз, эндометриоз, многие злокачественные опухоли, нервно-психические и сердечно-сосудистые заболевания, так как они возникают в результате взаимодействия многих генов с неблагоприятными факторами внешней среды [Баранов *и др.,* 2000]. Досимптоматическое выявление лиц из групп высокого риска по олигогенной или мультифакториальной патологии, а также ее первичная профилактика являются основными задачами предиктивной медицины [Баранов *и др.,* 2009; 2021].

В настоящее время в международной классификации болезней и причин смерти насчитывают более 55 000 нозологических единиц [Petersen, 2021]. Их подавляющее большинство относится к МФЗ. Каждое МФЗ характеризуется выраженной генетической гетерогенностью, обусловленной в значительной мере особенностями мутаций кандидатных генов, их сочетаниями, действием генов-

модификаторов и внешних факторов. Полиэтиологичность любого МФЗ доказывает уникальность сочетанного действия и свидетельствует об актуальности разработки новых вариантов их классификации. На 30 июля 2022 г. показана ассоциация с генетическими вариантами для более чем 12000 болезней человека (рис. 20). На данный момент GWAS-каталог включает 5876 публикаций описывает 220322 SNPs и 402121 ассоциации [https://www.ebi.ac.uk/gwas].

Рисунок 20. GWAS-каталог ассоциаций МФЗ и генов [https://www.ebi.ac.uk/gwas].

В то же время известно, что практически все широко распространенные заболевания, включая почти 90% всех злокачественных опухолей, в той или иной мере связаны с неблагоприятными внешними факторами, среди которых видное место принадлежит курению и неправильному питанию. Различные химические токсины, воздействуя на организм, могут провоцировать начало этих заболеваний.

Известно также, что многие гены способны модифицировать повреждающие эффекты ксенобиотиков, включая экзотоксины. Такие гены кодируют белки (ферменты, нрецепторы, сигнальные молекулы), которые по-

разному взаимодействуют с канцерогенными веществами. Поэтому в зависимости от особенностей генома, различные индивидуумы могут сохранять устойчивость или, напротив, обнаруживать повышенную чувствительность к различным повреждающим агентам [Баранов *и др.,* 2000]. Так, нами было показана связь между полиморфными аллелями *CYP1A1*, *GSTM1* и *CYP2C9* генов и риском развития неходжкинской лимфомы и/или хронического лимфолейкоза [Gra *et al.*, 2008], а также между множеством других заболеваний (бронхиальная астма, невынашивание беременности, эндометриоз и др.) и генами системы биотрансформации [Баранов *и др.,* 2021].

Выяснение молекулярно-генетических причин мультифакториальных заболеваний является достаточно сложной задачей и состоит из нескольких последовательных этапов (рис. 21). Прежде всего, необходимо провести соответствующий литературный поиск и определить генную сеть заболевания. Не менее важными являются и популяционные исследования, позволяющие дать объективную оценку генетического груза популяции, то есть определить частоту встречаемости редких функционально неполноценных аллелей соответствующих генов. На следующем этапе важно сравнить частоты аллелей соответствующих генов в популяции и у конкретных больных с данным клиническим диагнозом. Эти исследования должны быть проведены на достаточно репрезентативных выборках больных и параллельно среди заведомо здоровых в отношении изучаемого заболевания лиц. Только после этого, при наличии четко установленной ассоциации определенных аллелей с соответствующей патологией возможно проведение тестирования наследственной предрасположенности в семьях высокого риска, имеющих данное заболевание. В сравнительно недалеком будущем станет возможным рандомизированное тестирование наследственной предрасположенности [Баранов *и др.* 2005].

Сегодня для поиска генов-кандидатов применяют три основных подхода: метод функционального картирования (анализ кандидатных генов), метод

генетического сцепления в семьях высокого риска, метод полногеномного анализа ассоциаций (метод GWAS), в том числе основанный на секвенировании генома.

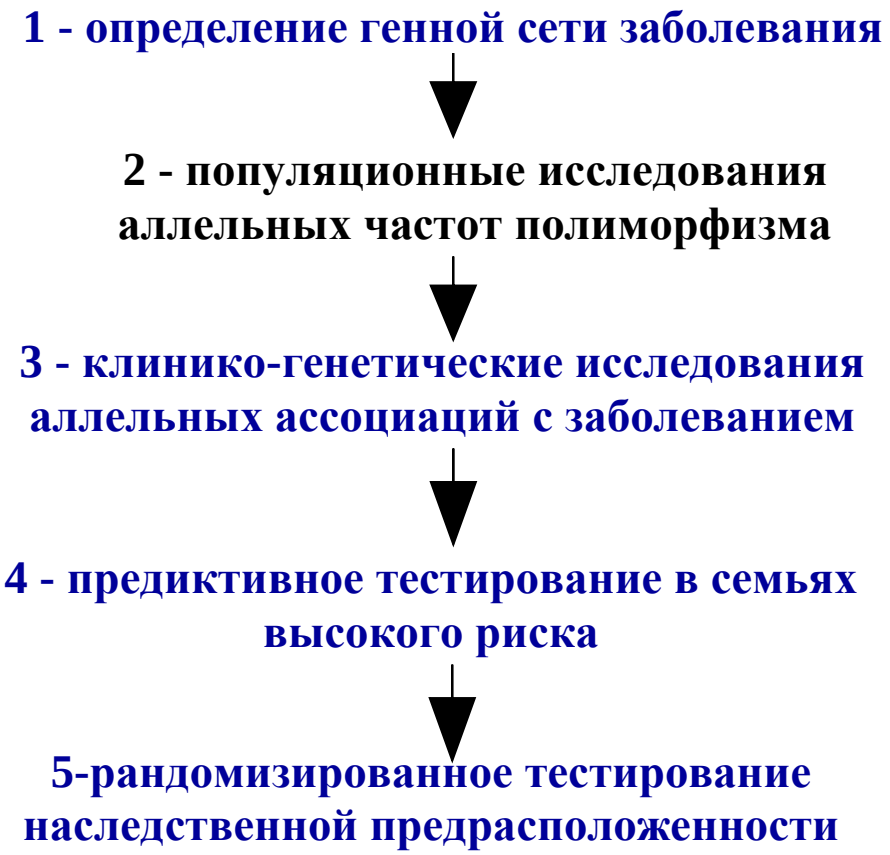


Рисунок 21. Последовательные этапы генетического тестирования мультифакториальных заболеваний [Глотов О.С., 2007].

Анализ работ по генетическому тестированию убеждает в том, что технология GWAS стала основной для поиска генов-кандидатов при МФЗ. Ниже приведены основные этапы метода GWAS.

1. Создание коллекций образцов ДНК (N>1000) индивидов с интересуемым проявлением признака и контрольной группы. Группы должны быть подобраны с учетом популяционных особенностей и хорошо фенотипически охарактеризованы.

2. Генотипирование образцов ДНК при помощи биочипов высокой плотности.

3. Сравнительный анализ аллельных частот и распределения генотипов соответствующих генетических маркеров у индивидуумов с заданным или другим проявлением признака. Выявление генетических маркеров, ассоциированных с признаком.

4. Процедуры репликации (проверки) необходимы для подтверждения результатов первичного сканирования. Для этого используются более точные или альтернативные системы генетического тестирования, проводятся дополнительные исследования на вновь созданных более маленьких выборках или чаще всего на выборках из других популяций.

Технологию GWAS активно применяют не только для проведения научных исследований, но и для скрининга образцов биобанков разных стран, включая биобанк Великобритании [https://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array]. Появление данных по генотипированию образцов различных биобанков и геномных проектов, существенное увеличение выборок способствуют увеличению мощности исследований и переоценке результатов GWAS, полученных ранее.

Существует ряд международных консорциумов, суммирующих усилия отдельных групп и институтов; их деятельность направлена на изучение генетики и проведение GWAS различных заболеваний. К ним относятся, например, психиатрический геномный консорциум (PsychiatricGenomicsConsortium) и международный генетический консорциум по изучению воспалительных заболеваний кишечника (International Inflammatory Bowel Disease Genetics Consortium — IIBDGC) [O'Donovan, 2015; Pierik *et al.*, 2005]. Однако сегодня не только GWAS подходы используют для поиска ассоциаций генетических маркеров и МФЗ.

2.4. Полноэкзомное секвенирование для оценки генетической предрасположенности к сахарному диабету типа 2

Примером использования новых подходов для идентификации генов кандидатов является наша статья, в которой мы при помощи экзомного секвенирования анализировали предрасположенность к сахарному диабету 2 типа и ожирению [Barbitoff *et al.,* 2018].

Известно, что диабет 2 типа (инсулинорезистентный диабет, СД2) и ожирение являются распространенными хроническими заболеваниями с многофакторной этиологией. В последние годы идентифицировано более 128 генетических маркеров предрасположенности к СД2 и более 700 - к ожирению [Wang *et al.,* 2016; Scott *et al.,* 2018; Yengo *et al.,* 2018]. В основном эти исследования были проведены с использованием GWAS [Scott *et al.,* 2018; Fuchsberger *et al.,* 2016; Yengo *et al.,* 2018]. Несмотря на высокую статистическую мощность, выявленные SNP сами по себе обычно не оказывают никакого влияния на сложные признаки, и, скорее всего, они находятся в неравновесном сцеплении (LD) с реальными маркерами заболеваний. Важно отметить, что многие многофакторные признаки формируются в результате сложного взаимодействия между распространенными и редкими вариантами, причем последние обычно упускаются из виду обычными подходами GWAS исследований. Поэтому секвенирование экзома в больших когортах дало новую важную информацию о роли редких вариантов при СД2 и ожирении [Lohmueller *et al.,* 2013]. Однако исследования ассоциаций, основанные на секвенировании всего экзома, обычно страдают от ограничений, связанных с размером выборки, поскольку для выявления значимых локусов в масштабах всего экзома обычно требуются тысячи секвенированных геномов, особенно для высокополигенных признаков. Таким образом, применение секвенирования экзома для анализа сложных признаков требует крупномасштабных исследовательских усилий и/или разработки специальных методов биоинформатического анализа. С другой

стороны, традиционный подход GWAS, предполагающий использование массивов генотипирования, дешевле, но требует дополнительных исследований, таких как точное картирование причинно-следственных вариантов, чтобы получить представление о патогенезе сложных заболеваний. Учитывая эти ограничения современных подходов, мы разработали и применили перспективный подход использования биологически значимых критериев фильтрации для выявления новых вариантов-кандидатов и локусов для СД2 и ожирения в репрезентативной когорте российских пациентов [Barbitoff *et al.*, 2018].

В нашем исследовании мы провели анализ секвенирования экзома у 110 пациентов русской национальности вместе с многоаспектным подходом, основанным на биологически значимых (функциональных) критериях фильтрации для выявления новых вариантов-кандидатов и локусов для СД2 и ожирения. Мы идентифицировали несколько известных SNP в качестве маркеров ожирения (rs11960429), СД2 (rs9379084, rs1126930) и индекса массы тела (ИМТ) (rs11553746, rs1956549 и rs7195386) ($p<0,05$). Используя метод, основанный на подсчете вариантов, специфичных для конкретного случая, вместе с выбором вариантов, изменяющих белок, мы идентифицировали rs328 в гене *LPL* ($p=0,023$), rs11863726 в гене *HBQ1* ($p=8\times10^{-5}$), rs112984085 в гене *VAV3* ($p=4,8\times10^{-4}$) для СД2 и ожирения, rs6271 в гене *DBH* ($p=0,043$), rs62618693 в гене *QSER1* ($p=0,021$), rs61758785 в гене *RAD51B* ($p=1,7\times10^{-4}$), rs34042554 в гене *PCDHA1* ($p=1\times10^{-4}$) и rs144183813 в гене *PLEKHA5* ($p=1,7\times10^{-4}$) для ожирения; и rs9379084 в гене *RREB1* ($p=0,042$), rs2233984 в гене *C6orf15* ($p=0,030$), rs61737764 в гене *ITGB6* ($p=0,035$), rs17801742 в гене *COL2A1* ($p=8,5\times10^{-5}$) и rs685523 в гене *ADAMTS13* ($p=1\times10^{-6}$) для СД2 как значимые локусы восприимчивости в российской популяции [Barbitoff *et al.*, 2018].

Так, мы обнаружили, что вариант rs328 в гене *LPL*, кодирующем липопротеинлипазу, ассоциирован с СД2 и ожирением одновременно. Ранее было показано, что минорная аллель rs328 связана с повышением уровня ЛПНП,

снижением уровня ЛПВП и играет роль в патогенезе СД2 [Mahajan*etal.*, 2018]. Мы также идентифицировали rs6271 в гене *DBH* и rs62618693 в гене *QSER1* в качестве специфических маркеров ожирения. *DBH* — это ген, кодирующий β-гидроксилазу дофамина (DβH), которая катализирует превращение дофамина в норадреналин, который является гормоном и основным нейромедиатором симпатической нервной системы. Ранее было показано, что определенные аллели rs6271 изменяют активность DβH в плазме крови [Zabetian *et al.*, 2003] и вовлечены в регуляцию уровня артериального давления [Ehret *et al.*, 2016]. Важно отметить, что вариант rs62618693 в гене *QSER1* также недавно был обнаружен в качестве маркера СД2 [Mahajan *et al.*, 2018]. Только один вариант среди известных причинных генов (rs2233984 в гене *C6orf15*) был идентифицирован как специфический маркер СД2 при сравнении с контрольной группой. Показана также значимая ассоциация этого полиморфизма с ростом человека [Shungin *et al.*, 2015].

Кроме того, при сравнении больных СД2 и контрольной группы, и пациентов с ожирением были обнаружены три дополнительных варианта (rs9379084 в гене *RREB1*, rs61737764 в гене *ITGB6* и rs17801742 в гене *COL2A1*). В частности, ген *RREB1* кодирует фактор транскрипции, который связывается с RAS-чувствительными элементами (RRES) промоторов генов. Ранее было продемонстрировано, что RREB-1 оказывает репрессивное действие на HLA-G, а также был описан как коактиватор генов кальцитонина, c-ErbB2 и секретина [Flajollet *et al.*, 2009]. Недавние исследования показали связь варианта rs9379084 в гене *RREB1* с распределением жира, уровнем глюкозы натощак [Liu *et al.*, 2013] и с СД2 [Fuchsberger *et al.*, 2016; Mahajan *et al.*, 2018].

Остальные результаты показали наличия ассоциаций СД2 с другими генными вариантами. Так, ген *ITGB6* кодирует интегрин β-6, который является трансмембранным рецептором гликопротеина. Вариант rs61737764 в гене *ITGB6* не был описан как маркер СД2, однако он находится в умеренном LD с другим ранее описанным некодирующим вариантом СД2 - rs7593730. Из редких

вариантов, специфичных для конкретного случая, мы обнаружили rs139972217 в гене *TMC8*, rs61758785 в гене *RAD51B*, rs34042554 в гене *PCDHA1* и rs144183813 в гене *PLEKHA5* в качестве наиболее значимых кандидатов ($p<0,001$). Ген *TMC8* кодирует трансмембранный белок, играющий важную роль в различных кожных заболеваниях. Варианты в локусе TMC6-TMC8 были связаны с уровнями гликированного гемоглобина (HbA1c) - распространенного биомаркера, который используется для диагностики СД2 [Hachiya *et al.*, 2017].

Уровни экспрессии другого гена, несущего ассоциативный сигнал, *PLEKHA5*, связаны с сероконверсией, лежащей в основе диабета 1 типа [Mehdi *et al.*, 2018]. Эти данные указывают на потенциальную высокую значимость выявленных вариантов для патогенеза СД2 и ожирения. Из вариантов экзома с промежуточной частотой ($0,02<SPBUMAF<0,1$) с высоким показателем специфичности для конкретного случая и статистической поддержкой мы обнаружили rs11863726 в гене *HBQ1* и rs112984085 в гене *VAV3*, которые были связаны с СД2 и ожирением по сравнению с контролем, и rs685523 в гене *ADAMTS13* в качестве специфического маркера для СД2. Ген *HBQ1* кодирует субъединицу гемоглобина тета 1, которая экспрессируется только в эритроидной ткани плода человека. Функция этого гена плохо изучена. Ранее не было описано никакой связи полиморфизма гена *HBQ1* с СД2 или другими эндокринными расстройствами. Ген факторов обмена гуаниновых нуклеотидов *VAV3* является членом семейства протоонкогенов VAV. Ген *VAV3* оказывает влияние на ангиогенез, организацию и функцию цитоскелета, регуляцию иммунной системы, что делает его потенциально релевантным геном для молекулярной патологии, лежащей в основе СД2 [Tsuboi *et al.*, 2016]. Ген *ADAMTS13* кодирует мультимерный плазменный гликопротеин, который играет решающую роль в адгезии и агрегации тромбоцитов при сосудистых поражениях. Ранее было показано, что концентрации циркулирующего фактора Виллебранда повышены у пациентов с СД2, и долгосрочные исследования пациентов с СД2 связали фактор Виллебранда с развитием как

микрососудистых, так и макрососудистых заболеваний [Skeppholm *et al.*, 2009]. Было установлено, что фактор Виллебранда является маркером риска раннего летального исхода при СД2 [Stehouwer *et al.*, 2002]. Механизмы, лежащие в основе повышенных концентраций фактора Виллебранда при СД2, остаются неясными, однако эти факты предполагают потенциальную роль гена *ADAMTS13* в патогенезе заболевания. В целом, все три гена, описанные выше, кажутся подходящими объектами для дальнейших генетических и клинических исследований.

Хотелось бы подчеркнуть, что применённый в нашем исследовании многоаспектный подход (рис. 22) для выявления потенциальных маркеров СД2 и ожирения в когорте российских пациентов с СД2 и ожирением является одним из способов поиска вариантов в небольших когортах. Данный подход использует как обычные тесты ассоциации на уровне SNP, так и анализ на уровне генов, а также новые стратегии для выявления вариантов, связанных с СД2, ожирением и соответствующими количественными признаками (ИМТ, ЧСС, концентрация глюкозы и триглицеридов).

Этот подход основан на рациональной фильтрации генетических вариантов, изменяющих белок, и приоритизации генетических вариантов, специфичных для конкретного случая или контроля, т.е. тех, которые являются наиболее вероятным фактором развития заболевания. Мы показываем, что, хотя этот подход обладает низкой мощностью для выявления распространенных вариантов с низкимOR, он эффективно определяет приоритеты вариантов средней и низкой частоты с более высоким OR (рис. 22 (a), (b)). Чтобы уменьшить вероятность ложного обнаружения, которая довольно существенна без каких-либо дополнительных фильтров (P (оценка $\geq$ 20) = 0,047 для варианта с MAF = 0,02 и OR=1), мы также применяем дополнительные корректировки p-значения, которые позволяют выбирать статистически обоснованные редкие варианты с низкой или умеренной вероятностью ошибки 1 типа.

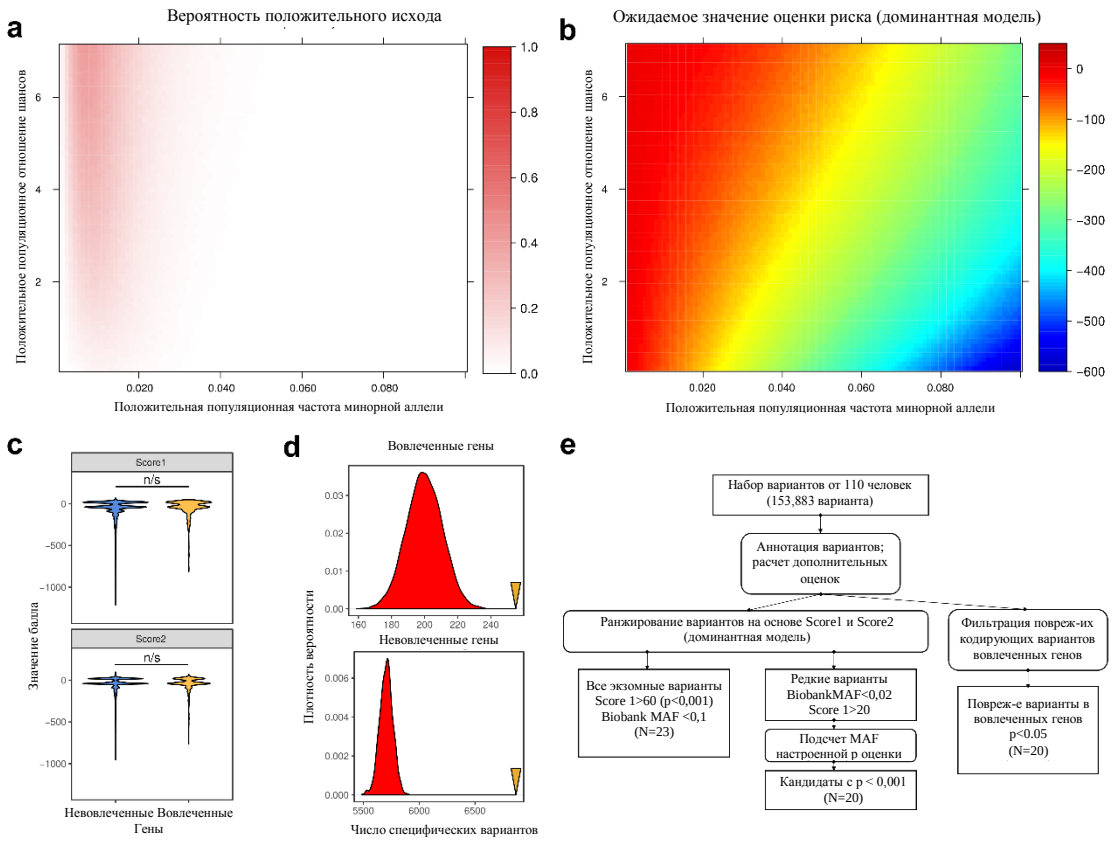


Рисунок 22. Использование методов оценки и фильтрации для выявления маркеров-кандидатов сахарного диабета 2 типа (СД2) и ожирения в российской популяции [Barbitoff *et al.*, 2018].

Примечания: (a, b). Вероятность положительного результата теста (Score1> 10) (a) и ожидаемое значение Score1 (b) для вариантов с разной истинной частотой второстепенных аллелей в популяции (MAF) и истинным отношением шансов (TOR) по оценке с помощью моделирования insilico (c). Распределение значений двух аддитивных оценок (Score1 и Score2) в рамках модели доминантного наследования для повреждающих вариантов кодирования внутри генов, вовлеченных в СД2, и других генов n/s — несущественная разница в U-тесте. (d). Распределения чисел случайных ожиданий, уникальных для конкретного случая вариантов, изменяющих белок, внутри вовлеченных (вверху) и не вовлеченных (внизу) генов. Желтые наконечники стрелок указывают наблюдаемые значения. (e). Схематическое представление анализа данных секвенирования всего экзома (WES), использованного в настоящем

исследовании. Закругленные прямоугольники представляют собой манипуляции с данными.

Применение нашего подхода позволило выявить потенциальные общие и специфические маркеры диабета и ожирения. Мы представляем доказательства потенциальной связи вариантов генов *RREB1*, *ITGB6*, *COL2A1*, *TMC8* и *ADAMTS13* с СД2, а также вариантов генов *HBQ1*, *LPL* и *VAV3* с ожирением в сочетании с диабетом, а также при отсутствии симптомов СД2 [Barbitoff *et al.*, 2018]. Важно отметить, что некоторые варианты (а именно rs685523 в гене *ADAMTS13* и rs61737764 в гене *ITGB6*) были специфичны для группы пациентов с СД2. Вполне вероятно, что эти маркеры контролируют процессы, инициируемые специфическими метаболическими каскадами, которые менее актуальны для недиабетического ожирения. Более того, мы наблюдали ассоциацию rs6271 в гене *DBH*, rs62618693 в гене *QSER1* и вариантов в генах *PCDHA1*, *RAD51B* и *PLEKHA5* с ожирением, не связанным с СД2. Можно предположить, что эти маркеры потенциально участвуют в развитии ожирения, как самостоятельного состояния. Однако эта гипотеза требует дальнейшего подтверждения. Важно отметить, что наша стратегия анализа позволила нам идентифицировать несколько маркеров-кандидатов СД, которые, как было показано, в значительной степени связаны с фенотипом, как было показано методом точного картирования генов (например, rs328 в гене *LPL*, rsrs62618693 в гене *QSER1* и rs9379084 в гене *RREB1*) [Mahajan *et al.*, 2018]. Таким образом, наша стратегия, основанная на фильтрации вариантов, изменяющих белок, внутри вовлеченных генов, может улучшить идентификацию маркеров-кандидатов в небольших группах пробандов. С другой стороны, мы заметили, что частота вариантов, изменяющих структуру белков, значительно повышена в исследуемых группах по сравнению с контрольными индивидуумами как внутри, так и вне известных генов, связанных с заболеванием (рис. 22 Б). Хотя этот результат может быть, частично объяснен слабой генетической связью,

которая не может быть доказана при небольшом размере выборки, недавние результаты [Mahajan *et al.,* 2018] указывают на то, что многочисленные кодирующие маркеры для СД2 на самом деле локализованы за пределами известных генов заболевания. Следовательно, представляется полезным рассмотреть варианты как среди вовлеченных, так и не вовлеченных генов в исследованиях, основанных на секвенировании экзома.

Важно отметить, что большинство редких генных вариантов, выявленных при оценке конкретных случаев заболевания, на самом деле известны как миссенс-мутации, т.е. они сопровождаются изменениями соответствующего белка. В нашем исследовании мы наблюдали несколько специфичных для конкретного случая вариантов в генах, для которых ранее не показана прямая ассоциация с СД2 и/или ожирением (например, в генах *TMC8, PCDHA1, PLEKHA5, HBQ1, VAV3 и ADAMTS13*). Хотя многие из этих вариантов пока не имеют функциональной валидации и не описаны в других исследованиях, изменения кодируемого белка или экспрессии некоторых из этих генов ассоциированы с гликемическими признаками, связанными с диабетом (например, уровни HbA1c для гена *TMC8* [Hachiya *et al.,* 2017] или сероконверсией, связанной с СД1 для гена *PLEKHA5* [Mehdi *et al.,* 2018]). Поэтому данные эти гены могут быть выбраны в качестве потенциальных кандидатов для дальнейших функциональных исследований и проверки воспроизводимости выявленных клинико-генетических ассоциаций.

Подводя итог, можно сказать, что предлагаемый нами подход может помочь в идентификации генов заболеваний по полигенным признакам и демонстрирует эффективность технологий секвенирования всего экзома (WES) для поиска новых маркеров многофакторных заболеваний в когортах ограниченного размера в малоизученных популяциях.

2.5. Поиск оптимальных статистических подходов к оценке генетической предрасположенности

В настоящее время для анализа факторов риска заболеваний и оценки взаимосвязи генотип-фенотип используют различные методы статистики, в том числе классический анализ нулевой гипотезы с применением критерия Хи-квадрат или точного критерия Фишера с подсчетом коэффициента соотношения шансов (OR) и некоторые другие [Реброва, 2003]. Эти методы хорошо зарекомендовали себя при исследовании генетических причин наследственных болезней, однако они недостаточно информативны для оценки риска наследственной предрасположенности. Кроме того, сила данных методов быстро уменьшается при оценке результатов множественного тестирования (использование поправки Бонферрони для уменьшения вероятности ошибки I рода), при небольшом размере выборки, её генетической гетерогенности и скоррелированных переменных [Lvovs *et al.*, 2012].

Существуют разные методы оценки предсказательной силы или качества статистического теста. Необходимо выделить несколько из них. Это и проверка данных с использованием ROC-кривых, анализ ложноположительных и ложноотрицательных результатов, оценка специфичности и чувствительности, и оценка клинической значимости теста. Судить о качестве теста можно по экспертной шкале для значений AUC. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию) [http://www.biometrica.tomsk.ru] - чем ближе полученная оценка чувствительности к 1, тем лучше проверяемый тест диагностирует наличие болезни у пациентов [Власов, 2004]. Если у теста высокая специфичность, то его положительный результат дает основания включить подозреваемую болезнь в дальнейшую дифференциальную диагностику [Власов, 2004]. Несмотря на значимость оценки чувствительности и специфичности, они являются лишь

операционными характеристиками теста и не отражают вероятность наличия болезни после выполнения диагностического теста. На практике врача в основном интересует, какова вероятность болезни у лиц с положительным и у лиц с отрицательным результатом теста [Власов, 2004]. Поэтому, кроме ROC-кривых, еще одной мерой оценки качества теста является его предсказательная полезность, которую оценивают, как для положительного, так и для отрицательного результата исследования [Власов, 2004]. Прогностическая ценность положительного результата диагностического исследования (PPV) равна вероятности наличия заболевания при положительном результате теста. Прогностическая ценность отрицательного результата диагностического исследования (NPV) равна вероятности отсутствия заболевания при отрицательном результате теста. Если PPV=0,8, то у больного с положительным результатом теста вероятность болезни 0,8 или 80%. К получению показателя такого типа направлена вся диагностическая деятельность [Власов, 2004]. Однако на практике нельзя ограничиваться оценкой показателей PPV и NPV, а надо оценивать диагностический тест по его чувствительности и специфичности. Поэтому высокая надежность окончательного диагноза (высокая прогностичность конечного результата) - не доказательство эффективности теста. Возможно, что высокая прогностичность теста является простым следствием того, что на результат в изученной группе влияет фактор предварительного отбора пациентов [Власов, 2004].

Сегодня не существует единого подхода к обработке результатов научных исследований в области изучения генетической предрасположенности, и, более того, отсутствует единая оценка целесообразности применения того или иного метода и его эффективности при оценке риска МФЗ. Поэтому требуется всесторонний и взвешенный подход как для поиска маркеров заболевания, так и для оценки их пригодности для точного прогноза риска.

Несколько лет назад для оценки вклада нескольких полиморфных генов в патологический процесс нами была разработана оригинальная балльная оценка –

оценка «суммы баллов комбинации аллелей» и «суммы неблагоприятных баллов» по генам ренин-ангиотензиновой и кинин-брадикининовой систем. Данный подход показал свою эффективность ранее при исследовании формирования стабильной АГ у детей, в котором было установлено, что аллельные варианты генов ренин-ангиотензиновой и брадикининовой систем ассоциированы с развитием стабильной гипертонии у девочек [Глотов *и др.,* 2007; Баранов *и др.,* 2009б]. Подсчет баллов проводится по формуле ∑m/k, где m – сумма баллов всех индивидуальных генотипов, а k – число изученных генов. Подсчет ведется раздельно для каждой генной сети (метаболической системы) - анализируется сумма баллов в каждой группе, после чего, число баллов в каждой группе суммируется. Важно отметить, что метод «суммы баллов» предполагает равный вклад каждого гена в риск развития заболевания, а само заболевание понимается как сумма аддитивных патологических реакций, контролируемых генами предрасположенности. Между тем, пример с Лейденовской мутацией V фактора свертывания крови показывает, что это далеко не так. Более того, в процессе клинических и лабораторных проверок реальная балльная оценка каждого гена постоянно уточняется и может существенно меняться [Баранов *и др.,* 2009б].

Каждый человек имеет свой характерный фенотип, отличающий его от других ему подобных индивидов. Под фенотипом принято понимать сумму всех признаков, причем не только таких как рост или цвет глаз, число пальцев на руках и ногах и т. д., но и различные физиологические и биохимические признаки. Большинство фенотипических проявлений относятся к сложным признакам, которые контролируются множественными генетическими и средовыми факторами [Инге-Вечтомов, 2010]. Согласно современным представлениям, влияние генетических факторов на экспрессию и пенетрантность фенотипических признаков в значительной мере обусловлено наличием «качественного» генетического полиморфизма, обусловленного, главным образом, однонуклеотидными вариантами или заменами, single

nucleotide polymorphism (SNP). В настоящее время общепринятой является точка зрения, что в контроле сложных признаков принимают участие как редкие SNP с сильным фенотипическим эффектом, так и относительно частые SNP с менее выраженным эффектом [Reich, Lander 2001; Pritchard, Cox, 2002]. Обнаружено большое число SNP в различных генах и/или участках генома, которые связаны с разнообразными количественными фенотипическими признаками человека, такими как рост, вес, спектр липопротеинов и другими [Аульченко, 2010]. Понимая, как генотип определяет фенотип, проще осуществлять поиск генов-кандидатов, ассоциированных с проявлениями мультифакториальных заболеваний (МФЗ) [Баранов, 2009].

2.6. Полигенные эффекты при анализе показателей антропометрии, липидного метаболизма и физиологического обмена

Длина тела (рост) является классическим полигенным признаком, важным для идентификации личности. С позиции генетики рост человека является полигенно наследуемым признаком. Многочисленные исследования показали, что доля дисперсии роста, обусловленная наследственными факторами, составляет 80–90% [Weedon *et al.,* 2007; Аульченко, 2010]. Для оценки генетического вклада в фенотип используют разные математические модели, в том числе и регрессионную модель.

Одним из примеров применения регрессионной модели для прогнозирования фенотипа на основе генетических маркеров является наша работа по оценке роста человека [Glotov *et al.,* 2014]. При построении модели предсказания роста мы использовали анкетные данные индивидов (актуальный рост человека, пол) и данные о генотипах по 13-ти вариантам в генах *EFEMP1, ZBTB38, HHIP, LCORL, ADAMTSL3, CDH13, JAZF1, IGF1R, GHSR, CABLES1, IFNG, VDR3, IGFBP3*. Наиболее высокие показатели коэффициентов детерминации получены для мужчин. Используя разработанную модель для индивидов мужского пола, рост «стоя» можно предсказывать с ошибкой + 4,6

см, а рост «сидя» — + 2,9 см. Полученные коэффициенты детерминации для женщин (0,013 и 0,006 для роста «стоя» и «сидя» соответственно) значительно ниже, чем для мужчин (0,109 и 0,127, для роста «стоя» и «сидя» соответственно), однако ошибка прогноза роста практически не отличается. Наши результаты свидетельствуют о большей генетической детерминированности роста среди мужчин. Нельзя не отметить, что для предсказания роста «сидя» получены лучшие характеристики, чем для предсказания роста «стоя». При измерении роста «стоя» пациент становится спиной к вертикальной стойке, касаясь ее пятками, ягодицами и межлопаточной областью. Тогда как при измерении роста «сидя» пациент садится на скамейку, касаясь вертикальной стойки ягодицами и межлопаточной областью. Таким образом, при измерении роста в положении сидя не учитывается истинная длина нижних конечностей, дающая необходимое представление о пропорциях тела. Исходя из вышесказанного, можно предположить, что рост «стоя» является более сложным признаком, чем рост «сидя» (а известно, что модели гораздо лучше описывают простые признаки, чем сложные) [Аульченко, 2010]. Необходимо отметить, что предлагаемая нами линейная модель позволяет проводить оценку роста как количественного признака, дальнейшее совершенствование которой путем включения новых генетических параметров и параметров внешней среды - представляется весьма перспективным [Glotov *et al.,* 2014].

Похожий подход мы использовали при анализе ассоциации полиморфизма генов метаболизма липидов с индексом массы тела, обхватом талии и параметрами липидограммы крови у женщин [Тарковская *и др.,* 2012]. Для построения модели предсказания количественных признаков (ИМТ, ОБ, ХС, ХС ЛПНП, ХС ЛПОНП) применяли также логистическую регрессию. В качестве исходных параметров модели использовались данные об аллелях индивидуумов по рассматриваемымSNP: гомозиготы по аллели, имеющей большую частоту (в исследуемой популяции), получали код «0», гетерозиготы — код «1», гомозиготы по аллели, имеющей меньшую частоту — код «2». Далее методом

пошагового включения/удаления параметров (stepwisealgorithm) в случае каждого признака была определена наиболее точная модель и, соответственно, наиболее значимые параметры (SNP). В качестве критерия точности модели использовали информационный критерий Акайке (AIC). При построении модели предсказания для анализируемых признаков мы использовали анкетные данные индивидов (ИМТ, ОТ), биохимические параметры (ХС, ХС ЛПНП и ХС ЛПОНП) и данные об их генотипах по 36 маркерам. Нами получены модели с различными параметрами и определены наиболее точные из них, позволяющие предсказывать ряд параметров человека с учетом его генетических особенностей. Известно, что эффективность данных моделей можно оценить на основании скорректированного коэффициента детерминации (adjusted R2), который позволяет сравнивать между собой модели, учитывающие разное число факторов [Mc Cullagh and Nelder, 1989]. Коэффициент детерминации является одним из основных критериев, по которым проводится оценка эффективности модели. Его значение изменяется в интервале [0;1]. Данный критерий показывает, насколько полно учитываемые факторы и их комбинации, присутствующие в прогностической модели, описывают изменение значений признака. При значении коэффициента 0 модель считается неэффективной. Основу оценки эффективности модели составляет анализ наблюдаемых и прогнозируемых значений. Расхождение между этими значениями называется ошибкой прогноза. Поскольку значение ошибки прогноза непосредственно зависит от конкретной выборки, по которой строится модель (получение списка значимых параметров (из всех рассматриваемых) и значений коэффициентов для них), то и сама ошибка прогноза является случайной величиной. Для анализа степени влияния слагаемых ошибки на ее значение и применяют коэффициент детерминации R2, который, по сути, и отображает общий вклад параметров в изучаемый признак.

По нашим данным наиболее высокие показатели коэффициента детерминации получены для параметра ХС ЛПОНП (R2=0.101). Более того, для

этого показателя выявлена наименьшая ошибка прогноза значения (± 0,21 ммоль/литр) по сравнению с ошибкой уровней общего ХС (± 0,95 ммоль/литр) и ЛПНП (±0,83 ммоль/литр). Полученный нами результат свидетельствует о большей генетической детерминированности ЛПОНП по сравнению с другими показателями [Тарковская *и др.,* 2012].

Для верификации данных, полученных с помощью регрессионных моделей, мы использовали метод корреляционного анализа. Для изучения корреляции между генами и анализируемыми признаками был применен метод ранговой корреляции по Кендаллу (τ). Данный метод позволяет сравнивать качественные порядковые признаки между собой (генотипы) и качественные порядковые признаки (генотипы) с количественными [Реброва, 2003]. Корреляционный анализ по Кендаллу частично подтвердил ассоциацию между генотипами проанализированных генов и ИМТ, ОТ, уровнями ХС, ХС ЛПНП и ХС ЛПОНП, которые были выявлены с помощью регрессионной линейной модели. Следует так же отметить, что метод линейной регрессии позволил установить ряд новых связей, не выявленных при использовании метода корреляционного анализа.

Таким образом, гены, выявленные с помощью регрессионной модели и верифицированные корреляционным анализом, безусловно, участвуют в определении фенотипических и биохимических особенностей индивидуума. Для некоторых показателей (в группе женщин) скорректированный коэффициент детерминации (adjusted R2) более 0,1, что свидетельствует о достаточно большой роли генетической компоненты в определении изучаемых признаков. Однако еда, режим питания, физическая активность, стресс, вредные привычки, экология, лекарства — все это может сыграть большую роль в жизни и «перекрыть» влияние генетических факторов [Тарковская *и др.,* 2012].

Еще одним примером применения регрессионной модели для оценки фенотипических признаков на основе анамнестических данных, генетических и лабораторных анализов является проведенное нами исследование спортсменов

[Glotov O. *et al.,* 2015]. Известно, что важными показателями оценки риска развития сердечно-сосудистых заболеваний у спортсменов являются результаты многократного измерения (мониторинга) различных физиологических показателей, таких как жизненная емкость легких (ЖЕЛ), пульс, артериальное давление, индекс массы тела, представляющих собой типичные сложные полигенные признаки человека [Wood *et al.,* 1998; Puthucheary *et al.,* 2011]. Учитывая большую индивидуальную и временную вариабельность этих физиологических параметров, представляет интерес выяснить, в какой мере они коррелируют с индивидуальными особенностями генома и возможно ли при наличии таких корреляций построение различных моделей здоровья [Пузырев, 2011; Xu and Hu, 2010]. Нами была поставлена задача по поиску корреляционных связей и построению регрессионных моделей для раннего предсказания уровня развития важных для спортивной деятельности физиологических показателей, используя различные методы отбора регрессоров, и разработка методической базы для дальнейших исследований по тестированию многих генов на больших выборках. В качестве объекта исследования была выбрана ЖЕЛ, имеющая высокую наследственную компоненту и меняющаяся под влиянием специальных физических нагрузок. Моделирование функции ЖЕЛ представляется особенно важным в профилактической и спортивной медицине, прогнозировании спортивных успехов и развитии физических возможностей организма, так как возможность ее предсказания позволит принимать меры, направленные на своевременную коррекцию этого существенного параметра [Allen *et al.,* 2010]. Регрессионный анализ оказался значительно более чувствительным, чем другие статистические методы. При построении регрессионной модели анализируемых признаков были использованы анкетные данные индивидов, биохимические параметры и результаты генотипирования 26-ти генов [Glotov O. *et al.,* 2015]. В регрессионной модели выявлены статистически значимые связи между ЖЕЛ и аллелями следующих генов: *AGTR2* (отрицательная, минорная аллель A), *NOS3*

(положительная, минорная аллель 4), *CNB1* (отрицательная, минорная аллель D), *ADRB2*_81CG (отрицательная, минорная аллель G). Значимый коэффициент корреляции по Кендаллу отмечен только для гена *NOS3*. Полученные результаты свидетельствуют о большей чувствительности регрессионного анализа по сравнению с корреляционным анализом. Регрессионная модель позволяет проводить первичную оценку фенотипических показателей индивидов (ЖЕЛ) на основании результатов генетического тестирования, с учетом дополнительных параметров. В связи с тем, что ЖЕЛ имеет высокую наследственную составляющую и является показателем риска мультифакторных заболеваний, разработанные модели могут найти применение для оценки риска развития этих заболеваний, а также для прогноза спортивных успехов и развития физических возможностей организма [Glotov O. *et al.,* 2015].

2.7. Перспективы комплексной индивидуальной диагностики полигенных факторов МФЗ

В целом, на основании анализа сбалансированной точности моделей и экспертной шкалы значений AUC лучшие модели получены при одновременном учете генетических, анамнестических и клинических данных не только для предсказания фенотипа, но также и для оценки риска МФЗ. Включение в расчёты преморбидного фона в качестве одного из показателей не позволяет применять такую модель для оценки риска патологии на доклиническом этапе, и оставляет лишь возможность разработки тактики коррекции заболевания на начальных стадиях. При этом, большинство исследователей использует метод общей линейной модели (GLM), так как его можно применять для оценки как качественных, так и количественных признаков [McCullagh, Nelder, 1989; Yi, Banerjee, 2009; Глотов *и др.,* 2012; Huang *et al.,* 2014]. Основным достоинством метода является то, что предсказание фенотипа может базироваться не только на данных о генотипе пациента, но и на анализе таких переменных как анамнез и клинико-лабораторные данные. Суммируя вышесказанное, стандартная

процедура оценки эффективности геномного предиктора заключается в построении ROC кривой и вычислении площади под кривой ROC (AUC).

Ярким примером интерактивного расчета риска развития заболевания является работа De Haan с коллегами [De Haan *et al.*, 2012]. На 2012 год не существовало доступных моделей риска, которые точно предсказывали бы риск развития венозного тромбоза у человека. Поэтому цель их работы состояла в том, чтобы выяснить, улучшает ли включение однонуклеотидного полиморфизма, связанного с тромбозом (SNP), в модель прогнозирования риска венозного тромбоза. Из 31 SNP, связанных с венозным тромбозом, они рассчитали оценки генетического риска путем подсчета аллелей, повышающих риск для участников крупного исследования «случай-контроль», включающего 2712 пациентов и 4634 человек контрольной группы. Оценки генетического риска, основанные на всех 31 SNP или на 5 наиболее сильно ассоциированных SNP, выполнялись аналогичным образом (области под кривыми [AUCs] 0,70 и 0,69 соответственно). Для оценки риска по 5 SNP коэффициенты вероятности венозного тромбоза варьировались от 0,37 (95% доверительный интервал (ДИ), 0,25-0,53) для лиц с 0 аллелями риска до 7,48 (95% ДИ, 4,49-12,46) для лиц с более чем 6 аллелями риска или равными им. AUC модели риска, основанной на известных негенетических факторах риска, составила 0,77 (95% ДИ 0,76-0,78). Сочетание негенетической и генетической моделей риска улучшило AUC до 0.82 (95% ДИ 0,81-0,83), что указывает на хорошую диагностическую точность [de Haan *et al.*, 2012]. В этом исследовании отдельные SNP не были существенно связаны с рецидивирующим венозным тромбозом. Однако, когда аллели риска отдельных SNP были объединены, оценка риска, а также значимость ассоциации возросли (табл. 17 и рис. 23).

Прогностическая способность множественного SNP-анализа не изучалась для первичных случаев венозного тромбоза. Гораздо важнее оценивать риск тромбоза при основных других заболеваниях. Таким образом, генетическое профилирование может определять решения о профилактических мерах в

группах высокого риска, таких как больные раком, лица, перенесшие операцию, лица, нуждающиеся в гипсовой повязке, или лица, подвергающиеся длительной иммобилизации.

Таблица 17. Оценка генетического риска, основанная на 5 наиболее сильно ассоциированныхSNP (по De Haan *et al.*, 2012).

Ген	SNP	Хромосома	MEGA				Усредненные литературные данные, OR
			Частота аллели риска, %		OR	95% CI	
			Опыт	Контроль			
F5	rs6025	1	10	3	4.30	3.70-4.99	3.79
F2	rs1799963	11	6	2	3.01	2.36-3.85	2.78
ABO	rs8176719	9	47	34	1.74	1.63-1.87	1.85
FGG	rs2066865	4	34	27	1.41	1.32-1.51	1.56
F11	rs2036914	4	59	52	1.35	1.26-1.44	1.32

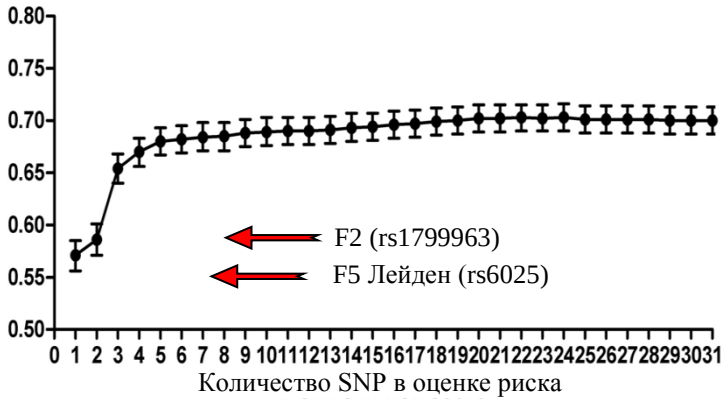


Рисунок 23. ROC кривые для оценки предсказания риска тромбофилии (по De Haan *et al.*, 2012).

Чтобы изучить клиническое применение генетического профилирования, De Haan с коллегами [De Haan *et al.*, 2012] более подробно изучили группы, подверженные воздействию известных негенетических факторов. Дискриминационная точность оценок генетического риска в этих подгруппах была аналогична дискриминационной точности в общей исследуемой популяции, за исключением больных раком (табл. 18). Субанализ у больных раком в зависимости от терапии (химиотерапия, хирургия, облучение) или класса опухоли (солидная по сравнению с другими) не улучшил точность определения взвешенной оценки риска по 5 SNP.

Чтобы оценить, работает ли оценка генетического риска лучше, чем текущая клиническая практика оценки семейного анамнеза, были сравнены точность оценки генетического риска с оценкой риска только по семейному анамнезу. AUC показателя риска 5-SNP (0,68, 95% ДИ, 0,67-0,70) был значительно выше, чем AUC семейного анамнеза (0,58, 95% ДИ, 0,57-0,60), при этом аналогичная тенденция наблюдалась среди всех подгрупп лиц с высоким риском. AUC для оценки негенетического риска, включая семейный анамнез, составил 0,77 (95% ДИ 0,76-0,78). Когда de Haan с коллегами добавили оценку генетического риска к негенетическому баллу, AUC значительно увеличилась до 0,82 (95% ДИ, 0,81-0,83) - по сравнению только с оценкой негенетического риска (P< 0,0001) с использованием либо 31-SNP, либо 5-SNP оценки риска (рис.24).

Таблица 18. Оценка генетического, основанная на 5 наиболее сильно ассоциированных SNP, семейного риска, негенетических факторов и комбинированная (по De Haan *et al.*, 2012).

Группы риска	Пациенты, N	Контроль, N	Семейный risk score, AUC (95% CI)	5-SNP risk score, AUC (95% CI)	Негенетический risk score, AUC (95% CI)	Комбинированный risk score, AUC (95% CI)
Хирургические вмешательства	292	111	0.60 (0.55-0.66)	0.66 (0.60-0.72)	0.67 (0.61-0.72)	0.73 (0.67-0.78)
Травмы конечностей, требующие иммобилизации	111	18	0.61 (0.48-0.73)	0.73 (0.59-0.87)	0.70 (0.56-0.84)	0.78 (0.64-0.91)
Госпитализация	278	93	0.57 (0.50-0.63)	0.66 (0.59-0.72)	0.60 (0.53-0.66)	0.66 (0.59-0.72)
Оральные контроцептивы*	513	327	0.58 (0.54-0.62)	0.71 (0.68-0.75)	0.73 (0.69-0.76)	0.81 (0.78-0.84)
Заместитетльная гормональная терапия	58	90	0.59 (0.49-0.68)	0.71 (0.63-0.80)	0.74 (0.66-0.82)	0.80 (0.72-0.87)
Беременность/послеродовый период*	67	46	0.54 (0.44-0.65)	0.70 (0.60-0.79)	0.68 (0.57-0.79)	0.76 (0.66-0.85)
Возраст > 50 лет	944	1534	0.57 (0.55-0.60)	0.68 (0.66-0.70)	0.73 (0.71-0.75)	0.79 (0.77-0.81)
Путешествия	379	610	0.58 (0.54-0.62)	0.70 (0.67-0.73)	0.77 (0.73-0.80)	0.82 (0.80-0.85)
Отягощенная наследственность	659	551	-	0.68 (0.65-0.71)	0.74 (0.71-0.76)	0.81 (0.78-0.83)
Злокачественные новообразования	156	65	0.57 (0.49-0.65)	0.60 (0.52-0.68)	0.71 (0.64-0.78)	0.72 (0.65-0.80)

- * Женщины моложе 50 лет.

Как негенетическая, так и комбинированная модели оценки риска показали лучшие результаты у женщин, чем у мужчин (оценка негенетического риска: AUC = 0,81, 95% ДИ, 0,80-0,83 для женщин и AUC = 0,74, 95% ДИ, 0,72-0,75 для мужчин; комбинированная оценка риска: AUC = 0,85, 95% ДИ, 0,83-0,86 для женщин и AUC = 0,80, 95% ДИ, 0,78-0,81 для мужчин).

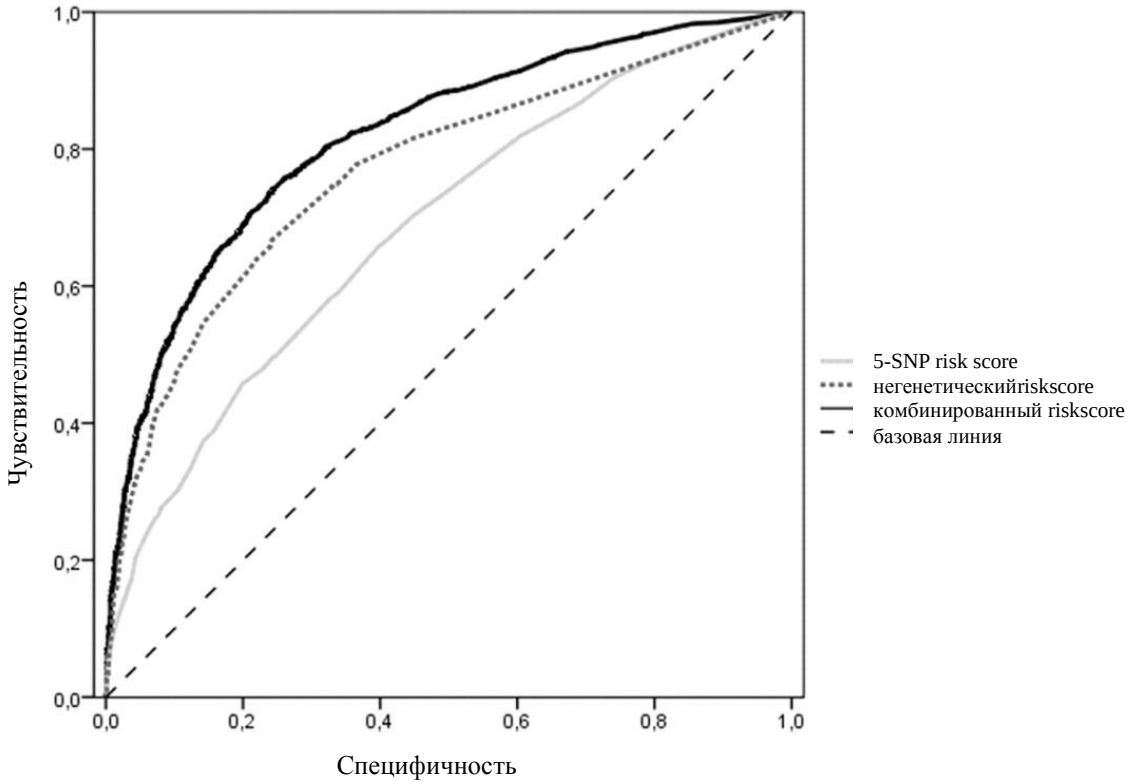


Рисунок 24. AUC кривые оценки рисков (по de Haan *et al.*, 2012).

Дискриминационная точность оценки риска как по 5 SNP, так и по 31 SNP была воспроизведена в другом исследовании (LETS) (табл.19), что свидетельствует о надежности генетических моделей [De Haan *et al.*, 2012].

Таблица 19. Оценка риска венозного тромбоза, используя генетические, негенетические и комбинированные модели (по De Haan *et al.*, 2012).

	MEGA (N = 7092)		LETS (N = 881)	
	AUC (95% CI)	r2	AUC (95% CI)	r2
31-SNP risk score	0.71 (0.69-0.72)	0.161	0.69 (0.65-0.72)	0.149
5-SNP risk score	0.69 (0.67-0.70)	0.135	0.67 (0.64-0.71)	0.138
Генетический risk score	0.77 (0.76-0.78)	0.288	0.71 (0.68-0.74)	0.200
Комбинированный risk score	0.82 (0.81-0.83)	0.378	0.77 (0.74-0.80)	0.292

Важность исследования генетических маркеров связана с тем, что профилактические меры после положительного теста являются инвазивными или могут иметь вредные побочные эффекты, поэтому требуется строгая дискриминация между лицами с высоким и низким риском развития конкретного заболевания. В случае венозного тромбоза неразборчивость может привести к повышенному риску тромбоза у лиц с высоким риском, получающих недостаточное профилактическое лечение антикоагулянтами, в то время как лица с низким риском, получающие лечение, подвергаются повышенному риску серьезных кровотечений. Исследование показало, в какой степени оценки генетического риска могут повысить точность оценки риска тромбоза с помощью ROC-кривых. Хотя доля вариабельности, объясняемая оценкой риска по 5 SNP, меньше, чем по шкале риска по 31 SNP, De Haan с коллегами [De Haan *et al.*, 2012] показали, что дискриминационная точность оценок риска по 5 SNP и 31 SNP была одинаковой. Оценка генетического риска по 5 SNP показала лучшие результаты, чем оценка семейного анамнеза, которая является текущей клинической практикой оценки риска у лиц, подверженных воздействию известных негенетических факторов риска. Однако оценка генетического риска по 5 SNP показала худшие результаты, чем оценка риска негенетических факторов риска. Таким образом, добавление оценки генетического риска по 5 SNP к модели оценки негенетического риска значительно улучшило AUC до 0,82, что указывает на хорошую диагностическую точность. Во всех подгруппах лиц с высоким риском комбинированный показатель риска показал лучшие результаты, чем только негенетический показатель, что может указывать на потенциальную клиническую ценность генетического профилирования у этих лиц с высоким риском.

Недавно [Khera *et al.*, 2018] создали предикторы риска для фибрилляции предсердий, диабета 2 типа, рака молочной железы, воспалительных заболеваний кишечника и ишемической болезни сердца (ИБС). Они получили следующие AUCs 0,77, 0,72, 0,68, 0,63 и 0,81 соответственно. Однако, для

получения этих результатов используются дополнительные переменные, такие как возраст и пол. Когда в предикторах используются только общие SNP, соответствующие AUCs меньше [Lello *et al.*, 2019] и мы получаем AUCs в диапазоне 0,580–0,707 (табл. 20), используя только данные SNP.

Таблица 20. Генетические AUCs с использованием только SNP без учета возраста или пола. Обучение и валидация проводятся с использованием UKBB [Lello *et al.*, 2019].

Заболевание	Тренировочный набор	Тестируемая выборка	AUC	Активные SNPs	λ*
Гипотериоз	impute	UKBB	0.705 (0.009)	3704 (41)	1.406e-06 (1.33e-7)
Гипотериоз	impute	eMERGE	0.630 (0.006)		
СД2	impute	UKBB	0.640 (0.015)	4168 (61)	6.93e-06 (1.73e-6)
СД2	impute	eMERGE	0.633 (0.006)		
Гипертония	impute	UKBB	0.667 (0.012)	9674 (55)	4.46e-6 (4.86e-7)
Гипертония	impute	eMERGE	0.651 (0.007)		
Устойчивая гипертония	impute	eMERGE	0.6861 (0.001)		
Астма	calls	AA	0.632 (0.006)	3215 (16)	2.37e-6 (0.35e-6)
СД1	calls	AA	0.647 (0.006)	50 (7)	7.9e-7 (0.1e-7)
РМЖ	calls	AA	0.582 (0.006)	480 (62)	3.38e-6 (0.05e-6)
Рак простаты	calls	AA	0.6399 (0.0077)	448 (347)	3.07e-6 (0.08e-8)
Рак яичек	calls	AA	0.65 (0.02)	19 (7)	1.42e-6 (0.04e-6)
Глаукома	calls	AA	0.606 (0.006)	610 (114)	8.69e-7 (0.71e-7)
Подагра	calls	AA	0.682 (0.007)	1010 (35)	9.41e-7 (0.03e-7)
Фибрилляция предсердий	calls	AA	0.643 (0.006)	181 (39)	8.61e-7 (0.94e-7)
Желчекаменная болезнь	calls	AA	0.625 (0.006)	981 (163)	1.01e-7 (0.02e-7))
Инфаркт	calls	AA	0.591 (0.006)	1364 (49)	1.181e-6 (0.002e-7)
Повышенный уровнь холистерина	calls	AA	0.628 (0.006)	3543 (36)	2.4e-6 (0.2e-6)
Злокачественная меланома	calls	AA	0.580 (0.006)	26 (15)	9.5e-7 (0.8e-7)
Базальноклеточная карцинома	calls	AA	0.631 (0.006)	76 (22)	9.9e-7 (0.3e-7)

Существенно более высокие AUCs получаются за счет включения дополнительных переменных, таких как возраст и пол (табл. 21). Ожидается быстрое улучшения геномного прогнозирования по мере того, как для анализа станет доступно больше данных о случаях контроля. Представляется вероятным, что геномное предсказание риска заболевания для ряда важных заболеваний вскоре станет достаточно хорошим, чтобы широко применяться в клинических условиях [Lello *et al.*, 2019].

Таблица 21. AUCs, полученные с использованием только пола и возраста, только SNP и всех трех вместе взятых [Lello *et al.*, 2019].

Заболевание	Тестируемая выборка	Возраст + Пол	Только генетика	Возраст + Пол + Генетика
Гипертония	UKBB	0.638 (0.018)	0.667 (0.012)	0.717 (0.007)
Гипотиреоз	UKBB	0.695 (0.007)	0.705 (0.009)	0.783 (0.008)
СД2	UKBB	0.672 (0.009)	0.640 (0.015)	0.651 (0.013)
Гипертония	eMERGE	0.818 (0.008)	0.651 (0.007)	0.851 (0.009)
Устойчивая гипертония	eMERGE	0.817 (0.008)	0.686 (0.007)	0.864 (0.009)
Гипотиреоз	eMERGE	0.643 (0.006)	0.630 (0.006)	0.697 (0.007)
СД2	eMERGE	0.565 (0.006)	0.633 (0.006)	0.651 (0.007)

Значительная наследуемость большинства распространенных заболеваний подразумевает, что по крайней мере некоторые различия в риске обусловлены генетическими эффектами. Располагая достаточным количеством обучающих данных, современные методы машинного обучения позволяют нам создавать полигенные предикторы риска. Алгоритм обучения с достаточным количеством примеров для обучения может в конечном итоге, основываясь только на генотипе идентифицировать людей, которые подвергаются необычно высокому риску этого заболевания. Это имеет очевидное клиническое применение: скудные ресурсы для профилактики и диагностики могут быть более эффективно распределены, если лица с высоким риском могут быть идентифицированы превентивно. Это можно сделать в раннем возрасте или даже до рождения [Lello *et al.*, 2019].

На основании вышеизложенного уже сегодня мы можем говорить о точной медицине (уточненной диагностике для задач клинической медицины). Несмотря на то что анамнестические данные предсказывают риск МФЗ несколько лучше, чем генетические и поскольку сбалансированная точность «генетических моделей» ниже «анамнестических», «генетические модели» позволяют выделять группы риска досимптоматически и когда нет какой-либо клинической информации. Однако если в «генетическую» модель «добавить» данные анамнеза, то такая GLM-модель уже может быть использована и для оценки риска патологии, т.к. обладает высокими показателями сбалансированной точности [Глотов А.С., 2017].

Таблица 22. Эффективность клинических прогнозов с учетом генетической, анамнестической, клинической информации [Глотов А.С., 2017].

Наличие клинической информации о пациенте:	**Категория маркеров**				**Рекомендуемый метод**
	Гены	**Анамнез**	**Клинические, биохимические, иммунологические и инструментальные данные («клиника»)**	**Биомаркеры**	
Нет данных	+	-	-	-	MDR
«Анамнез»	++	++	-	-	GLM
«Клиника»	+++	++	++	+++	
«Анамнез + преморбидный фон»	+++	++	+++	++++	
Диагноз	+++.	+++	++++	++++	

В настоящий момент применение геномики для улучшения диагностики, профилактики, лечения и прогнозирования не только широко обсуждается [Collins and Varmus, 2015; Баранов *и др.,* 2021], но и уже становится стандартом медицинской помощи, причем не только для онкологических и редких, но и для распространенных заболеваний.

Сегодня точная медицина, основанная на геномных данных («управляемая геномом») – это:

- дифференциальная диагностика;
- профилактика;

- лечение;
- прогноз.

Важным аспектом точной медицины является оценка клинической и технической готовности потенциальных тестов, которые могут принести пользу врачам и пациентам для диагностики сложных заболеваний. Хотя геномика, вероятно, будет играть ключевую роль в медицине будущего, ожидается, что она будет делать это в соответствии с демографическими и стандартными клиническими данными (например, возраст, пол, история болезни в прошлом, текущее состояние здоровья, семейный анамнез, негенетические биомаркеры и воздействие факторов окружающей среды). Вполне вероятно, что генетические данные не являются полезными или необходимыми для всех видов точной медицины. В ряде случаев мониторинг заболеваний, вероятно, будет основываться на повторных оценках других OMICs-технологий, например, на оценке экспрессии генов с помощью РНК-секвенирования.

Примером поиска потенциальных биомаркеров адаптации является наша статья [Glotov A. *et al.,* 2022]. Известно, что тренировки на выносливость на большой высоте становятся все более популярными среди спортсменов, в то время как молекулярные и клеточные основы этой адаптации остаются недостаточно изученными. В нашем исследовании мы стремились определить основные физиологические изменения и выявить потенциальные биомаркеры адаптации, используя транскрипционное профилирование цельной крови. Семь элитных конькобежек были отобраны на 18-й день высотной адаптации. РНК-секвенирование цельной крови до и после интенсивной 1-часовой тренировки на коньках использовали для измерения изменений экспрессии генов, связанных с физической нагрузкой. Чтобы идентифицировать гены, специфически регулируемые на больших высотах, мы использовали данные из восьми ранее опубликованных наборов данных микрочипов, изучающих изменения экспрессии в крови после тренировки на уровне моря. Используя сигнатуры, специфичные для конкретного типа клеток, мы смогли оценить изменения

численности типов клеток из изменений экспрессии отдельных генов. Среди них были ген *PHOSHO1*, играющий известную роль в эритропоэзе, и ген *MARC1*, играющий роль в эндогенном метаболизме NO. Мы обнаружили, что количество тромбоцитов и эритроцитов однозначно реагирует на высотные упражнения, в то время как изменения в нейтрофилах представляют собой более общий маркер интенсивных упражнений. Общедоступные данные как из атласов отдельных клеток, так и из профилирования крови, связанного с физической нагрузкой, значительно повышают ценность РНК-seq цельной крови для динамической оценки физиологических изменений в организме спортсмена [Glotov A. *et al.*, 2022].

Почти для всех сложных заболеваний генетический риск является вероятностным и недетерминированным (последнее справедливо для заболеваний, вызванных с высоко пенетрантными вариантами). Это создает сложности, поскольку степень риска оценить сложнее, чем четкое наличие или отсутствие известного патологического варианта. Потенциальное преимущество заключается в том, что повышенный генетический риск сложного заболевания можно профилактировать и назначить раннее лечение, поскольку варианты ядерной ДНК остаются стабильными на протяжении всей жизни [Глотов АС, 2017].

Как уже было упомянуто выше, использование генетических тестов для более распространенных сложных заболеваний продемонстрировало большой потенциал в некоторых исследовательских учреждениях, но пока еще не перешло от исследований к клинической практике. Распространенным способом характеристики генетического риска сложных заболеваний является использование *polygenic risks core (PRSs)* [Franks *et al.*, 2021]. PRS - это сумма множества (иногда тысяч) генетических вариантов, которые по отдельности дают небольшие эффекты. Недавние исследования показывают, что высокие PRSs создают большие и потенциально клинически значимые риски во взрослом возрасте для таких заболеваний, как сердечно-сосудистые заболевания и

сахарный диабет 2 типа [Khera *et al.,* 2018], а также для прогноза выживаемости при системной красной волчанке [Reid *et al.,* 2020]. Поскольку большинство сложных заболеваний неоднородны по этиологии, генетика может помочь в диагностике, определяя подгруппы/ подтипы в рамках нетрадиционной комплексной диагностики заболеваний, тем самым получая преимущество от целенаправленного лечения [Franks *et al.,* 2021]. Примером может служить ишемический инсульт с различными этиологическими подтипами (окклюзия крупных сосудов, окклюзия мелких сосудов, кардиоэмболический инсульт или расслоение артерий), каждый из которых может иметь различную генетическую архитектуру, требующую различной целевой терапии и клинического наблюдения. Геномная медицина может также помочь выявить редкие состояния, которые скрыты в рамках сложной диагностики заболевания. Например, примерно у 3% пациентов с хронической обструктивной болезнью легких наблюдается дефицит альфа-1- антитрипсина [Marciniuk *et al.,* 2012]. Дефицит альфа-1- антитрипсина чаще всего вызывается гомозиготностью по аллели *SERPINA1**Z* и ассоциирован как с хронической обструктивной болезнью легких, так и с циррозом печени и гепатоцеллюлярной карциномой [Silverman and Sandhaus, 2009], при раннем выявлении которых рекомендуется специальное клиническое наблюдение и лечение.

В идеале болезни следует предотвращать, а не лечить. Успешным примером является фенилкетонурия, тяжелые последствия которой (например, нарушения когнитивного развития) можно предотвратить путем раннего выявления и последующего соблюдения диеты, не содержащей фенилаланина. Другими хорошо известными примерами являются семейная гиперхолестеринемия, которую можно выявить и лечить с целью предотвращения различных коронарных событий [Sturm *et al.,* 2018], и генетические тесты генов *BRCA1-2* при раке молочной железы [Wooster *et al.,* 1995]. Некоторые специфические, высокоэффективные гены-мишени были выявлены в исследованиях наследственной гиперлипидемии и последующего

риска ИБС. К ним относятся такие гены, как *LDLR, PCSK9, APOB, LDLRAP1* и *ABCG8* [Miroshnikova *et al.*, 2021]. Например, инактивируя мутации в гене *PCSK9* можно вызывать снижение холестерина ЛПНП и, соответственно, снижение риска ИБС [Cohen *et al.*, 2006], также используя моноклональные антитела к PCSK9 можно резко снижать уровень ЛПНП и риск серьезных сердечно-сосудистых событий [Robinson *et al.*, 2015]. Методы лечения, нацеленные на PCSK9, в настоящее время внедрены в руководящие принципы клинической практики.

Ключевой потребностью общественного здравоохранения является внедрение геномной медицины и выявление лиц, подверженных высокому риску данного заболевания, для обеспечения более эффективного скрининга или профилактической терапии [Отева *и др.,* 1994; Khera *et al.*, 2018]. Хотя большая часть риска заболевания носит полигенный характер, до сих пор не было возможности использовать полигенные предикторы для выявления лиц, подверженных риску, сопоставимому с моногенными мутациями. Поэтому разработка полигенных оценок риска является важнейшей задачей. Так было показано более чем трехкратное повышение риска развития ишемической болезни сердца (ИБС), фибрилляции предсердий, диабета 2 типа, воспалительных заболеваний кишечника и рака молочной железы у 8.0%, 6.1%, 3.5%, 3.2% и 1,5% населения соответственно. При этом для ИБС эта распространенность в 20 раз превышает частоту носителей редких моногенных мутаций, дающих сопоставимый риск [Khera *et al.*, 2018]. Для различных распространенных заболеваний были идентифицированы гены, в которых редкие варианты у гетерозиготных носителей повышают риск в несколько раз. Важным примером является наличие вариантов риска семейной гиперхолестеринемии у 0,4% населения, что повышает риск развития ишемической болезни сердца (ИБС) в 3 раза [Abul-Husn *et al.*, 2016]. Интенсивное лечение для снижения уровня циркулирующего холестерина среди таких носителей может значительно снизить риск. Другим примером является

миссенс-мутация p.E508K в гене *HNF1A*, с частотой носителей 0,1% среди населения земли в целом и 0,7% среди латиноамериканцев [Lek *et al.,* 2016], что повышает риск развития диабета 2 типа в 5 раз [Estrada *et al.,* 2014]. Хотя выявление моногенных мутаций может иметь большое значение для носителей и их семей, подавляющее большинство заболеваний возникает у тех, у кого нет таких мутаций.

Таким образом, для большинства распространенных заболеваний полигенное наследование, включающее множество распространенных генетических вариантов с небольшим эффектом, играет большую роль, чем редкие моногенные мутации [Gibson *et al.,* 2012]. Однако было неясно, возможно ли создать общегеномную полигенную оценку (GPS) для выявления лиц с клинически значительно повышенным риском, например, сопоставимым с уровнями, присущими редким моногенным мутациям [Khera *et al.,* 2018].

Предыдущие исследования по созданию GPS имели лишь ограниченный успех, обеспечивая недостаточную стратификацию риска для клинической пользы (например, выявление 20% населения с повышенным в 1,4 раза риском по сравнению с остальной частью населения). Этим первоначальным усилиям препятствовали три проблемы: (i) небольшой размер первоначальных исследований общегеномных ассоциаций (GWAS), которые повлияли на точность оценки влияния отдельных вариантов на риск заболевания; (ii) ограниченные вычислительные методы для создания GPS; и (iii) отсутствие больших наборов данных, необходимых для проверки и тестирования GPS. Используя гораздо более масштабные исследования и улучшенные алгоритмы, Khera с коллегами решили вернуться к вопросу о том, может ли GPS идентифицировать подгруппы населения с риском, приближающимся или превышающим риск моногенной мутации. Они изучили пять распространенных заболеваний, оказывающих серьезное влияние на общественное здравоохранение: ИБС, фибрилляцию предсердий, диабет 2 типа, воспалительные заболевания кишечника и рак молочной железы. Для каждого из

заболеваний созданы несколько потенциальных GPS на основе сводной статистики недавних крупных GWAS у участников преимущественно европейского происхождения. Данные были проверены на массиве данных британского Биобанка. Предсказатели имели AUC в диапазоне от 0,79 до 0,81, причем лучший предсказатель CAD - coronaryarterydisease - (GPSCAD) включал 6630150 вариантов. Этот предиктор показал эквивалентные результаты в тестовом наборе данных с AUC 0,81 (табл. 23).

Таблица 23. Определение и тестирование GPS для пяти распространенных сложных заболеваний [Khera *et al.,* 2018].

Заболевание	**Объем GWAS, Случай/Контроль**	**Распространенность в наборе данных для проверки**	**Распространенность в тестируемом наборе данных**	**Количество SNP в GPS**	**Параметр настройки**	**AUC (95% CI) в наборе данных для проверки**	**AUC (95% CI) в тестируемом наборе данных**
ИБС (CAD)	60,801/123,504	3,963/120,280 (3.4%)	8,676/288,978(3.0%)	6,630,150	LD Pred(ρ = 0.001)	0.81 (0.80–0.81)	0.81 (0.81–0.81)
Фибрилляция предсердий	17,931/115,142	2,024/120,280 (1.7%)	4,576/288,978 (1.6%)	6,730,541	LDPred (ρ = 0.003)	0.77 (0.76–0.78)	0.77 (0.76–0.77)
СД2	26,676 / 132,532	2,785/120,280 (2.4%)	5,853/288,978 (2.0%)	6,917,436	LDPred (ρ = 0.01)	0.72 (0.72–0.73)	0.73 (0.72–0.73)
Воспалительное заболевание кишечника	12,882 / 21,770	1,360/120,280 (1.1%)	3,102/288,978 (1.1%)	6,907,112	LDPred (ρ = 0.1)	0.63 (0.62–0.65)	0.63 (0.62–0.64)
Рак молочной железы	122,977/105,974	2,576/63,347 (4.1%)	6,586/1576895 (4.2%)	5,218	Обрезка и пороговая обработка (r/2 < 0.2; P < 5x10-4	0.68 (0.67–0.69)	0.69 (0.68–0.69)

Преимущество GPSCAD состоит в том, что его можно оценить с момента рождения, задолго до того, как появится способность различать факторы риска (например, гипертонию или диабет 2 типа), используемые в клинической практике для прогнозирования ИБС. Более того, даже для исследуемой популяции среднего возраста практикующие клиницисты не смогли идентифицировать 8% лиц с ≥3-кратным риском на основе GPSCAD в отсутствие информации о генотипе [Khera *et al.,* 2018]. Например, обычные

факторы риска, такие как гиперхолестеринемия, присутствовали у 20% пациентов с ≥3-кратным риском, основанным на GPSCAD, по сравнению с 13% в остальной части распределения, гипертония - у 32% против 28%, и семейный анамнез сердечных заболеваний - у 44% против 35%. Информирование людей с высоким уровнем GPSCAD об их наследственной восприимчивости может способствовать интенсивным профилактическим мероприятиям. Например, ранее показано, что высокий полигенный риск ИБС может быть компенсирован одним из двух вмешательств: приверженностью здоровому образу жизни или терапией статинами, снижающей уровень холестерина [Khera *et al.,* 2018]. Сходные результаты были получены и по четырем другим заболеваниям.

Диабет 2 типа является ключевой причиной сердечно-сосудистых и почечных заболеваний, распространенность которых во всем мире быстро растет. Полигенный предиктор выявил 3,5% населения с риском ≥3 раз, а у 1% риск был 3,30-кратным [Khera *et al.,* 2018]. Доказано, что как лекарства, так и интенсивное вмешательство в образ жизни (в том числе хирургическое) предотвращают прогрессирование диабета 2 типа [Knowler *et al.,* 2002], но широкое их внедрение было ограничено побочными эффектами и стоимостью. Выявление лиц с высоким уровнем генетическим предсказателем диабета 2 типа (GPST2D) может предоставить возможность с большей точностью определять цели таких вмешательств.

Во многих областях медицины есть, по крайней мере, один геномный тест с сильным эффектом, который является частью стандарта медицинской помощи. Эффект сильно варьирует в зависимости от заболевания. При этом исследователи заметно продвинулись в этом отношении в онкологии, кардиологии, эндокринологии и пренатальном/неонатальном тестировании (акушерство и педиатрия) [Баранов *и др.* 2021]. Результаты таких тестов служат клиническим руководством в том смысле, что они позволяют идентифицировать/диагностировать более однородную подгруппу в рамках более крупной группы заболеваний. В качестве альтернативы они могут

идентифицировать подгруппу пациентов с различными терапевтическими потребностями и предлагать лекарства, которые могут быть эффективны и не иметь побочных эффектов. Некоторые из этих геномных тестов имеют убедительные подтверждающие доказательства, но еще не являются стандартом медицинской помощи часто потому, что процесс клинического внедрения еще не определен (включая инфраструктуру, образование для оказания медицинской помощи). Однако, даже когда эффективность геномной медицины будет доказана, необходимо будет оценить экономическую эффективность, безопасность, переносимость, доступность и приемлемость по сравнению с существующими лекарственными средствами для соответствующего клинического вопроса [Franks *et al.,* 2021]. Более того, поскольку подавляющее большинство исследований генетики человека проводилось у людей европейского происхождения, в будущем следует уделять приоритетное внимание изучению других этнических групп, особенно там, где открытие редких вариантов представляет интерес [Barbitoff *et al.,* 2021].

В странах, где доступны эффективные стратегии профилактики или раннего выявления, ключевые вопросы будут включать распределение внимания и ресурсов между лицами с различным уровнем генетического риска и интеграцию стратификации генетического риска с другими факторами риска, включая редкие моногенные мутации, клинические и экологические факторы. Там, где такие стратегии не существуют или являются неоптимальными, выявление лиц с высоким риском должно способствовать разработке эффективных исследований для выявления ранних маркеров начала заболевания и клинических испытаний для проверки стратегий профилактики. В обоих случаях важно признать, что риск, связанный с высоким показателем полигенности, может отражать не один основной механизм, а скорее совокупное влияние нескольких путей [Fry *et al.,* 2017]. Тем не менее, стратегии профилактики и своевременного выявления могут быть полезны, независимо от основного механизма — как в случае терапии статинами ИБС, назначения

антикоагулянтов для предотвращения инсульта у пациентов с фибрилляцией предсердий, или усиленного маммографического скрининга на рак молочной железы. Информирование о рисках потребует серьезного рассмотрения, в то время как показатели полигенного риска могут быть одновременно рассчитаны при рождении для всех распространенных заболеваний. При этом, полезность знаний и потенциальный вред от генетических тестов для человека могут варьировать в зависимости от заболевания и стадии жизни [Khera *et al.*, 2018]. Тем не менее, может оказаться целесообразным скрывать некоторую информацию, которая может быть легко вычислена на основе генетических данных.

Наконец, мы выделяем важнейший вопрос справедливости. Оценки полигенного риска, описанные здесь, были получены и протестированы у лиц преимущественно европейского происхождения - группы, в которой на сегодняшний день было проведено большинство генетических исследований. Поскольку частоты аллелей, закономерности неравновесия связей и размеры эффектов общих полиморфизмов варьируются в зависимости от происхождения, конкретный GPS здесь не будет обладать оптимальной прогностической способностью для других этнических групп [Khera *et al.*, 2018]. Важно, чтобы для биомедицинское сообщество обеспечило доступ всех этнических групп к прогнозированию генетического риска сопоставимого качества, что потребует проведения или расширения GWAS в неевропейских этнических группах.

Результаты показывают, что полигенные оценки действительно предсказывают риск сложных заболеваний. Точность предсказаний рисков значительно улучшится с учетом данных экзома и всего генома. Нет сомнения, что уже в течение нескольких лет комбинация данных секвенирования генома и общегеномного поиска аллельных ассоциаций для всех основных МФЗ позволит резко увеличить предсказательную (предиктивную) ценность досимптоматического генетического тестирования наследственной предрасположенности, и можно будет надеяться на быстрый прогресс в этой

области молекулярной медицины [Баранов *и др.,* 2021]. Считается, что есть веские основания для того, чтобы сделать недорогое генотипирование стандартом медицинской помощи в системы здравоохранения по всему миру [Lello*etal.,* 2019].

В заключение хотелось бы отметить, что в зависимости от этиологических факторов все наследственные болезни достаточно условно подразделяются на моногенные и мультифакторные. Тяжесть заболевания, время манифестации и клинические признаки наследственных болезней зависят от природы мутаций, повреждающих работу гена и генов-модификаторов, влияющих на проявление патологических признаков и факторов внешней среды. На долю типичных моногенных болезней приходится только около 1 % всей патологии человека. Семейные моногенные и полигенные формы достаточно редки, основные заболевания человека (сердечно-сосудистые, онкологические, психические, нейродегенеративные и др.) относятся к МФЗ, этиология которых складывается из сочетанного действия наследственной предрасположенности и внешней среды. Молекулярное тестирование генов этих болезней составляет методическую основу превентивной медицины. На анализе паттернов аллельного полиморфизма генов частых МФЗ базируется предиктивная (предсказательная) медицина.

Современные представления о генетическом варианте концептуально меняют наши знания о том, какие болезни являются моногенными, олигогенными и мультифакториальными. Назрела необходимость создания новой генетической классификации болезней с опорой не только на клинико-лабораторные данные, но и на результаты молекулярно-генетического анализа с учетом пенетрантности и экспрессивности. Требуются поиск оптимальных статистических подходов к оценке генетической предрасположенности и разработка новых полигенных факторов риска МФЗ.

ГЛАВА III. Секвенирование нового поколения и инфекционные болезни человека. Генетические факторы риска развития коронавирусной инфекции COVID-19

3.1. Общие сведения о SARS-CoV-2 и его геномной изменчивости

В предыдущих главах нашего исследования был установлен значительный вклад методов NGS в изучение моногенных и мультифакториальных заболеваний человека в рамках современной концепции генетического паспорта здоровья человека.

Сегодня важнейшим фактором социального стресса, возникшим в конце 2019 г и влияющим на здоровье человека всего мирового сообщества, включая РФ, стала новая коронавирусная инфекция [https://coronavirus.jhu.edu; Glotov *et al.,* 2021a]. На август 2022 года пандемия новой коронавирусной инфекции COVID-19 распространилась на 230 стран, во всем мире число с более случаев заражения составило более чем 589 миллионов, число летальных исходов - 6,4 миллионов [https://coronavirus.jhu.edu].

В разгар COVID-19 инфекции крайне актуальным для борьбы с данной пандемией было выявление новых белковых и генных мишеней, которые могут оказаться высокочувствительными диагностическими и прогностическими маркерами тяжести и исхода заболевания. Для этого необходимо изучение как генома вируса SARS-CoV-2, так и генома самого пациента.

Общеизвестно, что коронавирусы (CoV) являются возбудителями острого тяжелого респираторного синдрома (SARS-CoV), впервые вызвавшего вспышку глобальной эпидемии в 2002 г, когда было инфицировано 8000 человек, 10% из которых имели летальный исход [Freund *et al.,* 2015]. Позднее, в 2012 г произошла вспышка ближневосточного респираторного синдрома (MERS-CoV), когда эпидемия охватила 26 стран [Mackay, Arden, 2015]. К сожалению, эпидемиология и генетика данных вирусов была недостаточно изучена. Поэтому, когда в Китае в декабре 2019 г. была зафиксирована новая форма

коронавируса (SARS-CoV-2), ставшая причиной глобальной пандемии COVID-19 в 2020 г. [Liu *et al.*,2020], оставалось много «белых пятен» в понимании того, как может проходить заболевание, каковы основные факторы риска, методы профилактики инфекции и т.д.

Коронавирусы представляют собой крупные сферические частицы (диаметр 120 нм), состоящие из двухслойной липидной оболочки, включающей 4 белка: мембранный (M, E), шипиковый (S) и гемагглютининовую эстеразу (HE) вокруг нуклеокапсида (N), образованного множеством копий этого белка, связанных с одноцепочечной РНК [Cascella *et al.*, 2020] (рис.25).

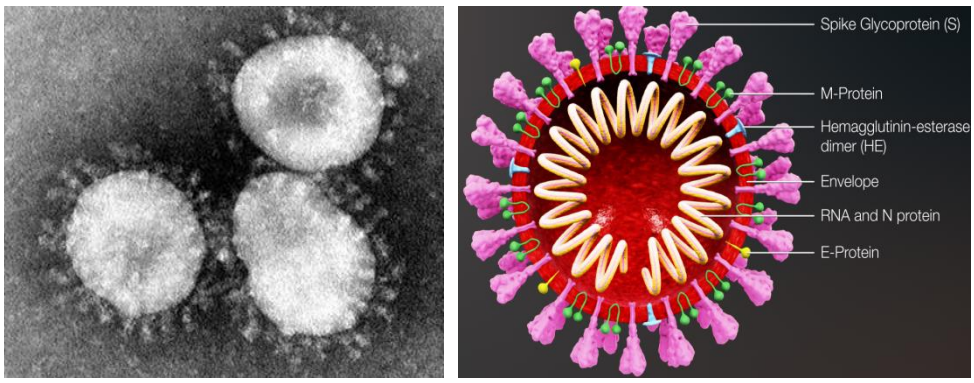


Рисунок 25. Микрофотография (А) и схематическая структура коронавируса SARS-CoV-2 (Б). Spikeglycoprotein (S) – S белок (шипиковый белок); Mprotein – мембранный М-белок; Hemagglutinin-esterasedimer (HE) – димер гемаглютинина и эстеразы; Envelope – оболочка; RNAandNprotein – РНК и N-белок нуклеокапсида; Eprotein – Е-белок оболочки [Cascella *et al.*, 2020].

S-белки образуют выросты на оболочке вируса, создавая вид «короны», благодаря которой вирус и получил свое название [Ashour *et al.*, 2019]. С помощью этих шипов вирусы прикрепляются к белкам-рецепторам клеток – хозяина, которые обеспечивают слияние вирусной и клеточной мембран, а также проникновение вирусной РНК в клетку. S-белки содержат рецептор-связывающий домен (RBD, аминокислоты N318-T509), обеспечивающий взаимодействие с рецептор-связывающим мотивом (RBM, аминокислоты S432-T486) ангиотензин-превращающего фермента 2 (ACE2) – клеточным рецептором для SARS-CoV-2 [Li *et al.*, 2005; Chen *et al.*, 2020]. Кроме того, S-гликопротеин

содержит фурин-подобный сайт рестрикции [Coutard *et al.*, 2020], который необходим для распознавания S-гликопротеина при пиролизе и, следовательно, способствует зоонозной инфекции вируса. Геном SARS-CoV-2 представлен одноцепочечной РНК длиной около 30 тыс. п.н., содержащей на 5'-конце cap-область и на 3'-конце poly-A-последовательность, которые позволяют вирусной РНК транслироваться на рибосомах клеток-хозяина. Вирусная РНК включает регуляторные последовательности, в которых происходит терминация транскрипции, и 10 открытых рамок считывания (ORF), которые транскрибируются с образованием мРНК (рис. 26).

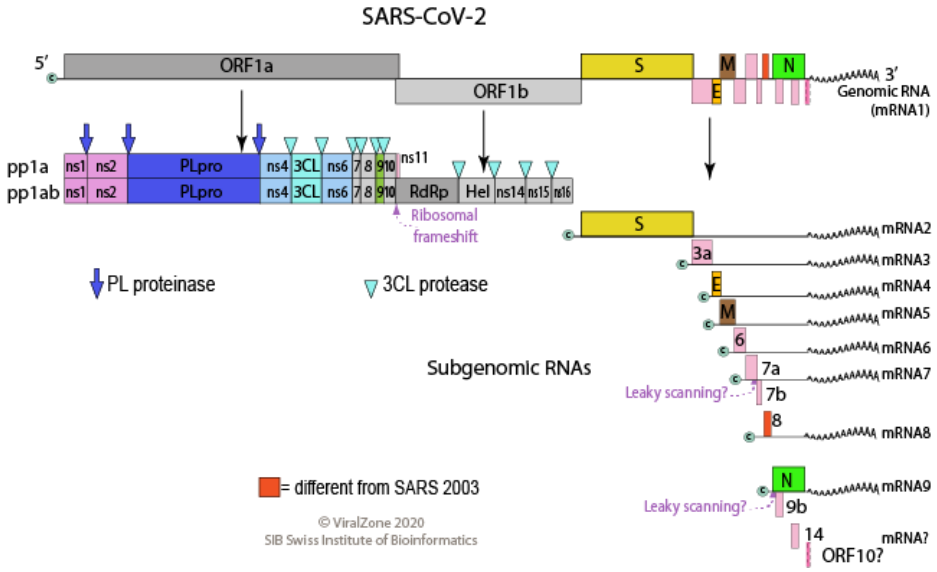


Рисунок 26. Структура генов РНК коронавируса SARS-CoV-2 [Schidtke, 2020].

