

На правах рукописи

БУГ ДМИТРИЙ СЕРГЕЕВИЧ

**СТРАТИФИКАЦИЯ РИСКОВ ТРАНСФОРМАЦИИ В ОСТРЫЙ ЛЕЙКОЗ
И СМЕРТИ У ПАЦИЕНТОВ С МИЕЛОДИСПЛАСТИЧЕСКИМ
СИНДРОМОМ НА ОСНОВАНИИ ДАННЫХ МУТАЦИОННОГО
АНАЛИЗА**

3.3.9. Медицинская информатика (медицинские науки)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата медицинских наук

Санкт-Петербург – 2025

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова» Министерства здравоохранения Российской Федерации

Научный руководитель:

доктор медицинских наук, доцент **Наркевич Артём Николаевич**

Официальные оппоненты:

доктор медицинских наук **Владзимирский Антон Вячеславович**

Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», заместитель директора по научной работе

доктор медицинских наук **Семочкин Сергей Вячеславович**

Московский научно-исследовательский онкологический институт имени П.А. Герцена – филиал федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр радиологии» Министерства здравоохранения Российской Федерации, заведующий группой высокодозной химиотерапии и трансплантации костного мозга отдела лекарственного лечения опухолей

Ведущая организация:

федеральное государственное бюджетное образовательное учреждение высшего образования «Самарский государственный медицинский университет» Министерства здравоохранения Российской Федерации

Защита диссертации состоится «09» декабря 2025 года в 14.00 часов на заседании объединенного диссертационного совета 99.0.122.02 на базе ГНЦ РФ – ИМБП РАН и ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России (Пироговский Университет) по адресу: 117513, г. Москва, ул. Островитянова, д. 1, стр. 6

С диссертацией можно ознакомиться на сайте <http://rsmu.ru> и в научной библиотеке ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России (Пироговский Университет) по адресу: 117513, г. Москва, ул. Островитянова, д. 1, стр. 5

Автореферат разослан «____» _____ 20__ г.

Ученый секретарь

диссертационного совета,

доктор медицинских наук, доцент



Антонов Андрей Евгеньевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В настоящее время для выбора тактики ведения пациентов с онкогематологическими, нейроонкологическими и другими заболеваниями необходимо применять молекулярно-генетические методы, которые позволяют обнаруживать нарушения дезоксирибонуклеиновой кислоты (ДНК) на геномном, хромосомном и генном уровнях (M. S. Tallman et al., 2019; B. A. Orr, 2020). Технология секвенирования следующего поколения (NGS) отличается от классического секвенирования по Сэнгеру значительным увеличением пропускной способности метода, позволяя обнаруживать большее число нарушений ДНК и использовать дополнительные сведения при выборе тактики ведения пациентов (J. Gagan, E. M. Van Allen, 2015). С внедрением метода NGS в клиническую практику объемы данных мутационного анализа ДНК пациентов с различными патологиями значительно увеличились. При этом выросло и число находок, клиническое значение (клинический эффект) которых остаётся неясным (C. E. Cook et al., 2015). В частности, большинство генетических вариантов, обнаруживаемых в некодирующих областях генома, являются неинтерпретируемыми (варианты неясного значения) (G. Federici, S. Soddu, 2020). В то же время, согласно базе данных ClinVar, около 75% генетических вариантов неясного значения представлено миссенс-заменами (S. Iqbal et al., 2020). В структуре генетических вариантов, обнаруженных при анализе опухолевых тканей, миссенс-замены также составляют большинство (Y. Sharma et al., 2019). Таким образом, наиболее остро проявляется проблема определения клинической значимости генетических вариантов в некодирующих областях генома и миссенс-замен.

Миелодиспластический синдром (МДС) — это группа заболеваний с поражением стволовой клетки крови, характеризующихся цитопенией и риском трансформации в острый миелоидный лейкоз (ОМЛ). Спектр возможных опций при определении тактики ведения пациентов варьирует от динамического наблюдения до системной противоопухолевой терапии с возможностью выполнения трансплантации гемопоэтических стволовых клеток. Для того чтобы определить, какой метод будет наиболее эффективным, используются специальные шкалы, позволяющие оценить индивидуальный риск развития осложнений: IPSS (P. L. Greenberg et al., 1997), WPSS (E. P. Alessandrino et al., 2008), IPSS-R (P. L. Greenberg et al., 2012) и другие. Эти шкалы учитывают выраженность цитопении, уровень бластных клеток в костном мозге, а также наличие геномных и хромосомных нарушений ДНК. По мере прогрессирования заболевания у пациентов с МДС повышается частота возникновения мутаций в гене *TP53*, что ассоциировано с увеличенным риском развития осложнений и смерти (E. Bernard et al., 2022, N. G. Daver et al., 2022). В целом, мутации в этом

гене наиболее часто встречаются при злокачественных новообразованиях и связаны с неблагоприятным прогнозом (Н. Ю. Цветков и др., 2020), а у пациентов с МДС и ОМЛ они возникают в 5–10% случаев и приводят к ухудшению ответа на химиотерапию и снижению выживаемости (N. G. Daver et al., 2022). Следует подчеркнуть, что это касается только патогенных мутаций, что обуславливает необходимость определения клинического эффекта генетических вариантов *TP53*, многие из которых попадают в категорию неясного значения. Таким образом, на сегодняшний день существует проблема определения патогенности генетических вариантов неясного значения в гене *TP53*, которая не позволяет корректно оценивать риски трансформации в ОМЛ и смерти у пациентов с МДС с учётом факторов, выявляемых на молекулярно-генетическом уровне.

