Кардиоваскулярная терапия и профилактика. 2025;24(12):4597.
doi: 10.15829/1728-8800-2025-4597
ISSN 1728-8800 (Print)
ISSN 2619-0125 (Online)

Анализ инделов без сдвига рамки считывания в белках, ассоциированных с сердечно-сосудистыми заболеваниями

Раменский В.Е.[1,2], Зайченока М.[1,3], Киселева А.В.[1], Букаева А.А.[1], Ершова А.И.[1], Мешков А.Н.[1], Драпкина О.М.[1]

[1]ФГБУ "Национальный медицинский исследовательский центр терапии и профилактической медицины" Минздрава России. Москва; [2]ФГБОУ ВО "Московский государственный университет им. М.В. Ломоносова". Москва; [3]ФГАОУ ВО "Московский физико-технический институт (национальный исследовательский университет)". Московская область, Долгопрудный, Россия

Цель. Описание распространенности и локализации, а также патогенности и пенетрантности инделов без сдвига рамки считывания в клинически значимых генах, ассоциированных с сердечно-сосудистыми заболеваниями (ССЗ).

Материал и методы. Использовались базы данных ClinVar, dbSNP, а также результаты секвенирования образцов из различных российских когорт. Аннотация вариантов генома проводилась с помощью программы ENSEMBL VEP.

Результаты. Отобраны 42 гена, ассоциированные с 22 ССЗ, проанализированы описанные в базе ClinVar инделы в генах из этой выборки. Наблюдается большой разброс числа инделов в этих генах и их распределения по типам клинической значимости. Инделы значительно менее распространены, но при этом более патогенны, чем несинонимичные варианты (миссенсы). В случае кардиомиопатий, миопатий и сосудистых заболеваний инделы, как каузальные варианты, наиболее редки, а в случае аритмий и гиперхолестеринемии, напротив, встречаются чаще. Показано, что патогенные инделы редко наблюдаются в повторах последовательности или в участках низкой сложности, а доброкачественные инделы реже попадают в участки последовательности, аннотированные как функциональные домены Pfam. На основе анализа >6800 секвенированных образцов из различных российских когорт был сделан вывод о достаточно высокой пенетрантности исследованных патогенных инделов, описаны потенциально специфичные для российской популяции примеры патогенных инделов.

Заключение. Впервые систематически описаны свойства определенного типа вариантов генома — инделов без сдвига рамки считывания — в ключевых генах ССЗ. Полученные результаты подчеркивают клиническую важность инделов в белках с точки зрения их каузальности и пенетрантности, несмотря на их меньшую распространенность по сравнению с несинонимичными вариантами.

Ключевые слова: варианты генома человека, инделы, сердечно-сосудистые заболевания, пенетрантность.

Отношения и деятельность. Государственное задание "Разработка модели предсказания пенетрантности и экспрессивности причинных вариантов наследственных моногенных заболеваний сердечно-сосудистой системы".

Поступила 20/09-2025
Рецензия получена 10/10-2025
Принята к публикации 28/11-2025

Для цитирования: Раменский В.Е., Зайченока М., Киселева А.В., Букаева А.А., Ершова А.И., Мешков А.Н., Драпкина О.М. Анализ инделов без сдвига рамки считывания в белках, ассоциированных с сердечно-сосудистыми заболеваниями. *Кардиоваскулярная терапия и профилактика.* 2025;24(12):4597. doi: 10.15829/1728-8800-2025-4597. EDN: CUZOEV

*Автор, ответственный за переписку (Corresponding author):
e-mail: vramensky@gnicpm.ru

[Раменский В.Е.* — к.ф.-м.н., доцент, руководитель лаборатории геномной и медицинской биоинформатики, в.н.с., руководитель научной группы "ИИ в биоинформатике и медицине" Института перспективных исследований проблем искусственного интеллекта и интеллектуальных систем, факультет биоинженерии и биоинформатики, ORCID: 0000-0001-7867-9509, Зайченока М. — аспирант, ORCID: 0000-0002-2798-9811, Киселева А.В. — к.б.н., руководитель лаборатории молекулярной генетики, в.н.с., ORCID: 0000-0003-4765-8021, Букаева А.А. — н.с. лаборатории клиномики, ORCID: 0000-0002-5932-1744, Ершова А.И. — д.м.н., руководитель лаборатории клиномики, зам. директора по фундаментальной науке, ORCID: 0000-0001-7989-0760, Мешков А.Н. — д.м.н., руководитель Института персонализированной терапии и профилактики, ORCID: 0000-0001-5989-6233, Драпкина О.М. — д.м.н., профессор, академик РАН, директор, ORCID: 0000-0002-4453-8430].