Высокая скорость распространения и тяжесть заболевания, с одной стороны, обусловлены высокой скоростью мутирования вируса, а с другой– генетической предрасположенностью человека к заболеванию (наличие аллелей риска). Мутации генов вирусной РНК-полимеразы (ORF8,241C>T, S84L, 14408C>T, C251T), РНК-праймазы (P323L), S-белка (23403A>G, D614G), а также структурных белков (N, E) повышают иммуногенность белков в механизмах Т-клеточного иммунного ответа, что может быть ассоциировано с повышенной трансмиссивностью и тяжестью COVID-19 инфекции в странах Европы [Yin *et al.*, 2020]. Наличие или отсутствие двух вариантов (8782C>T) в ORFL1ab и

28144 (28144T>C) в ORF8 вирусного генома позволило кластеризировать два основных генетических (L или S) типа COVID-19, различающихся по вирулентности, скорости репликации и тяжести заболевания [Koyama *et al.,* 2020]. Мутации в генах ключевых вирусных белков могут вести к изменениям аффинности и специфичности таргетных препаратов, являясь молекулярной основой различий заболеваемости и смертности, а также реакции организма на противовирусные препараты или вакцины.

Многочисленные исследования по секвенированию образцов РНК SARS-CoV-2 установили различные генетические подтипы коронавируса. На основании анализа SNP в образцах из эндемического региона Чанчунь, Yin и соавт. определили четыре генотипа SARS-CoV-2: генотип I (11083G>T), генотип II (26144G> T), генотип III (8782C> T, 28144T>C), генотип IV (241C>T, 3037C>T, 14408C>T, 23403A>G) [Yin *et al.,* 2020]. Позже в октябре 2020 года, в образцах от пациентов из штата Рио-де-Жанейро (Бразилия) были обнаружены мутации S-белка, из которых мутация E484K оказалась наиболее клинически значимой, вследствие чего данный вариант и получил свое название 484K.V2 [Toovey *et al.,* 2021]. В декабре 2020 г. были зарегистрированы варианты B.1.1.7 (альфа) и B.1.351 (501Y.V2, бета) коронавируса из Великобритании и Южной Африки, содержащие соответственно нуклеотидные замены 69-70del, Y144del и N501Y, K417N, E484K в S-белке [World Health Organization.SARS-CoV-2 Variants https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/].При анализе последовательностей 405871 образцов из базы GISAID в конце 2020 г. были выявлены варианты B.1.427 и B.1.429 коронавируса из Южной и Северной Калифорнии [Zhang W. *et al.,* 2021]. В январе 2021г. были обнаружены P.1 и P.2 (подлиния B.1.128, гамма) варианты, включающие 10 уникальных мутаций (E484K и N501K и др.) в S-белке в 42% протестированных образцах из Манауса, Бразильского штата Амазонас [Sabino *et al.,* 2021]. Вариант дельта (B 1.617.2), содержащий мутации G478K и 417N в S-белке, появился в октябре 2020 г. в Индии. В начале 2021 г. в Нью-Йорке появился вариант под названием B.1.526,

содержащий мутации E484K и S477N в S-белке [Bernal *et al.,* 2021]. Но вскоре был обнаружен еще более новый вариант SARS-CoV-2, который ВОЗ назвала как Mu. Он был впервые зафиксирован в Колумбии в январе 2021 г. и к середине года он был уже обнаружен в 39 странах [https://www.weforum.org/agenda]. В сентябре 2021 г. в Великобритании обнаружили мутантную форму дельта варианта (AY.4.2 дельта плюс), когда в течение недели (27 сентября - 3 октября) каждый день регистрировалось по 40-50 тыс. новых случаев инфицирования. Считается, что дельта-плюс на 10-15% более контагиозен, чем вариант дельта [https://www.weforum.org/agenda]. В декабре 2020 г. появились данные о распространении линий B.1.1.31 и B.1.1.317 SARS-CoV-2, имеющих российское происхождение, в Великобритании, США, Японии, Сингапуре, Турции Таиланде, Швейцарии и Бразилии [PANGO lineages. https://cov-lineages.org/lineage_designation.html].

3.2. Ассоциация генных вариантов вируса SARS-CoV-2 с тяжестью и исходами коронавирусной инфекции

Большое значение для понимания тактики лечения имеет исследование особенностей генетических разновидностей вируса, их корреляции с клиническими симптомами и степенью тяжести заболевания. Для решения этой задачи мы изучили образцы вирусной РНК, выделенные от 56 пациентов с COVID-19 инфекцией, находившихся на лечении в ГБУЗ «Городская больница №40» Санкт-Петербурга в период с 18.04.2020 по 31.01.2021 г., имевших положительный результат теста на наличие РНК SARS-CoV-2 методом амплификации нуклеиновых кислот в полимеразной цепной реакции (ПЦР) [Glotov *et al.,* 2021b]. Нами было выявлено всего 389 вариантов в РНК SARS-CoV-2, из них 263 (67,6%) были связаны с аминокислотной заменой в белках, 126 синонимичных мутаций (32,4%), причем 139 (35,7%) − в образцах пациентов с летальным исходом. Количество несинонимичных, молчащих мутаций и

мутаций в Sбелке SARS-CoV-2 со степенью тяжести коронавирусного заболевания представлено в таблице 24.

Показано, что количество мутаций S-белка в образцах от пациентов с заболеванием 3-й степени тяжести было достоверно выше (p<0,001), чем число мутаций вируса в образцах пациентов с заболеванием 1-й степени тяжести [Glotov *et al.,* 2021b]. Подобного рода исследования ранее были проведены японскими исследователями. Они обнаружили, что количество несинонимичных мутаций (особенно Pro108Ser в 3 химотрипсиноподобной протеазе (3CLpro) и Pro151Leu в N-белке нуклеокапсида) обратно коррелирует (OR=0,24, 95%CI=0,07-0,88, P=0,032) со степенью тяжести COVID-19 инфекции и потребностью в кислородной терапии [Abe *et al.,* 2021].

Таблица 24. Количество мутаций в геноме SARS-CoV-2 в зависимости от степени тяжести заболевания [Glotov *et al.*, 2021b].

Степень тяжести COVID-19	Среднее количество мутаций	Среднее количество несинонимичных мутаций	Среднее количество молчащих мутаций	Среднее количество мутаций в S-белке
1 (n=4)	11,0±0,2	7,0±1,4	4,0±1,4	1,0
2 (n=14)	11,4±1,9	8,3±1,9	3,3±1,6	2,0±1,0
3 (n=13)	13,0±3,6	9,0±2,4	4,0±2,1	2,2±1,0***
Умершие, 3 / dead, 3 (n=25)	17,1±5,0	10,6±3,0	6,6±2,2	1,6±0,9

Примечание: *** статистическая значимость при p<0,001. Легкая степень тяжести обозначена I, среднетяжелая степень – II, тяжелая степень – III, все умершие пациенты имели тяжелую III степень тяжести COVID-19 инфекции.

Для выявленных мутаций устанавливали их потенциальную опасность, т.е. оценивали ассоциации с степенью тяжести заболевания, рассчитывая частоты и OR в группах выписанных и умерших пациентов с применением кластерного анализа. Сравнительный анализ частот для каждого SNP в группах образцов пациентов с риском летального исхода позволил сгруппировать SNP в 3 кластера (группы). В первую были группу включены мутации, потенциально

ассоциированные с тяжестью коронавирусной инфекции и летальным исходом (OR>1). Они локализованы в генах, кодирующих лидерную 5'-концевую cap-область и белки Nsp1, Nsp2, Nsp3, Nsp4, РНК-зависимую РНК-полимеразу (Nsp12), геликазу (Nsp13), S–белок, Orf3a и 3'-концевую poly-A-последовательность. Ко второй группе относились нейтральные мутации (OR=1), локализованные в генах, кодирующих Nsp3 и N-белки. Третий кластер образовали потенциально протекторные мутации (OR<1), локализованные в генах, кодирующих белки: Nsp2, Nsp3, Nsp7, эндонуклеазу (Nsp14), эндоРНКазу (Nsp15), S- и N-белки [Glotov *et al.*, 2021b].

Следующей задачей была оценка эпидемиологических характеристик вируса на основе анализа генома SARS-CoV-2 в популяции наших пациентов. Нами было установлено, что пациенты из Санкт-Петербурга инфицированы 14 линиями SARS-CoV-2. Из 47 образцов 14 (29,8%) были инфицированы 5 линиями вируса российского происхождения. В целом, OR для 5 линий вируса, имеющих российское происхождение, составил 0,441 (95%CI=0,116-1,684). Это свидетельствует о том, что мутации в линиях SARS-CoV-2 нероссийского происхождения ассоциированы с повышенным риском летального исхода (OR=2,267, 95% CI=0,1594-8,653) в обследованной группе российских пациентов [Glotov *et al.*, 2021b].

Представляет интерес тот факт, что 75,8% образцов наших пациентов были инфицированы вирусом IV генотипа (241C>T, Nsp3: 3037C>T; Nsp12, РНК полимераза: 14408C>T, Pro323Leu; S-белок: 23403A>G, Asp614Gly) SARS-CoV-2, который был идентифицирован Zhang с соавторами [Zhang W *et al.*, 2021]. Гаплотип 3037C>T, 14408C>T, 23403A>G и SNP 241C>T наиболее часто встречаются в образцах из Европы, что свидетельствует о европейском происхождении вируса у наших пациентов. В исследовании, проведенном международным коллективом авторов из Вьетнама, Австралии, Великобритании и США, данный мутантный гаплотип обнаружен в 90% (40 из 44) образцов пациентов из Великобритании (n=15), России (n=6), Германии (n=5), Франции

(n=4), Италии (n=2), Испании (n=2), Нидерландов (n=1), Вьетнама (n=6) и стран Азии (n=3). Вместе с этой группой мутаций у данных пациентов в 75% геномов (33 из 44) также наблюдались SNP в N-белке 28881G>A, 28882G>A, 28883G>C [Wang *et al.,* 2021]. Данные мутации локализуются в небольшом участке (194-204) аминокислотной последовательности данного белка. Поскольку N-белок участвует в упаковке вирусной РНК в спиральный рибонуклеокапсид, мутации 28881G>A, 28882G>A, 28883G>C могут повышать эффективность транскрипции субгеномных РНК, способствуя выживанию клеток, репликации и персистенции вируса [Caccuri *et al.,* 2020]. В наших образцах SNP 28881G>A, 28882G>A, 28883G>C присутствовали в 100% образцов, что, скорее всего, указывает на нейтральный характер этих вариантов и на происхождение вируса наших пациентов, относящегося ккладе 20B линии B1.1 [Abe *et al.,* 2021]. В 47 из 50 наших образцов (94%) также наблюдался SNP 241C>T, локализованный в 5'-нетранслируемой области (5'UTR) вирусной РНК. Эта замена имеет одну из самых высоких частот: 0,758 (9673/12754) и 0,809 (36786/45494) соответственно в США и мире, влияя на регуляцию транскрипции и экспрессии генов вируса, а также его репликацию [Wang *et al.,* 2021]. SNP 14408C>T (Pro323Leu) и мутация 13554C>T, ассоциированные с тяжелым течением и летальным исходом, также обнаружены в образцах вируса от наших пациентов в РНК-зависимой РНК-полимеразе (белок RdRp или Nsp12) с частотами 0,940 (47 из 50) и 0,06 (3 из 50). Первая из этих мутаций также является наиболее распространённой в образцах из США (AF 0,464, 5918/12754) [Wang *et al.*, 2021]. О биологической роли этой мутации до сих пор ведется дискуссия. Wang R. с соавторами считают, что поскольку обе аминокислоты - пролин (Pro) и лейцин (Leu) - являются неполярными и алифатическими, то, в связи с этим Pro323Leu может не оказывать влияния на функцию белка Nsp12 [Wang *et al.*, 2021]. В 94% образцов наших пациентов (47/50) встречается SNP 23403A>G (Asp614Gly) в гене S-белка, связанный с заменой аспарагиновой кислоты в положении 614 на глицин и ассоциированный с повышенным риском тяжелого течения или летального

исхода пациентов. Как уже упоминалось, этот SNP часто сочетается с тремя другими SNP (241C>T, 3037C>T и 14408C>T) и наследуется как гаплотип [Korber *et al.,* 2020]. Исследователи из Гарвардского, Вашингтонского (США) и Шеффилдского университетов (Великобритания), изучив клинические данные и генетические последовательности SARS-CoV-2 у 999 пациентов с COVID-19, установили, что вариант Arg614Gly коррелирует с высоким уровнем вирусной РНК в верхних дыхательных путях и скоростью передачи вируса у пациентов по сравнению с диким типом. Это указывает на более высокую инфицированность Gly614. Однако Korber с коллегами не смогли установить ассоциации между D614G и длительностью госпитализации, т.е. степенью тяжести заболевания [Korber *et al.*, 2020].

Нами также установлены корреляционные зависимости между общим количеством несинонимичных мутаций (SNP) в S-белке линий SARS-CoV-2 с одышкой, риском летального исхода, уровнями ферритина, Д-димера и глюкозы в крови [Glotov *et al.*, 2021b].

3.3. Маркеры тяжести течения COVID-19

Известно, что при инфицировании SARS-CoV-2 у пациента по истечении 2-14 суток инкубационного периода характерны симптомы острого респираторного вирусного заболевания: повышенной температуры тела (90%); кашля (в 80% случаев), одышки (в 30% случаев), утомляемости (в 40% случаев), ощущения заложенности в грудной клетке (в 20%), боли в горле, насморка, снижения обоняния и вкуса, конъюнктивита. Данные симптомы могут свидетельствовать о развитии пневмонии без дыхательной недостаточности, острого респираторного дистресс-синдрома – ОРДС (пневмонии с острой дыхательной недостаточностью), сепсиса, септического шока или полиорганной недостаточности. Данные симптомы были характерны для альфа- и дельта-штаммов, тогда как для штамма «омикрон» они несколько другие [Временные

методические рекомендации: профилактика, диагностика и лечение новой коронавирусной инфекции (COVID-19). Версия 14, 27.12.2021].

Развитие септического шока или полиорганной недостаточности часто приводят к летальности пациентов трудоспособного возраста (59,7±13,3 года) с рядом хронических заболеваний: артериальной гипертензией (23,7-30%), сахарным диабетом (16,2%), метаболическим синдромом, ишемической болезнью сердца (5,8%), хронической обструктивной болезнью легких (ХОБЛ), никотиновой зависимостью, воспалительными заболеваниями кишечника и онкологической патологией [Yang X. *et al.,* 2020; Fang *et al.,* 2020; Shitao *et al.,* 2020]. Кроме того, группы риска инфицирования COVID-19 могут составлять пациенты с некоторыми генетическими заболеваниями. Например, в исследовании, проведенном учеными из университетского медицинского центра Утрехта (Нидерланды), 180 из 395 пациентов с синдромом Дауна (45,6%), заболели тяжелой респираторно-синцитиальной вирусной инфекцией [Beatrijs *et al.,* 2007].

Клиническая картина у пациентов, входящих в группы риска, характеризуется развитием «синдрома взаимного отягощения», сопровождающегося прогрессирующей дыхательной и сердечной недостаточностью, что в конечном итоге утяжеляет их состояние и приводит к трудовым потерям, ранней инвалидности и высокой летальности. В связи с этим инфицированные COVID-19 пациенты с хроническими и генетическими заболеваниями особенно остро нуждаются в немедленной диагностике и последующей реабилитации.

Центральной патофизиологической проблемой - COVID-19 является иммунная дисфункция с выраженной неконтролируемой генерализованной системной воспалительной реакцией в виде усиленной продукции воспалительных цитокинов. Этот «цитокиновый шторм» (ЦШ) может проявляться в двух клинических формах: в виде вторичного гемофагоцитарного лимфогистиоцитоза (ГЛГ) и синдрома активации макрофагов (САМ) [Costela-

Ruiz *et al.,* 2020, Blanco-Melo *et al.,* 2020] – синдромов, ранее описанных при некоторых инфекциях (вирус Эпштейна–Барр, грипп) и системных аутоиммунных заболеваниях (системная красная волчанка, болезнь Стилла), а также при лечении цитостатическими и иммуносупрессивными лекарственными препаратами, после аллогенной трансплантации органов и тканей, в 3-4% случаев сепсиса. ЦШ, таким образом, имеет некоторое сходство с вторичным гемофагоцитарным синдромом (ГФС), проявляющимся в виде лихорадки, цитопении, гиперферритинемии, аномальных печеночных показателей, коагулопатии и поражения легких (в т.ч. ОРДС) [Caso *et al.,* 2020]. При всех этих состояниях цитокины IL-1β, IL-18, IFN-γ и IL-6 являются основными медиаторами избыточного воспалительного ответа иммунной системы.

Результаты недавних исследований показывают, что ЦШ, связанный с COVID-19, является уникальной формой подобной гипервоспалительной реакции, требующей дальнейшей углубленной клинической и лабораторной оценки, а также разработки критериев его диагностики [Zachariah *et al.,* 2020]. Высказано предположение, что критерии классификации САМ (MAS-2016) не применимы к пациентам с COVID-19, а критерии шкал оценки ГЛГ (HLH-2004 и HScore) нуждаются в адаптации [Leverenz, Tarrant, 2020, McGonagle *et al.,* 2020]. Ясно обозначилась актуальность разработки прогностической модели развития ЦШ у пациентов сдиагностированной COVID-19, определения рутинных и дополнительных маркеров для оценки риска её развития [Caricchio *et al.,* 2020; Moore, June, 2020].

Известен ряд лабораторных биомаркеров, уровень которых патологически изменяется при ЦШ, независимо от фактора, запускающего процесс. При этом пороговые величины и сочетания основных биомаркеров могут лечь в основу дифференциальной диагностики состояний и тяжести течения заболевания. Так, например, внезапное и быстро прогрессирующее клиническое ухудшение на поздних стадиях COVID-19 (7–10 дни) коррелирует с повышением уровней показателей острой фазы (СРБ, ферритин) [Huang *et al.,* 2020; Grasselli *et al.,*

2020], а также с клиническими и лабораторными показателями ЦШ [Chen *et al.*, 2020; McGonagle *et al.*, 2020; Wiersinga *et al.*, 2020]. Концентрация ферритина >500 мкг/л выявлена у 55,9% пациентов с нетяжелыми формами COVID-19 и у 81,7% – с тяжелыми (p<0,0001) [Li *et al.*, 2020]. Гиперферритинемия наблюдается при синдроме активации макрофагов и ОРДС [Giamarellos-Bourboulis *et al.*, 2020] и может идентифицировать пациентов с высоким риском развития тяжелой COVID-19-ассоциированной пневмонии [McGonagle *et al.*, 2020; Kivela, 2020].

Таким образом, повышенные уровни С-реактивного белка (СРБ), интерлейкина-6, D-димера в крови являются ключевыми для диагностики системного гипервоспаления, цитокинового шторма и тяжелых случаев COVID-19-ассоциированной пневмонии [McGonagle *et al.*, 2020]. Эти показатели, наряду с лимфопенией, являются ключевыми при поступлении пациента в госпиталь для своевременного принятия решений по - выявлению пациентов с неблагоприятным прогнозом [Zhou *et al.*, 2020; Li *et al.*, 2020].

Для поиска ключевых показателей ЦШ нами было проведено исследование, где пациенты были разделены на две группы, сравнимые по возрасту. Первую группу составили 100 (21,8%) пациентов с клинико-рентгенологическими особенностями, характеризующими стабильное течение заболевания средней степени тяжести; вторую — 358 (78,2%) человек с прогрессирующим среднетяжелым, тяжелым и крайне тяжелым течением болезни (табл. 24) [Shcherbak *et al.*, 2021].

Были установлены различия по шкале NEWS: в 1-й группе индекс NEWS при поступлении составил в среднем 2 балла, средняя продолжительность госпитализации — 11 дней; во 2-й группе индекс NEWS при поступлении составил в среднем 4 балла, к началу терапии антицитокиновым препаратом, антиковидной плазмой, гемосорбции — 5 баллов, средняя продолжительность госпитализации — 12 дней (рис.27) У пациентов 2-й группы с тяжелым и крайне тяжелым течением заболевания зарегистрирована самая высокая смертность от

осложнений (28,8% в группе, 22,5% во всей когорте). Такие пациенты исходно имели неблагоприятный прогноз заболевания в силу возраста, коморбидности, клинической тяжести по степени дыхательной недостаточности, величины индекса по шкале NEWS, распространенности и последующей негативной динамики изменений легочной ткани по данным КТ. Методом построения классификационных деревьев (Classification Trees) идентифицированы пороговые уровни факторов риска развития ЦШ.

Таблица 24. Характеристика тяжести течения COVID-19 в двух группах пациентов г. Санкт-Петербурга [Shcherbak *et al.*, 2021].

Показатель	**1 группа**		**2 группа**		**Всего**	**p**
	n	%	n	%		
Женщины	58	58,00	159	44,41	217	p=0,016
Мужчины	42	42,00	199	55,59	241	
Всего:	100	21,83	358	78,17	458	
Степень тяжести течения заболевания:						
Легкая	0	0,00	0	0,00	0	p<0,001
Средняя	100	100,00	153	42,74	253	
Тяжелая и крайне тяжелая	0	0,00	205	57,26	205	
Всего:	100	21,83	358	78,17	458	
Форма заболевания по КТ 1-4 при поступлении:						
КТ-1	57	57,00	82	22,91	139	
КТ-2	43	43,00	223	62,29	263	p<0,001
КТ-3	0	0,00	44	12,29	47	
КТ-4	0	0,00	9	2,51	9	
Всего:	100	21,83	358	78,17	458	
Исходы заболевания:						
Выжившие	100	100,00	255	71,23	355	p<0,001
Умершие	0	0,00	103	28,77	103	
Всего	100	21,83	358	78,17	458	

В дальнейшем мы выполнили комплексную оценку риска ЦШ с ранжированием показателей, которыми в соответствии с рангом прогностической значимости, полученным методом построения деревьев классификации, к началу терапии ЦШ оказались:

1) динамика индекса по шкале NEWS*;*

2) уровень IL-6 крови выше 23 пг/мл;

3) уровень СРБ крови равный или выше 50 мг/л;

4) абсолютное количество лимфоцитов меньше0,72×109/л;

5) положительный результат теста на РНК коронавируса (SARS-CoV-2);

6) возраст пациентов 40 лет и старше.

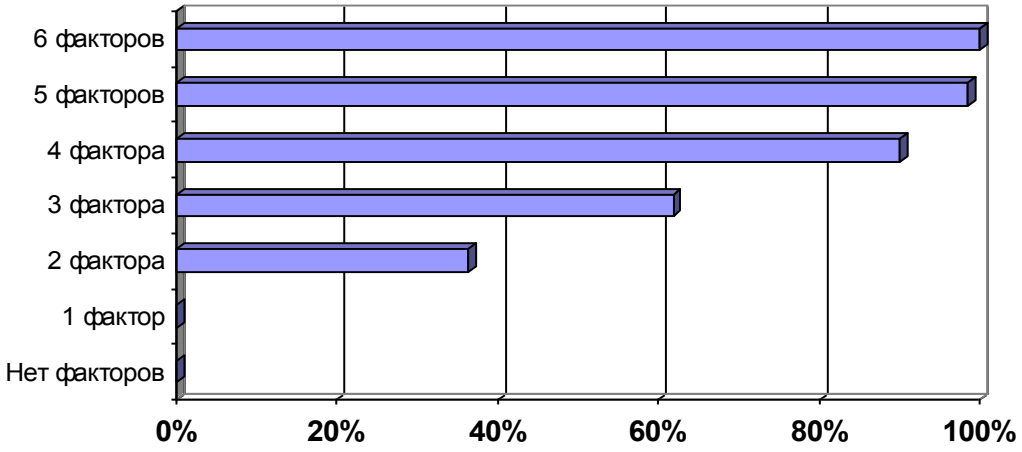


Рисунок 27. Частота случаев цитокинового шторма при различном количестве факторов риска COVID-19 в двух группах пациентов г. Санкт-Петербурга [Shcherbak *et al.*, 2021].

Выявленные показатели могут быть использованы в качестве критериев для оценки риска ЦШ. Необходимо отметить, что гендерные различия несущественны в дальнейшей комплексной оценке риска развития ЦШ. Для практического применения нашей прогностической модели намивыделены следующие категории риска:

- 1-я категория (0–1 фактор): риск ЦШ практически отсутствует;
- 2-я категория (2–3 фактора): риск ЦШ резко возрастает до 55%, увеличивается в 35,5 раз по сравнению с 1-й категорией;
- 3-я категория (4 и более факторов): риск ЦШ достигает 96%, увеличивается в 718 раз по сравнению с 1-й категорией.

Полученные результаты согласуются с оценкой факторов риска ЦШ при COVID-19 других авторов [Caricchio *et al.*, 2021; Moore el al., 2020] и позволяют

обосновать выбор лечебной тактики с ранним назначением упреждающей противовоспалительной терапии и антиковидной плазмы реконвалесцентов для пациентов с высоким риском развития ЦШ, что было подтверждено на практике работы. К основным факторам риска развития цитокинового шторма у больных COVID-19 относятся мужской пол, возраст старше 40 лет, положительный тест на РНК SARS-CoV-2, лимфопения, уровни лактатдегидрогеназы (ЛДГ), D-димера, ферритина и IL-6, динамика индекса по шкале NEWS (рис.28). Лабораторными критериями для диагностики и динамического контроля за течением цитокинового шторма являются абсолютное количество лимфоцитов, уровни ЛДГ, СРБ, ферритина, D-димера и IL-6.

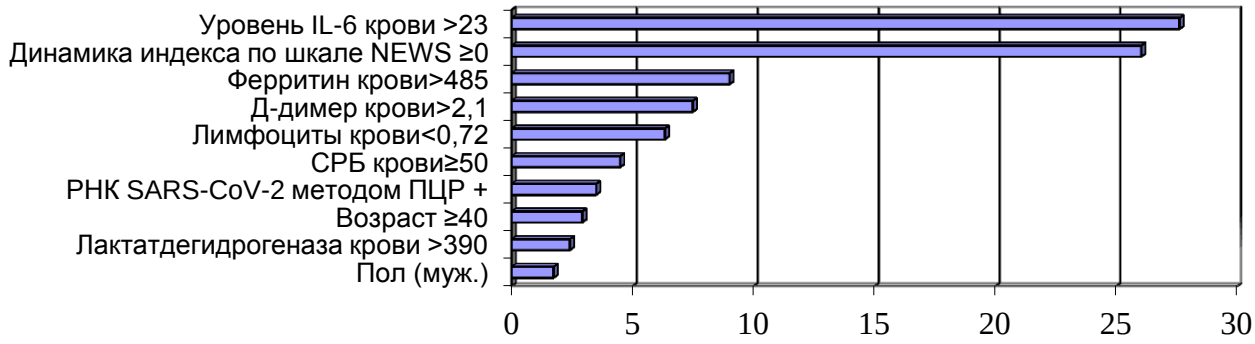


Рисунок 28. Увеличение риска цитокинового шторма (OR) при неблагоприятных значениях показателей [Shcherbak *et al.*, 2021].

3.4. Поиск генных вариантов, предрасполагающих к тяжелому течению COVID-19: роль ACE-2

Информация о лабораторных критериях риска коронавирусной инфекции позволяет корректно сформировать группы риска для проведения дальнейших генетических исследований [Glotov *et al.,* 2021a; Shcherbak *et al.*,2022].

Любые исследования генетической предрасположенности к вирусным инфекциям начинаются с изучения рецепторного аппарата, необходимого вирусу для проникновения в клетку хозяина. Установлено, что ангиотензин-превращающий фермент-2 (ACE2) является клеточным рецептором SARS-CoV-2. Вирус SARS-CoV2, используя рецепторсвязывающий домен (RBD) белка

Spike на поверхности частиц, присоединяется к рецепторсвязывающему мотиву (RBM) ACE2 и проникает в клетку [Yan R. *et al.*, 2020]. В организме человека этот белок кодируется геном *ACE2*. Полиморфизм гена *ACE2* может влиять на аффинность и специфичность связывания S-белка с ACE2, а, следовательно, определять наследственную предрасположенность к риску инфицирования и летальности от SARS-CoV-2. Показано, что восприимчивость человека к COVID-19 может быть результатом сочетанного влияния терапии и особенностей полиморфизма гена *ACE2* [Fang L. *et al.*, 2020]. Установлены корреляции между наиболее часто встречающимися (38) вариантами в гене*ACE2* с уровнем экспрессии белка в различных тканях человека [Chen J. *et al.*, 2020]. Показано, что возраст, пол, раса и курение достоверно (P=0,008) влияют на экспрессию гена *ACE2* [Chen J. *et al.*, 2020]. Таким образом очевидно, что мутации в гене *ACE2* могут нарушать взаимодействия рецептор-лиганд. Мутации в гене *ACE2*, устойчивые к связыванию с S-белком, отсутствуют в разных популяциях, однако предполагается, что редкие миссенс-варианты могут влиять на тяжесть инфекции [Cao *et al.*, 2020]. Если миссенс-вариант влияет на рецепторную функцию ACE2, то более высокая частота носителей миссенс-аллелей в некоторых популяциях может приводить к снижению уровня инфицирования и смертности. А более низкая частота носителей миссенс-вариантов, в свою очередь, может повысить восприимчивость и тяжесть течения COVD-19. Действительно, в недавнем итальянском исследовании [Benetti *et al.*, 2020] были обнаружены, казалось бы, защитные миссенс-варианты в гене *ACE2*. Хотя распространенные миссенс-варианты в гене *ACE2* могут и не влиять на взаимодействие вируса-хозяина, они могут оказывать косвенное влияние на восприимчивость к COVID-19 и исход заболевания, например, повышать уровень окислительного стресса, тем самым ухудшая исход заболевания [Devaux *et al.*, 2020].

Проведенный нами анализ 2754 вариантов в гене *ACE2* из базы данных геномов популяции Европы выявил более низкое соотношение миссенс-

вариантов в популяции Южной Европы по сравнению с другими регионами Европы, что может, отчасти, объяснить более высокий уровень смертности от COVID-19 в Испании и Италии [Shikov *et al.*, 2020]. Для изучения относительных различий в спектре вариантов в гене *ACE2* между популяциями мы применили анализ основных компонентов (PCA). Важно отметить, что все субпопуляции имели очень неравномерные размеры выборки. Учитывая это наблюдение, мы сузили набор данных, выбрав варианты с ненулевыми частотами аллелей в каждой популяции. Полученный PCA-график, основанный на 60 вариантах, показал определенные различия между популяциями; в то же время наблюдалось мало различий между европейскими популяциями. Мы предположили, что различия могут быть еще более выраженными при рассмотрении вариантов, имеющих функциональное значение. Чтобы проверить эту гипотезу, мы провели аналогичный анализ с использованием 229 миссенс вариантов. Важно отметить, что все популяции имели достаточно высокий охват (число аллелей) на этих сайтах. Анализ PCA показал, что население Южной Европы, а также население Эстонии и Болгарии, как правило, отделяются от других европейских популяций, и различия между другими европейскими популяциями также были более очевидными [Shikov *et al.*, 2020].

Далее мы провели сравнительный анализ частот пяти вариантов (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) в гене *ACE2* в русской и европейских популяциях (табл.25). Было установлено, что русские похожи на другие европейские популяции, что предполагает сходный уровень инфицирования и тяжести заболевания. Это была важно для понимания эпидемиологической ситуации в начале эпидемии в марте-апреле 2020 года.

Хотя влияние вариантов в гене *ACE2* на экспрессию соответствующего белка в легких не обнаружено, влияние cis-eQTLs на функцию ACE2 в различных тканях головного мозга может быть связано с неврологическими осложнениями у пациентов с COVID-19 [Strafella *et al.*, 2020]. Мы

предположили, что даже незначительно повышенная экспрессия ACE2 может привести к увеличению количества молекул рецептора на поверхности клетки, что, в свою очередь, может повысить восприимчивость к COVID-19. Статистическое сравнение частот отдельных вариантов между пациентами с легкой и тяжелой формой COVID-19 не выявило каких-либо существенных различий. Поэтому мы пришли к выводу, что, хотя существуют определенные различия в частотах ACE2 eQTL в разных популяциях, эти различия либо не оказывают влияния, либо очень мало влияют на восприимчивость и тяжесть COVID-19 [Shikov *et al.*, 2020].

Таблица 25. Варианты в гене *ACE2* в нашей когорте российских экзомов [Shikov *et al.*, 2020].

Позиция	rsID	Реф	Вар	Эффект	Патогенность	Белок	Гомозигота	Гетерозигота	AF
15582209	rs35803318	C	T	синонимичная	Патогенная	p.Val749Val	10	13	0.031
15582298	rs41303171	T	C	миссенс	Патогенная	p.Asn720Asp	2	15	0.016
15596143	rs113691336	C	CATAAG	интронная	-	-	232	82	0.609
15606024	-	T	TTC	интронная	-	-	1	1084	0.001
15606028	-	A	ATTGT	интронная	-	-	1	1084	0.001
15606029	-	A	ATTACTTT	интронная			1	1084	0.001
15607650	rs971249	T	C	интронная	Benign	-	278	142	0.671
15610348	rs2285666	C	T	сплайсинг	Benign	-	66	89	0.205

Тот же механизм может быть справедлив для редких гаплотипов в гене *ACE2*, которые, как было установлено, чрезмерно представлены у российских пациентов с тяжелой формой COVID-19 (рис. 29). Однако вероятно, что влияние этих вариантов на фенотип не связано с рецепторной функцией ACE2. Тем не менее редкие варианты могут косвенно играть роль в патологии COVID-19, влияя на важные нормальные функции белка, что позволяет объяснять найденную корреляцию тяжести болезни - наличием хронических заболеваний [Shikov *et al.*, 2020].

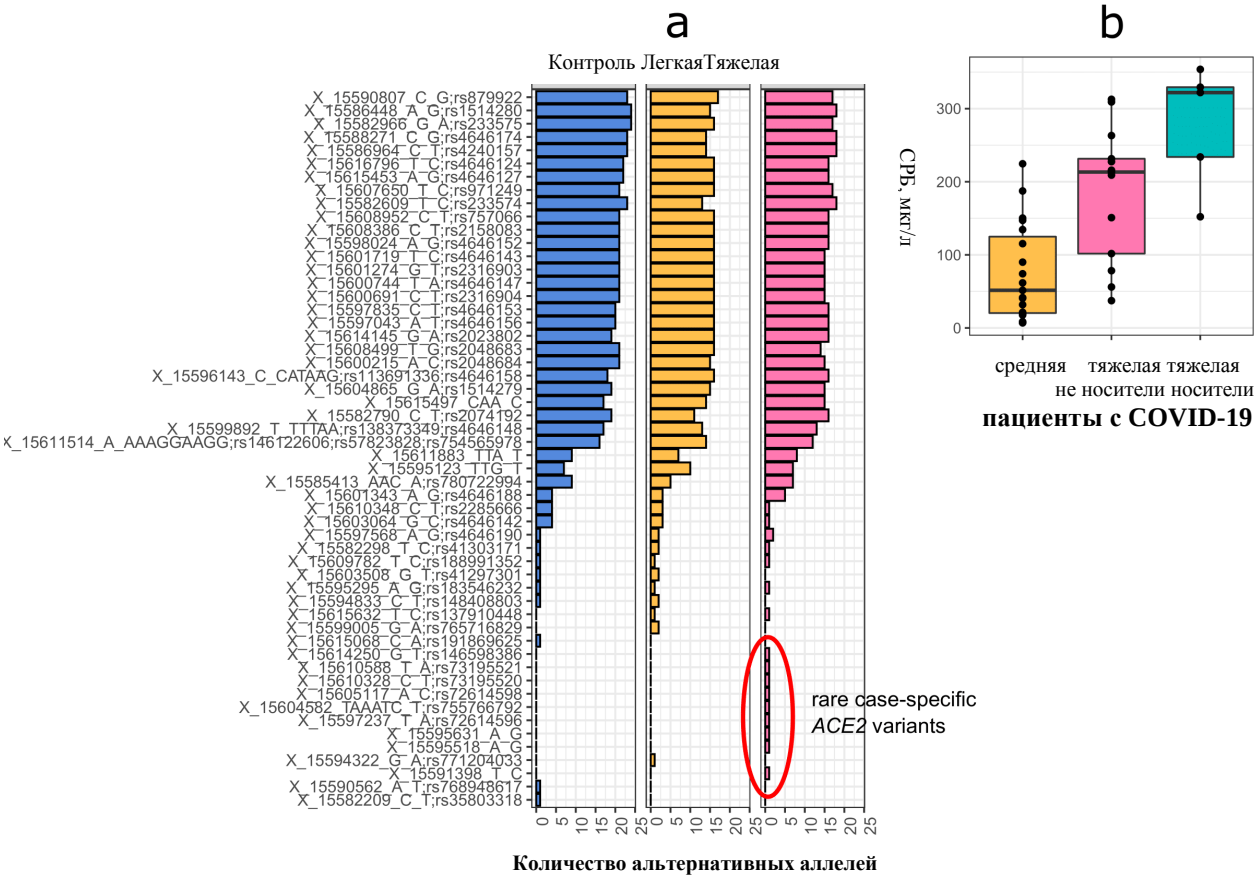


Рисунок 29. Распространенные и редкие варианты в гене *ACE2* у разных пациентов с COVID-19 [Shikov *et al.*, 2020]. (a). Альтернативная аллель включает 54 SNP и indels среди пациентов с легкими и тяжелыми формами COVID-19. (b) Средняя концентрация С-реактивного белка у пациентов с COVID-19.

3.5. Мультигенный характер предрасположенности к COVID-19

Важно отметить, что не только полиморфизм в гене *ACE2* может влиять на тяжесть протекания заболевания. Пептидазы ANPEP, DPP4 и ENPEP, TMPRSS2 также могут служить корецепторами ACE2 при заражении вирусом SARS-CoV2 [Qi *et al.*, 2020; Rossi *et al.*, 2021]. Исследования показывают, что для активации вирусного S-белка необходим фермент TMPRSS2, взаимодействие с которым облегчает внедрение вируса в клетку [Shulla *et al.*, 2011].

Кроме этого, для проявлений контагиозности вируса важны набор антигенов в системе главного локуса гистосовместимости организма, группа крови пациента, интенсивность продукции цитокинов и других факторов

воспаления и противовирусного иммунитета, в частности, секреция интерлейкинов и хемокинов.

Еще в начале эпидемии была установлена ассоциация тяжести COVID-19 инфекции с локусами (*HLA-B**4601), генами *FcγRIIA, MBL, TMPRSS2, TNF-α, IL-6*, антигеном А групп крови человека и другими [Asselta *et al.*, 2020; Feldmann *et al.*, 2020; Wu B. *et al.*, 2020; Anisenkova *et al.*, 2021].

Эпидемия стимулировала генетические исследования, и стало появляться все больше информации о том, что прогнозирование генетической восприимчивости к COVID-19 может помочь клиницистам выбрать правильные методы лечения для пациентов [Prakrithi *et al.*, 2021]. В недавних исследованиях сообщалось о нескольких десятках ассоциаций между генетическими вариантами и заболеваемостью, тяжестью и смертностью COVID-19 среди различных этнических групп [Suh *et al.*, 2022]. Например, одно геномное ассоциативное исследование (GWAS), проведенное в Объединенных Арабских Эмиратах на выборке из 600 участников, выявило 8 локусов восприимчивости к тяжелому COVID-19. Было обнаружено, что локусы в этих генах ассоциированы с воспалением, опосредованным Т-клетками, влияющим на выработку воспалительных цитокинов [Mousa *et al.*, 2021]. В другом исследование из Европы (семь больниц из Италии и Испании) были выявлены две перекрестно реплицирующиеся ассоциации тяжелого течения COVID-19 с вариантами в локусах 3p21.31 и 9q34.2 (группа GWAS тяжелого COVID-19). Эти локусы охватывают несколько генов, включая *SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6, CCR1* для 3p21.31 и *ABO* - для 9q34.2. Связь этих локусов с COVID-19 была также обнаружена Wu с коллегами в исследовании когорты китайских пациентов [Wu *et al.*, 2021].

Несмотря на большое количество зарегистрированных ассоциаций, варианты, которые были обнаружены в одном исследовании, могут быть не подтверждены в другом [Suh *et al.*, 2022]. Для того чтобы выявить локусы, которые демонстрируют связь между различными когортами и этническими

группами, проводятся крупномасштабные мета-анализы, одним из которых является проект COVID-19 HG project (COVID-19 Host Genetics Initiative, 2021), где сегодня представлена большая часть современных данных по генетической предрасположенности к COVID-19 [COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature 600, 472–477 (2021).https://doi.org/10.1038/s41586-021-03767-x].

Результаты метаанализа, опубликованные COVID-19 HG, содержат 23 локуса восприимчивости [COVID-19 Host Genetics Initiative, 2021]. Локусы, выявленные в этом мета-анализе, включают вышеупомянутые 3p21.31 и 9q34.2, а также несколько других важных локусов. Например, локус на хромосоме 6, который охватывает ген *FOXP4*. FOXP4 играет ключевую роль в регулировании регенерации секреторных эпителиальных клеток легких и, таким образом, может влиять на выработку слизи для защиты легких от патогенов и загрязнений. Другой локус (12q24.13) содержит набор паралогичных генов семейства *OAS* (*OAS1, OAS2, OAS3*). Гены *OAS* кодируют олигоаденилатсинтетазы, которые участвуют в деятельности врожденной иммунной системы. В дополнение к этим регионам также было обнаружено, что локус 21q22.11 связан со степенью тяжести и госпитализацией у пациентов с COVID-19. Этот локус включает ген *IFNAR2*, продукт которого работает как часть - в цепи интерферонового иммунитета. Наиболее достоверные ассоциации, выявленные в локусах 6p21.1, 12q24.13 и 21q22.11, также ассоциированы с генами, экспрессия которых значительно меняется в легких.

В то время как общегеномные исследования в целом были успешными в выявлении генетических факторов риска COVID-19, поиск новых ассоциаций в плохо изученных популяциях может быть затруднен из-за ограниченного размера когорты и/или в случае ограниченной широты охвата генома (т.е. в исследованиях, основанных на CES или таргетном секвенировании). В отечественной литературе, помимо пилотного исследования по генетике COVID-19 [Shikov *et al.,* 2020], отсутствуют публикации по

широкомасштабному изучению современными молекулярно-генетическими полногеномными методами предикторов различного течения COVID-19 инфекции, в том числе в разные «волны» заболевания.

Поэтому в продолжение предыдущего исследования [Shikov *et al.*, 2020], мы приступили к выявлению дополнительных локусов восприимчивости, связанных с тяжелой формой COVID-19. Для этого мы собрали, описали и исследовали когорту из 840 российских пациентов с COVID-19, используя ранее установленные критерии обследования пациентов [Glotov *et al.*, 2021a; Shcherbak *et al.*, 2021]. Секвенирование проводили на платформах для NGS секвенирования Illumina и MGI, используя панель зондов для клинического секвенирования экзома - CES. Перед всеми последующими анализами данные секвенирования были равномерно обработаны и совместно генотипированы для получения полного набора вариантов в пределах целевых интервалов экзома. В нашей выборке было обнаружено в общей сложности 727 656 генетических вариантов. 98 382 из этих вариантов были несинонимичными вариантами (включая миссенс варианты и варианты предполагаемой потери функции (pLOF). После фильтрации вариантов с низким качеством и/или частотой осталось 13 983 распространенных (AF>=5%). Из оставшихся редких (AF<5%) вариантов 1884 варианта были аннотированы как варианты в канонических транскриптах 1121 белок кодирующих генах. Также все пациенты были оценены на наличие моногенных нарушений иммунной системы, при этом у них не были выявлены патогенные варианты, находящиеся в базе данных ClinVar v. 20211130 [Shcherbak *et al.*,2022].

Идентификация значимых ассоциаций на уровне генома или экзома может быть затруднена в когортах с ограниченным размером. Поэтому мы решили применить более систематический подход и проанализировать генетические факторы COVID-19, используя обширный набор фенотипических данных, доступных для нашей когорты пациентов. Для каждого пациента был собран широкий набор из более чем 100 количественных и бинарных признаков

[Shcherbak *et al.*,2022]. Набор признаков включал основные параметры, которые служат прогностическими факторами риска тяжелого COVID-19 в соответствии с нашими недавними публикациями [Shikov *et al.*, 2020; Shcherbak *et al.*,2021]: уровень С-реактивного белка, интерлейкина-6 (IL-6), ферритина, D-димера, ЛДГ, глюкозы и креатина в сыворотке; количество клеток крови (лимфоцитов, лейкоцитов, нейтрофилов на мл образца крови); оценка поражения легких, полученная по изображениям компьютерной томографии; шкала NEWs. Большинство показателей регистрировалось каждые два дня во время госпитализации. Как и ожидалось, зарегистрированные значения большинства этих признаков существенно различались для пациентов с различными исходами (смерть или выздоровление) или тяжестью заболевания. Поскольку нормализация фенотипических данных с использованием подхода по преобразования нормализации на основе обратного ранга (IRNT) может повысить эффективность анализа ассоциаций в масштабах всего генома [Goh и Yap, 2009], мы нормализовали все количественные признаки в наборе данных с помощью IRNT с дополнительной фильтрацией и предсказанием недостающих точек данных. Полученные таким образом предварительно обработанные данные о количественных признаках были использованы для дальнейшего анализа ассоциаций (рис. 30).

Сначала мы провели анализ ассоциаций распространенных и редких вариантов с бинарными признаками (смерть и тяжесть). Анализ ассоциации распространенных вариантов не выявил значимых ассоциаций и не выявил доказательств сигнала ассоциации по всему экзому. Затем мы проверили вовлеченность редких вариантов в клинически значимых генах, проведя серию тестов ассоциации редких вариантов, используя агрегацию подсчетов вариантов как на уровне генов, так и на уровне метаболических путей (стратегия, аналогичная той, которую использовали Povysil с коллегами. [Povysil *et al.*, 2021]. Чтобы улучшить наш анализ, мы провели как внутренние сравнительные тесты (анализ ассоциаций, основанный на сравнении пациентов с различными

исходами или тяжестью COVID-19), так и сравнение с популяционными частотами аллелей [Barbitoff *et al.*, 2019]. В соответствии с результатами, полученными Povysil с коллегами, мы не обнаружили генов и метаболических путей, демонстрирующих значительную связь с тяжестью заболевания или исходом в нашем наборе данных.

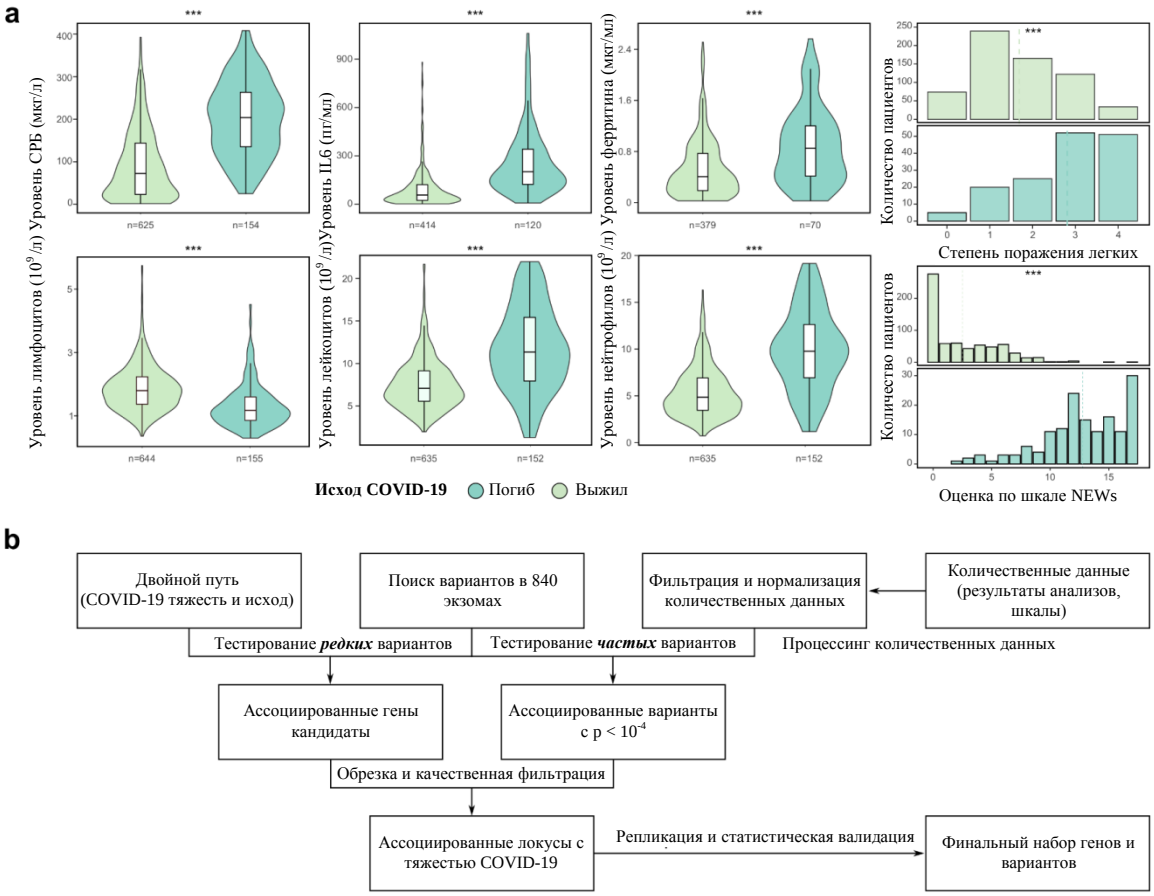


Рисунок 30. Идентификация потенциальных генетических маркеров-предикторов тяжелого COVID-19 с использованием глубоко фенотипированной когорты [Shcherbak *et al.*, 2022]. **(a)** Распределение выбранных количественных признаков в когорте из 840 пациентов с COVID-19 из России. Показаны распределения уровней сывороточного С-реактивного белка (СРБ), интерлейкина-6 и D-димера, оценка поражения легких на основе компьютерной томографии (от 0 до 4), количество лимфоцитов, лейкоцитов и нейтрофилов в образцах крови, а также шкалы NEWS. Все показанные значения соответствуют максимальным значениям, зарегистрированным во время госпитализации.

Значения, превышающие три стандартных отклонения от среднего значения по совокупности, опускаются. *** - p< 0,001 по критерию Уилкоксона-Манна-Уитни (для количественных признаков) или по критерию хи-квадрат (для качественных признаков). **(b)** Схематическое представление протокола анализа данных, используемого в исследовании.

Следующим этапом анализа был поиск ассоциаций SNP с большим набором количественных признаков. Мы проанализировали ассоциацию 13 983 распространенных (MAF>0,05) вариантов, обнаруженных в нашем массиве генотипов, с набором из 53 предварительно обработанных количественных признаков с низкой частотой пропусков. После первоначального раунда GWAS результаты для каждого признака были обработаны дополнительно, путем проверки графиков Q-Q. В общей сложности мы обнаружили 5 количественных признаков, которые демонстрировали сигналы ассоциаций. К ним относятся уровни сывороточного С-реактивногобелка (СРБ), количество лимфоцитов, лейкоцитов и нейтрофилов, а также степень поражения легких по данным компьютерной томографии (рис. 31).

В общей сложности 15 вариантов показали ассоциацию при $p<10^{-4}$ для выбранных количественных признаков [Shcherbak *et al.*, 2022]. Только два из идентифицированных вариантов достигли порога значимости для всего экзома на уровне ($3,5*10^{-6}$) (порог, соответствующий стандартному уровню значимости p<0,05 с поправкой на количество протестированных вариантов). Эти варианты показали ассоциации как с количеством лейкоцитов, так и с количеством нейтрофилов, что можно объяснить высокой степенью корреляции между этими признаками (табл. 26). Кластеризация найденных вариантов по неравновесию сцепления (LD) выявила 11 независимых локусов: 1 - для уровня СРБ в сыворотке крови; 2 - для количества лимфоцитов, лейкоцитов и нейтрофилов; и 5 - для оценки поражения легких на основе компьютерной томографии (табл. 26). Четыре из этих замен были расположены в кодирующих

последовательностях, в то время как остальные варианты были интронными или другими некодирующими вариантами [Shcherbak *et al.*, 2022].

Из 11 независимых вариантов, выявленных в нашем анализе, 9 соответствовали значимым cis-eQTLs в соответствии с данными Genotype Tissues Expression (GTEx). Четыре из этих вариантов соответствовали cis-eQTLs, влияющим на экспрессию множества генов в нескольких тканях.

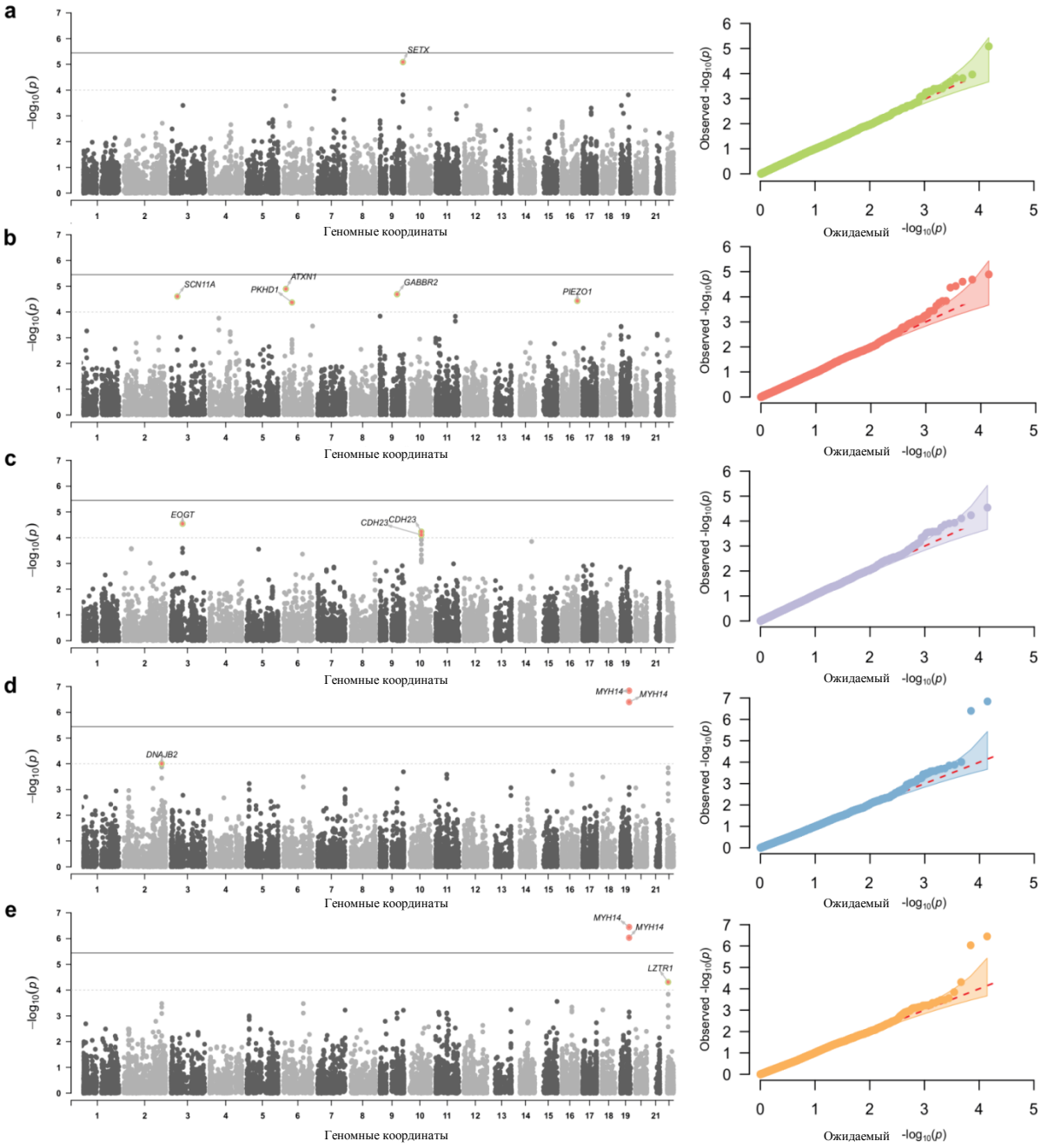


Рисунок 31. Результаты геномной ассоциации отдельных количественных признаков у пациентов с COVID-19 [Shcherbak *et al.*, 2022]. Показаны графики Манхэттена (слева) и квантиль-квантиль (справа) p-значений ассоциации для (сверху вниз) уровней сывороточного СРБ **(a)**, оценки поражения легких на основе компьютерной томографии **(b)**, уровня сывороточных лимфоцитов **(c)**, лейкоцитов **(d)** и нейтрофилов **(e)**. Пороговые значения на манхэттенских графиках соответствуют пороговому значению для всего экзома ($4x10^{-6}$) и более слабому пороговому значению $p=10^{-4}$, используемому для отбора ассоциированных генов-кандидатов.

Таблица 26. Возможные генетические варианты, ассоциированные с количественными признаками, связанными с COVID-19, в когорте российских пациентов [Shcherbak *et al.*, 2022].

Локус	**rsID**	**Замена**	**AF***	**Признак (и)**	**Ген**	**Результат**	***B*****	**p-value**	**GTEx eQTLs****
2:219280564	rs2276638	6247C>G	0.135	Число лейкоцитов	*DNAJB2*	Интронный вариант	-0.29	9.84E-05	Множество генов и тканей
3:38894643	rs33985936	c.2725G>T (p.Val909Phe)	0.241	КТ баллы	*SCN11A*	Миссенс вариант	-0.24	2.50E-05	Множество генов и тканей
3:68997990	rs4855544	g.20905C>A	0.332	Число лимфоцитов	*EOGT*	Интронный вариант	0.23	2.88E-05	Множество генов и тканей
6:16306520	rs16885	c.2257C>T (p.Pro753Ala)	0.197	КТ баллы	*ATXN1*	Миссенс вариант	0.27	1.28E-05	неизвестно
6:51830849	rs1571084	g.261777T>A	0.333	КТ баллы	*PKHD1*	Интронный вариант	0.21	4.30E-05	*PKHD1* (кожа)
9:98299383	rs41273925	g.414815C>G	0.081	КТ баллы	*GABBR2*	Интронный вариант	0.38	2.06E-05	*TBC1D2* (щитовидная железа)
9:132278286	rs11243705	g.81700A>G	0.180	СРБ	*SETX*	Интронный вариант	0.30	8.18E-06	*SETX* (множество тканей)
10:71799129	rs4747194	c.7073G>T (p.Arg2358Gln)	0.243	Число лимфоцитов	*CDH23*	Миссенс вариант	0.25	5.84E-05	*CDH23* (кишечник, тестикулы), *PSAP* (множество

									тканей)
16:88738516	rs34600315	c.*648_*649del	0.657	КТ баллы	*PIEZO1*	Некодирующий транскрипт экзонный вариант	0.21	3.73E-05	*PIEZO1* (клетки крови)
19:50259161	rs1651553	c.2127A>G	0.770	Число лейкоцитов, нейтрофилов	*MYH14*	Синонимичный вариант	0.32 0.31	1.45E-07 3.55E-07	неизвестно
22:20992196	rs112544	g.14928T>G	0.709	Число нейтрофилов	*LZTR1*	Интронный вариант	0.23	4.88E-05	Множество генов и тканей

* - частота аллели указана по отношению к нереференсной аллели; ** - размеры эффекта указаны по отношению к преобразованным IRNT значениям количественных признаков; *** - данные для выпуска анализа GTExv8 (доступен полный список значимых cis-EQTL.

DNAJB2 – член B2 семейства белков теплового шока Днк J (Hsp40); *SCN11A* – альфа-субъединица 11 натриевых каналов с регулируемым напряжением; *EOGT* – домен EGF, специфичный для O-связанной N-ацетилглюкозаминтрансферазы; *ATXN1* – атаксин 1; *PKHD1* – PKHD1 цилиарный IPT–домен, содержащий фиброцистин/полидуктин; *GABBR2* - гамма-аминомасляная кислота субъединица рецептора типа B 2; *SETX* – сенатаксин; *CDH23* – связанный с кадгерином 23; *PIEZO1* – компонент 1 механочувствительного ионного канала пьезо–типа; *MYH14* – тяжелая цепь миозина 14; *LZTR1* - лейциноподобный регулятор транскрипции 1.

Из них 3 генных варианта оказали наиболее значительное влияние на соседние гены: rs2276638 вариант в интроне гена *DNAJB2* влияет на экспрессию гена *PTPRN* в цельной крови, в соответствии с данными GTEx (p-значение = $2*10^{-27}$); rs33985936 вариант в гене *SCN11A* ассоциирован с экспрессией гена *TTC21A* в пищеводе, и вариант rs112544 в гене *LZTR1* влияет на экспрессию THAP7-AS1 антисмыслового транскрипта. Из оставшихся 5 вариантов со значительным сигналом cis-eQTLs, 4 оказывали значительное влияние на

экспрессию гена, несущего соответствующий вариант, и только 1 влиял на экспрессию соседнего гена. Предполагается, что варианты в генах *ATXN1, PKHD1, SETX, PIEZO1* и *CDH23* оказывают прямое влияние на фенотип путем изменения функции (в случае миссенсных вариантов в генах *ATXN1* и *CDH23*) или уровней экспрессии своего соответствующего гена.

Хотя мы и выявили 11 независимых генетических вариантов, которые ассоциированы с количественными признаками, которые в свою очередь непосредственно связанными с тяжестью и исходом заболевания, важно отметить, что уровень значимости этих ассоциацийнедостаточен для того, чтобы сделать уверенный вывод о влиянии найденных вариантов на фенотип пациента. Это предопределяет необходимость дополнительной репликации наблюдаемых ассоциаций и подтверждения их истинной роли в патогенезе COVID-19 [Shcherbak *et al.,* 2022].

Перед тем как получить такую валидацию, мы сначала задались вопросом, можно ли использовать выявленные варианты для прямого прогнозирования тяжести заболевания и/или исхода в нашей когорте. Мы начали с построения простой оценки риска путем вычисления взвешенной суммы аллелей риска в генотипе каждого пациента. Оценка имела нормальное распределение (рис. 32 (a)). Чтобы проверить, может ли такой показатель предсказать тяжесть или исход госпитализации у пациентов с COVID-19, мы отобрали пациентов, принадлежащих к верхнему децилю распределения баллов (т.е. 10% всех пациентов с наивысшими значениями баллов). Затем мы использовали статистику хи-квадрат для сравнения тяжести заболевания и исходов у этих пациентов и остальной части нашей выборки. Мы обнаружили значительные различия во всех сравнениях (рис. 32 (b)), при этом пациенты, относящиеся к децилю с наивысшим уровнем риска, имели большую вероятность смерти и большую тяжесть заболевания.

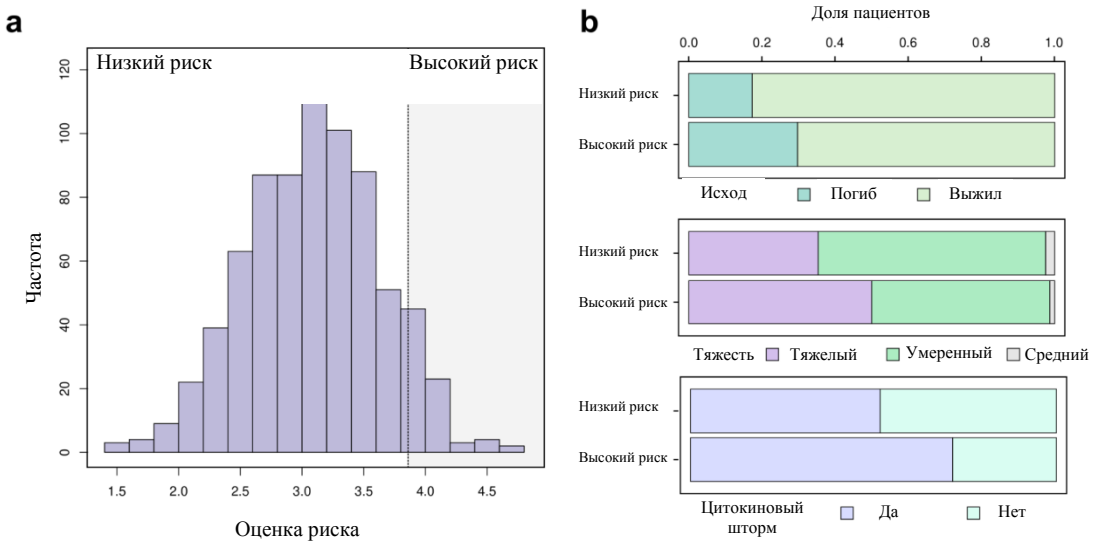


Рисунок 32. Оценка риска предсказания тяжести заболевания и исхода, основанная на 11 выявленных вариантах [Shcherbak *et al.*, 2022].

(a) Распределение оценки риска, рассчитанной на основе 11 основных SNP, показанных в таблице 26. Заштрихованная область указывает на дециль наивысшего балла, соответствующий лицам с высоким риском.

Логистическая регрессия, основанная на 11 идентифицированных маркерах, предсказывает исход COVID-19 с ROC/AUC=0,59. Полученные результаты подтверждают, что выявленные варианты можно рассматривать как генетические факторы риска тяжелой формы COVID-19.

Далее мы продолжили воспроизводить наблюдаемые ассоциации в независимых исследованиях. Для этого мы использовали данные COVID-19-HG (COVID-19 HG Initiative, 2021 (release 6), а также другие исследования GWAS в когортах европейского происхождения. При использовании данных COVID-19-HG мы успешно воспроизвели только одну из 11 ассоциаций-кандидатов (rs33985936 в гене *SCN11A*), которая показала умеренную значимость при анализе пациентов с COVID-19 по сравнению с популяцией (сравнение C2 в COVID-19-HG). Мы также попытались воспроизвести наши результаты в данных GWAS с тяжелой формой COVID-19 (исследование с участием пациентов и контрольной группы испанского и итальянского происхождения). К

сожалению, не было успешно воспроизведено ни одного эффекта данных генетических вариантов. Возможно, это связано с популяционными отличиями между нашими группами [Shcherbak *et al.,* 2022].

В дополнение к воспроизведению ассоциаций, найденных в других исследованиях COVD-19, мы стремились идентифицировать прочие (не связанные с COVID-19) сложные признаки, связанные с выявленными нами вариантами. С этой целью мы провели анализ ассоциаций по всему фенотипу (база PheWAS) с использованием данных Global Biobank Engine [McInnes *et al.,* 2018]. Все ассоциации при $p<10^{-5}$ считались значимыми в PheWAS. Мы смогли определить совпадения PheWAS для 3 из 11 протестированных вариантов. Примечательно, что все выявленные ассоциации были миссенс вариантами и они были идентифицированы по признакам, связанным с кроветворением. Вариант rs33985936 в гене *SCN11A* - единственный вариант, который был воспроизведен в когорте COVID-19-HG - показал значительную связь с уровнем и количеством тромбоцитов в данных Биобанка Великобритании. В дополнение к этому варианту, rs16885 в гене *ATXN1* показал значительную ассоциацию PheWAS со средним уровнем корпускулярного гемоглобина, а rs4747194 вариант в гене *CDH23* был связан с процентным содержанием моноцитов в крови. Эти результаты обеспечивают дополнительную поддержку биологической роли выявленных миссенс вариантов в формировании признаков, связанных с COVID-19. Таким образом, мы определили набор из 11 генетических вариантов, демонстрирующих умеренную связь с количественными и номинальными признаками, связанными с тяжестью COVID-19. Для трех из этих вариантов мы смогли найти подтверждающие доказательства их роли в патогенезе COVID-19 [Shcherbak *et al.,* 2022].

Несколько генов, информативность которых была показана в нашем исследовании заслуживают подробного обсуждения. Это, во-первых, наиболее значимый (и единственный значимый на уровне всего экзома) вариант в гене *MYH14,* который кодирует немышечный миозин II C (NMIIC), преимущественно

экспрессируемый во внутреннем ухе, включая орган Корти [Donaudy *et al.*, 2004]. Большое разнообразие генетических исследований показывает прямую связь между патогенными вариантами в гене *MYH14* с аутосомно-доминантной несиндромной наследственной потерей слуха (ADNSHL) в европейских [Lerat *et al.*, 2019; del Castillo *et al.*, 2022] и азиатских популяциях [Hiramatsu *et al.*, 2021; Wang M *et al.*, 2020]. Мутации в гене *MYH14* могут быть факторами риска метастазирования у пациентов с раком поджелудочной железы [Surcel *et al.*, 2019] и нейробластомы [Giwa *et al.*, 2020] и обеспечивать уклонения раковых клеток в полости рта от иммунного ответа [Pérez-Valencia *et al.*, 2018], что подтверждается тем, что нарушение NMIIC вызывает гиперпластическую пролиферацию эпителиальных клеток [Nguyen-Ngoc *et al.*, 2017]. Наконец, несинонимичные мутации в гене *MYH14* увеличивают развитие стеноза канала улиткового нерва [Liang *et al.*, 2021], неврологических осложнений диабета 2 типа [Rahman *et al.*, 2020], периферической невропатии [Almutawa *et al.*, 2019] и болезни Шарко-Мари-Тута (CMT) [Kanwal *et al.*, 2018]. Вышеупомянутые данные указывают на то, что продукт гена *MYH14* является важнейшим плейотропным регулятором гомеостаза - следовательно, наличие вариантов в этом гене ассоциированныхс различными заболеваниями, закономерно. Однако следует отметить, что аннотация вариантов не дает четкой функциональной основы для понимания влияния вариантов на фенотип. Эффекты вариантов в гене *MYH14* на фенотип могут быть косвенными и могут вызываться продуктами других генов в том же локусе.

Вторым ключевым геном в нашем исследовании является ген *DNAJB2*, который кодирует важный белок группы шаперонов Hsp40. Известно, что такие белки способствуют субстратной специфичности других шаперонов и опосредуют реакцию на стресс [Kampinga and Craig, 2011; Craig and Marszalek, 2017]. Найденная ассоциация вариантов в гене *DNAJB2* с уровнем повреждения легких при инфекции SARS-CoV-2 представляет интерес и может указывать на роль, которую играют пути реагирования на стресс при формировании тяжести

заболевания, что позволяет выдвинуть гипотезу, что реакция на тепловой стресс важна для смягчения негативного воздействия воспаления на структуру ткани. Таким образом, отклонение в регуляции реакции на тепловое воздействие может спровоцировать разрушение легочной ткани с последующим развитием фиброза легких и проблемы с дыханием у пациентов с COVID-19. Примечательно, что мутации в гене *DNAJB2* способствуют более тяжелому характеру протекания болезни Шарко-Мари-Тута [Yalcintepe *et al.*, 2021], двигательных невропатий (например, у Frasquet *et al.*, 2021) и других похожих заболеваний.

Ассоциация локуса в гене *LZTR1* также представляет интерес, поскольку этот ген кодирует важный leucine-zipper транскрипционный регулятор, связанный с пролиферацией клеток и различными типами рака [Zhang Z. *et al.*, 2021]. Патогенные варианты в гене *LZTR1* влияют на прогрессирование шванном, вызывая опухоли периферических нервов [Ishigami *et al.*, 2021; Piotrowski *et al.*, 2022], нейрофиброматоз [Perez-Becerril *et al.*, 2021], синдром Нунана [Talley et al., 2021], неиммунные отеки [Hurni *et al.*, 2021; Zhou *et al.*, 2021]. Показано, что rs112544 регулирует экспрессию гена *THAP7* и его антисмыслового транскрипта THAP7-AS1. В свою очередь, ген *THAP7* кодирует репрессор транскрипции, который деацетилирует гистоны [Macfarlan *et al.*, 2005]. Интересно отметить, что сверхэкспрессия THAP7 индуцирует восприимчивость клеток гепатомы Huh7 человека к вирусной инвазии гепатитом С [Dächert *et al.*, 2019]. Эти данные свидетельствуют о том, что как ген *LZTR1* сам по себе, так и гены-мишени, на которые влияет rs112544 этого гена, могут быть вовлечены в регуляцию функции иммунной системы и противовирусный ответ.

Ген *PIEZO1* кодирует механически управляемый ионный канал с плейотропными эффектами в организме человека. Мутации в этом гене сочетаются с патологическими состояниями, а именно - наследственной анемией [Zhou *et al.*, 2021], врожденной лимфедемой [Mustacich *et al.*, 2021],

лимфатической дисплазией [Li *et al.*,2021] и др. Таким образом, изменения в гене *PIEZO1* могут повлиять на иммунный ответ при SARS-CoV-2.

Варианты фиброцистина (кодируемого геном *PKHD1*) усугубляют последствия аутосомно-рецессивной поликистозной болезни почек у детей и взрослых [Zhang Z. *et al.*, 2021]. Также было обнаружено, что полиморфизм этого гена связан с умеренными когнитивными нарушениями [Mahmud *et al.*, 2021] и метахронным раком печени [Ohni *et al.*, 2022].

Аналогичным образом варианты в других 11 идентифицированных генах, например, *ATXN1, GABBR2, SETX, CDH23* также участвуют в патологии нервной системы, но не имеют четкой связи с иммунитетом и/или реакцией на инфекционные заболевания [Kumaran *et al.*, 2014; Wallace and Bird, 2018; Miyazawa *et al.*, 2021; Hadjinicolaou *et al.*, 2021; Saleem *et al.*, 2021]. Полученный результат может указывать на определенную взаимосвязь между функцией нервной системы и тяжестью COVID-19.

3.6. Методологические проблемы и оценки клинико-генетических ассоциаций

Пандемия COVID-19 привлекла значительное внимание к исследованию генома человека и вирусов. За последние два года в ряде публикаций была рассмотрена роль наследственной предрасположенности к инфекции SARS-Cov-2 и тяжелым ее осложнениям [Suh *et al.*, 2022]. Количество исследований GWAS, связанных с COVID-19, как крупномасштабных, так и локальных, постоянно растет. Исследования направлены как на выявление факторов восприимчивости хозяина и анализ тяжести заболевания, так и на изучение ассоциации конкретных симптомов с генетическими маркерами. Эти работы позволяют посмотреть на эпидемиологию по-новому. К примеру, установлена генетическая корреляция между более низким уровнем образования и более высоким риском COVID-19 [Jiang *et al.*, 2021]. С помощью GWAS найдены различные локусы, ассоциированные с риском потери запаха во время

инфекции: *UGT2A1/UGT2A2* [Shelton *et al.*, 2022], *MUC5B* [van Moorsel *et al.*, 2021], *LZTFL1* и *RAVER1* [Fink-Baldauf *et al.*, 2022] и другие. Как и во многих других исследованиях GWAS, репликация наблюдаемых ассоциаций и идентификация генетических вариантов, связанных с фенотипом, в различных человеческих популяциях остается важной и сложной проблемой. Например, Чжан с коллегами выявили возможную роль редких pLoF вариантов в генах, связанных с иммунитетом к интерферону I типа [Zhang H. *et al.*, 2020], с риском COVID-19. Однако эти ассоциации не были воспроизведены другими авторами [Povysil *et al.,* 2021]. Похожий результат мы видим при репликации локусов, идентифицированных в китайской когорте, в других внешних когортах [Li Y. *et al.,* 2021] и в исследовании COVID-19-HG. Такая низкая репликация, хотя и характерна для генетики сложных признаков, подчеркивает особую роль дизайна исследования и структуры популяции на выявление генетических факторов риска инфекционных заболеваний.

Анализ генетических ассоциаций в когортах ограниченного размера, особенно когда отсутствуют общегеномные генотипы, может также препятствовать открытию новых локусов восприимчивости в недопредставленных популяциях. Следовательно, для решения проблемы низкой статистической мощности анализа необходимо использовать сложные подходы, а также квалифицированно разделять группы для последующего анализа с учетом негенетических факторов риска при COVID-19, которые мы описали ранее [Shcherbak *et al.*,2021]. Несмотря на то, что в нашем исследовании только один из выявленных вариантов успешно прошел репликацию во внешних когортах, два дополнительных варианта продемонстрировали номинальную значимость в независимых исследованиях. Наши результаты аналогичны данным, полученным Li с соавторами [Li Y. *et al.*, 2021]. Как утверждалось ранее, низкая частота репликации может отражать как различия в дизайне исследования, так и различия между популяциями. Возможно, что еще более важно, мы наблюдали значительные PheWAS для трех наших вариантов в

данных Биобанка Великобритании. Важно отметить, что все ключевые PheWAS соответствовали признакам, связанным с кроветворением, подтверждая актуальность выявленных ассоциаций. Кроме того, анализ GTEx eQTLs также показывает, что многие из идентифицированных вариантов влияют на экспрессию генов в иммунных клетках или в цельной крови (например, rs2276638 в гене *DNAJB2*, rs34600315 в гене *PIEZO1*). Примечательно, что лишь немногие из выявленных нами вариантов влияют на экспрессию генов в легких. Это наблюдение может быть объяснено спецификой стратегии анализа, которая в основном сосредоточена на различных показателях в крови у пациентов с COVID-19.

Следует отметить, что общая сила наблюдаемых ассоциаций в нашем исследовании является умеренной, так, как только один из локусов воспроизводится в независимых когортах. Это наблюдение может быть объяснено либо небольшим размером выборки и, из-за этого, слабым сигналом ассоциации в нашем исследовании, либо специфическими для популяции эффектами вариантов. Поэтому, учитывая различия в дизайне разных исследований, мы не ожидаем множества репликацийнаших данных. Тем не менее, полученные нами результаты демонстрируют полезность глубокого лабораторного фенотипирования пациентов с COVID-19 для выявления новых генетических вариантов, влияющих на тяжесть и/или исход заболевания. Следовательно, мы считаем, что наша работа может служить примером успешной косвеннойоценки факторов риска тяжелой формы COVID-19.

Поиск генетических основ инфекционных заболеваний в настоящее время является основополагающим для понимания индивидуального механизма протекания болезни. Рассмотренный в данной главе поиск связи между экзомом и различными состояниями и исходами COVID-19 показывает, насколько важными для понимания патогенеза любого заболевания, в том числе и инфекционного процесса, являются подобного рода исследования. Нет

сомнения, что уже в течение нескольких лет данные секвенирования генома для всех основных инфекционных заболеваний позволят резко увеличить предсказательную (предиктивную) ценность досимптоматического генетического тестирования наследственной предрасположенности/устойчивостик инфекциям, и можно будет надеяться на быстрый прогресс в этой области молекулярной медицины. Тестирование на генетическую предрасположенность к отдельным инфекционным болезням войдет в общую структуру генетического тестирования, станет составляющей Предиктивной медицины и генетического клинического паспорта здоровья человека.

ЗАКЛЮЧЕНИЕ

Внедрение новых геномных технологий в последние годы немало способствовало прогрессу молекулярной медицины, прежде всего - технологии полногеномного секвенирования NGS и микроматричного анализа aCGH. С появлением каждой из этих технологий существенно расширились возможности и изменились приоритеты внедрения генетических лабораторных тестов в клинической практике. С развитием новых молекулярно-биологических подходов происходят и соответствующие изменения в базовой терминологии. Наметился четкий переход от изучения индивидуальных генов и их вариантов (мутаций) к исследованиям патогеномики заболеваний, активному поиску биомаркеров, досимптоматической профилактике и персонализированному лечению мультифакториальных заболеваний. Возникла и развивается концепция генетического паспорта здоровья человека [Баранов *и др.,* 2000, 2009, 2021].

В настоящее время в зависимости от поставленной задачи, используют несколько определений понятия «ген» и «мутация». Так, в классической генетике понятие «ген» определяется как картируемый на хромосоме локус, ответственный за тот или иной фенотипический признак [Инге-Вечтомов, 1998]. В молекулярной биологии ген рассматривают как ассоциированный с регуляторными последовательностями фрагмент ДНК, соответствующий определенной единице транскрипции [Сингер, Берг, 1998]. Мутациями называют любые наследуемые изменения (альтерации) в последовательности ДНК.

Принципиально важно, что разница между мутациями и генетическими вариантами (генетический полиморфизм) весьма относительна. Обычно генетический полиморфизм нейтрален, встречается чаще и может присутствовать у значительной части популяции — более 1 % [Баранов *и др.,* 2000]. Мутации встречаются реже, как правило, выключают работу гена, ведут к значительному снижению синтеза и уменьшению количества его белкового

продукта (минус-эффект, loss of function), или к его избытку (плюс-эффект gain of function), или к появлению аномального белка, следствием чего являются те или иные моногенные болезни. В отличие от генетического полиморфизма, фенотипический эффект большинства известных мутаций проявляется достаточно четко в виде того или иного наследственного заболевания. Таким образом, генетический полиморфизм и мутации - суть явления одного порядка. Грань между этими понятиями весьма условна, как зачастую условно разделение понятия нормы и патологии [Cotton and Scriver, 1998].

Сегодня использование NGS существенно меняет биологический смысл данной терминологии. Начиная с 2015 г. в США, а затем в 2017 в России [Рыжкова *и др.,* 2017, 2019] были разработаны рекомендации по интерпретации вариантов нуклеотидной последовательности, согласно которым, вместо терминов «мутация» и «полиморфизм» рекомендовано использовать «вариант нуклеотидной последовательности» со следующими пятью характеристиками: патогенный (pathogenic); вероятно патогенный (likely pathogenic); неопределенного значения (uncertain significance); вероятно доброкачественный (likely benign); доброкачественный (benign) [Рыжкова *и др.,* 2019]. Данную терминологию активно стали применять не только в диагностическом плане [Рыжкова *и др.,* 2019], но и при проведении различных исследований. Наши работы демонстрируют целесообразность применения этой терминологии.

Важно подчеркнуть, что паттерн (спектр) и частота различных вариантов обладают выраженной популяционной специфичностью. Это означает, что варианты, характерные для населения одного региона или этноса, существенно отличаются от таковых в других географических ареалах или в других этнических группах. Экологическая адаптация, различия в продуктах питания, тяжелые инфекции (оспа, чума, холера, СПИД), селективное преимущество гетерозигот (эффект гетерозиса), эффект основателя (founder effect, особенно в замкнутой популяции), дрейф генов (случайные колебания в популяции числа аллелей) — вот основные популяционные факторы и механизмы, определяющие

естественные колебания частоты и наличия тех или иных вариантов в различных популяциях и регионах мира [Инге-Вечтомов, 1998; Иващенко, Баранов, 2002].

Таким образом, знание о структуре гена, особенностях генетического полиморфизма и функциях различных вариантов в геноме с учетом популяционной специфичности дает понимание о наследственной природе того или иного моногенного или МФ заболевания и способствует диагностике, профилактике и лечению этих болезней. Все это в совокупности способствует развитию концепции предиктивной медицины и генетического паспорта, с последующей эволюцией его в генетический клинический паспорт (ГКП), в основе которого лежит экзомное секвенирование. ГКП уже сегодня является инструментом NGS для поиска патогенных вариантов у пробандов в семьях высокого риска и у пар, планирующих рождение здоровых детей, а также позволяет ответить на вопросы риска МФЗ. Поэтому ключевыми направлениями для повышения эффективности внедрения ГКП является развитие популяционных баз данных (в том числе отечественных) об относительных частотах генных вариантов, играющих роль в патогенезе наследственных заболеваний, поиск вариантов, ассоциированных с разными заболеваниями, совершенствование биоинформатических и статистических протоколов обработки и анализа данных секвенирования.

Важно отметить, что многие варианты, ранее перечисленные как патогенные, встречаются у здоровых людей слишком часто, чтобы вызывать заболевание, наследуемое по Менделю. Это стало наиболее мощным фактором снижения ложноположительных ассоциаций вариантов и фенотипов [Lek *et al.*, 2016]. В связи с этим информация о частоте популяционных аллелей (AF) широко используется для интерпретации клинической значимости генных вариантов.

Для решения вышеперечисленных задач мы провели популяционное исследование, где был получен набор данных из 5268 образцов и 2 092 456 вариантов [Barbitoff *et al.*, 2021]. Из них 349 811 вариантов совпадали с теми, о

которых сообщалось в нашей предыдущей публикации [Barbitoff *et al.,* 2019]. Из всех вариантов 75,7 % были известны (найдены в последней сборке dbSNP), а 24,3% (509 409) были новыми. Результаты исследования показали, что по крайней мере для рецессивных патологий, большая часть генетических детерминант является общей для России и других групп населения нашей планеты. Однако для некоторых заболеваний отмечается свой специфический спектр патогенных вариантов [Glotov O. *et al.,* 2019].

Наши результаты продемонстрировали необходимость создания генетических баз данных для конкретных популяций. Это крайне необходимо для интерпретации вариантов и выявления факторов риска заболеваний, особенно - в менее изученных популяциях. Нынешний размер выборки пока не позволяет делать более объективные выводы относительно генетической структуры населения России. Кроме того, еще можно ожидать большого количества редких генетических вариантов в остальной части населения, которые не были охвачены нашим анализом. Поэтому дальнейшее объединение данных из центров секвенирования генома по всей России, обследование большего числа здоровых доноров, а также включение в базы данных пациентов из разных регионов необходимы для более полной характеристики спектра генетических вариаций в населении современной России, чтобы обосновать предположения о распространенности популяционных частот аллелей моногенных заболеваний [Barbitoff *et al.,* 2019; 2021].

Важно отметить, что на основе популяционно-генетических работ появляется возможность не только оценить частоту того или иного заболевания, но и проанализировать распространенность некоторых вариантов, например, в гене *ACE2*, полиморфизм которого может играть роль в патологии COVID-19, влияя на важные функции этого белка в нормальных клетках. Так, проведенный нами сравнительный анализ частот пяти вариантов (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) в гене *ACE2* в русской и европейских популяциях установил, что русские по этим характеристикам сходны с другими

европейскими популяциями, что позволило предположить сходный уровень инфицирования и тяжести COVID-19 в данных регионах. Это было важно для понимания эпидемиологической ситуации в начале эпидемии в марте-апреле 2020 года [Shikov *et al.,* 2020].

Биоинформатическая обработка также является важнейшей частью технологии NGS. Поэтому установление генетической природы заболевания во многом зависит от качественного биоинформатического протокола анализа данных секвенирования [Barbitoff *et al.,* 2017; 2020]. Необходимо учитывать, что в референсной последовательности встречаются ошибки, связанные с так называемыми референсными минорными вариантами — RMA (позициями референсного генома, в которые инкорпорирован редкий или даже патогенный вариант). Такие ошибки необходимо корректировать при проведении биоинформатического анализа. Для коррекции данных ошибок была разработана специальная программа [Barbitoff *et al.,* 2018], которую мы впоследствии применяли для всех своих исследований.

Разработанные биоинформатические алгоритмы вместе с новыми подходами к интерпретации данных позволяют нам описывать не только известные, но также и новые генные варианты. С помощью молекулярно-генетического тестирования с использованием методов секвенирования следующего поколения найдены новые патогенные варианты в генах *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2* [Fedyakov *et al.,* 2019; Miroshnikova *et al.,* 2021; Glotov O. *et al.,* 2019; Balashova *et al.,* 2020]. Показана необходимость углубленного анализа клинической картины у пациентов с более чем одним генетическим вариантом в одном или разных генах-мишенях для синдрома MODY [Glotov O. *et al.,* 2019]. Исходя из проведенных исследований, нами предложены алгоритмы эффективной генетической диагностики для MODY, болезни Вильсона-Коновалова и других заболеваний [Fedyakov *et al.,* 2019; Miroshnikova *et al.,* 2021; Glotov O. *et al.,* 2019; Balashova *et al.,* 2020].

Еще одной важной особенностью использования результатов современных методов исследования является то, что мы можем выявлять сразу несколько наследственных заболеваний у одного человека. Например, описаны клинические случаи совместного наследования Х-сцепленной и аутосомно-доминантной форм ихтиоза [Alaverdian *et al.,* 2019], болезни Вильсона-Коновалова и гемохроматоза [Тулзуновская *и др.,* 2017; Balashova *et al.*, 2020]. Информацию о молекулярных дефектах, приводящих к развитию болезни, важно знать для клинического сопровождения пациента.

Сегодня ни одна сфера медицины не обходится без применения современных методов секвенирования ДНК. Так, важной задачей клинической медицины является решение проблем репродукции. Нами показана эффективность использования NGS для НИПТ и ПГД [Pendina *et al.,* 2019; Saifitdinova *et al.,* 2020]. Результаты исследований демонстрируют необходимость комплексного подхода с использованием всего арсенала молекулярно-генетических, цитогенетических, эмбриологических методов при планировании беременности. Так, успешность использования различных молекулярно-генетических технологий и возможностей их клинического применения продемонстрирована в нашем исследовании на примере семьи с наследственной патологией [Лязина *и др.,* 2017]. В данной работе представлен сложный и длительный путь клинико-генетического обследования и диагностики наследственных заболеваний. Данный пример демонстрирует необходимость внедрения ГКП и нового алгоритма преконцепционного обследования семей с использованием всего арсенала молекулярно-генетических методов, включая секвенирование нового поколения как метода первого звена при планировании беременности, а также метода ПГТ и НИПТ для последующего мониторинга беременности.