Степень разработанности темы исследования. В настоящее время для оценки клинической значимости генетического варианта используются методы *in vivo*, *in vitro* или *in silico*. К методам *in silico* относится определение клинической значимости генетических вариантов при помощи программ-предикторов, опирающихся на эволюционную консервативность гена и структурные характеристики продуцируемого им белка. Данные программы рекомендуются к использованию при обнаружении герминальных и соматических мутаций (P. Horak et al., 2022, M. M. Li et al., 2016, S. Richards et al., 2015). Согласно «Руководству по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования (MPS)» рекомендуется использовать компьютерные прогностические программы в случае, если генетический вариант не был описан ранее и не представлен ни в одной из баз данных или сведения о нём недостаточны (О. П. Рыжкова и др., 2020). То же касается и соматических мутаций, обнаруженных у пациентов с онкологическими заболеваниями (М. А. Спектор и др., 2020). Внедрение стандарта оценки клинического значения генетических вариантов позволяет реклассифицировать их, уточняя представление о пациенте с учётом новых данных, а также скорректировать тактику ведения больного (J. Mersch et al., 2018).

Одним из основных принципов оценки клинической значимости *in silico* является поиск обнаруженных у пациента генетических вариантов среди последовательностей-гомологов изменённого гена. Он лежит в основе ConSurf (H. Ashkenazy et al., 2016), EVE (J. Frazer et al., 2021), PROVEAN (Y. Choi, A. P. Chan, 2015) и других программ. В AlphaMissense и других программах используется дополнительная информация о структуре и функции белкового продукта исследуемого гена (J. Cheng et al., 2023). «Ансамблевые» программы (например, APF2) осуществляют оценку патогенности генетических вариантов

на основе результатов работы других программ-предикторов (Y. Zhou et al., 2024).

Однако, из-за недостаточной специфичности результаты работы прогностических программ могут использоваться для применения лишь дополнительных критериев для определения нейтральных (критерий ВР4) и патогенных (критерий РР3) вариантов (О. П. Рыжкова и др., 2020). При этом патогенность генетического варианта может подтверждаться только при условии его соответствия другим критериям, которые предполагают проведение функциональных и клинических исследований. Следовательно, оценка патогенности генетического варианта, полученная с помощью программ-предикторов, не может использоваться в качестве основополагающего критерия для классификации генетических вариантов согласно общепринятым руководствам для интерпретации данных секвенирования. Более того, программы-предикторы ограничиваются классификацией вариантов в целом, без учёта нозологии и персонализированного подхода. Кроме того, во многих программах не осуществляется обновление информации о характеристиках патогенных вариантов.

Таким образом, в настоящее время существует потребность в создании новой постоянно актуализируемой информационной системы для оценки клинической значимости генетических вариантов с выделением группы высоких рисков трансформации в ОМЛ и летального исхода у взрослых пациентов с МДС на основе механизма сбора из различных баз данных для получения информации об изменении функции и структуры белка, эволюционной консервативности гена, популяционной распространённости генетического варианта и его онкогенности.

Цель исследования: разработка информационной системы для стратификации рисков трансформации в острый лейкоз или летального исхода у пациентов с миелодиспластическим синдромом на основе данных мутационного анализа *TP53*.

Задачи исследования

1. Определить требования к информационной системе с учётом особенностей алгоритмов, лежащих в основе современных программ-предикторов.
2. Разработать программу для поиска ортологов в целях определения эволюционной распространённости генетических вариантов среди ортологичных последовательностей.
3. Выявить информативные факторы для дифференциации патогенных и нейтральных вариантов.
4. Разработать алгоритмическое и программное обеспечение информационной системы для определения групповых и индивидуальных рисков

трансформации в ОМЛ и летального исхода больных с МДС, определить её операционные характеристики.

Научная новизна исследования. В результате выполнения диссертационного исследования получены следующие результаты, характеризующиеся научной новизной:

1. Разработан метод исследования эволюционной variability мультидоменных белков, основанный на множественном выравнивании субпоследовательностей ортологов.
2. Впервые в России создана программа для поиска ортологов, позволяющая изучать эволюционную историю произвольных генов с использованием как кластеров Маркова, так и наибольших максимальных клик.
3. Предложена новая концепция, основанная на методе генетического выравнивания «псеворидов» гомологичных последовательностей, предназначенная для определения нейтрального клинического эффекта генетических вариантов.

В рамках выполнения диссертационной работы было осуществлено решение актуальной научной задачи — стратификация пациентов с миелодиспластическим синдромом на основании данных мутационного анализа гена *TP53*, что имеет значение для медицинской информатики.

Теоретическая и практическая значимость работы

Созданная информационная система позволяет определять группы высокого и низкого риска трансформации в лейкоз и смерти среди пациентов с МДС на основе данных мутационного анализа *TP53*. По мере обновления информации из независимых баз данных может производиться реклассификация генетических вариантов.

Информационная система предназначена для использования лабораторным генетиком при интерпретации данных мутационного анализа гена *TP53* у пациентов с МДС для оценки патогенности обнаруженных мутаций. Для интерпретации миссенс-замен целесообразно применять информационную систему в виде веб-приложения, где в качестве входных параметров указываются координаты мутации и альтернативная аминокислота.

Положения, выносимые на защиту

1. Информационная система обладает наилучшей точностью определения патогенных мутаций генов *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA* среди программ-предикторов, а по чувствительности и специфичности не уступает им.
2. Информационная система обеспечивает стратификацию рисков трансформации в лейкоз и летального исхода у пациентов с миелодиспластическим синдромом на основании данных мутационного анализа.

Степень достоверности результатов исследования. Достоверность исследования обеспечена корректностью применения современных методик сбора и обработки исходной информации, правильным подбором объектов наблюдения, достаточным объемом исследуемой выборочной совокупности, использованием апробированного математического аппарата и подтверждается проверкой, обсуждением результатов исследования на международных и всероссийских научных конференциях, публикациями результатов исследования в рецензируемых научных изданиях.

Внедрение результатов работы в практическое здравоохранение. Результаты внедрены в работу Лаборатории трансплантологии и молекулярной гематологии клиники «НИИ детской онкологии, гематологии и трансплантологии им. Р.М. Горбачевой» ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России и в учебный процесс кафедры физики, математики и информатики ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России.