Большинство болезней не являются моногенными, поэтому, прежде чем оценивать риск заболевания, нужно установить его природу (моногенное, олигогенное или мультифакториальное состояние). Иногда ответить на этот

вопрос достаточно сложно. Мы попробовали это сделать в наших работах по изучению наследственной кардиомиопатии [Glotov *et al.,* 2015; Komissarova *et al.,* 2016], семейной гиперхолестеринемии [Miroshnikova *et al.,* 2021] и МОДИ [Glotov O. *et al.,* 2019]. Мы пришли к выводу, что последние успехи в секвенировании генома человека позволяют заключить, что олигогенными, строго говоря, можно назвать практически все наследственные болезни, в том числе и моногенные, в клиническое развитие которых существенный вклад вносят аллельные варианты многих других генов - так называемых генов-модификаторов, ассоциированных с данным заболеванием [Agarwaland Moorchung, 2005; Kousiand Katsanis, 2015]. Суммарный фенотип определяется вкладом определенных вариантов генов, их экспрессивностью и пенетрантностью.

Несколько сложнее - ситуация с МФЗ, поскольку в их этиологии важна роль изменений в геноме, а предрасположенность к болезни зависит от большого числа генов (феномен аддитивности). Предрасположенность к заболеванию реализуется под влиянием большого числа факторов внешней среды, и характер наследования не объясняется только менделевскими законами [Баранов и др. 2021]. Действительно, пациенты с одним и тем же диагнозом могут отличаться по факторам риска и этиологии.

Однако ситуация с оценкой риска мультифакториальных заболеваний сильно изменилась за последние пять лет. Производительность, разрешающая способность и стоимость секвенирования генома сегодня таковы, что мы можем регулярно применять эти технологии в больших масштабах. Это открывает беспрецедентные возможности для медицинской практики и изучения этиологии заболеваний [Franks*etal.,* 2021]. Для оценки риска МФЗ используют современные математические модели, в первую очередь линейную регрессию, учитывающую и генетические, и клинические параметры [Khera *et al.,* 2018].

Наше исследование показывает, что анализ всего экзома может служить рациональным подходом для выявления генетических маркеров сложного

заболевания даже в ограниченных выборках. Используя стратегию многоаспектного анализа, мы обнаружили несколько приемлемых локусов-кандидатов и SNP, которые могут играть важную роль в патогенезе СД2 и ожирения в населении России. В целом, рациональная фильтрация и ранжирование потенциально причинных вариантов может помочь в идентификации генов заболеваний по полигенно обусловленным клиническим признакам и демонстрирует эффективность технологий секвенирования всего экзома для поиска новых маркеров многофакторных заболеваний в когортах ограниченного размера в малоизученных популяциях [Barbitoff *et al.*, 2018].

Геномная медицина может также помочь выявить редкие состояния, которые скрыты в рамках сложной многоэтапной и многокомпонентной диагностики заболевания. Более того, для различных распространенных заболеваний были идентифицированы гены, в которых редкие варианты у гетерозиготных носителей повышают риск МФЗ в несколько раз. Примером является наличие вариантов риска семейной гиперхолестеринемии у 0,4% населения, которыеповышают риск развития ИБС в 3 раза [Abul-Husn *et al.*, 2016]. Поэтому именно использование экзомного секвенирования (ГКП) позволит ответить на вопросы оценки риска не только моногенных патологий, но и МФЗ.

В то время как выявление носителей редких моногенных мутаций требует секвенирования конкретных генов и тщательной интерпретации функциональных эффектов обнаруженных мутаций, полигенные оценки могут быть легко рассчитаны одновременно для многих заболеваний на основе данных из одного массива генотипирования [Khera *et al.*, 2018; Barbitoff *et al.*, 2018]. Возможность выявления лиц со значительно более высоким генетическим риском по широкому спектру распространенных заболеваний (диабет, сердечно-сосудистые заболевания и др.) в любом возрасте открывает ряд перспектив и проблем для клинической медицины.

Сегодня применение предлагаемых нами ГКП и генетической карты репродуктивного здоровья с использованием технологий NGS [Баранов *и др.*, 2000; 2021] включает следующие направления: исследование на носительство моногенных и олигогенных заболеваний, планирование беременности, дифференциальная диагностика и лечение, подтверждающая диагностика, а в будущем оно будет дополнено расчетом риска МФ и инфекционных заболеваний, описанием фенотипических особенностейу человека (рис. 33).

Суммируя вышесказанное, мы можем предполагать следующий порядок действий по внедрению геномной медицины [ChatterjeeandGarcía-Closas, 2016].

1. Выявление факторов риска - высококачественные эпидемиологические исследования с большими размерами выборки, точечными и объективными измерениями фенотипов и воздействий необходимы для выявления новых факторов риска (включая генетические вариации, факторы риска окружающей среды, биомаркеры и др.).

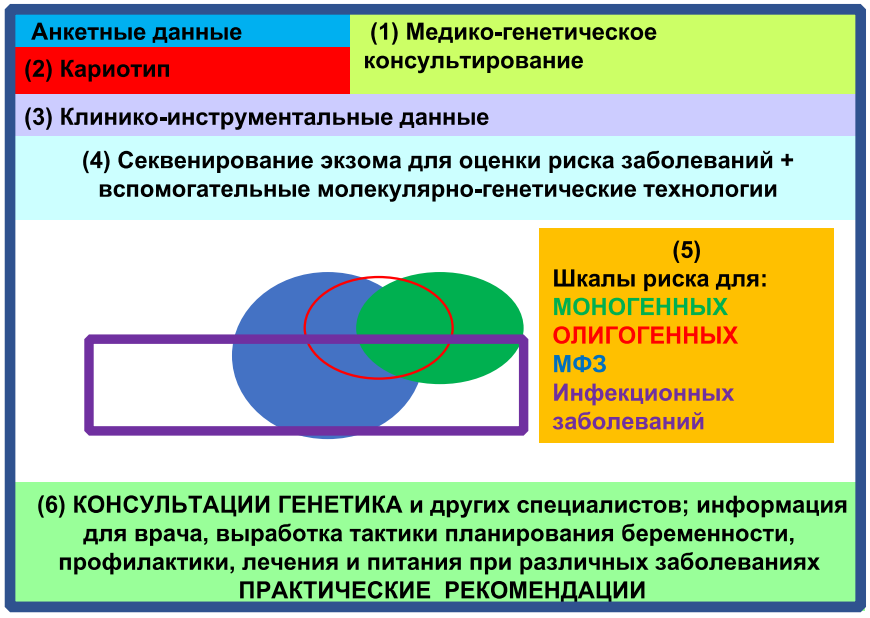


Рисунок 33. Блок-схема клинического генетического паспорта для задач репродукции.

2. Характеристика относительного риска - построение моделей относительного риска, которые объединяют информацию о многочисленных

факторах риска (включая оценки полигенного риска, факторы риска окружающей среды и их взаимодействие).

3. Оценка абсолютного риска - прогнозирование риска развития заболевания в течение определенного периода времени на основе факторов риска субъекта (с использованием моделей относительного риска, распределения факторов риска, общей возрастной заболеваемости и смертности в целевой популяции).

4. Оценка калибровки модели - сравнение количества прогнозируемых и наблюдаемых диагнозов заболеваний за определенный период времени в группах людей с различным прогнозируемым риском в проспективных когортных исследованиях.

5. Оценка полезности для общественного здравоохранения - оценка эффективности стратегий первичной и вторичной профилактики, разработанных в соответствии с уровнями прогнозируемого риска для людей.

Заключительным и не менее важным этапом является этический вопрос о том, как оценивать абсолютные, так и относительные риски и как информировать врача и пациента об этих рисках, например, поощрять изменения образа жизни или проводить дальнейший скрининг заболеваний. Это предстоит нам узнать в обозримом времени с учетом существующих положений о защите персональной информации.

Сегодня биобезопасность любого государства в значительной степени определяется уровнем развития фундаментальных и прикладных исследований в области иммунологии и изучения инфекционных заболеваний [Хаитов *и др.,* 2017]. Катализатором интереса к этим темам явилась пандемия новой коронавирусной инфекции.

К сожалению, ни новые антибиотики, ни новые вакцины принципиально не могут стать решением проблемы борьбы с инфекционными заболеваниями. Еще Л. Пастер говорил: «Микроб ничто, субстрат (т.е. человек) – все» [Полетаев, Чурилов, 2021]. Сегодня к этому высказыванию можно добавить:

вирус сегодня один, завтра - другой, а человек остается. Поэтому изучать нужно не только и не сколько сам вирус и его геном, но и геном его носителя (человека) и его иммунную систему. И.И. Мечников подчеркивал, что основное предназначение иммунной системы – это поддержание динамического постоянства молекулярно-клеточного состава организма, включая репарацию любых повреждений и участие в регуляции самых разных физиологических процессов [Полетаев, Чурилов, 2021].

Иммунная дисфункция с выраженной неконтролируемой генерализованной системной воспалительной реакцией в виде усиленной продукции воспалительных цитокинов, приводящих к цитокиновому шторму (ЦШ), является центральной проблемой в патофизиологии COVID-19. Полученные нами результаты позволяют обосновать выбор лечебной тактики с ранним назначением своевременной противовоспалительной терапии и антиковидной плазмы реконвалесцентов для пациентов с высоким риском развития ЦШ и согласуются с оценкой другими авторами факторов риска ЦШ при COVID-19 [Caricchio *et al.,* 2021; Moore *el al.,* 2020]. К основным факторам риска развития цитокинового шторма у больных COVID-19 относятся: мужской пол, возраст старше 40 лет, положительный тест на РНК SARS-CoV-2, лимфопения, уровни ЛДГ, D-димера, ферритина и IL-6, динамика индекса по шкале NEWS. Информация о клинико-инструментальных и лабораторных критериях риска короновирусной инфекции позволяет корректно сформировать группы риска для проведения генетических исследований и выяснения наследственных основ заболевания COVID-19 [Glotov *et al.,* 2021a; Shcherbak *et al.,* 2022].

Любые исследования генетической предрасположенности к вирусным инфекциям начинаются с изучения рецепторов клетки, необходимы для проникновения вируса в клетку хозяина. Важно отметить, что сравнительный анализ частот отдельных вариантов в гене *ACE2* не выявил каких-либо существенных различиймежду пациентами с легкой и тяжелой формой COVID-

19 [Shikov *et al.*, 2020]. Однако нами было показано, что редкие гаплотипы в гене *ACE2* избыточно представлены у российских пациентов с тяжелой формой COVID-19. Возможно, редкие варианты могут играть роль в патологии COVID-19, в норме влияя на важные функции белка [Shikov *et al.*, 2020]. Полученные данные говорят о необходимости более тщательного наблюдения пациентов с редкими патогенными вариантами, и не только в гене *ACE2*, но и в целом геномов случае заражения COVID-19.

Важно отметить, что эпидемия COVID-19 стимулировала не только изучение гена *ACE2* (рис. 34), но и широкомасштабные полногеномные исследования. Стало появляться все больше информации о том, что прогнозирование генетической восприимчивости к COVID-19 может помочь клиницистам выбрать оптимальные схемы лечения для пациентов [Prakrithi *et al.*, 2021].

Рисунок 34. Пандемия COVID-19 как двигатель анализа геномных данных.

Во многих исследованиях сообщается о нескольких десятках ассоциаций между генетическими вариантами и заболеваемостью, тяжестью и смертностью COVID-19 среди различных этнических групп [Suh *et al.*, 2022]. Несмотря на

большое количество зарегистрированных клинико-генетических ассоциаций, варианты, которые были обнаружены в одном исследовании, могут быть не подтверждены в другой выборке [Suh *et al.,* 2022]. Частично данный вопрос решают крупномасштабные мета-анализы, одним из которых является проект COVID-19 Host Genetics Initiative, 2021, где сегодня представлена большая часть современных данных по генетической предрасположенности к COVID-19 [COVID-19 Host Genetics Initiative].

Важно отметить, что поиск новых ассоциаций в малоизученных популяциях может быть затруднен из-за ограниченного размера когорты и/или в случае ограниченной широты охвата генома (т.е. в исследованиях, основанных на CES или таргетном секвенировании). Поэтому для решения данной проблемы мы осуществили оригинальный дизайн исследования [Shikov *et al.,* 2020; Shcherbak *et al.,* 2022], используя агрегацию подсчетов вариантов, как на уровне генов, так и на уровне метаболических путей. Это стратегия, аналогичная той, которую использовали Povysil с коллегами [Povysil *et al.,* 2021]. Мы выявили 11 независимых генетических вариантов, которые ассоциированы с количественными признаками, которые, в свою очередь, непосредственно связаны с тяжестью и исходом заболевания. Для трех из этих вариантов мы смогли найти подтверждающие доказательства их роли в патогенезе COVID-19. Однако уровень значимости этих ассоциаций недостаточен для того, чтобы сделать уверенный вывод о влиянии найденных вариантов на фенотип пациента. Необходимо дополнительное подтверждение наблюдаемых ассоциаций и подтверждение их истинной роли в патогенезе COVID-19 [Shcherbak *et al.,* 2022].

Учитывая различия в дизайне разных исследований, мы не ожидаем дальнейшего воспроизведения наших данных в ряде других работ. Тем не менее, полученные нами результаты демонстрируют полезность глубокого лабораторного фенотипирования пациентов с COVID-19 для выявления новых генетических вариантов, влияющих на тяжесть и/или исход заболевания.

Именно молекулярная медицина и ее основные направления (предиктивная медицина, генная терапия, фармакогеномика и др.) будут определять все многообразие фундаментальных и прикладных наук о человеке в XXI в., а возможно, и в третьем тысячелетии. Таким образом, концепция Предиктивной медицины - генетического клинического паспорта здоровья для решения задач преконцепционного скрининга, ПГД, рождения здорового потомства, постановки диагноза, профилактики МФЗ и инфекционных заболеваний - должна описаться на секвенирование нового поколения как базовый метод одновременном использовании специализированных собственных баз данных, алгоритмов и биоинформатики (рис. 35).

Рисунок 35. Исследование экзома в рамках концепция Предиктивной медицины - генетического клинического паспорта здоровья человека.

В заключение хочется подчеркнуть, что разработка научных основ точной медицины как для изучения, диагностики и лечения моногенных болезней, так и олигогенных, мультифакториальных и инфекционных заболеваний будет определяться эффективностью использования NGS технологий с учетом современных алгоритмов анализа и классических генетических понятий

экспрессивности и пенетрантности. Дальнейшие этапы практического внедрения генетического тестирования включают: досимптоматическое (упреждающее) генетическое тестирование (ГТ) в семьях высокого риска; проспективное ГТ с обязательным последующим мониторингом состояния лиц групп высокого риска по результатам тестирования; рандомизированное предиктивное тестирование [Barnoy, 2007, Баранов *и др.,* 2021].

ВЫВОДЫ

1. В обследованной популяции Северо-Западного региона России наиболее часто встречаются следующие моногенные болезни с рецессивным типом наследования: фенилкетонурия, недостаточность фактора VII4, синдром Элерса-Данлоса, кифосколиотический тип, 2, тирозиназонегативный кожно-глазной альбинизм и болезнь Вильсона-Коновалова.

2. В случайной выборке Северо-Западного региона России наиболее часто встречаются патогенные варианты ряда генов: *ABCA4* (дистрофия сетчатки, болезнь Штаргардта) и *CFTR* (муковисцидоз).

3. На примере популяции Северо-Западного региона России показано, что использование NGS-технологии повышает эффективность описания ранее неизвестных вариантов в мировой литературе до 25%, и поэтому целесообразно создание отечественных баз патогенных вариантов.

4. Сопоставление результатов NGS ДНК пациентов с моногенными заболеваниями по оригинальным биоинформатическим протоколам, как по международным базам данных, так и по анализу локальной популяции, позволило определить новые патогенные варианты в генах *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT* и выявлять патологические фенотипы, обусловленные их сочетанием.

5. Применение NGS, по сравнению с классическими молекулярно-генетическими методами на основе ПЦР для поиска патогенных вариантов у пациентов с моногенным сахарным диабетом и болезнью Вильсона-Коновалова увеличивает выявляемость патогенных вариантов в 1,5-2 раза.

6. В Северо-Западном регионе России вариативность проявления патологий как олигогенной, так и мультифакториальной природы обусловлена не только экспрессивностью, пенетрантностью, но и комплексными гаплотипами, ассоциированных с заболеваниями генов, выявленных с помощью экзомного

секвенирования и оригинального биоинформатического анализа, адаптированного для небольших когорт.

7. Для оценки клинических проявлений и прогнозирования фенотипа на основе данных о патогенных вариантах отдельных генов и их сочетаний, регрессионный метод является наиболее эффективным, в том числе для создания полигенных предикторов риска олигогенных и мультифакториальных заболеваний.

8. Варианты в генах *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX,* в том числе редкие варианты в гене *ACE2*, выявленные с помощью NGS, ассоциированы с тяжестью, оцененной в соответствии с оригинальной разработанной шкалой, и клиническим исходом новой коронавирусной инфекции COVID-19.

9. Впервые установлено, что частоты пяти вариантов (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) в гене *ACE2* не различаются в русской и европейских популяциях.

10. Разработан комплекс генетических обследований для использования в предиктивной медицине - «генетический клинический паспорт здоровья человека», включающий экзомное секвенирование и позволяющий предсказать возможное развитие олигогенной и мультифакториальной патологии, тяжесть протекания некоторых инфекционных заболеваний и объяснить патогенетический характер клинических проявлений болезней разной природы.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Glotov A.S., Kazakov S.V., Zhukova E.A., Alexandrov A.V., Glotov O.S., Pakin V.S., Danilova M.M., Tarkovskaya I.V., Niyazova S.Sh., Chakova N.N., Komissarova S.M., Kurnikova E.A., Sarana A.M., Sherbak S.G., Sergushichev A.A., Shalyto A.A., Baranov V.S. Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group // Clinica Chimica Acta, 2015, V.446, P.132–140. https://doi.org/10.1016/j.cca.2015.04.014

2. Komissarova S.M., Chakova N.N., Niyazova S.S., Kazakov S.V., Zhukova E.A., Aleksandrov A.V., Glotov O.S., Glotov A.S. The specifics of hypertrophic cardiomyopathy clinical presentation in patients with various mutations of sarcomere genes // Russian Journal of Cardiology (In Russ), 2016, V.1, P.20-25. https://doi.org/10.15829/1560-4071-2016-1-20-25

3. Bliznetz E., Lalayants M., Markova T., Balanovsky O., Balanovska E., Skhalyakho, R., Pocheshkhova E., Nikitina N., Voronin S., Kudryashova E., Glotov O., and Polyakov A. Update of the GJB2/DFNB1 mutation spectrum in Russia: a founder Ingush mutation del(GJB2-D13S175) is the most frequent among other large deletions // J Hum Genet, 2017, V.62, P.789–795. https://doi.org/10.1038/jhg.2017.42

4. Тулзуновская И.Г., Жученко Н.А., Балашова М.С., Филимонов М.И., Розина Т.П., Глотов О.С., Асанов А.Ю. Болезнь Вильсона-Коновалова: внутрисемейный клинический полиморфизм // Педиатрия. Журнал им. Г.Н. Сперанского, 2017, Т.96, №6, С.215-216. http://doi.org/10.24110/0031-403X-2017-96-6-215-216.

5. Barbitoff Y.A., Bezdvornykh I.V., Polev D.E., Serebryakova E.A., Glotov A.S., Glotov O.S., Predeus A.V. Catching hidden variation: systematic correction of reference minor alleles in clinical variant calling» // Genet. Med, 2018, V.20, P.360-364. http://doi.org/10.1038/gim.2017.168

6. Barbitoff Y.A., Serebryakova E.A., Nasykhova Y.A., Predeus A.V., Polev D.E., Shuvalova A.R., Vasiliev E.V., Urazov S.P., Sarana A.M., Scherbak S.G., Gladyshev D.V., Pokrovskaya M.S., Sivakova O.V., Meshkov A.N., Drapkina O.M., Glotov O.S., Glotov A.S. Identification of Novel Candidate Markers of Type 2 Diabetes and Obesity in Russia by Exome Sequencing with a Limited Sample Size // Genes, 2018,V.9(8), 415. https://doi.org/10.3390/genes9080415

7. Glotov O.S., Romanova O.V., Eismont Y.A., Sarana A.M., Scherbak S.G., Kuzmich E.V., Alyanskiy A.L., Ivanova N.E., Teplyashina V.V., Serov Y.A., Zubarovskaya L.S., Afanasyev B.V. Comparative analysis of NGS and Sanger sequencing methods for HLA typing at a Russian university clinic // Cellular Therapy and Transplantation (CTT), 2018, Vol.7, №4(25), P.72-82. http://doi.org/10.18620/ctt-1866-8836-2018-7-4-72-82

8. Glotov O.S., Serebryakova E.A., Turkunova M.S., Efimova O.A., Glotov A.S., Barbitoff Y.A., Nasykhova Y.A., Predeus A.V., Polev D.E., Fedyakov M.A., Polyakova I.V., Ivashchenko T.E., Shved N.Yu., Shabanova E.S., Romanova O.M., Sarana A.M., Pendina A.A., Scherbak S.G., Musina E.V., Petrovskaya-Kaminskaya A.V., Lonishin L.R., Ditkovskaya L.V., Zhelenina L.A., Tyrtova L.V., Berseneva O.S., Suspitsin E.N., Bashnina E.B., Baranov V.S. Whole-exome sequencing for monogenic diabetes in Russian children reveals wide spectrum of genetic variants in MODY-related and unrelated genes // Molecular Medicine Reports,2019, V.20, №6, 4905-4914. https://doi.org/10.3892/mmr.2019.10751

9. Balashova M.S., Tulzunovskaya I.G., Glotov O.S., Glotov A.S., Barbitoff Y.A., Fedyakov M.A., Alaverdian D.A., Ivashchenko T.E., Romanova O.V., Sarana A.M., Scherbak S.G., Baranov V.S., Filimonov M.I., Skalny A.V., Zhuchenko N.A., Ignatova T.M., Asanov A.Y. The spectrum of pathogenic variants of the ATP7B gene in Wilson disease in the Russian Federation // J Trace Elem Med Biol, 2020, V.59, 126420. https://doi.org/10.1016/j.jtemb.2019.126420

10. Barbitoff Y.A., Skitchenko R.K., Poleshchuk O.I., Shikov A.E., Serebryakova E.A., Nasykhova Y.A., Polev D.E., Shuvalova A.R., Shcherbakova I.V., Fedyakov M.A., Glotov O.S., Glotov A.S., Predeus A.V. Whole exome sequencing provides insights into monogenic disease prevalence in Northwest Russia // Mol Genet Genomic Med, 2019, V.7(11), e964. https://doi.org/10.1002/mgg3.964

11. Pendina A.A., Shilenkova Y.V., Talantova O.E., Efimova O.A., Chiryaeva O.G., Malysheva O.V., Dudkina V.S., Petrova L.I., Serebryakova E.A., Shabanova E.S., Mekina I.D., Komarova E.M., Koltsova A.S., Tikhonov A.V., Tral T.G., Tolibova G.K., Osinovskaya N.S., Krapivin M.I., Petrovskaia-Kaminskaia A.V., Korchak T.S., Ivashchenko T.E., Glotov O.S., Romanova O.V., Shikov A.E., Urazov S.P., Tsay V.V., Eismont Y.A., Scherbak S.G., Sagurova Y.M., Vashukova E.S., Kozyulina P.Y., Dvoynova N.M., Glotov A.S., Baranov V.S., Gzgzyan A.M. and Kogan I.Y. Reproductive History of a Woman With 8p and 18p Genetic Imbalance and Minor Phenotypic Abnormalities // Front. Genet, 2019, V.10, 1164. http://doi.org/10.3389/fgene.2019.01164

12. Alaverdian D.A., Fedyakov M., Polennikova E., Ivashchenko T., Shcherbak S., Urasov S., Tsay V., Glotov O.S. X-linked and autosomal dominant forms of the ichthyosis in coinheritance // Drug Metabolism and Personalized Therapy, 2019, V.34, №4, 20190008. https://doi.org/10.1515/dmpt-2019-0008

13. Barbitoff Y.A., Polev D.E., Shcherbakova I.V., Serebryakova E.A., Kiselev A.M., Kostareva A.A., Glotov O.S., Predeus A.V. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage // Sci Rep, 2020, V.10, 2057 https://doi.org/10.1038/s41598-020-59026-y

14. Shikov A.E., Barbitoff Y.A., Glotov A.S., Danilova M.M., Tonyan Z.N., Nasykhova Y.A., Mikhailova A.A., Bespalova O.N., Kalinin R.S., Mirzorustamova A.M., Kogan I.Y., Baranov V.S., Chernov A.N., Pavlovich D.M., Azarenko S.V., Fedyakov M.A., Tsay V.V., Eismont Y.A., Romanova O.V., Hobotnikov D.N.,

Vologzhanin D.A., Mosenko S.V., Ponomareva T.A., Talts Y.A., Anisenkova A.U., Lisovets D.G., Sarana A.M., Urazov S.P., Scherbak S.G. and Glotov O.S. Analysis of the Spectrum of ACE2 Variation Suggests a Possible Influence of Rare and Common Variants on Susceptibility to COVID-19 and Severity of Outcome // Front. Genet, 2020, V.11, 551220. http://doi.org/ 10.3389/FGENE.2020.551220

15. Goloshchapov O.V., Bakin E.A., Kucher M.A., Stanevich O.V., Suvorova M.A., Gostev V.V., Glotov O.S., Eismont Yu.A., Polev D.E., Lobenskaya A.Yu., Klementeva R.V., Goloshchapova M.O., Zubarovskaya L.S., Sidorenko S.V., Suvorov A.N., Moiseev I.S., Chukhlovin A.B. Bacteroides fragilis is a potential marker of effective microbiota transplantation in acute graft -versus-host disease treatment // Cellular Therapy and Transplantation (CTT), 2020, V.9(2), P.47-59. doi: 10.18620/ctt-1866-8836-2020-9-2-47-59

16. Miroshnikova V.V., Romanova O.V., Ivanova O.N., Fedyakov M.A., Panteleeva A.A., Barbitoff Y.A., Muzalevskaya M.V., Urazgildeeva S.A., Gurevich V.S., Urazov S.P., Scherbak S.G., Sarana A.M., Semenova N.A., Anisimova I.V., Guseva D.M., Pchelina S.N., Glotov A.S., Zakharova E.Y., Glotov O.S. Identification of novel variants in the LDLRgene in Russian patients with familial hypercholesterolemia using targeted sequencing // Biomedical Reports,2021, V.14.1, 15. http://doi.org/10.3892/BR.2020.1391

17. Glotov O.S., Chernov A.N., Scherbak S.G. and Baranov V.S. Genetic Risk Factors for the Development of COVID-19 Coronavirus Infection // Russian Journal of Genetics, 2021, V.57, №8, P.878–892. http://doi.org/10.1134/S1022795421080056

18. Shikov A., Tsay V., Fedyakov M., Eismont Y., Rudnik A., Urasov S., Sherbak S., and Glotov O. The application of Nanopore sequencing for variant calling on the human mitochondrial DNA // Bio. Comm., 2021, V.66(2), P.109–123. https:// doi.org/10.21638/spbu03.2021.202

19. Shcherbak S.G., Anisenkova A.Y., Mosenko S.V., Glotov O.S., Chernov A.N., Apalko S.V., Urazov S.P., Garbuzov E.Y., Khobotnikov D.N., Klitsenko O.A., Minina E.M. and Asaulenko Z.P. Basic Predictive Risk Factors for Cytokine Storms in COVID-19 Patients // Front. Immunol,2021, V.12, 745515. http://doi.org/ 10.3389/fimmu.2021.745515

20. Glotov O.S., Chernov A.N., Korobeynikov A.I., Kalinin R.S., Tsai V.V., Anisenkova A.Yu., Urazov S.P., Lapidus A.L., Mosenko S.V., Shcherbak S.G. The lineage of coronavirus SARS-CoV-2 of Russian origin: Genetic characteristics and correlations with clinical parameters and severity of coronavirus infection // The Siberian Journal of Clinical and Experimental Medicine (In Russ.), 2021, V.36(4), P.132–143. https://doi.org/10.29001/2073-8552-2021-36-4-132-143

21. Shcherbak S.G., Changalidi A.I., Barbitoff Y.A., Anisenkova A.Y., Mosenko S.V., Asaulenko Z.P., Tsay V.V., Polev D.E., Kalinin R.S., Eismont Y.A., Glotov A.S., Garbuzov E.Y., Chernov A.N., Klitsenko O.A., Ushakov M.O., Shikov A.E., Urazov S.P., Baranov V.S., Glotov O.S. Identification of Genetic Risk Factors of Severe COVID-19 Using Extensive Phenotypic Data: A Proof-of-Concept Study in a Cohort of Russian Patients // Genes, 2022, V.13(3), 534. https://doi.org/10.3390/genes13030534

22. Turkunova M.E., Barbitoff Y.A., Serebryakova E.A., Polev D.E., Berseneva O.S., Bashnina E.B., Baranov V.S., Glotov O.S. and Glotov A.S. Molecular Genetics and Pathogenesis of the Floating Harbor Syndrome: Case Report of Long-Term Growth Hormone Treatment and a Literature Review // Front. Genet, 2022, V.13, 846101. https://doi.org/10.3389/fgene.2022.846101

23. Glotov A.S., Zelenkova I.E., Vashukova E.S., Shuvalova A.R., Zolotareva A.D., Polev D.E., Barbitoff Y.A., Glotov O.S., Sarana A.M., Shcherbak S.G., Rozina M.A., Gogotova V.L., Predeus A.V. RNA Sequencing of Whole Blood Defines the Signature of High Intensity Exercise at Altitude in Elite Speed Skaters // Genes, 2022, V.13(4), 574. https://doi.org/10.3390/genes13040574

24. Koshevaya Y.S., Kusakin A.V., Buchinskaia N.V., Pechnikova V.V., Serebryakova E.A., Koroteev A.L., Glotov A.S., and Glotov O.S. Description of first registered case of the Lopes-Maciel-Rodan syndrome in Russia // Int. J. Mol. Sci, 2022, V.23, 12437. https://doi.org/10.3390/ijms232012437

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

АГ	Артериальная гипертензия
АД	Аутосомно-доминантный
АТФ	Аденозинтрифосфорная кислота
БВК	Болезнь Вильсона-Коновалова
ВМП	Высокие медицинские технологии
ВОГиС	Вавиловское общество генетиков и селекционеров
ВОЗ	Всемирная организация здравоохранения
ВРТ	Вспомогательные репродуктивные технологии
ГКП	Генетический клинический паспорт здоровья человека
ГЛГ	гемофагоцитарного лимфогистиоцитоза
ГП	Генетический паспорт
ГТ	Генетическое тестирование
ГФС	Гемофагоцитарный синдром
ДИ	Доверительный интервал
ДЛКЛ	Синдром дефицита лизосомной кислой липазы
ДНК	Дезоксирибонуклеиновая кислот
ЖЕЛ	Жизненная емкость легких
ИБС	Ишемическая болезнь сердца
ИМТ	Индекс массы тела
КТ	Компьютерная томография
ЛДГ	Лактатдегидрогеназа
ЛПВП	Липопротеины высокой плотности

ЛПНП	Липопротеины низкой плотности
ЛПОНП	Липопротеины очень низкой плотности
МОДИ	Моногенный диабет
МФ	Мультифакторное
МФЗ	Мультифакторное заболевание
мРНК	Матричная РНК
НИПТ	Неинвазивный пренатальный тест
НСД	Неонатальный сахарный диабет
ОРДС	Острый респираторный дистресс-синдром
ОМС	Обязательное медицинское страхование
ОТ	Объем талии
ПГТ	Пренатальное генетическое тестирование
ПГТ-А	Пренатальное генетическое тестирование хромосомной патологии
ПГД	Предимплантационная генетическая диагностика
ПДРФ	Полиморфизм длины рестрикционных фрагментов
ПМ	Предиктивная, превентивная, персонифицированная медицина
ПЦР	Полимеразная цепная реакция
ПЦР-РВ	Полимеразная цепная реакция в реальном времени
п.н.	Пар нуклеотидов
РНК	Рибонуклеиновая кислота
САМ	Синдром активации макрофагов
СГХ	Семейная гиперхолестеринемия
СД	Сахарный диабет

CMT	Болезнь Шарко-Мари-Зуба
СЗ	Сердечно-сосудистые заболевания
СРБ	С-реактивный белок
ТЕ	Трофэктодерма
ХОБЛ	Хроническая обструктивная болезнь легких
ТС	Триглицериды
УЗИ	Ультразвуковое исследование
ХС	Холестерин
ЦШ	Цитокиновый шторм
ЭКО	Экстракорпоральное оплодотворение
ACMG	American College of Medical Genetics and Genomics
AD	Autosomal dominant, аутосомно-доминантный
AF	Frequency of population alleles, частота популяционных аллелей
ADNSHL	Аутосомно-доминантная несиндромная наследственная потеря слуха
AFR	African population, африканская популяция
AIC	Информационный критерий Акайке
AMR	American Mixed population, смешанная американская популяция
ARPKD	Аутосомно-рецессивный поликистоз почек
ASJ	The population of Ashkenazi Jews, популяцияевреев-ашкенази
AUC	AreaUnderCurve, площадь под кривой
ARVC	Arrhythmogenic cardiomyopathy/right ventricular dysplasia, аритмогеннаякардиомиопатия/дисплазияправогожелудочка
CES	Таргетное секвенирование
ClinVar	ClinVardatabase, база клинических вариантов

CNV	Copy number variation, вариация числа копий
CoV	Коронавирусы
COVID-19	Новая кононавирусная инфекция
COVID-19 HG	COVID-19 Host Genetics Initiative project, проект по изучению генома COVID-19
dbNSFP	dbNSFP database, база данных фнкциональной аннотации вариантов
dbSNP	dbSNP database, база данных геномных вариантов
DECIPHER	DECIPHER database, базаданных DECIPHER
EAS	East Asian population, восточноазиатскаяпопуляция
EPMA	Европейское общество предиктивной, превентивной и персонализированной медицины
EVS	Exome variant sequencing, база данных вариантов, найденных вовремя экзомного секвенирования
ExAC	Exome Aggregation Consortium, экзомный консорциум
FDA	Food and Drug Administration, управление по санитарному надзору за качеством пищевых продуктов и медикаментов
FHS	Floating Harbor syndrome, синдромFloating Harbor
FIN	Finnish population, финская популяция
GATK	Genome Analysis ToolKit, протокол обработки геномных данных
GLM	Общая линейная модель
gnomAD	gnomAD database, базаданных gnomAD
GPS	Genome-wide polygenic score, общегеномная полигенная оценка риска
GRC	Genome Reference Consortium, консорциум исследования генома
GRCh38	Версия генома человека 38

GTEx	Genotype Tissues Expression database, база данных экспрессии генотипов в тканях
GWAS	Genome-wide association studies, полногеномные ассоциативные исследования
HapMap	Haplotype Map, карта генетических вариантов
HCM	гипертрофической кардиомиопатией
HGP	The Human Genome Project, проект геном человека
HGVS	Human Genome Variation Society, общество генетических вариантов
HLA	Human Leukocyte Antigens, антигены тканевой совместимости
HUGE	HUGE database, базаданных HUGE
IIBDGC	International Inflammatory Bowel Disease Genetics Consortium, консорциум по изучению воспалительных заболеваний кишечника
ICM	Internal cell mass, внутреннея клеточная масса
IRNT	Подход по преобразования нормализации на основе обратного ранга
LD	Linkage disequilibrium, неравновесиепо сцеплению
MAF	Minor allele frequency, частотаминорной аллели
MERS-CoV	Ближневосточный респираторный синдром
NCBI	National Center for Biotechnological Information, национальный центр биотехнологической информации США
NEWS	National Early Warning Score, индивидуальная шкала ранней оценки исхода
NFE	NonFinnish population, не финская популяция Европы
NGS	Next generation sequencing, секвенирование нового поколения
NHLBI	NHLBI exome project, проект по секвенированию экзома национального института сердца легких и крови США

NMIIC	Немышечный миозин II C
NPV	Negative Predicted Values, вероятность не заболеть при отсутствии маркера
NWR	North-West region population, популяция Северо-Запада РФ
OMIM	OMIM database, базаданных OMIM
OR	Соотношение шансов
ORF	Open reading frame, открытая рамка считывания
QTLs	Quantitative Trait Loci, локусы количественных признаков
pLoF	Putative loss-of-function, варианты с предполагаемой потерей функции
PCA	Principal component analysis, анализглавныхкомпонент
PheWAS	Phenome-wide association study database, базаданныхфенотипа
PPV	Positive Predicted Values, вероятность наличия заболевания при положительном результате теста
PRSs	Polygenic risk score, полигенныеоценкириска
RMA	Reference minor allele, референсный минорныйаллель (вариант)
RNA-seq	Данные секвенирования РНК
ROC	Receiver operating characteristic, ROC-кривая
RUSeq	Российский консорциум по секвенированию
SARS-CoV	Острый тяжелый респираторный синдром
SARS-CoV-2	Вирус COVID-19
SAS	South Asian population, южноазиатскаяпопуляция
SCD	Sudden cardiac death, внезапнаясердечнаясмерть

SNP	Single nucleotide polymorphism, однонуклеотидный полиморфизм
SNPSIFT	Genomic variant annotations and functional effect prediction toolbox, программа по обработки геномных вариантов и функциональных эффектов
SNV	Single-nucleotide variants, Однонуклеотидные варианты
TS	Targeted sequencing, таргетное секвенирование
WES	Whole-exomesequencing, полноэкзомноесеквенирование
WGS	Whole-genome sequencing, полногеномное секвенирование

СПИСОК ЛИТЕРАТУРЫ

1. Аульченко Ю.С. Разработка и применение методов полногеномного анализа генетических ассоциаций сложных признаков: дис. док. биол. наук: 03.02.07/ Аульченко Юрий Сергеевич. - Н., 2010.- 291 с.

2. Баранов В.С. *и др.* Геном человека и гены «предрасположенности» (Введение в предиктивную медицину) // СПб. Издательство «Интермедика». 2000. 272 с.

3. Баранов В.С., Хавинсон В.Х. Определение генетической предрасположенности к наследственным и мультифакториальным заболеваниям. Генетический паспорт // Методические рекомендации. СПб.: ИКФ «Фолиант». 2001. 48с.

4. Баранов В.С., Баранова Е.В., Иващенко Т.Э. Геном человека как научная основа предиктивной медицины. Геномика – медицине // ред. Иванов В.И. и Киселев Л.Л., Москва, Академкнига. 2005. С. 361-380.

5. Баранов В.С. *и др.* Генетический паспорт основа индивидуальной и предиктивной медицины. Под ред. В. С. Баранова — СПб.: «Изд-во Н-Л», ООО. 2009. 527 с.

6. Баранов В.С. *и др.* Определение наследственной предрасположенности к некоторым частым заболеваниям при беременности. Генетическая карта репродуктивного здоровья. Методические рекомендации / Санкт-Петербург, 2009. Сер. Ex libris «Журнал акушерства и женских болезней». 68 с.

7. Баранов В.С., Баранова Е.В. Генетический паспорт: состояние проблемы сегодня и завтра // Вестн. Росздравнадзора. 2018. №6. С.16-23.

8. Баранов В.С. *и др.* Эволюция предиктивной медицины /под. Ред. В.С. Баранова. – Санкт-Петербург: Эко-Вектор. 2021. 359 с

9. Бархатов И.М. *и др.* Секвенирование нового поколения и области его применения в онкогематологии // Онкогематология. 2016. Т. 11. С. 56-63.

10. Власов В.В. Эпидемиология - М.: ГЭОТАР – МЕД. 2004. 464 с.

11. Временные методические рекомендации: профилактика, диагностика и лечение новой коронавирусной инфекции (COVID-19). Версия 12, М, 21.09.2021. М.: 2021. 231 с.

12. Глотов А.С. *и др.* Зависимость между возникновением стабильной артериальной гипертензии у детей и полиморфизмом генов ренин-ангиотензиновой и кинин-брадикининовой систем // Молекулярная биология. – 2007. Т. 41. No 1. С. 18–25.

13. Глотов А.С. *и др.* Исследование молекулярно-генетических маркеров роста человека // Экологическая генетика. 2012. Т. 10. №. 4. С. 77–84.

14. Глотов А.С. Генетические и средовые факторы риска развития гестоза у женщин, артериальной гипертензии и метаболического синдрома у детей. Автореф. дис. докт. биол. наук. СПб. 2017. 34с.

15. Глотов О.С. Анализ полиморфизма генов сердечно-сосудистой системы и системы детоксикации в различных возрастных группах Санкт-петербурга. дис. канд. биол. наук. СПб. 2007. 188с.

16. Глотов О.С. *и др.* Исследование функционально – значимого полиморфизма *ACE, AGTR1, ENOS, MTHFR, MTRR и APOE* генов в популяции Северо-Западного региона России // Экологическая генетика. 2004. Т. 2. №. 3. С. 32-35.

17. Иващенко Т.Э., Баранов В.С. Биохимические и молекулярно-генетические основы патогенеза муковисцидоза. СПб.: Интермедика. 2002. 256 с.

18. Иващенко Т.Э. *и др.* Молекулярно-генетические методы // в кн: Медицинские лабораторные технологии: руководство по клинической лабораторной диагностике: в 2 т. // под ред. А. И. Карпищенко. 3-е изд., перераб. и доп. Т. 2. М.: ГЭОТАР-Медиа. 2013. С. 658-687.

19. Инге-Вечтомов С.Г. Генетика с основами селекции, М.: Высшая школа. 1998. 592 с.

20. Инге-Вечтомов С.Г. Генетика с основами селекции. СПб.: Изд-во Н-Л. 2010. 720 с.

21. Лязина Л.В. *и др.* Возможности оказания медицинской помощи в современных условиях на примере семьи с наследственной патологией // Медицинская генетика. 2017. Т. 16. №. 10. С. 51-54.

22. Мандельштам М.Ю. *и др.* Молекулярная генетика семейной гиперхолестеринемии: современное состояние вопроса в России // Тихоокеанский медицинский журнал. 2002. №. 1(8).С. 10-11.

23. Масленников А.Б. Взаимосвязь аллельных вариантов генов APOA1, APOB, APOC1, атерогенных дислипопротеидемий и осложненного течения инфаркта миокарда. Автореф. дис. канд. мед. наук. Томск. 1999. 26 с.

24. Отева Э.А., Масленников А.Б., Николаева А.А. Ускоренное развитие атеросклероза // Врач. 1994. №. 3. С. 50-52.

25. Полетаев А.Б., Чурилов Л.П. Иммунология здоровья и болезни: простые ответы на сложные вопросы. СПБ: Фолиант. 2021. 264 с.

26. Пузырев В.П. Феномно-геномные отношения и патогенетика многофакторных заболеваний // Вестн. РАМН. 2011. Т. 9. С. 17-27.

27. Пузырев В.П., Фрейдин М.Б., Кучер А.Н. Генетическое разнообразие народонаселения и болезни человека. Томск: Печатная литература. 2007. 319 с.

28. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA // М. МедиаСфера. 2003. 312 с.

29. Рыжкова О.П. *и др.* Руководство по интерпретации данных, полученных методами массового параллельного секвенирования (MPS). Медицинская генетика. 2017. Т. 16(7). С. 4-17.

30. Рыжкова О.П. *и др.* Руководство по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования (MPS) (редакция 2018, версия 2). Медицинская генетика. 2019. Т. 18(2). С. 3-23.

31. Сингер М., Берг П. Гены и геномы: В 2-х томах. Том 1. Пер. с англ. М.: Мир. 1998. 391 с.

32. Степанов В.А. Этногеномика неселения Северной Евразии // Томск: Изд-во «Печатная мануфактура». 2002. 244 с.

33. Степанов В.А., Пузырев В.П. Анализ аллельных частот семи микросателлитных локусов Y-хромосомы в трех популяциях тувинцев // Генетика. 2000. Т. 36. С. 241-248.

34. Тайц Б.М. Практическая предиктивная, превентивная и персонализированная медицина. «10П медицина в решении вопросов профилактики, активного долголетия, снижения смертности и увеличения продолжительности жизни населения // М-во здравоохранения Российской Федерации, Федеральное гос. бюджет. образов. Учреждение высшего образования «Северо-Западный гос. мед. ун-т им. И.И. Мечникова». Санкт-Петербург: ИПК Береста. 2019. 380 с.

35. Тарковская И.В. *и др.* Анализ ассоциации полиморфизма генов метаболизма липидов с индексом массы тела, обхватом талии и параметрами липидограммы крови у женщин. // Экологическая генетика. 2012. Т. 10. №. 4. С. 66-77.

36. Тулзуновская И.Г. *и др.* Болезнь Вильсона-Коновалова: внутрисемейный клинический полиморфизм // ПЕДИАТРИЯ. ЖУРНАЛ ИМ. Г.Н. СПЕРАНСКОГО. 2017. Т. 96. №. 6. С. 215-216.

37. Федяков М.А. *и др.* Анаукзетическая дисплазия: клиника, молекулярно-генетическая диагностика и лечение // Молекулярно-биологические технологии в медицинской практике / Под ред. чл.-корр. РАЕН А.Б. Масленникова. Вып. 32. Новосибирск: Академиздат. 2021. С. 81-92.

38. Хуснутдинова Э.К. *и др.* Рестрикционно-делеционный полиморфизм V-области митохондриальной ДНК в популяциях народов Волго-Уральского региона // Генетика. 1997. Т. 33. С. 996-1000.

39. Шиков А.Е. *и др.* Применение биоинформатики в анализе клинических данных// Молекулярно-биологические технологии в медицинской практике / Под ред. чл.-корр. РАЕН А.Б. Масленникова. Вып. 29. Новосибирск: Академиздат. 2019. С. 119-136.

40. Abe K. *et al.* Pro108Ser mutant of SARS-CoV-2 3CLpro reduces the enzymatic activity and ameliorates COVID-19 severity in Japan // medRxiv and bioRxiv. 2021. doi: https://doi.org/10.1101/2020.11.24.20235952

41. Abramov D.D. *et al.* High carrier frequency of *CFTR* gene mutations associated with cystic fibrosis, and *PAH* gene mutations associated with phenylketonuria in Russian population // Bulletin of Russian State Medical University. 2015. V. 4. P. 32– 35.

42. Abramov D.D. *et al.* Carrier frequency of *GJB2* and *GALT* mutations associated with sensorineural hearing loss and galactosemia in the Russian population // Bulletin of Russian State Medical University. 2017. V. 6. P. 20– 23.

43. Abramov A., Schorr S., Wolman M. Generalized xanthomatosis with calcified adrenals // Am J Dis Child. 1956. V.91(3). P. 282–286.

44. Abul-Husn N.S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system // Science. 2016. V. 354.

45. Agarwal S., Moorchung N. Modifier genes and oligogenic disease // J. Nippon Medical School. 2005. V. 72. N 6. P. 326–334.

46. Alaverdian D.A. *et al.* X-linked and autosomal dominant forms of the ichthyosis in coinheritance // Drug Metab Pers Ther. 2019.

47. Allen H.L. *et al.*Hundreds of variants clustered in genomic loci and biological pathways affect human height // Nature. 2010. V. 467. № 7317. C. 832-838.

48. AllSeq. WGS vs. WES. Available at: http://allseq.com/kb/wgsvswes/ [accessed November 16, 2018].

49. Almutawa W. *et al.* The R941L mutation in *MYH14* disrupts mitochondrial fission and associates with peripheral neuropathy // EBioMedicine. 2019.V. 45. P. 379–92.

50. Anisenkova A. *et al.* Immunoinformatics in COVID-19 Vaccine Development: The Role of HLA System // Cellular Therapy and Transplantation (CTT). V. 10.№. 1. 2021.

51. Ashour M.H. *et al.* Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks // Pathogens. 2020. V. 9. № 3. P. 186.

52. Aslanidis C. *et al.* Genetic and biochemical evidence that CESD and Wolman disease are distinguished by re sidual lysosomal acid lipase activity // Genomics. 1996. V. 33(1). P. 85–93.

53. Asselta R. *et al. ACE2* and *TMPRSS2* variants and expression as candidates to sex and country differences in COVID-19 severity in Italy // Aging (Albany NY). 2020. V. 12. №. 11. P. 10087–98.

54. Auton A. *et al.* A Global Reference for Human Genetic Variation // Nature. 2015. V. 526 (7571). P. 68–74.

55. Balashova M.S. *et al.* The spectrum of pathogenic variants of the *ATP7B* gene in Wilson disease in the Russian Federation // J Trace Elem Med Biol. 2020. V. 59. P. 126420.

56. Baranov A.A. *et al.* Deficiency of lysosomic acid lipase: clinical recommendations for child health care delivery // Pediatric pharmacology. 2016. V. 13(3). P. 239–243.

57. Barbetti F. and D'Annunzio G. Genetic causes and treatment of neonatal diabetes and early childhood diabetes // Best Pract Res Clin Endocrinol Metab. 2018. V. 32. P. 575-591.

58. Barbitoff Y.A. *et al.* Catching hidden variation: systematic correction of reference minor alleles in clinical variant calling» // Genet. Med. 2018. V. 20. P. 360-364.

59. Barbitoff Y.A. *et al.* Identification of Novel Candidate Markers of Type 2 Diabetes and Obesity in Russia by Exome Sequencing with a Limited Sample Size // Genes. 2018. V. 9(8). P. 415.

60. Barbitoff Y.A. *et al.* Whole exome sequencing provides insights into monogenic disease prevalence in Northwest Russia // Mol Genet Genomic Med. 2019 V. 7(11). e964.

61. Barbitoff Y.A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage // Sci Rep. 2020. V. 10 P. 2057.

62. Barbitoff Y.A. *et al.* Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples // medRxiv preprint.

63. Barnoy S. Genetic testing for late-onset diseases: effect of disease controllqability, test predictivity, and gender on the decision to take the test // Genetic testing. 2007. V. 11. №. 2. P. 187-193.

64. Beatrijs L.P. *et al.* Strengers and Louis J. Bont. Down Syndrome: A Novel Risk Factor for Respiratory Syncytial Virus Bronchiolitis — A Prospective Birth-Cohort Study // Pediatrics. 2007. V. 120. №. 4. e1076–e1081.

65. Benetti E. *et al. ACE2* gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population // Eur. J. Hum. Genet. 2020.

66. Bennett K. *et al.* Four novel cases of permanent neonatal diabetes mellitus caused by homozygous mutations in the glucokinase gene // Pediatr Diabetes. 2011. V. 12. P. 192-P196.

67. Bernal J.L. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant // N Engl J Med. 2021. V. 385. P. 585-594.

68. Bick A.G. *et al.* Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes // Nature. 2020. V. 586. №. 7831. P.763–68.

69. Biesecker L.G., Green R.C. Diagnostic clinical genome and exome sequencing // N Engl J Med. 2014. V. 370 (25). P. 2418-25.

70. Blanco-Melo D. *et al.* Imbalanced Host Response to SARSCoV-2 Drives Development of COVID-19 // Cell. 2020. V. 181(5). P. 1036–1045.

71. Bliznetz E. *et al.* Update of the *GJB2/DFNB1* mutation spectrum in Russia: a founder Ingush mutation del (*GJB2*-D13S175) is the most frequent among other large deletions // J Hum Genet. 2017. V. 62. P. 789–795.

72. Boomsma D.I. *et al.* The Genome of the Netherlands: Design, and Project Goals // European Journal of Human Genetics. 2014. V. 22 (2). P. 221–27.

73. Caccuri F. *et al.* A persistently replicating SARS-CoV-2 variant derived from an asymptomatic individual // J Transl Med. 2020. V. 18. P. 362.

74. Cao Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor *ACE2* in different populations // Cell Discov. 2020. V. 6. 11.

75. Capriotti E., Fariselli P. and Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure // Nucleic Acids Res. 2005. V. 33. P. 306-310.

76. Carere D.F. *et al.* Consumers report lower confidence I their genetics knowledge fooling direct-to-consumer personal genomic testing // Genet. Med. 2016. 2016. Vol. 18. № 1. P. 65-72.

77. Caricchio R. *et al.* Preliminary predictive criteria for COVID-19 cytokine storm // Ann Rheum Dis. 2020.

78. Caricchio R. *et al.* Preliminary Predictive Criteria for COVID-19 Cytokine Storm // Ann Rheum Dis. 2021. V. 80(1). P. 88–95.

79. Cascella M. *et al.* Features, Evaluation and Treatment Coronavirus (COVID-19) // StatPearls. 2020. PMID: 32150360.

80. Caso F. *et al.* Could Sars-coronavirus-2 trigger autoimmune and/or autoinflammatory mechanisms in genetically predisposed subjects? // Autoimmun Rev. 2020. V. 19(5). P. 102524.

81. Chatterjee N., Shi J., García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention // Nat Rev Genet. 2016. V. 17(7). P. 392-406.

82. Chen J. *et al.* Individual variation of the SARS-CoV2 receptor *ACE2* gene expression and regulation // Aging Cell. 2020. V.19. e13168.

83. Chen Y., Guo Y., Pan Y., Zhao Z.J. Structure analysis of the receptor binding of 2019-nCoV // Biochem. Biophys. Res. Commun. 2020. V. 525. №. 1. P. 135–140.

84. Cingolani P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3 // Fly. 2012. V. 6. P. 80–92.

85. Cohen J.C. *et al.* Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease // N Engl J Med. 2006. V. 354. P. 1264–72.

86. Collins F.S., McKusick V.A. Implication of Human Genome Project for Medical Science // JAMA. 2001. V. 285. №. 5. P. 1–11.

87. Collins F.S., Varmus H. A new initiative on precision medicine //N Engl J Med. 2015. V. 372. P. 793–5.

88. Coutard B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade // Antiviral. Res. 2020. V. 176. P. 104742.

89. Costela-Ruiz V.J. *et al.* SARS-CoV-2 infection: The role of cytokines in COVID-19 disease // Cytokine Growth Factor Rev. 2020. V. 54. P. 62–75.

90. Cotton R.G., Scriver C.R.Proof of "disease causing" mutation // Hum Mutat. 1998. V. 12(1). P. 1-3.

91. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 // Nature. 2021. V. 600. P. 472–477.

92. Craig E.A., Marszalek J. How Do J-Proteins Get Hsp70 to Do So Many Different Things? //Trends Biochem. Sci. 2017. V. 42. P. 355–368.

93. Dächert C., Gladilin E., Binder M. Gene Expression Profiling of Different Huh7 Variants Reveals Novel Hepatitis C Virus Host Factors // Viruses. 2019. V. 12. P. 36.

94. de Bie P. *et al.* Molecular pathogenesis of Wilson and Menkes disease: Correlation of mutations with molecular defects and disease phenotypes // Journal of Medical Genetics. 2007. V. 44(11). P. 673– 688.

95. de Haan *et al.* Multiple SNP testing improves risk prediction of first venous thrombosis // Blood. 2012. V. 120 (3). P. 656–663.

96. Deelen J. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes // Nat. Commun. 2019. V. 10. № 1. P. 3669.

97. del Castillo I. *et al.* Genetic etiology of non-syndromic hearing loss in Europe // Hum Genet. 2022.

98. De Pristo M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data // Nature Genetic senetics. 2011. V. 43(5). P. 491– 498.

99. Devaux C., Rolain J.M., and Raoult D. *ACE2* receptor polymorphism: susceptibility to SARS-CoV-2, hypertension, multi-organ failure, and COVID- 19 disease outcome // J. Microbiol. Immunol. Infect. 2020. V. 53. P. 425–435.

100. Donaudy F. *et al.* Nonmuscle Myosin Heavy-Chain Gene *MYH14* Is Expressed in Cochlea and Mutated in Patients Affected by Autosomal Dominant Hearing Impairment (DFNA4) // Am J Hum Genet. 2004. V. 74. P. 770–6.

101. Ehret G.B. *et al.* The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals // Nat. Genet.2016. V. 48. P. 1171–1184.

102. Estrada K. *et al.* Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population // JAMA. 2014. V. 311. P. 2305–14.

103. Fakhro K.A. *et al.* The Qatar genome: A population-specific tool for precision medicine in the middle east // Human Genome Variation. 2016. V. 3(1). P. 16016.

104. Fanale D. *et al.* Breast cancer genome-wide association studies: there is strength in numbers // Oncogene. 2012. V. 31. №. 17. P. 2121-8.

105. Fang L. Karakiulakis G., Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? // Lancet Respir Med. 2020. V. 8. №. 4: e21.

106. Fedyakov M.A. *et al.* The Incidence of Lysosomal Acid Lipase Deficiency in the Russian Population // Pediatricheskaya farmakologiya — Pediatric pharmacology. 2018. V. 15 (2). P. 184–185.

107. Fedyakov M.A. *et al.* New frameshift mutation found in *PKP2* gene in arrhythmogenic right ventricular cardiomyopathy/dysplasia: a family case study // Vestnik of Saint Petersburg University. Medicine. 2019. V. 14(1). P. 3–13.

108. Feldmann M. *et al.* Trials of anti-tumour necrosis factor therapy for COVID-19 are urgently needed // The Lancet. 2020. V. 395. P. 1407-1409.

109. Fink-Baldauf I.M. *et al.* CRISPRi links COVID-19 GWAS loci to *LZTFL1* and *RAVER1* // eBioMedicine. 2022. V. 75. P. 103806.

110. Franks P.W.*et al.* Technological readiness and implementation of genomic-driven precision medicine for complex diseases // J Intern Med. 2021. V. 290(3). P. 602-620.

111. Flajollet S.*et al. RREB-1* is a transcriptional repressor of HLA-G // J. Immunol.2009. V. 183. P. 6948–6959.

112. Fraser F.C. and Gunn T. Diabetes mellitus, diabetes insipidus, and optic atrophy. An autosomal recessive syndrome? // J Med Genet. 1977. V. 14. P. 190-193.

113. Frazer K.A. et al. The International Hapmap Consortium. A second generation human haplotype map of over 3.1 million SNPs // Nature. 2007. V. 449. P. 851–861.

114.

115. Fredrickson D.S. Newly recognized disorders of cholesterol metabolism // Ann Intern Med. 1963. V. 58(4). P. 718.

116. Freund M.K. *et al.* Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits // Am. J. Hum. Genet. 2018. V. 103. №. 4. P. 535-552.

117. Freund N.T. *et al.* Reconstitution of the receptor-binding motif of the SARS coronavirus // Protein Eng. Des. Sel. 2015. V. 28. №. 12. P. 567–575.

118. Fry A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population // Am J Epidemiol. 2017. V. 186. P. 1026–34.

119. Fuchsberger C. *et al.* The genetic architecture of type 2 diabetes // Nature. 2016. V. 536. P. 41–47.

120. Fu W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants // Nature. 2013. V. 493(7431). P. 216–220.

121. Gao Yang *et al.* PGG.Han: The Han Chinese Genome Database and Analysis Platform // Nucleic Acids Research. 2020. V. 48(D1). P. 971–76.

122. Genome Reference Consortium. Human Genome Overview https://www.ncbi.nlm.nih.gov/grc/human. Data access: 05/27/2021

123. Giamarellos-Bourboulis E.J. *et al.* Complex immune dysregulation in COVID-19 patients with severe respiratory failure // Cell Host Microbe. 2020. V. 27(6). P. 992–1000.

124. Giwa A. *et al.* Identification of novel prognostic markers of survival time in high-risk neuroblastoma using gene expression profiles // Oncotarget. 2020. V. 11. P. 4293–305.

125. Glotov A.S. *et al.* Identification and Analysis of Genetic Markers of Human Height // Russian Journal of Genetics: Applied Research. 2014. V. 4. №. 2. P. 98–104.

126. Glotov A.S. *et al.* Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group // Clinica Chimica Acta. 2015. V. 446. P. 132–140.

127. Glotov A.S. *et al.*Targeted sequencing analysis of *ACVR2A* gene identifies novel risk variants associated with preeclampsia // J. Matern. Fetal. Neonatal. Med. 2018. V. 5. P. 1-131.

128. Glotov A.S. *et al.* RNA Sequencing of Whole Blood Defines the Signature of High Intensity Exercise at Altitude in Elite Speed Skaters // Genes. 2022. V. 13(4). P. 574.

129. Glotov O.S. *et al.* Comparative analysis of NGS and Sanger sequencing methods for HLA typing at a Russian university clinic // Cellular Therapy and Transplantation (CTT). 2018. V. 7. №. 4(25). P. 72-82.

130. Glotov O.S. *et al.* Whole-exome sequencing for monogenic diabetes in Russian children reveals wide spectrum of genetic variants in MODY-related and unrelated genes // Molecular Medicine Reports. 2019. V. 20. №. 6. P. 4905-4914.

131. Glotov O.S. *et al.* Correlation and regression analysis of athletes` complex traits, based on their personal data, genetic and biochemical parameters // Theory and Practice of Physical Culture. 2015. №. 10. P. 18.

132. Glotov O.S. *et al.* Genetic Risk Factors for the Development of COVID-19 Coronavirus Infection // Russian Journal of Genetics. 2021. V. 57. №. 8. P. 878–892. (а).

133. Glotov O.S. *et al.* The lineage of coronavirus SARS-CoV-2 of Russian origin: Genetic characteristics and correlations with clinical parameters and severity of coronavirus infection // The Siberian Journal of Clinical and Experimental Medicine. 2021. V. 36(4). P. 132–143. (б).

134. Gibson G. Rare and common variants: twenty arguments // Nat Rev Genet. 2012. V. 18. P. 135–45.

135. Goh L., Yap V.B. Effects of normalization on quantitative traits in association test // BMC Bioinformatics. 2009. V. 10. P. 415.

136. Goloshchapov O.V. *et al.* Bacteroides fragilis is a potential marker of effective microbiota transplantation in acute graft-versus-host disease treatment // Cell Ther Transplant. 2020. V. 9(2). P. 47-59.

137. Golubnitschaja O. *et al.* Medicine in the early twenty-first century: paradigm and anticipation - EPMA position paper 2016 // EPMA J. 2016. V. 25. №. 7(1). P. 23.

138. Gonzalez-Garay M.L. The road from next-generation sequencing to personalized medicine // Per. Med. 2014. V. 11. №. 5. P. 523–544.

139. Gra O.A. *et al.* Polymorphisms in xenobiotic-metabolizing genes and the risk of chronic lymphocytic leukemia and non-Hodgkin's lymphoma in adult Russian patients // American Journal of Hematology. 2008. V. 83(4). P. 279-287.

140. Greens K. FDA OK’s 23 and Me Test // Scientists. 2015.

141. Greeley S.A. *et al.* Neonatal diabetes: An expanding list of genes allows for improved diagnosis and treatment // Curr Diab Rep. 2011. V. 11. P. 519-532.

142. Guo Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis // Genomics. 2017. V. 109. №. 2. P. 83-90.

143. Hachiya T. *et al.* Genome-wide meta-analysis in Japanese populations identifies novel variants at the TMC6-TMC8 and SIX3-SIX2 loci associated with HbA1c // Sci. Rep.2017. V. 7. P. 1–11.

144. Hadjinicolaou A. *et al.* De novo pathogenic variant in *SETX* causes a rapidly progressive neurodegenerative disorder of early childhood-onset with severe axonal polyneuropathy // Acta Neuropathol Commun. 2021. V. 9. P. 194.

145. Hamosh A. *et al.* Online Mendelian Inheritance in Man (OMIM®): Victor McKusick's magnum opus // American Journal of Medical Genetics. 2022. Part A. 185A. P. 3259–3265.

146. Hattersley A.T. *et al.* ISPAD clinical practice consensus guidelines 2018: The diagnosis and management of monogenic diabetes in children and adolescents // Pediatr Diabetes. 2018. V. 19 (27). P. 47-63.

147. Hiramatsu K. *et al.* Prevalence and Clinical Characteristics of Hearing Loss Caused by *MYH14* Variants // Genes (Basel). 2021. V. 12. P. 1623.

148. Hofmann A.L. *et al.* Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers // BMC Bioinformatics. 2017. V. 18. P. 8.

149. Huang H.H., Xu T., Yang J. Comparing logistic regression, support vector machines, and permanental classification methods in predictinghypertension // BMC Proc. 2014. V. 8(1). P. 96.

150. Huang C., Wang Y., Li X. Clinical features of patients infected with 2019 novel coronavirus in Wuhan. China // Lancet. 2020. V. 395(10223). P. 497–506.

151. Huber O. Structure and function of desmosomal proteins and their role in development and disease // Cell Mol Life Sci. 2003. V. 60. P. 1872-1890.

152. Hurni Y. *et al.* Spontaneous resolution of nonimmune hydrops fetalis in a fetus with *TP63* gene mutation and *LZTR1* gene variants // Clin Case Reports. 2021. V. 9.

153. Jiang Y. *et al.* The Effect of the Online and Offline Blended Teaching Mode on English as a Foreign Language Learners' Listening Performance in a Chinese Context // Front Psychol. 2021. V. 16 (12). P. 742742.

154. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome // Nature. 2004. V. 431. P. 931– 945.

155. Ishigami D. *et al.* Brainstem intraparenchymal schwannoma with genetic analysis: a case report and literature review // BMC Med Genomics. 2021. V. 14. P. 205.

156. Kampinga H.H., Craig E.A. The HSP70 chaperone machinery: J proteins as drivers of functional specificity // Nat. Rev. Mol. Cell Biol. 2010. V. 11. P. 579–592.

157. Kanwal S., Perveen S., Arshad H.M. Role of Alpha-methylacyl-CoA racemase gene in pathogenecity of CMT patients // J Pak Med Assoc. 2018. V. 68. P. 1039–42.

158. Karczewsk K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes // BioRxiv. 531210. 2019.

159. Khera A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations // Nat Genet. 2018. V. 50. P. 1219–24.

160. Kivela P. Paradigm shift for COVID-19 response: identifying high-risk individuals and treating inflammation // West J Emerg Med. 2020. V. 21(3). P. 473–476.

161. Kleinberger J.W. and Pollin T.I. Undiagnosed MODY: Time for Action // Curr Diab Rep. 2015. V. 15. P. 110.

162. Knowler W.C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin // N Engl J Med. 2002. V. 346. P. 393–403.

163. Komissarova S.M. *et al.* The specifics of hypertrophic cardiomyopathy clinical presentation in patients with various mutations of sarcomere genes // Russian Journal of Cardiology. 2016. V. (1). P. 20-25.

164. Korber B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus // Cell. 2020. V. 182. P. 812–827.

165. Koshevaya Y.S. *et al.* Description of first registered case of the Lopes-Maciel-Rodan syndrome in Russia // Int. J. Mol. Sci. 2022. V. 23. 12437. https://doi.org/10.3390/ijms232012437

166. Kousi M., Katsanis N. Genetic modifiers and oligogenic inheritance // Cold Spring Harb Perspect Med. 2015. V. 5. № 6. P. a017145.

167. Koyama T., Parida L., Platt D.E. Variant analysis of COVID-19 genomes // Bull World Health Organ. 2020. V. 98. №. 7. P.495–504.

168. Kumaran D. *et al.* Genetic characterization of Spinocerebellar ataxia 1 in a South Indian cohort // BMC Med Genet. 2014. V. 15. P. 114.

169. Labay V. *et al.* Mutations in SLC19A2 cause thiamine-responsive megaloblastic anaemia associated with diabetes mellitus and deafness // Nat Genet. 1999. V. 22. P. 300-304.

170. Lazarin G.A. *et al.* An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: Results from an ethnically diverse clinical sample of 23,453 individuals // Genetics in Medicine. 2013. V. 15(3). P. 178– 186.

171. Lek Monkol *et al.* Analysis of Protein-Coding Genetic Variation in 60,706 Humans // Nature. 2016. V. 536 (7616). P. 285–91.

172. Lello L. *et al.* Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer // Sci Rep. 2019. V. 9. P. 15286.

173. Lemelman M.B., Letourneau L. and Greeley S. Neonatal diabetes mellitus: An update on diagnosis and management // Clin Perinatol. 2018. V. 45. P. 41-59.

174. Lerat J. *et al.* Hearing loss in inherited peripheral neuropathies: Molecular diagnosis by NGS in a French series // Mol Genet Genomic Med. 2019. V. 7.

175. Leverenz D.L., Tarrant T.K. Is the HScore useful in COVID-19? // Lancet. 2020. V. 395(10236). P. e83.

176. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics (Oxford, England). 2009. V. 25(14). P. 1754– 1760.

177. Li F., Li W., Farzan M., Harrison S.C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor // Science. 2005. V. 309. P. 1864–1868.

178. Li X. *et al.* Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan // J Allergy Clin Immunol. 2020. V. 146(1). P. 110–118.

179. Li Y. *et al.* Genome-wide association study of COVID-19 severity among the Chinese population // Cell Discov. 2021. V. 7.

180. Liang W. *et al.* Cochlear Nerve Canal Stenosis: Association with *MYH14* and *MYH9* Genes // Ear Nose Throat J. 2021. V. 100. P. 343-346.

181. Lightbody G. *et al.* Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application // Briefings in Bioinformatics. 2019. V. 20 (5). P. 1795–1811.

182. Liu C.T. *et al.* Genome-wide association of body fat distribution in African ancestry populations suggests new loci // PLoS Genet.2013. V. 9. e1003681.

183. Liu X. *et al.* dbNSFP v3. 0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs // Human Mutation. 2016. V. 37(3). P. 235– 241.

184. Liu Z. *et al.* Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2 // J. Med. Virol. 2020. V. 92. №. 6. P. 595–601.

185. Lohmueller K.E. *et al.* Whole-exome sequencing of 2000 Danish individuals and the role of rare coding variants in type 2 diabetes // Am. J. Hum. Genet.2013. P. 1072–1086.

186. Lohse P. *et al.* Compound heterozygosity for a Wolman mutation is frequent among patients with cholesteryl ester storage disease // J Lipid Res. 2000. V. 41(1). P. 23–31.

187. Lopes F. *et al.* Identification of novel genetic causes of Rett syndrome-like phenotypes // J. Med Genet. 2016. V. 53. P. 190.

188. Lvovs D., Favorova O.O., Favorov A.V. A Polygenic Approach to the Study of Polygenic Diseases // Acta Naturae. 2012. V.4. №. 3. P. 59–71.

189. MacDonald *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes // Cell. 1993. V. 72. P. 971–983.

190. Macfarlan T. *et al.* Human THAP7 Is a Chromatin-associated, Histone Tail-binding Protein That Represses Transcription via Recruitment of HDAC3 and Nuclear Hormone Receptor Corepressor // J Biol Chem. 2005. V. 280. P. 7346–58.

191. Mackay I.M., Arden K.E. MERS coronavirus: diagnostics, epidemiology and transmission // Virol J. 2015. V. 12. P. 222.

192. Mahajan A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article // Nat. Genet.2018. V. 50. P. 559–571.

193. Mahmud S. *et al.* Whole Exome Sequence Study of Mild Cognitive Impairment in African and European Americans; the Atherosclerosis Risk in Communities-Neurocognitive Study // Alzheimer's Dement. 2021. V. 17.

194. Majewski J., Schwartzentruber J., Lalonde E. What can exome sequencing do for you? // J. Med. Genet. 2011. V. 48. №. 9. P. 580–589.

195. Manolio T.A. *et al.* Finding the missing heritability of complex diseases // Nature. 2010. V. 461. P. 747-753.

196. Mannucci P.M., Duga S., Peyvandi F. Recessively inherited coagulation disorders // Blood. 2004. V. 104(5). P. 1243–1253.

197. Marciniuk D.D. *et al.* Alpha-1 antitrypsin deficiency targeted testing and augmentation therapy: a Canadian Thoracic Society clinical practice guideline // Can Respir J. 2012. V. 19. P. 109–16.

198. Marino P. *et al.* Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study // Europ. J. of Human Genetics. 2018. V. 26. №. 3. P. 314-323.

199. Martin A.R. *et al.* The Critical Needs and Challenges for Genetic Architecture Studies in Africa // Current Opinion in Genetics & Development. 2018. V. 53. P. 113–20.

200. McGonagle D. *et al.* The role of cytokines including interleukin-6 in COVID-19 induced pneumonia and macrophage activation syndrome-like disease // Autoimmun Rev. 2020. V. 19(6). P. 102537.

201. McCullagh P., Nelder J.A. Generalized Linear Models, Second Edition // Boca Raton: Chapman and Hall/CRC. 1989. 532 p.

202. Mcinnes G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics // Bioinformatics. 2018. December:1–3.

203. Mehdi A.M. *et al.* A peripheral blood transcriptomic signature predicts autoantibody development in infants at risk of type 1 diabetes // JCI Insight.2018. V. 3. e98212.

204. Millat G., Chanavat V., Rousson R. Evaluation of a new NGS method based on a custom AmpliSeq library and Ion Torrent PGM sequencing for the fast detection of genetic variations in cardiomyopathies // Clin Chim Acta. 2014. V. 433. P. 266–71.

205. Miroshnikova V.V. *et al.* Identification of novel variants in the *LDLR*gene in Russian patients with familial hypercholesterolemia using targeted sequencing // Biomedical Reports. 2021. V. 14 (1). P. 15.

206. Miyazawa A. *et al.* A preliminary genetic association study of *GAD1* and *GABAB* receptor genes in patients with treatment-resistant schizophrenia // Mol Biol Rep. 2021.

207. Mohnike K. *et al.* Clinical and genetic evaluation of patients with *KATP* channel mutations from the German registry for congenital hyperinsulinism // Horm Res Paediatr. 2014. V. 81. P. 156-168.

208. Moore J., June C. Cytokine release syndrome in severe COVID-19 // Science. 2020. V. 368(6490). P. 473–474.

209. Morgant S. *et al.* Chapter 8. Role of Next-Generation Sequencing Technologies in Personalized Medicine // P5 eHealth: An Agenda for the Health Technologies of the Future // Eds. by G. Pravettoni, S. Triberti. 2020. P. 125-145.

210. Mousa M. *et al.* Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates // eBioMedicine. 2021. V.74. P. 103695.

211. Munne S. *et al.* Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing // Fertil Steril. 2017. V. 108(1). P. 62-71.

212. Muntoni S. *et al.* Prevalence of cholesteryl ester storage disease // Arterioscler Thromb Vasc Biol. 2007. V. 27(8). P. 1866–1868.

213. Mustacich D.J. *et al.* Digenic Inheritance of a *FOXC2* Mutation and Two *PIEZO1* Mutations Underlies Congenital Lymphedema in a Multigeneration Family // Am J Med. 2021.

214. Nebert D.W., Carvan M.J. Ecogenetics: from Biology to Health // Toxicol. Industr. Health. 1997. V. 13. P. 163-192.

215. Nguyen-Ngoc K.V. *et al.* Mosaic loss of non-muscle myosin IIA and IIB is sufficient to induce mammary epithelial proliferation // J Cell Sci. 2017.

216. Ng S.B.*et al.* Targeted capture and massively parallel sequencing of 12 human exomes // Nature. 2009. V. 461(7261). P. 272-6.

217. Ng S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder // Nat Genet. 2010. V. 42(1). P. 30-5.

218. Nykamp Keith *et al.* Sherloc: A Comprehensive Refinement of the ACMG–AMP Variant Classification Criteria // Genetics in Medicine. 2017. V. 19(10). P.1105–17.

219. O'Donovan M.C. What have we learned from the Psychiatric Genomics Consortium // World Psychiatry. 2015. V. 14. №. 3. P. 291–293.

220. Ohni S. *et al.* Direct molecular evidence for both multicentric and monoclonal carcinogenesis followed by transdifferentiation from hepatocellular carcinoma to

cholangiocarcinoma in a case of metachronous liver cancer // Oncol Lett. 2021. V. 23. P. 22.

221. Oleksyk T., Brukhin V., O'Brien S.J. The Genome Russia Project: Closing the Largest Remaining Omission on the World Genome Map // GigaScience. 2015. V. 4(1). P. 53.

222. PANGO lineages. https://cov-lineages.org/lineage_designation.html

223. Pendina A.A. *et al.* Reproductive History of a Woman With 8p and 18p Genetic Imbalance and Minor Phenotypic Abnormalities // Front. Genet. 2019. V. 10. P. 1164.

224. Perez-Becerril C., Evans D.G., Smith M.J. Pathogenic noncoding variants in the neurofibromatosis and schwannomatosis predisposition genes // Hum Mutat. 2021. V. 42. P. 1187–207.

225. Pérez-Valencia J.A. *et al.* Angiogenesis and evading immune destruction are the main related transcriptomic characteristics to the invasive process of oral tongue cancer // Sci Rep. 2018. V. 8. P. 2007.

226. Petersen I. Classification and Treatment of Diseases in the Age of Genome Medicine Based on Pathway Pathology // Int J Mol Sci. 202.1 V. 30. №. 22(17). P. 9418.

227. Pierik M. *et al.* The IBD international genetics consortium provides further evidence for linkage to IBD4 and shows gene-environment interaction // Inflammatory Bowel Diseases. 2005. V. 11. №.1. P. 1–7.

228. Piotrowski A. *et al.* Targeted massively parallel sequencing of candidate regions on chromosome 22q predisposing to multiple schwannomas: An analysis of 51 individuals in a single-center experience // Hum Mutat. 2022. V. 43. P. 74–84.

229. Peltonen L., McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era // Science. 2001. V. 16. №. 291(5507). P. 1224-9.

230. Povysil G. *et al.* Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19 // J Clin Invest. 2021. 131.

231. Prakrithi P. *et al.* Genetic Risk Prediction of COVID-19 Susceptibility and Severity in the Indian Population // Frontiers in Genetics. 2021. V. 12.

232. Pritchard J.K., Cox N.J. The allelic architecture of human disease genes: common disease-common variant...or not? // Hum Mol Genet. 2002. V. 11(20). P. 2417-23.

233. Puthucheary Z. *et al.* The *ACE* gene and human performance: 12 years on // Sports Med. 2011. V. 41(6). P. 433-48.

234. Qi F. *et al.* Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses // Biochem. Biophys. Res. Commun. 2020. V. 526. №. 1. P. 135–140.

235. Rabbani B. *et al.* Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders // J. Hum. Genet. 2012. V. 57. P. 621-632.

236. Rahman M.H. *et al.* A Network-Based Bioinformatics Approach to Identify Molecular Biomarkers for Type 2 Diabetes that Are Linked to the Progression of Neurological Diseases // Int J Environ Res Public Health. 2020. V.17. P. 1035.

237. Ramensky V.E. *et al.* 2021. Targeted Sequencing of 242 Clinically Important Genes in the Russian Population from the Ivanovo Region // Frontiers in Genetics. V. 12. 709419.

238. Richards S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology // Genet Med. 2015. V. 17. №. 5. P. 405–424.

239. Reich D.E., Lander E.S. On the allelic spec- trum of human disease // Trends Genet. 2001. V. 17. №.9. P. 502–510.

240. Reid S. *et al.* High genetic risk score is associated with early disease onset, damage accrual and decreased survival in systemic lupus erythematosus // Ann Rheum Dis. 2020. V. 79. P. 363–9.

241. Robinson J.G. *et al.* Efficacy and safety of alirocumab in reducing lipids and cardiovascular events // N Engl J Med. 2015. V. 372. P. 1489–99.

242. Rodan L.H. *et al.* A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene // Eur. J. Hum. Genet. 2016. V. 24. P. 1826–1827.

243. Rodriguez-Flores J.L. *et al.* Exome sequencing identifies potential risk variants for Mendelian disorders at high prevalence in Qatar // Human Mutation. 2014. V. 35(1). P. 105– 116.

244. Rossi Á.D. *et al.* Association between *ACE2* and *TMPRSS2* nasopharyngeal expression and COVID-19 respiratory distress // Sci Rep. 2021. V. 11. P. 9658.

245. Rubio-Cabezas O. *et al:* Wolcott-Rallison syndrome is the most common genetic cause of permanent neonatal diabetes in consanguineous families // J Clin Endocrinol. Metab. 2009. V. 94. P. 4162-4170.

246. Sachdev N.M. *et al.* The rate of mosaic embryos from donor egg as detected by next generation sequencing (NGS) varies by IVF laboratory // Fertil Steril. 2016. V. 106(3). e156–7.

247. Sabino E.C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence // The Lancet. 2021. V. 397. P. 452-455.

248. Saifitdinova A.F. *et al.* Mosaicism in preimplantation human embryos // Integrative Physiology. 2020. V. 1. №. 3. P. 225–230.

249. Saleem I.B. *et al.* Identification and Computational Analysis of Rare Variants of Known Hearing Loss Genes Present in Five Deaf Members of a Pakistani Kindred // Genes (Basel). 2021. V.12. P. 1940.

250. Schidtke P. SARS-CoV-2 - part 2 - From the viral genome to protein structures. MARCH 27, 2020. https://www.discngine.com/blog?author=52850d39e4b0b817d0c61ff9

251. Schmidt B., Hildebrandt A. Next-generation sequencing: big data meets high performance computing // Drug Discov Today. 2017. V. 22(4). P. 712-717.

252. Scott R.A. *et al.* An expanded genome-wide association study of type 2 diabetes in Europeans // Diabetes. 2018. V. 66. P. 2888–2902.

253. Shcherbak S.G. *et al.* Basic Predictive Risk Factors for Cytokine Storms in COVID-19 Patients // Front. Immunol. 2021. V. 12. P. 745515.

254. Shcherbak S.G. *et al.* Identification of Genetic Risk Factors of Severe COVID-19 Using Extensive Phenotypic Data: A Proof-of-Concept Study in a Cohort of Russian Patients // Genes. 2022. V. 13(3). P. 534.

255. Shelton J.F. *et al.* The UGT2A1/UGT2A2 locus is associated with COVID-19-related loss of smell or taste // Nat Genet. 2022.

256. Shields B.M. *et al.* Maturity-onset diabetes of the young (MODY): How many cases are we missing? // Diabetologia. 2010. V. 53. P. 2504-2508.

257. Shitao R.A.O., Alexandria L.A.U., Hon-Cheong S.O. Exploring diseases/traits and blood proteins causally related to expression of *ACE2*, the putative receptor of 2019-nCov: A Mendelian Randomization analysis // Diabetes Care. 2020. V. 43. №. 7. P. 1416–1426.

258. Shaw-Smith C. *et al.* Recessive *SLC19A2* mutations are a cause of neonatal diabetes mellitus in thiamine-responsive megaloblastic anaemia // Pediatr Diabetes. 2012. V. 13. P. 314-321.

259. Sheremet N.L. *et al.* Molecular genetic diagnosis of Stargardt disease // Vestnik Oftalmologii. 2017. V. 133(4). P. 4– 11.

260. Shikov A.E. *et al.* Analysis of the Spectrum of *ACE2* Variation Suggests a Possible Influence of Rare and Common Variants on Susceptibility to COVID-19 and Severity of Outcome // Front. Genet. 2020. V. 11. P. 551220.

261. Shikov A. *et al.* The application of Nanopore sequencing for variant calling on the human mitochondrial DNA // Bio. Comm. 2021. V. 66(2). P. 109–123.

262. Shulla A.*et al.* Transmembrane Serine Protease Is Linked to the Severe Acute Respiratory Syndrome Coronavirus Receptor and Activates Virus Entry // Journal of Virology. 2011. V. 85. №. 2. P. 873–882.

263. Shungin D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution // Nature.2015. V. 518. P. 187–196.

264. Silverman E.K., Sandhaus RA. Clinical practice. Alpha1-antitrypsin deficiency. // N Engl J Med. 2009. V. 360. P. 2749–57.

265. Skeppholm M. *et al.* ADAMTS13 and von Willebrand factor concentrations in patients with diabetes mellitus // Blood Coagul. Fibrinol.2009. V. 20. P. 619–626.

266. Sladek R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes // Nature. 2007. V. 445. P. 881-885.

267. Stankov K., Benc D., Draskovic D. Genetic and epigenetic factors in etiology of diabetes mellitus type 1 // Pediatrics. 2013. V. 132. №. 6. P. 1112-22.

268. Stehouwer C.D. *et al.* Increased urinary albumin excretion, endothelial dysfunction, and chronic low-grade inflammation in type 2 diabetes: Progressive, interrelated, and independently associated with risk of death // Diabetes.2002. V. 51. P. 1157–1165.

269. Strafella C. *et al.* Analysis of *ACE2* genetic variability among populations highlights a possible link with COVID19-related neurological complications // Genes. 2020. V. 11 P. 741.

270. Strokova T.V., Bagaeva M.E., Matinyan IA. Defitsit lizosomnoi kisloi lipazy // Russkii meditsinskii zhurnal. 2017. V. 25(19). P. 1346–1351.

271. Sturm A.C. *et al.* Clinical genetic testing for familial hypercholesterolemia: JACC scientific expert panel // J Am Coll Cardiol. 2018. V. 72. P. 662–80.

272. Suh S. *et al.* A systematic review on papers that study on Single Nucleotide Polymorphism that affects coronavirus 2019 severity // BMC Infect Dis. 2022. V. 22. P. 1–11.

273. Surcel A. *et al.* Targeting Mechanoresponsive Proteins in Pancreatic Cancer: 4-Hydroxyacetophenone Blocks Dissemination and Invasion by Activating MYH14 // Cancer Res. 2019. V. 79. P. 4665.

274. Suwinski P. *et al.* Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics // Front. Genet. 2019. V. 10. P. 49.

275. Talley M.J. *et al.* Generation of a Mouse Model to Study the Noonan Syndrome Gene Lztr1 in the Telencephalon // Front Cell Dev Biol. 2021. V. 9.

276. Tan O. *et al.* Application of next-generation sequencing to improve cancer management: A review of the clinical effectiveness and cost effectiveness // Clinical Genetics. 2018. V. 93. № 3. P. 533–544.

277. Di Taranto M.D., Giacobbe C. and Fortunato G. Familial hypercholesterolemia: A complex genetic disease with variable phenotypes // Eur J Med Genet. 2020. V. 63. P. 103831.

278. Teekakirikul P. *et al.* Inheritedcardiomyopathies: molecular genetics and clinical genetic testing in the postgenomic era // J Mol Diagn. 2013. V.15(2). P. 158–70.

279. Toovey O.R. *et al.* Introduction of Brazilian SARS-CoV-2 484K.V2 related variants into the UK // J Infect. 2021. V. 82(5). e23-e24.

280. Tighe O. *et al.* Genetic diversity within the R408W phenylketonuria mutation lineages in Europe // Human Mutation. 2003. V. 21(4). P. 387– 393.

281. Turkunova M.E. *et al.* Molecular Genetics and Pathogenesis of the Floating Harbor Syndrome: Case Report of Long-Term Growth Hormone Treatment and a Literature Review // Front. Genet. 2022. V. 13. P. 846101.

282. Van der Graaf A. *et al.* Molecular basis of autosomal dominant hypercholesterolemia: Assessment in a large cohort of hypercho- lesterolemic children // Circulation. 2011. V. 123. P. 1167-1173.

283. van Moorsel CHM. *et al.* The *MUC5B* Promoter Polymorphism Associates with Severe COVID-19 in the European Population // Front Med. 2021. V. 8.

284. Wallace S.E., Bird T.D. Molecular genetic testing for hereditary ataxia // Neurol Clin Pract. 2018. V. 8. P. 27–32.

285. Walter K. *et al.* The UK10K project identifies rare variants in health and disease // Nature. 2015. V. 526(7571). P. 82– 89.

286. Wang M. *et al.* A novel MYH14 mutation in a Chinese family with autosomal dominant nonsyndromic hearing loss // BMC Med Genet. 2020. V. 21. P. 154.

287. Wang Q. *et al.* Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity // Sci. Rep. 2017. V. 7. №. 1. P. 885.

288. Wang R. *et al.* Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants // Commun Biol. 2021. V. 4(1). P. 228.

289. Wang X. *et al.* Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction // J. Diabetes.2016. V. 8. P. 24–35.

290. Weedon M.N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population // Nat Genet. 2007. V. 39(10). P. 1245-50.

291. Weissman A. *et al.* Chromosomal mosaicism detected during preimplantation genetic screening: results of a worldwide Web-based survey // Fertil Steril. 2017. V. 107(5). P. 1092–7.

292. Wiegman A. *et al.* European atherosclerosis society consensus panel. Familial hypercholester- olaemia in children and adolescents: Gaining decades of life by optimizing detection and treatment // Eur Heart J. 2015. V. 36. P. 2425-2437.

293. Wong K.H.Y. *et al.* Towards a Reference Genome That Captures Global Genetic Diversity // Nature Communications. 2020. V. 11 (1). P. 5482.

294. Wood D., De Backer G., Faergeman O. Prevention of Coronary Heart Disease in Clinical Practice. Recommendations of the Second Joint Task Force of the European and other Societies on Coronary Prevention // Eur Heart J. 1998.№. 19. P. 1434–1503.

295. Wooster R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2* // Nature. 1995. V. 378. P. 789–92.

296. World Health Organization. SARS-CoV-2 Variants https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/

297. Wulff K., Herrmann F.H. Twentytwo novel mutations of the factor VII gene in factor VII deficiency // Human Mutation. 2000. V. 15(6). P. 489– 496.

298. Wu P. *et al.* Trans-ethnic genome-wide association study of severe COVID-19 // Communications Biology. 2021. V. 4. P. 1034.

299. Wu B.B. *et al.* Association between ABO blood groups and COVID-19 infection, severity and demise: A systematic review and meta-analysis // Infect. Genet. Evol. 2020. V. 84. P. 104485.

300. Xu S., Hu Z. Generalized Linear Model for Interval Mapping of Quantitative Trait Loci // Theor. Appl. Genet. 2010.V. 121. №. 1.P.47–63.

301. Yalcintepe S. *et al.* The importance of multiple gene analysis for diagnosis and differential diagnosis in CharcotMarie tooth disease // Turk Neurosurg. 2021.

302. Yang X. *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study // Lancet Respir Med. 2020. V. 8. №. 5. P. 475–481.

303. Yan R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2 // Science. 2020. V. 367. P. 1444–1448.

304. Yengo L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry // bioRxiv.2018. 274654.

305. Yi N., Banerjee S. Hierarchical generalized linear models for multiple quantitative trait locus mapping // Genetics. 2009. V. 181. №. 3. P. 1101–13.

306. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications // Genomics. 2020. V. 112. №. 5. P. 3588–96.

307. Zabetian C.P. *et al.* A revised allele frequency estimate and haplotype analysis of the DBH deficiency mutation IVS1+2T->C in African- and European-Americans // Am. J. Med. Genet. Part A. 2003. V. 123. P. 190–192.

308. Zachariah P. *et al.* Epidemiology, clinical features, and disease severity in patients with coronavirus disease 2019 (COVID-19) in a children's hospital in New York City, New York // JAMA Pediatr. 2020. V. 174(10). e202430.

309. Zhang H. *et al.* LZTR1: A promising adaptor of the CUL3 family (Review) // Oncol Lett. 2021. V. 22. P. 564.

310. Zhang Z. *et al.* Detection of *PKD1* and *PKD2* Somatic Variants in Autosomal Dominant Polycystic Kidney Cyst Epithelial Cells by Whole-Genome Sequencing // J Am Soc Nephrol. 2021. V. 32. P. 3114–29.

311. Zhang W. *et al.* Emergence of a Novel SARS-CoV-2 Variant in Southern California // JAMA Pediatr. 2021. V. 325(13). P. 1324-1326.

312. Zhao S.*et al.* Pilot study of expanded carrier screening for 11 recessive diseases in China: Results from 10,476 ethnically diverse couples // European Journal of Human Genetics. 2019. V. 27(2). P. 254– 262.

313. Zhernakova D.V. *et al.* Analytical “bake-off” of whole genome sequencing quality for the genome Russia project using a small cohort for autoimmune hepatitis // PLoS ONE. 2018. V. 13(7). P. 1– 18.

314. Zhernakova D.V. *et al.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. 2020. V. 112. №. 1. P. 442-458.

315. Zhou F. *et al.*Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study // Lancet. 2020. V. 395(10229). P. 1054–1062.

316. Zhou Z. *et al.* Loss-of-Function Piezo1 Mutations Display Altered Stability Driven by Ubiquitination and Proteasomal Degradation // Front Pharmacol. 2021. V. 12.

317. Zou Y. *et al.* Multiple gene mutations, not the type of mutation, are the modifier of left ventricle hypertrophy in patients with hypertrophic cardiomyopathy // Mol Biol Rep. 2013. V. 40(6). P. 3969–76.

318. http://www.biometrica.tomsk.ru

319. https://www.weforum.org/agenda

320. https://www.omim.org/statistics/geneMap

321. https://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array

322. http://www.internationalgenome.org/

323. https://genomics.ut.ee/en

324. https://mutalyzer.nl.

325. http://www.ncbi.nlm.nih.gov/RefSeq/
326. http://www.lrg-sequence.org
327. https://www.ebi.ac.uk/gwas
328. https://phgkb.cdc.gov/PHGKB/hNHome.acton
329. http://www.hgmd.cf.ac.uk/ac/index.php
330. http://www.ncbi.nlm.nih.gov/snp
331. http://www.hgvs.org/dblist/dblist.html
332. http://www.lovd.nl
333. https://decipher.sanger.ac.uk
334. http://browser.1000genomes.org/index.html
335. https://gnomad.broadinstitute.org/
336. https://evs.gs.washington.edu/EVS/
337. http://exac.broadinstitute.org
338. http://gnomad.broadinstitute.org
339. https://allofus.nih.gov/news-events
340. https://www.genomicsengland.co.uk
341. https://coronavirus.jhu.edu

БЛАГОДАРНОСТИ

Автор выражает свою глубокую благодарность своим наставникам и научным руководителям и консультантам: Владиславу Сергеевичу Баранову и Татьяне Эдуардовне Иващенко, к сожалению, безвременно ушедшими от нас в 2022 году, без которых данная диссертация вряд ли состоялась бы. Хотелось бы выразить благодарность всем сотрудникам ФГБНУ «НИИ АГиР им.Д.О.Отта», которые прямо или косвенно помогали в создании, рождении и становлении данной диссертации, особенно: руководителю отдела геномики Андрею Сергеевичу Глотову, с.н.с. Михаилу Владимировичу Асееву, с.н.с. Антону Вячеславовичу Киселеву, н.с. Юрию Александровичу Барбитову, с.н.с. Анне Андреевне Пендиной за вопрос, который позволил более четко определить цель диссертации, а также руководителям института: директору - Игорю Юрьевичу Когану и заместителю директора по научной работе - Олесе Николаевне Беспаловой. Рад поблагодарить всех сотрудников Кафедры генетики СПБГУ за вклад в мое образование и становление как ученого. Хотелось бы поблагодарить коллектив лаборатории генетики и вирусологического центра СПб ГБУЗ «Городская больница № 40», который с 2013 по 2021 год не только помогал эффективно бороться с Covid-19, но и помогал мне делать замечательные научные исследования во время пандемии, врачей клиницистов СПб ГБУЗ «Городская больница № 40»: Уразова Станислава Петровича, Анисенкову Анну Юрьевну, Мосенко Сергея Викторовича. Рад поблагодарить заместителя директора по науке Наталью Викторовну Скрипченко и коллег НИО экспериментальной медицинской вирусологии, молекулярной генетики и биобанкинга ФГБУ «ДНК ЦИБ» ФМБА РФ: Алексея Борисовича Чухловина, Ольгу Владимировну Голеву и Юрия Александровича Эйсмонта за ценные советы при подготовке диссертации.Автор рад выразить благодарность коллегам, долговременное сотрудничество с которыми внесло значительный вклад в выполнение исследований и публикацию работ в ведущих

периодических изданиях, на которых основана эта диссертация: всем вышеперечисленным коллегам, а также А.Н. Чернову, Д.Е. Полеву, М.А. Федякову, А.В. Предеусу, А.Е. Шикову, В.В. Цай, Р.С. Калинину, М.В. Москаленко, Е.Б. Башниной, И.В. Поляковой.

В заключение автор выражает глубокую благодарность всем своим родным и близким, а именно родителям – Сергею Александровичу Глотову, Ольге Владимировне Глотовой, супруге – Горбачевой Валерии Юрьевне, тестю – Горбачеву Юрию Евгеньевичу, детям: Виталию и Юлиане Глотовым, и брату – Андрею Сергеевичу Глотову за многолетнее терпение, моральную поддержку и создание благоприятного и позитивного фона для написания этой работы.

THE RESEARCH INSTITUTE OF OBSTETRICS, GYNECOLOGY AND REPRODUCTOLOGY NAMED AFTER D.O.OTT

Manuscript

Oleg S. Glotov

HUMAN EXOME SEQUENCING AND PROSPECTS FOR PREDICTIVE MEDICINE

Scientific Specialty 1.5.7. GENETICS

A DISSERTATION
submitted for the degree of
Doctor of Biological Sciences

Translation from Russian

Saint Petersburg, 2023

This dissertation is dedicated to the bright memory of my teacher, mentor, supervisor and consultant, Doctor of Medical Sciences, Professor, Correspondent Academician RAS Vladislav Sergeevich Baranov.

CONTENT

INTRODUCTION. Studies on Genetic Basis of Human Health: from Function to Pathology, from Mutation to Variation. Predictive Medicine and Clnical Genetic Passport as Concepts of the New Generation Sequencing Era

In the late 20th and early 21st centuries scientific and technological advances in biology and medicine have produced new high-tech methods for early diagnostics and paved the way to identification of numerous target protein and genetic markers and introduction of new screening strategies and targeted therapy protocols in clinical practice. This enabled clinicians to precisely identify causes of rare monogenic diseases, improve prevention, and boost effeciency of treatment for multifactorial socially significant diseases, ultimately contributing to better health quality and life expectancy among the population of economically developed countries [Baranov *et al.*, 2021]. All these achievements have given an impetus to a paradigm shift in the overall healthcare system and enabled the transition from group-based to predictive or preventive personalized medicine (PM) and therapy that relies on the disease clinical diagnosis and stage, patient's gender and age, as well as individual molecular genetic biomarker profiles associated with pathology development, prognosis, outcome, and treatment efficacy. Advances in genetics and information technology allowed to redefine 'mutation' as a term, while emerging new disciplines, such as genomics (proteomics, metabolomics, transcriptomics, pharmacogenomics) shaped the development of standard criteria for large data set processing using bioinformatics, once high-throughput methods, in particular the new generation DNA sequencing (NGS), were introduced to study gene structure [Baranov *et al.*, 2021].

Once invented in the 1980s, first generation DNA sequencing technology enabled deciphering of the human genome sequence. For this purpose, in 1990 the US National Institutes of Health launched the Human Genome Project (HGP), that brought together the UK, Japan, France, Germany, Spain and China in addition to the US. The project ended up in 2003, when the US National Center for Biotechnology Information (NCBI) published the first complete human genome assembly (hg17)

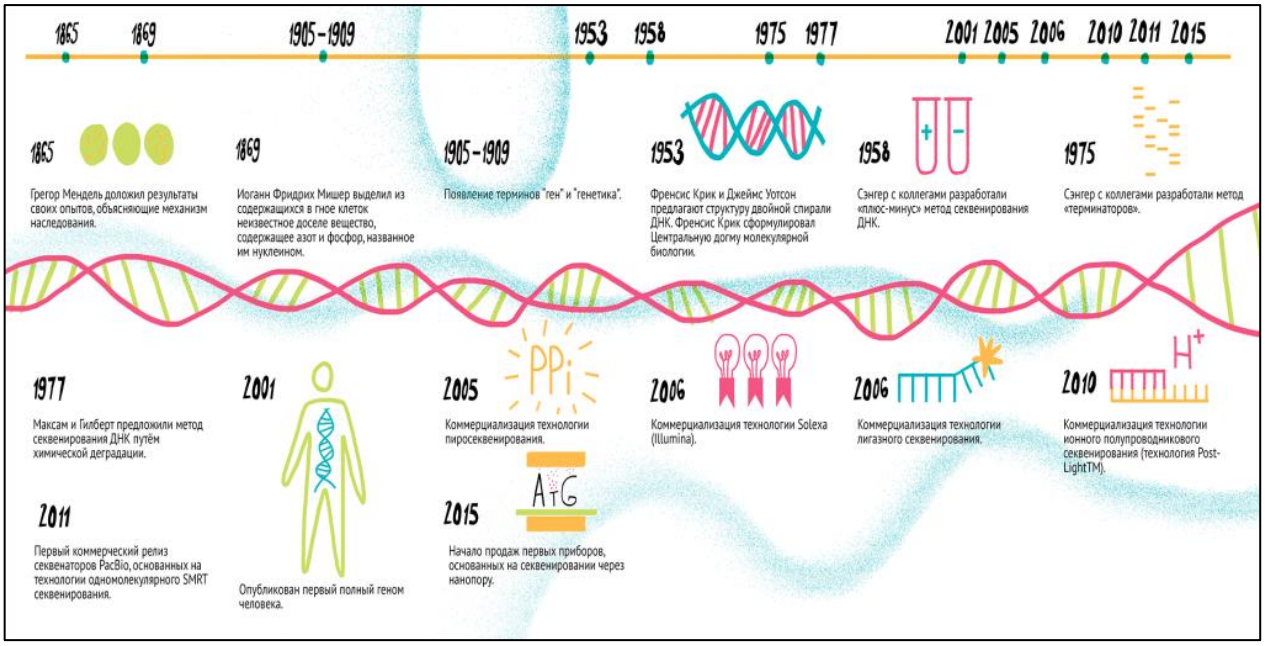
[IHGSC. Finishing the euchromatic, 2004]. However, this assembly contained numerous omissions that were resequenced later. On May 09, 2022 the Genome Research Consortium (GRC) published the latest human genome assembly GRCh38.p14 [Genome Reference Consortium, 2021].

The next step after the human genome publication was to establish a map of human genetic variation, or haplotype map (Project Haplotype Map, HapMap), for 270 individuals representing four races. In 2007 the second HapMap assembly of 3.1 million variants was published [The International Hapmap, 2007]. The HapMap project was followed up by the 1000 Human Genomes Sequencing Initiative (1000 Genomes Project) [Genomes Project Consortium, 2012]. This project examined 1,092 individuals belonging to 14 ethnic populations to identify 38 million single nucleotide variants (SNVs, SNPs), 1.4 million biallelic insertions or deletions (indels), and 14,000 major deletions. The HapMap and 1000 Genomes projects allowed to identify rare high-penetrance SNPs, apparently responsible for human monogenic diseases [Freund *et al.*, 2018]. Identification of SNVs led to further challenges to annotate pathological variants and associate them with diseases respectively. The HapMap and 1000 Genomes projects data were used to develop methodology for genome-wide association studies (GWAS). GWAS allowed to characterize population frequencies of multiple SNV/SNPs, associated with multifactorial diseases, such as type 1 diabetes [Stankov *et al.*, 2013], type 2 diabetes [Sladek *et al.*, 2007], breast cancer [Fanale *et al.*, 2012], etc., as well as human aging and longevity [Deelen *et al.*, 2019]. As soon as by now we understand the type of inheritance, disease pathogenesis, and SNV/SNPs population frequencies, NGS-identified pathogenic gene variants can be annotated and clinically interpreted [Rabbani *et al.*, 2012]. Notably, identification of disease (condition) associated genes and their variants enables us to study their pathogenetic role. In 2009 the Shendure group pursued the first attempt to use NGS to detect genetic aberrations and eventually discover Miller-Fisher syndrome, a rare recessive inflammatory (autoimmune) demyelinating polyradiculoneuropathy, by full-exome sequencing [Ng *et al.*, 2009; Ng *et al.*, 2010].

Recent years have seen the publication of numerous studies devoted to specific allele frequencies of functionally significant polymorphisms in various populations. An obligatory initial stage of research investigates population frequencies of functionally significant gene alleles. The importance of population studies is explained by frequency differences in functional polymorphic gene variants across different populations, which may depend on geographical conditions, region of residence, dietary features, race and ethnicity, and many other factors [Khusnutdinova *et al*., 1997; Nebert and Carvan, 1997; Stepanov and Puzyrev, 2000; Stepanov, 2002]. Therefore, there is little doubt that in different populations or ethnic groups, identical genetic polymorphisms may impact differently on etiology and pathogenesis of a particular disease [Baranov et al., 2000; Stepanov, 2002; Glotov O. *et al*., 2004].

This data proves that new high-throughput methods, including NGS, are still highly demanded for further population-oriented studies of the fundamentals of human health. Therefore, most crucial prospects to boost practical efficiency of NGS in predictive medicine include the development of population databases (including those registered in Russia) to estimate frequencies of gene variants, responsible for the pathogenesis of hereditary and multifactorial diseases. In addition, bioinformatic and statistical protocols for DNA sequencing data processing and analysis require further improvement.

The background for genomic research and nucleic acid sequencing is presented in Figure 1. In contrast to earlier sequencing methods, all NGS technology platforms are based on simultaneous reading of multiple genome fragments. NGS can generate over a few million nucleotide sequences in a single run to undergo further analysis.



1865	Gregor Mendel presented experimental results explaining transmission of genetic traits.
1869	Johannes Friedrich Miescher was the first to isolate unknown phosphate-rich chemicals from pus cells, which he called nuclein.
1905-1909	Terms ‘gene’ and ‘genetics’ emerged.
1953	Francis Crick and James Watson uncovered the double-helix structure of DNA. Francis Crick formulated the central dogma of molecular biology.
1958	Sanger et. al. developed the plus and minus DNA sequencing technique.
1975	Sanger et. al. developed the terminator sequencing method.
1977	Maxam and Gilbert developed a DNA sequencing method based on chemical degradation.
2005	Commercial availability of pyrosequencing.
2006	Commercial availability of Solexa (Illumina) technology.

2006	Commercial availability of ligase-mediated sequencing.
2010	Commercial availability of ion semiconductor sequencing (Post-LightTM technology).
2011	Pacific Biosciences announced the first commercial release of PacBio Single-Molecule Real-Time Sequensing Systems (SMRT)
-	The first complete sequence of a human genome was published.
2015	First commercial shipments of nanopore-based DNA sequencing systems.

Figure 1. Historical milestones in genome research and nucleic acid sequencing (biomolecula.ru).

NGS Platforms – Advantages and Drawbacks

NGS technology can be divided into short (100 to 600 bp) and long (up to 900,000 bp) read sequencing. Currently, short reads of sequences are the most broadly used methods, as they are cheaper and have higher accuracy than long reads. However, short-read technologies cannot be used for repetitive or heterozygous sequences, where long-read sequencing is the technology of choice [Morganti *et al*., 2020]. Hence, sequencing technologies formally distinguish second- and third-generation NGS. Second generation NGS is represented by platforms generating large numbers of short reads (25-800 bp), such as the 454 Life Sciences (now discontinued), Illumina, Ion Torrent, and MGI. Third generation NGS includes Pacific Biosciences and Oxford Nanopore sequencers with long fragment reading capacity [Barkhatov *et al*., 2016]. Currently, Illumina sequencers are the most widely used; Ion Torrent devices (for noninvasive prenatal testing (NIPT) and prenatal genetic testing (PGT)) are used for specific clinical applications. In addition, MGI machines are actively deployed as well. In Russia, Oxford Nanopore technology is widely spread, though applied to fundamental matters exclusively (these machines are more prominent in bioinformatics, than in laboratory genetics) [Shikov *et al*., 2019].

Oxford Nanopore technology is generally considered less accurate than second generation sequencing techniques. However, there is no rule of thumb. For instance, our pioneer investigation provided breakthrough data demonstrating that when applied to mitochondrial genome variants at large coverage (about 4000x), nanopore and Illumina technologies generate consistent results [Shikov *et al.*, 2021].

NGS Applications

NGS applications in medical research are versatile. By their intended use high-throughput sequencing techniques can be split in the following groups:

1) analysis of entire genome (whole-genome sequencing, WGS);
2) analysis of all protein-coding regions in a genome (whole-exome sequencing, WES);
3) analysis of particular disease-causing gene sequences (from clinical exomes embracing some 4-5000 clinically relevant genes, to kits for small target regions of 1–3 genes or loci);
4) transcriptome sequencing (RNA-seq);
5) analysis of bacterial microbiome biological diversity [Goloshchapov *et al.*, 2020; Баранов *и др.*, 2021].

Specific aspects regarding the application of various NGS techniques in medical practice are presented in Table 1.

Table 1. Advantages and drawbacks of different NGS techniques.

Technology	Highlights	Advantages	Drawbacks	Import-substitution options
Targeted NGS (genetic variation kits to evaluate particular genes of interest)	Investigates only select gene variations of interest (SNP)	Fast and cost-effective 100% efiiciency	Gene panels fails to include widely-spread gene variations, typical for some or other population Fails to detect CNV, mini- and macro-satellite sequences	Available
NGS human exome sequencing	Includes over 20,000 gene	Ideal for closely related pairs	Expensive, 50% diagnostic efficiency.	Solutions are under way

('exome')	variations (nearly all monogenic diseases)	Not sensitive to population traits	Fails to detect CNV, mini-and macro-satellites Requires bioinformatic data processing	
NGS human genome sequencing ('genome')	Entire human genome	Allows to identify molecular origin of almost all diseases Not sensitive to population traits	Expensive, unknown efficiency. Requires bioinformatic processing for bulky databases	Available
Comprehensive approach (NGS, MLPA, RT-PCR)	Includes select gene variations	Cost-effective 100% efficient	Sensitive to population traits	Available

Targeted sequencing (TS) of specific genes pioneered NGS practical application. Based on long-range PCR to amplify sections up to 50-100 thousand bp, this technology is able to read particular genes at 1000 times cheaper cost than before. Firstly, this approach is particularly effective for small genes [Glotov A. *et al.*, 2018]. Based on Ampliseq technology, targeted sequencing also allows to amplify up to 27 thousand PCR products of target transcripts for mass parallel sequencing. This technique is therefore helpful to study a relatively small number of genes [Glotov *et al.*, 2015]. Thirdly, targeted sequencing can analyze target gene (or genes) fragments, with samples prepared via enrichment technology utilizing target-specific probes to capture the desired gene fragments of interest in the genome for sequencing. This technique is employed to diagnose all cystic fibrosis gene mutations, as well as gene mutations responsible for other monogenic diseases, such as Wilson disease [Balashova *et al.*, 2020]. Low cost is a notable advantage of TS over WGS and WES (more than 50-fold) and shorter timespan for sequencing and bioinformatic analysis [Gonzalez-Garay, 2014].

Whole-genome sequencing allows to entirely investigate all human genes, including both structural and regulatory, as well as some 3 billion bp in the haploid set. However, the latest published version of human GRCh38 genome showed that protein-encoding structural genes (exome) account for 3.09% (90 million bp) only of the total number of genes, though contain about 85% of functional variants associated

with clinically manifest diseases [Guo *et al*., 2017; Majewski *et al*., 2011]. Therefore, if the purpose is to examine protein-coding nucleotide sequences only, full-exome sequencing (WES) is a more cost-efficient technique than WGS to entirely examine 22000-25000 genes when looking for rare pathological genetic variants (Single Nucleotide Polymorphism, SNP, insertions, deletions) that may underlie a disease [Suwinski *et al*., 2019].

Though WGS is believed to have apparent advantages over WES, the diagnostic value and ability of these two methods to find clinically significant mutations does not exhibit any critical difference. Versatile investigations show that genome-wide tests detect 44 to 50% of all pursued mutations, whereas the performance of exome tests is only 2% inferior. This is largely conditioned by the fact that exome analysis utilizes longer fragments and more advanced probe designs [Barbitoff *et al*., 2020]. Considering its relatively low cost, WES is currently more attractive in clinical setting.

In technical terms, it is the choice of a probe set for the targeted hybridization (capture) of protein-coding genes (exome), rather than the platform, that is critically crucial for WES. Various kits are available on the market – Agilent SureSelect XT, Agilent SureSelect QXT, NimbleGen SeqCap EZ, and Illumina Nextera Rapid Capture Exome. While all of them include biotinylated DNA that hybridizes with DNA library fragments, they operate different DNA fragmentation methods, as well as coverage lengths and areas of target genes [Suwinski *et al*., 2019]. In addition, WES allows to significantly minimize the size of the analyzed database to 5-6 GB, in contrast to WGS (90 GB) [AllSeq, 2018].

In contrast to clinicians, researchers usually pursue somewhat different goals, focusing on how to obtain extensive information about the genome overall. Therefore, large-scale genomic research projects would prefer whole-genome sequencing – i.e. 100,000 Genomes, Russian Genomes, etc. [Zhernakova *et al*., 2020]. The near future is likely to see a combination of mass parallel sequencing techniques in routine practice, making it possible to identify up to 99% of all known, as well as earlier undetected mutations.

Since the introduction of NGS, diagnostics of orphan diseases has made a significant leap forward. New technologies have clarified the prevalence rates of numerous monogenic diseases. For instance, Stargardt disease was found to be the most prevalent monogenic pathology in the Northwestern region of Russia, rather than cystic fibrosis, also called mucoviscidosis [Barbitoff *et al*., 2019]. Diagnostics of certain oligogenic diseases has seen improvement as well. For example, exome sequencing is 10-fold more efficient in diagnosing MODY mutations [Glotov *et al*., 2019].

In addition, exome sequencing is gaining prominence as a tool to find new genetic markers of multifactorial diseases. For instance, in combination with a unique bioinformatic approach, this method enabled us to identify over 10 new markers for type 2 diabetes mellitus [Barbitoff *et al*., 2018].

NGS Technical Highlights

NGS, like any other teconology, shows a set of drawbacks due its technical characteristics. Thus, considering the significant size of generated WGS or WES data, data processing and analysis becomes a bottleneck, making it challenging to differentiate small mutations from random errors generated during sequencing [Hofmann *et al.*, 2017]. In addition, a major limitation of WES is the uneven coverage of sequence reads over the exome targets, contributing to many low coverage regions, which affect the downstream analysis and hinder accurate variant annotation, causing missed variant calls [Wang Q. *et al.*, 2017]. WES data can include inconsistencies, such as anomalies and outliers, or inconsistent speed at which data is loaded into the repository, alongside with inherent limitations, such as GC bias, difficulties in discriminating paralogous sequences or in phasing alleles, or linking sequence variants with biological data and phenotype. Translation of sequencing findings into easily understood medical standards, similarly to clinical diagnostic scoring, may present another potential limitation [Suwinski *et al.*, 2019].

Bioinformatic processing constitutes one of NGS dimensions. Therefore, to identify potential pathogenic genetic variants, protocol settings shall be configured appropriately [Barbitoff *et al*., 2017; 2020]. Importantly, the reference human genome sequence can occasionally include the so-called reference minor alleles - RMAs (allels of the reference genome that include rare pathogenic variants) and thus is prone to misinterpretation. Bioinformatic analysis allows to rectify such errors [Barbitoff *et al.*, 2018].

During last 20 years, the automated Sanger technique has become a prevalent approach to genome sequencing in humans, animals, bacteria, and viruses. However, a need for more rapid routine genome screening required some novel technologies of multiplex DNA sequencing. In our research we compared of two methods (Sanger and NGS) and their efficiency evaluation. We selected DNA samples from potential hematopoietic cells donors and conducted a comparative analysis by Sanger and NGS method. NGS method allowed detecting rare or novel variants of alleles. This approach is confirmed to be more sensitive and more cost-effective, especially in large HLA-typing laboratories [Glotov O. *et al.*, 2018].

Nevertheless, despite lower productivity and higher cost, Sanger sequencing remains relevant for a variety of practical tasks and is commonly used to find new or known mutations within a relatively small DNA molecule region (100 to 1200 bp), still considered the ‘golden standard’ to verify NGS-detected mutations in patients (Fig. 2).

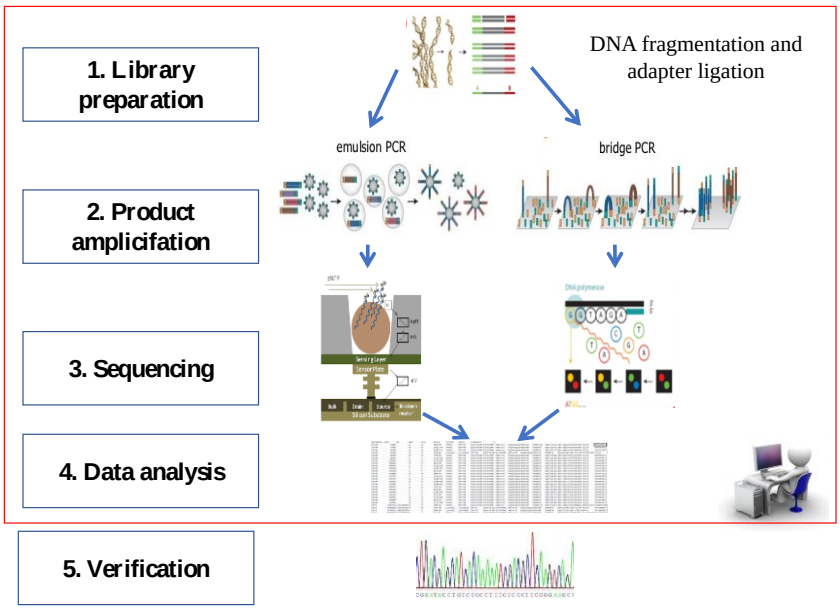


Figure 2. New generation sequencing procedure shown step-wise.

Sanger sequencing is effective to search for clinically relevant mutations in hot spots of well-studied genes, to determine nucleotide sequences of short (with few exons) and noncontiguous genes, to study short repeats [Baranov *et al*., 2021], or to identify a 'second' pathogenic variant when used together with or following NGS [Fedyakov *et al*., 2021].

Genetic terminology

Of note is are critically important updates in genetic terminology regarding the newly emerged methods. Instead of common terms 'mutation' and 'polymorphism', in 2015 and in 2017 the American College of Medical Genetics and Genomics (Eng. American College of Medical Genetics and Genomics (ACMG) and the Russian Society of Medical Genetics, respectively, recommended to use the term 'nucleotide sequence variant' with the following modifiers: (1) pathogenic, (2) likely pathogenic, (3) uncertain significance, (4) likely benign, or (5) benign [Richards *et al*., 2015; Ryzhkova *et al*., 2017, 2019]. A specific website [https://mutalyzer.nl.] recommends tools for correct description nucleotide sequence variants according to HGVS nomenclature. Since then, these terms have been actively used both in diagnostics [Ryzhkova et al., 2019], and in research. The term 'variant' and its five-tier

classification are relevant to explicitly characterize a particular variant and its function in the genome.

Today, there is a clear understanding that genetic variants are the main carriers of predictive information on disease pathogenicity and possess two main characteristics:

- penetrance (the percentage of carriers of the corresponding genotype who exhibit the trait);
- expressivity (varying manifestation of the trait in individuals with the same genotype) [Inge-Vechtomov, 2010].

Thus, the terms proposed back in 1925 by Timofeyev-Ressovsky turned out to be of such massive importance and so much ahead of their time [Inge-Vechtomov, 2010] that today they offer an explanation why ‘mutation’ is receding into obscurity as a term, replaced by the term ‘variant’ with its five modifications [Ryzhkova *et al*., 2019].

Genome Medicine and Its Prospects

Knowledge of gene structure, peculiarities of genetic polymorphism, and functions of different variants in the genome with regard to population specificity provide insight into the hereditary nature of a particular monogenic or multifactorial disease (MFD), and contribute to its diagnosis, prevention and treatment, as well as promote us to revise the classification of human diseases (Figure 3); this in its turn leads to a paradigm shift and propels the advent and rapid development of molecular medicine as a new science [Collins, McKusick, 2001; Peltonen, McKusick, 2001].

The current health care system operates in such a way that traditional medical care is usually provided to an ailing individual. Throughout historical development, information about healthy individuals and the so-called pre-nosology diagnostics prior to the disease manifestation has often remained obscure. Today, health care is on the cusp of a massive transformation. The main focus of medical care should shift towards disease risk prediction, early (potentially preclinical) diagnostics defining the disease

stage, followed by timely adequate intervention (pharmacological, nutritive, etc.) in order to prevent the disease onset or its transition to a more advanced stage. These principles shape the foundation of a fundamentally new 'three Ps' strategy - predictive, preventive and personalized medicine (PM) [Baranov *et al*., 2000; Taitz, 2019]. It should be noted that the terms PM and genetic passport (GP) arrived to St. Petersburg as early as in 2000 and are gradually becoming rooted in the popular mindset [Baranov *et al*., 2021].

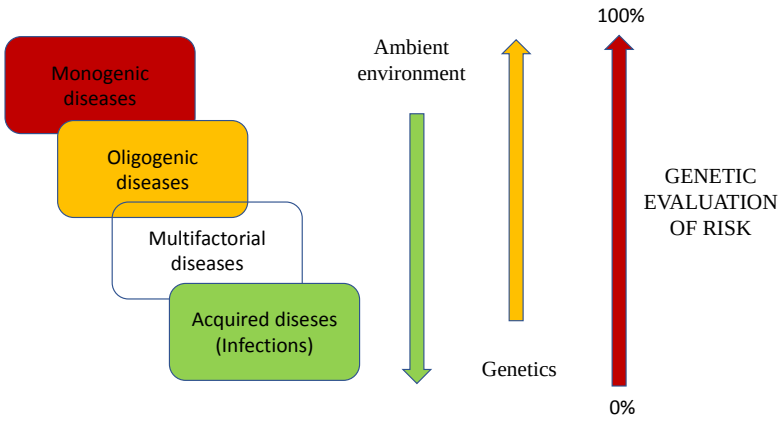


Figure 3. Classification of human diseases in genetic perspective.

Figure 4 shows the main stages of PM evolution from 3P (predictive, preventive, personalized) to 4P medicine (participatory) and then over to translational integrative medicine, eventually to precision medicine.

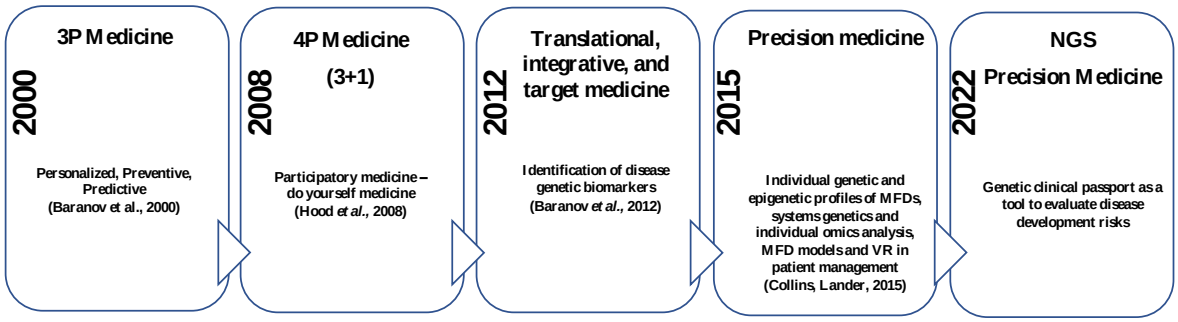


Figure 4. From the genetic passports to the NGS clinical genetic passport (by Baranov *et al.*, 2021 w. modifications).

Since 2008 physical mapping allowed to identify over 1,500 human MFD-associated genes. Russian laboratories and centers have analyzed multiple MFD risk candidate genes; for additional information see monographs and reviews by [Baranov *et al.*, 2000; Baranov and Khavinson, 2001; Puzyrev *et al.*, 2007]. Information on candidate genes and MFD-associated variants is available in various international databases and catalogs: OMIM [http://omim.org], HUGE NAVIGATOR [https://phgkb.cdc.gov/PHGKB/hNHome.action], Human Gene Mutation Database [http://hgmd.cf.ac.uk/ac/index.php], GWAS [https://ebi.ac.uk/gwas], ClinVar [http://et al, .ncbi.nlm.nih.gov/clinvar], dbSNP [http://ncbi.nlm.nih.gov/snp], Human Genome Variation Society [http://hgvs.org/dblist/dblist.html], Leiden Open Variation Database [http://lovd.nl], DECIPHER [https://decipher.sanger.ac.uk], 1000 Genomes Project Database [http://browser.1000genomes.org/index.html], Exome Aggregation Consortium Exome Database [http://exac.broadinstitute.org], Genome Aggregation Database [http://gnomad.broadinstitute.org], etc.

According to such luminary experts studying the human genome as Francis Collins, Victor McKusick, Leena Peltonen and others, the golden age of predictive medicine is to arrive in 5-10 years, with 4Ps medicine gradually evolving into 10Ps medicine to actively involve patients themselves [Taitz, 2019]. 10 Ps medicine shall comprise the following parameters [Taiz, 2019]:

11. predictive;
12. preventive;
13. personalized;
14. participatory (involving patients);
15. practical;
16. permanent;
17. proactive;
18. positive;
19. precision;
20. promotional.

Thus, molecular medicine and daughter sciences - predictive medicine (genetic passport, pharmacogenetics, gene therapy) – are the result of extensive expansion of genetic knowledge into medical science. Molecular medicine is specific for its individualized approach, since it is based on molecular structure of human genome. Due to the uniqueness of the individual genome, molecular medicine is aimed at rectifying pathological process in a particular patient, based on their unique genetic features [Baranov, 2000; Baranov, Baranova 2018]. Another most important aspect of molecular medicine is prevention. Complete genome information can be obtained long before the disease onset and, if necessary, even before birth. Hence, if properly set-up, prevention can complete eliminate a severe disease or largely prevent its development.

The contribution of genomics to medicine can hardly be overestimated. Its most significant achievements include:

- accurate, effective and universal methods to diagnose hereditary diseases at any stage of ontogenesis, including before birth (prenatal diagnosis);
- new scientific and practical fields of genomic medicine - oncogenomics, cardiogenomics, immunogenomics, reproductive genomics, aging genomics, nutrigenomics, sports genomics, metabolomics, proteomics, microbiomics and others; molecular tests for personal identification, i.e. genomic dactyloscopy; microbiome studies;
- experimental and clinical gene therapy framework for hereditary and non-hereditary diseases;
- cellular models of hereditary diseases and development of effective genome editing methods;
- diagnosis and personalized treatment (pharmacogenetics and pharmacogenomics) of monogenic and prevalent MFDs;
- NGS for non-invasive diagnostics of rare (orphan) diseases, genetic or chromosomal diseases at all stages of fetal development;

- preconceptional genetic tests and molecular framework of preventive (predictive) medicine.

Therefore, now we can already talk about NGS technologies standing on the guard of human health:

- Exome/Clinical Exome for NGS sequencing of a diseased child's DNA sample and family screening (carrier);
- Genetic Passport;
- PGD;
- NIPT.

In clinical practice the critical limitations of targeted sequencing technologies (WGS, WES, RNA-seq, ChIP-seq) are high cost, which includes the cost of data storage, transfer, processing, and bioinformatic analysis in addition to the cost of reagents and equipment [Schmidt and Hildebrandt, 2017]. For example, depending on the number of genes analyzed, in France gene panel sequencing may cost €376 to €968 [Marino et al., 2018]. Tan and colleagues calculated the average cost of gene panel sequencing per sample standing at $1,609 ($488-3443) based on 10 studies in the United States [Tan O. *et al*., 2018]. Another problem is the clinical interpretation of sequencing data. Knowledge of sequencing data is insufficient to explain etiology, pathogenesis, and symptoms of multifactorial diseases due to the strong influence of external factors and lifestyle, in addition to genetic factors [Lightbody *et al*., 2019; Suwinski *et al*., 2019].

Today, under the initiative of Francis Collins, head of the Human Genome Program and now director of the US National Institutes of Health, a special institute (Patient-Centered Outcomes Research Institute) was established to implement the concept of predictive medicine. To improve the quality of diagnostics, prevention and treatment of most prevalent diseases, the institute compiles patient medical records with lab tests results and individual genome data. In 2015, under involvement and approval by U.S. President Barack Obama, the Precision Medicine program was launched in the United States to include 1 million people and conduct extensive

clinical and genomic research. According to Francis Collins, the results of research in such a large group would give evidence to support the concept of precision medicine. The American project ‘All of Us’ is an ideological follow-up to the Precision Medicine initiative [https://allofus.nih.gov/news-events], aimed at solving the problems of personalized medicine. Its goal is to find and improve ways to interpret genomic and medical data. The project involves major American universities to collect health status data of over 200,000 volunteers using DNA samples, tested for 59 genes of severe hereditary diseases. Along with the England Genomes Projects, We All project should represent a world's largest genomic and clinical database. By March 2022 it featured over 100,000 complete genomes and 165,000 chip datasets [https://allofus.nih.gov/news-events/announcements/research-roundup-genomic-data-release-opens-new-paths-discovery].

It is noteworthy that two years earlier (in 2013), the European Society for Predictive, Preventive and Personalized Medicine (EPMA) was founded, while in 2015 the Personalized Medicine for Europeans: Toward More Accurate Medicine to Diagnose, Treat and Prevent Disease program was published [Golubnitschaja [Golubnitschaja *et al*., 2016; https://epmanet.eu].

Unfortunately, neither the American PM program nor the European Society mention the pioneering PM research efforts in Russia, though their results are regularly reported and published both in Russia and globally [Baranov *et al*., 2021].

Today, despite the persisting PM challenges, all programs emphasize that NGS has played a particularly important role in genomic medicine advancements, greatly increasing the clinical significance of testing. In 2016 ‘at consumers’ request’ the FDA (USA) lifted the ban on prognostic genetic testing (GT) for ten prevalent MFDs, including Crohn’s, Alzheimer’s, Parkinson’s diseases, prostate cancer, breast cancer, etc., imposed on 23andMe company [Greens, 2015]. Currently, many large Western companies (e.g. Systemas Genomicos) offer genetic tests to assess hereditary predisposition to 100 MFDs. The GT prediction accuracy of some diseases (Crohn’s disease, prostate cancer) has improved to 20% (prostate cancer) and even to 80%

(Crohn’s disease). The success relies on comprehensive kits with high marker density, which allow to study several hundreds of susceptibility genes at a time or specific loci with relevant gene clusters. DNA sequencing followed by bioinformatic analysis allows to identify pathogenic variants [Manolio *et al.,* 2010; Carere *et al.,* 2016].

Thus, current advances in PM and their practical value depend on genome sequencing quality and functional analysis within systems genetics paradigm. The EPMA program suggests a PM roadmap that predominantly relies on mass sequencing of individual genomes to elucidate their population, ethnic, social, and even interstitial features. Integrative analysis of gene expression and protein-protein interactions allows to yield individual omics profiles to be compared with the patient’s clinical and lab data. Based on such data, integrated gene networks are identified for the patient's organs and systems, most susceptible to pathological processes, and different future developmental scenarios are analyzed. As a result, patients are both the source of information and PM data users [Baranov *et al.*, 2021].

Today, medical genetics and PM evolution can be visualized as a timeline of pivotal advancements in genetic research and technology (Fig. 7), driven by the rigorous efforts of Russian and international research teams to study the genetic implications of monogenic, multifactorial and infectious human diseases and secure transition to the next stage of research.

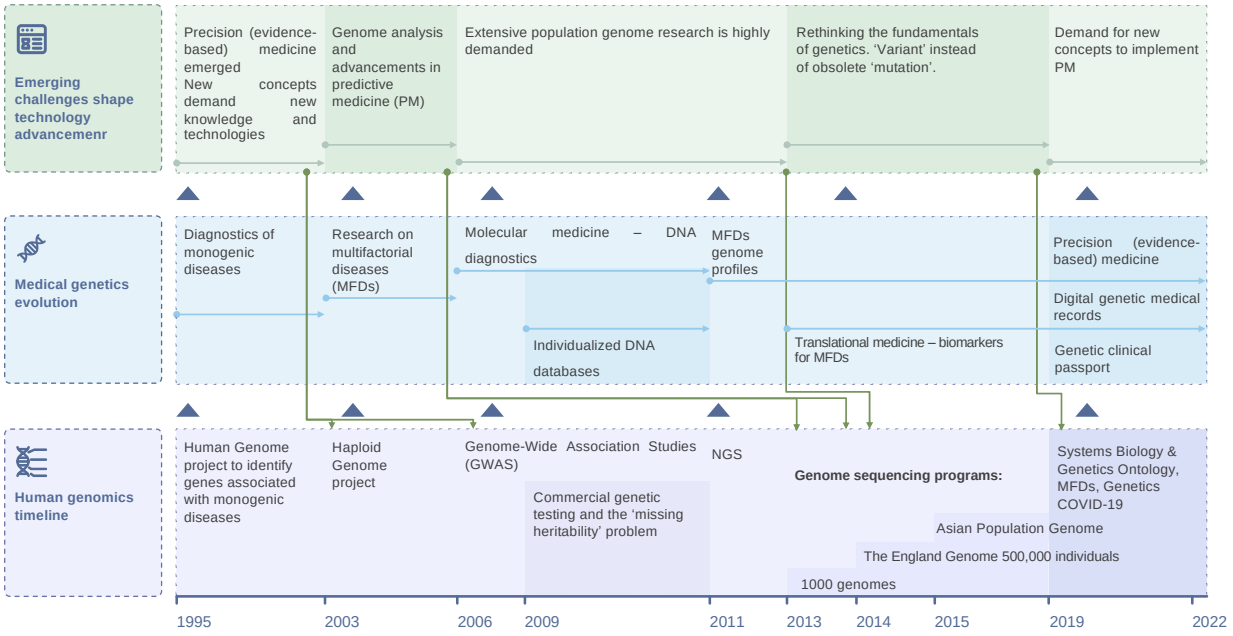


Figure 5. Timeline of pivotal advancements in medical genetics research and technology.

It is important to emphasize that the development of precision medicine research framework to study, diagnose, and treat monogenic, oligogenic, multifactorial and infectious diseases depends on the efficiency of NGS technologies, including modern analysis algorithms and classical genetic concepts of expression and penetrance. Further stages of practical implementation of genetic testing include: presymptomatic (pre-emptive) genetic testing (GT) in high-risk families; prospective GT and mandatory follow-up monitoring of high-risk individuals based on testing results; randomized predictive testing [Barnoy, 2007, Baranov *et al.*, 2021].

The purpose of our study is to determine the risk factors for socially significant diseases based on exome sequencing results and to develop methods allowing identification of clinically relevant gene variants in order to assess the risk of monogenic, oligogenic and multifactorial pathologies, as well as the severity of some viral infectious diseases in the population of Northwest Russia, as a case in point.

The study pursues the following objectives:

8. based on exome sequencing results, to characterize the structure of monogenic diseases and responsible gene variants in the population of Northwest Russia;
9. to evaluate the NGS efficiency in detecting previously undescribed pathogenic gene variants associated with mono- and oligogenic diseases, including combined pathologies;
10. to compare the efficacy of NGS technology with other PCR molecular genetic methods in detecting pathogenic variants in patients with monogenic diabetes and Wilson disease;
11. based on exome sequencing results, to assess manifestation variability of oligogenic and multifactorial pathologies in small patient cohorts of Northwest Russia;
12. to evaluate the efficiency of patient phenotype prognostic modelling using regression models and based on diagnosed gene variants and their combinations;
13. to describe the NGS-detected spectrum of genetic variants associated with different grade of severity and outcomes of the new coronavirus infection COVID-19;
14. to develop a set of genetic examinations, including exome sequencing, to predict the risk of oligogenic and multifactorial pathology, as well as severity of some infectious diseases and to explain the pathogenetic nature of clinical manifestations for a few human genetic diseases.

Research novelty

The study yields new DNA NGS data on the prevalence of monogenic diseases in the Northwestern region of the Russian Federation, obtained using original bioinformatic protocols applied to international databases and local cohort data. The results show that 24.3% of variants in the Russian population have never been analyzed before. The study shows that the most frequent monogenic diseases with recessive inheritance type are phenylketonuria, factor VII4 deficiency, kyphoscoliotic type 2 Ehlers-Danlos syndrome, tyrosinase-negative oculocutaneous albinism, and Wilson disease. Pathogenic variants in the *ABCA4* (retinal dystrophy, Stargardt’s

disease) and *CFTR* (cystic fibrosis) genes demonstrate utmost frequency in Northwest Russia.

The study provides the pioneering frequency characterization of the rs554847663 variant in the *OTOG* gene associated with autosomal recessive deafness and the rs119473033 variant in the *SMARCAL1* gene causing Schimke's immune-bone dysplasia. More than 100 putative loss-of-function (pLoF) variants present in healthy patients at the time of examination are identified. The analysis estimates the prevalence of variant c.894G>A (1/262) and the frequency of heterozygous carriage of all pathogenic variants (1/130) in the *LIPA* gene in the Russian population.

Errors in the reference genome are identified and solutions are provided to boost accuracy of bioinformatic analysis. NGS allowed to describe new pathogenic variants in the *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT* genes; the paper also reveals unknown clinical scenarios for monogenic diabetes mellitus (MODI) and Wilson disease in patients having more than one genetic variant in a single or several target genes. For the first time in the Russian Federation, the paper sheds light on clinical cases when single individuals present the following combinations of hereditary diseases – joint inheritance of X-linked and autosomal dominant forms of ichthyosis; Wilson disease and hemochromatosis. Evidence shows that NGS increases diagnostic accuracy from 15% to 50% (for MODI) and from 75% to 96% for Wilson disease.

Moreover, the study is a breakthrough attempt to analyze clinical polymorphism of oligogenic diseases affecting the Russian population, as well as the oligogenic inheritance mechanism of hypertrophic cardiomyopathy, monogenic diabetes, and familial hypercholesterolemia. Four previously unknown inheritance models for hereditary cardiomyopathies are revealed, with only 2 informative dominant models. Evidence of incomplete penetrance is provided for the following variants: *MYBPC3* (c.977G>A and c.2678G>T), *CASQ2* (c. 1014+12delG). Cardiomyopathy protective action of *TNNT2* gene variants (c.97+151delC, c.223+92G>C and c.223+93C>G) has been discovered.

For the first time in Russian population, genetic variants in different target genes have been detected in a single patient: *GCK* and *HNF1A*; *GCK* and *BLK*; *GCK*, *BLK* and *WFS1*, with clinical presentation highly typical of MODI2. In addition, the study provides evidence for small cohorts of patients to demonstrate the value of exome sequencing and original bioinformatics approaches for identification of type 2 diabetes mellitus and obesity candidate genes.

The study discovered that the loci of rs328 in the *LPL* gene, rs11863726 in the *HBQ1* gene, and rs112984085 in the *VAV3* gene can be associated with DM2 and obesity, with rs6271 in the *DBH* gene, rs62618693 in the *QSER1* gene, rs61758785 in the *RAD51B* gene, rs34042554 in the *PCDHA1* gene and rs144183813 in the *PLEKHA5* gene associated with obesity and rs9379084 in the *RREB1* gene, rs2233984 in the *C6orf15* gene, rs61737764 in the *ITGB6* gene, rs17801742 in the *COL2A1* gene, and rs685523 in the *ADAMTS13* gene associated with DM2.

The link between rs328 in the *LPL* gene and both DM2 and obesity is for the first time revealed for the Russian population, with rs6271 in the *DBH* gene and rs62618693 in the *QSER1* gene being specific markers for obesity and rs2233984 in the *C6orf15* gene being a specific DM2 marker. Earlier unknown DM2-specific variants have been identified – rs9379084 in the *RREB1* gene, rs61737764 in the *ITGB6* gene and rs17801742 in the *COL2A1* gene, rs139972217 in the *TMC8* gene, rs61758785 in the *RAD51B* gene, rs34042554 in the *PCDHA1* gene and rs144183813 in the *PLEKHA5* gene.

The study also justifies the value of regression model for human growth phenotype prediction, based on genotype data in the *EFEMP1, ZBTB38, HHIP, LCORL, ADAMTSL3, CDH13, JAZF1, IGF1R, GHSR, CABLES1*, *IFNG*, *VDR3*, and *IGFBP3* genes, as well as 36 genetic markers to predict quantitative traits (BMI, WHR, CHL, LDL-CHL, VLDL-CHL), and *AGTR2, NOS3, CNB1, ADRB2* genes to assess vital lungs capacity (VLC). For the first time the critical genetic potential of VLDL-CHL compared to BMI, WHR, CHL, and LDL-CHL has been shown. Novel evidence is provided for the Russian population to substantiate the value of exome

sequencing in detecting monogenic disease risk factors and its role as a tool to reduce the risk of both monogenic diseases and MFDs, which proves that monogenic, oligogenic and multifactorial diseases are closely intertwined.

The study suggests a breakthrough prognostic model for cytokine storm (CS) development in COVID-19 patients allowing to discover genetic risk factors of COVID-19 infection in the Russian population. The number of S-protein mutations is shown to be higher in grade 3 severity compared to milder symptoms. Mutations in non-Russian SARS-CoV-2 lines are found to be associated with an increased risk of lethal outcome in the group of Russian patients. The frequencies of five variants (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) in the *ACE2* gene in Russian and European populations are proved to bear no difference. It is shown that rare variants in the ACE2 gene may play an indirect role in COVID-19 pathology by affecting critical healthy protein functions and exacerbating disease severity.840 exomes have been sequenced in Russian patients suffering from COVID-19. Eleven independent genetic variants in the *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX* genes were identified that are associated with quantitative traits, which in turn are directly related to COVID-19 severity and outcome. It is suggested that variants in the *ATXN1, PKHD1, SETX, PIEZO1,* and *CDH23* genes can directly impact the COVID-19 phenotype by altering the function (in the case of missense variants in the *ATXN1* and *CDH23* genes) or expression levels of respective genes; for three of the variants, their contribution to COVID-19 pathogenesis was detected.

The paper elaborates the rationale behind the term ‘variant’ preferred over the term ‘mutation’ with regard to monogenic and oligogenic diseases, MFDs, and COVID-19 studies. The demand for a novel genetic classification of diseases is justified on the basis of clinical and laboratory data, as well as molecular genetic analysis, considering the penetrance and expressiveness of genetic variants among the Russian population; hence, the concept of the clinical genetic passport (CGP) is

proposed as a practical Predictive Medicine (PM) tool for subsequent risk assessment of all types of diseases.

Significance of the study results in theory and practice

The obtained data expand existing knowledge of molecular genetic mechanisms behind monogenic (DLCL, hearing loss, Wilson disease, ichthyosis, Noah-Lax syndrome, Floating Harbor syndrome, anauxtetic dysplasia), oligogenic (arrhythmogenic cardiomyopathy/right ventricular dysplasia, familial hypercholesterolemia, monogenic and neonatal diabetes), and multifactorial diseases (DM2, obesity), as well as COVID-19 new coronavirus infection resistance.

The present research elaborates the mechanisms underlying disease development due to specific variants in a single or several target genes of the genotype. The information obtained on allele frequency of known and new pathogenic variants in the genes *ABCA4, ABCC8, ACE2, ADAMTSL3, ADAMTS13, ADRB2, AGTR2, ALDOB, ALMS1, ALOXE3, APOB,ATP7B,ATXN1, BCKDHB, BLK, BLM, C6orf15, CABLES1, CASQ2, CCNO, CDH13, CDH23, CFTR, COL2A1, COL7A1, CPLANE1, DBH, DNAJB2, EFEMP1, EIF2AK3, EOGT, F7, F8, FGG, FKBP14, FLG, GAA, GABBR2, GALT, GATA6, GCK, GDAP1, GHSR, GJB2, HBQ1, HHIP, HNF1A, HFE, HTT, JAZF1, IFNG, IGF1R, IGFBP3, ITGB6, KCNJ11, LCORL, LDLR, LIPA, LIPC, LPL, LZTR1, MSH2, MTO1, MYBPC3, MYH14, NEB, NOS3, NPC1, NPHS2, OTOG, PAH, PAX4, PCDHA1, PHGDH, PIEZO1, PKHD1, PKP2, PLEKHA5, QSER1, RAD51B, RMRP, RREB1, SBF1, SCN11A, SETX, SLC19A2, SLC26A2, SMARCAL1, SRCAP6, STAC3 STS, SURF1,TGM5, TNNT2, TYR, VAV3, VDR3, WFS1, ZBTB38* in the Russian population can be used to interpret NGS results in clinical practice and supplement national and international databases.

Bioinformatic approaches based on reference genome errors, population frequencies and other features are suggested to accurately interpret sequencing data.

Our study provides an insight into genetic diagnostic algorithms for MODI, Wilson disease, arrhythmogenic cardiomyopathy/right ventricular dysplasia, familial

hypercholesterolemia, and other diseases. The new findings regarding protective variants and their detection in the *TNNT2* gene (c.97+151delC, c.223+92G>C and c.223+93C>G), in particular, expand our understanding of pathogenesis.

Genetic variants detected in different target genes of an individual patient allow us to adjust counseling, surveillance and treatment strategies. For example, the *GCK* gene shows that its pathogenic variants can be associated with various nosologies, which allows to assume that nosology forms merely represent different modified variants of the same basic set of genes, without underestimating a certain impact of environmental factors.

In small cohorts of patients, original algorithms and bioinformatic approaches have demonstrated efficiency of whole exome sequencing (WES) technologies in finding new MFD markers in limited cohorts of understudied populations.

DM and obesity specific and nonspecific markers pave the road towards differentiated diagnostics and individual therapy modifying disease-specific metabolic pathways. The study proposes and validates new GLM models to predict height, BMI, WHR, CHL, LDL-CHL, HDL-CHL, and VLC. We demonstrate that the regression model is an efficient tool for phenotype investigation, based on the medical history, genetic and clinical data; the obtained results can be useful to produce polygenic MFD risk prediction scenarios. The research offers additional arguments in favor of exome sequencing to be deployed in genomic medicine as a tool to identify rare conditions that are often obscured by the general complexity of MFD diagnostics. In addition, the paper elaborates NGS algorithms for NIPT and PGD to provide a comprehensive approach encompassing the entire array of molecular, genetic, cytogenetic, and embryological methods for successful pregnancy planning. As an advanced concept, the genetic variant profoundly transforms our understanding of monogenic, oligogenic, and multifactorial diseases.

The suggested prognostic model for cytokine storm development in COVID-19 patients can be clinically practical to identify patients with poor prognosis at an early stage. The presented clinical and genetic correlations between COVID-19 cytokine

storm scenarios can serve as disease outcome predictors and indications for medical and genetic counseling.

Rare variants in the *ACE2* gene may indirectly impact COVID-19 pathology; therefore, careful monitoring and follow-up in COVID-19 patients with rare pathogenic variants is pivotal. Pioneering data suggest a correlation between new coronavirus infection severity and lethality, on the one hand, and the number of SARS-COV-2 S-protein mutations (both in Russian and non-Russian population), on the other hand, which offers an extra prognostic tool for the disease. The discovered frequency similarities in the five variants (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) of the *ACE2* gene in Russian and European populations explain common infection rates and disease severity. We identified 11 novel variants in the *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX* genes associated with quantitative traits and disease severity, with COVID-19 outcome predictive potential of ROC/AUC = 0.59. Low replication rates of results, although typical of genetic complexity, highlight the specific role of the study design and population structure in identification of genetic risk factors for infectious diseases.

Hereby, we have developed the concept of the clinical genetic passport and a new set of methods to identify genetic determining factors and interpret the detected variants. Overall, this confirms the clinical significance of genetic predictors in detecting groups of highest risk for monogenic and multifactorial diseases, as well as new coronavirus infection COVID-19.

Key findings presented for defence:

6. DNA NGS results for monogenic diseases compared across original bioinformatic protocols, based on both international databases and local data for the population of Northwestern Russia, enabled identification of new pathogenic variants in the *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT* genes, as well as pathologies associated with variant combinations (joint

inheritance of X-linked and autosomal dominant forms of ichthyosis, Wilson disease and hemochromatosis).

7. Once underpinned by classical molecular genetic analysis (using PCR, PCR-PFLP, Sanger sequencing), NGS demonstrated higher efficiency in patients with clinically manifest monogenic diabetes mellitus and Wilson disease, with the rate of pathogenic variant detection increasing from 15% to 50% and 75% to 96%, respectively.

8. Variable manifestation of both oligogenic and multifactorial pathologies in Northwest Russia is associated with complex haplotypes of disease-associated genes, as well as their expression and penetrance, identified using exome sequencing and the original bioinformatic analysis adapted for small cohorts.

9. NGS-identified variants in the *ATXN1, CDH23, DNAJB2, EOGT, GABBR2, LZTR1, MYH14, PIEZO1, PKHD1, SCN11A, SETX* genes, including rare variants rs35803318, rs41303171, rs113691336, rs971249, rs2285666 in the *ACE2* gene, are associated with novel coronavirus infection COVID-19 severity and clinical outcomes assessed according to the original scale.

10. The study developed a set of predictive genetic examinations paving the way towards advanced predictive medicine and the concept of ‘human health genetic clinical passport’ based on exome sequencing data, thus enabling investigators to predict potential oligogenic and multifactorial pathologies and severe course of certain infectious diseases, as well as to explain the versatile pathogenetic mechanisms behind clinical manifestations of various diseases, taking into account expression, penetrance, pathogenic variant combinations in genes.

Reliability and Appraisal of Results

Reliability of results is ensured by the variety of implemented methods in conformity with the study purpose and tasks, statistical significance of findings, consistency of obtained data with clinical and experimental outcomes, as well as a

representative selection of samples (total number of biological samples collected from over 4670 patients).

The main results of the dissertation were presented and discussed within reports presented at Russian and international conferences and congresses, including: Biologie Prospective - Santorini Conference, 2004, Santorini Island; II International Conference 'Medicine of longevity and quality of life', 2006, Moscow, Russia; VI European Congress Internation Association of Gerontology and Geriatrics, 2007, St. Petersburg, Russia; 6th International Conference on Bioinformatics of Genome Regulation and Structure – BGRS 2008, Novosibirsk, Russia; 2nd International Conference on Genetics of Aging and Longevity, 2012, Moscow, Russia; 3nd International Conference on Genetics of Aging and Longevity, 2014, Sochi, Russia; congresses of Vavilov Society of Geneticists and Breeders, 2004, 2009 Moscow, 2014 Rostov-on-Don, 2019 St. Petersburg, Russia;

First International Scientific-Practical Conference MEDBIOTECH 2005, Moscow, Russia; All-Russian Conference 'Perspectives of Fundamental Gerontology', 2006, St. Petersburg, Russia; III International Anti-Aging Conference on Medicine of Longevity and Quality of Life', 2007, Moscow, Russia; International School Conference devoted to the 100th Birthday Anniversary of M. E. Lobashev on Systems Control of Genetic and Cytogenetic Processes, 2007, St. Petersburg, Russia; III International Scientific Conference 'Donozology 2007', 'Donozology 2009', St. Petersburg, Russia; All-Russian Seminar on 'Genetics of Longevity and Aging', 2008, 2009, Syktyvkar, Russia; 12th Puschino International School Conference for Young Scientists (Biology as a Science of the 21st century), 2008, Puschino, Russia; 5th All-Russian Conference with international participation on Prenatal Diagnostics and Genetic Passport as the Basis of Preventive Medicine amid the Age of Nanotechnology', 2012, Saint Petersburg, Russia; Russian Congress with international participation on Molecular Fundamentals of Clinical Medicine – Possibilities and Real Data, 2012, 2020, 2021, 2022 St. Petersburg, Russia; II International Congress on Medicine of Longevity and Life Quality, 2013, Moscow, Russia; All-Russian

Scientific and Practical Conference with international participation MOLECULAR DIAGNOSTICS, 2014, 2017, 2021, Moscow, Russia; II All-Russian Scientific Conference with international participation on Preventive Medicine 2014. Innovative methods to socially significant diseases diagnostics, treatment, and rehabilitation", 2014, Moscow, Russia; International Scientific Congress on 'Sport, Man, and Health', 2015, Saint-Petersburg, Russia; Conference of the Institute of Translational Biomedicine SPbSU (ITBM SPbSU) on Actual Problems of Translational Biomedicine, 2017, 2019, Saint-Petersburg, Russia; XI Scientific Conference on Human Genetics and Pathology dedicated to the 35th anniversary of the Research Institute of Medical Genetics, 2017, Tomsk, Russia; International Scientific and Practical Conference on NGS in Medical Genetics, 2018, 2019, 2021, Suzdal, Russia; XI All-Russian Congress of Neurologists, 2019, St. Petersburg, Russia; XXI Winter Youth School on Biophysics and Molecular Biology, 2020, St. Petersburg, Russia; All-Russian Scientific and Practical Conference on Fundamental and Applied Problems of Human Health Protection in the North, 2020, Surgut, Russia; III LABRIN-2021 National Congress with international participation, 2021 Moscow, Russia; Molecular Diagnostics and Biosafety Online Congress with international participation, 2021, Moscow, Russia; VI All-Russian Scientific and Practical Conference with international participation on Genetics of Hematopoietic Tumors - From Diagnostics to Therapy, 2021, St. Petersburg, Russia; 1st International Forum on Genomic and Biomedical Technologies 'From Birth To Active Longevity', 2021, Surgut, Russia.

In addition, The study results have been regularly presented at the poster session at the European Congress of Human Genetics (ESHG), 2002, Strasbourg, France; 2003 Birmingham, England; 2004, Munich, Germany; 2006, Amsterdam, The Netherlands; 2007, Nice, France; 2008, Barcelona, Spain; 2010, Gothenburg, Sweden; 2013, Paris, France; 2014, Milan, Italy; 2015, Glasgow, Scotland, United Kingdom; 2018, Milan, Italy; 2019, Gothenburg, Sweden; 2020, Virtual conference); Biologie Prospective - Santorini Conference, 2008, 2018, Santorini Island, Greece; the Anti-

Aging Medicine World Congress 2006, Paris, France; the 12th International Conference on Advanced Technologies & Treatments for Diabetes (ATTD 2019) Berlin, Germany.

The results of the work are presented and discussed in 35 research papers, including: 24 most relevant papers (whereof 24 papers are published in peer-reviewed journals indexed in international WoS and Scopus databases and 24 papers are published in peer-reviewed journals included in list of recommended academic journals by the Higher Attestation Commission (the VAK) of the Russian Federation Ministry of Education and Science), 1 monograph book, and 1 methodology guideline for physicians.

The dissertation results have been appraised in medical practice and education at the D.O. Ott Research Institute Obstetrics, Gynecology, and Reproduction, the Pediatric Research and Clinical Center of Infectious Diseases, St. Petersburg State University, the Engelhardt Institute of Molecular Biology (EIMB), St. Petersburg State Medical Institution City Hospital 40, N. L. Bochkov Medical Genetic Research Center, St. Petersburg State Center for Medical and Genetic Diagnostics, V.A. Almasov National Medical Research Center of the Ministry of Healthcare of the Russian Federation, St. Petersburh State Medical Referral and Diagnostics Center for Children, the I.I. Mechnikov North-West Medical University, the St. Petersburg State First Medical University of the Ministry of Healthcare of the Russian Federation, the I.M. Sechenov First State Moscow Medical University of the Ministry of Healthcare of the Russian Federation, the I.P. Pavlov First St. Petersburg State Medical University Ministry of Healthcare of the Russian Federation, the Institute of Experimental Medicine.

The dissertation relies on experimental and analytical results obtained by the author in person or under his direct supervision. The defendant coordinated and took part in the sampling of biologic material and clinical database compilation, prepared samples for molecular genetic analysis using NGS, Sanger sequencing, and PCR-PFLP methods, as well as supervised bioinformatic processing and statistical analysis

of obtained data. The author independently developed the study design and the framework of methods, selected indications for analysis in databases, identified genome regions for investigation, performed analysis of academic papers, summarized the clinical, laboratory, and molecular genetic results, presented the study results in papers and abstracts, and reported presentations at research conferences.

The results presented in the dissertation received supported via the following academic grants: Grant No. 09-04-13849-ofi_ts of the Russian Foundation for Base Research, Grant No. 14-50-00069 of the Russian Science Foundation, Grant No. MK-4113.2009.7 by the President of the Russian Federation, IAS SPbU Grant No. 1.38.79.2012 – as well as subsidies of the Committee for Science and Higher Education of the Government of St. Petersburg dated 2008, 2009, and 2011.

The dissertation comprises an introduction, three chapters and Conclusion, 242 pages in total, including 35 figures, 26 tables, and 341 references.

CHAPTER I. First Generation Sequencing and Human Monogenic Diseases

By 2003 the Human Genome Program and likewise projects identified genes and examined mutations associated with 1,485 hereditary human diseases. Later, mutations for more 3,000+ hereditary syndromes and diseases of pronounced genetic origin were identified. At the same time, so far, no molecular genetic associations have been detected for over 1500 phenotypes [Hamosh *et al.*, 2021].

The OMIM database (as of August 02,2022) includes entries for 7221 hereditary diseases and syndromes and their molecular associations [https://www.omim.org/statistics/geneMap]. These include 6152 phenotypes associated with one single gene, i.e. showing monogenic nature of a genetic trait or syndrome. Presumably, further genome decoding and identification of new genes may not critically change these numbers, though providing more clarity regarding candidate genes underlying specific hereditary diseases, as well as modifier genes that can significantly affect the disease phenotype, if present in some or other variant.

Advances in medical genetics are largely driven by rapid technological progress, associated with whole-genome DNA coverage and reading the exact nucleotide sequence both in the entire genome and in exome, i.e. its protein-coding genes only. The arrival of DNA sequencing technology in the 1980s enabled us to decipher human genes sequence. Later, parallel DNA sequencing – i.e. Next Generation Sequencing (NGS) – also contributed to the development of medical genetics.

Rapid implementation of NGS in a relatively short period of time (1.5 years) allowed to sequence the genomes of almost 2,000 people in Europe (2013). Programs were launched to sequence 500,000 genomes of indigenous population of Great Britain (2015) and 200,000 people of North America (2015). Genomic projects were initiated in Asia (Genomes Asia (Japan, South Korea, China and others), Germany, France, USA, and the Netherlands). By now Estonia has already finalized its genomic project, having sequenced over 2,000 genomes and over 10,000 human exomes [https://genomics.ut.ee/en].

Genetic testing and national Biobanks have fundamentally transformed the profile of healthcare in European countries. The UK is expecting to develop a new genetic map for each citizen once the program to sequence 500,000 genomes is finalized. This is going to change the principles of healthcare organization, as well as individual therapeutic approaches for each patient. Genome editing technology allows to correct defective genes, paving the way for unparalleled business processes in insurance medicine [https://www.genomicsengland.co.uk].

Projects on the assembly of human genome(s), Haplotype Map, 1000 Genomes, etc. are presented in detail in the Introduction (see the section above) and in Table 2. Some projects are already completed and the results are available in publications: 1000 Genomes Project [http://www.internationalgenome.org/], NHLBI Exome Sequencing Project [https://evs.gs.washington.edu/EVS/], The Genome Aggregation Database [https://gnomad.broadinstitute.org/], The 100,000 Genomes Project [https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/].

In addition to the above-mentioned projects, annotating causative variants for multifactorial and rare monogenic diseases is the main focus of projects on exome sequencing (The NHLBI (MD, USA) Exome Sequencing, CHARGE consortia), personal genome sequencing (The Personal Genome), or genotyping of variants from Icelandic volunteers (NextCode Health) (Table 2).

Table 2. Most prominent genome and exome projects [by Gonzalez-Garay, 2014].

Project name	License	Description	Source
HapMap	Free access	HapMap project focuses on SNPs with a minor allele frequency of ≥5%	The International Hapmap, 2007
1000 Genomes	Free access	1000 Genome project captured up to 98% of the SNPs with a minor allele frequency of ≥1% in 1092 individuals from 14 populations	Genomes Project Consortium, 2012
The NHLBI (MD, USA) Exome Sequencing	Free access	The project is directed to discover protein-coding genes responsible for heart, lung and blood disorders. It analyses the allele frequency of each	www.evs.gs.washington.edu/EVS/

Project name	License	Description	Source
		SNP.	
The Personal Genome	Free access	The Personal Genome Project has the genomes of 174 individuals and the exomes of over 400 volunteers.	www.personalgenomes.org
NextCode Health	Commercial	The project has 40 million validated variants collected from the genotype of 140,000 volunteers from Iceland.	www.nextcode.com
CHARGE consortia	Free access	1000 whole exome data sets of well-phenotyped individuals from the CHARGE consortium	www. dnanexus.com/usecases-charge

Importantly, new technologies have made it possible to re-assess the prevalence rates for multiple monogenic diseases [Ng *et al.*, 2009]. Developments in software and bioinformatic approaches have underpinned efficient accurate analysis of sequencing data and interpretation of gene variants, i.e. classification according to assumed pathogenicity. This step is considered most challenging and critical on the way from raw genomic sequencing data to the sophisticated molecular diagnostics of diseases [Nykamp *et al.*, 2017; Richards *et al.*, 2015].

1.1. Population Genetics Research

As noted above, global human DNA sequencing projects among the most significant advances to evaluate clinical effects associated with gene variants – e.g. 1000 Genomes Project [Auton *et al.*, 2015], Genome Aggregation Database (gnomAD) or the National Heart, Blood and Lung Institute (NHLBI) TopMed program [Bick *et al.*, 2020]. The plain argument that many variants previously listed as pathogenic occur too frequently in healthy individuals to cause Mendelian-inherited disease has become the most powerful tool for reducing false-positive associations between gene variants and phenotypic manifestations. To this end, population allele frequency (AF) information is now widely utilized to interpret variants in clinical practice [Lek *et al.*, 2016].

Genetic structure of human populations has been extensively studied

worldwide. Non-reference (i.e. non-wild type) allele frequency in particular population is a most important factor influencing clinical interpretation of a genetic variant. Despite extremely large number of samples in the Gnomad data (125,748 for version v. 2.1), genetic variability in many regions of the world is still poorly understood. Many current large-scale genome projects aim to characterize variants prevalent in a particular country or region, such as Genomics England [Walter *et al.*, 2015]. Genetic variation in many regions around the globe is still poorly studied, and it is unlikely that theses gaps will be eliminated in the years to come. Many countries take efforts to fill the gap by running national genome projects (e.g., two recent studies of Qatari population) [Fakhro *et al.*, 2016; Rodriguez Flores *et al.*, 2014]. One of such initiatives, the Genomes Russia project launched in 2015, aimed at characterizing the spectrum of genetic variation in diverse ethnic groups across Russia [Oleksyk *et al.*, 2015; Zhernakova *et al.*, 2018]. However, the project is currently far from being completed, while Russia remains among the poorly studied regions in this respect [Oleksyk *et al.*, 2015].

The number of samples is insufficient to make assumptions about the prevalence of monogenic disorders. Therefore, we decided to analyze genetic variation in the Northwest Russia [Barbitoff *et al.*, 2019]. We used a set of 694 samples sequenced with both whole-exome sequencing kits (Agilent SureSelect V6, Illumina Nextera Rapid Capture, Roche SeqCap EZ MedExome, and Illumina TruSeq Exome) and clinical exome panel (CES; Illumina TruSight One sequencing kit), utilized in clinical molecular diagnostics and/or research in the labs of St. Petersburg.

Bioinformatic analysis of exome sequencing data was performed using a custom pipeline based on the bwa aligner [Li & Durbin, 2009], Genome Analysis Toolkit v. 3.5., and Picard tools v. 2.2.2. The pipeline was constructed according to the GATK Best Practices workflow [De Pristo *et al.*, 2011]. All samples were processed with identical pipeline settings and genotyped jointly using the cohort genotyping method in GATK. Variants were annotated with SnpEff/SnpSift [Cingolani *et al.*, 2012] using the following resources: 1000 Genomes project allele frequencies [Auton *et al.*, 2015],

gnomAD r.2.1 allele frequencies [Karczewski *et al.*, 2019; Lek *et al.*, 2016], ESP6500 allele frequencies [Fu *et al.*, 2013], pathogenicity predictions from dbNSFP [Liu *et al.*, 2016], ClinVar database v. 2019.

The entire protocol for NGS analysis deployed in all our published research projects, including read quality assessment, sequence alignment, variant detection, annotation, visualization, data interpretation, as well as tools for genome analysis are elaborated in our review paper [Shikov *et al.*, 2019].

For the complete set of 694 study participants, we identified a total of 463,100 variant sites inside targeted exome regions. We demonstrated that the vast number of these variants constitute a specific component of the genetic structure of the population of Northwest Russia. For example, 9.3% (42,913) of the identified variants were not reported in the latest dbSNP build 151.

As anticipated, the population of the Northwest of Russia is in close proximity to the non-Finnish Europeans in terms of the spectrum of genetic variants [Barbitoff *et al.*, 2019]. Overall, our analysis of the gnomAD dataset allowed us to observe the highest concordance of allele frequencies in the Northwest Russia population with the allele frequencies derived from the Non-Finnish European population (Fig. 6).

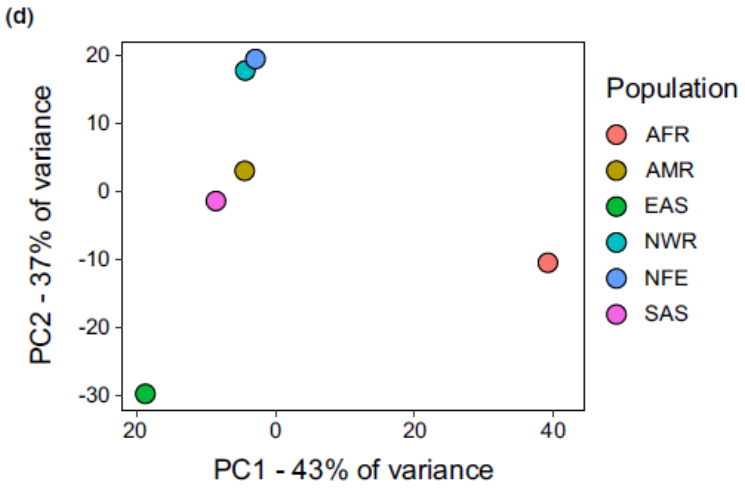


Figure 6. Analysis of allele frequencies for 121,171 exome variants present in Northwest Russia and all of the gnomAD (Genome Aggregation Database) populations (AFR, African; AMR, Ad Mixed American; EAS, East Asian; NFE, non-Finnish European; NWR, Northwest Russia; SAS, South Asia [Barbitoff *et al.*, 2019].

These findings support the need for a large-scale national genetic variation database in Russia, which would support both local and global clinical genetics research.

We also show that many of the previously reported pathogenic alleles that are highly prevalent in European population are also overrepresented in residents of Northwest Russia (Table 3), with the allele frequencies for many these alleles in Russia being substantially higher than in non-Finnish Europeans. These include a dominant rs76151636 allele in the gene *ATP7B* [de Bie *et al.*, 2007] and the common R408W (rs5030858) mutation in the *PAH* gene [Tighe *et al.*, 2003].

Таблица 3. Prevalent monogenic recessive disease alleles in the Northwest Russia dataset [Barbitoff *et al.*, 2019].

Rs ID	Gene	gnomAD AF	gnomAD NFE AF	p-value	Disease/ condition
rs5030858	*PAH*	7.6 × 10−4	0.0015	7.9 × 10−4 (1.1 × 10−5)	Phenylketonuria
rs36209567	*F7*	5.6 × 10−4	0.0010	0.0010 (5.7 × 10−4)	Factor VII deficiency
rs542489955	*FKBP14*	5.5 × 10−4	0.0010	0.0061 (9.6 × 10−5)	Ehlers-Danlos syndrome, kyphoscoliotic type, 2
rs61754365	*TYR*	2.9 × 10−4	3.2 × 10−4	1.1 × 10−4 (2.4 × 10−8)	Tyrosinase-negative oculocutaneous albinism
rs76151636	*ATP7B*	9.2 × 10−4	0.0013	0.0159 (0.0095)	Wilson disease
rs200488568	*SBF1*	3.3 × 10−4	1.4 × 10−4	1.2 × 10−4 (5.0 × 10−5)	Charcot-Marit-Tooth disease, type 4B3
rs549794342	*NEB*	2.7 × 10−4	4.7 × 10−4	0.0050 (5.9 × 10−5)	Nemaline myopathy
rs201544686	*MTO1*	1.7 × 10−4	2.0 × 10−4	4.9 × 10−4 (5.4 × 10−6)	Combined ox. phos. deficiency 10
rs386834233	*BCKDHB*	5.5 × 10−4	3.9 × 10−4	3.0 × 10−3 (2.3 × 10−3)	Maple syrup urine disease
rs775051461	*CCNO*	9.8 × 10−5	4.7 × 10−5	4.6 × 10−4 (0.0015)	Ciliary dyskinesia
rs121434233	*ALOXE3*	1.5 × 10−4	2.8 × 10−4	0.018 (5.2 × 10−5)	Autosomal recessive congenital ichthyosis 3
rs543206298	*NPC1*	7.6 × 10−5	1.1 × 10−4	0.0033 (5.2 × 10−4)	Polyneuropathy, Charcot-Marie-Tooth intermediate A

rs104894080	*GDAP1*	3.2 × 10−5	7 × 10−5	0.0013 (0.0044)	Polyneuropathy, Charcot-Marie-Tooth intermediate A
rs1307458231	*ALMS1*	2.0 × 10−5	4.4 × 10−5	5.0 × 10−4 (1.8 × 10−3)	Alstrom syndrome

Notably, we identified no prevalent pathogenic variants missing from ClinVar or dbSNP in autosomal recessive disease-related genes. This suggests that the majority of disease alleles are shared between Russian and European populations, at least for disorders with recessive inheritance pattern. These results allowed us to suggest preliminary estimates for the prevalence of monogenic disorders, based on the identified exome variants for the region (Table 4).

Table 4. Most disease prevalence in the Northwest Russia estimated from known pathogenic variants' frequencies [Barbitoff *et al.*, 2019].

Disease/ condition	**Gene**	**Allele count**	**Carrier frequency (lower/upper CI)**	**Disease frequency (lower/upper CI)**	**Known frequency**	**Comments/ references**
Retinal dystrophy, Stargardt disease	*ABCA4*	13 (23)	0.0350 (0.0206/0.0589)	3.1×10^{-4} (1.1×10^{-4}/8.8×10^{-4})	1 in 10,000	Zol'nikova, 2016; Sheremet *et al.*, 2017
Cystic fibrosis	*CFTR*	11 (19)	0.0296 (0.0167/0.0522)	2.2×10^{-4} (6.9×10^{-5}/6.9×10^{-4})	1 in 10,000	Reported carrier frequency of 0.032 (Abramov *et al.*, 2015)
Phenylketonuria	*PAH*	11 (18)	0.0296 (0.0167/0.0522)	2.2×10^{-4} (6.9×10^{-5}/6.9×10^{-4})	1 in 10,000	Reported carrier frequency of 0.029 (Abramov *et al.*, 2015)
Afibrinogenemia, congenital	*FGG*	7 (10)	0.0190 (0.0092/0.0387)	9.0×10^{-5} (2.1×10^{-5}/3.8×10^{-4})	n.a.	1in 1,000,000 (Mannucci *et al.*, 2004)
Hepatic lipase deficiency	*LIPC*	6 (14)	0.0162 (0.0075/0.0359)	6.6×10^{-5} (1.4×10^{-5}/3.1×10^{-4})	n.a.	—
Tyrosinase-negative oculocutaneous albinism	*TYR*	6 (12)	0.0162 (0.0075/0.0359)	6.6×10^{-5} (1.4×10^{-5}/3.1×10^{-4})	1 in 39,000	—
Peeling skin syndrome	*TGM5*	5 (8)	0.0135 (0.0058/0.0311)	4.5×10^{-5} (8.3×10^{-6}/2.5×10^{-4})	n.a.	—
Factor VII deficiency	*F7*	5 (7)	0.0135	4.6×10^{-5}	n.a.	1 in 500,000 (Wulff et

			(0.0058/0.0311)	(8.3×10^{-6}/2.5×10^{-4})		al., 2000)
Wilson disease	*ATP7B*	4 (6)	0.0108 (0.0042/0.0274)	2.9×10^{-5} (4.3×10^{-6}/1.9×10^{-4})	1 in 30,000.	Similar global incidence reported (Ala, Walker, Ashkan, Dooley, & Schilsky, 2007)
Ehlers-Danlos syndrome, kyphoscoliotic type, 2	*FKBP14*	4 (8)	0.0108 (0.0042/0.0274)	2.9×10^{-5} (4.3×10^{-6}/1.9×10^{-4})	n.a.	—
Fructose intolerance, hereditary	*ALDOB*	4 (7)	0.0108 (0.0042/0.0274)	2.9×10^{-5} (4.3×10^{-6}/1.9×10^{-4})	n.a.	—
Galactosemia	*GALT*	4 (5)	0.0108 (0.0042/0.0274)	2.9×10^{-5} (4.3×10^{-6}/1.9×10^{-4})	1 in 20,000	Reported carrier frequency of 0.006 (Abramov *et al.,* 2015)

Although the sample size limitation did not allow us to reliably estimate the degree of this discordance, our findings for the two most prevalent disorders – cystic fibrosis and phenylketonuria – are concordant with earlier estimates for individual genes [Barbitoff *et al.*, 2019]. Remarkably, our findings show that in Northwest Russia Stargardt disease is more prevalent, that cystic fibrosis, as was the belief earlier [Barbitoff *et al.*, 2019]. Then we decided to estimate the prevalence of every identified high-frequency variant in gnomAD populations, different from the Northwest Russia. Interestingly, we found that one of the variants, rs38683423 in *BCKDHB*, is also overrepresented in the Finnish population, possibly indicating either gene flow between Northwest Russia and Finnish population or region-specific selection against the related condition.

The research also looks into pathogenic variants for a number of human diseases. The results are formulated in Table 4, showing the following disorders with the most prevalence: (a) Stargardt disease *ABCA4* (MIM#601691) as the major gene, as also reported previously [Sheremet *et al.*, 2017], incidence at least 1:3226; (b) cystic fibrosis (the gene *CFTR*, MIM#602421, with the F508del (7:117199644:ATCT>A) mutation being the dominant variant), estimated incidence 1:5263; (c) phenylketonuria (the gene *PAH*, with incidence of up to 1:5556); (d) hepatic lipase deficiency (the gene *LIPC*, MIM#151670, 1:10000, with one pathogenic

variant rs113298164), and (e) tyrosinase-negative oculocutaneous albinism (the gene *TYR*, 1:13158). Our results are concordant with an earlier large-scale research into incidences of pathogenic alleles associated with cystic fibrosis in the non-Finnish European population [Lazarin *et al.*, 2013]. Moreover, the suggected incidence of phenylketonuria in Russia also shows similarity with the rresults reported for various other populations, including a recent study in China [Zhao et al., 2019]. [Zhao *et al.*, 2019]. Our estimates of cystic fibrosis, galactosemia, and phenylketonuria incidence were concordant with previous gene-level estimates [Abramov *et al.*, 2015; Abramov *et al.*, 2017]. On the other hand, estimated incidences of other diseases (factor VII deficiency and congenital afibrinogenemia) were approximately 20 to 100 times higher than the reported global ones [Mannucci, Duga, & Peyvandi, 2004; Wulff *et al.*, 2000]. Presumably, this discordance occurred due to the sample size limitation at the time of research [Barbitoff *et al.*, 2019].

With global databases on allele prevalence nevertheless critical, reference databases for individual populations contribute an extra value. The original ExAC paper by Lek et al. demonstrated that filtering candidate variants by highest allele frequency in various populations allows to significantly reduce potential disease-causing variants, observed in smaller exome data sets [Lek *et al.*, 2016].

Significant recent efforts in genomics were devoted to establishing more diverse and specific population reference data and other resources covering different ethnic and racial groups [Wong *et al.*, 2020; Martin *et al.*, 2018]. Many countries have conducted nationwide sequencing projects, including the Genome of the Netherlands (GoNL, [Boomsma *et al.* 2014]) or the Han Chinese Genome Database [Gao *et al.*, 2020]. As mentioned above, studies on the genomic variability of the Russian population are scarce and include the pilot Genomes of Russia project [Zhernakova *et al.*, 2020], our study on the prevalence of monogenic diseases in 694 patients [Barbitoff *et al.*, 2019], and target sequencing of 242 known disease genes in 1658 healthy people from the Ivanovo region [Ramensky *et al.*, 2021].

These works revealed some important aspects of genome variability in Russian patients. However, all these studies fail to provide a comprehensive analysis, either due to a limited sample size (e.g., Genomes of Russia) or due to a narrow set of analyzed genes [Ramensky *et al.*, 2021]. In order to increase the amount of data, 6096 samples collected from two the major Russian cities – Moscow and St. Petersburg – were analyzed to establish an extensive reference set of genetic variants [Barbitoff *et al.*, 2021].

Until now a most common approach to reliable allele frequency analysis relies on standardized and uniform processing of genetic data produced by multiple sequencing centers and genomic laboratories. At the same time, to successfully integrate the data, a centralized data-intensive framework, time and computational resources are required, as well as a potential procedure to share confidential or protected information. An aggregated and uniform pre-analytical procedure followed by different laboratories offers a possible solution to such challenges. Then, the variant analysis data is brought together and integrated for each sample. We implemented thus approach to bring together the data of two laboratories [Barbitoff *et al.*, 2021].

Following initial data aggregation and genotyping, the dataset was subjected to extensive sample- and variant-level quality control [Barbitoff *et al.*, 2021], leaving 5,268 samples and 2,092,456 variant sites. Of these, 349,811 variant sites overlapped with the ones reported in our previous publication [Barbitoff *et al.*, 2019]. Out of all variants, 75.7 % were known (found in the latest dbSNP build), and 24.3% (509,409) were novel. In total, 1,459,530 variants (69.8%) were either non-coding or silent coding variants; 579,974 (27.7%) were missense mutations or other moderate-impact variants; and 52,952 (2.5%) variants were putative loss-of-function (pLoF) variants. Similar to previous findings, rare and protein-damaging variants were greatly overrepresented among the novel variants. For example, only 23.6% of non-coding and silent coding variants were novel compared to as much as 41.3% of all pLoF variants. Likewise, 89.6% (1,875,600) of all variant sites were rare (MAF < 1% in the total sample) compared to as much as 99.2% (505,427) of novel variants. Finally,

allele frequencies of variants in samples of each of the participating parties showed a perfect correlation (r2 = 0.999), see Fig. 7. In their shape the clusters resembled different parts of a foot. Hence, we labelled them ‘heel’, ‘toes’, and ‘ankle’ accordingly.