Методология и методы исследования. Объектом настоящего исследования являются данные молекулярной диагностики пациентов с миелоидной патологией, включая МДС, — одного из наиболее актуальных направлений развития современной онкологии. В качестве предмета исследования рассматривается оценка индивидуальных и групповых рисков трансформации в ОМЛ и смерти у пациентов с МДС. В данной работе для обозначения любого отличия последовательности ДНК пациента от референсной последовательности генома человека 2013 года версии GRCh38 используется термин «генетический вариант». Ретроспективно были проанализированы результаты мутационного анализа гена *TP53* пациентов с МДС и ОМЛ, обследованных в рамках деятельности Международной рабочей группы изучения МДС (2343 человека) (E. Bernard et al., 2022) и проходивших лечение в ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России в период с 2018 по 2023 год (195 человек). В качестве материала для выделения ДНК использовался костный мозг или кровь. Методы настоящего исследования включают в себя обработку биологических, а также деперсонифицированных клинических и лабораторных данных с использованием языка Python 3 и библиотеки Pandas, формирование баз данных последовательностей типа Lightning Memory-Mapped Database (LMDB) при помощи программного пакета BLAST+ (C. Camacho et al., 2009), графовый анализ, визуализацию графов в виде файлов HTML (HyperText Markup Language, язык гипертекстовой разметки) с внедрёнными функциями JavaScript, множественное выравнивание последовательностей, построение филогенетических деревьев, анализ данных,

полученных с помощью NGS, методы дескриптивной и сравнительной статистики, включая отбор признаков и анализ выживаемости.

Апробация результатов. Основные положения диссертации доложены и обсуждены на:

1. XXIV Всероссийской конференции молодых учёных с международным участием «Актуальные проблемы биомедицины - 2018» (г. Санкт-Петербург, 12–13 апреля 2018 г.);
2. XVI Санкт-Петербургской международной конференции «Региональная информатика (РИ-2018)» (г. Санкт-Петербург, 24–26 октября 2018 г.);
3. Конференции «Вычислительная биология и искусственный интеллект для персонализированной медицины 2022» (г. Москва, 2–4 августа 2022 г.);
4. XXVII конгрессе Европейской Ассоциации Гематологии «ЕНА2022 Congress» (г. Вена, 9–12 июня 2022 г.);
5. XVIII международном Симпозиуме памяти Раисы Максимовны Горбачевой «Трансплантация гемопоэтических стволовых клеток. Генная и клеточная терапия» (г. Санкт-Петербург, 19–21 сентября 2024 г.);
6. Заседаниях кафедры медицинской кибернетики и информатики им. С. А. Гаспаряна медико-биологического факультета ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России (Пироговский Университет) 26 марта 2024 г., 24 сентября 2024 г., 24 июня 2025 г.

Апробация диссертационной работы была проведена 3 июня 2025 г. (протокол заседания № 2) на Проблемной комиссии №18 «Биомедицинская статистика» ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России.

Публикации по теме диссертации. По теме диссертации опубликовано 16 научных работ, в том числе в рецензируемых научных изданиях — 2, входящих в международные реферативные базы данных и системы цитирования — 6, входящих в сборники конференций — 4, свидетельств о регистрации программы для электронно-вычислительных машин (ЭВМ) — 3, иных входящих в перечень Высшей аттестационной комиссии — 1. В опубликованных работах в полной мере изложены основные положения диссертации. В диссертации отсутствуют недостоверные сведения об опубликованных соискателем ученой степени работах, в которых изложены основные научные результаты диссертации.

Личный вклад автора. Диссертация является результатом самостоятельной научно-исследовательской работы, выполненной на кафедре физики, математики и информатики ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России. Автором лично выполнен весь объем исследования, произведен обзор отечественной и иностранной литературы по изучаемой

проблеме, проведен логический и статистический анализ результатов исследования, разработаны методика, методы и компьютерные программы, сформулированы выводы и практические рекомендации, оформлены диссертация и автореферат.

Соответствие диссертации паспорту специальности. Диссертация выполнена по специальности медицинской отрасли науки 3.3.9. Медицинская информатика (медицинские науки) и соответствует пунктам 3 и 5 паспорта специальности: 3. Разработка компьютерных методов, баз данных и программных средств для получения, накопления, обработки, передачи и систематизации медицинских и экологических данных с целью использования в лечебно-диагностическом, реабилитационном, профилактическом, образовательном процессах, 5. Разработка алгоритмов профилактики, диагностики, прогнозирования, лечения; моделей и алгоритмов для оценки индивидуального и группового риска.

Объем и структура диссертации. Диссертация изложена на 221 странице машинописного текста (191 страница без приложений) и состоит из введения, обзора литературы, материалов и методов, главы собственных исследований, заключения, выводов, практических рекомендаций, списка сокращений и условных обозначений, списка литературы и списка иллюстративного материала. Иллюстрирована 20 таблицами и 27 рисунками. Библиографический указатель включает 12 отечественных и 253 зарубежных источника.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В **первой главе** при использовании данных отечественной и зарубежной литературы представлен обзор современных компьютерных технологий, которые применяются для определения патогенности генетических вариантов, а также общепринятый алгоритм интерпретации значения генетических вариантов, позволяющий учитывать различные факторы, включая молекулярные, при выборе тактики ведения пациентов.

Во **второй главе** представлены материалы и методы, которые были использованы в ходе исследования. На первом этапе на основе критического анализа современных программ-предикторов были определены требования к разрабатываемой информационной системе.

На втором этапе разрабатывалась программа для поиска ортологов методом кластера ортологичной группы (COG, cluster of orthologous group) с использованием языков Bash, Python и JavaScript. В первом модуле осуществляется создание базы данных типа LMDB на основе геномных сборок с наибольшими показателями полноты BUSCO. Во втором модуле реализован поиск ортологов на основе графового анализа. Выравнивания ортологичных последовательностей формировались с учётом мультидоменной структуры некоторых белков.