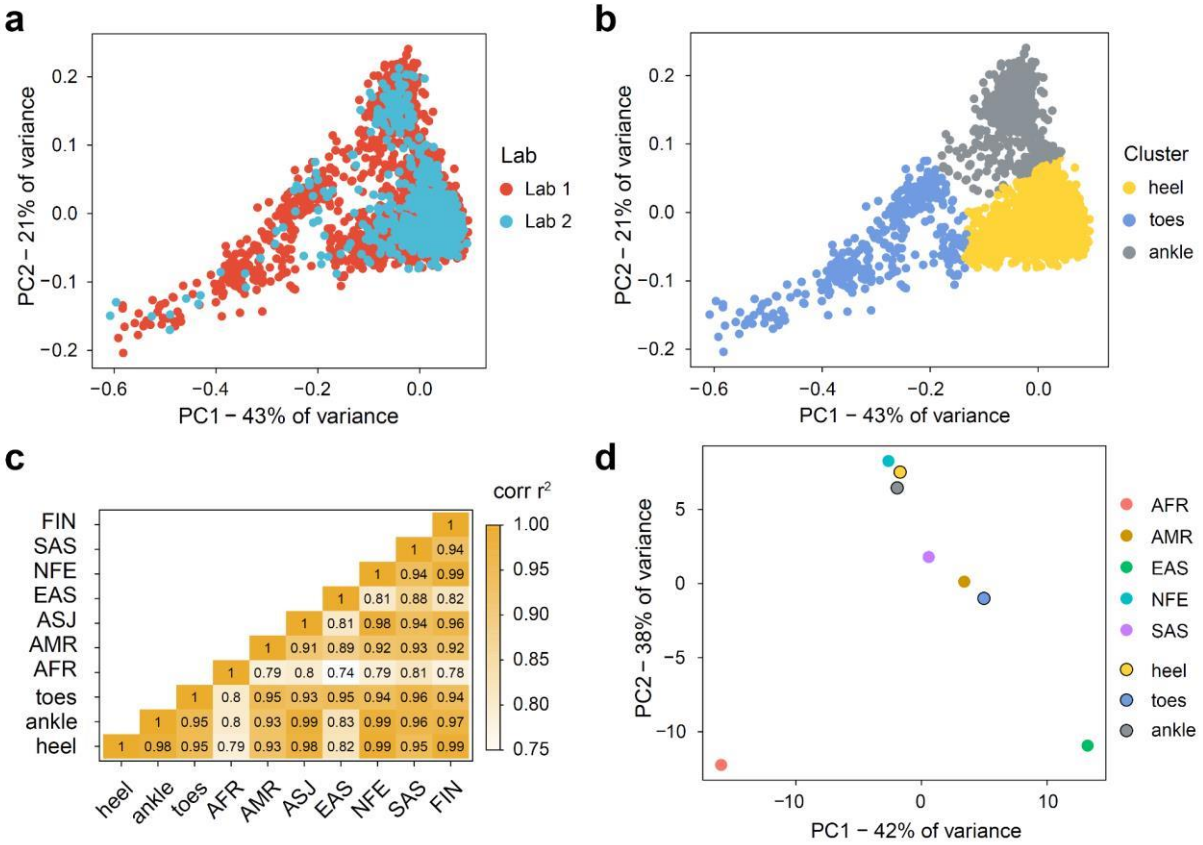


Figure 7. Analysis of the sub-structure of the admixed Russian population. (a, b) [Barbitoff *et al.*, 2021]. Note: Principal component analysis of the genotypes colored by the genetic centre **(a)** or the results of k-means clustering in the space of first 10 principal components **(b)**. **(c)** A heatmap showing Pearson's correlation between common variant allele frequencies in gnomAD ancestral groups and three clusters of individuals identified by k-means in **(b)**. **(d)** Principal component analysis of allele frequencies of common variants in gnomAD and three clusters (‘heel’, ‘toes’, ‘ankle’) of individuals identified by k-means in **(b)**.

Having identified the three distinct subgroups of samples (‘heel’, ‘toes’, ‘ankle’), we then decided to find out which of the global ancestral groups are closest to these clusters. To answer this question, we first analyzed the correlation between

common variant allele frequencies between each cluster and the main seven populations of gnomAD (African (AFR), Ad Mixed American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), European - Finnish (FIN) and non-Finnish (NFE), and South Asian (SAS)). The main cluster of Russian individuals (‘heel’) had the greatest correlation with the gnomAD NFE group (r2 = 0.991), closely followed by the Finnish population (r2 = 0.988). At the same time, the second (‘ankle’) cluster had a high degree of AF correlation with the NFE group (r2 = 0.987), was closer to Ashkenazi Jews than to the Finnish individuals from gnomAD (r2 = 0.987 and r2 = 0.980, respectively), and had a much higher AF correlation with both EAS and SAS. Finally, the third (‘toes’) cluster showed the greatest correlation with the SAS subpopulation from gnomAD (r2 = 0.975), and was equally distinct from EAS and NFE (r2 = 0.964 for both, Fig. 7(c)). In addition to the observations made by the analysis of AF correlation, we performed principal component analysis of common variant allele frequencies using data from the three clusters and five major gnomAD ancestral groups (AFR, AMR, EAS, NFE, and SAS). This analysis showed that, as expected, both the ‘heel’ and ‘ankle’ clusters were close to the NFE group, with the ‘ankle’ found to be closer to SAS/EAS. At the same time, the heterogeneous ‘toes’ cluster was much more distant and appeared closer to the gnomAD SAS (Fig. 7(d)).

Given these observations, we conclude that the first cluster represents the individuals of European ancestry, i.e. native residents of the Central and Northwest Russia; the second cluster represents Southern Russia and Northern Caucasus populations, while the third cluster corresponds to patients originating from the Siberian regions and/or Asian republics of the former USSR. These assumptions were further validated by the available patient information from both sequencing centers [Barbitoff *et al.*, 2021].

Nearly two dozen prevalent and overrepresented pathogenic variants were reported in the two major sequencing-based studies of the Russian population [Barbitoff *et al.*, 2019; Ramensky *et al.*, 2021]. We first questioned if the variants identified as overrepresented in earlier studies are also confirmed in our dataset.

Overall, 22 variants that were reported in these two publications: 14 in Barbitoff et al. and 10 in Ramensky et al., with 2 overlapping variants [Barbitoff *et al.*, 2019; Ramensky *et al.*, 2021, respectively]. Of these, overrepresentation of 10 variants was successfully validated using the healthy donor subset, and of 15 - in the complete set of 5,268 samples.

We next went on to identify all variants that showed overrepresentation in our dataset. As a result, 19 disease alleles were identified as overrepresented in the healthy donor subset (Table 5). These included both known high-frequency variants, such as the rs5030654 variant in *PAH* and the rs1555287300 in *ATP7B* linked to phenylketonuria and Wilson disease respectively, and variants that have not been previously reported as overrepresented. The latter category included such variants as rs554847663 in *OTOG* linked to autosomal recessive deafness, and variant rs119473033 in *SMARCAL1* causing Schimke immuno-osseous dysplasia. High incidence of these variants has not been previously noted [Barbitoff *et al.*, 2021].

Table 5. Known pathogenic variants present at high frequency in the expanded allele frequency reference of the Russian population [Barbitoff *et al.*, 2021].

Variant ID	Gene	gnomAD NFE AF	Allele count	RUSeq AF*	p-value	Disease
rs200482683	*NPHS2*	0.02%	5	0.35%	2.07E-05	Nephrotic syndrome
rs549794342	*NEB*	0.05%	8	0.56%	5.86E-07	Muscular dystrophy
rs119473033	*SMARCAL1*	0.01%	4	0.28%	4.62E-05	Schimke immuno-osseous dysplasia
rs775288140	*COL7A1*	0.00%	3	0.21%	8.84E-06	Recessive dystrophic epidermolysis bullosa
rs777686211	*CPLANE1*	0.02%	4	0.29%	3.41E-04	Nephronophthisis
rs104893924	*SLC26A2*	0.02%	5	0.35%	2.91E-05	Osteochondrodysplas ia
rs386834233	*BCKDHB*	0.04%	5	0.35%	2.83E-04	Maple syrup urine disease
rs782316919	*SURF1*	0.02%	4	0.28%	7.59E-05	Cerebellar ataxia
rs554847663	*OTOG*	0.08%	6	0.46%	5.86E-04	Intellectual disability and autosomal recessive deafness
rs371720347	*STAC3*	0.01%	3	0.21%	1.06E-04	Bailey-Bloch congenital myopathy
rs5030858	*PAH*	0.15%	10	0.70%	7.45E-05	Phenylketonuria
rs76151636	*ATP7B*	0.13%	9	0.63%	1.26E-04	Wilson disease
rs36209567	*F7*	0.10%	7	0.49%	8.87E-04	Abnormal bleeding
rs200389141	*BLM*	0.03%	5	0.35%	6.55E-05	Bloom syndrome
rs375470378	*GAA*	0.03%	4	0.28%	7.21E-04	Glycogen storage disease
rs387906455	*F8*	0.00%	3	0.21%	9.34E-07	Hereditary factor VIII deficiency disease

* - AF is given with reference to the healthy donor subgroup.

These estimates, therefore, have a significant clinical value for the Russian population.

Having characterized the spectrum of overrepresented pathogenic alleles in our dataset, we then went on to identify potentially clinically significant variants that are present in healthy patients but are not found in gnomAD. We began by identifying known pathogenic variants missing from gnomAD v2.1.1 data that are reported in the ClinVar database. In total, we discovered 72 such variants, with 25 of them located in genes with autosomal-dominant disease inheritance [Barbitoff *et al.*, 2021]. In addition to these variants, we also searched for potentially clinically significant variants that are absent from gnomAD but are present in the healthy subgroup. In total, more than 100 putative loss-of-function (pLoF) variants present in healthy patients were identified. Of these, 27 variants localized in genes with high degree of evolutionary conservation according to gnomAD-derived metrics (pLI, LOEUF) and with known connection to autosomal dominant phenotypes [Barbitoff *et al.*, 2021].

Among the known and expected pathogenic variants present in our sample, the most notable examples include variant rs397516471 in the *TNNT2* gene. This variant is reported as pathogenic for left ventricular noncompaction and familial restrictive cardiomyopathy; however, it is present in 1 healthy heterozygous subject in our dataset who has not yet displayed any symptoms of the disease. Another notable example is variant rs1064793825 in *MSH2* possibly causing hereditary colorectal cancer. Similar to the case of rs397516471 variant in *TNNT2*, a carrier subject has also not yet displayed any symptoms of the disorder. However, in the case of both rs397516471 and rs1064793825, family history of variant carriers listed several cases of related disorders (cardiomyopathy or cancer, respectively), suggesting that the disease is likely to manifest in the near future.

Overall, our results demonstrate the urgent need for population-specific genetic databases, which is essential for interpretation of variants and identification of disease risk factors in poorly studies populations. While the current sample size allows for more unbiased conclusions regarding the genetic structure of the Russian population,

we can still expect a large number of rare genetic variants in the rest of the population that were not covered by our analysis. Hence, further aggregation of data from sequencing centers across Russia, sequencing of more healthy donors, and inclusion of patients from distinct regions are all required to fully characterize the genetic variation spectrum of present-day Russia and bring forward evidence-based ideas regarding the prevalence of monogenic diseases and population frequencies of disease-associated alleles [Barbitoff *et al.*, 2019; 2021]. We nevertheless aspire that these findings would enable further research in medical and clinical genetics both in Russia and globally.

1.2. Variant Frequency Interpretation Using Population Research

As earlier said, population genetic studies provide an insight into the incidence of a particular disease and enable us to analyze the prevalence of certain disease-associated variants, e.g. lysosomal acid lipase deficiency in the Russian population [Fedyakov *et al*., 2018]. Lysosomal acid lipase deficiency (LALD) is a rare inherited progressive lipid metabolic disorder conducive of atherosclerosis, hepatosplenomegaly, cirrhosis, malabsorption and other symptoms. In absence of dedicated treatment, the forecast is unfavorable. Hence, timely diagnosis is critical. Different types of the disease are distinguished depending on the residual activity of the lysosomal acid lipase enzyme. Wolman disease is an early severe type, the enzyme activity is less than 1% [Abramov *et al.*, 1956; Aslanidis *et al.*, 1996]. Cholesterol ester storage disease is late type of LAL deficiency with a milder course and enzyme activity within 1 to 12% [Fredrickson *et al*., 1963]. The cumulative incidence of LALD is 1/40,000 to 300,000 for different populations [Lohse *et al*., 2000; Muntoni *et al*., 2007]. Prior to our research data on the disease incidence in Russia was missing. Expected numbers varied within 1/100,000 to 150,000 [Baranov *et al*., 2016; Strokova *et al*., 2017]. To estimate the LALD incidence in the Russian population, we used exome and genomic sequencing data of 523 Russian individuals from the Northwestern and Central regions. Samples in the dataset included blood samples of patients with monogenic (monogenic types of diabetes mellitus, hereditary connective

tissue disorders, hereditary metabolic disorders, and other orphan diseases) and multifactorial diseases (obesity, type 2 diabetes mellitus), as well as blood samples in the control population. The dataset included no patients with the clinically diagnosed LALD. The samples were genotyped to identify the pathogenic variant c.894G>A, allowing to detect the variant c.894G>A in 2 heterozygous samples out of total 523 in the dataset, which is concordant with the frequency of 1/262 in the study population. Thus, the estimated frequency for all pathogenic variants in the *LIPA* gene of heterozygous carriers can reach 1/130. According to the Hardy-Weinberg law, this corresponds to an estimated incidence of 1/67 600 [Fedyakov *et al*., 2018].

Remarkably, both a particular disease and a set of variants can be estimated in terms prevalence using population genetic analysis allows; for example, polymorphism in the *ACE2* gene associated with severe COVID-19 pathology and affects important normal protein functions. Frequencies for five variants (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) in the *ACE2* gene were compared for Russian and European populations. Similarity between Russians and other European populations suggest the disease may have similar incidence and severity. These findings were able to shed light on the epidemiological situation in March-April 2020, when the COVID-19 started to unfold [Shikov *et al*., 2020]. This research is embraced in detail in Chapter III of the dissertation.

Population analysis makes it possible to assess the risk of a pathogenic variant to develop a particular disease. Thought large genomic databases are still missing, this does not invalidate studies analyzing different groups of Russian patients, e.g. our study of tuberculosis [Bliznetz *et al.,* 2017]. Although mutations in the *GJB2* gene sequence account for the majority of variants causing autosomal recessive nonsyndromic hearing loss, several large deletions in the *DFNB1* locus can apparently contribute to deafness as well. Current genetic testing for hearing loss includes *GJB2* gene sequencing and analysis of two common major deletions – del (*GJB6-D13S1830*) and del (*GJB6-D13S1854*). In addition, we were able to identify another new 101 kb deletion, del (*GJB2-D13S175*). A multinational cohort of 1104 unrelated hearing loss

patients with bi-allelic mutations at the *DFNB1* locus exhibited the del (*GJB2-D13S175*) allele frequency at 0.5% (11/2208). This allele was found to be predominantly associated with profound neurosensory hearing loss. In addition, eight unknown mutations in the *GJB2* gene were described [Bliznetz *et al*., 2017]. All del (*GJB2-D13S175*) carriers were of Ingush origin. Among people with normal hearing, del (*GJB2-D13S175*) was observed in the Republic of Ingushetia with a carrier frequency of ~ 1% (2/241). Analysis of the haplotypes associated with the deletion revealed a common Ingush ancestor, with the deletion age ~3000 years. Analysis for del (*GJB2-D13S175*) was added to the routine hearing testing procedure [Bliznetz *et al*., 2017]. Thus, today detection of deletions, as well as point mutations entails huge diagnostic value, making a high-quality bioinformatic data analysis protocol a collateral.

1.3. Bioinformatic NGS Data Processing

Bioinformatic NGS data processing is encorporated in the NGS technology. Therefore, to identify the genetic background of a disorder, a high-quality bioinformatic protocol is a prerequisite [Barbitoff *et al.,* 2018; 2020]. At the onset, sequence mapping is initiated and the experimental reads are mapped to the reference genome from the RefSeq database of the National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/RefSeq/] (with the version specified) or Locus Reference Genomic [http://www.lrg-sequence.org/]. It is noteworthy that the reference sequence may occasionally contain inaccuracies due to the so-called reference minor alleles – RMAs (reference genome loci that include rare pathogenic variants that are occasionally misinterpreted). Bioinformatic analysis shall correct such errors. To this end a dedicated procedure was developed [Barbitoff *et al.,* 2018]. Step 1 in NGS data bioinformatic processing is variant calling – the process identifying variants from sequence data. It is followed by variant annotation that allows to assign functional information to DNA variants, including their pathogenicity (based on ClinVar, OMIM, etc.) (see Fig. 8 and Table 6). Scarcely known variants are

investigated using one of the sequence-based prediction tools (PROVEAN, SIFT, Polyphen, MutPred, etc.) (Table 7).

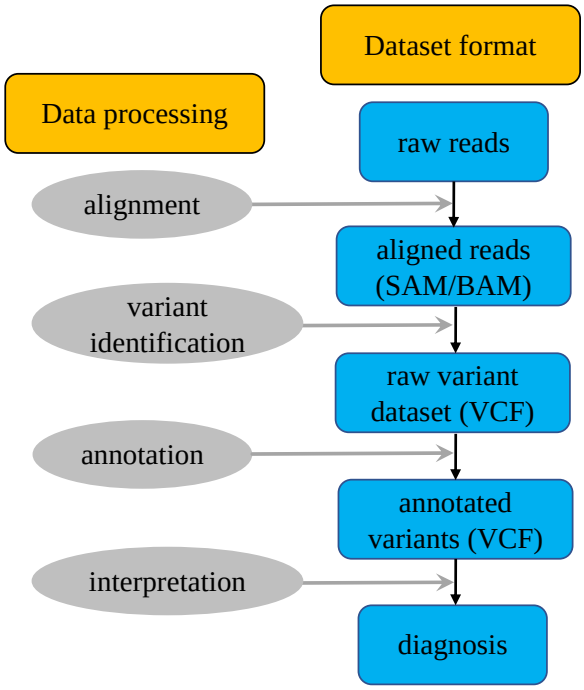


Figure 8. Algorithm for bioinformatics of sequencing data.

Variant ranking is the last stage of dataset processing. The identified and annotated variants are ranked according to a set of criteria, such as allele frequency in public databases, pathogenicity, substitute homo-/heterozygosity, etc. The major challenges at stage 1 are international genome databases and RMA detection, while stage 2 largely depends on availability of well-characterized databases and stage 3 depends on the performance of prediction tools and a multitude of training factors, i.e. high-capacity software, a dedicated database, and clinical diagnostics accuracy. While stages 1 and 2 depend on how skillful the bioinformatician is, at stage 3 clinical expertise and professionalism the geneticist or genetic consultant is of equal or even overriding importance.

1.4. NGS Data Interpretation

All current variant interpretation strategies and recommendations, e.g. the American College of Medical Genetics and Genomics (ACMG) recommendations

[Richards *et al*., 2015], the Russian Recommendations for Variant Interpretation [Ryzhkova *et al.*, 2018], or the SHERLOCK technique [Nykamp *et al.*, 2017], utilize AF in healthy populations to classify variant-associated effects. This becomes particularly critical for autosomal dominant (AD) diseases.

Implementation of NGS techniques brings up changes in the language of genetics as well. Instead of the custom term's 'mutation' and 'polymorphism', the term 'nucleotide sequence variant' is now preferrable with the following five modifications: pathogenic; likely pathogenic; uncertain significance; likely benign; benign [Ryzhkova *et al*., 2019]. More information on the accurate terminology to describe sequence variants and HGVS nomenclature is available at [https://mutalyzer.nl.]. All detected variants are classified by their pathogenicity. This involves studies on medical and scientific research papers and databases. Below we provide a list of recommended databases to search earlier described variants, see Table 6.

After curation using several platforms (except ClinVar), all assertions should be classified with respect to a disease and inheritance pattern. Curation of other variants is performed using two sets of evidence criteria: one for likely pathogenic variants and the other for likely benign variants. Each pathogenic evidence criterion lends to classification depending on the strength of evidence for pathogenicity: very strong (PVS1), strong (PS1-4); moderate (PM1-5), or supporting (PP1-5). Each benign criterion as very strong (independent) (BA1), strong (BS1-4), or supporting (BP1-6).

A points-based system has been developed to enable easier implementation of the criteria. Each variant is assessed based on the expert judgement with regard to specific parameters, that all together allow to assign the variant one of the five modifications: pathogenic, likely pathogenic, in, probably benign, benign. If a variant has conflicting evidence, some in support and some against pathogenicity, it is classified as a variant of uncertain significance [Ryzhkova *et al*., 2019].

Table 6. Recommended databases to assess pathogenicity of nucleotide sequence variants [Ryzhkova *et al*., 2019].

Population	
Exome Aggregation Consortium http://exac.broadinstitute.org/	A publicly accessible database of variants from exome sequencing spanning 61,486 of individuals for use as a global "reference set". While the individuals in the reference set aren't necessarily healthy and many participated in various disease-specific and population-based genetic studies, individuals with hereditary diseases manifesting in childhood were excluded from the sample.
genome Aggregation Database http://gnomad.broadinstitute.org/	An extensive genomic variant database based on the Exome Aggregation Consortium platform includes 123,136 exomes and 15,496 genomes.
Exome Variant Server http://evs.gs.washington.edu/EVS/	A database of variants from exome sequencing of several large human cohorts of European and African-American descent includes coverage data to account for missing variants.
1000 Genomes Project http://browser.1000genomes.org/index.html	A database of variants represents low and high coverage genome and target sequencing data from 26 populations. Though of greater diversity than Exome Variant Server, it contains data of lower quality and occasionally on related individuals in some cohorts.
dbSNP http://www.ncbi.nlm.nih.gov/snp	A database of short genetic variants (typically ≤50 bp) from various sources, that indicates population polymorphisms, as well as numerous pathogenic variants.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	A database of structural variants (typically >50 bp) from a multitude of sources.
Phenotypes	
OMIM http://www.omim.org/	A database of human genes and genetic states with a representative sample of disease-associated variants.
Human Gene Mutation Database http://www.hgmd.cf.ac.uk/ac/index.php	A database of annotated variants published in literature. Payment of fees is required to access the bulk of content. The dataset indicates polymorphisms, while clinical significance needs reference to literature for clarification.
ClinVar http://www.ncbi.nlm.nih.gov/clinvar/	A database of reports of the relationships among human variations and phenotypes, with assertions made regarding their clinical significance Includes low-quality data and is generally ***not recommended***. Use should be limited to the search of literature references.
Study-specific databases (loci-/disease- /ethnicity-specific, etc.)	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html	The website of the Human Genome Variation Society includes a list of thousands of databases that suggest annotation options for human variants of specific types. Significant share of databases is included in

Leiden Open Variation Database http://www.lovd.nl	Leiden Open Variation Database system.
DECIPHER https://decipher.sanger.ac.uk/	Molecular and cytogenetic database for doctors and investigators, that utilizes Ensembl genomic browser to bridge microchip genomic data and phenotypes.
Coding sequences	
NCBI Genome http://www.ncbi.nlm.nih.gov/genome	A resource for whole-genome reference sequences
RefSeqGene http://www.ncbi.nlm.nih.gov/refseq/rsg/	A resource for reference sequences of clinically relevant genes.
MitoMap http://www.mitomap.org/MITOMAP/	A corrected Cambridge reference sequence of human mitochondrial DNA.

Application of databases requires lab verification of the following information:

5. updates frequency, database curation (latest version and/or curation by a reputable institution);
6. confirmed application of HGVS nomenclature, specifying reference sequences for genome and transcript assembly in variant naming;
7. assessment of all quality parameters to evaluate data accuracy (may need reference to dedicated publications);
8. assessment of the study reliability and level of evidence to obtain variant-specific data.

In case a variant has never been described earlier in the literature and is missing from every database or no sufficient information about it is available, its relevance can be determined based on pathogenicity prediction findings. Table 7 below includes references to most common prediction tools, accompanied by a brief description.

Table 7. Most common pathogenicity prediction tools for sequence variants [Ryzhkova *et al*., 2019].

Website	**Input for prediction**
Missens substitutions	
Align GVGD -http://agvgd.hci.utah.edu/agvgd_input.php	Protein structure/function and evolutionary conservatism

MAPP – http://mendel.stanford.edu/SidowLab/ downloads/ MAPP/index.html	
MutationTaster - http://www.mutationtaster.org/	
MutPred - http://mutpred.mutdb.org/	
PolyPhen-2 - http://genetics.bwh.harvard.edu/pph2/	
PROVEAN -http://provean.jcvi.org/index.php	Variant and homologous protein sequence alignment and similarity measurement
SIFT- http://provean.jcvi.org/index.php	
nsSNPAnalyzer - http://snpanalyzer.uthsc.edu/	Multiple sequence alignment and protein structure analysis
Condel - http://bg.upf.edu/fannsdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
Splice site disruptions	
GeneSplicer - http://ccb.jhu.edu/software/genesplicer/	Markov Models
Human Splicing Finder - http://www.umd.be/HSF/	Variant position input
MaxEntScan – http://genes.mit.edu/burgelab/maxent/ Xmaxentscanscoreseq.html	Maximum entropy principle
NetGene2 – http://www.cbs.dtu.dk/services/NetGene2/	Neural Networks
NNSplice - http://www.fruitfly.org/seq_tools/splice.html	Neural Networks
ASSP- http://wangcomputing.com/assp/	Neural Networks
FSPLICE - http://www.softberry.com/berry.phtml? topic=fsplice& group=programs&subgroup=gfind	Species-specific site-splicing predictor based on the position weight matrix

1.5. NGS Application to Identify New Variants in Patients' Genome

New approaches to data interpretation and advanced bioinformatic algorithms allow us to describe earlier unknown, as well as well-studied variants. We refer to a clinical case of a new frameshift mutation in the *PKP2* gene detected using NGS molecular genetic testing [Fedyakov *et al.*, 2019].

Arrhythmogenic right ventricular cardiomyopathy/dysplasia (ARVC) is a progressive myocardial disorder primarily in the right ventricle. It develops mostly at a young age, manifesting often as a sudden cardiac death (SCD) due to malignant ventricular arrhythmias. Diagnosis via standard evaluation of cardiac function can be difficult because of minor and nonspecific clinical signs at an early stage, especially in the patient's relatives. Molecular genetic testing may provide more information for clinical decision making. Our study shows the proband and his three children to carry a pathogenic variant in the *PKP2* gene [Fedyakov *et al.*, 2019].

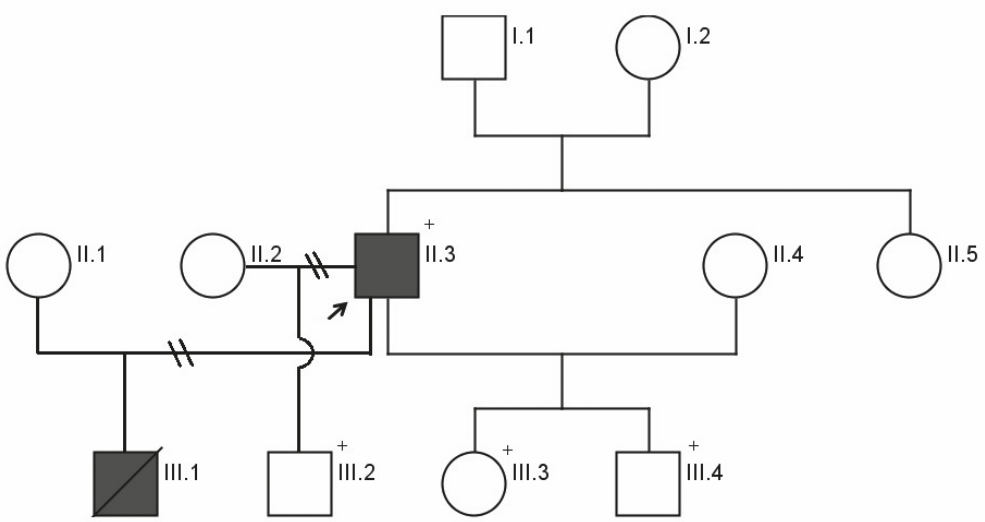


Figure 9. Pedigree. The arrow indicates the proband; figures painted black show ARVC+ patients; + shows individuals carrying the c.355delT mutation.

Until now (as of year 2018) the variant has never been reported in databases or research papers; it was also missing from the Exome Aggregation Consortium, Exome Variant Server, and the 1000 Genomes project. The *PKP2* gene encodes plakophilin-2 – a desmosomal protein found in the intercalated discs of cardiac cells. Desmosomes are major mediators of intercellular adhesion in the epidermis and myocardium. Impaired intercellular interactions are the major cause underlying ARVCs [Huber, 2003]. A single thymine nucleotide deletion at position 355 in the cDNA causes a frameshift, leading to a premature chain termination of p.Y119fs*23 (which eventually produces a truncated protein). According to the ACMG criteria [Richards *et al*., 2015], this mutation is classified as likely pathogenic. Molecular diagnostics allowed us to

assess the risk of ARVC and SCD in the proband's relatives and produce individual cardiac risk assessment protocols every 2 to 3 years until and every 12 months after the age of 10 years. Our results underscore the value of family screening to identify a pathogenic mutation in the proband and demonstrate how to implement genetic testing in clinical cardiology.

Another case of NGS success is identification of new variants in the *LDLR* gene in Russian patients with a history of familial hypercholesterolemia (FH) [Miroshnikova *et al.,* 2021]. FH is caused by mutations in various genes, including the *LDLR*, *APOB* and *PCSK9* genes [Maslennikov, 1999; Mandelstam *et al.,* 2002]. This study included mutation screening on the *LDLR* gene and other FH-associated genes in patients with definite FH, using NGS [Miroshnikova *et al.,* 2021]. In total, 59 unrelated patients were recruited and split into two separate groups depending on their age: adults (n=31; median age, 49; age range, 23-70) and children/adolescents (n=28; median age, 11; age range, 2-21). FH-associated variants were identified in 18 adults and 25 children, demonstrating mutation detection rates of 58 and 89% for the adults and children/adolescents respectively. In the adult group, 13 patients had FH-associated mutations in the *LDLR* gene, including two novel variants NM_000527.4: c.433_434dupG p.(val145glyfS*35) and c.1186G>C (GLY396ARG); 3 patients had *APOB* mutations and two had *ABCG5*/*G8* mutations. In the children/adolescent group, 21 patients had FH-causing mutations in the *LDLR* gene, including five novel variants NM_000527.4: c.325T>G p.(Cys109Gly), c.401G>C p.(Cys134Ser), c.616A >C p.(Ser206Arg), c.1684_1691delTGGCCCA A p.(Pro563His), and c.940+1_c.940+4delgtga. Our study reported seven novel *LDLR* variants considered to be pathogenic or likely pathogenic(Table 8). Among them, four missense variants were located in the coding regions, which corresponded to functional protein domains, and two frameshifts were identified that produced truncated proteins. These variants were observed only once in different patients, whereas a splicing variant in intron 6 (c.940+1_c.940+4delGTGA) was detected in four unrelated individuals. Variant p.Gly592Glu in the *LDLR* gene was detected in 6 patients, representing 10% of the FH

cases reported in the present study. Thus, it may be a major FH variant present in the Russian population.

Table 8. Pathogenicity of novel *LDLR* gene variants.

Gene	Patient ID	Exon/ intron	Variant	Allele frequency in GnomAD	Allele frequency in [Barbitoff *et al.*, 2019]	Variant pathogenicity classification by ACMG
LDLR	G31	4	c.316_328delCCCAAGACG TGCT p.(Lis107Argfs*95)	Not found	Not found	Pathogenic (PVS1 PS1 PM1 PM2 PP3)
LDLR	G29	4	c.325T>G p.(Cys109Gly)	Not found	Not found	Likely pathogenic (PS1 PM1 PM2 PM5 PP3)
LDLR	G36	4	c.401G>C (p.Cys134Ser)	Not found	Not found	Likely pathogenic (PS1 PM1 PM2 PM5 PP3)
LDLR	1	4	c.433_434insG p(Val145Glyfs*35)	Not found	Not found	Pathogenic (PVS1 PM2 PP3)
LDLR	G18	4	c.616A>C (p.Ser206Arg)	Not found	Not found	Uncertain significance (PM2 PP1 PP3)
LDLR	G21	IVS6	c.940+1_c.940+4 delGTGA (g.18154_18157delGTGA)	Not found	Not found	Pathogenic (PVS1 PM1 PM2 PP3)
LDLR	32	8	c.1186G>C p.(Gly396Arg)	Not found	Not found	Pathogenic (PVS1 PM1 PM2 PM5 PP3)
LDLR	G26	IVS8	c.1186+1G>T (g.22279G>T)	Not found	Not found	Pathogenic (PVS1 PM2 PP3)
LDLR	G17	11	c.1684_1691delTGGCCCAA p.(Pro563Hisfs*14)	Not found	Not found	Pathogenic (PVS1 PM1 PM2 PP3)

Despite being one of the most common genetic disorders, FH still remains largely undetected and untreated worldwide [Taranto *et al.,* 2020; Wiegman *et al.,* 2015]. Screening during childhood may enhance the potential identification of individuals with the condition before establishment of cardiovascular pathologies [Van der Graaf *et al.,* 2011]. As lipoprotein metabolism in children is influenced by fewer environmental factors than it is for adults, the difference in LDL-CHL levels between children with and without FH is more pronounced [Santos *et al.,* 2016]. Pediatric FH is diagnosed phenotypically by the presence of elevated LDL-CHL levels, in addition to a family history of premature CAD, high baseline TC levels in one parent and/or a FH-causing mutation [an der Graaf *et al.,* 2011]. It should be noted that there are no universal criteria for LDL-CHL cut-offs in the case of pediatric diagnosis of FH. Hence, early FH detection is critical for timely prevention and lower FH-associated disability rates in the future, whereas studies like ours are of paramount importance for

population genetics for having identified seven novel *LDLR* gene variants and an expanded spectrum of mutations in FH-associated genes in the Russian population.

A large, clinically heterogeneous group of dominantly inherited disorders linked to primary β-cell dysfunction is classified as maturity onset diabetes in the young (MODY). To date, 13 genes causative of 13 types of MODY are known [Barbetti *et al.,* 2018]. MODY is typically diagnosed before 25 years of age; it is non-insulin dependent and its symptoms are usually mild. However, due to the variety of clinical forms caused by a wide spectrum of mutations in MODY-related genes, different treatment strategies are used: From appropriate diet and physical activity to oral and/or insulin therapy. Monogenic diabetes also includes a number of non-MODY transient or permanent neonatal forms occurring under 6 months of age. More than 20 genes are known to be related to congenital neonatal diabetes [Lemelman *et al.,* 2018]. Depending on the gene involved, neonatal diabetes may follow patterns of dominant or recessive inheritance and may be isolated or associated with a variety of syndromic features [Greeley *et al.,* 2011]. However, due to a very early onset of diabetes, hyperglycemia is often diagnosed prior to other syndromic features. The treatment strategy for non-MODY neonatal diabetes depends on the specific genetic defect causing the diabetic phenotype. Our study suggests a panel that includes 13 MODY-causative genes (*HNF4A* (MODY1), *GCK* (MODY2), *HNF1A* (MODY3), *PDX1* (MODY4), *HNF1B* (MODY5), *NEUROD1* (MODY6), *KLF11* (MODY7), *CEL* (MODY8), *PAX4* (MODY9), *INS* (MODY10), *BLK* (MODY11), *ABCC8* (MODY12), and *KCNJ11* (MODY13)) and 22 genes causative of transient or permanent neonatal diabetes, including the ones related to specific syndromes (*EIF2AK3, RFX6, WFS1, ZFP57, FOXP3, AKT2, PPARG, APPL1, PTF1A, GATA4, GATA6, GLIS3, IER3IP1, LMNA, NEUROG3, PAX6, PLAGL1, SLC19A2, SLC2A2, SH2B1, SERPINB4* и *MADD*). Whole-exome sequencing (WES) was used to analyze the 35 genes causative of early age monogenic diabetes (MODY) and ransient or permanent neonatal diabetes. The WES results were verified using Sanger sequencing. Overall, genetic variants in target genes (see Table 9; 21-40) were identified in 33 out of 60 patients

(55%). For 12 patients, parents were available for genetic testing and origins of genetic variants were determined. In 11 cases, genetic variants had been inherited from the parents, and in one case, a de novo genetic variant was confirmed. A total of 38 genetic variants were identified in 33 out of 60 patients (55%). The majority of patients (27/33, 81.8%) had variants in MODY-related genes: *GCK* (n=19), *HNF1A* (n=2), *PAX4* (n=1), *ABCC8* (n=1), *KCNJ11* (n=1), *GCK+HNF1A* (n=1), *GCK+BLK* (n=1) и *GCK+BLK+WFS1* (n=1). A total of 6 patients (6/33, 18.2%) had variants in MODY-unrelated genes: *GATA6* (n=1), *WFS1* (n=3), *EIF2AK3* (n=1), and *SLC19A2* (n=1). A total of 15 out of 38 variants were novel, including *GCK*, *HNF1A*, *BLK*, *WFS1*, *EIF2AK3*, and *SLC19A2* [Glotov O. *et al.*, 2019]. The wide spectrum of genetic variants in *GCK* gene is demonstrated in Fig. 10 and Table 9.

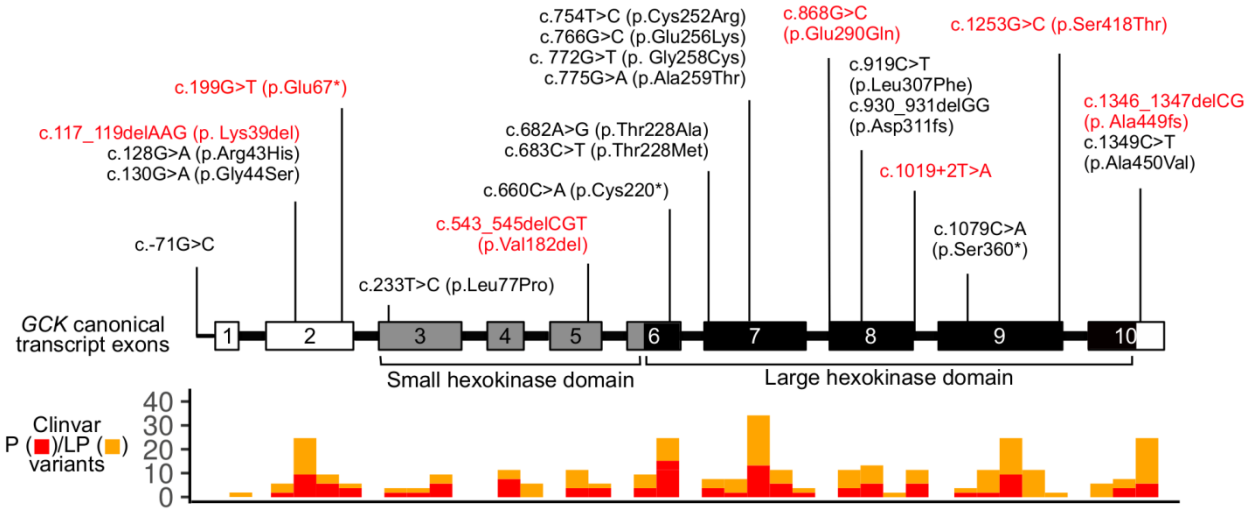


Figure 10. The spectrum of genetic variants in the *GCK* gene [Glotov O. *et al.*, 2019]. Novel pathogenic variants are highlighted in red.

Missens mutations in the *HNF1A* (MODY3) gene were discovered in 2 patients. Other MODY-associated genetic variants included 3 cases of missens mutations in the *PAX4* (MODY9), *ABCC8* (MODY12), and *KCNJ11* (MODY13) genes.

Ethnic differences play an important role in determining the epidemiology of monogenic diabetes, especially of MODY. Large population studies in European Caucasians showed a general trend of increased *HNF1A*-MODY frequency in Northern Europe, while *GCK*-MODY is prevalent in Southern European populations

[Kleinberger and Pollin, 2015]. Here, we report *GCK*-MODY in 19 and *HNF1A*-MODY in only 2 out of 27 MODY-positive Russian patients. These mutation rates appeared to be closer to those in Southern European populations than to those in Northern Europe residents. Our finding may indicate the population-specific frequency MODY types in Russian patients [Glotov O. *et al.*, 2019].

Three patients demonstrated genetic variants in different target genes. One patient exhibited deletion in the *GCK* gene accompanied by a missens mutation in the *HNF1A* gene (patient #226). Case 2 showed two missens mutations *GCK* and *BLK* genes (patient #529). Case 3 (patient #662) had a splicing defect in the *GCK* genes and missens mutations in the *BLK* and *WFS1* genes. the remaining variants are present at low frequencies in 9 different genes, altogether amounting to ~50% of the cases and highlighting the efficiency of using WES in non-GCK-MODY cases [Glotov O. *et al.*, 2019]. Ultimately, WES has significantly outperformed Sanger sequencing in its ability to identify pathogenic variants, the latter limited to analyzing a few MODY-associated genes and able to verify as few as some 15% of all MODY cases [Shields *et al.,* 2010]. The higher mutation detection rate in our study is achieved by increasing the number of genes tested and a thorough clinical selection of patients with possible monogenic diabetes. In this regard, one more advantage of WES should be mentioned: DNA sequencing data may be easily stored for further analysis of newly discovered candidate genes. Considering that monogenic diabetes may be associated with deletions and duplications, we analyzed the possible presence of CNVs in the target genes. We found no evidence for CNVs in the target genes in either sample. However, it should be noted that the limitations of WES technology do not allow for confident detection of small-scale CNVs. We analyzed the relationship of the detected genetic variants to the patients' diabetic phenotypes. Among the 38 detected genetic variants, 23 had been previously reported as linked to monogenic diabetes and 15 were novel ones (Table 9).

Table 9. Genetic variants identified in Russian children with non-type 1 diabetes mellitus [Glotov O. *et al.*, 2019].

Patient ID	Gene	Nucleotide change (protein change)	Mutation type	Mutation origin	Known /novel variant	Pathogenicity according to ACMG
59	*GCK*	c.772G>T (p.Gly258Cys)	Missense	Unknown	Known	Likely pathogenic
62	*GCK*	c.930_931delGG (p.Asp311fs)	Frameshift	Unknown	Known	Pathogenic
83	*GCK*	c.930_931delGG (p.Asp311fs)	Frameshift	Unknown	Known	Pathogenic
95	*GCK*	c.130G>A (p.Gly44Ser)	Missense	Father	Known	Likely pathogenic
167	*GCK*	c.128G>A (p.Arg43His)	Missense	Mother	Known	Likely pathogenic
197	*GCK*	c.233T>C (p.Leu77Pro)	Missense	Father	Known	Likely pathogenic
426	*GCK*	c.683C>T (p.Thr228Met)	Missense	Unknown	Known	Likely pathogenic
460	*GCK*	c.682A>G (p.Thr228Ala)	Missense	Mother	Known	Likely pathogenic
580	*GCK*	c.775G>A (p.Ala259Thr)	Missense	Unknown	Known	Likely pathogenic
663	*GCK*	c.1079C>A (p.Ser360*)	Nonsense	Unknown	Known	Pathogenic
665	*GCK*	c.660C>A (p.Cys220*)	Nonsense	Unknown	Known	Pathogenic
176	*GCK*	c.1349C>T (p.Ala450Val)	Missense	Unknown	Known	Likely pathogenic
661	*GCK*	c.1349C>T (p.Ala450Val)	Missense	Unknown	Known	Likely pathogenic
118	*GCK*	c.117_119delAAG (p.Lys39del)	In-frame deletion	Unknown	Novel	Uncertain significance
119	*GCK*	c.1346_1347delCG (p.Ala449fs)	Frameshift	Unknown	Novel	Pathogenic
434	*GCK*	c.868G>C (p.Glu290Gln)	Missense	Mother	Novel	Uncertain significance
578	*GCK*	c.1253G>C (p.Ser418Thr)	Missense	Unknown	Novel	Pathogenic
27	*GCK*	c.754T>C (p.Cys252Arg)	Missense	Unknown	Known	Likely pathogenic
		c.-71G>C	Promoter	Unknown	Known	Likely pathogenic
78	*GCK*	c.199G>T (p.Glu67*)	Nonsense	Mother	Novel	Pathogenic
		c.766G>C (p.Glu256Lys)	Missense	Mother	Known	Likely pathogenic
153	*HNF1A*	c.709A>G (p.Asn237Asp)	Missense	Unknown	Known	Uncertain significance
422	*HNF1A*	c.485T>G (p.Leu162Arg)	Missense	Unknown	Novel	Uncertain significance
215	*PAX4*	c.574C>A (p.Arg192Ser)	Missense	Unknown	Known	Uncertain significance
114	*ABCC8*	c.4139G>A (p.Arg1380His)	Missense	Unknown	Known	Likely pathogenic
134	*KCNJ11*	c.406C>A (p.Arg136Ser)	Missense	Unknown	Known	Uncertain significance
68	*GATA6*	c.1477C>T (p.Arg493*)	Nonsense	de novo	Known	Pathogenic

266	*WFS1*	c.2452C>T (p.Arg818Cys)	Missense	Mother	Known	Likely benign
408	*WFS1*	c.2327A>T (p.Glu776Val)	Missense	Mother	Known	Likely benign
133	*WFS1*	c.1124G>A (p.Arg375His)	Missense	Unknown	Novel	Uncertain significance
411	*EIF2AK3*	c.1912C>T (p.Arg638*)	Nonsense	Father	Novel	Pathogenic
	EIF2AK3	c.1912C>T (p.Arg638*)	Nonsense			
432	*SLC19A2*	c.164delC (p.Pro55fs)	Frameshift	Mother	Novel	Pathogenic
	SLC19A2	c.161C>A (p.Thr54Asn)	Missense	Father	Novel	Uncertain significance
226	*GCK*	c.543_545delCGT (p.Val182del)	In-frame deletion	Unknown	Novel	Uncertain significance
	HNF1A	c.92G>A (p.Gly31Asp)	Missense	Unknown	Known	Likely pathogenic
529	*BLK*	c.939G>C (p.Glu313Asp)	Missense	Unknown	Novel	Uncertain significance
	GCK	c.919C>T (p.Leu307Phe)	Missense	Unknown	Novel	Uncertain significance
662	*GCK*	c.1019+2T>A	Splicing defect	Unknown	Novel	Pathogenic
	BLK	c.1148G>A (p.Arg383Gln)	Missense	Unknown	Novel	Uncertain significance
	WFS1	c.1957C>T (p.Arg653Cys)	Missense	Unknown	Known	Likely pathogenic

According to the American College of Medical Genetics and Genomics (ACMG) guidelines [Richards *et al.,* 2015], most of the detected genetic variants (18 previously reported and 6 novel ones) were classified as pathogenic or likely pathogenic and thus were considered as causative of the diabetic phenotypes in the studied patients. However, the relationship of the detected KCNJ11 missense mutation to the diabetic phenotype was not apparent, because earlier it had been shown to be associated with hyperinsulinism [Mohnike *et al.,* 2014], which was not present in patient #134. Three previously reported and 9 novel genetic variants were classified as those of uncertain significance, and two genetic variants were likely benign (Table 9). These variants included 12 missense mutations; for them, we performed an additional in silico analysis using I-Mutant 2.0 [Capriotti *et al.,* 2005] (see Table 10).

In all but one case, the *in silico* modeling attested to a decrease of protein stability, thus suggesting the pathogenic effect of the checked genetic variants. Of special interest were two novel *WFS1* genetic variants, initially classified as likely benign. Patient #266 inherited the genetic variant from a non-diabetic mother, while patient #408 inherited the genetic variant from a mother with diabetes. Homozygous

mutations in *WFS1* lead to the development of Wolfram syndrome, an autosomal recessive disorder characterized by a list of clinical signs including a bilateral progressive optic atrophy, deafness, and diabetes mellitus [Fraser and Gunn, 1977].

Таблица 10. In silico prediction of increase/decrease in the protein stability caused by missense mutations with uncertain significance and by benign missense mutations. [Glotov O. *et al.*, 2019].

Patient ID	Gene	Nucleotide change (protein change)	Pathogenicity according to ACMG	Protein stability predicted by I-Mutant
434	*GCK*	c.868G>C (p.Glu290Gln)	Uncertain significance	Decrease
153	*HNF1A*	c.709A>G (p.Asn237Asp)	Uncertain significance	Decrease
422	*HNF1A*	c.485T>G (p.Leu162Arg)	Uncertain significance	Decrease
215	*PAX4*	c.574C>A (p.Arg192Ser)	Uncertain significance	Decrease
134	*KCNJ11*	c.406C>A (p.Arg136Ser)	Uncertain significance	Decrease
266	*WFS1*	c.2452C>T (p.Arg818Cys)	Likely benign	Decrease
408	*WFS1*	c.2327A>T (p.Glu776Val)	Likely benigh	Increase
133	*WFS1*	c.1124G>A (p.Arg375His)	Uncertain significance	Decrease
432	*SLC19A2*	c.161C>A (p.Thr54Asn)	Uncertain significance	Decrease
529	*BLK*	c.939G>C (p.Glu313Asp)	Uncertain significance	Decrease
	GCK	c.919C>T (p.Leu307Phe)	Uncertain significance	Decrease
662	*BLK*	c.1148G>A (p.Arg383Gln)	Uncertain significance	Decrease

Heterozygous carriers of *WFS1* mutations have been reported to have risk of early-onset diabetes mellitus [Bennett *et al.,* 2011]. The latter cannot be excluded in our patients. However, an intriguing point is that the *WFS1* genetic variant in patient #408, who inherited it from a diabetic mother, appeared to not decrease the protein stability according to I-Mutant, which makes its pathogenicity questionable. Finally, we analyzed the clinical picture in patients with more than one genetic variant in one or different target gene (Table 11).

A simultaneous presence of two *GCK* genetic variants in patient #27 raised the question of their location in one or both alleles. The parents were not available for analysis.

Table 11. Clinical characteristics of the patients with multiple genetic variants in monogenic diabetes-related genes [Glotov O. *et al.*, 2019].

Patient ID	Gene Nucleotide change (amino acid change)	Age at diagnosis, months	Diabetic ketoacidosis	C-peptide ng/ml	HbA1C, %	SDS BMI	Treatment
27	*GCK* c.754T>C (p.Cys252Arg)	3	No	0.7	6	-0,63	Diet
	GCK c.-71G>C						
78	*GCK* c.199G>T (p.Glu67*)	39	No	0.63	6.4	+0,83	Diet
	GCK c.766G>C (p.Glu256Lys)						
226	*GCK* c.543_545delCGT (p.Val182del)	36	No	1.1	6	-1,69	Diet
	HNF1A c.92G>A (p.Gly31Asp)						
411	*EIF2AK3* c.1912C>T (p.Arg638*)	3	Ketonuria	0.2	9.2	-0,72	Insulin
	EIF2AK3 c.1912C>T (p.Arg638*)						
432	*SLC19A2* c.164delC (p.Pro55fs)	48	Ketonuria	1.1	5.3	-1.0	Insulin/ Diet
	SLC19A2 c.161C>A (p.Thr54Asn)						
529	*BLK* c.939G>C (p.Glu313Asp)	10	No	0.43	6.7	-0,46	Diet
	GCK c.919C>T (p.Leu307Phe)						
662	*GCK* c.1019+2T>A	22	No	1.1	6.82	-1,32	Diet
	BLK c.1148G>A (p.Arg383Gln)						
	WFS1 c.1957C>T (p.Arg653Cys)						

Clinical manifestations were mild and common of MODY2. This was in stark contrast with the severity of disease, usually reported in patients with both affected *GCK* alleles, suggesting that, in patient #27, both genetic variants were present in the same allele and thus had no accumulative effect. In patient #78, who was also a carrier of two *GCK* genetic variants, the clinical picture was typical for MODY2. As both genetic variants were inherited from the mother, we concluded that only one allele was affected. Moreover, only nonsense mutation c.199G>T seemed to be clinically significant, because the resulting stop-codon terminates translation before the c.766G>C site. The clinical picture in patient #226, who had genetic variants in *GCK*

and *HNF1A*, was more typical for MODY2 than for MODY3. He had mild fasting and postprandial hyperglycemia, had no glucosuria, and was successfully being treated by a diet. Patient #411 had a homozygous *EIF2AK3* nonsense mutation, inherited from consanguineous parents and associated with Wolcott-Rallison syndrome, which, in turn, has been reported to be the most common genetic cause of permanent neonatal diabetes in consanguineous families [Rubio-Cabezas *et al.,* 2009]. Patient #432 had two novel genetic variants affecting both *SLC19A2* alleles. Homozygous mutations in *SLC19A2* cause Rogers syndrome – thiamine-responsive megaloblastic anaemia associated with diabetes mellitus and deafness [Labay *et al.,* 1999]. Among other clinical signs are congenital heart defects, retinal degeneration, ketonuria, dwarfism, and neurologic symptoms [Shaw Smith *et al.*, 2012]. Of note, patient #432 had only diabetes mellitus, retinal degeneration, ketonuria, and neurological symptomatology and thus did not manifest a typical clinical picture. Both patients #529 and #662 had typical clinical signs of *GCK*-MODY rather than *BLK*-MODY, suggesting an absence of strong accumulation of the pathogenic effect of the detected genetic variants. Among genetic variants detected in our study, 60.5% had already been reported in diabetic patients and 39.5% were novel ones. On the one hand, these results point towards a significant recurrent variation within monogenic-diabetes-related genes. On the other hand, they suggest that, in spite of the multitude of monogenic diabetes studies, many variants still remain unidentified. Identification of novel genetic variants as well as accumulating data on previously known causes of monogenic diabetes is of high importance, both for fundamental understanding of the disease pathogenesis and for clinical practice. To summarize, our data show a high rate of genetic variants causative of monogenic diabetes in Russian children with non-type 1 diabetes mellitus. The use of a WES-based panel allowed us to identify a variety of previously known and novel genetic variants in MODY-related and unrelated genes, including multiple variants in a number of patients. The revealed variety is characterized by the prevalence of *GCK* genetic variants (MODY2), suggesting that genetic analysis for monogenic diabetes in Russian children may start with testing for *GCK* variants,

which may not necessarily be performed by WES. Hence, genetic analysis in Russian children with suspected monogenic diabetes may start with MODY2 testing.

Yet another case of demonstrates how advanced sequencing techniques are successfully deployed to investigate mutations ifn Wilson disease (WD) patients and detect novel pathogenic variants [Balashova *et al.*, 2020]. Wilson disease, or hepatolenticular degeneration, is an autosomal-recessive genetic disorder associated with *ATP7B* gene mutations. *ATP7B* gene codes for a P-type copper-transporting ATPase providing copper incorporation into ceruloplasmin and excreting it into bile through apical membrane of hepatocytes. ATP7B dysfunction results in hepatocyte damage due to copper overload. Both timely diagnosis and copper chelation significantly improve prognosis of Wilson disease. Oppositely, untreated Wilson disease results in early disability and death due to liver cirrhosis and its complications. the objective of the present study was estimation of the frequency of *ATP7B* gene mutations in the Russian population of Wilson disease patients. A targeted NGS panel was developed to detect pathogenic variants in all *ATP7B* gene exons, except exons 1, 3, 5, 9, 10, 12, and 21 (Fig. 11). The observed variants were considered as: frameshift (5 alleles, 16.7%), nonsense (5 alleles, 16.7%), missense (15 alleles, 50%), splice site mutations (4 alleles, 13.3%), and indel (1 allele, 3.3%) mutations.

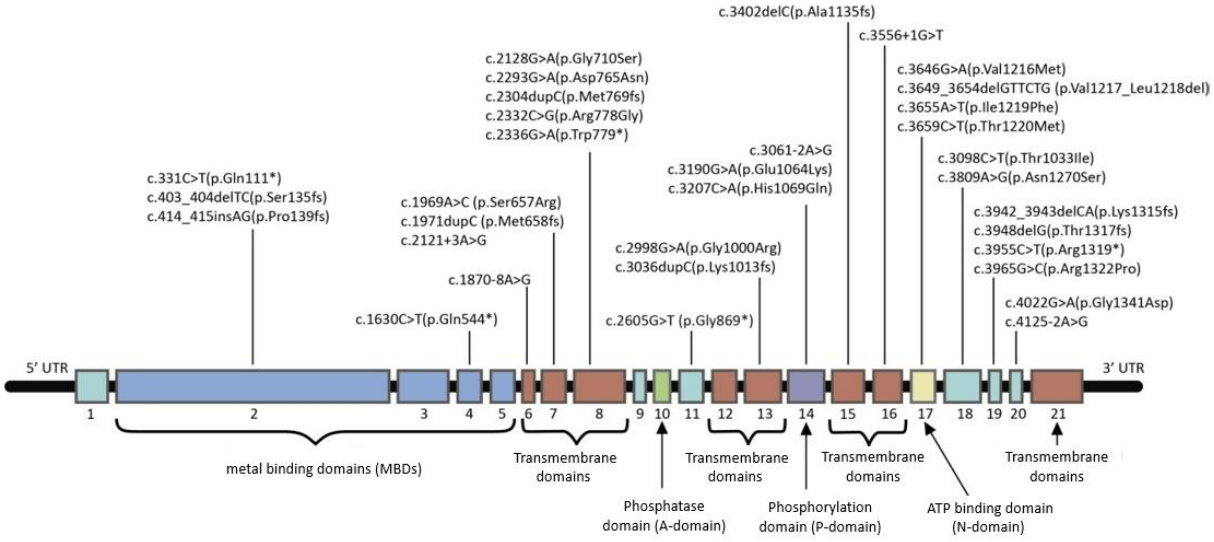


Figure 11. Distribution of identified mutations by gene *ATP7B* [Balashova *et al.*, 2020].

Thus, c.3207C > A (p.His1069Gln) mutation was observed in 73.3% of the examinees. However, its homozygous form was detected only in 28% cases, whereas in the remaining patients it was observed in compound-heterozygous form. A total of 30 mutations were revealed using NGS. The list of detected mutations (including the unknown before), as well as their characteristics and the number of alleles see in Table 12.

Table 12. Genetic variants revealed in Russian patients with Wilson disease [Balashova *et al.*, 2020].

Domain	Exon	Variant	Effect	No. of alleles detected	%	Earlier described as pathogenic in WD
Metal-binding	2	c.331C>T (p.Gln111*)	Stop	2	1.34%	Yes
	4	c.1630C>T (p.Gln544*)	Stop	1	0.67%	Yes
Transmembrane	6	c.1870-8A>G	Splicing	1	0.67%	No
	7	c.1969A>C (p.Ser657Arg)	Missense	1	0.67%	Yes
		c.1971dupC (p.Met658fs)	Frameshift	1	0.67%	No
		c.2121+3A>G	Splicing	1	0.67%	Yes
	8	c.2128G>A (p.Gly710Ser)	Missense	1	0.67%	Yes
		c.2293G>A (p.Asp765Asn)	Missense	1	0.67%	Yes
		c.2304dupC (p.Met769fs)	Splicing	7	4.70%	Yes
		c.2332C>G (p.Arg778Gly)	Missense	3	2.01%	Yes
		c.2336G>A (p.Trp779*)	Stop	1	0.67%	Yes
Phosphatase	11	c.2605G>T (p.Gly869*)	Stop	1	0.67%	Yes
Transmembrane	13	c.2998G>A (p.Gly1000Arg)	Missense	1	0.67%	Yes
		c.3036dupC (p.Lys1013fs)	Frameshift	2	1.34%	No
Phosphorylation	14	c.3098C>T(p.Thr1033Ile)	Missense	1	0.67%	Yes
		c.3190G>A (p.Glu1064Lys)	Missense	12	8.05%	Yes
		c.3207C>A (p.His1069Gln)	Missense	84	56.38%	Yes
Transmembrane	15	c.3402delC (p.Ala1135fs)	Frameshift	10	6.71%	Yes
	16	c.3556+1G>T	Splicing	1	0.67%	Yes
ATP-binding	17	c.3646G>A (p.Val1216Met)	Missense	1	0.67%	Yes
		c.3649_3654delGTTCTG (p.Val1217_Leu1218del)	Indel	6	4.03%	Yes
		c.3655A>T (p.Ile1219Phe)	Missense	1	0.67%	No
		c.3659C>T (p.Thr1220Met)	Missense	1	0.67%	Yes
Phosphorylation	18	c.3809A>G (p.Asn1270Ser)	Missense	1	0.67%	Yes
		c.3948delG (p.Thr1317fs)	Frameshift	1	0.67%	Yes

Transmembrane	19	c.3955C>T (p.Arg1319*)	Stop	1	0.67%	Yes
		c.3965G>C (p.Arg1322Pro)	Missense	1	0.67%	Yes
		c.4022G>A (p.Gly1341Asp)	Missense	1	0.67%	Yes
C-terminal (stabilization)	20	c.4125-2A>G	Splicing	3	2.01%	Yes

We have compared the spectrum of the revealed mutations with those included into the most frequently used PCR panel in Russia. It has been demonstrated that only 4 of 12 mutations were detected in the studied cohort (c.2304insC, c.3207C>A, c.3402delC, c.3649_3654del6). Moreover, 3 mutations were frequently revealed in the studied cohort, although being not included into the frequent mutations panel (c.2332C>G (p.Arg778Gly), c.4125-2A>G, c.3190G>A (p.Glu1064Lys). The estimated informativity for the most frequent mutation (c.3207C > A) is 75% patients and 82% alleles, respectively. Application of NGS allowed to increase the informativity to 96% [Balashova *et al.*, 2020]. Detection of a mutant *ATP7B* gene is overarching to verify WD diagnosis, considering that clinical appearance of WD may be rather variable at manifestation and genetic profiling at this step is the only way to confirm the presence of WD before it progresses to the advanced stage.

Current sequencing methods allow detection of a bundle of hereditary diseases in an individual, thus gaining unprecedented significance. Such cases are not as rare as they may seem. For instance, we would like to refere to a case of co-inheritance of X-linked and autosomal dominant forms of ichthyosis [Alaverdian *et al.*, 2019]. Abstract According to modern classification, there are two forms of inherited ichthyoses: syndromic and non-syndromic, each of them consists of more than ten different nosologies. The commonest types of the ichthyosis are X-linked recessive (prevalence 1/2000–6000 in men) and autosomal dominant, or ichthyosis vulgaris with incomplete penetrance (1/250–1000). The X-linked form is associated with mutations in steroid sulfatase *STS* gene, it is noteworthy that there is a full deletion of the gene in 90% of cases. Ichthyosis vulgaris is caused by heterozygous mutations in the *FLG* gene encoding filaggrin. It is important to note that the clinical forms of these diseases are indistinguishable. The aim of this study was to search for pathogenic or likely

pathogenic mutations which are associated with various forms of the inherited ichthyosis such as other inherited diseases with similar phenotypic signs. The identified mutation p.Arg2037Ter in the heterozygous condition in the *FLG* gene has been described before in databases as being pathogenic. Also, our patient has a full deletion of the *STS* gene and it was found that our patient carries two pathogenic mutations which are related to different forms of the inherited ichthyosis ((Fig. 12).

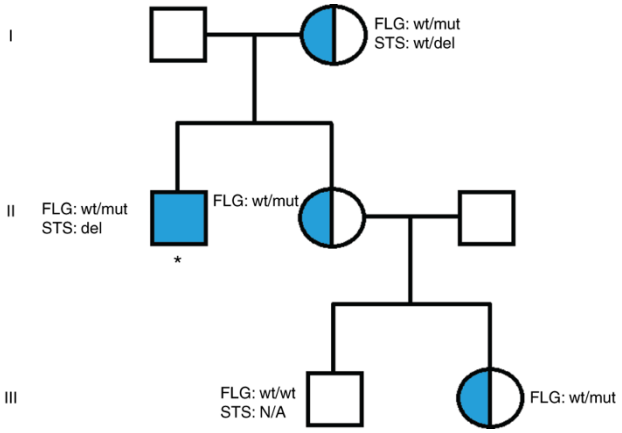


Figure 12. Genetic pedigree map of the proband (*). Mother, sister, sister's daughter show symptoms of vulgar ichthyosis caused by the *FLG* gene mutation (p.Arg2037Ter: c.6109C>T) [Alaverdian *et al.*, 2019].

This information may be valuable for genetic counselling because of similar clinical signs. It is therefore recommended to analyze both *STS* and *FLG* genes to exclude combined forms of ichthyosis.

Clinical polymorphism in patients with Wilson disease is another example of combined inherited pathology [Tulzunovskaya *et al.*, 2017; Balashova *et al.*, 2020]. In one family, 3 different *ATP7B* mutations were revealed in 3 siblings suffering from WD. All three brothers were characterized by c.3649_3654delGTTCTG (p.Val1217_Leu1218del) mutation on one of the copies. c.3036dupC (p.Lys1013fs) mutation was revealed on the second gene copy in the older brother. At the same time, the younger brothers being dizygous twins (as assessed by phenotype, blood group, and NGS data) carried c.3207C > A (p.His1069Gln) mutation. Clinical appearance of WD in each patient was significantly variable at manifestation. The elder brother alone showed a WD pathognomonic symptom – Kayser-Fleischer ring and neurological

manifestations, i.e. unsteady gait, speech and handwriting disorders. At screening one of the two younger brothers, both having no complaints at the time of examination, revealed initial manifestations of low-stage chronic hepatitis (without cirrhosis); the other one had chronic hepatitis progressing to cirrhosis. This allowed clinicians to formulate the WD diagnosis. Manifestations identified in all the 3 patients included liver blood tests and decreased ceruloplasmin. Thus, the elder brother was diagnosed with a mixed form of WD with liver cirrhosis, portal hypertension syndrome (splenomegaly), and CNS tremor and rigidity disorder. The brothers were diagnosed with WD following screening of the patient's relatives. Their diagnosis featured abdominal WD of milder course. In addition to pathogenic variants in the main gene, variants in the modifier *HFE* gene were detected, associated with hereditary hemochromatosis (Fig. 13).

Overall, WD clinical diversity may stem from the variety of molecular defects, causative of the disease. Various mutations and polymorphisms in other genes, however, as well as diverse exogenous and endogenous factors, may significantly modify the course and outcome of this hereditary disease [Tulzunovskaya *et al.*, 2017].

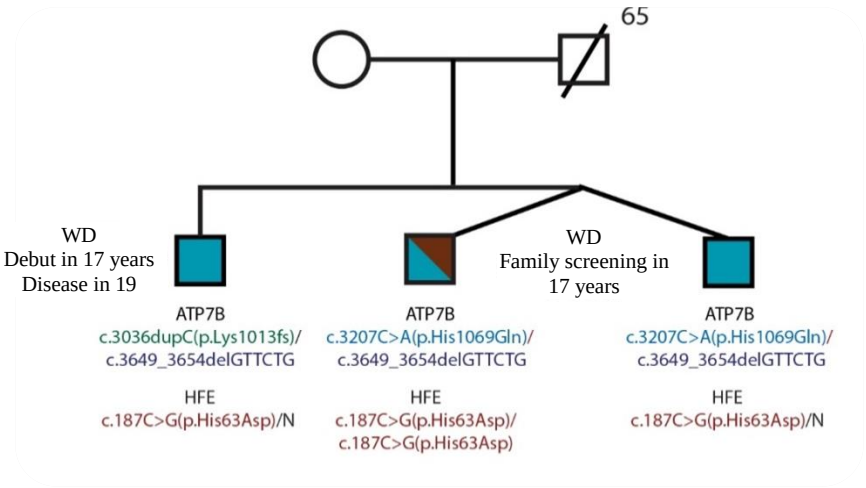


Figure 13. Genetic pedigree map for a WD patient [Tulzunovskaya *et al.*, 2017].

Notably, NGS allows us to identify pathogenic or likely pathogenic variants in genes that were earlier believed to possess mutations of a single type [Koshevaya *et*

al., 2022]. For example, Lopes–Maciel–Rodan syndrome (LOMARS; OMIM 617435) is an extremely rare autosomal recessive disorder, caused by a compound heterozygous LoF mutation in the huntingtin (*HTT*) gene. Toxic gain-of-function (GoF) expansion of the unstable CAG repeat on the *HTT* (OMIM 613004) gene causes Huntington's disease – a neurodegenerative disorder with the autosomal dominant mode of inheritance [MacDonald *et al.*, 1993]. However, little is still known about huntingtin function in humans. By now, investigators have described a few family cases, carrying two putative compound heterozygous LoF mutations in the HTT gene, responsible for the development of Lopes–Maciel–Rodan syndrome – a rare congenital disorder with Rett-like neurological symptoms [Lopes *et al.*, 2016; Rodan *et al.*, 2021]. So far, LOMARS syndrome is known to be associated with a variety of symptoms: limb spasticity, decreased muscle tone (hypotonia), stereotyped arm movements, dystonia, ataxia, epilepsy, myopia, bruxism, etc. Whole exome sequencing revealed two earlier unknown c.8440C>A (chr4:4:3233337 (GRCh38), chr4:3235064 (GRCh37)) and c.2350C>T (chr4:3132675 (GRCh38), chr4:3134402 (GRCh37)) mutations in the *HTT* gene. Parental re-examination showed that the c.2350C>T variant was passed on by the father whereas c.8440C>A was inherited from the mother. Given the presented clinical and molecular genetic data, we recommend considering the probability of LOMARS in children with Rett-like neurological symptoms and perform molecular genetic testing to search for putative LoF mutations in the *HTT* gene sequence [Koshevaya *et al.*, 2022].

Elaborate diagnostic assessment and adequate interpretation of findings are pivotal for the rapid accurate diagnosis of inherited neuropsychiatric diseases. The present study is an attempt to raise awareness among physicians regarding this rare condition and facilitate its diagnosis and molecular genetic confirmation in the future [Koshevaya *et al.*, 2022].

1.6. General Strategy and Algorithm of NGS Implementation in Human Genetic Pathology Diagnostics

For a set of disorders, adequate therapy is the most critical outcome of NGS examination. Floating Harbor syndrome (FHS) is an extremely rare disease, with slightly more than a hundred cases reported worldwide. FHS is caused by heterozygous mutations in the *SRCAP* gene; however, little is known about the pathogenesis of FHS or the effectiveness of its treatment. In our study we report the first case of FHS in the Russian Federation [Turkunova *et al.*, 2022]. The male proband presented with most of the typical phenotypic features of FHS, including short stature, skeletal and facial features, delayed growth and bone age, high pitched voice, and intellectual impairment. The proband also had partial growth hormone deficiency. WES revealed a pathogenic c.7466C>G (p.Ser2489*) mutation in the last exon of the FHS-linked *SRCAP* gene. A systematic literature review and analysis of available genetic variation datasets highlighted an unusual distribution of pathogenic variants in the *SRCAP* gene and confirmed the lack of pathogenicity for variants outside of exons 33 and 34. Finally, we suggested a new model of FHS pathogenesis which provides possible basis for the dominant negative nature of FHS-causing mutations and explains limited effects of growth hormone treatment in FHS. Our findings expand the number of reported FHS cases and provide new insights into disease genetics and the efficiency of growth hormone therapy for FHS patients [Turkunova *et al.*, 2022].

In terms of efficiency (see Table 13), we suggest the following cost-effective strategy of genetic diagnosis for MODI, WD, and other monogenic diseases associated with major mutations (Fig. 14).

Table 13. Most efficient diagnostic strategies for hereditary diseases.

Nosology	Efficiency of diagnostics prior to NGS, %	Efficiency of diagnostics after NGS, %	Efficiency of diagnostics with novel variants considered, %	Reference
Cystic fibrosis	45-55 (1 mutation) 58 (35 mutations)	67-80	-	Unpublished
WD	Up to 75 (4 mutations) Up to 86 (12 mutations)	Up to 96	97	Balashova et al., 2020
MODY	15-35	40-50	55	Glotov O. et al., 2019

It should be noted that NGS is not always sufficient to formulate the diagnosis; hence, in some cases, concurrent or subsequent Sanger sequencing is required to detect the other pathogenic varian.

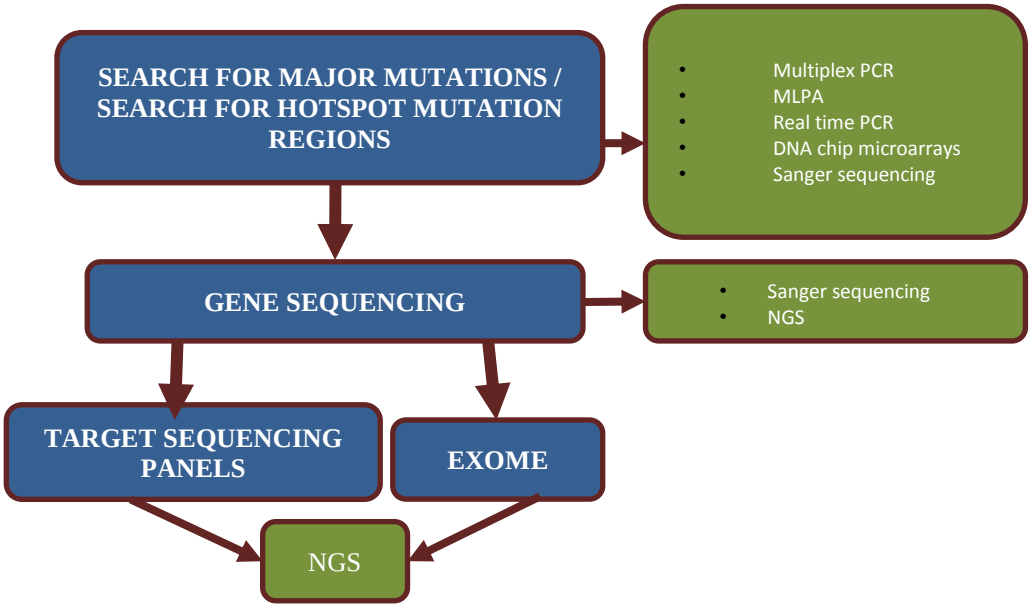


Figure 14. Roadmap to select most adequate diagnostic procedure.

For example, NGS performed to a patient with anauxetic dysplasia (AD), an extremely rare form of autosomal recessive skeletal chondrodysplasia, detected the heterozygous rs387906533 variant (n.91 _92delinsGC) of the nucleotide sequence (chr9:35657924-35657925delCTinsGC) in the first exon of the RNA processing endoribonuclease (*RMRP*) gene [Fedyakov *et al*., 2021]. Due to the absence of the second mutation, NGS was used to analyze the hotspot region in the *RMRP* gene promoter, followed by direct automatic Sanger sequencing of the *RMRP* gene in the proband (Fig. 15).

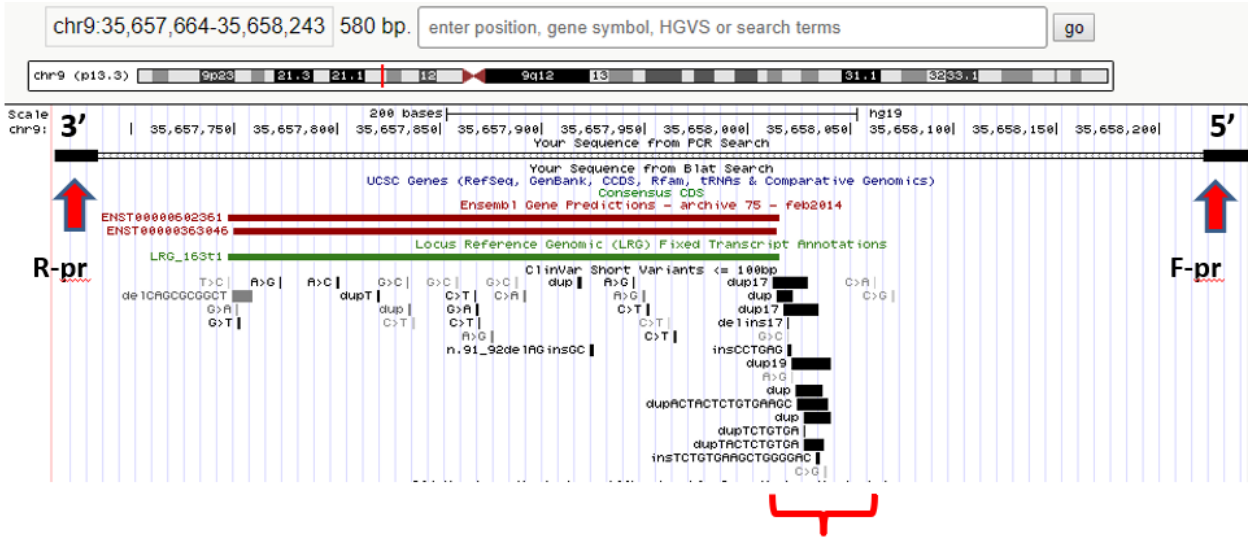


Figure 15. Hotspot RMRP gene promoter region in the proband.

Our study confirmed the heterozygous n.91_92delinsGC mutation. In addition, we detected an unknown n.-6_-5insTCTCAGCTTCAC substitution (chr9:g.35658020-35658021insTCTCAGCTTCAC) in the gene promoter region (Fig. 16).

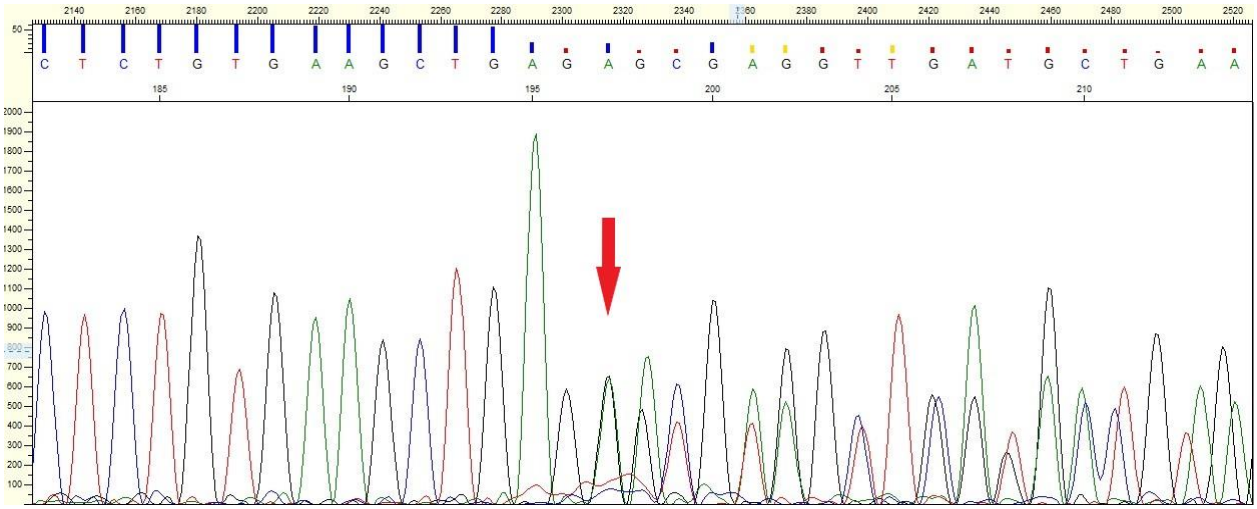


Figure 16: Electropherogram. Pathogenic heterozygous n.-6_-5insTCTCTCAGCTTCAC variant in the *RMRP* gene in proband [Fedyakov *et al.*, 2021].

The variant is a 12-nucleotide insertion between the TATA box and the transcription start site. Sanger direct automatic sequencing was used to analyze the RMRP gene in the proband's parents. It was found that the n.-6_-

5insTCTCTCAGCTTCAC mutation was of paternal origin and the n.91_92delinsGC mutation was of maternal origin. The insertion in the *RMRP* gene promoter region has been unknown before for AD, with no extraskeletal manifestations (typical of similar mutations carriers) observed in the patient so far [Fedyakov *et al*., 2021].

1.7. NGS in Family Planning As a Tool to Prevent Severe Hereditary Disorders

Today modern sequencing methods are indispensable for all areas of medicine, especially reproductive medicine. Our study reports the phenotype and reproductive history of an adult patient with an unbalanced karyotype – 8p23 and 18p11.3 terminal deletions and 8p22 a duplication [Pendina et al., 2019]. The indication for karyotyping for the 28-year-old patient was structural rearrangement in her miscarriage specimen: 45,XX,der(8;18)t(8;18)(p23;p11.3). Unexpectedly, the patient had the same karyotype with only one normal chromosome 8, one normal chromosome 18, and a derivative chromosome, which was a product of chromosomes 8 and 18 fusion with loss of their short arm terminal regions. The patient had minor facial and cranial dysmorphia and no pronounced physical or mental abnormalities. She was socially normal, had higher education and had been married since the age of 26 years. After four unsuccessful IVF/PGT-SR cycles, the patient conceived naturally. Non-invasive prenatal testing showed additional chromosome 18. The prenatal cytogenetic analysis of chorionic villi revealed an abnormal karyotype: 46,XX,der(8;18)t(8;18)(p23;p11.3)mat,+18. The pregnancy was terminated for medical reasons. It should be emphasized that inheritance of the patient's derivative chromosome by her offspring together with normal chromosomes 8 and 18 from the patient's spouse is not desirable. Even though in this case the fetus would have the same karyotype as the patient, the phenotypic effect of the aberration may be unpredictable. This case demonstrates the overriding importance of a comprehensive approach using the entire array of molecular, genetic, cytogenetic, and embryologic methods in pregnancy planning.

Another case shows NGS application for PGT, especially to detect chromosomal mosaicism [Saifitdinova *et al.*, 2020]. Postzygotic mitotic errors can

produce cell clones with unequal sets of chromosomes within the same embryo. This phenomenon has been described as embryonic mosaicism. Mosaicism should be carefully considered when examining preimplantation embryos at the blastocyst stage. Numerous studies report a relatively high frequency of mosaicism in the blastocyst [Weissman *et al*., 2017]. Meanwhile, mosaic of trophectoderm (TE) cells are not necessarily concordant with mosaic cells of the inner cell mass (ICM) [Munne *et al*., 2017]. This fact can cause significant challenges to interpret preimplantation genetic testing (PGT) results. The observation that in human preimplantation embryos, preimplantation genetic testing for aneuploidy (PGT-A) allows to detect mosaicism in almost every IVF cycle inevitably questions reliability of mosaicism assessment. Rapid development of NGS and whole-cell genome amplification techniques enabled to implement these tools for PGT-A as well. High NGS sensitivity allows to identify mosaicism in single abnormal cells of TE DNA samples, showing a high degree of confidence (20% for samples of 5 cells). NGS demonstrates that the frequency of mosaicism in preimplantation blastocysts varies between 17% and 47% in 9 different assisted reproductive technology (ART) tertiary centers [Sachdev *et al*., 2016]. However, available comparative data on chromosomal mosaicism values in TE and ICM cells remain limited and inconsistent [Munne *et al*., 2017]. Six human blastocysts with TE cell mosaicism were split into three groups to further investigation, two containing TE cells and the other ICM cells predominantly. Our data show that the proportion of aneuploid cells in the biopsy taken for PGT-A analysis does not necessarily reflect the true chromosomal status of the entire embryo and cannot be extrapolated to that of ICM cells. Notably, qualitative and quantitative characteristics of mosaic status may vary across different parts of the same embryo. In samples containing TE adjacent to the ICM, mosaicism tends to increase, which may have physiological significance for implantation. The results of our study strongly suggest that mosaicism revealed in blastocyst reduces the likelihood of finding the euploid chromosome set in the other parts of the embryo [Saifitdinova *et al*., 2020]. Chromosomal abnormalities in mosaic embryos are unpredictably diverse, which may

lead not only to loss of conception, but also to the development of genetic disease in the offspring. This greatly complicates the interpretation of preimplantation genetic testing results and requires extra research to improve clinical recommendations for embryo transfer.

Successful application of various molecular genetic technologies is demonstrated in our next study on familial hereditary pathology [Lyazina *et al*., 2017]. This paper presents a complicated and time-consuming pathway towards hereditary pathology diagnosis. The proband had compound heterozygous variants: c.851T>G and c.242A>T in the *PHGDH* gene. Mutations in this gene are associated with Noah-Lax syndrome (OMIM:#256520) and phosphoglycerate dehydrogenase deficiency (OMIN:#601815). The child did not have the common form of Neu-Laxova syndrome [Lyazina *et al*., 2017]. Based on diagnostics results, the family received medical genetic counseling in December 2016 and prenatal diagnostics options. A dedicated test system to rapidly diagnose the identified mutations was developed. The family reapplied to hospital in February 2017 due to a naturally occurring pregnancy. The first-trimester ultrasound scan diagnosed dichorionic diamniotic twins. One fetus had the same genotype as the diseased proband. The forecast was extremely unfavorable and treatment was unknown. The other fetus was healthy [Lyazina *et al*., 2017]. Reduction of the sick fetus at 16/17 weeks of gestation was performed. The healthy female fetus completed gestation successfully and was delivered without complications.

This case demonstrates the demand for a new comprehensive preconception screening algorithm based on molecular genetic methods, including NGS in the first line, as well as PGT and NIPT methods for subsequent pregnancy monitoring. Thus, advanced technologies are changing the paradigm of family planning and prevention of severe hereditary human diseases. Preconception genetic screening to ensure a healthy first pregnancy is required (Fig. 17).

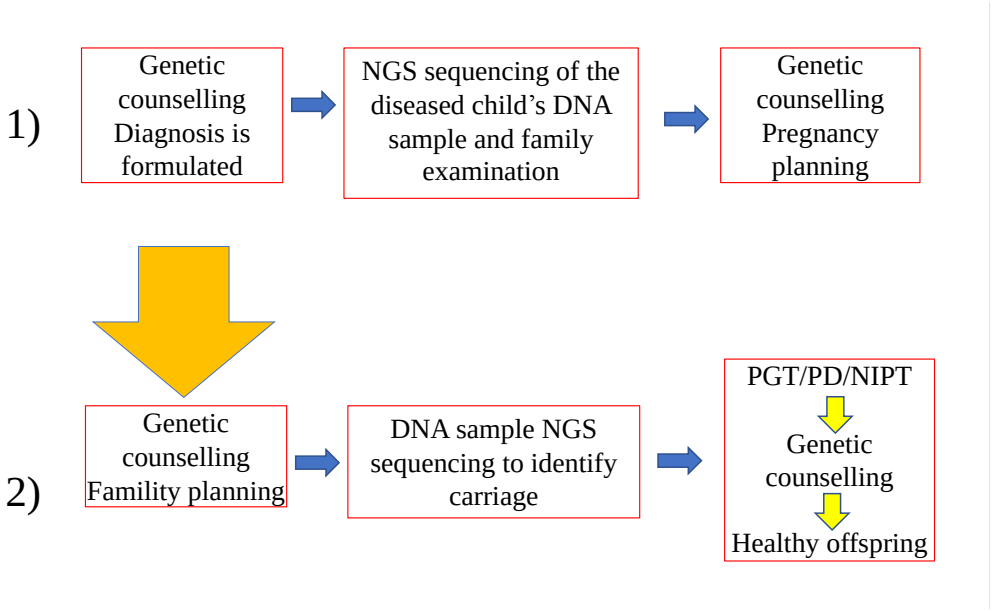


Figure 17: Transformation of family planning and severe disease prevention.

Preconception testing:

6. optimizes pregnancy follow-up, offering:
- *choice of diagnostic procedures;*
- *recommendations regarding medical termination of pregnancy;*
- *counselling;*
- *multidisciplinary approach;*

7. can be included into pregnancy planning scheme:
- *oocyte donation;*
- *PGT;*

8. reduces the risk of perinatal fetal loss;
9. plays an important role in psychology support of future parents (allows to relieve self-blame in those who suffered spontaneous abortion, including due to monogenic diseases);
10. safe and simple for patients.

Monogenic diseases associated with reproductive loss are split in the following groups:

- *maternal monogenic diseases;*

- *fetal monogenic diseases;*
- *other genetic factors (predisposition genes, RNA).*

Table 14 refers to set of maternal monogenic diseases with a high risk of reproductive losses.

Table 14. A set of maternal monogenic diseases with a high risk of reproductive losses.

Disease (OMIM)	Gene (OMIM)	Inheritance type	Description
Myotonic dystrophy, type I (160900)	*DMPK* (605377)	AD	Increased risk of miscarriage and obstetric complications at any term, including miscarriage, preterm labor, edema, fetal death.
Congenital Adrenal Hyperplasia (201910)	*CYP21A2* (613815)	AR	Failure to conceive in 25% of women; incidence decreases significantly after treatment of the disease.
Factor XIII deficiency in the A subunit (613225)	*F13A1* (134570)	AD	FXIIID deficiency leads to bleeding, spontaneous abortions and other complications during pregnancy. High probability of early pregnancy loss.
Polycystic kidney disease, type 1 (173900)	*PKD1* (601313)	AD	The disease develops hypertension and pre-eclampsia, and the likelihood of pregnancy termination increases.
Long QT syndrome, types 1-3 (192500, 613688, 603830)	*KCNQ1, KCNH2, SCN5A* (607542, 152427, 600163)	AD	Increased risk of fetal death, and growth retardation is observed in surviving newborns.
Glycogen storage disease, type Ib (232220)	*SLC37A4* (602671)	AR	Increased risk of spontaneous abortions and fetal death.

Fetal lethal phenotypes are another important problem, associated with monogenic diseases (Table. 15): autosomal recessive diseases (α-thalassemia; multiple pterygium syndrome, lethal type; galactosialidosis; cystic fibrosis, type VII),

autosomal dominant diseases (thanatophoric dysplasia; osteogenesis imperfecta, type II; achondroplasia; tuberous sclerosis, type I); X-linked diseases (incontinentia pigmenti or Bloch-Sulzberger syndrome; Holtz syndrome (focal cutaneous hypoplasia); Rett syndrome; immune dysregulation syndrome, polyendocrinopathies and enteropathies).

Table 15. Table 15. Fetal lethal phenotypes associated with novel genomic variants.

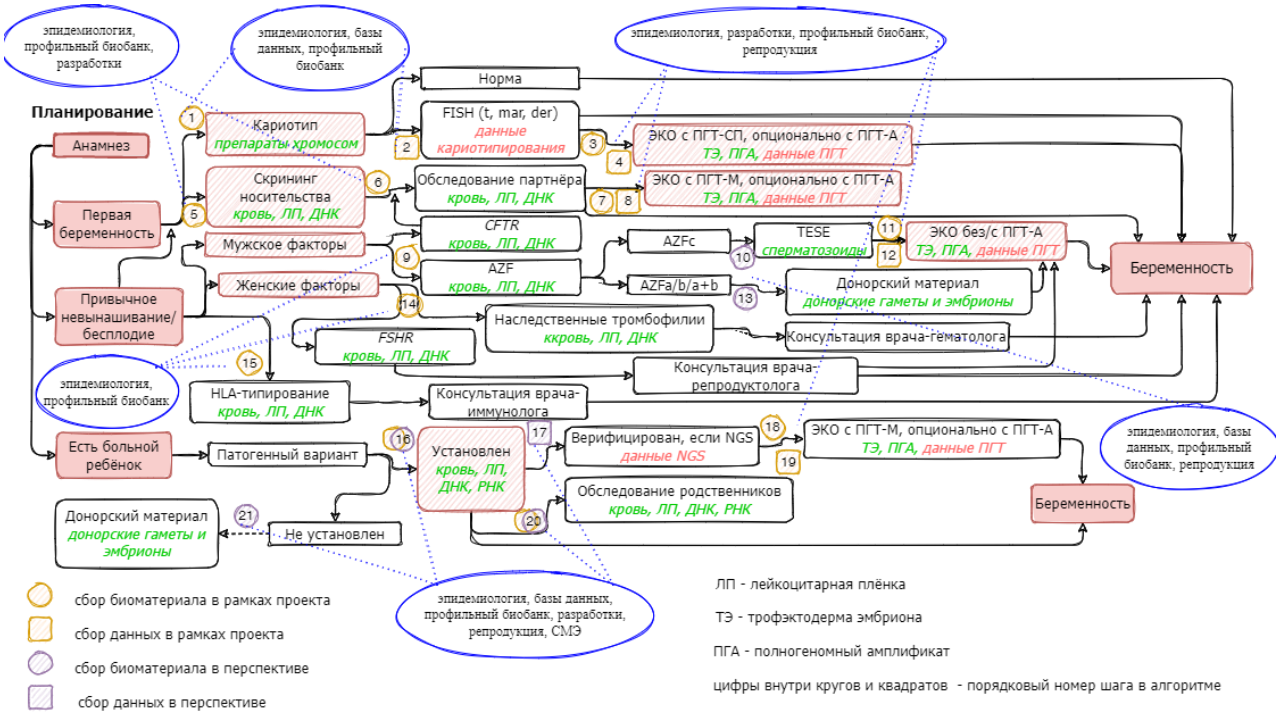
Code	Genes	Pathology/ mechanism	Reference
2011	*KIF7*	Hydrolethalus syndrome, acrocallosal syndrome, Joubert syndrome	Putoux *et al.,* 2011
2013	*WDR60*	Short rib syndrome polydactyly syndrome, type 2	McInerney-Leo AM *et al.,* 2013
2014	*FGFR3, COL2A1, OFD1, PRKDC, DLC1, RERE, ACF1, FRAS1*	Fatal skeletal dysplasia with severe fragility of the occipital bone, tricuspid insufficiency and lower limb anomalies, ventriculomegaly and agenesis of the corpus callosum, etc.	Carss *et al.,* 2014
2014	*KIF14*	Ciliopathy	Filges *et al.,* 2014
2015	*THSD1*	Vascular permeability disorder and defective vascular integrity	Shamseldin *et al.,*2015
2015	*GLE1, RYR1*	Fetal akinesia and arthrogryposis	Ellard *et al.,* 2015
2016	*DYNC2H1, ALOX15*	Ciliopathy and placental dysfunction	Qiao *et al.,* 2016
2016	*FOXP3*	X-linked immune dysregulation syndrome, polyendocrinopathies and enteropathies (IPEX)/disorders of development and maintenance of CD3 + CD4 + CD25+ regulatory T cells	Rae *et al.,* 2016
2018	*ASPM, ATAD3A, ATRX, B3GLCT, BBS9, BBS10, CENPJ, DYNC2H1, ERCC5, ETFA, EXOSC3, FRAS1, GLE1, IFT122, ITGA8, LRP4, MKS1, MRPS22, NEK9, POMGNT1, RYR1, SASS6, TMEM67, TRIP11*	Skeletal dysplasia, fetal akinesia, congenital microcephaly, Bardet-Biedl syndrome, Fraser syndrome, etc.	Stals *et al.,* 2018

In today's Russia, if both spouses are carriers of genetic mutations and pursue having a healthy child, in vitro fertilization (IVF) followed by preimplantation genetic testing (PGT) of embryos is recommended to ensure healthy offspring.

Based on our own studies and literature data, we propose the following algorithm for genetic examination of patients with impaired reproductive function (Figure 18).

Pre-conceptional screening may include the following steps:

1. expert consensus decision (based on clear criteria for disease selection and carrier screening, as well as identification of high-risk individuals who require testing);
2. settlement of ethical and legal issues;
3. pilot projects;
4. settlement of costs of sophisticated medical technology and options for public medical insurance coverage to prevent genetic diseases in couples with high risk of delivering a child with severe monogenic pathology; implement preconception screening, preimplantation genetic testing, and prenatal diagnostics;
5. comprehensive patient guidance and counseling support for families having children with monogenic diseases or mutation carriers identified by neonatal screening.



Сбор биоматериала в рамках проекта Сбор данных в рамках проекта Сбор биоматериала в перспективе Сбор данных в перспективе ЛП – лейкоцитарная пленка ТЭ – трофэктодерма эмбриона ПГА – полногеномный амплификат Цифры внутри круга и квадратов – порядковый номер шага в алгоритме	Specimen collection for the project Data collection for the project Data collection for long-term Specimen collection for long-term BF - blood film ET - embryo trophectoderm WGA - whole genome amplicon Numbers in circles and squares correspond to the step number in the algorithm
Эпидемиология, профильный биобанк, разработки	Epidemiology, cohort biobank, R&D
Эпидемиология, базы данных, профильный биобанк	Epidemiology, datasets, cohort biobank
Эпидемиология, разработки, профильный биобанк, репродукция	Epidemiology, R&D, cohort biobank, reproduction
Эпидемиология, профильный биобанк	Epidemiology, cohort biobank
Эпидемиология, базы данных, профильный биобанк, разработки, репродукция, СМЭ	Epidemiology, datasets, cohort biobank, R&D, reproduction, forensic medicine
Эпидемиология, базы данных, профильный биобанк, репродукция	Epidemiology, datasets, cohort biobank, reproduction
Планирование	Planning
Анамнез	Patient medical history
Первая беременность	Gravida 1
Привычное невынашивание / бесплодие	Recurrent pregnancy loss / persistent infertility
Есть больной ребенок	History of giving birth to an unhealthy child
Донорский материал донорские гаметы и эмбрионы	Donor cells gamete or embryo donation
Кариотип препараты хромосом	Karyotype chromosomes in a sample of cells
Скрининг носительства кровь, ЛП, ДНК	Screening for hereditary carriage Blood, BF, DNA
Мужские факторы	Male factors
Женские факторы	Female factors
HLA типирование кровь, ЛП, ДНК	HLA typing Blood, BF, DNA

Патогенный вариант	Pathogenic variant
Не установлен	Not identified
FSHR кровь, ЛП, ДНК	FSHR expression Blood, BF, DNA
Норма	Reference level
FISH (t, mar, der) данные кариотипирования	FISH (t, mar, der) Karyotype test results
Обследование партнёра кровь, ЛП, ДНК	Partner screening Blood, BF, DNA
CFTR кровь, ЛП, ДНК	CFTR Blood, BF, DNA
AZF кровь, ЛП, ДНК	AZF Blood, BF, DNA
Наследственные тромбофилии кровь, ЛП, ДНК	Hereditary thrombophilias Blood, BF, DNA
Консультация врача-иммунолога	Refer to immunologist
Установлен кровь, ЛП, ДНК	Identified Blood, BF, DNA
ЭКО с ПГТ-СП, опционально с ПГТ-А ТЭ, ПГА, данные ПГТ	PGT-SR for IVF, PGT-A as option ET, WGA, PGT results
ЭКО с ПГТ-М, опционально с ПГТ-А ТЭ, ПГА, данные ПГТ	PGT-M for IVF, PGT-A as option ET, WGA, PGT results
AZFc	AZFc
AZFa/b/a+b	AZFa/b/a+b
Консультация врача-репродуктолога	Refer to fertility specialist
Верифицирован, если NGS Данные NGS	Verified, if NGS NGS data
Обследование родственников кровь, ЛП, ДНК, РНК	Screening of relatives Blood, BF, DNA, RNA
TESE сперматозоиды	Sperm cells obtained by TESE
ЭКО без /с ПГТ-А, ТЭ, ПГА, данные ПГТ	IVF wo/with PGT-A ET, WGA, PGT results
Донорский материал донорские гаметы и эмбрионы	Donor cells gamete or embryo donation
Консультация врача- гематолога	Refer to hematologist
ЭКО с ПГТ-М, опционально с ПГТ-А ТЭ, ПГА, данные ПГТ	PGT-M for IVF, PGT-A as option ET, WGA, PGT results
Беременность	Pregnancy

Figure 18. Genetic screening algorithm for patients with reproductive disorders.

Overall, predictive medicine as a concept shall rely on the clinical genetic passport and allows to resolve challenges of preconception screening and PGT to ensure birth of healthy offspring, adequate diagnosis, etc. Moreover, once applied to monogenic diseases, predictive medicine shall be underpinned by next-generation sequencing as a collateral, as well as dedicated databases, algorithms, bioinformatics, and numerous other adjuvant methods.

CHAPTER II. New Generation Sequencing, Phenotypic Screening, Oligogenic and Multifactorial Diseases

Chapter I stipulated the significant contribution of NGS to our knowledge of human monogenic diseases and the impact of specific genome variants in context of the human genetic health passport. Hereditary diseases caused by pathogenic variants in a few genes are a lot more prevalent – i.e. the so-called oligogenic hereditary diseases [Agarwal and Moorchung, 2005, Kousi and Katsanis, 2015]. Oligogenic diseases are an interim condition between monogenic diseases, associated with one specific defected gene, and polygenic diseases, caused by several genes and exogenic factors. Moreover, according to Lap-Chee Tsui, a luminary scientist, all monogenic diseases shall be considered oligogenic [Baranov *et al.*, 2021].

2.1. Oligogenic Etiology of Cardiomyopathies

Understanding the disease origin is therefore critical. Whether the disorder is mono- or oligogenic, or multifactorial? The answer is often far from evident. Our research on hereditary heart diseases is an attempt to prove this fact [Glotov *et al.*, 2015; Komissarova *et al.*, 2016]. The main goal of this study was to identify genetic markers in patients with cardiomyopathies and students in an at-risk group that were not detected in the student control group and to identify markers present only in the control group. Two additive models were considered. We made the assumption, that clinical effect depends on the presence of alternative alleles or their combinations (dominant models) and alternative genotypes or their combinations between detected variants (recessive models).

These assumptions resulted in four different models. For each model a single table was generated. Our recessive models did not show any significance for determining the at-risk group [Glotov *et al.*, 2015]. However, our dominant models were more informative for these purposes, which was concordant with existing data on autosomal dominant inheritance pattern for cardiomyopathy [Millat *et al.*, 2014]. For each model a single table was generated. Dominant models produced different results

[Glotov *et al.,* 2015]. We can therefore assume that a knowledge of the type of inheritance is pivotal forthis kind of diseases.

The following mutations were found only in the patients with cardio- myopathy and in the at-risk group of students: *MYBPC3* (c.977G>A and c.2678G>T) and *CASQ2* (c.1014+12delg) respectively; *MYBPC3* (c.977G>A) was in two patients and one at-risk student; *MYBPC3* (c.2678G>T) as identified in two patients and four at-risk students; *CASQ2* (c.1014+12delG) as identified in five patients and one at-risk student. These variants were recognized as meaningful for screening individuals at risk for cardiomyopathy, therefore predictive genetic testing for variant c.977G>A in the *MYBPC3* gene has been proposed to participants and their family members to determine individual predispositions for cardiomyopathies [Glotov *et al.*, 2015].

The prevalence of HCM in the general population challenges cardiomyopathy diagnostics due to heterogeneity of pathogenic mutations in the population and their incomplete penetrance [Teekakirikul *et al.,* 2013]. This turns diagnostic screening of cardiomyopathies, and HCM in particular, into a complicated exercise.

According to the SNPSIFT analysis, the most significant mutations were substitutions in the *TNNT2* gene (Table 16). Some of them (c.97+151delC, c.223+92G>C и c.223+93C>G) occurred only in the student control group. These variants can be consid- ered as protective against cardiomyopathy [Glotov *et al.,* 2015]. Meanwhile, the analysis of the complex mutations in other genes, including *MYBPC3* (c.706A>G) - *MYH7* (c.3973-30A>G), *MYBPC3* (c.3288G>A) - *MYH7* (c.1095G>A), *MYBPC3* (c.3815-66C>T) - *MYH7* (c.1128C>T), *MYBPC3* (c.706A>G) - *MYH7* (c.3853+27T>A), *MYBPC3* (c.706A>G) - *CASQ2* (c.939+23C>T) and *MYBPC3* (c.1223+29G>A) - *MYH7* (c.1095G>A) [Glotov *et al.,* 2015], may also be useful, especially because the number of mu- tations in an individual may influence disease severity [Zou *et al.*, 2013]. We speculate that HCM is more likely a complex rather than a single-gene disease. Novel genetic and environmental causes of HCM should be revealed in order to fully determine the pathogenic mechanisms of HCM [Glotov *et al.*, 2015].

Table 16. Main HCM genetic variants identified in patients and the at-risk group compared to the control group [Glotov *et al.*, 2015].

Gene	**Nucleotide change**	**Diseased/risk /healthy, %**	**Risk**	**Risk2**	**p-value**	**Polyphen 2**	**SIFT**	**Clinical verification**
MYBPC3	**c.977G>A (NM_000256.3)**	5/4/0	19	-99	0.41	BENIGN	Damaging	Jääskeläinen *et al.*, 2014
MYBPC3	**c.2678G>T (NM_000256.3)**	5/17/0	16	-96	-	PROBABLY DAMAGING	Damaging	-
CASQ2	**c.1014+12delG (NM_001232.3)**	13/4/0	49	-249	8.62E-05	-	-	-
TNNT2	**c.97+151delC (NM_000364.3)**	0/0/10	-100	20	1.80E-05	-	-	-
TNNT2	**c.223+92G>C (NM_000364.3)**	0/0/29	-300	60	1.902E-07	-	-	-
TNNT2	**c.223+93C>G (NM_000364.3)**	0/0/33	-350	70	2.535E-04	-	-	-

2.2. Monogenic Diabetes Mellitus

The so-called monogenic maturity onset diabetes of the young (MODY) is yet another disorder with multifactorial etiology, accounting for 1 to 6% of severe diabetes mellitus (DM) cases in children and adolescents [Hattersley *et al.,* 2018]. In contrast to T1DM of autoimmune origin, MODY is associated with clinically heterogeneous group of dominantly disorders. To date, 13 genes causative of 13 types of MODY are known leading to moderate or manifest hyperglycemia [Barbetti *et al.*, 2018].

MODY is typically diagnosed before 25 years of age; it is non-insulin dependent and its symptoms are usually mild. However, due to the variety of clinical forms caused by a wide spectrum of mutations in MODY-related genes, different treatment strategies are used: From appropriate diet and physical activity to oral and/or insulin therapy. Today whole-exome sequencing (WES) enables investigators to identify pathogenic variants in other non-*GCK* genes, alongside with the glucokinase protein gene [Glotov O. *et al.*, 2019].

Our data suggest that patients with more than one genetic variant in one or different target genes require extra attention. The presence of genetic variants in different target genes was detected in three patients. In one of them, a *GCK* in-frame deletion was accompanied by an *HNF1A* missense mutation (patient #226). In another one, two missense mutations were present – in *GCK* and *BLK* (patient #529). In the third patient (#662), a splicing defect in *GCK* and missense mutations in *BLK* and *WFS1* were present. The clinical picture in patient #226, who had genetic variants in *GCK* and *HNF1A*, was more typical for MODY2 than for MODY3: He had mild fasting and postprandial hyperglycemia, had no glucosuria, and was successfully being treated by a diet. Both patients #529 and #662 had typical clinical signs of *GCK*-MODY rather than *BLK*-MODY, suggesting an absence of strong accumulation of the pathogenic effect of the detected genetic variants. The clinical picture in patients with more than one genetic variant in one or different target genes is elaborated in Chapter I (see ‘NGS Application to Identify New Variants in Patients’). Our data show an absence of strong accumulation of the pathogenic effect of the detected non-*GCK* genetic variants [Glotov O. *et al.,* 2019]. However, manifestation with age is plausible, considering that MODY3 clinical onset occurs at more senior age that MODY2. This fact shows that patients require further follow-up.

Identification of novel genetic variants as well as accumulating data on previously known causes of monogenic diabetes is of high importance, both for fundamental understanding of the disease pathogenesis and for clinical practice. Whether to attribute this disease to oligogenic disorders (caused by a set of genes) is still a matter of debate. The same stands for familial hypercholesterolemia [Miroshnikova *et al.,* 2021] and congenital heart diseases [Glotov *et al.,* 2015].

Another important aspect is that pathogenic variants in an individual gene can result in different disorders. Figure 19 shows the spectrum of pathogenic variants associated with neonatal diabetes mellitus (NDM) and MODY. Pathogenic variants in the *GCK* gene can be associated with different clinical manifestations of the disease [Glotov O. *et al.*, 2019].

Monogenic diabetes also includes a number of non-MODY transient or permanent neonatal forms occurring under 6 months of age. More than 20 genes are known to be related to congenital neonatal diabetes [Lemelman *et al.*, 2018]. Depending on the gene involved, neonatal diabetes may follow patterns of dominant or recessive inheritance and may be isolated or associated with a variety of syndromic features [Greeley *et al.*, 2011]. However, due to a very early onset of diabetes, hyperglycemia is often diagnosed prior to other syndromic features. The treatment strategy for non-MODY neonatal diabetes depends on the specific genetic defect causing the diabetic phenotype.

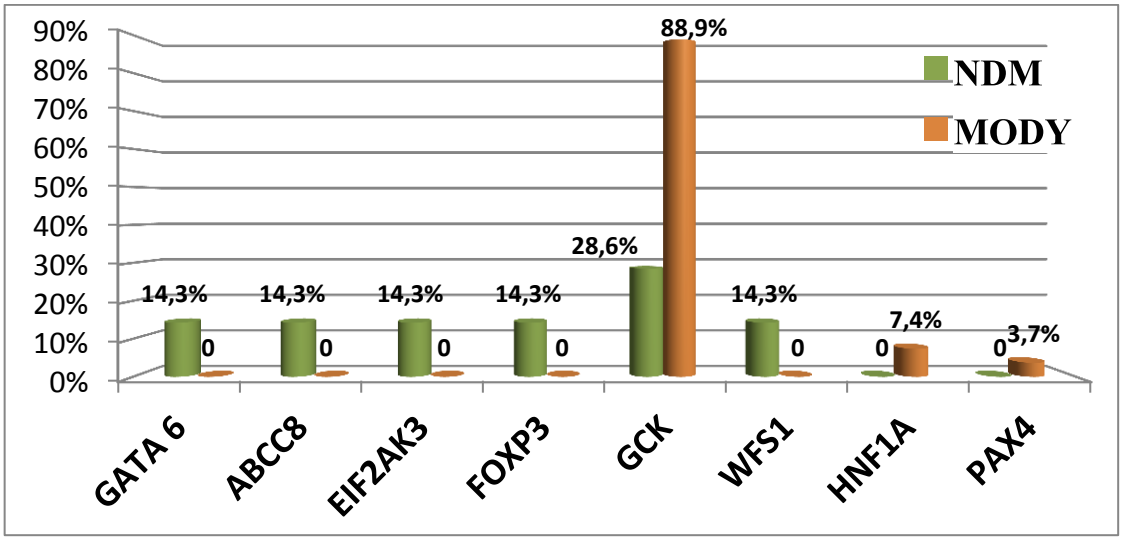


Figure 19. Differences in the prevalence of genetic variants causative of NDM and MODY.

Our data on regarding pathogenic variants in a single gene and their link with various diseases or DM manifestations (MODI and NDM) emphasize the importance and cautious implementation of ‘penetrance’ and ‘expression’ as concepts describing such genetic syndromes in clinical practice. Nosology varieties could as well originate from different variants of transformed function in the same genes, exacerbated by environmental factors.

2.3. Predictive Medicine Framework

The realities of today enable us to rely on verified diagnostics, prevention, treatment, and prediction of complex diseases [Franks *et al*., 2021], but widespread chronic diseases have a complicated, multifactorial etiology that implies interaction between genetic susceptibility and environmental risk factors, broadly referred to as lifestyles, behavior, occupational or environmental impacts, that cannot be underestimated risk assessment [Chatterjee and García-Closas, 2016]. Historically, investigation of familial disorders allowed to detect rare highly penetrant variants underlying some complex diseases, such as familial hypercholesterolemia. These discoveries made genetic testing part of clinical management in individuals from high-risk families. Obviously, the contribution of environmental factors is minimal in monogenic and oligogenic diseases, while in late manifesting diseases and multifactorial diseases, in particular, their role is overriding.

Multifactirial diseases (MFD) include almost all most prevalent chronic disorders, i.e. atherosclerosis, diabetes, obesity, bronchial asthma, osteoporosis, endometriosis, many malignant tumors, neuropsychiatric and cardiovascular diseases, since they result from the interaction of many genes with unfavorable environmental factors [Baranov *et al*., 2000]. Pre-symptomatic detection of individuals in high-risk groups for oligogenic or multifactorial pathology, as well as its primary prevention, are the main tasks of predictive medicine [Baranov *et al.*, 2009; 2021].

Currently, the International Classification of Diseases and Causes of Death includes more than 55,000 nosological units [Petersen, 2021]. The vast majority of them belong to MFDs. Each MFD is characterized by pronounced genetic heterogeneity, largely due to specific candidate gene mutations, their combinations, and the impact of modifier genes and external factors. MFD polyetiology demonstrates specific individual patterns and justifies the need to detect new variants and update the existing classification. As of July 30, 2022, over 12,000 human diseases revealed genome associations (Fig. 20). Currently, the GWAS Catalog includes

5876 publications, 220322 SNPs, and 402121 associations [https://www.ebi.ac.uk/gwas].

Figure 20. GWAS Catalog of Human Genome-Wide Association Studies for MFDs [https://www.ebi.ac.uk/gwas].

Meanwhile, almost all widespread diseases, including some 90% of all malignant tumors, are apparently associated with adverse external factors, such as smoking, unhealthy diet, etc. Exposure to various chemical toxins can provoke the disease onset as well. Numerous genes are known to modify the damaging effects of xenobiotics, including exotoxins. These genes encode proteins (enzymes, n-receptors, and signaling molecules) that interact differently with carcinogens. Therefore, depending on the genome, different individuals can remain resistant or, on the contrary, show increased sensitivity to various damaging agents [Baranov *et al*., 2000]. Thus, we showed the association between polymorphic alleles of *CYP1A1*, *GSTM1*, and *CYP2C9* genes, on the one hand, and the risk of non-Hodgkin lymphoma and/or chronic lympholeukosis [Gra *et al*., 2008], on the other hand, as well as many other diseases (bronchial asthma, pregnancy failure, endometriosis, etc.) and biotransformation system genes [Baranov *et al.*, 2021].

Molecular genetic causes of multifactorial diseases are rather difficult to reveal and require a step-wise approach (Fig. 21). First of all, literature search shall be performed to determine the disease gene network. Population studies are equally important in order to provide an objective assessment of the genetic burden in the population, i.e., to determine the frequency of rare functionally defective alleles of corresponding genes. Next, allele frequencies in the population shall be compared against those in individual diagnosed patients. Such studies require a rather representative cohort of patients and unquestionably healthy individuals, considering the investigated disease. Once specific alleles are associated with the corresponding pathology, hereditary predisposition testing shall be performed in high-risk families. In the relatively near future, randomized testing of hereditary predisposition is going to be routinely possible [Baranov *et al*. 2005].

Today there are three main approaches to the identification of candidate genes: functional mapping method (candidate genes analysis), genetic linkage in high-risk families, whole genome association analysis (GWAS), including genome sequencing.

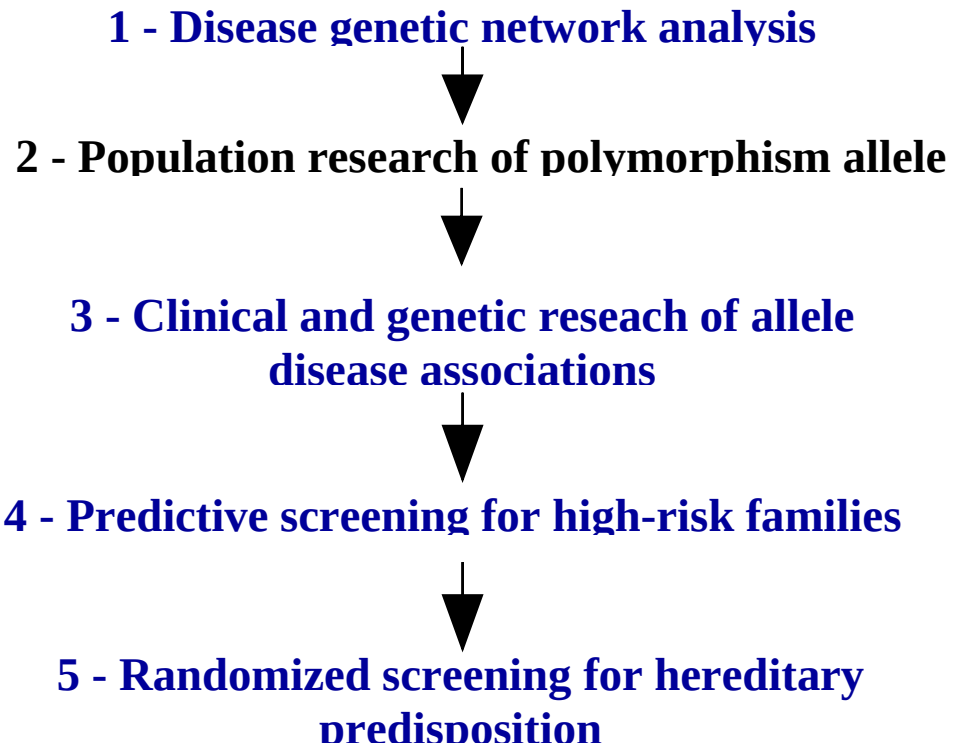


Figure 21. Steps for MFD genetic testing [Glotov, 2007].

Research analysis on genetic testing demonstrates that GWAS has become the key technique to detect MFD candidate genes. It includes the following steps:

1. Genetic testing research demonstrates that GWAS has become the key technique to detect MFD candidate genes. It includes the following steps:

1. collect DNA samples (N>1000) and trait for the study cohort and the control group. The groups shall match specific population characteristics, with elaborate phenotype description;

2. genotype DNA samples using high-density biochips;

3. compare allele frequencies and genotype distributions for dedicated genetic markers in individuals exhibiting a specific trait; identify trait-associated genetic markers;

4. perform verification to confirm the primary screening results. For this purpose, more accurate or alternative genetic testing systems can be used. Extra studies may be required for smaller collections of samples or rather samples obtained from other populations.

GWAS is actively used to analyze and test samples from different national biobanks, including the UK biobank [https://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array]. Data on genotyping samples in various national biobanks and genomic projects, as well as extensive collections of samples empower the research and verification of earlier GWAS results.

International consortia bring together the results produced by individual groups and institutions to further implement GWAS to study genetics of numerous diseases. Those include, for example, the Psychiatric Genomics Consortium and the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) [O’Donovan, 2015; Pierik *et al*., 2005]. However, in addition to GWAS, other approaches are also practical to investigate associations between genetic markers and MFDs.

2.4. Full Genome Sequencing to Assess Genetic Predisposition to Type 2 Diabetes Mellitus

Our study shows that whole exome analysis can serve as a reasonable approach for identifying novel genetic candidate markers of type 2 diabetes mellitus and obesity in limited samples [Barbitoff *et al.*, 2018].

Type 2 diabetes (insulin-resistant diabetes, T2D) and obesity are common chronic disorders with multifactorial etiology. In the recent years the understanding of etiology of these disorders has improved dramatically and genetic high-resolution technologies allowed the identification of 128 susceptibility genetic markers of T2D and more than 700 markers for increased body mass and obesity [Wang *et al.*, 2016; Scott *et al.*, 2018; Yengo *et al.*, 2018]. The genetic architecture of T2D and obesity was elucidated mostly using GWAS [Scott *et al.*, 2018; Fuchsberger *et al.*, 2016; Yengo *et al.*, 2018]. Despite having high statistical power to detect genetic associations, GWAS-derived single nucleotide polymorphisms (SNPs) themselves do not usually have any impact on complex traits; rather, they are in high linkage disequilibrium (LD) with the real causal variants for the disease. Importantly, many complex traits are shaped by a complex interplay between common and rare variants, with the latter usually being missed by conventional GWAS approaches. Next-generation sequencing (NGS) technologies have become an important instrument in identifying genetic causes of obesity and T2D [Lohmueller *et al.*, 2013]. However, WES-based association studies usually suffer from sample size limitations, as thousands of sequenced individuals are usually required to detect exome-wide significant loci, especially for highly polygenic traits. As such, the application of exome sequencing to the analysis of complex traits requires large-scale research efforts and/or development of special methods of bioinformatics analysis. On the other hand, a traditional GWAS approach implying the use of genotyping arrays is cheaper but requires additional studies, such as fine-mapping of causal variants, to get insights into the pathogenesis of complex diseases. Given these limitations of modern approaches, we developed and applied a multi-

perspective approach together with biologically meaningful filtering criteria to detect novel candidate variants and loci for T2D and obesity in a moderate-sized cohort of Russian patients [Barbitoff *et al.*, 2018].

The study comprised 110 patients of Russian ethnicity together with a multi-perspective approach based on biologically meaningful filtering criteria to detect novel candidate variants and loci for T2D and obesity. We have identified several known single nucleotide polymorphisms (SNPs) as markers for obesity (rs11960429), T2D (rs9379084, rs1126930), and body mass index (BMI) (rs11553746, rs1956549 and rs7195386) ($p < 0.05$). Using this method, we identified rs328 in *LPL* ($p = 0.023$), rs11863726 in *HBQ1* ($p = 8 \times 10{-}5$), rs112984085 in *VAV3* ($p = 4.8 \times 10{-}4$) for T2D and obesity, rs6271 in *DBH* ($p = 0.043$), rs62618693 in *QSER1* ($p = 0.021$), rs61758785 in *RAD51B* ($p = 1.7 \times 10{-}4$), rs34042554 in *PCDHA1* ($p = 1 \times 10{-}4$), and rs144183813 in *PLEKHA5* ($p = 1.7 \times 10{-}4$) for obesity; and rs9379084 in *RREB1* ($p = 0.042$), rs2233984 in *C6orf15* ($p = 0.030$), rs61737764 in *ITGB6* ($p = 0.035$), rs17801742 in *COL2A1* ($p = 8.5 \times 10{-}5$), and rs685523 in *ADAMTS13* ($p = 1 \times 10{-}6$) for T2D as important susceptibility loci in Russian population [Barbitoff *et al.*, 2018].

Thus, we found rs328 in the LPL gene, coding for lipoprotein lipase, associated with T2D and obesity simultaneously. The minor allele of rs328 was previously shown to be associated with elevated LDL and decreased HDL, as well as to play its role in the T2D pathogenesis [Mahajan *et al.*, 2018]. We also identified rs6271in the *DBH* gene and rs62618693 in the *QSER1* gene as specific markers for obesity. *DBH* is a gene encoding a dopamine β-hydroxylase (DβH) that catalyzes conversion of dopamine to norepinephrine, which functions both as a hormone and as the main neurotransmitter of the sympathetic nervous system. Earlier it was shown that the rs6271 polymorphic variant affects the plasma DβH activity [Zabetian *et al.*, 2003] and is involved in blood pressure regulation [Ehret *et al.*, 2016]. Importantly, the rs62618693 variant in *QSER1* has also been recently discovered as a T2D marker using fine-mapping of coding variants [Mahajan *et al.*, 2018]. Only one variant inside known causal genes (rs2233984 in the *C6orf15* gene) was identified as a specific T2D

marker when comparing T2D and control groups. This variant is also significantly associated with height [Shungin *et al.*, 2015].

In addition, three additional marker variants (rs9379084 in *RREB1*, rs61737764 in *ITGB6*, rs17801742 in *COL2A1*) were discovered when comparing T2D patients over control and obese groups together. The *RREB1* gene encodes a transcription factor that binds to RAS-responsive elements (RRES) of gene promoters. Earlier it was demonstrated that RREB-1 exerts a repressive activity on the HLA-G and it was also described as a coactivator of calcitonin, c-erbB2, and secretin genes [Flajollet *et al.*, 2009]. Recent studies have shown the association of variant rs9379084 of *RREB1* gene with fat distribution, fasting glucose [Liu *et al.*, 2013] and strong association with T2D [Fuchsberger *et al.*, 2016; Mahajan *et al.*, 2018].

Other results have shown the association between T2D and gene variants. *TGB6* gene encodes an integrin β-6 that is a transmembrane glycoprotein receptor. The rs61737764 variant in *ITGB6* has not been described as a T2D marker; however, it is in modest LD with another previously described non-coding T2D variant, rs7593730. Out of rare case-specific variants, we discovered rs139972217 in *TMC8*, rs61758785 in *RAD51B*, rs34042554 in *PCDHA1*, and rs144183813 in *PLEKHA5* as the most significant candidates ($p < 0.001$). The *TMC8* gene encodes for a transmembrane protein, playing a role in diverse skin diseases. Variants at the TMC6-TMC8 locus have been associated with the levels of glycated haemoglobin (HbA1c), a common biomarker that is used for diagnostics of T2D [Hachiya *et al.*, 2017].

Expression levels of another gene harboring association signal, *PLEKHA5*, are linked to seroconversion behind type 1 diabetes [Mehdi *et al.*, 2018]. These data indicate potential high relevance of the identified variants for pathogenesis of T2D and obesity. Out of exome variants with intermediate frequency ($0.02 <$ SPBU MAF < 0.1) with high case-specificity score and statistical support, we found rs11863726 in *HBQ1* and rs112984085 in *VAV3* which were associated with T2D and obesity compared to controls, and rs685523 in *ADAMTS13* as a specific marker for T2D. *HBQ1* gene encodes the hemoglobin subunit theta 1 that is expressed only in human fetal erythroid

tissue. The function of this gene is poorly understood. No association for polymorphism of *HBQ1* gene with T2D or other endocrine disorders has been described previously. The gene of guanine nucleotide exchange factors *VAV3* is a member of the VAV family of proto-oncogenes. The *VAV3* gene has an impact on angiogenesis, cytoskeleton organization and function, regulation of immune system which renders it a potentially relevant gene for molecular pathology behind T2D [Tsuboi *et al.*, 2016]. The *ADAMTS13* gene codes a multimeric plasma glycoprotein that plays a critical role in platelet adhesion and aggregation on vascular lesions. Previously it was shown that the circulating von Willebrand factor (VWF) concentrations are elevated in T2D patients, and long-term studies of T2D patients have linked VWF to the development of both microvascular and macrovascular disease [Skeppholm *et al.*, 2009]. VWF has been found to be a risk marker for death in T2D [Stehouwer *et al.*, 2002]. The mechanisms behind elevated VWF concentrations in T2D remain unclear, however, these facts imply a potential role of *ADAMTS13* in the pathogenesis of the disease. Overall, all three genes described above seem as relevant targets for further genetic and clinical studies.

Notably, in our study we applied a multi-perspective approach (Fig. 22) to identify candidate markers of T2D and obesity in a cohort of Russian patients with T2D and obesity, which turned out tobe a practical approach for limited cohort. We applied both conventional SNP-level and gene-level association tests, as well as novel strategies to identify variants associated with T2D, obesity, and relevant quantitative traits (BMI, WHR, glucose, and triglyceride concentration).

This approach is based on rational filtering of protein-altering genetic variants and prioritization of case- or control-specific genetic variants, i.e., the ones with the highest potential of being causal for the disease. We show that, while this approach has low power to identify common variants with low OR, it efficiently prioritizes variants of intermediate and low frequency with higher OR (Figure 22 (a), (b)). In order to decrease false discovery probability, which is quite substantial without any additional filters (P (Score $\geq$ 20) = 0.047 for a variant with MAF = 0.02 and OR = 1),

we also apply additional p-value adjustments that allow for selection of statistically justified rare variants with low or moderate type 1 error probability.

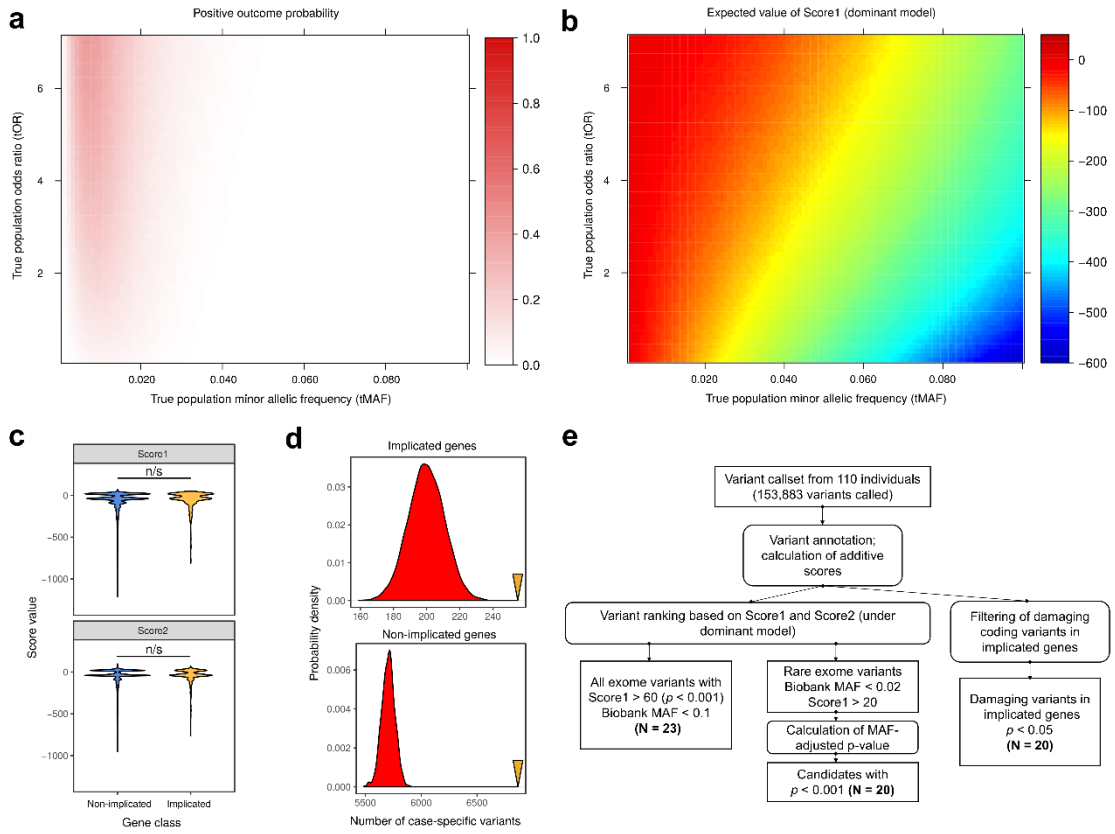


Figure 22. Usage of the scoring and filtering approaches to identify candidate markers of type 2 diabetes (T2D) and obesity in the Russian population [Barbitoff *et al.*, 2018]. **Note**: (a,b). Probability of positive test outcome (Score1 > 10) (a) and the expected value of Score1 (b) for variants with different true population minor allele frequency (MAF) and true odds ratio (TOR), as estimated by in silico simulation. (c). Distribution of values of two additive scores (Score1 and Score2) under dominant inheritance model for damaging coding variants inside genes implicated in T2D and other genes n/s—non-significant difference in U-test. (d). Distributions of the random expectation numbers of case-unique protein-altering variants inside implicated (top) and non-implicated (bottom) genes. Yellow arrowheads indicate observed values. (e). Schematic representation of whole exome sequencing (WES) data analysis used in the present study. Rounded rectangles represent data manipulations.

The application of our approach allowed us to identify potential common and specific markers for diabetes and obesity. We present evidence for potential association of variants in *RREB1*, *ITGB6*, *COL2A1*, *TMC8* and *ADAMTS13* genes with T2D, and of variants in *HBQ1*, *LPL* and *VAV3* with both diabetic and non-diabetic obesity [Barbitoff *et al.*, 2018]. Importantly, some variants (namely, rs685523 in *ADAMTS13* and rs61737764 in *ITGB6*) were specific to the group of T2D patients. It is probable that these markers control the processes initiated by specific metabolic cascades which are less relevant for non-diabetic obesity. Moreover, we observed candidate association of rs6271 in the *DBH* gene, rs62618693 in the *QSER1* gene, and variants in *PCDHA1*, *RAD51B* and *PLEKHA5* with non-T2D-linked obesity. It can be assumed that these markers are potentially involved in the development of obesity as an independent condition. However, this hypothesis requires further confirmation. It is important to note that our analysis strategy allowed us to identify several candidate markers of T2D which have been showed to be significantly associated with the phenotype in one of the most recent fine-mapping studies (e.g., rs328 in *LPL*, rsrs62618693 in *QSER1*, and rs9379084 in *RREB1*) [Mahajan *et al.*, 2018]. As such, our strategy based on filtering for protein-altering variants inside implicated genes can enhance identification of candidate coding markers in smaller samples. On the other hand, we observed that protein-altering variants are significantly overrepresented in cases compared to control individuals both inside and outside known disease-relevant genes (Fig. 22 Б). While this result may be at least partially explained by weak genetic linkage which could not be resolved given a small sample size, recent results [Mahajan *et al.*, 2018] indicate that numerous coding markers for T2D actually lie outside of known disease genes. Hence, it seems useful to consider damaging variants inside both implicated and non-implicated genes in exome sequencing-based studies.

Importantly, most of the rare variants identified by case-specific scoring approach are actually annotated as damaging missense mutations in non-implicated genes, i.e., they belong to the protein-altering class. In our study, we observed several highly case-specific variants in genes previously not directly linked to T2D and/or

obesity (e.g., *TMC8*, *PCDHA1*, *PLEKHA5*, *HBQ1*, *VAV3* and *ADAMTS13*). While many of these variants lack functional validation and were not reported in other studies, genetic alterations or expression changes of some of the corresponding genes are associated withdiabetes-related glycemic traits (e.g., HbA1c levels for *TMC8* [Hachiya *et al.*, 2017] and T1D-related seroconversion for *PLEKHA5* [Mehdi *et al.*, 2018]). Thus, these genes may be selected as potential candidates for further functional investigations and replication of association results.

In conclusion, our study shows that whole exome sequencing (WES) can serve as a reasonable approach for identifying genetic markers of a complex disease in limited cohorts of poorly studied populations. As such, this approach may assist in disease gene identification for polygenic traits.

2.5. Most Optimal Statistical Approaches to Genetic Predisposition Assessment

Various statistical methods are currently used to analyze disease risk factors and evaluate the genotype-phenotype relationship, including classical null hypothesis analysis using Chi-square or Fisher's exact test with calculation of the odds ratio (OR) and some others [Rebrova, 2003]. These methods have proven well in the study of genetic causes of hereditary diseases, but they are insufficiently informative to assess the risk of hereditary predisposition. In addition, the value of these methods is dramatically impaired in case of multiple testing (using Bonferroni correction to reduce the probability of type 1 error), when the sample size is limited, genetically heterogeneous and includes correlated variables [Lvovs *et al.*, 2012].

There are different methods to assess predictive power or quality of statistical tests. Let's focus on a few of them. They include data validation using ROC curves, analysis of false positive and false negative results, assessment of specificity, sensitivity, and the test clinical significance. The test quality can be verified using an expert scale for AUC values. The higher the AUC value, the better the quality of the classifier, with a value of 0.5 demonstrating the unsuitability of the implemented classification method (corresponds to random guessing)

[http://www.biometrica.tomsk.ru], the closer is the sensitivity score to 1, the more efficient is the test of choice in diagnosing a particular disease [Vlasov, 2004]. In case of high specificity, a positive result justifies further differential diagnostics [Vlasov, 2004]. Despite the importance of sensitivity and specificity evaluation, those are merely operational test parameters and do not mean any likelihood of the disease once the diagnostic test is conducted. In practice, physicians are mainly interested in disease probability in case of both positive and negative test result [Vlasov, 2004]. Therefore, ROC curves are elaborated by measuring the predictive value of a test for both positive and negative results [Vlasov, 2004]. The predictive value of a positive test result (PPV) is equal to the probability of disease for a positive test result. The prognostic value of a negative result (NPV) equals to the probability of disease absence for a negative test result. If PPV=0.8, then a positive patient has a 0.8, or 80% probability of disease. All diagnostic efforts pursue to obtain values within that range [Vlasov, 2004]. In practice, however, evaluation of PPV and NPV indices requires further verification against the test sensitivity and specificity. Therefore, high reliability of the final diagnosis (i.e. high predictive value of the final result) yet not a proof of test efficacy. It is equally feasible that a test of a high predictive value merely produced by the distorting impact of pre-selection factors in the patient cohort [Vlasov, 2004].

There is currently no one-size-fit-all approach to genetic predisposition studies. Moreover, there is no uniform criteria to assess how adequately and efficiently the MFD risk is evaluated. Therefore, a comprehensive and balanced approach is required to detect markers of a disease and assess their risk prediction power.

A few years earlier we proposed the Total Allele Combination score and the Total Unfavorable Score for renin-angiotensin and kinin-bradykinin genes to assess the impact of a set of polymorphic genes on the pathological process. Earlier this approach proved efficient in the study stable arterial hypertension (AH) in children, revealing that allelic variants of renin-angiotensin and bradykinin genes are associated with stable hypertension in girls [Glotov *et al*., 2007; Baranov *et al*., 2009b]. The

score was calculated by the formula $\sum m/k$, where m is the sum of scores for all individual genotypes and k is the number of genes studied. Calculation was performed for every networks of genes (metabolic system), the total sum of points for each group was analyzed. It is important to note that by calculating the total sum of points we assume that every gene equally contributes to the risk of disease development, while the disease is the sum of additive pathological responses controlled by predisposition genes. Meanwhile, the clotting factor V Leiden mutation shows that other scenarios are not impossible. Moreover, further clinical and lab data compelled us to continuously update the real score for every gene, sometimes with significant modifications [Baranov *et al*., 2009b].

Every human, as any living being, has his own unique phenotype, different from any other similar organism. Phenotype is defined as the sum of all characteristics, both external, such as height, size, eye color, and the number of fingers on their hands and feet, as well as physiological and biochemical characteristics. The majority of phenotypic traits pertain to complex properties, controlled by multiple genetic and environmental factors [Inge-Bechtomov, 2010]. According to modern concepts, the main effect of genetic factors on the expression and penetrance of various phenotypic traits is largely due to the existence of genetic polymorphism caused by point mutations, Single Nucleotide Polymorphism (SNP). The regulation of complex phenotypic traits depends both on rare SNPs with strong phenotypic effects and on the relatively common SNPs with mild effects [Reich, Lander 2001; Pritchard, Cox, 2002].

In humans, a large number of SNPs has been discovered in a variety of genes and genetic regions which control various quantitative phenotypic characteristics, such as height, weight, and the spectrum of produced lipoprtoteins [Aulchenko, 2010]. To understand the genotype–phenotype correlation, the preferable approach is to investigate candidate genes associated with multifactorial diseases (MFDs) [Baranov, 2009].

2.6. Polygenic Effects Underlying Anthropometric Analysis, Lipid and Physiological Metabolism

The length of the human body (height) represents a classic polygenic trait, vital for a person's identification. From a genetic point of view, a person's height is a polygenic familial trait. Multiple studies have shown that genetic factors contribute to 80–90% of the variability in an individual's height [Weedon et al., 2007; Aulchenko, 2010]. Various mathematical models, including regression model, are used to estimate the genetic contribution to the phenotype.

We developed a novel algorithm to predict human height based on the linear regression method [Glotov *et al.*, 2014]. During development of the prediction algorithm, we used the individual data from the personal questionnaire form (height and sex) and the genotypic parameters for the 13 gene markers – *EFEMP1, ZBTB38, HHIP, LCORL, ADAMTSL3, CDH13, JAZF1, IGF1R, GHSR, CABLES1, IFNG, VDR3, IGFBP3*. The highest coefficient values were obtained for the male population. Using the developed model, we were able to accurately predict the height of men in the standing position within accuracy of up to 4.6 cm, and in the sitting position with accuracy of up to 2.9 cm. The obtained values for the coefficient for women were 0.013 in the standing position and 0.006 in the sitting position. These values are considerably lower than the ones obtained for men: 0.109 for height measurements in the standing position, and 0.127, in the sitting position. However, despite these differences in the determination coefficient, the accuracy of height prediction in women was practically the same as in men. These results indicate a greater effect of genetic factors in determining height in men. Notably, the predictions for the male height in the sitting position were more precise than in the standing position. During measurements of the height in the standing position, the person is standing up with his back to a vertical post and touching it with the heels, buttocks, and shoulder blades. In contrast, during the measurement of height in the sitting position, a person sits on a bench, touching the post with the buttocks and shoulder blades. Therefore,

measurements of the height in the sitting position do not account for the length of the lower extremities, and this value is necessary for proper determination of the proportions of the human body. Therefore, we suggest that a person's height in the standing position is a more complex trait than the height in the sitting position. Notably, most mathematical models provide better predictions for the simple rather than complex traits [Aulchenko, 2010]. Thus, we developed and tested a novel algorithm for predicting a person's height as a quantitative trait. Further amplification of this model by including new genetic parameters and accounting for environmental factors seems very promising [Glotov *et al.*, 2014].

We used a similar approach to analyze the association between polymorphism in lipid metabolism genes, on the one hand, and body mass index (BMI), waist circumference, and blood lipidogram parameters, on the other hand, in women [Tarkovskaya *et al*., 2012]. Logistic regression was implemented to build a model to predict quantitative traits (BMI, WHR, CHL, LDL-CHL, VLDL-CHL). The input parameters relied on individual SNP-associated alleles: homozygous high frequency alleles (in the studied cohort) were labelled '0', heterozygous as '1', homozygous low frequency alleles as '2'. We implemented the stepwise inclusion and removal method (stepwise algorithm) the select the most adequate approach relying on relevant parameters (SNPs). The Akaike Information Criterion (AIC) was used to boost model accuracy. Our prediction model was based on individual questionnaire data (BMI, WHR), biochemical parameters (CHL, LDL-CHL and VLDL-CHL) and genotype data for 36 markers. Once models for different parameters were obtained, we had to identify the most accurate one in terms of its ability to predict specific parameters in humans based on their genetic traits. It is common knowledge that adjusted R2 coefficient allows to compare and evaluate the efficiency of different models, relying on a discordant set of factors [McCullagh and Nelder, 1989]. The R2 coefficient is therefore of paramount importance. With its value ranging within [0;1], the R2 coefficient estimates the predictive power of factors and their combinations to reflect value change for a particular trait. When the R2 coefficient is 0, the model is

considered void. The analysis of observed and predicted values is fundamental for model efficiency evaluation. Discrepancy between values is called prediction error. Since prediction error value directly depends on a specific sample (with the set of most significant parameters of all the parameters under consideration, as well as their coefficient values), prediction error is therefore considered a random value. In order to measure the impact of error, the R2 coefficient is deployed to display the total contribution of selected parameters.

According to our data, the R2 coefficient produced highest indices for VLDL-CHL (R2=0.101). Moreover, this parameter had the lowest value prediction error (± 0.21 mmol/liter), opposite to the total cholesterol error (± 0.95 mmol/liter) and LDL (± 0.83 mmol/liter). The result we obtained indicates that VLDL-CHL has a greater genetic determination compared to the other parameters [Tarkovskaya *et al*., 2012].

To verify the data obtained using regression models, we used the method of correlation analysis. The Kendall rank correlation method (τ) was applied to study the correlation between the genes and the analyzed traits. This method allowed us to analyze qualitative ordinal traits (genotypes), as well as compare qualitative ordinal traits (genotypes) to quantitative ones [Rebrova, 2003]. In part, Kendall correlation analysis confirmed the association between the genotypes and BMI, WHR, CHL, LDL-CHL, and VLDL-CHL identified using the regression linear model. We should also note that the linear regression method allowed us to establish a number of new correlations that were overlooked when using correlation analysis.

Thus, the genes identified by the regression model and verified by the correlation analysis certainly impact individual phenotypic and biochemical traits. Some parameters (in women) produced the adjusted R2 coefficient over 0.1, which is in line with the critical role of genetics in determining the traits underconsideration. However, food, diet, physical activity, stress, hazardous habits, environment, medications can play an equally critical role and therefore override the impact of genetic factors [Tarkovskaya *et al*., 2012].

Our study of athletes is another attempt to apply the regression model to assess phenotype traits based on medical history, genetic and laboratory results [Glotov O. *et al.*, 2015]. It is known that repeated monitoring of various physiological parameters, reflecting traits of genotype – such as lung vital capacity (LVC), pulse, blood pressure, body mass index – allows to assess cardiovascular risk in athletes [Wood *et al*., 1998; Puthucheary *et al*., 2011]. Given the broad individual and temporal variability, it is worth analyzing the extent of correlation with the genome and, if any, identifying different health patterns [Puzyrev, 2011; Xu and Hu, 2010]. Our objective was to analyze correlations and produce regression models to measure the forecasting value of physiological parameters relevant for athletes. We implemented regressor selection methods to establish a framework to further genetic testing of multiples genes in large cohort of patients. We pursued the investigation of LVC as a predominantly hereditary trait that may transform under extensive physical activity. LVC modeling utterly important for preventive and sports medicine, prediction of athletic success and physical endurance in order to duly undertake adequate mitigating measures [Allen *et al*., 2010]. Regression analysis proved significantly more sensitive than other statistical methods. The regression model for dedicated traits was based on individual questionnaire responses, blood chemistry, and genotyping of 26 genes [Glotov O. *et al*., 2015]. The regression model revealed statistically significant correlations between LVC and alleles of the following genes: *AGTR2* (negative, minor allele A), *NOS3* (positive, minor allele 4), *CNB1* (negative, minor allele D), *ADRB2_81CG* (negative, minor allele G). The Kendall correlation coefficient was remarkable for the *NOS3* gene only. The results indicate greater sensitivity of regression analysis compared to correlation analysis. The regression model allows to perform preliminary evaluation of individual phenotypic traits (e.g. LVC) based on genetic testing and a set of extra parameters. Since LVC is a predominantly hereditary indicator of MFD risk, our technique is a feasible risk-assessment tool, allowing to formulate forecasts about individual athletic success and physical abilities [Glotov O. *et al.*, 2015].

2.7. Prospects of Comprehensive Individualized Screening for MFD Polygenic Factors

Overall, reinforced by balanced accuracy analysis and the AUC measurement, most accurate models are to encompass medical history, genetic and clinical data to produce reliable phenotype prediction and MFD risk forecasts. However, once we include the premorbid background in our calculations, it is no longer possible to use any of the above models to assess the pathology risk at the preclinical stage, leaving strategies for disease correction at the initial stage as a refuge tool. Meanwhile, most researchers apply the general linear model (GLM) for it allows to assess both qualitative and quantitative traits [McCullagh, Nelder, 1989; Yi, Banerjee, 2009; Glotov *et al*., 2012; Huang *et al*., 2014]. The main advantage is that phenotype prediction relies both on patient genotype data and medical history, clinical and lab data and other variables. In total, to assess genomic predictive ability, investigators shall build the ROC curve and measure the area under the curve (AUC).

A striking attempt to calculate the disease development risk is suggested in the paper by de Haan and colleagues [de Haan *et al*., 2012]. In 2012 no handy risk models were available to accurately predict the risk of venous thrombosis in humans. Therefore, their objective was to find out whether the inclusion of a single nucleotide polymorphism (SNP) in a venous thrombosis risk model improves the risk prediction. The investigators calculated genetic risk scores by counting risk-increasing alleles from 31 venous thrombosis-associated SNPs for subjects of a large case-control study (2,712 patients and 4,634 controls). Genetic risk scores based on all 31 SNPs or on the 5 most strongly associated SNPs performed similarly (areas under receiver-operating characteristic curves [AUCs] of 0.70 and 0.69, respectively). For the 5-SNP risk score, the odds ratios for venous thrombosis ranged from 0.37 (95% confidence interval [CI], 0.25-0.53) for persons with 0 risk alleles to 7.48 (95% CI, 4.49-12.46) for persons with more than or equal to 6 risk alleles. The AUC of a risk model based on known nongenetic risk factors was 0.77 (95% CI, 0.76-0.78). Combining the nongenetic and

genetic risk models improved the AUC to 0.82 (95% CI, 0.81-0.83), indicating good diagnostic accuracy. To become clinically useful, subgroups of high-risk persons must be identified in whom genetic profiling will also be cost-effective [de Haan *et al*., 2012]. In this study, individual SNPs were not significantly associated with recurrent venous thrombosis. However, when risk alleles of individual SNPs were put together, the risk score went up, as well as the significance of the association itself (see Table 17 and Fig. 23).

Table 17. Genetic risk score prediction based on the 5 strongest SNP associations (by de Haan *et al.*, 2012).

			MEGA				
			Risk allele frequency, %				
Gene	SNP	Chromosome	Cases	Controls	OR	95% CI	Literature average, OR
F5	rs6025	1	10	3	4.30	3.70-4.99	3.79
F2	rs1799963	11	6	2	3.01	2.36-3.85	2.78
ABO	rs8176719	9	47	34	1.74	1.63-1.87	1.85
FGG	rs2066865	4	34	27	1.41	1.32-1.51	1.56
F11	rs2036914	4	59	52	1.35	1.26-1.44	1.32

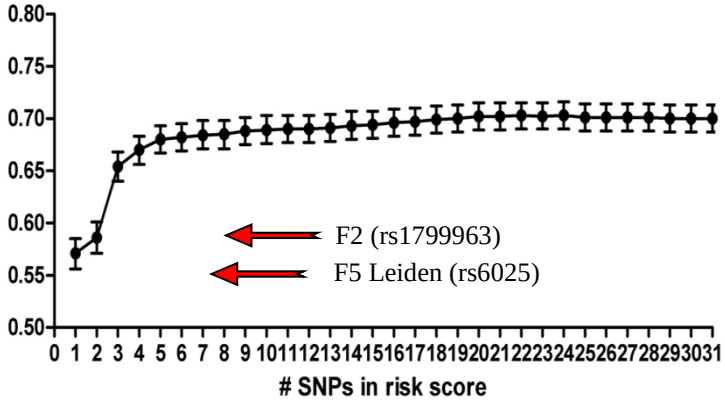


Figure 23. Area under the ROC of genetic risk scores for venous thrombosis (by de Haan *et al.*, 2012).

Predictive ability of multiple SNP analysis had not been studied for first events of venous thrombosis before that publication. Genetic profiling may guide decisions on prophylactic measures in high-risk groups, such as cancer patients, persons undergoing surgery, persons requiring a plaster cast, or those subject to prolonged immobilization.

To explore to what extent venous-thrombosis associated SNPs can be used as predictors for a first venous thrombosis in the general population and in high-risk groups, de Haan and colleagues [de Haan *et al*., 2012] investigated 31 SNPs in 2 large population-based case-control studies, of which one was used as a validation set. They created genetic risk scores based on these SNPs and a risk score based on nongenetic risk factors.

They also compared and combined the genetic risk score with the nongenetic risk score to determine whether genetic profiling with the currently known SNPs will improve the assessment of venous thrombosis risk. The AUC of the 5-SNP risk score (0.68, 95% CI, 0.67-0.70) was significantly higher than the AUC of family history (0.58, 95% CI, 0.57-0.60), with a similar trend among all high-risk subgroups. The AUC for non-genetic risk, including family history, was 0.77 (95% CI 0.76-0.78). When de Haan and colleagues added the genetic risk score to the nongenetic score, the AUC increased significantly to 0.82 (95% CI, 0.81-0.83) compared with the nongenetic risk score alone ($P < 0.0001$) for either 31-SNP or 5-SNP risk scores (Fig. 24).

Table 18. Genetic assessment based on the 5 strongest SNP associations, family history, nongenetic, and combined risk scores (by de Haan *et al.*, 2012).

Risk groups	Patients, N	Controls, N	Family history risk score, AUC (95% CI)	5-SNP risk score, AUC (95% CI)	Nongenetic risk score, AUC (95% CI)	Combined risk score, AUC (95% CI)
Surgery	292	111	0.60 (0.55-0.66)	0.66 (0.60-0.72)	0.67 (0.61-0.72)	0.73 (0.67-0.78)
Paster cast	111	18	0.61 (0.48-0.73)	0.73 (0.59-0.87)	0.70 (0.56-0.84)	0.78 (0.64-0.91)
Hospitalization	278	93	0.57 (0.50-0.63)	0.66 (0.59-0.72)	0.60 (0.53-0.66)	0.66 (0.59-0.72)
Oral contraceptives*	513	327	0.58 (0.54-0.62)	0.71 (0.68-0.75)	0.73 (0.69-0.76)	0.81 (0.78-0.84)
HRT	58	90	0.59 (0.49-0.68)	0.71 (0.63-0.80)	0.74 (0.66-0.82)	0.80 (0.72-0.87)
Pregnancy/postpartum *	67	46	0.54 (0.44-0.65)	0.70 (0.60-0.79)	0.68 (0.57-0.79)	0.76 (0.66-0.85)
Age > 50 y	944	1534	0.57 (0.55-0.60)	0.68 (0.66-0.70)	0.73 (0.71-0.75)	0.79 (0.77-0.81)
Travel	379	610	0.58 (0.54-0.62)	0.70 (0.67-0.73)	0.77 (0.73-0.80)	0.82 (0.80-0.85)
Family history	659	551	-	0.68 (0.65-0.71)	0.74 (0.71-0.76)	0.81 (0.78-0.83)

Malignancies	156	65	0.57 (0.49-0.65)	0.60 (0.52-0.68)	0.71 (0.64-0.78)	0.720.65-0.80)

*Women younger than 50 years.

Women performed better than men in both non-genetic and combined risk scores (non-genetic risk score: AUC = 0.81, 95% CI, 0.80-0.83 in women and AUC = 0.74, 95% CI, 0.72-0.75 in men; combined risk score: AUC = 0.85, 95% CI, 0.83-0.86 in women and AUC = 0.80, 95% CI, 0.78-0.81 in men).

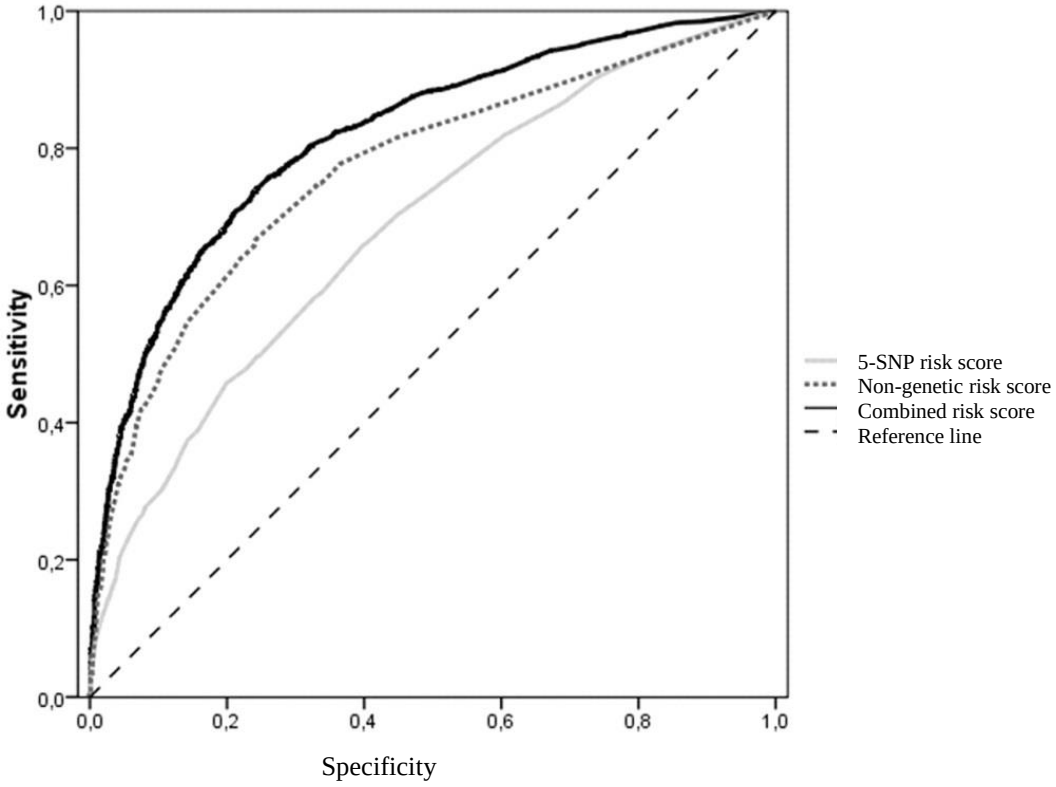


Figure 24. AUC curves of the weighted risk score (by de Haan *et al.*, 2012).

The discriminative accuracy of both the 5 SNP and the 31 SNP risk score in subjects from another population, the LETS population was similar (Table 19), indicating that both risk scores are reliable enought [de Haan et al., 2012].

Table 19. Risk assessment of venous thrombosis using genetic, non-genetic, and combined risk scores (by de Haan *et al.*, 2012).

	MEGA (N = 7092)		LETS (N = 881)	
	AUC (95% CI)	r2	AUC (95% CI)	r2
31-SNP risk score	0.71 (0.69-0.72)	0.161	0.69 (0.65-0.72)	0.149
5-SNP risk score	0.69 (0.67-0.70)	0.135	0.67 (0.64-0.71)	0.138
Genetic risk score	0.77 (0.76-0.78)	0.288	0.71 (0.68-0.74)	0.200
Combined risk score	0.82 (0.81-0.83)	0.378	0.77 (0.74-0.80)	0.292

When preventive measures after a positive test are invasive or can have harmful side effects, strict discrimination is required between those at high risk and low risk of developing a specific disease. In the case of venous thrombosis, indiscrimination may lead to an increased risk of thrombosis in high-risk persons receiving insufficient prophylactic anticoagulant treatment, whereas persons at low risk receiving treatment are at an increased risk of major bleeding. The paper investigated the extent to which genetic risk scores can improve the accuracy of thrombosis risk assessment by ROC curves. Although the proportion of variability explained by the 5-SNP risk score is smaller than by the 31-SNP risk score, de Haan with colleagues [de Haan *et al.*, 2012] showed that the discriminative accuracy of the 5-SNP and 31-SNP risk scores was similar. The 5-SNP genetic risk score performed better than family history assessment, which is the current clinical practice of risk assessment in persons exposed to known nongenetic risk factors. However, the 5-SNP genetic risk score performed worse than a risk score of nongenetic risk factors. Thus, addition of the 5-SNP genetic risk score to the nongenetic risk score model significantly improved the AUC to 0.82, indicating good diagnostic accuracy. Information on the nongenetic risk factors was less complete, which explains the lower discriminative accuracy of both the nongenetic risk score and the combined risk score. Identification of persons at risk of developing venous thrombosis is most useful in high-risk populations.

Lately [Khera *et al.*, 2018] developed risk scores for atrial fibrillation, type 2 diabetes, breast cancer, inflammatory bowel disease, and ischemic heart disease (IHD), with AUCs standing at 0.77, 0.72, 0.68, 0.63, and 0.81, respectively. However, these results also rely on additional variables such as age and gender. If the risk score relies on generic SNPs only, respective AUCs are likely to be lower [Lello *et al*., 2019], whereas the obtained AUCs are in the range of 0.580-0.707 (Table 20) using SNP data only.

Table 20. Genetic AUCs for SNP only (age or sex not considered). Training and validation are performed using UKBB [Lello *et al.*, 2019].

Condition	Training set	Test set	AUC	Active SNPs	λ*
Hypothyroidism	impute	UKBB	0.705 (0.009)	3704 (41)	1.406e-06 (1.33e-7)
Hypothyroidism	impute	eMERGE	0.630 (0.006)		
T2D	impute	UKBB	0.640 (0.015)	4168 (61)	6.93e-06 (1.73e-6)
T2D	impute	eMERGE	0.633 (0.006)		
Hypertension	impute	UKBB	0.667 (0.012)	9674 (55)	4.46e-6 (4.86e-7)
Hypertension	impute	eMERGE	0.651 (0.007)		
Resistant hypertension	impute	eMERGE	0.6861 (0.001)		
Asthma	calls	AA	0.632 (0.006)	3215 (16)	2.37e-6 (0.35e-6)
T1D	calls	AA	0.647 (0.006)	50 (7)	7.9e-7 (0.1e-7)
Breast cancer	calls	AA	0.582 (0.006)	480 (62)	3.38e-6 (0.05e-6)
Prostate cancer	calls	AA	0.6399 (0.0077)	448 (347)	3.07e-6 (0.08e-8)
Testicular cancer	calls	AA	0.65 (0.02)	19 (7)	1.42e-6 (0.04e-6)
Glaucoma	calls	AA	0.606 (0.006)	610 (114)	8.69e-7 (0.71e-7)
Gout	calls	AA	0.682 (0.007)	1010 (35)	9.41e-7 (0.03e-7)
Atrial fibrillation	calls	AA	0.643 (0.006)	181 (39)	8.61e-7 (0.94e-7)
Gallstone disease	calls	AA	0.625 (0.006)	981 (163)	1.01e-7 (0.02e-7))
Heart attack	calls	AA	0.591 (0.006)	1364 (49)	1.181e-6 (0.002e-7)
High cholesterol	calls	AA	0.628 (0.006)	3543 (36)	2.4e-6 (0.2e-6)
Malignant melanoma	calls	AA	0.580 (0.006)	26 (15)	9.5e-7 (0.8e-7)
Basal cell carcinoma	calls	AA	0.631 (0.006)	76 (22)	9.9e-7 (0.3e-7)

Substantially higher predictor AUCs are obtained when incorporating additional variables such as age and sex (see Table 21). Rapid improvement in genomic prediction is anticipated as more case-control data become available for analysis. The results indicate that substantial improvements in predictive power are attainable using training sets with larger case populations [Lello *et al.*, 2019].

Table 21. AUCs obtained using sex and age alone, SNPs alone, and all three together [Lello *et al.*, 2019].

Condition	Teest set	Age + Sex	Genetic only	Age + Sex + Genetic
Hypertension	UKBB	0.638 (0.018)	0.667 (0.012)	0.717 (0.007)

Condition	Teest set	Age + Sex	Genetic only	Age + Sex + Genetic
Hypothyroidism	UKBB	0.695 (0.007)	0.705 (0.009)	0.783 (0.008)
T2D	UKBB	0.672 (0.009)	0.640 (0.015)	0.651 (0.013)
Hypertension	eMERGE	0.818 (0.008)	0.651 (0.007)	0.851 (0.009)
Resistant hypertension	eMERGE	0.817 (0.008)	0.686 (0.007)	0.864 (0.009)
Hypothyroidism	eMERGE	0.643 (0.006)	0.630 (0.006)	0.697 (0.007)
T2D	eMERGE	0.565 (0.006)	0.633 (0.006)	0.651 (0.007)

The significant heritability of most common disease conditions implies that at least some of the variance in risk is due to genetic effects. With enough training data, modern machine learning techniques enable us to construct polygenic predictors of risk. A learning algorithm with enough examples to train on can eventually identify individuals, based on genotype alone, who are at unusually high risk for the condition. This has obvious clinical applications: scarce resources for prevention and diagnosis can be more efficiently allocated if high risk individuals can be identified while still negative for the disease condition. This identification can occur early in life, or even before birth [Lello *et al.,* 2019].

Based on the above, we can already talk about precision medicine (advanced clinical diagnostics) today. Though medical history predicts the MFD risk somewhat better than genetic data with their low weighed accuracy, it is genetic models that allow us to allocate risk groups pre-symptomatically or when clinical information is missing. However, if we combine medical history and the genetic model, such GLM-model is feasible enough for pathology risk assessment and has high weighed accuracy [Glotov A.S., 2017].

Currently, the use of genomics to improve diagnosis, prevention, treatment and prognosis is widely discussed [Collins and Varmus, 2015; Baranov *et al*., 2021] to already become standard of medical care in cancer, rare disorders, as well as common diseases.

Today, precision medicine is based on genomic data (i.e. genome-guided medicine) and includes:

- differential diagnostics;

- prevention;
- treatment;
- forecasting.

Table 22. Efficiency of clinical forecasting based on medial history, genetic and clinical data [Glotov A.S., 2017].

Availability of patient's clinical data:	Category of markers				Recommended method
	Genes	Medical history	Clinical data, blood chemistry, immunological data, instrumental tests (clinical records)	Biomarkers	
None	+	-	-	-	MDR
Medical history	++	++	-	-	GLM
Clinical records	+++	++	++	+++	
Medical history + premorbid background	+++	++	+++	++++	
Diagnosis	+++.	+++	++++	++++	

Prompt clinical and technical timeframes of screening is an important aspect of precision medicine that can benefit physicians and patients in diagnosing complicated diseases. Although genomics is likely to play a key role in future medicine, demographic and standard clinical data (e.g., age, sex, past medical history, current health status, family history, non-genetic biomarkers, and exposure to environmental factors) cannot be overlooked. It is likely that genetic data are useless for certain types of precision medicine. In some cases, disease monitoring is likely to be based on reassessments of other OMICs technologies, e.g., RNA sequencing to measure gene expression.

Our paper is an attempt to detect such adaptation biomarkers [Glotov A. *et al.*, 2022]. Although high altitude training has been increasingly popular among endurance athletes, the molecular and cellular bases of this adaptation remain poorly understood. We aimed to define the underlying physiological changes and screen for potential biomarkers of adaptation using transcriptional profiling of whole blood. Seven elite female speed skaters were profiled on the 18th day of high-altitude adaptation. Whole blood RNA-seq before and after an intense 1 h skating bout was used to measure gene

expression changes associated with exercise. In order to identify the genes specifically regulated at high altitudes, we have leveraged the data from eight previously published microarray datasets studying blood expression changes after exercise at sea level. Using cell type-specific signatures, we were able to deconvolute changes of cell type abundance from individual gene expression changes. Among these were *PHOSPHO1*, with a known role in erythropoiesis, and *MARC1* with a role in endogenic NO metabolism. We find that platelet and erythrocyte count uniquely responds to altitude exercise, while changes in neutrophils represent a more generic marker of intense exercise. Publicly available data from both single cell atlases and exercise-related blood profiling dramatically increases the value of whole blood RNA-seq for the dynamic evaluation of physiological changes in an athlete's body [Glotov A. *et al.*, 2022].

For almost all complex diseases, genetic risk is probabilistic and nondeterministic (the latter is true for diseases caused by highly penetrant variants). This incurs challenges since the risk score is more difficult to assess than presence or absence of a known pathological variant. A potential advantage is that an increased MFD genetic risk is preventable and early treatment can be administered thanks to lifelong stability of nuclear DNA variants [Glotov A.S., 2017].

As mentioned above, in more common multifactorial diseases several research centers were able to shown great potential of genetic tests, although the path from research to clinical practice is yet to be covered. A common way to characterize the genetic risk of complex diseases is to use *polygenic risk score* (PRSs) [Franks *et al*., 2021]. A PRS is the sum of multiple (sometimes thousands of) genetic variants that produce little effect if taken individually. Recent studies show that high PRSs create large and potentially clinically significant risks in adulthood for such diseases as cardiovascular disease and type 2 diabetes [Khera *et al*., 2018] or are associated with impaired survival in systemic lupus erythematosus [Reid *et al.*, 2020]. Since most complex diseases are heterogeneous in etiology, genetics can aid in diagnostics by identifying subgroups/subtypes within unconventional comprehensive diagnostics

procedures, promising targeted treatment [Franks *et al.*, 2021]. An example would be ischemic stroke and its subtypes of different origin (large vessel occlusion, small vessel occlusion, cardioembolic stroke, or arterial dissection). Each may have a different genetic architecture, requiring different targeted therapy and clinical monitoring. Genomic medicine can also help to identify rare conditions that are unveiled by comprehensive diagnostics. For example, approximately 3% of patients with chronic obstructive pulmonary disease have alpha-1-antitrypsin deficiency [Marciniuk *et al.*, 2012]. Alpha-1-antitrypsin deficiency is most commonly caused by homozygous *SERPINA1*Z* allele and is associated with chronic obstructive pulmonary disease cirrhosis, and hepatocellular carcinoma [Silverman and Sandhaus, 2009]. In case of early detection specific clinical follow-up and treatment are recommended.

Ideally, diseases shall be prevented rather than treated. A successful example is phenylketonuria, with its severe, though preventable consequences (e.g., cognitive impairment) in case of early detection and adherence to a phenylalanine-free diet. Other well-known examples are familial hypercholesterolemia, which can be identified and treated to prevent various coronary events [Sturm *et al*., 2018], or *BRCA1-2* genetic tests in breast cancer [Wooster *et al*., 1995]. Some specific, high-impact target genes have been identified in studies on hereditary hyperlipidemia and subsequent risk of CHD, including *LDLR, PCSK9, APOB, LDLRAP1* and *ABCG8* genes [Miroshnikova *et al*., 2021]. For example, inactivating mutations in the PCSK9 gene can cause a decrease in LDL cholesterol and, consequently, reduce the CHD risk [Cohen *et al*., 2006]; using monoclonal antibodies to PCSK9 can dramatically reduce LDL levels and the risk of serious cardiovascular events [Robinson *et al*., 2015]. PCSK9 target therapies are now incorporated into clinical practice.

A key public health need is the introduction of genomic medicine and the identification of individuals at high risk for a particular disease to provide more effective screening or preventive therapy [Oteva *et al*., 1994; Khera *et al*., 2018]. Although most disease risk is polygenic, it has not yet been possible to use polygenic predictors to identify individuals at risk, as opposite to monogenic mutations.

Therefore, the development of polygenic risk scores is a critical challenge, considering the triple and over risk of coronary artery disease (CAD), atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer in 8.0%, 6.1%, 3.5%, 3.2%, and 1.5% of the population, respectively. CAD prevalence is 20 times higher than that of rare monogenic mutations associated with comparable risk [Khera *et al.*, 2018]. Genes have been identified for various common diseases exhibiting rare heterozygous variants and thus increasing the risk manyfold. An important case is risk variants for familial hypercholesterolemia in 0.4% of the population, which causes a 3-fold increase in the risk of coronary artery disease (CAD) [Abul-Husn *et al*., 2016]. Intensive treatment to reduce blood cholesterol levels among such carriers can significantly drop the risk. Another case is the p.E508K missense mutation in the *HNF1A* gene, with a carrier rate of 0.1% in the general world population and 0.7% in Hispanics [Lek *et al.*, 2016], thus increasing the T2D risk by 5-fold [Estrada *et al*., 2014]. Although identification of monogenic mutations can be important for carriers and their families, the vast majority of diseases occur in individuals who do not possess these mutations.

For most common diseases, polygenic inheritance, involving many common genetic variants of small effect, plays a greater role than rare monogenic mutations [Gibson *et al*., 2012]. However, it has been unclear whether it is possible to create a genome-wide polygenic score (GPS) to identify individuals at clinically significantly increased risk—for example, comparable to levels conferred by rare monogenic mutations [Khera *et al.*, 2018].

Previous studies to create GPS had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population). These initial efforts were hampered by three challenges: (i) the small size of initial genome-wide association studies (GWAS), which affected the precision of the estimated impact of individual variants on disease risk; (ii) limited computational methods for creating GPS; and (iii) lack of large datasets needed to validate and test GPS. Using much larger studies and

improved algorithms, Khera and colleagues set out to revisit the question of whether a GPS can identify subgroups of the population with risk approaching or exceeding that of a monogenic mutation. They studied five common diseases with major public health impact – CAD, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer. For each of the diseases, they created several candidate GPS based on summary statistics and imputation from recent large GWAS in participants of primarily European ancestry, using the UK Biobank genotype data. The predictors had AUC ranging from 0.79 – 0.81 in the validation set, with the best predictor (GPSCAD) involving 6,630,150 variants. This predictor performed equivalently well in the testing dataset, with AUC of 0.81 (Table 23).

Table 23. GPS derivation and testing for five common, complex diseases [Khera *et al.*, 2018].

Disease	Discovery GWAS, case/ control	Prevalence in validation dataset	Prevalence in testing dataset	No. of SNPs in GPS	Tuning parameter	AUC (95% CI) in validation dataset	AUC (95% CI) in testing dataset
CAD	60,801/ 123,504	3,963/120,280 (3.4%)	8,676/ 288,978 (3.0%)	6,630,150	LDPred (ρ = 0.001)	0.81 (0.80–0.81)	0.81 (0.81–0.81)
Atrial fibrillation	17,931/ 115,142	2,024/120,280 (1.7%)	4,576/288,978 (1.6%)	6,730,541	LDPred (ρ = 0.003)	0.77 (0.76–0.78)	0.77 (0.76–0.77)
T2D	26,676 / 132,532	2,785/120,280 (2.4%)	5,853/288,978 (2.0%)	6,917,436	LDPred (ρ = 0.01)	0.72 (0.72–0.73)	0.73 (0.72–0.73)
Inflammatory bowel disease	12,882 / 21,770	1,360/120,280 (1.1%)	3,102/288,978 (1.1%)	6,907,112	LDPred (ρ = 0.1)	0.63 (0.62–0.65)	0.63 (0.62–0.64)
Breast cancer	122,977/ 105,974	2,576/63,347 (4.1%)	6,586/1576895 (4.2%)	5,218	Pruning and thresholding (r/2 < 0.2; P < 5x10-4	0.68 (0.67–0.69)	0.69 (0.68–0.69)

GPSCAD has the advantage that it can be assessed from the time of birth, well before the discriminative capacity emerges for risk factors (for example, hypertension or type 2 diabetes) used in clinical practice to predict CAD. Moreover, even for our middle-aged study population, practicing clinicians could not identify the 8% of

individuals at ≥3-fold risk based on GPSCAD in the absence of genotype information [Khera *et al.,* 2018]. For example, conventional risk factors such as hypercholesterolemia was present in 20% of those with ≥3-fold risk based on GPSCAD versus 13% of those in the remainder of the distribution, hypertension in 32% versus 28%, and family history of heart disease in 44% versus 35%. Making high GPSCAD individuals aware of their inherited susceptibility may facilitate intensive prevention efforts. For example, we previously showed that a high polygenic risk for CAD may be offset by either of two interventions: adherence to a healthy lifestyle or cholesterol-lowering therapy with statin medications [Khera *et al.,* 2018]. Similar results were reported for four other conditions.

Type 2 diabetes is a key driver of cardiovascular and renal disease, with rapidly increasing global prevalence. The polygenic predictor identified 3.5% of the population at ≥3-fold risk and the top 1% had 3.30-fold risk [Khera *et al.,* 2018]. Both medications and an intensive lifestyle intervention have been proven to prevent progression to type 2 diabetes [Knowler *et al.,* 2002], but widespread implementation has been limited by side effects and cost, respectively. Ascertainment of those with high GPST2D may provide an opportunity to target such interventions with increased precision.

In many areas of medicine, there is at least one genomic test with strong effects that is part of the standard of care. The effect varies greatly depending on the disease. That said, researchers have made notable advances in oncology, cardiology, endocrinology, and prenatal/neonatal testing (obstetrics and pediatrics) [Baranov *et al.* 2021]. The results of such tests serve as a clinical guide in the sense that they allow identification/diagnosis of a more homogeneous subgroup within a larger group of diseases. Alternatively, they can identify a subset of patients with different therapeutic needs and suggest medications that may be effective and have no side effects. Some of these genomic tests have strong supporting evidence but are not yet the standard of care, often because the clinical implementation process has not yet been defined (including infrastructure, education for care). However, even when genomic medicine

is proven to be effective, cost-effectiveness, safety, tolerability, accessibility, and acceptability compared to existing medications will need to be evaluated for the relevant clinical question [Franks *et al*., 2021]. Moreover, since the vast majority of human genetics research has been conducted in people of European descent, future studies of other ethnic groups should be prioritized, especially where the discovery of rare variants is of interest [Barbitoff *et al*., 2021].

In countries where effective prevention or early detection strategies are available, key issues will include allocating attention and resources to individuals at different levels of genetic risk and integrating genetic risk stratification with other risk factors, including rare monogenic mutations, clinical and environmental factors. Where such strategies do not exist or are suboptimal, identification of high-risk individuals should facilitate the development of effective studies to identify early markers of disease onset and clinical trials to test prevention strategies. In both cases, it is important to recognize that the risk associated with high polygenicity may not reflect a single underlying mechanism, but rather the combined effect of multiple pathways [Fry *et al*., 2017]. Nevertheless, prevention and timely detection strategies can be useful regardless of the underlying mechanism-as in the case of statin therapy for CAD, anticoagulants to prevent stroke in patients with atrial fibrillation, or enhanced mammography screening for breast cancer. Risk communication will require serious consideration. Whereas polygenic risk scores can be simultaneously calculated at birth for all common diseases. That said, the usefulness of knowledge and potential harm from genetic tests for an individual may vary by disease and life stage [Khera *et al*., 2018]. Nevertheless, it may be worthwhile to withhold some information that can be easily computed from genetic data.

Ultimately, the fair deal issue is another challenge at stake. The polygenic risk scores described in the paper were derived and tested in individuals of primarily European ancestry, the group in which most genetic studies have been undertaken to date. Because allele frequencies, linkage disequilibrium patterns, and effect sizes of common polymorphisms vary with ancestry, the specific GPS will not have optimal

predictive power for other ethnic groups [Khera *et al.,* 2018]. It will be important for the biomedical community to ensure that all ethnic groups have access to genetic risk prediction of comparable quality, which will require undertaking or expanding GWAS in non-European ethnic groups.

Polygenic scores provide a quantitative metric of an individuals inherited risk based on the cumulative impact of many common polymorphisms. Risk prediction accuracy shall improve considerably with the advancement of exome and whole-genome data. There is no doubt that within a few years, the combination of genome-wide sequencing data and genome-wide search for allelic associations for all major MFDs is bound to dramatically increase the predictive value of presymptomatic hereditary predisposition testing, considering the rapid progress in molecular medicine [Baranov *et al.,* 2021]. There is an apparent demand to promote inexpensive genotyping into a standard of care in health care systems around the world [Lello *et al.,* 2019].

In conclusion, we would like to note that, depending on the etiological factors, all hereditary diseases are conventionally divided into monogenic and multifactorial. The severity of disease, time of manifestation, and clinical signs of hereditary diseases depend on the mutations that damage the gene, as well as modifier genes affecting the manifestation of pathological traits and environmental factors. Common monogenic diseases account for only some 1% of all human pathologies. Familial monogenic and polygenic diseases are quite rare; most critical human diseases (cardiovascular, cancer, mental, neurodegenerative, etc.) belong to MFDs, where the etiology depends on a combination of hereditary predisposition and the external environmental impacts. Molecular gene testing is the key fundamental method of preventive medicine. Predictive medicine shall rely on investigations of allelic polymorphism patterns in prevalent MFD-associated genes.