На третьем этапе осуществлялось определение факторов для классификации. Проводился анализ рекомендаций по интерпретации данных NGS (О. П. Рыжкова и др., 2020, P. Horak et al., 2022). Для оценки эволюционной распространённости вариантов последовательности ортологов разделялись на фрагменты («псевдорида»), которые затем выравнивались на последовательность гена человека, после чего в выравнивании производился поиск вариантов. Для определения популяционных частот генетических вариантов использовалась база данных GnomAD, для оценки онкогенности генетических вариантов — OncoKB, для определения «горячих точек» — «Cancer Hotspots». Механизм нарушения функции гена определялся по стандартной оценке частоты возникновения миссенс-вариантов.

Набор данных для обучения включал 233 патогенных и 199 доброкачественных миссенс-вариантов генов *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA* базы данных ClinVar. В ходе трёхэтапной перекрёстной проверки определялись чувствительность, специфичность, точность и логарифм отношения шансов классификации генетических вариантов. Отбор признаков производился при помощи точного теста Фишера. Пороговое значение популяционной частоты определялось при помощи индекса Юдена.

На четвёртом этапе осуществлялась разработка информационной системы для определения групповых и индивидуальных рисков трансформации в ОМЛ и

летального исхода у пациентов с МДС на основании данных мутационного анализа гена *TP53*. В основе системы лежат методы биоинформатической обработки данных NGS, реализованные с помощью языков Python, Bash и R (D. S. Bug et al., 2020, I. S. Moiseev et al., 2021, D. Bug et al., 2022). В процессе научного поиска применялись методы молекулярной динамики и оценки термодинамических свойств молекул (D. S. Bug et al., 2021). Тестовая выборка состояла из 99 патогенных и 115 нейтральных вариантов.

В рамках валидации были ретроспективно проанализированы результаты мутационного анализа гена *TP53* 2343 пациентов из когорты Международной рабочей группы изучения МДС (E. Bernard et al., 2022). Проводился сплошной отбор информации. Критерии включения: диагноз МДС, известный мутационный статус гена *TP53*, возраст не менее 18 лет, количество бластов крови менее 20%, число лейкоцитов крови менее $13 \cdot 10^9/\text{л}$, отсутствие потери *TP53*. Определялись пол пациентов, возраст в дебюте, диагноз, гемоглобин и тромбоциты крови, бласты костного мозга, комплексный кариотип, делеции 5q, 7q и 17p без потери *TP53*, мутации гена *TP53*. Половозрастной состав обследованных отображён на рисунке 1.

Ретроспективно проанализированы результаты мутационного анализа *TP53* 195 пациентов с МДС и ОМЛ, обследовавшихся в ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России с 2018 по 2023 год. Проводился сплошной отбор информации. Критерии включения: возраст от 18 лет и установленный диагноз МДС или ОМЛ. Прогрессия МДС определялась как увеличение количества бластных клеток более чем на 50% и переход в более продвинутую нозологическую форму.

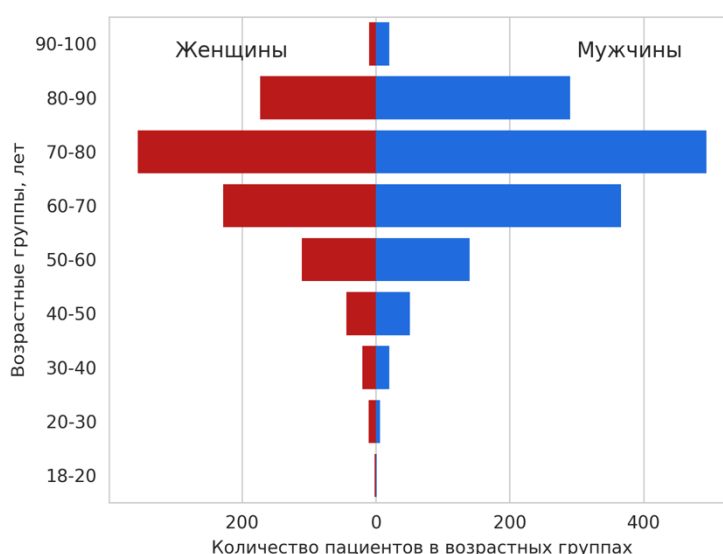


Рисунок 1 — Половозрастная диаграмма обследованных Международной рабочей группы изучения миелодиспластического синдрома

Время наблюдения составило до 5 лет с момента постановки диагноза в обеих когортах. Для оценки риска применялись шкалы IPSS-R и IPSS-M.

Для проверки нормальности распределения использовался критерий Колмогорова-Смирнова, для сравнения групп пациентов по качественным характеристикам — критерий Хи-квадрата Пирсона и точный тест Фишера, для сравнения групп пациентов по количественным признакам — критерии Краскела-Уоллиса, Манна-Уитни и t-критерий Стьюдента. Поправка Бенджамини-Хохберга применялась для решения проблемы множественных сравнений. Функция выживаемости определялась по методу Каплана-Мейера. Для оценки влияния факторов на выживаемость применялась регрессионная модель Кокса.

В **третьей главе** описаны полученные результаты. В первом разделе выявлены основные характеристики, которыми должна обладать информационная система:

1. Отбор наиболее полных геномных сборок;
2. Использование только ортологов для оценки патогенности;
3. Разделение мультидоменных белков на субпоследовательности;
4. Учёт качества эволюционных изменений нуклеотидных последовательностей;
5. Возможность одновременного выполнения анализа данных NGS;
6. Привлечение дополнительных сведений из регулярно обновляемых баз данных.

Информационная система, удовлетворяющая этим требованиям, будет лишена недостатков программ-предикторов, что позволит решить проблему генетических вариантов неясного значения при МДС.

Во втором разделе описана программа для поиска ортологов. В результате работы программы формируется граф, узлы которого представлены генами, при этом предполагаемые ортологи связаны рёбрами и определяются по кластерам Маркова или наибольшим максимальным кликам. Таким образом, была разработана программа для поиска ортологов в целях определения фактора эволюционной распространённости генетических вариантов, который может применяться для оценки патогенности.