Genetic variant as a cutting-edge concept changes our custom understanding of diseases splitting into monogenic, oligogenic, and multifactorial. There is an urgent

need for an updated genetic classification of diseases based on clinical and laboratory data, as well as molecular genetic analysis of penetrance and expression. Most optimal statistical approaches to assess genetic predisposition are required to develop a framework of risk factors for polygenic MFDs.

CHAPTER III. Next Generation Sequencing and Human Infectious Diseases. Genetic Risk Factors for COVID-19 Infection

3.1. General Information on SARS-CoV-2 and Its Genomic Variability

In previous chapters of our study, we demonstrated how significantly the NGS has contributed to our knowledge on monogenic and multifactorial human diseases within the framework of the human genetic health passport concept.

The COVID-19 pandemic outbreak in late 2019 caused by the new strain of coronavirus was a major stressor for human health and the society both in Russia and globally [https://coronavirus.jhu.edu; Glotov *et al.*, 2021a]. By August 2022, the COVID-19 pandemic swept over 230 countries with over 589 million cases and over 6.4 million deaths worldwide [https://coronavirus.jhu.edu].

Amid the pandemic it is extremely imoirtant to identifying specific single nucleotide polymorphisms (SNPs) and new protein and gene targets that may be highly sensitive diagnostic and prognostic markers of the severity and outcome of the disease for combating this pandemic. Identification of individual genetic predisposition allows personalizing programs of medical rehabilitation and therapy. Investigation of SARS-CoV-2 pathogenetic mechanisms must consider genetic aspects at the level of virus and human genome.

It is well known that coronaviruses (CoV) are agents that cause acute severe respiratory syndrome (SARS-CoV), which was responsible for a global outbreak in 2002, when 8000 people were infected, with 10% of fatalities [Freund *et al.*, 2015]. More recently, in 2012 an outbreak of Middle East respiratory syndrome (MERS-CoV) occurred spreading to 26 countries [Mackay, Arden, 2015]. Unfortunately, epidemiology and genetics of these viruses are poorly understood. When a new form of coronavirus (SARS-CoV-2) was reported in China in December 2019, causing the global COVID-19 pandemic in 2020 [Liu *et al*., 2020], numerous blank spaces remained in our understanding of the underlying mechanisms of the disease, its risk factors, prevention, etc.

Coronaviruses are large spherical particles (with a diameter of 120 nm) consisting of a bilayer lipid envelope containing four proteins: membrane proteins (M, E), spike protein (S), and hemagglutinin esterase (HE) around nucleocapsid (N) formed by multiple copies of this protein associated with single-stranded RNA [Cascella *et al.*, 2020] (Fig.25).

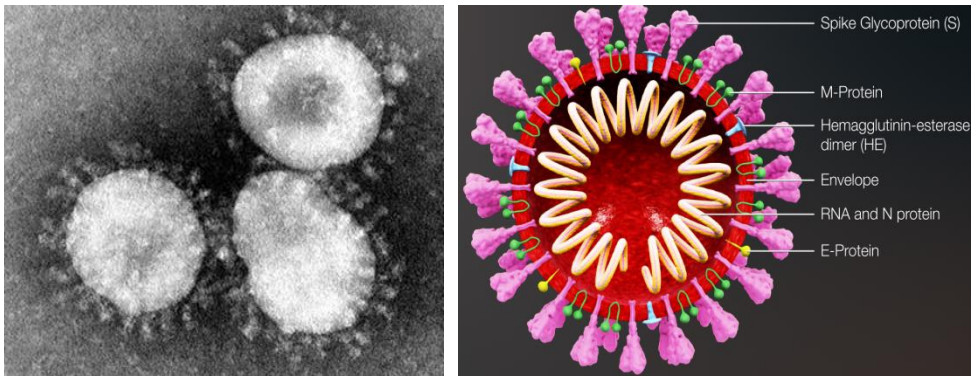


Figure 25. SARS-CoV-2 microimage (A) and scematic structure (B). Spike glycoprotein (S); M protein – membrane protein; Hemagglutinin-esterase dimer (HE); Envelope; RNA and N protein; E protein [Cascella *et al.*, 2020].

S proteins form outgrowths on the envelope of the virus, creating a kind of crown, owing to which the virus got its name [Ashour *et al.*, 2019]. Using these spikes, the viruses attach to receptor proteins of the host cells, which ensure fusion of the viral and cell membranes and the entry of the viral RNA into the cell. S proteins contain the receptor-binding domain (RBD, amino acids N318-T509), which interacts with the receptor-binding motif (RBM, amino acids S432-T486) of angiotensin converting enzyme 2 (ACE2), a cell receptor for SARS-CoV-2 [Li *et al.*, 2005; Chen *et al.*, 2020]. In addition, S glycoprotein contains a furin-like restriction site [Coutard *et al.*, 2020], which is required for recognition during pyrolysis and, therefore, contributes to zoonotic infection of the virus. The SARS-CoV-2 genome is represented by single-stranded RNA about 30 kb containing a cap region at the 5'-end and a poly-A sequence at the 3'-end, which allows viral RNA to be translated on the ribosomes of the host cells. Viral RNA includes regulatory sequences, in which transcription is terminated, and ten open reading frames (ORFs), which are transcribed to form mRNA (Fig. 26).

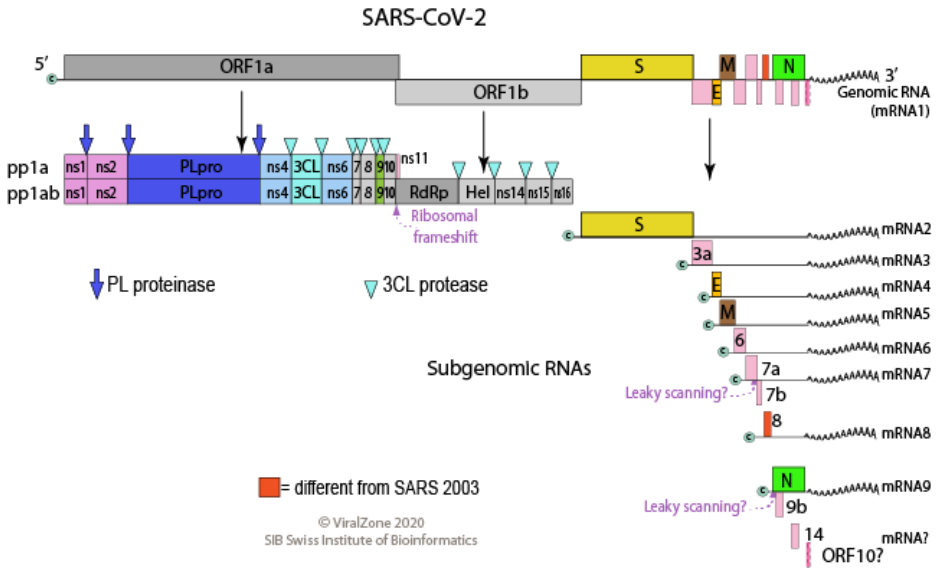


Figure 26. The structure of RNA genes of SARS-CoV-2 coronavirus [Schidtke, 2020].

The severity of COVID-19 infection depends on the polymorphism of the virus genome and some alleles of the human genome. Mutations in viral RNA polymerase (ORF8, 241C>T, S84L, 14408C>T, C251T), RNA primase (P323L), the S protein (23403A>G, D614G), and the structural proteins (N, E) increase the immunogenicity of proteins in relation to the immune T cell response, which may be associated with the increased transmissibility and severity of COVID-19 infection in European countries [Yin *et al.*, 2020]. Presence or absence of two comutations (8782C>T) in ORFL1ab and 28144 (28144T>C) in ORF8 of the SARS-CoV-2 genome allowed to isolate the two major genetic types of the virus – L and S, that differ in terms of virulence, replication rate, and disease severity [Koyama *et al.*, 2020]. Mutations in the viral proteins can change the affinity and specificity for the binding of targeted drugs to them, being the molecular basis of individual differences in the response of the human body to antiviral drugs and/or vaccines.

Analysis of SNPs in samples from the endemic region revealed various sybtypes of coronavirus. SNPs analysis from the endemic Changchuan region, Yin and colleagues identified four SARS-CoV-2 genotypes: genotype I (11083G>T), genotype II (26144G>T), genotype III (8782C>T, 28144T>C), genotype IV (241C>T, 3037C>T, 14408C>T, 23403A>G). Later in October 2020, S-protein mutations were

detected in patient samples from Rio de Janeiro State, Brazil, with the E484K mutation showing utmost clinical significance, hence the name 484K.V2 [Toovey *et al*., 2021]. In December 2020, coronavirus variants B.1.1.7 (alpha) and B.1.351 (501Y.V2, beta) from the UK and South Africa were reported, exhibiting nucleotide replacements of 69-70del, Y144del and N501Y, K417N, E484K respectively in the S protein [World Health Organization. SARS-CoV-2 Variants https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/]. Sequence analysis of 405871 samples from the GISAID database in late 2020 identified B.1.427 and B.1.429 coronavirus variants from Southern and Northern California [Zhang W. *et al*., 2021]. In January 2021, P.1 and P.2 (sublineage B.1.128, gamma) variants were detected including 10 unique mutations (E484K and N501K, etc.) in the S-protein in 42% in samples from Manaus, Brazilian Amazonas state [Sabino *et al*., 2021]. A delta variant (B 1.617.2) bearing G478K and 417N mutations in the S-protein appeared in October 2020 in India. In early 2021, a variant named B.1.526, bearing E484K and S477N mutations in the S-protein, was reported in New York [Bernal *et al*., 2021].

Shortly later an even newer variant of SARS-CoV-2 was discovered, for which the WHO coined the name Mu. It was first detected in Colombia in January 2021 and by mid-year it was already detected in 39 countries [https://www.weforum.org/agenda]. In September 2021, a mutant delta variant (AY.4.2 delta plus) was discovered in the UK, when during the week (September 27 to October 3) 40-50,000 new infections were reported each day. It is believed that delta plus is 10-15% more infectious than delta variant [https://www.weforum.org/agenda]. In December 2020, the spread of B.1.1.31 and B.1.1.317 SARS-CoV-2 lines of Russian origin was reported in the UK, USA, Japan, Singapore, Turkey, Thailand, Switzerland and Brazil [PANGO lineages. https://cov-lineages.org/lineage_designation.html].

3.2. SARS-CoV-2 Genetic Variates and Associated Risk of Coronavirus Infection Severity and Outcomes

The identification of new SARS-CoV-2 and human protein and gene targets, which may be markers of the severity and outcome of the disease, are extremely important during the COVID-19 pandemic. To this end we studiediral RNA samples isolated from 56 patients with COVID-19 infection who received treatment at the City Hospital No. 40 of St. Petersburg from 04/18/2020 to 01/31/2021 who reported PCR positive for SARS-CoV-2 RNA [Glotov *et al.*, 2021b]. Single nucleotide polymorphisms (SNPs) in viral RNA were identified through the creation of cDNA libraries by targeted sequencing (MiSeq Illumina). We identified 389 single nucleotide polymorphisms (SNPs) in viral RNA, with 263 (67.6%) amino acid substitutions in protein sequences, 126 synonymous mutations (32.4%), with 139 (35.7%) in fatal patients. Disease severy correlation with nonsynonymous, silent, and S-protein mutations is reported in Table 24.

The number of S-protein mutations in stage 3 disease severity patients was remarkably higher ($p<0.001$) than in stage 1 disease severity patients [Glotov et al., 2021b]. A similar study was earlier conducted by Japanese investigators to find out that the number of nonsynonymous mutations (especially Pro108Ser in 3-chymotrypsin-like protease (3CLpro) and Pro151Leu in the nucleocapsid N-protein) inversely correlated (OR=0.24, 95%CI=0.07-0.88, P=0.032) with COVID-19 severity and oxygen therapy demands [Abe *et al.*, 2021].

Potential severity of the detected mutations was identified, i.e. association with disease severity was assessed by based on frequency and OR values in discharged and fatal patients using cluster analysis. Comparative analysis of frequencies for each SNP in sample groups of patients at fatality risk allowed to split SNPs into 3 clusters (groups). The first group included mutations potentially associated with coronavirus infection severity and fatal outcome (OR>1). Such SNPs were localized in the genes encoding the leader 5'-end cap region and Nsp1, Nsp2, Nsp3, Nsp4 proteins, RNA-

dependent RNA polymerase (Nsp12), helicase (Nsp13), S -protein, Orf3a and in the 3'-end poly-A sequence. The second cluster included neutral mutations (OR=1) localized in the genes encoding Nsp3 and N-proteins. The third cluster included potentially protective mutations (OR<1) localized in genes encoding Nsp2, Nsp3, Nsp7, endonuclease (Nsp14), endoRNAase (Nsp15), S- and N-proteins [Glotov *et al.*, 2021b].

Table 24. Disease severy correlation with mutations in the SARS-CoV-2 genome [Glotov *et al.*, 2021b].

COVID-19 Disease Severity	Average no. of mutations	Average no. of nonsynonymous mutations	Average no. of silent mutations	Average no. of mutations in S-protein
1 (n=4)	11.0±0.2	7.0±1.4	4.0±1.4	1.0
2 (n=14)	11.4±1.9	8.3±1.9	3.3±1.6	2.0±1.0
3 (n=13)	13.0±3.6	9.0±2.4	4.0±2.1	2.2±1.0***
Dead, 3 (n=25)	17.1±5.0	10.6±3.0	6.6±2.2	1.6±0.9

Note: *** statistical significance at $p < 0.001$. Mild, moderate, and high severity degress are designated by I, II, and III, respectively. All patients who died had degree III severity of COVID-19 infection.

Then epidemiological characteristics of the virus were assessed based on SARS-CoV-2 genome analysis in patient cohort. We found that patients from St. Petersburg were infected with 14 SARS-CoV-2 lines. Out of 47 samples, 14 (29.8%) were infected with 5 lines of the virus of Russian origin. Overall, the OR for the 5 virus lines of Russian origin was 0.441 (95%CI=0.116-1.684). This indicates that mutations in SARS-CoV-2 lines of non-Russian origin are associated with an increased risk of lethal outcome (OR=2,267, 95% CI=0.1594-8.653) in the examined group of Russian patients [Glotov *et al.*, 2021b].

It is noteworthy that 75.8% of patient samples were infected with genotype IV SARS-CoV-2virus (241C>T, Nsp3: 3037C>T; Nsp12, RNA polymerase: 14408C>T, Pro323Leu; S-protein: 23403A>G, Asp614Gly), identified by Zhang and colleagues [Zhang W *et al.*, 2021]. Haplotypes 3037C>T, 14408C>T, 23403A>G and SNP

241C>T were the most frequent in samples from Europe, indicating European origin of the virus in the patients. In a study conducted by an international team of investigators from Vietnam, Australia, the UK and the US, this mutant haplotype was found in 90% (40 of 44) of patient samples from the UK (n=15), Russia (n=6), Germany (n=5), France (n=4), Italy (n=2), Spain (n=2), the Netherlands (n=1), Vietnam (n=6) and Asian countries (n=3). Together with this group of mutations, SNPs in N-protein 28881G>A, 28882G>A, and 28883G>C were also observed in these patients in 75% of genomes (33 of 44) [Wang et al., 2021]. These mutations are localized in a small region (194-204) of the amino acid sequence of this protein. Since the N-protein is involved in packaging viral RNA into the spiral ribonucleocapsid, 28881G>A, 28882G>A, and 28883G>C mutations can increase the efficiency of subgenomic RNA transcription, contributing to cell survival, replication, and viral persistence [Caccuri *et al.*, 2020]. In our samples, SNPs 28881G>A, 28882G>A, and 28883G>C were present in 100% of the samples, likely indicating the neutral nature of these variants and the virus originating from the B1.1 lineage 20B clade [Abe *et al*., 2021]. SNP 241C>T, localized in the 5'-untranslated region (5'UTR) of the viral RNA, was also observed in 47 of our 50 samples (94%). This substitution has one of the highest frequencies, 0.758 (9673/12754) and 0.809 (36786/45494), respectively, in the United States and worldwide, affecting the regulation of transcription and gene expression of the virus as well as its replication [Wang *et al*., 2021]. SNP 14408C>T (Pro323Leu) and mutation 13554C>T, associated with severe course and lethality, were also found in virus samples from our patients in RNA-dependent RNA polymerase (RdRp or Nsp12 protein) with frequencies of 0.940 (47 of 50) and 0.06 (3 of 50). The first of these mutations is also the most common in samples from the United States (AF 0.464, 5918/12754) [Wang *et al*., 2021]. The biological role of this mutation is still under discussion. Wang and colleagues believe that since both amino acids, proline (Pro) and leucine (Leu), are non-polar and aliphatic, Pro323Leu may therefore have no effect on the function of the Nsp12 protein [Wang *et al.*, 2021]. SNP 23403A>G (Asp614Gly) in the S-protein gene is found in 94% of our patient samples

(47/50) and is associated with a substitution of asparagic acid at position 614 for glycine and is associated with an increased risk of severe or fatal patient outcomes. As mentioned above, this SNP is often combined with three other SNPs (241C>T, 3037C>T, and 14408C>T) and is inherited as a haplotype [Korber *et al*., 2020]. Researchers from Harvard University, University of Washington (USA), and University of Sheffield (UK), studying clinical data and SARS-CoV-2 genetic sequences in 999 patients with COVID-19, found that the Arg614Gly variant correlated with high levels of viral RNA in the upper airways and transmission rates in patients compared to the wild type. This indicates a higher Gly614 infection rate. However, Korber and colleagues could not establish an association between D614G and duration of hospitalization, i.e., disease severity [Korber *et al.*, 2020].

We also established correlations between the total number of nonsynonymous mutations (SNPs) in the S-protein of SARS-CoV-2 lines with dyspnea, risk of death, ferritin, D-dimer and blood glucose levels [Glotov *et al.,* 2021b].

3.3. Markers of Severe Clinical Course of COVID-19

It is common knowledge, that when infected with SARS-CoV-2, the patient displays symptoms of an acute respiratory viral disease after 2–14 days of the incubation period: increased body temperature (90%); cough (in 80% of cases), shortness of breath (in 30% of cases), fatigue (in 40% of cases), chest congestion (in 20%), sore throat, runny nose, decreased sense of smell and taste, conjunctivitis. These symptoms may indicate the development of pneumonia without respiratory failure, acute respiratory distress syndrome (ARDS, pneumonia with acute respiratory failure), sepsis, septic shock, or multiple organ dysfunction syndrome. These symptoms were typical for alpha- and delta- strains, while o-micron strain is associated with somewhat different manifestations [Interim guidelines: prevention, diagnosis and treatment of new coronavirus infection (COVID-19). Version 14, 12/27/2021].

Septic shock and multiple organ dysfunction syndrome These pathological conditions most often lead to the mortality in patients of working age (59.7 ± 13.3

years) with chronic diseases: arterial hypertension (23.7–30%), diabetes mellitus (16.2%), metabolic syndrome, coronary heart disease (5.8%), chronic obstructive pulmonary disease (COPD), nicotine addiction, inflammatory bowel diseases, and cancers [Yang X. *et al.*, 2020; Fang *et al.*, 2020; Shitao *et al.*, 2020]. In addition, patients with genetic diseases may be a risk group for COVID-19 infection. For example, in a study conducted by scientists from the University Medical Center Utrecht (Netherlands), 180 (45.6%) of 395 patients with Down syndrome developed severe respiratory syncytial virus infection [Beatrijs *et al.*, 2007].

The clinical picture in patients of the risk groups is characterized by the development of mutual burdening syndrome, accompanied by progressive respiratory and heart failure, which ultimately worsens their condition and leads to labor losses, early disability, and high mortality. In this regard, patients infected with COVID-19 with chronic and genetic diseases especially urgently need immediate diagnosis and rehabilitation.

The central pathophysiological problem of COVID-19 is immune dysfunction with a pronounced uncontrolled generalized systemic inflammatory response in the form of increased production of inflammatory cytokines. The so-called cytokine storm (CS) can manifest in two clinical forms: secondary hemophagocytic lymphohistiocytosis (HLH) and macrophage activation syndrome (CAM) [Costela-Ruiz *et al*., 2020, Blanco-Melo *et al*., 2020], syndromes previously described in some infections (Epstein-Barr virus, influenza) and systemic autoimmune diseases (systemic lupus erythematosus, Still's disease), as well as during treatment with cytostatic and immunosuppressive drugs, after allogeneic organ and tissue transplantation, in 3-4% of sepsis. CS thus has some similarities with secondary hemophagocytic syndrome (HPS) manifested as fever, cytopenia, hyperferritinemia, abnormal liver values, coagulopathy and lung damage (including ARDS) [Caso *et al*., 2020]. In all of these conditions, the cytokines IL-1β, IL-18, IFN-γ, and IL-6 are major mediators of the excessive inflammatory response of the immune system.

The results of recent studies show that CS associated with COVID-19 is a unique form of such hyperinflammatory response that requires further in-depth clinical and laboratory evaluation as well as development of its diagnostic criteria [Zachariah *et al*., 2020]. It has been suggested that CAM classification criteria (MAS-2016) are not applicable to patients with COVID-19, and the criteria of HLH assessment scales (HLH-2004 and HScore) need adaptation [Leverenz, Tarrant, 2020, McGonagle *et al*., 2020]. The relevance of developing a prognostic model of CS in patients diagnosed with COVID-19, determination of routine and additional markers to assess the risk of its development has been clearly outlined [Caricchio *et al.,* 2020; Moore, June, 2020].

A number of laboratory biomarkers are known whose levels change pathologically in CSF, irrespective of the triggering factor. The threshold values and combinations of the main biomarkers can form the basis for differential diagnosis of the conditions and the severity of the disease course. For example, sudden and rapidly progressive clinical deterioration in the late stages of COVID-19 (days 7-10) correlates with increased levels of acute phase indicators (CRP, ferritin) [Huang *et al*., 2020; Grasselli *et al*., 2020] as well as CS clinical and laboratory indicators [Chen *et al*., 2020; McGonagle *et al*., 2020; Wiersinga *et al*., 2020]. Ferritin concentrations >500 μg/L were found in 55.9% of patients with COVID-19 nonsevere forms and 81.7% of patients with severe forms ($p<0.0001$) [Li et al., 2020]. Hyperferritinemia is observed in macrophage activation syndrome and ARDS [Giamarellos-Bourboulis *et al*., 2020] and can identify patients at high risk for severe COVID-19-associated pneumonia [McGonagle *et al.,* 2020; Kivela, 2020].

Thus, elevated levels of C-reactive protein (CRP), interleukin-6, and D-dimer in blood are key for the diagnosis of systemic hyperinflammation, cytokine storm, and severe COVID-19-associated pneumonia [McGonagle *et al*., 2020]. These indicators, along with lymphopenia, are key during patient admission to the hospital for timely decision-making to identify patients with poor prognosis [Zhou *et al.,* 2020; Li *et al.,* 2020].

To identify key CS markers, patients were split in two groups comparable by age. one group included 100 patients with moderate disease symptoms; the other group included 358 patients with progressive moderately severe, severe, and extremely severe disease (see Table 24) [Shcherbak *et al.*, 2021].

Table 24. Characteristics of the severity of COVID-19 in two groups of patients in St. Petersburg [Shcherbak *et al.*, 2021].

Parameter	Group 1		Group 2		Total	p-value
	n	%	n	%		
Females	58	58.00	159	44.41	217	p=0.016
Maes	42	42.00	199	55.59	241	
total:	100	21.83	358	78.17	458	
Severity of disease:						
Mild	0	0.00	0	0.00	0	p<0.001
Moderate	100	100.00	153	42.74	253	
Severe and extremely severe	0	0.00	205	57.26	205	
Total:	100	21.83	358	78.17	458	
Lung involvement evaluation at admission based on a 4-point CT score ranking:						
CT-1	57	57.00	82	22.91	139	
CT-2	43	43.00	223	62.29	263	p<0.001
CT-3	0	0.00	44	12.29	47	
CT-4	0	0.00	9	2.51	9	
Total:	100	21.83	358	78.17	458	
Outcomes:						
Survivals	100	100.00	255	71.23	355	p<0.001
Deaths	0	0.00	103	28.77	103	
Total	100	21.83	358	78.17	458	

There is a reliable difference in the NEWS score: group 1 had the average admission NEWS score of 2 and the average hospital stay of 11 days; group 2 had the average NEWS score of 4 at admission which exacerbated to 5 at the start of treatment, including anticytokine drugs, anti-COVID-19 convalescent plasma, and hemoadsorbtion; the average hospital stay in group 2 was 12 days (see Fig.27). The death rate due to comorbidities was highest in group 2 patients with severe and extremely severe course of disease (28.8% in group 2 and 22.5% in the total cohort). At baseline, such patients had an unfavorable forecast due to age, comorbidity, clinical

severity of acute respiratory failure, a high NEWS score, dynamic evaluation of extensive lung involvement, and exacerbation of lung injury based on chest CT. Classification Trees were utilized to identify cutoffs for CS-associated risk factor.

This was followed by a comprehensive evaluation of CS-associated risk based on the ranking of variables obtained at admission. These variables were ranked by prognostic relevance in accordance with Classification Trees by using CART-style method for Split selection and included the following:

1) NEWS score dynamics,

2) serum IL-6 greater than 23 pg/ml,

3) serum CRP 50 mg/l and greater,

4) absolute lymphocyte count less than 0.72 × 109/L,

5) positive test for replicative SARS-CoV-2 RNA,

6) age 40 years and over.

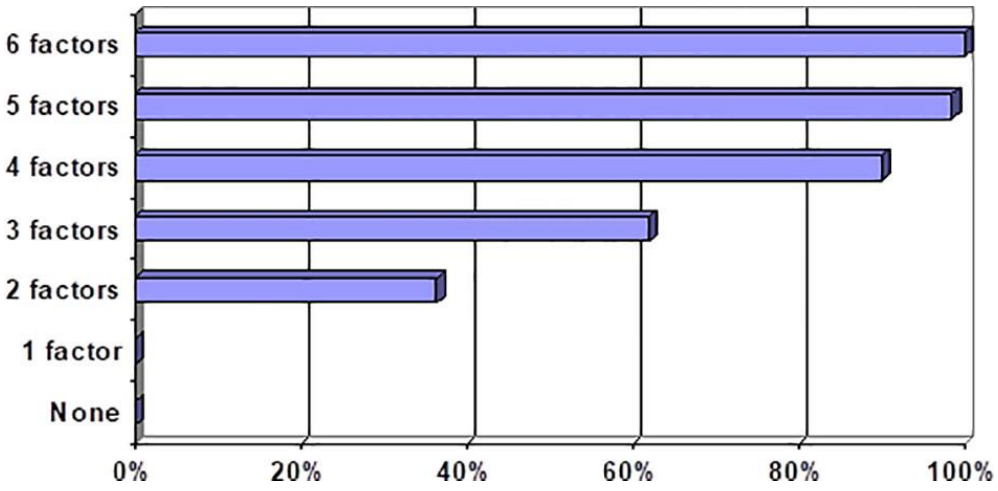


Figure 27. CS incidence rate depending on the number of risk factors in two groups of patients from St. Petersburg [Shcherbak *et al.*, 2021].

These biomarkers can be used as predictive criteria to determine the CS risk. We should note that there are no meaningful gender-related differences in the comprehensive examination of CS-related risk. We outlined the following risk categories of patients to enhance clinical relevance of our prognostic model:

- category 1 (0 to 1 risk factors) = CS-associated risk is near zero,

- category 2 (two to three risk factors) = CS-associated risk escalates dramatically to 55%, augmenting 35.5-fold vs. category 1,
- category 3 (four and more risk factors) = CS-associated risk exacerbates to 96%, augmenting 718-fold vs. category 1.

The obtained results are consistent with findings by other authors on the evaluation of risk factors for COVID-19-associated CS [Caricchio *et al.*, 2021; Moore el al., 2020]. Our results provide a justification for a treatment strategy with an early onset of proactive anti-inflammatory therapy and anti-COVID-19 convalescent plasma in patients at a high risk for CS development. The basic risk factors for cytokine storms in COVID-19 patients include male gender, lactate dehydrogenase level, age over 40 years, positive test result for replicative SARS-CoV-2 RNA, absolute lymphocyte count, D-dimer and ferritin levels, dynamics in the NEWS score, and plasma IL-6 concentration (Fig.28). Absolute lymphocyte count, LDH, CRP, ferritin, D-dimer, and IL-6 levels are the most critical lab parameters for diagnosis and dynamic monitoring of cytokine storms.

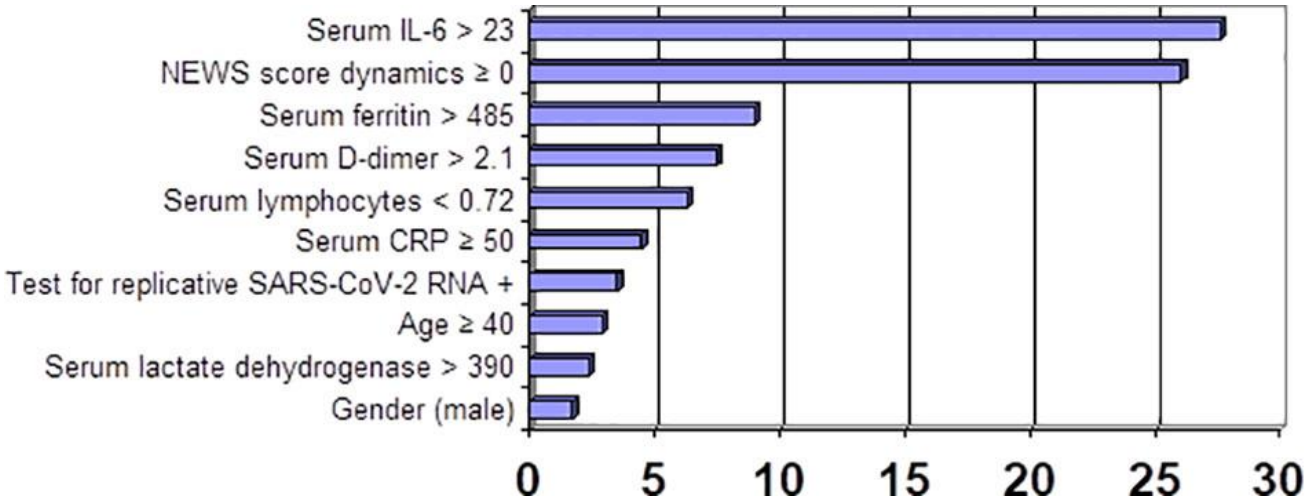


Figure 28. Incremental CS-associated risk (OR) and its correlation with unfavorable lab test values [Shcherbak *et al.*, 2021].

3.4. Identification of Genetic Variants Predisposing to Severe Course of COVID-19: ACE-2 and its Role

These lab findings can be used to accurately split patients into risk groups for further genetict testing [Glotov *et al.,* 2021a; Shcherbak *et al.,* 2022].

Any attempt to investigate genetic susceptibility to viral infections shall start with looking into receptor binding mechanisms allowing the virus to enter the host cell. It was found that angiotensin-converting enzyme-2 (*ACE2*) is the cellular receptor for SARS-CoV-2. S-proteins contain the receptor-binding domain (RBD), which interacts with the receptor-binding motif (RBM) of *ACE2* and thus enters the cell [Yan R. *et al.*, 2020]. In the human body, this protein is encoded by the *ACE2* gene. *ACE2* polymorphisms can affect the affinity and specificity of the S protein binding to *ACE2*, and, therefore, determine the hereditary predisposition to the risk of infection and lethality of SARS-CoV-2, which are associated with the development of arterial hypertension, diabetes mellitus, and cerebral stroke, which constitute a high-risk group of infection and lethality of COVID-19 infection. A person's susceptibility to SARS-CoV-2 may be a result of the combined effect of the therapy and characteristics of the *ACE2* gene polymorphism [Fang L. *et al.*, 2020]. Correlations between the most common (38) *ACE2* variants and the level of protein expression in various human tissues were established [Chen J. *et al.,* 2020]. It was shown that age, gender, race, and smoking significantly ($P = 0.008$) affect expression of the *ACE2* gene [Chen J. *et al.*, 2020]. It is thus evidents that mutations in the *ACE2* gene can disrupt receptor-ligand interaction. Mutations in the *ACE2* gene are resistant to S-protein binding and can be absent in a variety of populations. It is suggested, however, that rare missense-mutations may exacerbate disease severity [Cao *et al.*, 2020]. If missense variant affects *ACE2* receptor function, a higher frequency of missense allele carriers in some populations may result in lower infection rates and mortality; a lower frequency of carriers of missense variants, in turn, may increase the susceptibility and severity of the course of COVD-19. A recent work by Italian scientist [Benetti *et al.,* 2020]

detected apparently protective missense-variants in the *ACE2* gene. Although common missense variants in the *ACE2* gene may not affect the virus-host interaction, they can have indirect effects on COVID-19 susceptibility and disease outcome, for example, by increasing the level of oxidative stress, thereby worsening the disease outcome [Devaux *et al.,* 2020].

We demonstrate that the European populations slightly differ in alternative allele frequencies at the 2,754 variant sites in *ACE2*. A lower ratio of missense variants in the Southern European population compared to other European regions can partly explain higher mortality rates from COVID-19 in Spain and Italy [Shikov *et al.,* 2020]. In our study we implemented primary component analysis (PCA) to elaborate on the spectrum of variants in *ACE2* in different European populations. It is important to note that all subpopulations were disbalanced in sample size. Given this observation, we narrowed the data set by selecting variants with non-zero allele frequencies in each population. The resulting PCA plot for 60 variants showed some differences between populations; at the same time, few differences were observed between European populations. We assumed that the differences may be even more pronounced when considering variants of functional significance. To test this hypothesis, we performed a similar analysis using 229 missense variants. Importantly, all populations had fairly high coverage (number of alleles) at these sites. PCA analysis showed that Southern European populations, as well as those in Estonia and Bulgaria, tended to be separated from other European populations, and differences between other European populations were quite pronounced [Shikov *et al.*, 2020].

We next considered the five coding variants (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) in the *ACE2* gene for Russian and European populations (see Table 25). Remarkably, Russians falls in close proximity to other Europeans based on allele frequencies. This finding is crucial to understandthe epidemiologic setting at the start of the pandemic in March-April 2020.

Table 25. Variants detected in Russian exomes [Shikov *et al.*, 2020].

Position	rsID	Ref	Alt	Effect	Pathogeny prediction	Protein	Homozygotes	Heterozygotes	AF
15582209	rs35803318	C	T	SYNON	Pathogenic	p.Val749Val	10	13	0.031
15582298	rs41303171	T	C	MISS	Pathogenic	p.Asn720Asp	2	15	0.016
15596143	rs113691336	C	CATAAG	INTRN	-	-	232	82	0.609
15606024	-	T	TTC	INTRN	-	-	1	1084	0.001
15606028	-	A	ATTGT	INTRN	-	-	1	1084	0.001
15606029	-	A	ATTACTTT	INTRN			1	1084	0.001
15607650	rs971249	T	C	INTRN	Benign	-	278	142	0.671
15610348	rs2285666	C	T	SPLIR	Benign	-	66	89	0.205

Although the effect of variants in the *ACE2* gene on the expression of the corresponding protein in the lungs has not been found, the effect of cis-eQTLs on *ACE2* function in various brain tissues may be associated with neurological complications in patients with COVID19 [Strafella *et al*., 2020]. We assumed that even slightly elevated *ACE2* expression may lead to an increase in the number of receptor molecules on the cell surface, which, in turn, may increase susceptibility to COVID-19. A statistical comparison of the frequencies of individual variants between patients with mild and severe COVID-19 revealed no significant differences. Therefore, we concluded that while certain differences in *ACE2* eQTL frequencies across populations exist (Figure 1A), these differences have either no or very little effect on the COVID-19 susceptibility and severity [Shikov *et al.*, 2020].

The same mechanism might hold true for the rare *ACE2* haplotypes, which were found to be overrepresented in Russian patients with the severe form of COVID-19 (Fig. 29). However, it is also likely that the effect of these variants on the phenotype is also not related to the receptor function of *ACE2* [Shikov *et al.*, 2020].

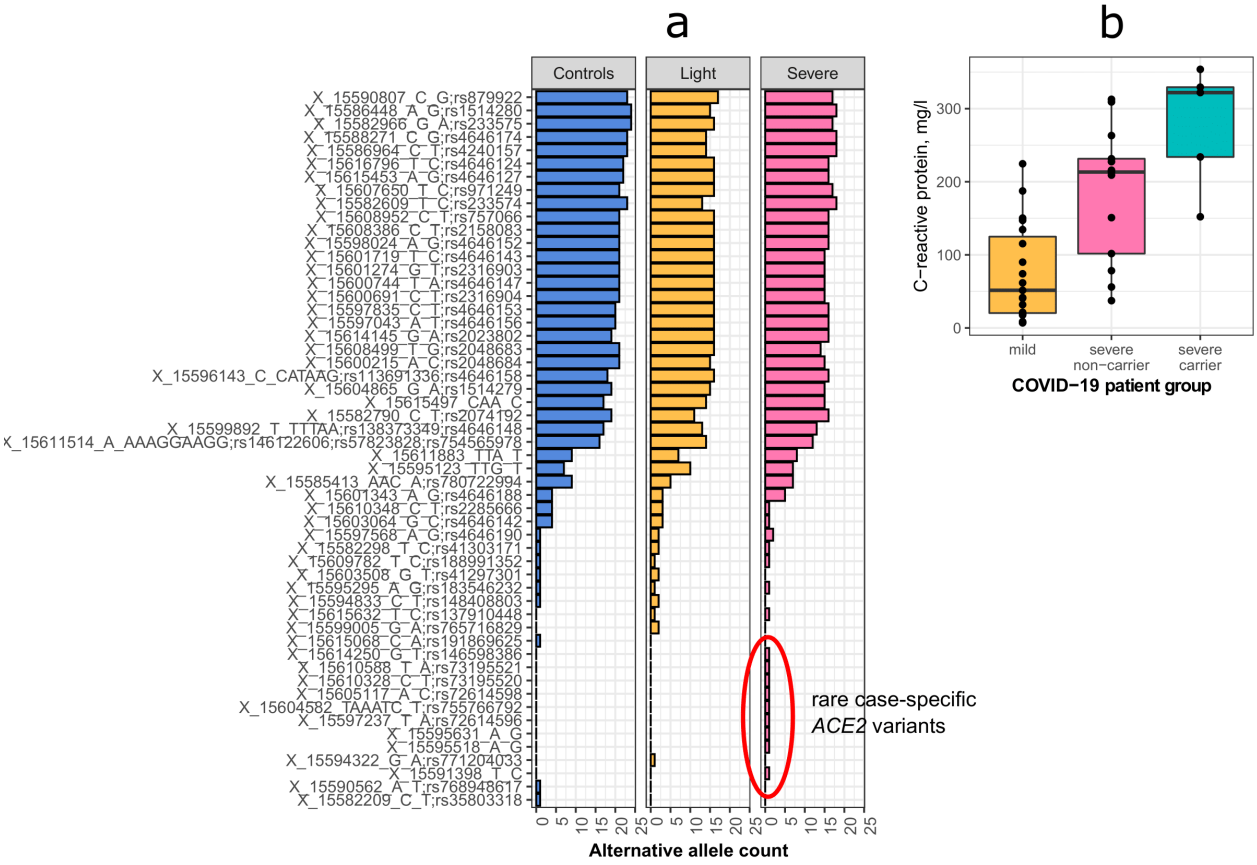


Рисунок 29. Rare and common variants of the *ACE2* gene in COVID-29 patients [Shikov *et al.*, 2020]. (a). Alternative allele counts for 54 single-nucleotide polymorphic substitutions (SNPs) and indels among patients with mild and severe forms of the COVID-19 infection. (b) Mean concentration of C-reactive protein in the blood of COVID-19 patients with either mild or severe type of disease.

3.5. Polygenic analysis of Predisposition to COVID-19

Notably, polymorphism in the ACE2 gene is by far not the only factor to influence the severity of the disease. The peptidases ANPEP, DPP4 and ENPEP, TMPRSS2 can also serve as *ACE2* co-receptors in SARS-CoV2 infection [Qi *et al.*, 2020; Rossi *et al.*, 2021]. Studies show that the TMPRSS2 enzyme is required for viral S-protein activation to facilitate the virus entry into the cell as a result of interaction [Shulla *et al.*, 2011].

In addition, a set of antigens in the main histocompatibility system, patient's blood group, intensity cytokine production and other inflammatory and antiviral

immunity factors, in particular the secretion of interleukins and chemokines, are important for viral infection manifestation.

As early as the beginning of the epidemic, the association of COVID-19 infection severity with loci (*HLA-B*4601*), *FcγRIIA, MBL, TMPRSS2, TNF-α, IL-6,* human blood group A antigen and other genes was established [Asselta *et al.,* 2020; Feldmann *et al.,* 2020; Wu B. *et al.,* 2020; Anisenkova *et al.,* 2021].

The epidemic stimulated genetic research; increasing evidence was produced to predict genetic susceptibility to COVID-19 and help clinicians choose the right treatment [Prakrithi *et al.,* 2021]. Recent studies reported several dozens of links between genetic variants and COVID-19 incidence, severity, and mortality among different ethnic groups [Suh *et al.,* 2022]. For example, one genomic association study (GWAS) in the United Arab Emirates investigated a sample of 600 participants to identify 8 susceptibility loci for severe COVID-19. Loci in these genes were found to be associated with T-cell-mediated inflammation affecting the production of inflammatory cytokines [Mousa *et al*., 2021]. Another study from Europe (seven hospitals from Italy and Spain) identified two cross-replicating associations of severe COVID-19 with variants at loci 3p21.31 and 9q34.2 (GWAS group of severe COVID-19). These loci encompass several genes including *SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6, CCR1 for 3p21.31* and *ABO* for 9q34.2. The association of these loci with COVID-19 was also found by Wu and colleagues in a cohort study of Chinese patients [Wu *et al.,* 2021].

Despite the large number of reported associations, variants that were found in one study may not be confirmed in another [Suh *et al.,* 2022]. In order to identify loci that demonstrate association between different cohorts and ethnic groups, large-scale meta-analyses are conducted. A case in point is the COVID-19 HG project (COVID-19 Host Genetics Initiative, 2021) representing the bulk of current data on genetic predisposition to COVID-19 [COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature 600, 472-477 (2021). https://doi.org/10.1038/s41586-021-03767-x].

The results of meta-analysis published by COVID-19 HG contain 23 susceptibility loci [COVID-19 Host Genetics Initiative, 2021]. Loci identified in this meta-analysis include the aforementioned 3p21.31 and 9q34.2 as well as several other – for example, the locus on chromosome 6, which covers the *FOXP4* gene. *FOXP4* plays a key role in regulating lung secretory epithelial cell regeneration and thus may influence mucus production to protect the lungs from pathogens and contaminants. Another locus (12q24.13) contains a set of paralogous genes of the *OAS* family (*OAS1, OAS2, OAS3*). The *OAS* genes encode oligoadenylate synthetases, involved in the innate immune system activity. In addition to these regions, locus 21q22.11 is also associated with severity and hospitalization rate in patients with COVID-19. This locus includes the *IFNAR2* gene, the product of which functions as part of the interferon immunity circuit. The most significant associations found at loci 6p21.1, 12q24.13, and 21q22.11 are also associated with genes with a largely different expression in lungs.

While genome-wide studies have generally been successful in identifying COVID-19 genetic risk factors, finding new associations in poorly studied populations can be difficult due to limited cohort size and/or when genome-wide coverage is limited (i.e., in studies based on CES or targeted sequencing). Except a pilot study on COVID-19 genetics [Shikov *et al*., 2020], Russian investigators have so far failed to produce a large-scale modern molecular genetic activity involving full genomic predictor analysis for different COVID-19 course patterns, including consideration of different waves of the disease.

For this reason, we followed-up on the earlier study [Shikov *et al.,* 2020] to identify extra susceptibility loci, associated with severe course of COVID-19. A cohort of 840 Russian COVID-19 patients underwent screening according to the already selected criteria [Glotov *et al.,* 2021a; Shcherbak *et al.,* 2021]. We used Illumina and MGI NGS platforms and probes for clinical exome sequencing (CES). Prior to association analysis, the phenotypic information was preprocessed and genotyped to yield a complete set of variants within exome target intervals. A total of

727,656 genetic variants were discovered in our sample. 98,382 of these variants were non-synonymous variants (including missense and putative loss-of-function (pLoF) variants). After filtering out variants with low quality and/or call rate, 13,983 of the remaining ones were common (AF>=5%). Out of the remaining rare (AF < 5%) variants, 1884 variants were annotated as pLoF variants in the canonical transcripts of 1121 protein-coding genes. All individuals were assessed for the presence of monogenic immune system disorders, with no pathogenic variants identified according to the ClinVar database (ClinVar v. 20211130) [Shcherbak *et al.,* 2022].

Identification of significant genome- or exome-wide associations can be difficult in cohorts with limited size; hence, we decided to undertake a more systematic approach and analyze the genetic factors of COVID-19 using an extensive collection of phenotypic data available for our cohort of patients. A broad set of more than 100 quantitative and binary traits were collected for each patient [Shcherbak *et al.,* 2022]. The set of traits included the major parameters that serve as predictive risk factors of severe COVID-19 according to a recent publications [Shikov *et al.,* 2020; Shcherbak *et al.,* 2021]: serum levels of key cytokines such as the C-reactive protein and interleukin-6 (IL-6); levels of ferritin, D-dimer, lactate dehydrogenase (LDH), glucose, and creatine in the serum; blood cell count (lymphocytes, leukocytes, neutrophils per mL of blood sample); lung involvement score derived from CT images, as well as the NEWS score. Most values were obtained every 2 days at admission. As expected, the recorded values of most of these traits differed substantially for patients with different outcomes (death or recovery) of hospitalization or disease severity. Considering that normalization of phenotypic data using the inverse rank-based inverse normal transformation IRNT may increase the power of genome-wide association analyses [Goh и Yap, 2009], qualitative parameters in the dataset were normalized using IRNT with extra filtration and prediction of missing data values. Thus, all quantitative traits were additionally pre-processed for further association analysis (Fig. 30).

We first conducted common- and rare-variant association analysis with binary traits (death and severity). Common variant association analysis identified no significant associations and no evidence for the exome-wide association signal. We next tested the involvement of rare variants in clinically significant genes by conducting a series of rare variant association tests using both gene- and pathway-level aggregation of variant counts (a strategy similar to the one used by Povysil and colleagues [Povysil *et al.*, 2021]). To enhance our analysis, we performed both within-cohort tests (i.e., association analysis based on comparison of patients with different COVID-19 outcome or severity) and a comparison with the populational allele frequencies [Barbitoff *et al.*, 2019]. Concordantly with the results obtained by Povysil and colleagues, we found no genes and pathways showing significant association with disease severity or outcome in our dataset.

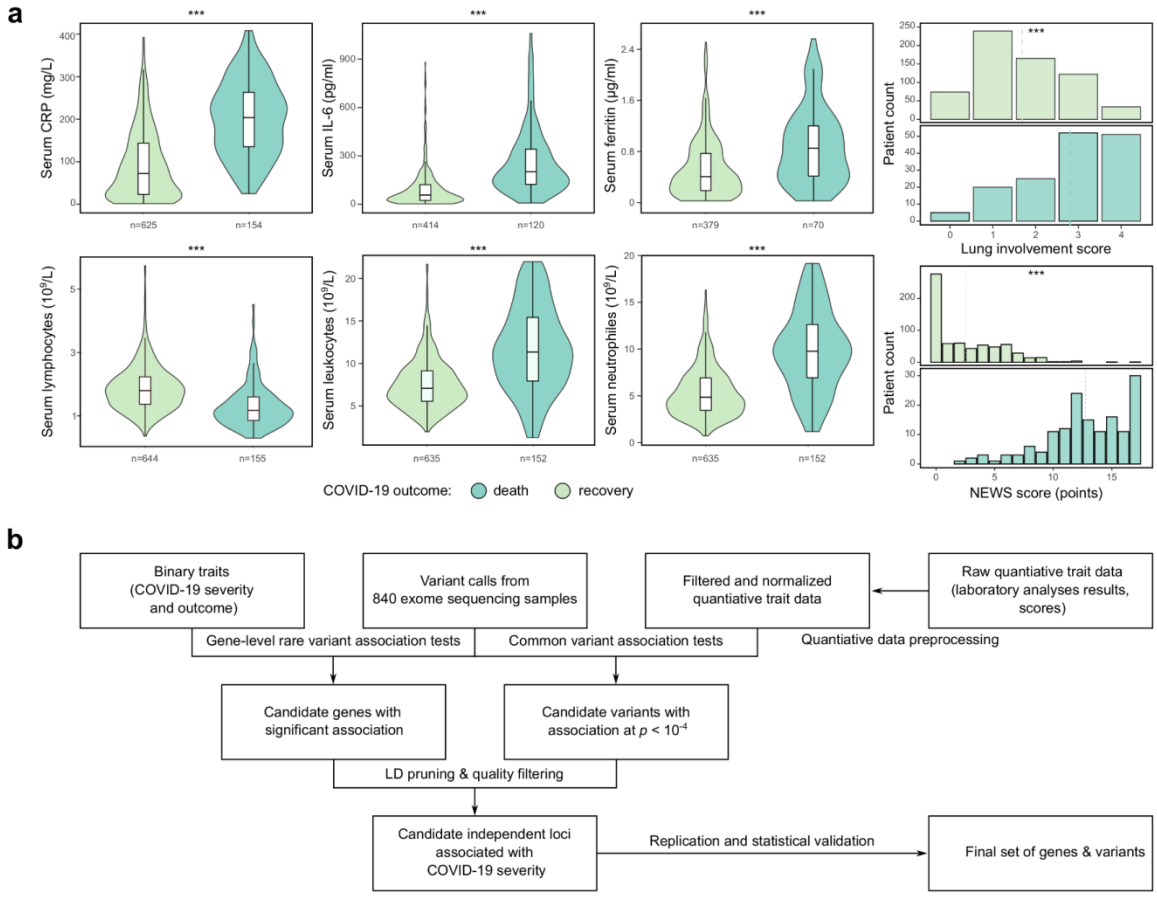


Figure 30. Identification of candidate genetic markers of severe COVID-19 using a deeply phenotype cohort. **(a)** Distributions of selected quantitative traits for

individuals with different disease outcome (death or recovery) in the cohort of 840 COVID-19 patients from Russia. Shown are the distributions of the serum C-reactive protein (CRP), interleukin-6, and D-dimer levels, CT-based lung involvement score (ranging from 0 to 4), counts of lymphocytes, leukocytes, and neutrophils in the blood samples, as well as the NEWS score. All values shown correspond to maximum values recorded during the course of hospitalization. Values exceeding three standard deviations from the population mean are omitted. ***— $p<0.001$ in Wilcoxon-Mann-Whitney test (for quantitative traits) or chi-squared test (for categorical traits). **(b)** A schematic representation of the data analysis pipeline employed in the present study.

We next turned to the analysis of single-variant associations using a broad panel of quantitative trait data. In this analysis, we performed exome-wide association analysis using a set of 13,983 common (MAF > 0.05) variants discovered in our genotype dataset and a set of 53 pre-processed quantitative traits with low missingness rate. After the initial round of GWAS, the results for each trait were manually curated by inspection of the Q-Q plots. In total, we found 5 quantitative traits that showed modest exome-wide association signals. These include the serum C-reactive protein (CRP) levels, lymphocyte, leukocyte, and neutrophil counts, and the lung involvement assessed using CT analysis (Fig. 31).

In total, 15 variants showed association at $p<10^{-4}$ for the selected quantitative variants [Shcherbak *et al.,* 2022]. Only two of the identified variants reached exome-wide significance threshold at ($3.5*10^{-6}$) (a threshold corresponding to the standard significance level of $p<0.05$ corrected for the number of variants tested). These variants showed significant associations with both leukocyte and neutrophil counts. This can be explained by a high degree of correlation between these traits (Table 26). Clustering of these variants by linkage disequilibrium (LD) identified 11 independent loci (1 - for the serum CRP levels; 2 - for lymphocyte, leukocyte, and neutrophil count; and 5 - for the CT-based lung involvement score). Four out of these

substitutions were located in the coding sequences, while the rest of the variants were intronic or other non-coding variants [Shcherbak *et al.*, 2022].

Of the 11 independent variants identified in our analysis, 9 corresponded to significant cis-eQTLs according to the Genotype Tissues Expression (GTEx) data. Four of these variants corresponded to cis-eQTLs affecting the expression of multiple genes across multiple tissues.

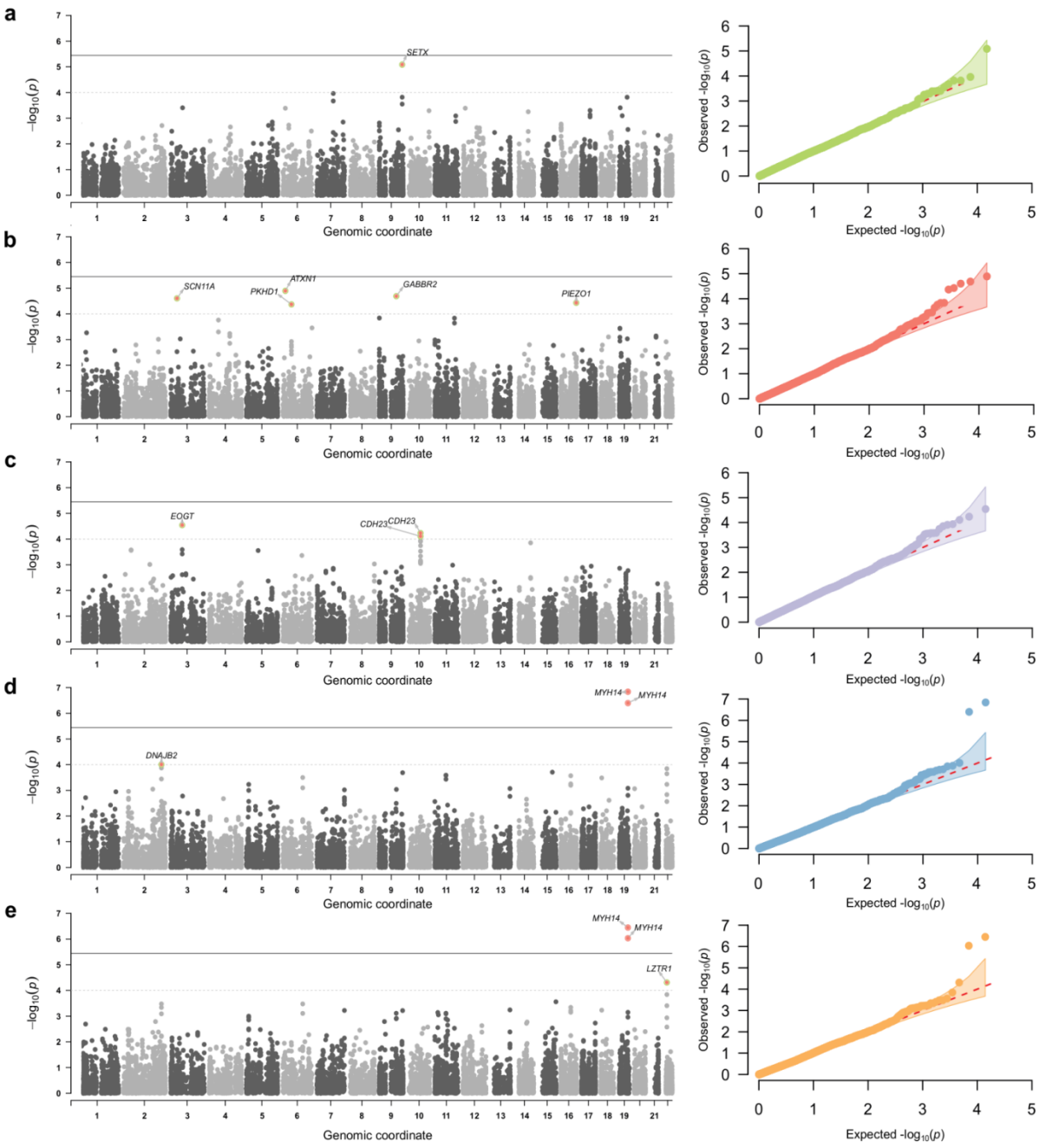


Figure 31. Genome-wide association results for selected quantitative traits in COVID-19 patients [Shcherbak *et al.*, 2022]. We can see Manhattan (left) and quantile-quantile (right) plots of association p-values for (from top to bottom) the serum CRP levels **(a)**, CT-based lung involvement score **(b)**, serum lymphocyte **(c)**, leukocyte **(d)**, and neutrophil **(e)** counts. Thresholds on the Manhattan plots correspond to the exome-wide significance cutoff ($4x10^{-6}$) and the sub looser cutoff $p=10^{-4}$ used to select candidate associations.

Table 26. Candidate genetic variants associated with COVID-19 related quantitative traits in a cohort of Russian patients [Shcherbak *et al.*, 2022].

Locus	rsID	Substitution	AF*	Trait(s)	Gene	Consequence	*B***	p-value	GTEx eQTLs**
2:219280564	rs2276638	6247C>G	0.135	Leukocytes	*DNAJB2*	Intron variant	-0.29	9.84E-05	Multiple genes and tissues
3:38894643	rs33985936	c.2725G>T (p.Val909Phe)	0.241	CT score	*SCN11A*	Missense variant	-0.24	2.50E-05	Multiple genes and tissues
3:68997990	rs4855544	g.20905C>A	0.332	Lymphocytes	*EOGT*	Intron variant	0.23	2.88E-05	Multiple genes and tissues
6:16306520	rs16885	c.2257C>T (p.Pro753Ala)	0.197	CT score	*ATXN1*	Missense variant	0.27	1.28E-05	None
6:51830849	rs1571084	g.261777T>A	0.333	CT score	*PKHD1*	Intron variant	0.21	4.30E-05	*PKHD1* (skin)
9:98299383	rs41273925	g.414815C>G	0.081	CT score	*GABBR2*	Intron variant	0.38	2.06E-05	*TBC1D2* (thyroid)
9:132278286	rs11243705	g.81700A>G	0.180	CRP	*SETX*	Intron variant	0.30	8.18E-06	*SETX* (multiple tissues)
10:71799129	rs4747194	c.7073G>T (p.Arg2358Gln)	0.243	Lymphocytes	*CDH23*	Missense variant	0.25	5.84E-05	*CDH23* (colon, testis), *PSAP* (multiple tissue)
16:88738516	rs34600315	c.*648_*649del	0.657	CT score	*PIEZO1*	Non-coding transcript exon variant	0.21	3.73E-05	*PIEZO1* (whole blood)

19:50259161	rs1651553	c.2127A>G	0.770	Leukocytes, neutrophiles	*MYH14*	Synonymous variant	0.32 0.31	1.45E-07 3.55E-07	none
22:20992196	rs112544	g.14928T>G	0.709	Neutrophiles	*LZTR1*	Intron variant	0.23	4.88E-05	Multiple gene and tissues

* - allele frequency is given with respect to the non-reference allele; ** - the effect sizes are given with respect to the IRNT-transformed values of quantitative traits; *** - data for the GTEx Analysis (full list of significant cis-eQTLs is available). *DNAJB2* - B2 member of the Dna J heat shock protein family (Hsp40); *SCN11A* - sodium voltage-gated channel alpha subunit 11; *EOGT* - EGF domain specific O-linked N-acetylglucosamine transferase; ATXN1 - ataxin 1; *PKHD1* - PKHD1 ciliary IPT domain containing fibrocystin/polyductin; *GABBR2* - gamma-aminobutyric acid type B receptor subunit 2; *SETX* - senataxin; *CDH23* - cadherin related 23; *PIEZO1* - piezo type mechanosensitive ion channel component 1; *MYH14* - myosin heavy chain 14; *LZTR1* - leucine zipper like transcription regulator.

Of these, three variants had the most significant effect on neighboring genes: the rs2276638 intron variant in the *DNAJB2* gene had the most significant effect on the expression of the *PTPRN* gene in whole blood according to the GTEx data (p-value = $2x10^{-27}$); the rs33985936 variant in *SCN11A* had the highest effect on the expression of the *TTC21A* gene in esophagus; and the rs112544 variant in *LZTR1* had the most significant and broad impact on the expression of the THAP7-AS1 antisense transcript. Of the remaining five variants with significant cis-eQTL signal, four had a significant effect on the expression of the gene bearing the corresponding variant, and only one affected the expression of the neighboring gene. It appears likely that the variants in *ATXN1, PKHD1, SETX, PIEZO1*, and *CDH23* have a direct impact on the phenotype by changing the function (in case of missense variants in *ATXN1* and *CDH23*) or expression levels of the corresponding gene.

While we identified 11 independent genetic variants that are associated with quantitative traits that are directly connected to the disease severity and outcome, it is

important to note that the significance level of these associations is not sufficient for making a confident conclusion about the effects on the patient phenotype. This predicates the need for additional replication of the observed associations and validation of their true role in the pathogenesis of COVID-19 [Shcherbak *et al.*, 2022].

To obtain such a validation, we first questioned whether the identified variants can be used to directly predict the severity of the disease and/or outcome in our cohort. We began by constructing a simple risk score by computing the weighted sum of risk alleles in the genotype of each patient. The score had a nearly normal distribution (Figure 32(a)). To test whether such a score has a power to predict the severity or outcome of the hospitalization in COVID-19 patients, we then selected the patients belonging to the top decile of the score distribution (i.e., 10% of all patients with the highest score values). We then used chi-squared statistics to compare disease severity and outcome in these patients and the rest of our sample. Indeed, we found significant differences in all comparisons (Figure 32(b)), with patients belonging to the top risk score decile having greater probability of death and greater disease severity.

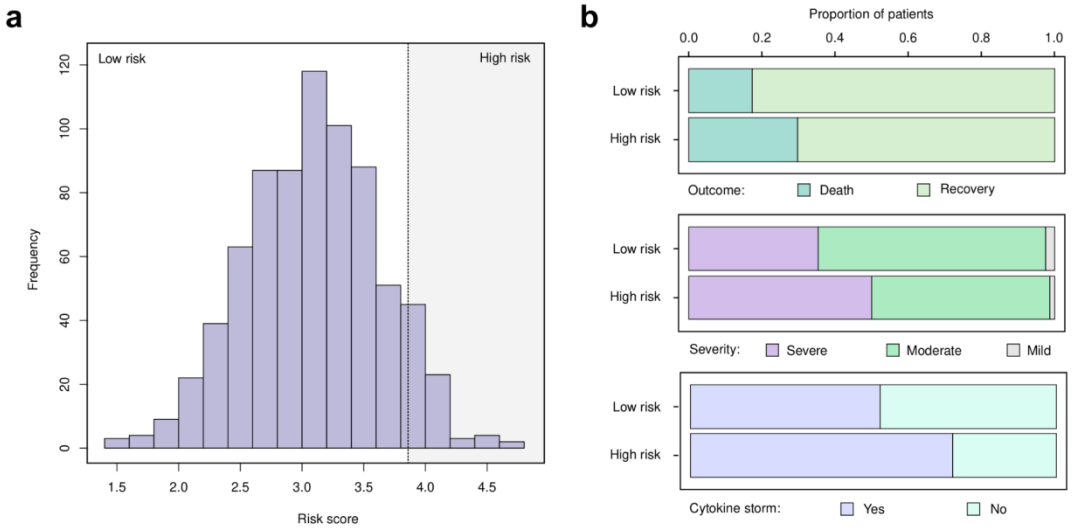


Figure 32. A risk score based on 11 identified variants predicts disease severity and outcome [Shcherbak *et al.*, 2022]. **(a)** Distribution of the risk score computed based on the 11 lead SNPs shown in Table 26. Shaded area indicates the top score decile corresponding to high-risk individuals.

Logistic regression based on the 11 identified markers predicts COVID-19 outcome with ROC/AUC = 0.59. These results confirm that the identified variants can be considered as genetic risk factors of severe COVID-19.

We next went on to replicate the observed associations in independent studies. When using the COVID-19 HG data (COVID-19 HG Initiative, 2021 (release 6) and other studies in cohorts of European ancestry, we successfully replicated only one of 11 candidate associations (rs33985936 in *SCN11A*) which showed modest significance in the analysis of COVID-19 patients vs. population (C2 comparison in COVID-19 HG). We also attempted replicating our findings in the results of the Severe COVID-19 GWAS Group (a study involving patients and controls of Spanish and Italian ancestry). No genetic variants were successfully replicated in this study, presumably due to population-related differences between the groups [Shcherbak *et al.*, 2022].

In addition to replicating the associations in other studies of the COVD-19 host genetics, we sought to identify other (not COVID-19-related) complex traits associated with the identified variants. To this end, we performed phenome-wide association analysis (PheWAS) using the Global Biobank Engine [McInnes *et al.*, 2018]. All associations at $p<10^{-5}$ were considered as significant PheWAS hits. We were able to identify PheWAS hits for 3 out of 11 tested variants. Remarkably, all identified phenome-wide associations corresponded to missense variants and were identified for blood-related traits. The rs33985936 variant in the *SCN11A* gene, the only variant that was replicated in the COVID-19 HG cohort, showed significant association with platelet crit and platelet count in the UK Biobank data. In addition to this variant, rs16885 in *ATXN1* showed significant PheWAS association with mean corpuscular hemoglobin levels, and the rs4747194 variant in *CDH23* was connected to the percentage of monocytes in the blood. These results provide additional support for the biological role of the identified missense variants in driving COVID-19 related traits. To sum up, we identified a set of 11 genetic variants showing modest association with quantitative and nominal traits linked to COVID-19 severity. For

three of these variants, we were able to find supporting evidence substantiating their role in the COVID-19 pathogenesis [Shcherbak *et al.*, 2022].

Several genes belonging to the top associated loci in our study deserve a detailed discussion. First, the most significant (and the only significant on the exome-wide level) variant corresponded to the *MYH14* gene encoding for nonmuscle myosin II C (NMIIC) predominantly expressed in the inner ear, including the organ of Corti [Donaudy *et al.*, 2004]. Pathogenic alterations in the *MYH14* gene are directly associated ausomal dominant nonsyndromic congenital hearing loss (ADNSHL) in European [Lerat *et al.*, 2019; del Castillo *et al.*, 2022] and Asian populations [Hiramatsu *et al.*, 2021; Wang M *et al.*, 2020]. Mutations in the *MYH14* gene may be risk factors for metastasis in patients with pancreatic cancer [Surcel *et al*., 2019] and neuroblastoma [Giwa *et al*., 2020] and provide for the ability of oral cavity cancer cells to escape immune response [Pérez-Valencia *et al*., 2018], as supported by the fact that NMIIC disruption causes hyperplastic epithelial cell proliferation [Nguyen-Ngoc *et al.*, 2017]. Finally, nonsynonymous mutations in the *MYH14* gene increase the development of cochlear nerve canal stenosis [Liang *et al.*, 2021], neurological complications of type 2 diabetes [Rahman *et al*., 2020], peripheral neuropathy [Almutawa *et al.*, 2019] and Charcot-Marie-Tooth disease (CMT) [Kanwal *et al.*, 2018]. The aforementioned data indicate that the *MYH14* gene product is a crucial pleiotropic regulator of homeostasis, hence the presence of variants in this gene associated with various diseases is legitimate. However, it should be noted that the annotation of variants does not provide a clear functional basis for understanding the effect of variants on the phenotype. The effects of variants in the *MYH14* gene on the phenotype may be indirect and caused by the products of other genes at the same locus.

Second, the *DNAJB2* gene encodes an important protein of the Hsp40 chaperone group. Such proteins are known to contribute to the substrate-specificity of other chaperones and mediate the stress response [Kampinga and Craig, 2011; Craig and Marszalek, 2017]. The involvement of the *DNAJB2* gene variation in the levels of lung

damage upon SARS-CoV-2 infection is interesting and may point to the role played by the stress response pathways and protein quality control in disease severity. Presumably, the heat stress response is important for alleviation of negative effects of inflammation on the structure of the tissue. Thus, deviation in the regulation of response to heat dresses may trigger the destruction of the lung tissue and cause lung fibrosis and respiratory problems in COVID-19 patients. Notably, mutations in the *DNAJB2* gene contribute to a more severe course of Charcot-Marie-Tooth disease [Yalcintepe *et al.,* 2021], motor neuropathies (e.g., Frasquet et al., 2021), and other similar diseases.

The association of the locus spanning the *LZTR1* gene is also notable as this gene encodes an important leucine-zipper transcriptional regulator that is linked to cell proliferation in cancer [Zhang Z. *et al.,* 2021]. Pathogenic variants in the *LZTR1* gene affect schwannomas progression, causing peripheral nerve tumors [Ishigami *et al.,* 2021; Piotrowski *et al.,* 2022], neurofibromatosis [Perez-Becerril *et al.,* 2021], Noonan syndrome [Talley *et al.,* 2021], nonimmune edema [Hurni *et al.,* 2021; Zhou *et al.*] The lead variant at the locus, rs112544, has a significant and broad impact on the expression of the *THAP7* gene and its antisense transcript, THAP7-AS1. The *THAP7* gene encodes a transcriptional repressor that acts via histone deacetylation [Macfarlan *et al.,* 2005]. Notably, *THAP7* overexpression induces permissiveness of human hepatoma Huh7 cells to hepatitis C viral invasion [Dächert *et al.,* 2019]. These data suggest that both *LZTR1* per se, as target genes affected by rs112544, may be involved in immune system function and anti-viral response.

The *PIEZO1* gene encodes for mechanically-gated ion channels with multiple roles in human organisms, mutations in this gene are comorbid with pathological conditions, namely, hereditary anemia [Zhou *et al.,* 2021], congenital lymphedema [Mustacich *et al.,* 2021], lymphatic dysplasia [Li *et al.,* 2021], and others, implying that alterations in *PIEZO1* may affect the immune response to SARS-CoV-2.

Variants in fibrocystin (encoded by the *PKHD1* gene) worsen the ramifications of the autosomal recessive polycystic kidney disease (ARPKD) in children and adults

[Zhang Z. *et al.*, 2021]. Notably, the gene's polymorphisms were found to be associated with mild cognitive impairment [Mahmud *et al.*, 2021] and metachronous liver cancer [Ohni *et al.*, 2022].

Similarly, other genes located at the 11 identified loci (e.g., *ATXN1, GABBR2, SETX, CDH23*) are also implicated in nervous system pathology but are not clearly linked to immunity and/or infectious disease response [Kumaran *et al.*, 2014; Wallace and Bird, 2018; Miyazawa *et al.*, 2021; Hadjinicolaou *et al.*, 2021; Saleem *et al.*, 2021]. This result might indicate a certain relationship between nervous system function and COVID-19 severity; however, it is important to note that no significant overrepresentation of nervous system genes was identified at the associated loci.

3.6. Assessment of Clinical Genetic Associations and Method Related Challenges

The COVID-19 pandemic has drawn considerable attention to human genome and virus research. Over the past two years, a number of publications have examined the role of hereditary predisposition to SARS-Cov-2 infection and its severe complications [Suh *et al*., 2022]. The number of GWAS studies related to COVID-19, both large-scale and local, is continuously increasing. Studies are aimed both at identifying host susceptibility factors and analyzing the severity of the disease, as well as at studying the association of specific symptoms with genetic markers. These works allow us to look at epidemiology in new ways. For example, a genetic correlation has been established between lower educational attainment and a higher risk of COVID-19 [Jian *et al.*, 2021]. Various loci associated with risk of odor loss during infection have been found using GWAS: *UGT2A1/UGT2A2* [Shelton *et al*., 2022], *MUC5B* [van Moorsel *et al.*, 2021], LZTFL1 and *RAVER1* [Fink-Baldauf *et al.*, 2022] and others. As in many other GWAS studies, replication of the observed associations and identification of genetic variants associated with the phenotype in different human populations remains an important and challenging problem. For example, Zhang and colleagues identified a possible role of rare pLoF variants in genes associated with type I interferon immunity [Zhang H. *et al.*, 2020] in COVID-19 risk. However, these

associations have not been replicated by other authors [Povysil *et al.*, 2021]. We see a similar result in the replication of loci identified in the Chinese cohort in other external cohorts [Li Y. *et al.*, 2021] and in the COVID-19-HG study. This low replication, although characteristic of complex trait genetics, highlights the special role of study design and population structure on the identification of genetic risk factors for infectious diseases.

Analysis of genetic associations in cohorts of limited size, especially when genome-wide genotypes are absent, can also hinder the discovery of new susceptibility loci in underrepresented populations. Consequently, to address the low statistical power of the analysis, sophisticated approaches must be used, as well as qualified to separate groups for subsequent analysis taking into account the non-genetic risk factors in COVID-19, which we described previously [Shcherbak *et al.*, 2021]. Although only one of the identified variants in our study successfully replicated in external cohorts, two additional variants demonstrated nominal significance in independent studies. Our results are similar to those obtained by Li and colleagues [Li Y. *et al.*, 2021]. As previously argued, the low replication rate may reflect both differences in study design and differences between populations [Li Y. *et al.*, 2021]. As previously argued, the low replication rate may reflect both differences in study design and differences between populations. Perhaps more importantly, we observed significant PheWAS for our three variants in the UK Biobank data. Importantly, all key PheWAS were consistent with traits associated with hematopoiesis, confirming the relevance of the associations identified. In addition, analysis of GTEx eQTLs also shows that many of the identified variants affect gene expression in immune cells or in whole blood (e.g., rs2276638 in the *DNAJB2* gene, rs34600315 in the *PIEZO1* gene). It is noteworthy that only a few of the variants we identified affect gene expression in the lungs. This observation may be explained by the specificity of the assay strategy, which mainly focused on different indices in the blood in patients with COVID-19.

It should be noted that the overall strength of the observed associations in our study is moderate because only one of the loci is reproduced in independent cohorts.

This observation may be explained either by the small sample size and, because of this, weak association signal in our study, or by population-specific variant effects. Therefore, given the differences in the design of the different studies, we do not expect many replications of our data. Nevertheless, our results demonstrate practical value of elaborate laboratory phenotyping of patients with COVID-19 to identify novel genetic variants affecting disease severity and/or outcome. Consequently, we believe that our work can serve as a benchmark for successful indirect assessment of severe COVID-19 risk factors.

The underlying genetic mechanisms of infectious diseases provide an insight into individual mechanisms of disease. The association between the exome and various COVID-19 conditions and outcomes discussed in this chapter demonstrates the importance of research to understand any disease pathogenesis, including infectious processes. There is no doubt that a few years later genome sequencing data would dramatically increase the predictive value of presymptomatic genetic testing for hereditary susceptibility/resistance to infections, with further advancements in molecular medicine yet to arrive. Tests for genetic predisposition to individual infectious diseases would become part of the general genetic testing within Predictive Medicine and the clinical genetic health passport framework.

SUMMARY

Introduction of new genomic technologies in recent years has contributed greatly to further advances in molecular medicine, especially whole genome sequencing and aCGH microarray analysis. These emerging technologies have significantly expanded horizons, with having changed the priorities for genetic lab tests in clinical practice. As new molecular biology approaches have evolved, so have the demands to update the terminology framework. The transition from analyzing individual genes and their variants (mutations) towards understanding pathogenomics and biomarkers of multifactorial diseases, as well as presymptomatic prevention and personalized treatment is more than evident. The emerging concept of a human genetic health passport is gaining prominence [Baranov *et al.,* 2000, 2009, 2021].

Depending on particular objectives, investigators refer to such concepts as ‘gene’ and ‘mutation’. For example, classical genetics defines the gene as a locus mapped on the chromosome and responsible for a particular phenotypic trait [Inge-Vechtomov, 1998]. In molecular biology, a gene is a DNA fragment associated with regulatory sequences and a particular transcription unit [Singer, Berg, 1998]. Mutations are any inherited changes (alterations) in the DNA sequence.

Importantly, the difference between mutations and genetic variants (genetic polymorphisms) is utterly relative. Usually, genetic polymorphism is neutral, more frequent to occur, and can be present in a significant part of the population, i.e. over 1% [Baranov *et al.,* 2000]. Mutations occur less frequently. They disrupt the functionality of a particular gene, leading to significant decrease in the synthesized protein product (minus-effect, loss of function), or surplus (plus-effect gain of function), or to the appearance of an abnormal protein, thus causing monogenic diseases. In contrast to genetic polymorphism, the phenotypic effect of most of the known mutations manifests as an unarguable hereditary disease. Thus, genetic polymorphism and mutations are in fact different dimensions of an identical

phenomenon. The boundary between these concepts is rather vague, as is that between normal physiology and pathology [Cotton and Scriver, 1998].

Today, NGS has critically transformed the underlying biological meaning of these terms. In 2015 the US, followed by Russia in 2017 [Ryzhkova *et al.*, 2017, 2019], developed recommendations regarding the interpretation of nucleotide sequence variants, suggesting to discard the terms ‘mutation’ and ‘polymorphism’ in favor of ‘nucleotide sequence variant’ with the following five modifications: pathogenic; likely pathogenic; uncertain significance; likely benign; benign [Ryzhkova *et al.,* 2019]. Since then, this terminology has become custom both in diagnostics [Ryzhkova *et al*., 2019] and in research. Our work demonstrates feasibility of the updated terminology.

Notably, the pattern (spectrum) and frequency of different variants have pronounced population specificity. This means that variants detected in the population or ethnicity of a particular region differs remarkably from those inhabiting other geographical areas or belonging to other ethnic groups. Environmental adaptation, nutrition differences, severe infections (smallpox, plague, cholera, AIDS), selective advantage of heterozygotes (heterosis effect), founder effect (especially for closed populations), gene drift (random variations in the number of alleles) are the main factors and mechanisms that determine natural variations in frequency and availability of variants in different populations and regions of the world [Inge-Vechtomov, 1998; Ivashchenko and Baranov, 2002].

Thus, knowledge of gene structure, genetic polymorphism, and functions of different variants in the genome of particular population allows to understand the hereditary mechanisms underlying monogenic or MF diseases, facilitating diagnostics, prevention, and treatment. Overall, this promotes predictive medicine and genetic passport framework as a stepping stone towards the Clinical Genetic Passport (CGP) as a result of exome sequencing. Even today CGP is an NGS tool to detect pathogenic variants in probands, high-risk families and couples planning to have healthy children, as well as to answer questions regarding the risk of MFDs. Therefore, development of

population databases (including domestic ones) for relative frequencies of gene variants associated with hereditary, as well as other diseases, accompanied with enhanced bioinformatic and statistical protocols to process and analyze sequencing data are crucial for efficient CGD implementation.

It is important to note that many variants previously listed as pathogenic occur increasingly frequently in healthy individuals to cause Mendelian-inherited disease. This has become a most powerful factor in reducing false positive associations in variants and phenotypes [Lek *et al*., 2016]. In this regard, information on population allele frequency (AF) is widely used to elucidate the clinical significance of gene variants.

To address the above challenges, we conducted a population-based study with a dataset of 5,268 samples and 2,092,456 variants [Barbitoff *et al*., 2021]. Those included 349,811 variants matched the ones reported in our previous publication [Barbitoff *et al.,* 2019]. The detected variants comprised 75.7% of already known (detected in the last dbSNP assembly) and 24.3% (509,409) new ones. The results of the study demonstrated that most genetic markers – at least for recessive pathologies – are common to Russia and other global populations. However, some diseases have a specific spectrum of pathogenic variants [Glotov O. *et al.*, 2019].

Our results demonstrated the demand for genetic databases covering specific populations, which is essential for interpretation of variants and identification of disease risk factors, especially in understudied populations. The current sample size does not yet allow us to draw promising unbiased conclusions regarding the genetic structure of the Russian population. In addition, one can still expect a large number of rare genetic variants in the rest of the population not included into our study. Therefore, further collection of data from genome sequencing centers throughout Russia and examination of healthier donors and patients from different regions followed by register of their results in databases shall be the stepping stone to a comprehensive investigation of genetic variations in the Russian population. These

efforts are bound to substantiate our assumptions regarding most prevalent monogenic disease alleles [Barbitoff *et al.*, 2019; 2021].

Population genetics enables us both to estimate the frequency of a particular disease but to analyze the prevalence of certain variants, for example, in the *ACE2* gene, whose polymorphism may play a role in COVID-19 pathology by affecting important functions of this protein in normal cells. Thus, our comparative analysis of the frequencies of five variants (rs35803318, rs41303171, rs113691336, rs971249, rs228566) in the *ACE2* gene in Russian and European populations found that Russians exhibited traits similar to other European populations, suggesting similar infection incidence and COVID-19 severity. This fact provided a vauable insight into the epidemiological situation at the early stage of the epidemic in March-April 2020 [Shikov *et al.*, 2020].

Bioinformatic processing is another essential part of NGS technology. Our understanding the disease genetics largely depends on the accuracy of bioinformatic sequencing protocols [Barbitoff *et al.*, 2017; 2020]. Errors in the reference sequence associated with the so-called reference minor variants (RMAs, i.e. reference genome loci encompassing a rare or even pathogenic variant) are another matter of concern. Such errors need correction within bioinformatic analysis. A special program to correct these errors [Barbitoff *et al.,* 2018] was utilized in all our studies.

Bioinformatic algorithms, enhanced by new approaches to data interpretation, allow us to describe known, as well as new gene variants. NGS molecular genetic testing allowed to detect new pathogenic variants in the *PKP2*, *LDLR*, *GCK*, *HNF1A*, *BLK, WFS1, EIF2AK3, SLC19A2* genes [Fedyakov *et al.*, 2019; Miroshnikova *et al.*, 2021; Glotov O. *et al*., 2019; Balashova *et al*., 2020]. We demonstrated the prospects of an in-depth analysis of MODY syndrome in patients with more than one genetic variant in a single or more target genes [Glotov O. *et al*., 2019]. Based on our studies, algorithms for effective genetic diagnostics for MODY, Wilson and other diseases have been proposed [Fedyakov *et al.*, 2019; Miroshnikova *et al.*, 2021; Glotov O. *et al.,* 2019; Balashova *et al.,* 2020].

Moreover, cutting edge research methods allow to instantaneously identify several hereditary diseases in any individual. Reported clinical cases refer to jointly inherited X-linked and autosomal dominant forms of ichthyosis [Alaverdian *et al.*, 2019], Wilson disease and hemochromatosis [Tulzunovskaya *et al.*, 2017; Balashova *et al.*, 2020]. By knowing the molecular defects, that lead to the development of the disease, patients may benefit from the most adequate follow-up.

Today modern DNA sequencing methods are present in all areas of medical science, enabling clinical medicine to solve reproductive problems among other issues. Here we have demonstrated the efficiency of NGS for NIPT and PGD [Pendina *et al.*, 2019; Saifitdinova *et al.*, 2020]. A comprehensive approach based on the entire array of molecular genetic, cytogenetic, embryological methods in pregnancy planning is desperately demanded today. Thus, the success of various molecular genetic techniques and their clinical application is elucidated by the case of a family with hereditary pathology [Lyazina *et al*., 2017]. This paper presents a comprehensive and lengthy roadway of clinical and genetic screening to diagnose inherited diseases. This example demonstrates a high demand for a clinical genetic passport (CGP) and a new algorithm for preconceptual screening of families using the entire scope of molecular genetic methods, including next-generation sequencing as a first-line method for pregnancy planning, as well as PGT and NIPT methods for subsequent pregnancy monitoring.

Most diseases are not monogenic, hence, prior to risk assessment, the nature of a disease shall be properly understood (i.e. monogenic, oligogenic or multifactorial condition)., which is not always easy. Our studies on hereditary cardiomyopathy [Glotov *et al.*, 2015; Komissarova *et al*., 2016], familial hypercholesterolemia [Miroshnikova *et al.*, 2021] and MODI [Glotov O. *et al*., 2019] are an attempt to step up to this challenge. Our conclusion is that recent advances in human genome sequencing show that, strictly speaking, almost all hereditary diseases can be labelled oligogenic, including monogenic ones, in terms of their clinical manifestation, largely affected by allelic variants of many other genes, or the so-called disease-associated

modifier genes [Agarwal and Moorchung, 2005; Kousi and Katsanis, 2015]. The entire phenotype is determined by expression and penetrance of an array of gene variants.

The situation with MFDs is somewhat more complicated, since changes in the genome affect disease etiology, with a large set of genes predisposing to the disease (the additivity phenomenon). The disease predisposition is shaped by a large number of environmental factors, while inheritance is not subject to explanation by Mendelian laws only [Baranov *et al*. 2021]. Indeed, identical diagnosis may trigger different risk factors and etiology in different individuals.

However, the situation with risk assessment of multifactorial diseases has changed dramatically in the past five years. Today Enhanced productivity, resolution parameters and costs of genome sequencing provide for routine and more extensive application of these technologies. The horizons of medical practice and disease etiology studies are truly unprecedented [Franks *et al*., 2021]. Cutting edge mathematical models and primarily linear regression bring together genetic and clinical parameters to assess the MFD risk. [Khera *et al.*, 2018].

Our study shows that whole exome sequencing offers rational algorithms to identify genetic markers of complicated diseases, even in limited samples. Multidimensional analysis strategy allowed us to elucidate several eligible candidate loci and SNPs that contribute T2DM and obesity pathogenesis in the Russian population. Overall, rational filtering and ranking of potentially causative variants enhances identification of disease-driving genes by implementing polygenic clinical features and, thus, demonstrate the efficiency of WES in finding novel markers of multifactorial diseases in cohorts of limited size in understudied populations [Barbitoff *et al.*, 2018].

Genomic medicine can also help identify rare conditions concealed behind a complicated multistep and multicomponent disease diagnosis. Moreover, various common diseases have reported rare genetic variants to increase the MFD risk by several times in heterozygous carriers, e.g. the presence of risk variants for familial hypercholesterolemia in 0.4% of the population, which increase the CAD risk 3-fold

[Abul-Husn *et al.,* 2016]. Therefore, risk of monogenic pathologies and MFDs is best assessed using exome sequencing.

While identification of carriers of rare monogenic mutations requires sequencing of specific genes and careful interpretation of the detected mutations functionality, polygenic estimates lend to simple concurrent calculations for many diseases, based on a single genotyping array (Khera *et al*., 2018; Barbitoff *et al*., 2018). The ability to identify individuals at significantly higher genetic risk for a wide range of common diseases (diabetes, cardiovascular disease, etc.) at any age is impregnated with both opportunities and challenges in clinical medicine.

Today, the CGP, genetic mapping, and NGS is applied for reproductive purposes [Baranov *et al.,* 2000; 2021] the following areas: screening for monogenic and oligogenic diseases, pregnancy planning, differential diagnostics and treatment, diagnosis confirmation, prospectively extending into MFD and infectious diseases risk assessment and identification of phenotypic traits in humans (Fig. 33).

Conclusively, the following order of action is assumed to ensure full-fledged implementation of genomic medicine [Chatterjee and García-Closas, 2016]:

6. to identify risk factors – to conduct high-quality epidemiological studies on large sample sizes, high-precision unbiased measurements of phenotypes and exposures are needed to identify new risk factors (including genetic variations, environmental risk factors, biomarkers, etc.);

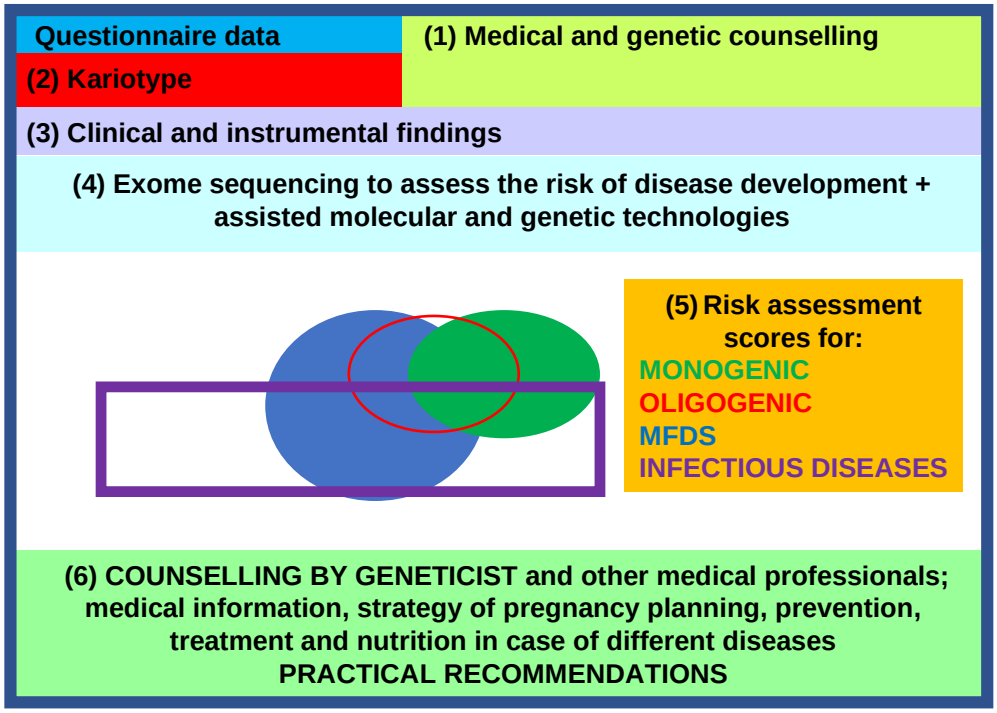


Figure 33. Block diagram for the clinical genetic passport to aid reproduction.