В третьем разделе представлены факторы, отобранные для оценки патогенности вариантов. Наиболее связанными с целевой переменной оказались признаки популяционной и эволюционной распространённости: для определения последней был предложен и реализован новый метод выравнивания для мультидоменных белков (D. S. Bug et al., 2024). Дополнительно для классификации использовался фактор онкогенности. Итак, были определены

наиболее информативные факторы, которые могут служить основой для оценки патогенности генетических вариантов в рамках информационной системы.

В четвёртом разделе представлена информационная система. Схема мутационного анализа с применением информационной системы представлена на рисунке 2. Генетический вариант относился к нейтральным при его отсутствии среди онкогенных и соответствии хотя бы одному из условий: (а) популяционная частота выше пороговой, (б) наличие генетического варианта среди ортологов, (в) наличие последовательности ортолога с заменой на остаток биохимической группы альтернативной аминокислоты при совпадении соседних аминокислот рассматриваемой позиции в последовательности ортолога и последовательности человека. Остальные генетические варианты классифицировались как патогенные.

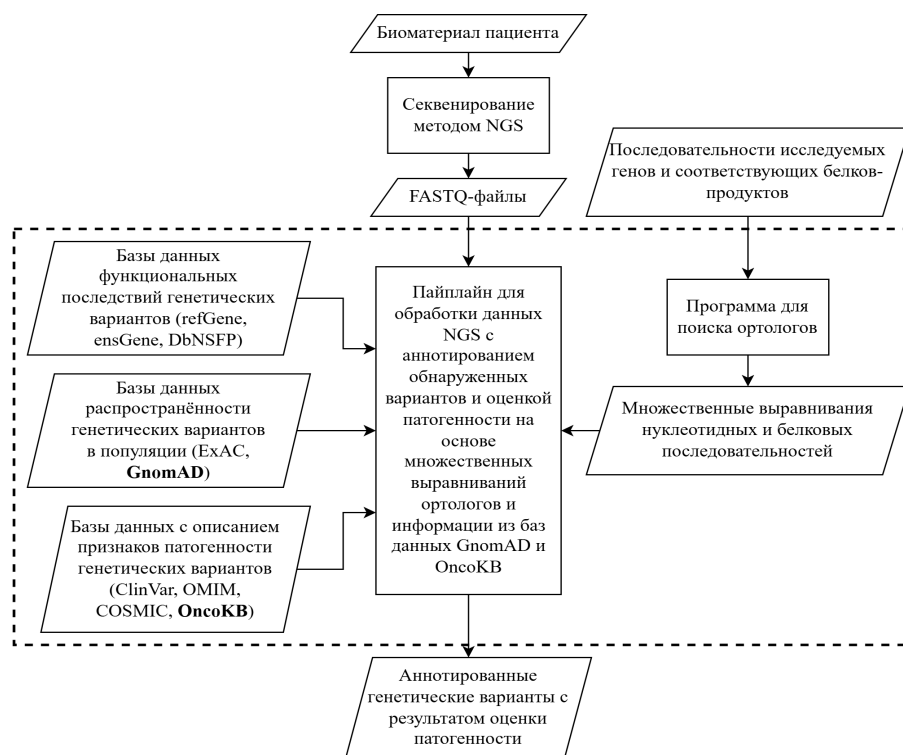


Рисунок 2 — Схема мутационного анализа с применением информационной системы для оценки патогенности генетических вариантов. Шаги, осуществляемые в рамках работы информационной системы, выделены пунктирной рамкой

Основные принципы, лежащие в основе информационной системы, были изложены в публикациях, посвящённых оценке патогенности генетических вариантов генов *CLCN7*, *MeCP2*, *UMOD* и *CYP21A2* (D. S. Bug et al., 2020, Д. С. Буг, И. А. Фильченко, М. В. Радионова, 2018, Д. С. Буг, 2018, Д. А. Кузовенкова и др., 2022). В качестве пороговой популяционной частоты было выбрано значение 0,00071% (рисунок 3). Представлено веб-приложение

для классификации миссенс-замен гена *TP53* (https://epicenter.lspbgmu.ru/shiny/tp53_missense/, рисунок 4).

Таким образом, была разработана информационная система, позволяющая осуществлять поиск ортологов в обновляемых базах данных для определения и актуализации показателя эволюционной распространённости генетических вариантов, а также проводить их классификацию.

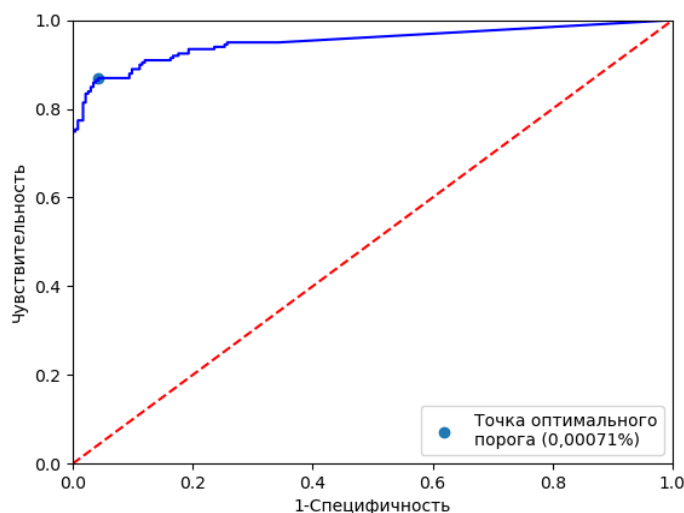


Рисунок 3 — ROC-кривая классификатора, основанного на популяционной частоте генетических вариантов

Классификатор миссенс-вариантов TP53

Классификатор миссенс-вариантов, основанный на эволюционном, популяционном критериях и данных об онкогенности.

Позиция в последовательности белка (P04637)

Альтернативная аминокислота:

Arg (R) ▾

Оценка патогенности P72R:
Нейтральный вариант
Нет в OncoKB; Высокое распространение в популяции; Не встречается в эволюции

Рисунок 4 — Веб-приложение для классификации миссенс-вариантов *TP53*

При помощи карт выравниваний «псевдоридов» ортологичных последовательностей удалось определить нейтральный эффект 72 505 генетических вариантов исследуемых генов (Д. С. Буг, А. Н. Наркевич, Н. В. Петухова, 2023).

Информационная система обладала наибольшими специфичностью и точностью в отношении оценки патогенности миссенс-вариантов среди программ-предикторов (рисунок 5).

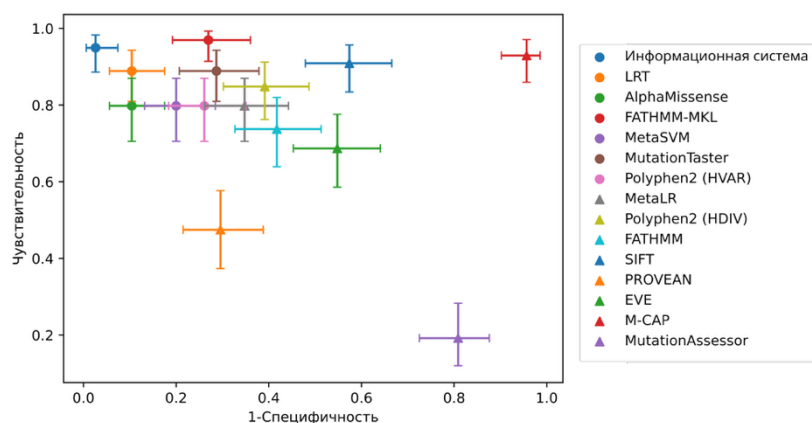
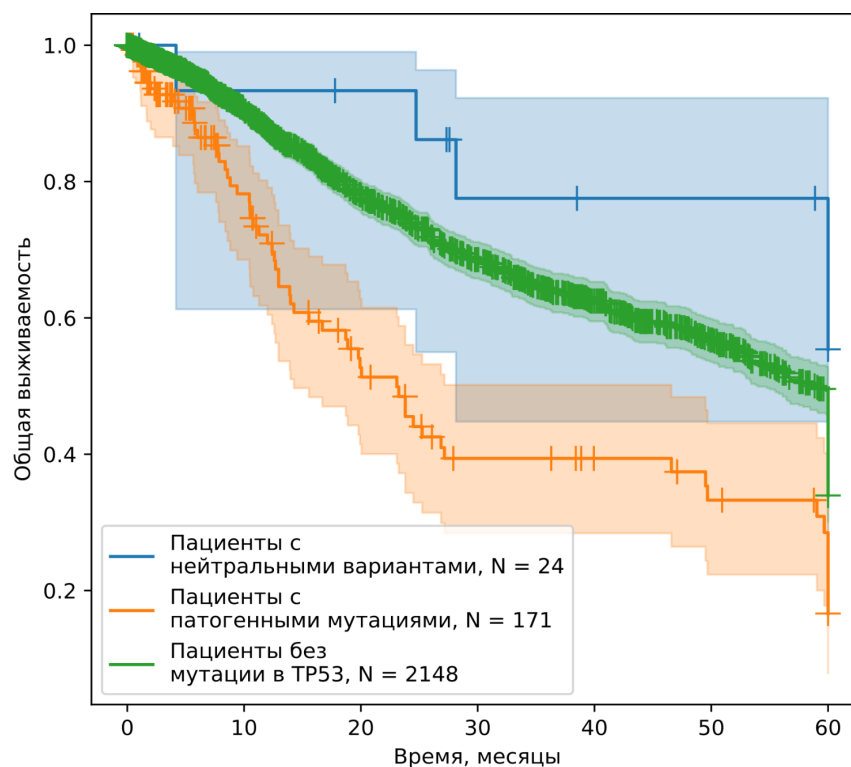


Рисунок 5 — Чувствительность и специфичность определения патогенных вариантов информационной системы и программ-предикторов

Проведена стратификация рисков трансформации в ОМЛ и смерти у пациентов Международной рабочей группы по изучению МДС. У 171 пациента обнаружены патогенные, а у 24 — нейтральные варианты в гене *TP53*. Наблюдались статистически значимые различия между группами пациентов с патогенными и нейтральными вариантами в пятилетней общей ($p_{adj} < 0,01$, рисунок 6) и безлейкозной выживаемости ($p_{adj} < 0,01$), в количественной оценке риска согласно шкалам IPSS-R ($p_{adj} = 0,04$) и IPSS-M ($p_{adj} < 0,01$), в распределении по группам высокого и низкого риска согласно шкале IPSS-M ($p_{adj} = 0,04$).

Дополнительно была исследована группа пациентов с МДС и ОМЛ, наблюдавшихся в ФГБОУ ВО ПСПбГМУ им. И.П. Павлова Минздрава России. К пациентам с патогенными мутациями были отнесены 28 человек, 27 пациентов имели нейтральные варианты. Найдены статистически значимые различия в беспрогрессивной пятилетней выживаемости между группами пациентов с нейтральными и патогенными вариантами ($p_{adj} = 0,04$).

Таким образом, при помощи информационной системы удалось осуществить стратификацию рисков трансформации в ОМЛ и летального исхода у пациентов с МДС.