7. **to assess the relative risk –** to develop relative risk models integrating information on multiple risk factors (including estimates of polygenic risk, environmental risk factors and their interaction);

8. **to assess the absolute risk** – to predict the risk of disease development over time considering individual risk factors (using relative risk models, risk factor distributions, total age-specific morbidity and mortality in the target population);

9. **to assess model calibration** – within prospective cohort studies to compare the number of predicted and observed diseases over time in groups of people with different predicted risk;

10. **to assess utility for public healthcare** – to evaluate the efficiency of primary and secondary prevention strategies considering the predicted risk for humans.

The final pivotal step is ethics of absolute and relative risk assessment, as well as informing the physician and patient about these risks, including encouraging lifestyle changes or further disease screening. This is to be elaborated in the foreseeable future, given the existing regulation on personal information protection.

Today national biosafety in any country is largely determined by advances in basic and applied research in immunology and infectious diseases [Khaitov *et al*., 2017]. The new coronavirus infection pandemic has catalyzed these concerns.

Regrettably, neither new antibiotics nor new vaccines can fundamentally solve the problem of combating infectious diseases. A lot earlier Louis Pasteur said, that 'the microbe is nothing; the substrate (i.e., the human) is everything' [Poletaev, Churilov, 2021]. Today we can add to this statement that a virus of today becomes a different virus tomorrow, while the human remains unchanged. It is therefore critical to study the genome and immune system of the host (human) first and foremost, as well as the virus and its genome to a lesser extent. Ilya I. Mechnikov emphasized that the main purpose of the immune system is to maintain the dynamic stability of the molecular and cellular body composition by repairing any damages and regulating a variety of physiological processes [Poletaev, Churilov, 2021].

Immune dysfunction with pronounced uncontrolled generalized systemic inflammatory response exhibiting increased production of inflammatory cytokines ultimately induces a cytokine storm (CS), which is a central problem in COVID-19 pathophysiology. Our results allow us to justify the treatment strategy with early and timely administration of anti-inflammatory and anticoagulation drugs in patients with a high risk of CS development. Such strategy is consistent with CS risk assessment in COVID-19 patients suggested by other investigators [Caricchio *et al*., 2021; Moore *et al*., 2020]. The main cytokine storm risk factors in COVID-19 patients include male gender, age over 40 years, positive SARS-CoV-2 RNA test, lymphopenia, LDH, D-dimer, ferritin, IL-6, and NEWS score dynamics. Information on clinical, instrumental and laboratory risks for coronavirus infection allows to adequately streamline risk groups for genetic studies to elucidate the COVID-19 hereditary profile [Glotov *et al.*, 2021a; Shcherbak *et al.,* 2022].

To study the genetic predisposition to viral infections, profound cell receptor profiling is a collateral to understand how the virus enters the host cell. Notably, a comparative frequency analysis of individual *ACE2* gene variants did not reveal any

significant differences between patients with mild and severe COVID-19 [Shikov *et al*., 2020]. However, our study showed that rare haplotypes in the *ACE2* gene are overrepresented in Russian patients with severe COVID-19. It is possible that rare variants may play a role in COVID-19 severity by impairing important protein functions [Shikov *et al*., 2020]. The evidence obtained suggests that more careful monitoring is needed for COVID-19 patients with rare pathogenic variants in the ACE2 gene, as well as in the whole genome.

The COVID-19 epidemic has enabled an array of large-scale whole-genome studies, including those of the *ACE2* gene (Fig. 34). Rapidly accumulating evidence reported that predictions regarding genetic susceptibility to COVID-19 helped clinicians choose most adequate treatment [Prakrithi *et al*., 2021].

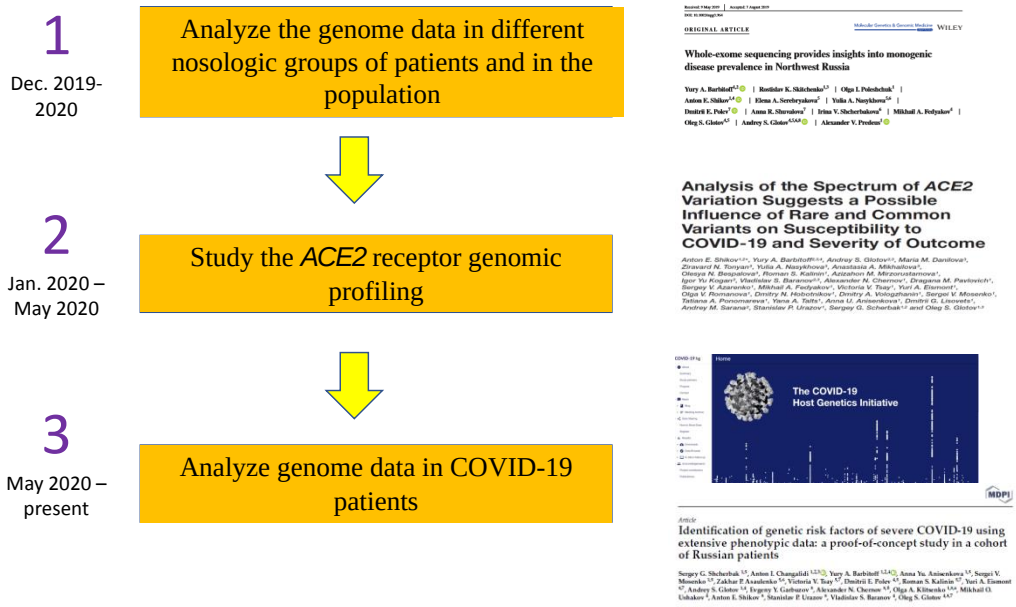


Figure 34. The COVID-19 pandemic as a driver of genome analysis.

Numerous studies have reported a few dozens of associations between genetic variants and COVID-19 incidence, severity and mortality in different ethnic groups [Suh *et al*., 2022]. Despite the large number of reported clinical and genetic associations, variants reported in one study may fail confirmation for a different sample [Suh *et al*., 2022]. Large-scale meta-analyses address this issue to some degree. Today the COVID-19 Host Genetics Initiative project, 2021, accumulates the

most of current data on genetic predisposition to COVID-19 [COVID-19 Host Genetics Initiative].

It is noteworthy that it may be difficult to find new associations in understudied populations due to the limited cohort size and/or limited genome-wide coverage (i.e., in studies relying on CES or target sequencing). To address this problem, we used an original study design [Shikov *et al*., 2020; Shcherbak *et al*., 2022] to look into the aggregation of variant counts at gene and metabolic pathway levels. This is a strategy similar to the one used by Povysil and colleagues [Povysil *et al.*, 2021]. We identified 11 independent genetic variants that are associated with quantitative traits, which, in turn, are directly related to disease severity and outcome. For three of these variants we were able to find supporting evidence for their role in the pathogenesis of COVID-19. However, the level of significance of these associations is insufficient to draw a confident conclusion about how variants affect the phenotype. Additional confirmation is required regarding the associations and confirmation of their role in COVID-19 pathogenesis [Shcherbak *et al.*, 2022].

Given the differences in study design, we do not expect our results to further replicate in other studies. Nevertheless, they demonstrate the utility of in-depth laboratory phenotyping in COVID-19 patients to identify novel genetic variants affecting disease severity and/or outcome.

Molecular medicine and its main areas (predictive medicine, gene therapy, pharmacogenomics, etc.) are to shape diverse base and applied human science in the 21st century and potentially the third millennium. Thus, the concept of Predictive Medicine - the clinical genetic passport for solving the problems of preconceptional screening, PGD, birth of healthy offspring, diagnostics, prevention of MFDs and infectious diseases - shall rely on the new generation sequencing as a fundamental tool using specialized proprietary databases, algorithms and bioinformatics (Fig. 35).

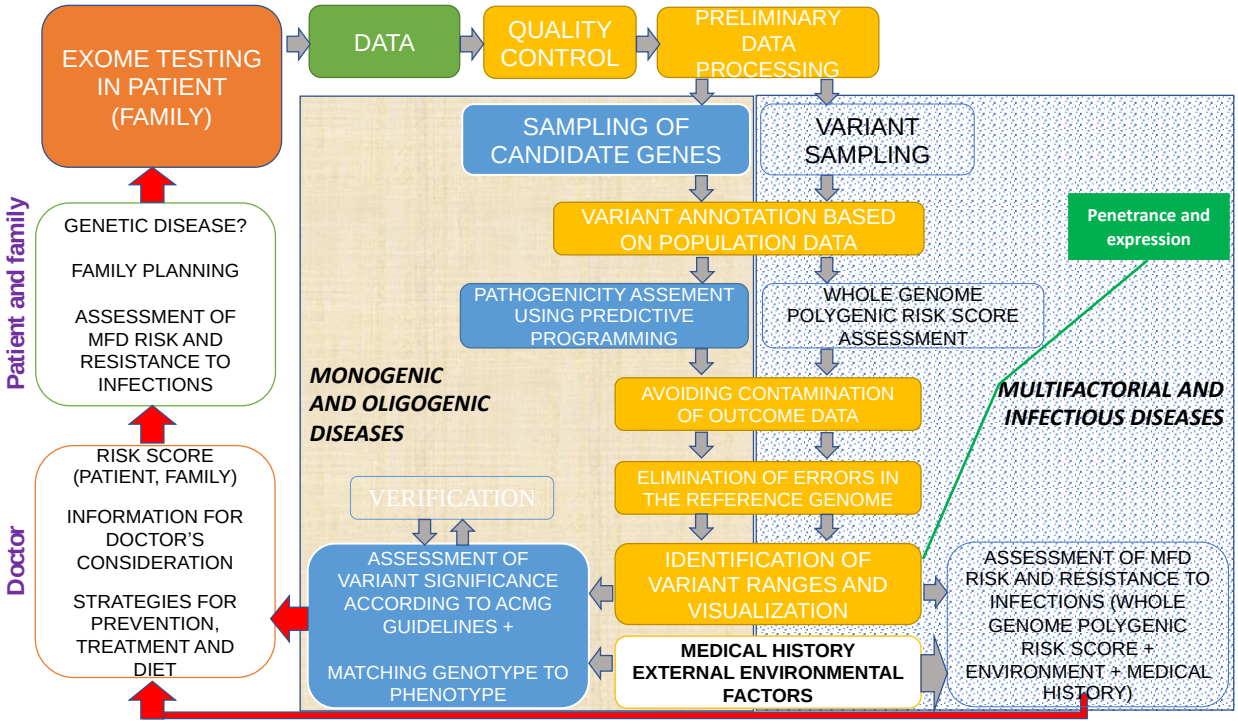


Figure 35. Exome study within Predictive Medicine framework of the clinical genetic human health passport.

In conclusion, we would like to emphasize that scientific foundations of precision medicine and advances in diagnostics and treatment of monogenic, oligogenic, multifactorial and infectious diseases are driven by efficient application of NGS technologies jointly with modern analysis algorithms and genetic custom concepts of expression and penetrance. Future horizons to apply genetic testing include: presymptomatic (pre-emptive) genetic testing (GT) in high-risk families, prospective GT with mandatory follow-up in high-risk individuals based on test findings, and randomized predictive testing [Barnoy, 2007, Baranov *et al*., 2021].

CONCLUSIONS

1. The following recessively inherited monogenic diseases are among the most frequent in the investigated Northwest Russia's population: phenylketonuria, factor VII4 deficiency, kyphoscoliotic type 2 Ehlers-Danlos syndrome, tyrosinase-negative oculocutaneous albinism, and Wilson disease.

2. A randomly sampled cohort of Northwest Russia's population revealed highest frequencies of pathogenic variants in the following genes: *ABCA4* (retinal dystrophy, Stargardt disease) and *CFTR* (cystic fibrosis).

3. The investigated Northwest Russia's population exemplified that NGS technology results in the 25% efficiency improvement in characterizing variants earlier unknown to global research, which underpins the rationale of establishing domestic databases for pathogenic variants.

4. NGS DNA results for monogenic diseases compared across original bioinformatic protocols, based on both international and local databases, enabled identification of new pathogenic variants in the *PKP2, LDLR, GCK, HNF1A, BLK, WFS1, EIF2AK3, SLC19A2, ATP7B, HTT* genes, revealing pathological phenotypes caused by pathogenic variant combinations.

5. Accompanied by classic PCR-based molecular genetic methods to detect pathogenic variants in patients with monogenic diabetes and Wilson disease, NGS provides a 1.5 to 2-fold improvement in the detection of pathogenic variants.

6. Variable manifestation of both oligogenic and multifactorial pathologies in Northwest Russia is associated with complex haplotypes of disease-associated genes, as well as their expression and penetrance, identified using exome sequencing and *original* bioinformatic analysis adapted for small cohorts.

7. Regression modelling demonstrated utmost ability to assess clinical manifestations and phenotype prediction based on pathogenic variant data for

individual genes and their combinations, including polygenic risk factor predictors for oligogenic and multifactorial diseases.

8. NGS-detected variants in the *ATXN1*, *CDH23*, *DNAJB2*, *EOGT*, *GABBR2*, *LZTR1*, *MYH14*, *PIEZO1*, *PKHD1*, *SCN11A*, *SETX* genes, including rare variants in the *ACE2* gene, are associated with novel coronavirus infection COVID-19 severity and clinical outcomes assessed according to the *original* scale.

9. The study revealed that Russian and European populations fail to show any difference in the frequencies for the following five variants (rs35803318, rs41303171, rs113691336, rs971249, rs2285666) in the *ACE2* gene.

10. The study developed a set of predictive genetic examinations paving the way towards the concept of ‘human health genetic clinical passport’ based on exome sequencing data, thus enabling investigators to predict potential oligogenic and multifactorial pathologies and severe course of certain infectious diseases, as well as to explain the versatile pathogenetic mechanisms behind clinical manifestations of various diseases.

MAIN PUBLICATIONS RELEVANT FOR THE DISSERTATION

1. Glotov A.S., Kazakov S.V., Zhukova E.A., Alexandrov A.V., Glotov O.S., Pakin V.S., Danilova M.M., Tarkovskaya I.V., Niyazova S.Sh., Chakova N.N., Komissarova S.M., Kurnikova E.A., Sarana A.M., Sherbak S.G., Sergushichev A.A., Shalyto A.A., Baranov V.S. Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group // Clinica Chimica Acta, 2015, V.446, P.132–140. https://doi.org/10.1016/j.cca.2015.04.014

2. Komissarova S.M., Chakova N.N., Niyazova S.S., Kazakov S.V., Zhukova E.A., Aleksandrov A.V., Glotov O.S., Glotov A.S. The specifics of hypertrophic cardiomyopathy clinical presentation in patients with various mutations of sarcomere genes // Russian Journal of Cardiology (In Russ), 2016, V.1, P.20-25. https://doi.org/10.15829/1560-4071-2016-1-20-25

3. Bliznetz E., Lalayants M., Markova T., Balanovsky O., Balanovska E., Skhalyakho, R., Pocheshkhova E., Nikitina N., Voronin S., Kudryashova E., Glotov O., and Polyakov A. Update of the GJB2/DFNB1 mutation spectrum in Russia: a founder Ingush mutation del(GJB2-D13S175) is the most frequent among other large deletions // J Hum Genet, 2017, V.62, P.789–795. https://doi.org/10.1038/jhg.2017.42

4. Tulzunovskaya I.G., Zhuchenko N.A., Balashova M.S., Filimonov M.I., Rozina T.P., Glotov O.S., Asanov A.V. Wilson disease: intrafamilial clinical polymorphism // Pediatrics. Journal im. G.N. Speransky, 2017, Vol. 96, No. 6, P.215-216. http://doi.org/10.24110/0031-403X-2017-96-6-215-216.

5. Barbitoff Y.A., Bezdvornykh I.V., Polev D.E., Serebryakova E.A., Glotov A.S., Glotov O.S., Predeus A.V. Catching hidden variation: systematic correction of reference minor alleles in clinical variant calling» // Genet. Med, 2018, V.20, P.360-364. http://doi.org/10.1038/gim.2017.168

6. Barbitoff Y.A., Serebryakova E.A., Nasykhova Y.A., Predeus A.V., Polev D.E., Shuvalova A.R., Vasiliev E.V., Urazov S.P., Sarana A.M., Scherbak S.G., Gladyshev D.V., Pokrovskaya M.S., Sivakova O.V., Meshkov A.N., Drapkina O.M., Glotov O.S., Glotov A.S. Identification of Novel Candidate Markers of Type 2 Diabetes and Obesity in Russia by Exome Sequencing with a Limited Sample Size // Genes, 2018, V.9(8), 415. https://doi.org/10.3390/genes9080415

7. Glotov O.S., Romanova O.V., Eismont Y.A., Sarana A.M., Scherbak S.G., Kuzmich E.V., Alyanskiy A.L., Ivanova N.E., Teplyashina V.V., Serov Y.A., Zubarovskaya L.S., Afanasyev B.V. Comparative analysis of NGS and Sanger sequencing methods for HLA typing at a Russian university clinic // Cellular Therapy and Transplantation (CTT), 2018, Vol.7, №4(25), P.72-82. http://doi.org/10.18620/ctt-1866-8836-2018-7-4-72-82

8. Glotov O.S., Serebryakova E.A., Turkunova M.S., Efimova O.A., Glotov A.S., Barbitoff Y.A., Nasykhova Y.A., Predeus A.V., Polev D.E., Fedyakov M.A., Polyakova I.V., Ivashchenko T.E., Shved N.Yu., Shabanova E.S., Romanova O.M., Sarana A.M., Pendina A.A., Scherbak S.G., Musina E.V., Petrovskaya-Kaminskaya A.V., Lonishin L.R., Ditkovskaya L.V., Zhelenina L.A., Tyrtova L.V., Berseneva O.S., Suspitsin E.N., Bashnina E.B., Baranov V.S. Whole-exome sequencing for monogenic diabetes in Russian children reveals wide spectrum of genetic variants in MODY-related and unrelated genes // Molecular Medicine Reports, 2019, V.20, №6, 4905-4914. https://doi.org/10.3892/mmr.2019.10751

9. Balashova M.S., Tulzunovskaya I.G., Glotov O.S., Glotov A.S., Barbitoff Y.A., Fedyakov M.A., Alaverdian D.A., Ivashchenko T.E., Romanova O.V., Sarana A.M., Scherbak S.G., Baranov V.S., Filimonov M.I., Skalny A.V., Zhuchenko N.A., Ignatova T.M., Asanov A.Y. The spectrum of pathogenic variants of the ATP7B gene in Wilson’s disease in the Russian Federation // J Trace Elem Med Biol, 2020, V.59, 126420. https://doi.org/10.1016/j.jtemb.2019.126420

10. Barbitoff Y.A., Skitchenko R.K., Poleshchuk O.I., Shikov A.E., Serebryakova E.A., Nasykhova Y.A., Polev D.E., Shuvalova A.R., Shcherbakova I.V., Fedyakov M.A., Glotov O.S., Glotov A.S., Predeus A.V. Whole exome sequencing provides insights into monogenic disease prevalence in Northwest Russia // Mol Genet Genomic Med, 2019, V.7(11), e964. https://doi.org/10.1002/mgg3.964

11. Pendina A.A., Shilenkova Y.V., Talantova O.E., Efimova O.A., Chiryaeva O.G., Malysheva O.V., Dudkina V.S., Petrova L.I., Serebryakova E.A., Shabanova E.S., Mekina I.D., Komarova E.M., Koltsova A.S., Tikhonov A.V., Tral T.G., Tolibova G.K., Osinovskaya N.S., Krapivin M.I., Petrovskaia-Kaminskaia A.V., Korchak T.S., Ivashchenko T.E., Glotov O.S., Romanova O.V., Shikov A.E., Urazov S.P., Tsay V.V., Eismont Y.A., Scherbak S.G., Sagurova Y.M., Vashukova E.S., Kozyulina P.Y., Dvoynova N.M., Glotov A.S., Baranov V.S., Gzgzyan A.M. and Kogan I.Y. Reproductive History of a Woman With 8p and 18p Genetic Imbalance and Minor Phenotypic Abnormalities // Front. Genet, 2019, V.10, 1164. http://doi.org/10.3389/fgene.2019.01164

12. Alaverdian D.A., Fedyakov M., Polennikova E., Ivashchenko T., Shcherbak S., Urasov S., Tsay V., Glotov O.S. X-linked and autosomal dominant forms of the ichthyosis in coinheritance // Drug Metabolism and Personalized Therapy, 2019, V.34, №4, 20190008. https://doi.org/10.1515/dmpt-2019-0008

13. Barbitoff Y.A., Polev D.E., Shcherbakova I.V., Serebryakova E.A., Kiselev A.M., Kostareva A.A., Glotov O.S., Predeus A.V. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage // Sci Rep, 2020, V.10, 2057 https://doi.org/10.1038/s41598-020-59026-y

14. Shikov A.E., Barbitoff Y.A., Glotov A.S., Danilova M.M., Tonyan Z.N., Nasykhova Y.A., Mikhailova A.A., Bespalova O.N., Kalinin R.S., Mirzorustamova A.M., Kogan I.Y., Baranov V.S., Chernov A.N., Pavlovich D.M., Azarenko S.V., Fedyakov M.A., Tsay V.V., Eismont Y.A., Romanova O.V., Hobotnikov D.N.,

Vologzhanin D.A., Mosenko S.V., Ponomareva T.A., Talts Y.A., Anisenkova A.U., Lisovets D.G., Sarana A.M., Urazov S.P., Scherbak S.G. and Glotov O.S. Analysis of the Spectrum of ACE2 Variation Suggests a Possible Influence of Rare and Common Variants on Susceptibility to COVID-19 and Severity of Outcome // Front. Genet, 2020, V.11, 551220. http://doi.org/ 10.3389/FGENE.2020.551220

15. Goloshchapov O.V., Bakin E.A., Kucher M.A., Stanevich O.V., Suvorova M.A., Gostev V.V., Glotov O.S., Eismont Yu.A., Polev D.E., Lobenskaya A.Yu., Klementeva R.V., Goloshchapova M.O., Zubarovskaya L.S., Sidorenko S.V., Suvorov A.N., Moiseev I.S., Chukhlovin A.B. Bacteroides fragilis is a potential marker of effective microbiota transplantation in acute graft -versus-host disease treatment // Cellular Therapy and Transplantation (CTT), 2020, V.9(2), P.47-59. doi: 10.18620/ctt-1866-8836-2020-9-2-47-59

16. Miroshnikova V.V., Romanova O.V., Ivanova O.N., Fedyakov M.A., Panteleeva A.A., Barbitoff Y.A., Muzalevskaya M.V., Urazgildeeva S.A., Gurevich V.S., Urazov S.P., Scherbak S.G., Sarana A.M., Semenova N.A., Anisimova I.V., Guseva D.M., Pchelina S.N., Glotov A.S., Zakharova E.Y., Glotov O.S. Identification of novel variants in the LDLRgene in Russian patients with familial hypercholesterolemia using targeted sequencing // Biomedical Reports, 2021, V.14.1, 15. http://doi.org/10.3892/BR.2020.1391

17. Glotov O.S., Chernov A.N., Scherbak S.G. and Baranov V.S. Genetic Risk Factors for the Development of COVID-19 Coronavirus Infection // Russian Journal of Genetics, 2021, V.57, №8, P.878–892. http://doi.org/10.1134/S1022795421080056

18. Shikov A., Tsay V., Fedyakov M., Eismont Y., Rudnik A., Urasov S., Sherbak S., and Glotov O. The application of Nanopore sequencing for variant calling on the human mitochondrial DNA // Bio. Comm., 2021, V.66(2), P.109–123. https:// doi.org/10.21638/spbu03.2021.202

19. Shcherbak S.G., Anisenkova A.Y., Mosenko S.V., Glotov O.S., Chernov A.N., Apalko S.V., Urazov S.P., Garbuzov E.Y., Khobotnikov D.N., Klitsenko O.A., Minina E.M. and Asaulenko Z.P. Basic Predictive Risk Factors for Cytokine Storms in COVID-19 Patients // Front. Immunol, 2021, V.12, 745515. http://doi.org/ 10.3389/fimmu.2021.745515

20. Glotov O.S., Chernov A.N., Korobeynikov A.I., Kalinin R.S., Tsai V.V., Anisenkova A.Yu., Urazov S.P., Lapidus A.L., Mosenko S.V., Shcherbak S.G. The lineage of coronavirus SARS-CoV-2 of Russian origin: Genetic characteristics and correlations with clinical parameters and severity of coronavirus infection // The Siberian Journal of Clinical and Experimental Medicine (In Russ.), 2021, V.36(4), P.132–143. https://doi.org/10.29001/2073-8552-2021-36-4-132-143

21. Shcherbak S.G., Changalidi A.I., Barbitoff Y.A., Anisenkova A.Y., Mosenko S.V., Asaulenko Z.P., Tsay V.V., Polev D.E., Kalinin R.S., Eismont Y.A., Glotov A.S., Garbuzov E.Y., Chernov A.N., Klitsenko O.A., Ushakov M.O., Shikov A.E., Urazov S.P., Baranov V.S., Glotov O.S. Identification of Genetic Risk Factors of Severe COVID-19 Using Extensive Phenotypic Data: A Proof-of-Concept Study in a Cohort of Russian Patients // Genes, 2022, V.13(3), 534. https://doi.org/10.3390/genes13030534

22. Turkunova M.E., Barbitoff Y.A., Serebryakova E.A., Polev D.E., Berseneva O.S., Bashnina E.B., Baranov V.S., Glotov O.S. and Glotov A.S. Molecular Genetics and Pathogenesis of the Floating Harbor Syndrome: Case Report of Long-Term Growth Hormone Treatment and a Literature Review // Front. Genet, 2022, V.13, 846101. https://doi.org/10.3389/fgene.2022.846101

23. Glotov A.S., Zelenkova I.E., Vashukova E.S., Shuvalova A.R., Zolotareva A.D., Polev D.E., Barbitoff Y.A., Glotov O.S., Sarana A.M., Shcherbak S.G., Rozina M.A., Gogotova V.L., Predeus A.V. RNA Sequencing of Whole Blood Defines the Signature of High Intensity Exercise at Altitude in Elite Speed Skaters // Genes, 2022, V.13(4), 574. https://doi.org/10.3390/genes13040574

24. Koshevaya Y.S., Kusakin A.V., Buchinskaia N.V., Pechnikova V.V., Serebryakova E.A., Koroteev A.L., Glotov A.S., and Glotov O.S. Description of first registered case of the Lopes-Maciel-Rodan syndrome in Russia // Int. J. Mol. Sci, 2022, V.23, 12437. https://doi.org/10.3390/ijms232012437

DESIGNATIONS AND ABBREVIATIONS

ACMG	American College of Medical Genetics and Genomics
AD	Autosomal dominant
ADNSHL	Autosomal dominant nonsyndromic hereditary hearing loss
AF	Frequency of population alleles
AFR	African population
AH	Arterial hypertension
AIC	Akaike information criterion
AMR	American Mixed population
ARDS	Acute respiratory distress syndrome
ARPKD	Autosoman recessive polycystic kidney diseases
ART	Assisted Reproductive Technologies
ARVC	Arrhythmogenic cardiomyopathy/right ventricular dysplasia
ASJ	The population of Ashkenazi Jews
ATP	Adenosine triphosphoric acid
AUC	Area Under Curve
b.p.	Base pairs
BMI	Body mass index
CAD	Coronary heart disease
CGP	Clinical Genetic Panel
CHL	Cholesterol
CI	Confidence interval
ClinVar	ClinVar database
CMT	Charcot-Marie-Tooth disease

CNV	Copy number variation
COPD	Chronic Obstructive Pulmonary Disease
CoV	Coroviruses
COVID-19	New coronavirus infection
COVID-19 HG	COVID-19 Host Genetics Initiative project – COVID-19 genome research project
CRP	C-reactive protein
CS	Cytokine storm
CT	Computed tomography
CVD	Cardiovascular diseases
dbNSFP	dbNSFP database
dbSNP	dbSNP database
DECIPHER	DECIPHER database
DM	Diabetes mellitus
DNA	Deoxyribonucleic acid
EAS	East Asian population
EPMA	European Society for Predictive, Preventive and Personalized Medicine
EVS	Exome variant sequencing
ExAC	Exome Aggregation Consortium
FDA	Food and Drug Administration
FHS	Floating Harbor syndrome
FIN	Finnish population
GATK	Genome Analysis ToolKit
GLM	General linear model

gnomAD	gnomAD database
GP	Genetic Passport
GPS	Genome-wide polygenic score
GRC	Genome Reference Consortium
GRCh38	Genome Reference Consortium Human Build 38
GT	Genetic Testing
GTEx	Genotype Tissues Expression database
GWAS	Genome-wide association studies
HapMap	Haplotype Map
HCM	Hypertrophic cardiomyopathy
HDL	High-density lipoproteins
HGP	The Human Genome Project
HGVS	Human Genome Variation Society
HLA	Human Leukocyte Antigens
HLH	Hemophagocytic lymphohistiocytosis
HPS	Hemophagocytic syndrome
HUGE	HUGE database
ICM	Internal cell mass
IIBDGC	International Inflammatory Bowel Disease Genetics Consortium
IRNT	Inverse rank-based normal transformation
IVF	In-vitro fertilization
LALD	Lysosomal acid lipase deficiency
LD	Linkage disequilibrium

LDH	Lactate dehydrogenase
LDL	Low-density lipoproteins
MAF	Minor allele frequency
MAS	Macrophage activation syndrome
MERS-CoV	Middle East respiratory syndrome
MF	Multifactorial
MFD	Multifactorial disease
MODY	Monogenic diabetes
mRNA	Matrix RNA
NCBI	National Center for Biotechnological Information, US
NEWS	National Early Warning Score
NFE	Non-Finnish population
NGS	Next generation sequencing
NHLBI	The US National Heart, Lung, and Blood Institute (NHLBI) exome project
NIPT	Noninvasive prenatal test
NMIIC	Nonmuscle myosin II CII C
NPV	Negative Predicted Values
NWR	North-West region population
OMIM	OMIM database
OR	Odds ratio
ORF	Open reading frame
PCA	Principal component analysis
PCR	Polymerase chain reaction

PGD	Preimplantation genetic diagnosis
PGT	Prenatal genetic testing
PGT-A	Prenatal genetic testing for chromosomal abnormalities
PheWAS	Phenome-wide association study database
pLoF	Putative loss-of-function
PM	Predictive, preventive, personalized medicine
PPV	Positive Predicted Values
PRSs	Polygenic risk score
QTLs	Quantitative Trait Loci
RMA	Reference minor allele
RNA	Ribonucleic acid
RNA-seq	RNA sequencing data
ROC	Receiver operating characteristic
RT-PCR	Real-time polymerase chain reaction
RUSeq	Russian sequencing consortium
SARS-CoV	Acute severe respiratory syndrome
SARS-CoV-2	COVID-19 virus
SAS	South Asian population
SCD	Sudden cardiac death
SNP	Single nucleotide polymorphism
SNPSIFT	Genomic variant annotations and functional effect prediction toolbox
SNV	Single-nucleotide variants
TE	Trophectoderm

TS	Target sequencing
US	Ultrasound
VLC	Vital lung capacity
VLDL	Very low-density lipoproteins
WD	Wilson disease
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WHO	World Health Organization
WHR	Waist hip ratio

REFERENCES

1. Abe K. *et al.* Pro108Ser mutant of SARS-CoV-2 3CLpro reduces the enzymatic activity and ameliorates COVID-19 severity in Japan // medRxiv and bioRxiv. 2021. doi: https://doi.org/10.1101/2020.11.24.20235952

2. Abramov A., Schorr S., Wolman M. Generalized xanthomatosis with calcified adrenals // Am J Dis Child. 1956. V. 91(3). P. 282–286.

3. Abramov D.D. *et al.* Carrier frequency of *GJB2* and *GALT* mutations associated with sensorineural hearing loss and galactosemia in the Russian population // Bulletin of Russian State Medical University. 2017. V. 6. P. 20– 23.

4. Abramov D.D. *et al.* High carrier frequency of *CFTR* gene mutations associated with cystic fibrosis, and *PAH* gene mutations associated with phenylketonuria in Russian population // Bulletin of Russian State Medical University. 2015. V. 4. P. 32– 35.

5. Abul-Husn N.S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system // Science. 2016. V. 354.

6. Agarwal S., Moorchung N. Modifier genes and oligogenic disease // J. Nippon Medical School. 2005. V. 72. N 6. P. 326–334.

7. Alaverdian D.A. *et al.* X-linked and autosomal dominant forms of the ichthyosis in coinheritance // Drug Metab Pers Ther. 2019.

8. Allen H.L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height // Nature. 2010. V. 467. № 7317. C. 832-838.

9. AllSeq. WGS vs. WES. Available at: http://allseq.com/kb/wgsvswes/ [accessed November 16, 2018].

10. Almutawa W. *et al.* The R941L mutation in *MYH14* disrupts mitochondrial fission and associates with peripheral neuropathy // EBioMedicine. 2019. V. 45. P. 379–92.

11. Anisenkova A. *et al.* Immunoinformatics in COVID-19 Vaccine Development: The Role of HLA System // Cellular Therapy and Transplantation (CTT). V. 10. №. 1. 2021.

12. Ashour M.H. *et al.* Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks // Pathogens. 2020. V. 9. № 3. P. 186.

13. Aslanidis C. *et al.* Genetic and biochemical evidence that CESD and Wolman disease are distinguished by re sidual lysosomal acid lipase activity // Genomics. 1996. V. 33(1). P. 85–93.

14. Asselta R. *et al. ACE2* and *TMPRSS2* variants and expression as candidates to sex and country differences in COVID-19 severity in Italy // Aging (Albany NY). 2020. V. 12. №. 11. P. 10087–98.

15. Aulchenko Y. S. Development and application of methods for genome-wide analysis of genetic associations of complex traits: Ph. Novosibirsk. 2010. 291 p.

16. Auton A. *et al.* A Global Reference for Human Genetic Variation // Nature. 2015. V. 526 (7571). P. 68–74.

17. Balashova M.S. *et al.* The spectrum of pathogenic variants of the *ATP7B* gene in Wilson disease in the Russian Federation // J Trace Elem Med Biol. 2020. V. 59. P. 126420.

18. Baranov A.A. *et al.* Deficiency of lysosomic acid lipase: clinical recommendations for child health care delivery // Pediatric pharmacology. 2016. V. 13(3). P. 239–243.

19. Baranov V.S. *et al.* Human Genome and "predisposition" genes (Introduction to Predictive Medicine) // SPb. Intermedica Publisher. 2000. 272 p.

20. Baranov V.S., Baranova E.V., Ivaschenko T.E. Human genome as a scientific basis of predictive medicine. Genomics to medicine // Ed. by Ivanov V.I. and Kiselev L.L., Moscow, Akademkniga. 2005. P. 361-380.

21. Baranov V.S. *et al.* Determination of hereditary predisposition to certain common diseases in pregnancy. Genetic map of reproductive health. Guidelines /

Saint-Petersburg, 2009. Ser. Ex libris ‘Journal of Obstetrics and Women's Diseases’. 68 p.

22. Baranov V.S. *et al.* Genetic passport the basis of individual and predictive medicine. Ed. by V. S. Baranov - SPb: ‘Izd vo N-L’, Ltd. 2009. 527 p.

23. Baranov V.S. *et al.* Evolution of predictive medicine / pod. Ed. V.S. Baranov. - St. Petersburg: Eko-Vektor. 2021. 359 p.

24. Baranov V.S., Baranova E.B. Genetic passport: the state of the problem today and tomorrow // Bulletin. Roszdravnadzor. 2018. №6. P.16-23.

25. Baranov V.S., Khavinson V.Kh. Determination of genetic predisposition to hereditary and multifactorial diseases. Genetic passport //Methodic guidelines. SPb.: Foliant. 2001. 48 p.

26. Barbetti F and D'Annunzio G. Genetic causes and treatment of neonatal diabetes and early childhood diabetes // Best Pract Res Clin Endocrinol Metab. 2018. V. 32. P. 575-591.

27. Barbitoff Y.A. *et al.* Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples // medRxiv preprint.

28. Barbitoff Y.A. *et al.* Catching hidden variation: systematic correction of reference minor alleles in clinical variant calling» // Genet. Med. 2018. V. 20. P. 360-364.

29. Barbitoff Y.A. *et al.* Identification of Novel Candidate Markers of Type 2 Diabetes and Obesity in Russia by Exome Sequencing with a Limited Sample Size // Genes. 2018. V. 9(8). P. 415.

30. Barbitoff Y.A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage // Sci Rep. 2020. V. 10 P. 2057.

31. Barbitoff Y.A. *et al.* Whole exome sequencing provides insights into monogenic disease prevalence in Northwest Russia // Mol Genet Genomic Med. 2019 V. 7(11). e964.

32. Barkhatov I.M. et al. New generation sequencing and its application in oncohematology // Oncohematology. 2016. T. 11. P. 56-63.

33. Barnoy S. Genetic testing for late-onset diseases: effect of disease controllqability, test predictivity, and gender on the decision to take the test // Genetic testing. 2007. V. 11. №. 2. P. 187-193.

34. Beatrijs L.P. *et al.* Strengers and Louis J. Bont. Down Syndrome: A Novel Risk Factor for Respiratory Syncytial Virus Bronchiolitis — A Prospective Birth-Cohort Study // Pediatrics. 2007. V. 120. №. 4. e1076–e1081.

35. Benetti E. *et al. ACE2* gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population // Eur. J. Hum. Genet. 2020.

36. Bennett K. *et al.* Four novel cases of permanent neonatal diabetes mellitus caused by homozygous mutations in the glucokinase gene // Pediatr Diabetes. 2011. V. 12. P. 192-P196.

37. Bernal J.L. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant // N Engl J Med. 2021. V. 385. P. 585-594.

38. Bick A.G. *et al.* Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes // Nature. 2020. V. 586. №. 7831. P.763–68.

39. Biesecker L.G., Green R.C. Diagnostic clinical genome and exome sequencing // N Engl J Med. 2014. V. 370 (25). P. 2418-25.

40. Blanco-Melo D. *et al.* Imbalanced Host Response to SARSCoV-2 Drives Development of COVID-19 // Cell. 2020. V. 181(5). P. 1036–1045.

41. Bliznetz E. *et al.* Update of the *GJB2/DFNB1* mutation spectrum in Russia: a founder Ingush mutation del (*GJB2*-D13S175) is the most frequent among other large deletions // J Hum Genet. 2017. V. 62. P. 789–795.

42. Boomsma D.I. *et al.* The Genome of the Netherlands: Design, and Project Goals // European Journal of Human Genetics. 2014. V. 22 (2). P. 221–27.

43. Caccuri F. *et al.* A persistently replicating SARS-CoV-2 variant derived from an asymptomatic individual // J Transl Med. 2020. V. 18. P. 362.

44. Cao Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor *ACE2* in different populations // Cell Discov. 2020. V. 6. 11.

45. Capriotti E., Fariselli P. and Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure // Nucleic Acids Res. 2005. V. 33. P. 306-310.

46. Carere D.F. *et al.* Consumers report lower confidence I their genetics knowledge fooling direct-to-consumer personal genomic testing // Genet. Med. 2016. 2016. Vol. 18. № 1. P. 65-72.

47. Caricchio R. *et al.* Preliminary predictive criteria for COVID-19 cytokine storm // Ann Rheum Dis. 2020.

48. Caricchio R. *et al.* Preliminary Predictive Criteria for COVID-19 Cytokine Storm // Ann Rheum Dis. 2021. V. 80(1). P. 88–95.

49. Cascella M. *et al.* Features, Evaluation and Treatment Coronavirus (COVID-19) // StatPearls. 2020. PMID: 32150360.

50. Caso F. *et al.* Could Sars-coronavirus-2 trigger autoimmune and/or autoinflammatory mechanisms in genetically predisposed subjects? // Autoimmun Rev. 2020. V. 19(5). P. 102524.

51. Chatterjee N., Shi J., García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention // Nat Rev Genet. 2016. V. 17(7). P. 392-406.

52. Chen J. *et al.* Individual variation of the SARS-CoV2 receptor *ACE2* gene expression and regulation // Aging Cell. 2020. V.19. e13168.

53. Chen Y., Guo Y., Pan Y., Zhao Z.J. Structure analysis of the receptor binding of 2019-nCoV // Biochem. Biophys. Res. Commun. 2020. V. 525. №. 1. P. 135–140.

54. Cingolani P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3 // Fly. 2012. V. 6. P. 80–92.

55. Cohen J.C. *et al.* Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease // N Engl J Med. 2006. V. 354. P. 1264–72.

56. Collins F.S., McKusick V.A. Implication of Human Genome Project for Medical Science // JAMA. 2001. V. 285. №. 5. P. 1–11.

57. Collins F.S., Varmus H. A new initiative on precision medicine //N Engl J Med. 2015. V. 372. P. 793–5.

58. Costela-Ruiz V.J. *et al.* SARS-CoV-2 infection: The role of cytokines in COVID-19 disease // Cytokine Growth Factor Rev. 2020. V. 54. P. 62–75.

59. Cotton R.G., Scriver C.R. Proof of "disease causing" mutation // Hum Mutat. 1998. V. 12(1). P. 1-3.

60. Coutard B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade // Antiviral. Res. 2020. V. 176. P. 104742.

61. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 // Nature. 2021. V. 600. P. 472–477.

62. Craig E.A., Marszalek J. How Do J-Proteins Get Hsp70 to Do So Many Different Things? //Trends Biochem. Sci. 2017. V. 42. P. 355–368.

63. Dächert C., Gladilin E., Binder M. Gene Expression Profiling of Different Huh7 Variants Reveals Novel Hepatitis C Virus Host Factors // Viruses. 2019. V. 12. P. 36.

64. de Bie P. *et al.* Molecular pathogenesis of Wilson and Menkes disease: Correlation of mutations with molecular defects and disease phenotypes // Journal of Medical Genetics. 2007. V. 44(11). P. 673– 688.

65. de Haan *et al.* Multiple SNP testing improves risk prediction of first venous thrombosis // Blood. 2012. V. 120 (3). P. 656–663.

66. Deelen J. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes // Nat. Commun. 2019. V. 10. № 1. P. 3669.

67. del Castillo I. *et al.* Genetic etiology of non-syndromic hearing loss in Europe // Hum Genet. 2022.

68. De Pristo M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data // Nature Genetic senetics. 2011. V. 43(5). P. 491– 498.

69. Devaux C., Rolain J.M., and Raoult D. *ACE2* receptor polymorphism: susceptibility to SARS-CoV-2, hypertension, multi-organ failure, and COVID- 19 disease outcome // J. Microbiol. Immunol. Infect. 2020. V. 53. P. 425–435.

70. Di Taranto M.D., Giacobbe C. and Fortunato G. Familial hypercholesterolemia: A complex genetic disease with variable phenotypes // Eur J Med Genet. 2020. V. 63. P. 103831.

71. Donaudy F. *et al.* Nonmuscle Myosin Heavy-Chain Gene *MYH14* Is Expressed in Cochlea and Mutated in Patients Affected by Autosomal Dominant Hearing Impairment (DFNA4) // Am J Hum Genet. 2004. V. 74. P. 770–6.

72. Ehret G.B. *et al.* The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals // Nat. Genet. 2016. V. 48. P. 1171–1184.

73. Estrada K. *et al.* Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population // JAMA. 2014. V. 311. P. 2305–14.

74. Fakhro K.A. *et al.* The Qatar genome: A population-specific tool for precision medicine in the middle east // Human Genome Variation. 2016. V. 3(1). P. 16016.

75. Fanale D. *et al.* Breast cancer genome-wide association studies: there is strength in numbers // Oncogene. 2012. V. 31. №. 17. P. 2121-8.

76. Fang L. Karakiulakis G., Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? // Lancet Respir Med. 2020. V. 8. №. 4: e21.

77. Fedyakov M.A. et al. Anauksetic dysplasia: clinic, molecular genetic diagnosis and treatment // Molecular biological technologies in medical practice / Ed. by A.B. Maslennikov, Corresponding Member of RANS. Vol. 32. Novosibirsk: Akademizdat. 2021. P. 81-92.

78. Fedyakov M.A. *et al.* New frameshift mutation found in *PKP2* gene in arrhythmogenic right ventricular cardiomyopathy/dysplasia: a family case study // Vestnik of Saint Petersburg University. Medicine. 2019. V. 14(1). P. 3–13.

79. Fedyakov M.A. *et al.* The Incidence of Lysosomal Acid Lipase Deficiency in the Russian Population // Pediatricheskaya farmakologiya — Pediatric pharmacology. 2018. V. 15 (2). P. 184–185.

80. Feldmann M. *et al.* Trials of anti-tumour necrosis factor therapy for COVID-19 are urgently needed // The Lancet. 2020. V. 395. P. 1407-1409.

81. Fink-Baldauf I.M. *et al.* CRISPRi links COVID-19 GWAS loci to *LZTFL1* and *RAVER1* // eBioMedicine. 2022. V. 75. P. 103806.

82. Flajollet S. *et al. RREB-1* is a transcriptional repressor of HLA-G // J. Immunol. 2009. V. 183. P. 6948–6959.

83. Franks P.W. *et al.* Technological readiness and implementation of genomic-driven precision medicine for complex diseases // J Intern Med. 2021. V. 290(3). P. 602-620.

84. Fraser F.C. and Gunn T. Diabetes mellitus, diabetes insipidus, and optic atrophy. An autosomal recessive syndrome? // J Med Genet. 1977. V. 14. P. 190-193.

85. Frazer K.A. et al. The International Hapmap Consortium. A second-generation human haplotype map of over 3.1 million SNPs // Nature. 2007. V. 449. P. 851–861.

86. Fredrickson D.S. Newly recognized disorders of cholesterol metabolism // Ann Intern Med. 1963. V. 58(4). P. 718.

87. Freund M.K. *et al.* Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits // Am. J. Hum. Genet. 2018. V. 103. №. 4. P. 535-552.

88. Freund N.T. *et al.* Reconstitution of the receptor-binding motif of the SARS coronavirus // Protein Eng. Des. Sel. 2015. V. 28. №. 12. P. 567–575.

89. Fry A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population // Am J Epidemiol. 2017. V. 186. P. 1026–34.

90. Fu W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants // Nature. 2013. V. 493(7431). P. 216–220.

91. Fuchsberger C. *et al.* The genetic architecture of type 2 diabetes // Nature. 2016. V. 536. P. 41–47.

92. Gao Yang *et al.* PGG.Han: The Han Chinese Genome Database and Analysis Platform // Nucleic Acids Research. 2020. V. 48(D1). P. 971–76.

93. Genome Reference Consortium. Human Genome Overview https://www.ncbi.nlm.nih.gov/grc/human. Data access: 05/27/2021

94. Giamarellos-Bourboulis E.J. *et al.* Complex immune dysregulation in COVID-19 patients with severe respiratory failure // Cell Host Microbe. 2020. V. 27(6). P. 992–1000.

95. Gibson G. Rare and common variants: twenty arguments // Nat Rev Genet. 2012. V. 18. P. 135–45.

96. Giwa A. *et al.* Identification of novel prognostic markers of survival time in high-risk neuroblastoma using gene expression profiles // Oncotarget. 2020. V. 11. P. 4293–305.

97. Glotov A.S. *et al.* Dependence between the occurrence of stable arterial hypertension in children and gene polymorphism of renin-angiotensin and kinin-bradykinin systems // Molecular Biology. - 2007. T. 41. No 1. P. 18-25.

98. Glotov A.S. *et al.* Study of molecular genetic markers of human growth // Ecological Genetics. 2012. T. 10. №. 4. P. 77-84.

99. Glotov A.S. *et al.* Identification and Analysis of Genetic Markers of Human Height // Russian Journal of Genetics: Applied Research. 2014. V. 4. №. 2. P. 98–104.

100. Glotov A.S. *et al.* Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group // Clinica Chimica Acta. 2015. V. 446. P. 132–140.

101. Glotov A.S. Genetic and environmental risk factors for gestosis in women, arterial hypertension and metabolic syndrome in children. D. in Biological Sciences. SPb. 2017. 34 p.

102. Glotov A.S. *et al.* Targeted sequencing analysis of *ACVR2A* gene identifies novel risk variants associated with preeclampsia // J. Matern. Fetal. Neonatal. Med. 2018. V. 5. P. 1-131.

103. Glotov A.S. *et al.* RNA Sequencing of Whole Blood Defines the Signature of High Intensity Exercise at Altitude in Elite Speed Skaters // Genes. 2022. V. 13(4). P. 574.

104. Glotov O.S. *et al.* Study of functionally significant polymorphism of *ACE, AGTR1, ENOS, MTHFR, MTRR* and *APOE* genes in population of North-West region of Russia // Ecological Genetics. 2004. Т. 2. №. 3. P. 32-35.

105. Glotov O.S. Analysis of cardiovascular and detoxification system gene polymorphism in different age groups of St. Petersburg. Ph.D. in Biology. SPb. 2007. 188 p.

106. Glotov O.S. *et al.* Correlation and regression analysis of athletes` complex traits, based on their personal data, genetic and biochemical parameters // Theory and Practice of Physical Culture. 2015. №. 10. P. 18.

107. Glotov O.S. *et al.* Comparative analysis of NGS and Sanger sequencing methods for HLA typing at a Russian university clinic // Cellular Therapy and Transplantation (CTT). 2018. V. 7. №. 4(25). P. 72-82.

108. Glotov O.S. *et al.* Whole-exome sequencing for monogenic diabetes in Russian children reveals wide spectrum of genetic variants in MODY-related and unrelated genes // Molecular Medicine Reports. 2019. V. 20. №. 6. P. 4905-4914.

109. Glotov O.S. *et al.* Genetic Risk Factors for the Development of COVID-19 Coronavirus Infection // Russian Journal of Genetics. 2021. V. 57. №. 8. P. 878–892. (a).

110. Glotov O.S. *et al.* The lineage of coronavirus SARS-CoV-2 of Russian origin: Genetic characteristics and correlations with clinical parameters and severity of coronavirus infection // The Siberian Journal of Clinical and Experimental Medicine. 2021. V. 36(4). P. 132–143. (б).

111. Goh L., Yap V.B. Effects of normalization on quantitative traits in association test // BMC Bioinformatics. 2009. V. 10. P. 415.

112. Goloshchapov O.V. *et al.* Bacteroides fragilis is a potential marker of effective microbiota transplantation in acute graft-versus-host disease treatment // Cell Ther Transplant. 2020. V. 9(2). P. 47-59.

113. Golubnitschaja O. *et al.* Medicine in the early twenty-first century: paradigm and anticipation - EPMA position paper 2016 // EPMA J. 2016. V. 25. №. 7(1). P. 23.

114. Gonzalez-Garay M.L. The road from next-generation sequencing to personalized medicine // Per. Med. 2014. V. 11. №. 5. P. 523–544.

115. Gra O.A. *et al.* Polymorphisms in xenobiotic-metabolizing genes and the risk of chronic lymphocytic leukemia and non-Hodgkin's lymphoma in adult Russian patients // American Journal of Hematology. 2008. V. 83(4). P. 279-287.

116. Greeley S.A. *et al.* Neonatal diabetes: An expanding list of genes allows for improved diagnosis and treatment // Curr Diab Rep. 2011. V. 11. P. 519-532.

117. Greens K. FDA OK’s 23 and Me Test // Scientists. 2015.

118. Guo Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis // Genomics. 2017. V. 109. №. 2. P. 83-90.

119. Hachiya T. *et al.* Genome-wide meta-analysis in Japanese populations identifies novel variants at the TMC6-TMC8 and SIX3-SIX2 loci associated with HbA1c // Sci. Rep. 2017. V. 7. P. 1–11.

120. Hadjinicolaou A. *et al.* De novo pathogenic variant in *SETX* causes a rapidly progressive neurodegenerative disorder of early childhood-onset with severe axonal polyneuropathy // Acta Neuropathol Commun. 2021. V. 9. P. 194.

121. Hamosh A. *et al.* Online Mendelian Inheritance in Man (OMIM®): Victor McKusick's magnum opus // American Journal of Medical Genetics. 2022. Part A. 185A. P. 3259–3265.

122. Hattersley A.T. *et al.* ISPAD clinical practice consensus guidelines 2018: The diagnosis and management of monogenic diabetes in children and adolescents // Pediatr Diabetes. 2018. V. 19 (27). P. 47-63.

123. Hiramatsu K. *et al.* Prevalence and Clinical Characteristics of Hearing Loss Caused by *MYH14* Variants // Genes (Basel). 2021. V. 12. P. 1623.

124. Hofmann A.L. *et al.* Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers // BMC Bioinformatics. 2017. V. 18. P. 8.

125. Huang C., Wang Y., Li X. Clinical features of patients infected with 2019 novel coronavirus in Wuhan. China // Lancet. 2020. V. 395(10223). P. 497–506.

126. Huang H.H., Xu T., Yang J. Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension // BMC Proc. 2014. V. 8(1). P. 96.

127. Huber O. Structure and function of desmosomal proteins and their role in development and disease // Cell Mol Life Sci. 2003. V. 60. P. 1872-1890.

128. Hurni Y. *et al.* Spontaneous resolution of nonimmune hydrops fetalis in a fetus with *TP63* gene mutation and *LZTR1* gene variants // Clin Case Reports. 2021. V. 9.

129. Inge-Vechtomov S.G. Genetics with the basics of breeding, Moscow: High School. 1998. 592 p.

130. Inge-Vechtomov S.G. Genetics with the basics of breeding. St. Petersburg: N-L Publishing House. 2010. 720 p.

131. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome // Nature. 2004. V. 431. P. 931– 945.

132. Ishigami D. *et al.* Brainstem intraparenchymal schwannoma with genetic analysis: a case report and literature review // BMC Med Genomics. 2021. V. 14. P. 205.

133. Ivashchenko T.E., Baranov V.S. Biochemical and molecular genetic basis of the pathogenesis of cystic fibrosis. SPb.: Intermedica. 2002. 256 p.

134. Ivashchenko T.E. *et al.* Molecular and genetic methods // in: Medical laboratory technology: Manual for clinical laboratory diagnostics: in 2 vol. 3rd ed., revised and supplemented. vol. 2. M.: GEOTAR-Media. 2013. P. 658-687.

135. Jiang Y. *et al.* The Effect of the Online and Offline Blended Teaching Mode on English as a Foreign Language Learners' Listening Performance in a Chinese Context // Front Psychol. 2021. V. 16 (12). P. 742742.

136. Kampinga H.H., Craig E.A. The HSP70 chaperone machinery: J proteins as drivers of functional specificity // Nat. Rev. Mol. Cell Biol. 2010. V. 11. P. 579–592.

137. Kanwal S., Perveen S., Arshad H.M. Role of Alpha-methylacyl-CoA racemase gene in pathogenecity of CMT patients // J Pak Med Assoc. 2018. V. 68. P. 1039–42.

138. Karczewsk K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes // BioRxiv. 531210. 2019.

139. Khera A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations // Nat Genet. 2018. V. 50. P. 1219–24.

140. Khusnutdinova E.K. *et al.* Restriction-deletion polymorphism of V-region of mitochondrial DNA in populations of peoples of Volga-Ural region // Genetics. 1997. T. 33. P. 996-1000.

141. Kivela P. Paradigm shift for COVID-19 response: identifying high-risk individuals and treating inflammation // West J Emerg Med. 2020. V. 21(3). P. 473–476.

142. Kleinberger J.W. and Pollin T.I. Undiagnosed MODY: Time for Action // Curr Diab Rep. 2015. V. 15. P. 110.

143. Knowler W.C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin // N Engl J Med. 2002. V. 346. P. 393–403.

144. Komissarova S.M. *et al.* The specifics of hypertrophic cardiomyopathy clinical presentation in patients with various mutations of sarcomere genes // Russian Journal of Cardiology. 2016. V. (1). P. 20-25.

145. Korber B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus // Cell. 2020. V. 182. P. 812–827.

146. Koshevaya Y.S. *et al.* Description of first registered case of the Lopes-Maciel-Rodan syndrome in Russia // Int. J. Mol. Sci. 2022. V. 23. 12437. https://doi.org/10.3390/ijms232012437

147. Kousi M., Katsanis N. Genetic modifiers and oligogenic inheritance // Cold Spring Harb Perspect Med. 2015. V. 5. № 6. P. a017145.

148. Koyama T., Parida L., Platt D.E. Variant analysis of COVID-19 genomes // Bull World Health Organ. 2020. V. 98. №. 7. P.495–504.

149. Kumaran D. *et al.* Genetic characterization of Spinocerebellar ataxia 1 in a South Indian cohort // BMC Med Genet. 2014. V. 15. P. 114.

150. Labay V. *et al.* Mutations in SLC19A2 cause thiamine-responsive megaloblastic anaemia associated with diabetes mellitus and deafness // Nat Genet. 1999. V. 22. P. 300-304.

151. Lazarin G.A. *et al.* An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: Results from an ethnically diverse clinical sample of 23,453 individuals // Genetics in Medicine. 2013. V. 15(3). P. 178– 186.

152. Lek Monkol *et al.* Analysis of Protein-Coding Genetic Variation in 60,706 Humans // Nature. 2016. V. 536 (7616). P. 285–91.

153. Lello L. *et al.* Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer // Sci Rep. 2019. V. 9. P. 15286.

154. Lemelman M.B., Letourneau L. and Greeley S. Neonatal diabetes mellitus: An update on diagnosis and management // Clin Perinatol. 2018. V. 45. P. 41-59.

155. Lerat J. *et al.* Hearing loss in inherited peripheral neuropathies: Molecular diagnosis by NGS in a French series // Mol Genet Genomic Med. 2019. V. 7.

156. Leverenz D.L., Tarrant T.K. Is the HScore useful in COVID-19? // Lancet. 2020. V. 395(10236). P. e83.

157. Li F., Li W., Farzan M., Harrison S.C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor // Science. 2005. V. 309. P. 1864–1868.

158. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics (Oxford, England). 2009. V. 25(14). P. 1754– 1760.

159. Li X. *et al.* Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan // J Allergy Clin Immunol. 2020. V. 146(1). P. 110–118.

160. Li Y. *et al.* Genome-wide association study of COVID-19 severity among the Chinese population // Cell Discov. 2021. V. 7.

161. Liang W. *et al.* Cochlear Nerve Canal Stenosis: Association with *MYH14* and *MYH9* Genes // Ear Nose Throat J. 2021. V. 100. P. 343-346.

162. Lightbody G. *et al.* Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application // Briefings in Bioinformatics. 2019. V. 20 (5). P. 1795–1811.

163. Liu C.T. *et al.* Genome-wide association of body fat distribution in African ancestry populations suggests new loci // PLoS Genet. 2013. V. 9. e1003681.

164. Liu X. *et al.* dbNSFP v3. 0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs // Human Mutation. 2016. V. 37(3). P. 235– 241.

165. Liu Z. *et al.* Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2 // J. Med. Virol. 2020. V. 92. №. 6. P. 595–601.

166. Lohmueller K.E. *et al.* Whole-exome sequencing of 2000 Danish individuals and the role of rare coding variants in type 2 diabetes // Am. J. Hum. Genet. 2013. P. 1072–1086.

167. Lohse P. *et al.* Compound heterozygosity for a Wolman mutation is frequent among patients with cholesteryl ester storage disease // J Lipid Res. 2000. V. 41(1). P. 23–31.

168. Lopes F. *et al.* Identification of novel genetic causes of Rett syndrome-like phenotypes // J. Med Genet. 2016. V. 53. P. 190.

169. Lvovs D., Favorova O.O., Favorov A.V. A Polygenic Approach to the Study of Polygenic Diseases // Acta Naturae. 2012. V.4. №. 3. P. 59–71.

170. Lyazina L.V. et al. Possibilities of medical care in modern conditions on the example of a family with hereditary pathology // Medical Genetics. 2017. T. 16. №. 10. P. 51-54.

171. MacDonald *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes // Cell. 1993. V. 72. P. 971–983.

172. Macfarlan T. *et al.* Human THAP7 Is a Chromatin-associated, Histone Tail-binding Protein That Represses Transcription via Recruitment of HDAC3 and Nuclear Hormone Receptor Corepressor // J Biol Chem. 2005. V. 280. P. 7346–58.

173. Mackay I.M., Arden K.E. MERS coronavirus: diagnostics, epidemiology and transmission // Virol J. 2015. V. 12. P. 222.

174. Mahajan A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article // Nat. Genet. 2018. V. 50. P. 559–571.

175. Mahmud S. *et al.* Whole Exome Sequence Study of Mild Cognitive Impairment in African and European Americans; the Atherosclerosis Risk in Communities-Neurocognitive Study // Alzheimer's Dement. 2021. V. 17.

176. Majewski J., Schwartzentruber J., Lalonde E. What can exome sequencing do for you? // J. Med. Genet. 2011. V. 48. №. 9. P. 580–589.

177. Mandelstam M.Y. *et al.* Molecular genetics of familial hypercholesterolemia: current status of the issue in Russia // Pacific Medical Journal. 2002. №. 1(8). P. 10-11.

178. Mannucci P.M., Duga S., Peyvandi F. Recessively inherited coagulation disorders // Blood. 2004. V. 104(5). P. 1243–1253.

179. Manolio T.A. *et al.* Finding the missing heritability of complex diseases // Nature. 2010. V. 461. P. 747-753.

180. Marciniuk D.D. *et al.* Alpha-1 antitrypsin deficiency targeted testing and augmentation therapy: a Canadian Thoracic Society clinical practice guideline // Can Respir J. 2012. V. 19. P. 109–16.

181. Marino P. *et al.* Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study // Europ. J. of Human Genetics. 2018. V. 26. №. 3. P. 314-323.

182. Martin A.R. *et al.* The Critical Needs and Challenges for Genetic Architecture Studies in Africa // Current Opinion in Genetics & Development. 2018. V. 53. P. 113–20.

183. Maslennikov A.B. Relationship of allelic variants of *AROA1, ARO, AROS1* genes, atherogenic dyslipoproteidemia and complicated myocardial infarction. Doct. Diss. Tomsk. 1999. 26 p.

184. McCullagh P., Nelder J.A. Generalized Linear Models, Second Edition // Boca Raton: Chapman and Hall/CRC. 1989. 532 p.

185. McGonagle D. *et al.* The role of cytokines including interleukin-6 in COVID-19 induced pneumonia and macrophage activation syndrome-like disease // Autoimmun Rev. 2020. V. 19(6). P. 102537.

186. Mcinnes G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics // Bioinformatics. 2018. December: 1–3.

187. Mehdi A.M. *et al.* A peripheral blood transcriptomic signature predicts autoantibody development in infants at risk of type 1 diabetes // JCI Insight. 2018. V. 3. e98212.

188. Millat G., Chanavat V., Rousson R. Evaluation of a new NGS method based on a custom AmpliSeq library and Ion Torrent PGM sequencing for the fast detection of genetic variations in cardiomyopathies // Clin Chim Acta. 2014. V. 433. P. 266–71.

189. Miroshnikova V.V. *et al.* Identification of novel variants in the *LDLR* gene in Russian patients with familial hypercholesterolemia using targeted sequencing // Biomedical Reports. 2021. V. 14 (1). P. 15.

190. Miyazawa A. *et al.* A preliminary genetic association study of *GAD1* and *GABAB* receptor genes in patients with treatment-resistant schizophrenia // Mol Biol Rep. 2021.

191. Mohnike K. *et al.* Clinical and genetic evaluation of patients with *KATP* channel mutations from the German registry for congenital hyperinsulinism // Horm Res Paediatr. 2014. V. 81. P. 156-168.

192. Moore J., June C. Cytokine release syndrome in severe COVID-19 // Science. 2020. V. 368(6490). P. 473–474.

193. Morgant S. *et al.* Chapter 8. Role of Next-Generation Sequencing Technologies in Personalized Medicine // P5 eHealth: An Agenda for the Health Technologies of the Future // Eds. by G. Pravettoni, S. Triberti. 2020. P. 125-145.

194. Mousa M. *et al.* Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates // eBioMedicine. 2021. V. 74. P. 103695.

195. Munne S. *et al.* Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing // Fertil Steril. 2017. V. 108(1). P. 62-71.

196. Muntoni S. *et al.* Prevalence of cholesteryl ester storage disease // Arterioscler Thromb Vasc Biol. 2007. V. 27(8). P. 1866–1868.

197. Mustacich D.J. *et al.* Digenic Inheritance of a *FOXC2* Mutation and Two *PIEZO1* Mutations Underlies Congenital Lymphedema in a Multigeneration Family // Am J Med. 2021.

198. Nebert D.W., Carvan M.J. Ecogenetics: from Biology to Health // Toxicol.Industr. Health. 1997. V. 13. P. 163-192.

199. Ng S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder // Nat Genet. 2010. V. 42(1). P. 30-5.

200. Ng S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes // Nature. 2009. V. 461(7261). P. 272-6.

201. Nguyen-Ngoc K.V. *et al.* Mosaic loss of non-muscle myosin IIA and IIB is sufficient to induce mammary epithelial proliferation // J Cell Sci. 2017.

202. Nykamp Keith *et al.* Sherloc: A Comprehensive Refinement of the ACMG–AMP Variant Classification Criteria // Genetics in Medicine. 2017. V. 19(10). P.1105–17.

203. O’Donovan M.C. What have we learned from the Psychiatric Genomics Consortium // World Psychiatry. 2015. V. 14. №. 3. P. 291–293.

204. Ohni S. *et al.* Direct molecular evidence for both multicentric and monoclonal carcinogenesis followed by transdifferentiation from hepatocellular carcinoma to cholangiocarcinoma in a case of metachronous liver cancer // Oncol Lett. 2021. V. 23. P. 22.

205. Oleksyk T., Brukhin V., O’Brien S.J. The Genome Russia Project: Closing the Largest Remaining Omission on the World Genome Map // GigaScience. 2015. V. 4(1). P. 53.

206. Oteva E.A., Maslennikov A.B., Nikolaeva A.A. Accelerated development of atherosclerosis // Phys. 1994. №. 3. P. 50-52.

207. PANGO lineages. https://cov-lineages.org/lineage_designation.html

208. Peltonen L., McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era // Science. 2001. V. 16. №. 291(5507). P. 1224-9.

209. Pendina A.A. *et al.* Reproductive History of a Woman With 8p and 18p Genetic Imbalance and Minor Phenotypic Abnormalities // Front. Genet. 2019. V. 10. P. 1164.

210. Pérez-Valencia J.A. *et al.* Angiogenesis and evading immune destruction are the main related transcriptomic characteristics to the invasive process of oral tongue cancer // Sci Rep. 2018. V. 8. P. 2007.

211. Perez-Becerril C., Evans D.G., Smith M.J. Pathogenic noncoding variants in the neurofibromatosis and schwannomatosis predisposition genes // Hum Mutat. 2021. V. 42. P. 1187–207.

212. Petersen I. Classification and Treatment of Diseases in the Age of Genome Medicine Based on Pathway Pathology // Int J Mol Sci. 202.1 V. 30. №. 22(17). P. 9418.

213. Pierik M. *et al.* The IBD international genetics consortium provides further evidence for linkage to IBD4 and shows gene-environment interaction // Inflammatory Bowel Diseases. 2005. V. 11. №.1. P. 1–7.

214. Piotrowski A. *et al.* Targeted massively parallel sequencing of candidate regions on chromosome 22q predisposing to multiple schwannomas: An analysis of 51 individuals in a single-center experience // Hum Mutat. 2022. V. 43. P. 74–84.

215. Poletaev A.B., Churilov L.P. Immunology of Health and Disease: Simple Answers to Difficult Questions. SPB: Foliant. 2021. 264 p.

216. Povysil G. *et al.* Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19 // J Clin Invest. 2021. 131.

217. Prakrithi P. *et al.* Genetic Risk Prediction of COVID-19 Susceptibility and Severity in the Indian Population // Frontiers in Genetics. 2021. V. 12.

218. Pritchard J.K., Cox N.J. The allelic architecture of human disease genes: common disease-common variant...or not? // Hum Mol Genet. 2002. V. 11(20). P. 2417-23.

219. Provisional guidelines: prevention, diagnosis and treatment of new coronavirus infection (COVID-19). Version 12, M, September 21, 2021. M.: 2021. 231 p.

220. Puthucheary Z. *et al.* The *ACE* gene and human performance: 12 years on // Sports Med. 2011. V. 41(6). P. 433-48.

221. Puzyrev V.P., Freidin M.B., Kucher A.N. Genetic diversity of population and human diseases. Tomsk: Printed Literature. 2007. 319 p.

222. Puzyrev V.P. Phenomic-genomic relationships and pathogenetics of multifactorial diseases // Vest. RAMS. 2011. T. 9. P. 17-27.

223. Qi F. *et al.* Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses // Biochem. Biophys. Res. Commun. 2020. V. 526. №. 1. P. 135–140.

224. Rabbani B. *et al.* Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders // J. Hum. Genet. 2012. V. 57. P. 621-632.

225. Rahman M.H. *et al.* A Network-Based Bioinformatics Approach to Identify Molecular Biomarkers for Type 2 Diabetes that Are Linked to the Progression of Neurological Diseases // Int J Environ Res Public Health. 2020. V.17. P. 1035.

226. Ramensky V.E. *et al.* 2021. Targeted Sequencing of 242 Clinically Important Genes in the Russian Population from the Ivanovo Region // Frontiers in Genetics. V. 12. 709419.

227. Rebrova O.Y. Statistical analysis of medical data. Application of applied software package STATISTICA // M. MediaSphere. 2003. 312 p.

228. Reich D.E., Lander E.S. On the allelic spec- trum of human disease // Trends Genet. 2001. V. 17. №. 9. P. 502–510.

229. Reid S. *et al.* High genetic risk score is associated with early disease onset, damage accrual and decreased survival in systemic lupus erythematosus // Ann Rheum Dis. 2020. V. 79. P. 363–9.

230. Richards S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology // Genet Med. 2015. V. 17. №. 5. P. 405–424.

231. Robinson J.G. *et al.* Efficacy and safety of alirocumab in reducing lipids and cardiovascular events // N Engl J Med. 2015. V. 372. P. 1489–99.

232. Rodan L.H. *et al.* A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene // Eur. J. Hum. Genet. 2016. V. 24. P. 1826–1827.

233. Rodriguez-Flores J.L. *et al.* Exome sequencing identifies potential risk variants for Mendelian disorders at high prevalence in Qatar // Human Mutation. 2014. V. 35(1). P. 105– 116.

234. Rossi Á.D. *et al.* Association between *ACE2* and *TMPRSS2* nasopharyngeal expression and COVID-19 respiratory distress // Sci Rep. 2021. V. 11. P. 9658.

235. Rubio-Cabezas O. *et al:* Wolcott-Rallison syndrome is the most common genetic cause of permanent neonatal diabetes in consanguineous families // J Clin Endocrinol. Metab. 2009. V. 94. P. 4162-4170.

236. Ryzhkova O.P. *et al.* Guidelines for the interpretation of data obtained by massively parallel sequencing (MPS). Medical Genetics. 2017. T. 16(7). P. 4-17.

237. Ryzhkova O.P. *et al.* Guidelines for interpretation of human DNA sequence data obtained by massively parallel sequencing (MPS) (revision 2018, version 2). Medical Genetics. 2019. T. 18(2). P. 3-23.

238. Sabino E.C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence // The Lancet. 2021. V. 397. P. 452-455.

239. Sachdev N.M. *et al.* The rate of mosaic embryos from donor egg as detected by next generation sequencing (NGS) varies by IVF laboratory // Fertil Steril. 2016. V. 106(3). e156–7.

240. Saifitdinova A.F. *et al.* Mosaicism in preimplantation human embryos // Integrative Physiology. 2020. V. 1. №. 3. P. 225–230.

241. Saleem I.B. *et al.* Identification and Computational Analysis of Rare Variants of Known Hearing Loss Genes Present in Five Deaf Members of a Pakistani Kindred // Genes (Basel). 2021. V.12. P. 1940.

242. Schidtke P. SARS-CoV-2 - part 2 - From the viral genome to protein structures. MARCH 27, 2020. https://www.discngine.com/blog?author=52850d39e4b0b817d0c61ff9

243. Schmidt B., Hildebrandt A. Next-generation sequencing: big data meets high performance computing // Drug Discov Today. 2017. V. 22(4). P. 712-717.

244. Scott R.A. *et al.* An expanded genome-wide association study of type 2 diabetes in Europeans // Diabetes. 2018. V. 66. P. 2888–2902.

245. Shaw-Smith C. *et al.* Recessive *SLC19A2* mutations are a cause of neonatal diabetes mellitus in thiamine-responsive megaloblastic anaemia // Pediatr Diabetes. 2012. V. 13. P. 314-321.

246. Shcherbak S.G. *et al.* Basic Predictive Risk Factors for Cytokine Storms in COVID-19 Patients // Front. Immunol. 2021. V. 12. P. 745515.

247. Shcherbak S.G. *et al.* Identification of Genetic Risk Factors of Severe COVID-19 Using Extensive Phenotypic Data: A Proof-of-Concept Study in a Cohort of Russian Patients // Genes. 2022. V. 13(3). P. 534.

248. Shelton J.F. *et al.* The UGT2A1/UGT2A2 locus is associated with COVID-19-related loss of smell or taste // Nat Genet. 2022.

249. Sheremet N.L. *et al.* Molecular genetic diagnosis of Stargardt disease // Vestnik Oftalmologii. 2017. V. 133(4). P. 4– 11.

250. Shields B.M. *et al.* Maturity-onset diabetes of the young (MODY): How many cases are we missing? // Diabetologia. 2010. V. 53. P. 2504-2508.

251. Shikov A.E. *et al.* Application of bioinformatics in the analysis of clinical data // Molecular biological technology in medical practice / Ed. by A.B. Maslennikov, Corresponding Member of RAAS. Vol. 29. Novosibirsk: Akademizdat. 2019. P. 119-136.

252. Shikov A.E. *et al.* Analysis of the Spectrum of *ACE2* Variation Suggests a Possible Influence of Rare and Common Variants on Susceptibility to COVID-19 and Severity of Outcome // Front. Genet. 2020. V. 11. P. 551220.

253. Shikov A. *et al.* The application of Nanopore sequencing for variant calling on the human mitochondrial DNA // Bio. Comm. 2021. V. 66(2). P. 109–123.

254. Shitao R.A.O., Alexandria L.A.U., Hon-Cheong S.O. Exploring diseases/traits and blood proteins causally related to expression of *ACE2*, the putative receptor of 2019-nCov: A Mendelian Randomization analysis // Diabetes Care. 2020. V. 43. №. 7. P. 1416–1426.

255. Shulla A. *et al.* Transmembrane Serine Protease Is Linked to the Severe Acute Respiratory Syndrome Coronavirus Receptor and Activates Virus Entry // Journal of Virology. 2011. V. 85. №. 2. P. 873–882.

256. Shungin D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution // Nature. 2015. V. 518. P. 187–196.

257. Silverman E.K., Sandhaus RA. Clinical practice. Alpha1-antitrypsin deficiency. // N Engl J Med. 2009. V. 360. P. 2749–57.

258. Singer M., Berg P. Genes and genomes: In 2 volumes. Volume 1. Translated from English: Mir. 1998. 391 p.

259. Skeppholm M. *et al.* ADAMTS13 and von Willebrand factor concentrations in patients with diabetes mellitus // Blood Coagul. Fibrinol. 2009. V. 20. P. 619–626.

260. Sladek R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes // Nature. 2007. V. 445. P. 881-885.

261. Stankov K., Benc D., Draskovic D. Genetic and epigenetic factors in etiology of diabetes mellitus type 1 // Pediatrics. 2013. V. 132. №. 6. P. 1112-22.

262. Stehouwer C.D. *et al.* Increased urinary albumin excretion, endothelial dysfunction, and chronic low-grade inflammation in type 2 diabetes: Progressive, interrelated, and independently associated with risk of death // Diabetes. 2002. V. 51. P. 1157–1165.

263. Stepanov V.A. Ethnogenomics of the population of Northern Eurasia // Tomsk: Publishing house "Pechatnaya Manufaktura". 2002. 244 p.

264. Stepanov V.A., Puzyrev V.P. Analysis of allelic frequencies of seven microsatellite loci of Y-chromosome in three populations of Tuvinians // Genetics. 2000. T. 36. P. 241-248.

265. Strafella C. *et al.* Analysis of *ACE2* genetic variability among populations highlights a possible link with COVID19-related neurological complications // Genes. 2020. V. 11 P. 741.

266. Strokova T.V., Bagaeva M.E., Matinyan IA. Defitsit lizosomnoi kisloi lipazy // Russkii meditsinskii zhurnal. 2017. V. 25(19). P. 1346–1351.

267. Sturm A.C. *et al.* Clinical genetic testing for familial hypercholesterolemia: JACC scientific expert panel // J Am Coll Cardiol. 2018. V. 72. P. 662–80.

268. Suh S. *et al.* A systematic review on papers that study on Single Nucleotide Polymorphism that affects coronavirus 2019 severity // BMC Infect Dis. 2022. V. 22. P. 1–11.

269. Surcel A. *et al.* Targeting Mechanoresponsive Proteins in Pancreatic Cancer: 4-Hydroxyacetophenone Blocks Dissemination and Invasion by Activating MYH14 // Cancer Res. 2019. V. 79. P. 4665.

270. Suwinski P. *et al.* Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics // Front. Genet. 2019. V. 10. P. 49.

271. Taitz B.M. Practical predictive, preventive and personalized medicine. "10P medicine in addressing issues of prevention, active longevity, reducing mortality and increasing life expectancy of the population // Ministry of Health of the Russian Federation, I.I. Mechnikov Northwestern State Medical University. Saint Petersburg: IPK Beresta. 2019. 380 p.

272. Talley M.J. *et al.* Generation of a Mouse Model to Study the Noonan Syndrome Gene Lztr1 in the Telencephalon // Front Cell Dev Biol. 2021. V. 9.

273. Tan O. *et al.* Application of next-generation sequencing to improve cancer management: A review of the clinical effectiveness and cost effectiveness // Clinical Genetics. 2018. V. 93. № 3. P. 533–544.

274. Tarkovskaya I.V. *et al.* Analysis of the association of lipid metabolism gene polymorphism with body mass index, waist circumference and blood lipidogram parameters in women. // Ecological Genetics. 2012. T. 10. №. 4. P. 66-77.

275. Teekakirikul P. *et al.* Inheritedcardiomyopathies: molecular genetics and clinical genetic testing in the postgenomic era // J Mol Diagn. 2013. V.15(2). P. 158–70.

276. Temporary guidelines: prevention, diagnosis and treatment of new coronavirus infection (COVID-19). Version 12, M, 09/21/2021. M.: 2021. 231 p.

277. Tighe O. *et al.* Genetic diversity within the R408W phenylketonuria mutation lineages in Europe // Human Mutation. 2003. V. 21(4). P. 387– 393.

278. Toovey O.R. *et al.* Introduction of Brazilian SARS-CoV-2 484K.V2 related variants into the UK // J Infect. 2021. V. 82(5). e23-e24.

279. Tulzunovskaya I.G. *et al.* Wilson-Conovalov disease: intrafamilial clinical polymorphism // PEDIATRIA. JOURNAL. G.N. SPERANSKY. 2017. T. 96. №. 6. P. 215-216.

280. Turkunova M.E. *et al.* Molecular Genetics and Pathogenesis of the Floating Harbor Syndrome: Case Report of Long-Term Growth Hormone Treatment and a Literature Review // Front. Genet. 2022. V. 13. P. 846101.

281. Van der Graaf A. *et al.* Molecular basis of autosomal dominant hypercholesterolemia: Assessment in a large cohort of hypercho- lesterolemic children // Circulation. 2011. V. 123. P. 1167-1173.

282. van Moorsel CHM. *et al.* The *MUC5B* Promoter Polymorphism Associates with Severe COVID-19 in the European Population // Front Med. 2021. V. 8.

283. Vlasov V.V. Epidemiology - Moscow: GEOTAR-MED. 2004. 464 p.

284. Wallace S.E., Bird T.D. Molecular genetic testing for hereditary ataxia // Neurol Clin Pract. 2018. V. 8. P. 27–32.

285. Walter K. *et al.* The UK10K project identifies rare variants in health and disease // Nature. 2015. V. 526(7571). P. 82– 89.

286. Wang M. *et al.* A novel MYH14 mutation in a Chinese family with autosomal dominant nonsyndromic hearing loss // BMC Med Genet. 2020. V. 21. P. 154.

287. Wang Q. *et al.* Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity // Sci. Rep. 2017. V. 7. №. 1. P. 885.

288. Wang R. *et al.* Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants // Commun Biol. 2021. V. 4(1). P. 228.

289. Wang X. *et al.* Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction // J. Diabetes. 2016. V. 8. P. 24–35.

290. Weedon M.N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population // Nat Genet. 2007. V. 39(10). P. 1245-50.

291. Weissman A. *et al.* Chromosomal mosaicism detected during preimplantation genetic screening: results of a worldwide Web-based survey // Fertil Steril. 2017. V. 107(5). P. 1092–7.

292. Wiegman A. *et al.* European atherosclerosis society consensus panel. Familial hypercholester- olaemia in children and adolescents: Gaining decades of life by optimizing detection and treatment // Eur Heart J. 2015. V. 36. P. 2425-2437.

293. Wong K.H.Y. *et al.* Towards a Reference Genome That Captures Global Genetic Diversity // Nature Communications. 2020. V. 11 (1). P. 5482.

294. Wood D., De Backer G., Faergeman O. Prevention of Coronary Heart Disease in Clinical Practice. Recommendations of the Second Joint Task Force of the European and other Societies on Coronary Prevention // Eur Heart J. 1998. №. 19. P. 1434–1503.

295. Wooster R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2* // Nature. 1995. V. 378. P. 789–92.

296. World Health Organization. SARS-CoV-2 Variants https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/

297. Wu B.B. *et al.* Association between ABO blood groups and COVID-19 infection, severity and demise: A systematic review and meta-analysis // Infect. Genet. Evol. 2020. V. 84. P. 104485.

298. Wu P. *et al.* Trans-ethnic genome-wide association study of severe COVID-19 // Communications Biology. 2021. V. 4. P. 1034.

299. Wulff K., Herrmann F.H. Twentytwo novel mutations of the factor VII gene in factor VII deficiency // Human Mutation. 2000. V. 15(6). P. 489– 496.

300. Xu S., Hu Z. Generalized Linear Model for Interval Mapping of Quantitative Trait Loci // Theor. Appl. Genet. 2010. V. 121. №. 1. P. 47–63.

301. Yalcintepe S. *et al.* The importance of multiple gene analysis for diagnosis and differential diagnosis in Charcot Marie tooth disease // Turk Neurosurg. 2021.

302. Yan R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2 // Science. 2020. V. 367. P. 1444–1448.

303. Yang X. *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study // Lancet Respir Med. 2020. V. 8. №. 5. P. 475–481.

304. Yengo L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry // bioRxiv. 2018. 274654.

305. Yi N., Banerjee S. Hierarchical generalized linear models for multiple quantitative trait locus mapping // Genetics. 2009. V. 181. №. 3. P. 1101–13.

306. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications // Genomics. 2020. V. 112. №. 5. P. 3588–96.

307. Zabetian C.P. *et al.* A revised allele frequency estimate and haplotype analysis of the DBH deficiency mutation IVS1+2T->C in African- and European-Americans // Am. J. Med. Genet. Part A. 2003. V. 123. P. 190–192.

308. Zachariah P. *et al.* Epidemiology, clinical features, and disease severity in patients with coronavirus disease 2019 (COVID-19) in a children's hospital in New York City, New York // JAMA Pediatr. 2020. V. 174(10). e202430.

309. Zhang H. *et al.* LZTR1: A promising adaptor of the CUL3 family (Review) // Oncol Lett. 2021. V. 22. P. 564.

310. Zhang W. *et al.* Emergence of a Novel SARS-CoV-2 Variant in Southern California // JAMA Pediatr. 2021. V. 325(13). P. 1324-1326.

311. Zhang Z. *et al.* Detection of *PKD1* and *PKD2* Somatic Variants in Autosomal Dominant Polycystic Kidney Cyst Epithelial Cells by Whole-Genome Sequencing // J Am Soc Nephrol. 2021. V. 32. P. 3114–29.

312. Zhao S. *et al.* Pilot study of expanded carrier screening for 11 recessive diseases in China: Results from 10,476 ethnically diverse couples // European Journal of Human Genetics. 2019. V. 27(2). P. 254– 262.

313. Zhernakova D.V. *et al.* Analytical "bake-off" of whole genome sequencing quality for the genome Russia project using a small cohort for autoimmune hepatitis // PLoS ONE. 2018. V. 13(7). P. 1– 18.

314. Zhernakova D.V. *et al.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. 2020. V. 112. №. 1. P. 442-458.

315. Zhou F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study // Lancet. 2020. V. 395(10229). P. 1054–1062.

316. Zhou Z. *et al.* Loss-of-Function Piezo1 Mutations Display Altered Stability Driven by Ubiquitination and Proteasomal Degradation // Front Pharmacol. 2021. V. 12.

317. Zou Y. *et al.* Multiple gene mutations, not the type of mutation, are the modifier of left ventricle hypertrophy in patients with hypertrophic cardiomyopathy // Mol Biol Rep. 2013. V. 40(6). P. 3969–76.

318. http://www.biometrica.tomsk.ru

319. https://www.weforum.org/agenda

320. https://www.omim.org/statistics/geneMap

321. https://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array

322. http://www.internationalgenome.org/

323. https://genomics.ut.ee/en

324. https://mutalyzer.nl.

325. http://www.ncbi.nlm.nih.gov/RefSeq/

326. http://www.lrg-sequence.org

327. https://www.ebi.ac.uk/gwas

328. https://phgkb.cdc.gov/PHGKB/hNHome.acton

329. http://www.hgmd.cf.ac.uk/ac/index.php

330. http://www.ncbi.nlm.nih.gov/snp

331. http://www.hgvs.org/dblist/dblist.html

332. http://www.lovd.nl

333. https://decipher.sanger.ac.uk

334. http://browser.1000genomes.org/index.html

335. https://gnomad.broadinstitute.org/

336. https://evs.gs.washington.edu/EVS/

337. http://exac.broadinstitute.org

338. http://gnomad.broadinstitute.org
339. https://allofus.nih.gov/news-events
340. https://www.genomicsengland.co.uk
341. https://coronavirus.jhu.edu

ACKNOWLEDGEMENTS

The author expresses his deep gratitude to his mentors, research supervisors and consultants: Vladislav S. Baranov and Tatiana E. Ivashchenko who have regrettably and untimely passed away in 2022 for without them this dissertation would hardly ever come to exist. I would like to express my gratitude to all the employees of the D.O. Ott Research Institute of Obstetrics, Gynecology and Reproduction who directly or indirectly helped to shape, conceptualize and develop the dissertation. My special gratitude goes to the Head of the Genomics Department Andrey S. Glotov, Senior Research Fellow Mikhail V. Aseev, Senior Research Fellow Anton V. Kiselev, Research Fellow Yuri A. Barbitov, and Senior Research Fellow Anna A. Pendina for her question that allowed us to better formulate the dissertation objective, as well as the Heads of the Institute Director Igor Yu. Kogan and Deputy Director for Science and Research Olesya N. Bespalova. I am glad to thank all the members of the Department of Genetics at St. Petersburg State University for their contribution to my education and my scientific up-bringing. I would like to pass over my gratitude to the members of the Genetics Laboratory and the Virology Center at St. Petersburg Hospital No. 40, who in 2013 to 2021 helped me to effectively fight COVID-19, as well as do remarkable research amid the pandemic. I am glad to thank Natalia V/ Skripchenko, Deputy Director for Science, and my colleagues from the Research Laboratory of Experimental Medical Virology, Molecular Genetics and Biobanking at the t. Petersburh State Medical Referral and Diagnostics Center for Children: Alexey B. Chukhlovin, Olga V. Goleva and Yury A. Eismont for valuable advice regarding the dissertation. The author is pleased to express his gratitude to the colleagues whose long-term cooperation has contributed significantly to the research and publication of papers in the high-rank periodicals, which constitute the framework of the dissertation – to all the above-mentioned colleagues, as well as A.N. Chernov, D.E. Polev, M.A. Fedyakov, A.V. Predeus, A.E. Shikov, V.V. Tsai, R.S. Kalinin, M.V. Moskalenko, E.B. Bashnina, I.V. Polyakova.

In conclusion the author expresses his deep gratitude to all his relatives and friends, namely to his parents - Sergey A. Glotov, Olga V. Glotova, wife Valeria Yu. Gorbacheva, father-in-law Yuri Ye. Gorbachev, children Vitaly and Uliana Glotov, and brother Andrey S. Glotov for their patience, spiritual support and a favorable positive environment while working on the study for many years.