Пациенты с нейтральными вариантами							
Осталось	20	14	13	9	8	8	0
Цензурировано	4	9	10	12	13	13	19
Событий	0	1	1	3	3	3	5
Пациенты с патогенными мутациями							
Осталось	164	66	38	24	20	16	0
Цензурировано	7	83	91	96	100	101	110
Событий	0	22	42	51	51	54	61
Пациенты без мутации в TP53							
Осталось	2063	1187	841	647	479	335	0
Цензурировано	85	817	1015	1113	1228	1333	1561
Событий	0	144	292	388	441	480	587

Рисунок 6 — Кривые пятилетней общей выживаемости групп пациентов с мутациями *TP53* и без них, отображены кривые выживаемости и 95%-е доверительные интервалы

ЗАКЛЮЧЕНИЕ

Разработана информационная система, позволяющая осуществлять стратификацию рисков смерти и трансформации в острый миелоидный лейкоз у пациентов с миелодиспластическим синдромом. Определены требования к разрабатываемой информационной системе, которые позволили избежать недостатков, присущих программам-предикторам. В частности, в рамках информационной системы осуществляется поиск ортологов в наиболее полных геномных сборках. Разработаны новые подходы для выравнивания мультидоменных белков с учётом эволюции отдельных доменов и определения эволюционной распространённости генетических вариантов на основе выравниваний псевдори́дов. Информационная система использует наиболее значимые факторы для оценки патогенности: признаки популяционной и эволюционной распространённости, а также критерий онкогенности.

Удалось определить эволюционную распространённость 646 известных патогенных и доброкачественных генетических вариантов генов *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA*. Положительная прогностическая ценность с поправкой на встречаемость патогенных вариантов (88%) составила 99,50% (99,03%, 99,98%), отрицательная — 72,27% (55,14%, 89,40%). Разработанная информационная система превосходит современные программы-предикторы по специфичности и точности определения патогенных генетических вариантов. С помощью новой концепции выравнивания «псевдори́дов» гомологичных генов определён нейтральный клинический эффект 72 505 генетических вариантов. Валидация подтвердила, что стратификация при помощи информационной системы позволяет разделять пациентов с миелодиспластическим синдромом на две подгруппы с достоверно различающимися рисками смерти ($p_{adj} < 0,01$) и трансформации в острый миелоидный лейкоз ($p_{adj} < 0,01$).

Следует использовать информационную систему для стратификации рисков трансформации в ОМЛ и смерти у пациентов с МДС на основании данных мутационного анализа при формировании заключения молекулярно-генетического исследования в рамках работы молекулярной лаборатории, а также для определения индивидуального риска осложнений МДС по шкале IPSS-M. Универсальность использованных подходов позволяет адаптировать информационную систему для анализа вариантов неясного значения в других генах и при других нозологиях, что открывает перспективы для персонализированной диагностики в онкологии.

ВЫВОДЫ

1. Определены требования к системе для стратификации рисков трансформации в ОМЛ и летального исхода при МДС на основе мутационного статуса гена *TP53*: отбор наиболее полных геномных сборок с фильтрацией паралогичных последовательностей, фрагментация мультидоменных белков и учёт эволюции нуклеотидных последовательностей генов при определении эволюционной распространённости генетических вариантов, возможность одновременного выполнения анализа данных NGS, привлечение дополнительных сведений из открытых, регулярно обновляемых баз данных.
2. В рамках информационной системы осуществляется поиск ортологов, позволивший определить эволюционную распространённость 646 известных патогенных и доброкачественных генетических вариантов генов *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA*.
3. Определены наиболее значимые факторы для классификации генетических вариантов генов *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA*: их эволюционная распространённость и популяционная частота (с пороговым значением 0,00071% в соответствии с индексом Юдена).
4. Разработана система, позволяющая обновлять данные об эволюционной распространённости генетических вариантов, классифицировать генетические варианты и определять групповые и индивидуальные риски трансформации в ОМЛ и смерти у пациентов с МДС.
5. Произведена оценка операционных характеристик информационной системы в отношении определения патогенности генетических вариантов на наборе миссенс-замен в генах *TP53*, *BRCA1*, *BRCA2*, *DICER1* и *PIK3CA*. Чувствительность информационной системы составила 94,95% (88,61%, 98,34%), специфичность — 96,52% (91,33%, 99,04%), точность — 95,79% (92,17%, 98,06%). С помощью новой концепции выравнивания «псевдори́дов» гомологичных генов определён нейтральный клинический эффект 72 505 генетических вариантов.
6. Валидация подтвердила, что стратификация при помощи информационной системы позволяет разделять пациентов с МДС на две подгруппы с достоверно различающимися рисками трансформации в ОМЛ ($p_{adj} < 0,01$) и летального исхода ($p_{adj} < 0,01$).

ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ

1. При определении клинического эффекта генетического варианта следует опираться на разработанные критерии, в которые входят распространённость варианта среди последовательностей ортологов, его популяционная частота и признаки онкогенности.
2. В процессе исследования эволюции генов для оценки распространённости генетических вариантов среди ортологичных последовательностей целесообразно создавать базу данных наиболее полных (по показателю BUSCO) сборок. Для создания такой базы данных типа LMDb следует использовать специальный модуль программы для поиска ортологов.
3. При поиске ортологов целесообразно пользоваться методом COG, который реализован в программе для поиска ортологов.
4. Следует использовать информационную систему для стратификации рисков трансформации в ОМЛ и смерти у пациентов с МДС на основании данных мутационного анализа при формировании заключения молекулярно-генетического исследования в рамках работы молекулярной лаборатории, а также для определения индивидуального риска осложнений МДС по шкале IPSS-M.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ДНК — дезоксирибонуклеиновая кислота;

МДС — миелодиспластический синдром;

ОМЛ — острый миелоидный лейкоз;

ЭВМ — электронно-вычислительная машина;

BLAST — basic local alignment search tool (Базовый инструмент поиска локального выравнивания);

BUSCO — benchmarking universal single-copy orthologs (эталонное оценивание универсальных однокопийных ортологов);

COG — cluster of orthologous group (кластер ортологичной группы);

HTML — HyperText Markup Language (язык гипертекстовой разметки);

IPSS — международная прогностическая балльная система;

IPSS-M — молекулярная международная прогностическая балльная система;

IPSS-R — пересмотренная международная прогностическая балльная система;

LMDB — Lightning Memory-Mapped Database;

NGS — Next-generation sequencing (секвенирование следующего поколения);

WPSS — прогностическая балльная система Всемирной организации здравоохранения.

СПИСОК НАУЧНЫХ РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России

1. Буг, Д. С. Обзор программ для оценки патогенности генетических вариантов / Д. С. Буг, А. Н. Наркевич, Н. В. Петухова // Учёные записки Первого Санкт-Петербургского государственного медицинского университета имени академика И. П. Павлова. — 2025. — Т. 32, № 1. — С. 11–20.
2. An algorithm for assessing the pathogenicity of genetic mutations in tumor based on a retrospective study of pathogenic and neutral genetic variants / **D.S.Bug**, A. N. Narkevich, A. V. Tishkov, N. V. Petukhova // Siberian Journal of Clinical and Experimental Medicine. — 2025. — Vol. 40, № 1. — P. 226–234.

Публикации в изданиях, входящих в Scopus

1. Evaluating the Effect of 3'-UTR Variants in DICER1 and DROSHA on Their Tissue-Specific Expression by miRNA Target Prediction / **D. S. Bug**, A. V. Tishkov, I. S. Moiseev [et al.] // Curr. Issues Mol. Biol. — 2021. — Vol. 43, № 2. — P. 605–617.
2. Evolutionary reconstruction of MT-RNR2 gene demonstrates a diverse compositional landscape of humanin in vertebrates / **D. S. Bug**, A. V. Tishkov, T. F. Subbotina [et al.] // Journal of Evolutionary Biochemistry and Physiology. — 2023. — Vol. 59, № 5. — P. 1566–1576.
3. High mutation burden in the checkpoint and micro-RNA processing genes in myelodysplastic syndrome / I. S. Moiseev, N. Y. Tsvetkov, I. M. Barkhatov, M. V. Barabanshikova, E. A. Bakin, E. A. Izmailova, A. I. Shakirova, A. D. Kulagin, E. V. Morozova, **D. S. Bug**, N. V. Petuhova, A. V. Tishkov // PLoS ONE. — 2021. — Vol. 16, № 3. — P.e0248430. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0248430>. Дата публикации: 17.03.2021.
4. Identification and Characterization of a Novel CLCN7 Variant Associated with Osteopetrosis / **D. S. Bug**, I. M. Barkhatov, Y. V. Gudozhnikova [et al.] // Genes. — 2020. — Vol. 11, № 11. — P. 1242. URL: <https://www.mdpi.com/2073-4425/11/11/1242>. Дата публикации: 22.10.2020.
5. Shedding light on the *DICER1* mutational spectrum of uncertain significance in malignant neoplasms / **D. S. Bug**, I. S. Moiseev, Yu. B. Porozov, N. V. Petukhova // Front. Mol. Biosci. — 2024. — Vol. 11. URL: <https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2024.1441180/full>. Дата публикации: 03.10.2024 (Scopus).
6. Towards Understanding the Pathogenicity of DROSHA Mutations in Oncohematology / **D. S. Bug**, A. V. Tishkov, I. S. Moiseev [et al.] // Cells. — 2021. — Vol. 10, № 9. — P. 2357. URL: <https://www.mdpi.com/2073-4409/10/9/2357>. Дата публикации: 08.09.2021 (Scopus).

Свидетельства о государственной регистрации программ для ЭВМ

1. Свидетельство о гос. Регистрации программы для ЭВМ 2021611083 Российская Федерация. Программа для оценки относительной экспрессии генов при определении минимальной остаточной болезни у пациентов с острыми лейкозами / **Д. С. Буг**, Е. А. Измайлова, А. В. Тишков, И. М. Бархатов; правообладатель: Общество с ограниченной ответственностью "Иноген". — №2021610274; заявл. 12.01.2021; опублик. 21.01.2021.
2. Свидетельство о гос. Регистрации программы для ЭВМ 2023685071 Российская Федерация. Компьютерная система для анализа данных NGS с прогнозированием клинического значения генетических вариантов / **Д. С. Буг**; правообладатель: ФГБОУ ВО «Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова» Министерства здравоохранения Российской Федерации. — №2023683316; заявл. 03.11.2023; опублик. 22.11.2023.
3. Свидетельство о гос. Регистрации программы для ЭВМ 2023684743 Российская Федерация. Компьютерная система для филогенетической реконструкции методом COG / **Д. С. Буг**, Н. В. Петухова; правообладатель: ФГБОУ ВО «Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова» Министерства здравоохранения Российской Федерации. — №2023683367; заявл. 03.11.2023; опублик. 20.11.2023.

Другие статьи и материалы конференций

1. **Буг, Д. С.** Поиск ортологов метил-СpG-связывающего белка в задаче определения мутации MeCP2 у больных синдромом Ретта / **Д. С. Буг**, И. А. Фильченко, М. В. Радионова // Материалы XXIV Всероссийской конференции молодых учёных с международным участием. — СПб.: РИЦ ПСПбГМУ. — 2018. — С. 50–51.
2. Исследование патогенности мутации V281I в гене CYP21A2 при врожденной гиперплазии надпочечников / Д. А. Кузовенкова, **Д. С. Буг**, В. Д. Назаров [и др.] // Сборник тезисов конференции «Вычислительная биология и искусственный интеллект для персонализированной медицины». — М.: НМИЦ эндокринологии. — 2022. — С. 50.
3. Классификация пациентов с миелодиспластическим синдромом и мутацией TP53 / **Д. С. Буг**, Д. К. Жоголев, А. М. Садыков [и др.] // Приложение к журналу «Клеточная терапия и трансплантация». — 2024. — Т. 13, № 3. — С. 100.
4. Assessment of miRNA biogenesis genes variants in MDS development and progression / **D. Bug**, A. Tishkov, I. Moiseev [et al.] // HemaSphere. — 2022. — Vol. 6. — P. 640–641.
5. Bug, D. S. Alignment of pseudoreads obtained from homologous sequences in identifying potentially tolerated genomic variants. / **D. S. Bug**, A. N. Narkevich, N. V. Petukhova // Journal of Bioinformatics and Genomics. — 2023. — Vol. 21, № 3. — P. 19–25.