Адреса организаций авторов: ФГБУ "Национальный медицинский исследовательский центр терапии и профилактической медицины" Минздрава России, Петроверигский пер., д. 10, стр.3, Москва, 101990, Россия; ФГБОУ ВО "Московский государственный университет им. М.В. Ломоносова", Ленинские горы, д.1, Москва, 119991, Россия; ФГАОУ ВО "Московский физико-технический институт (национальный исследовательский университет)", Институтский переулок, д. 9, г. Долгопрудный, Московская область, 141701, Россия.

Addresses of the authors' institutions: National Medical Research Center for Therapy and Preventive Medicine, Petroverigsky, 10, bld. 3, Moscow, 101990, Russia; Lomonosov Moscow State University, Leninskie gory, 1, Moscow, 119991, Russia; Moscow Institute of Physics and Technology (National Research University), Institutsky Ln., 9, Dolgoprudny, Moscow region, 141701, Russia.

Analysis of inframe indels in cardiovascular disease-associated proteins

Ramenskiy V. E.[1,2], Zaychenoka M.[1,3], Kiseleva A. V.[1], Bukaeva A. A.[1], Ershova A. I.[1], Meshkov A. N.[1], Drapkina O. M.[1]
[1]National Medical Research Center for Therapy and Preventive Medicine. Moscow; [2]Lomonosov Moscow State University. Moscow; [3]Moscow Institute of Physics and Technology (National Research University). Moscow Oblast, Dolgoprudny, Russia

Aim. To describe the prevalence, location, pathogenicity, and penetrance of inframe indels in clinically significant genes associated with cardiovascular diseases (CVD).
Material and methods. We used the ClinVar and dbSNP databases, as well as sequencing data from samples from various Russian cohorts. Genome variant annotation was performed using the ENSEMBL VEP program.
Results. Forty-two genes associated with 22 CVDs were selected, and indels in the genes from this sample, described in the ClinVar database, were analyzed. A wide range of indel numbers and their distribution by type of clinical significance was observed. Indels are significantly less common, but are more pathogenic than non-synonymous variants (missenses). Indels, as causal variants, are the rarest in cardiomyopathies, myopathies, and vascular diseases, while they are more common in arrhythmias and hypercholesterolemia. Pathogenic indels were shown to be rarely observed in sequence repeats or low-complexity regions, while benign indels were less frequently observed in sequence regions annotated as Pfam functional domains. Based on the analysis of >6800 sequenced samples from various Russian cohorts, we revealed that the studied pathogenic indels have a relatively high penetrance. Examples of pathogenic indels potentially specific to the Russian population are described.
Conclusion. For the first time, the characteristics of a certain type of genomic variant (inframe indels) in key CVD genes have been systematically described. The obtained results highlight the clinical importance of causality and penetrance of protein indels, despite their lower prevalence compared to nonsynonymous variants.

Keywords: human genome variants, indels, cardiovascular diseases, penetrance.

Relationships and Activities. State assignment "Development of a Model for Predicting the Penetrance and Expressivity of Causal Variants in Hereditary Monogenic Cardiovascular Diseases".

Ramenskiy V. E.* ORCID: 0000-0001-7867-9509, Zaychenoka M. ORCID: 0000-0002-2798-9811, Kiseleva A. V. ORCID: 0000-0003-4765-8021, Bukaeva A. A. ORCID: 0000-0002-5932-1744, Ershova A. I. ORCID: 0000-0001-7989-0760, Meshkov A. N. ORCID: 0000-0001-5989-6233, Drapkina O. M. ORCID: 0000-0002-4453-8430.

*Corresponding author:
vramensky@gnicpm.ru

Received: 20/09-2025
Revision Received: 10/10-2025
Accepted: 28/11-2025

For citation: Ramenskiy V. E., Zaychenoka M., Kiseleva A. V., Bukaeva A. A., Ershova A. I., Meshkov A. N., Drapkina O. M. Analysis of inframe indels in cardiovascular disease-associated proteins. *Cardiovascular Therapy and Prevention.* 2025;24(12):4597. doi: 10.15829/1728-8800-2025-4597. EDN: CUZOEV

ГМС — глубокое мутационное сканирование, ДИ — доверительный интервал, ДНК — дезоксирибонуклеиновая кислота, ССЗ — сердечно-сосудистые заболевания, ACMG — American College of Medical Genetics (Американская коллегия медицинской генетики и геномики), MONDO — Mondo Disease Ontology, OR — odds ratio (отношение шансов).

Введение

Термином "инделы" (англ. insertions and deletions) традиционно называют вставки (инсерции) и удаления (делеции) коротких участков генома, которые неизбежно происходят при репликации или репарации геномной дезоксирибонуклеиновой кислоты (ДНК) [1]. Инделы являются наиболее распространенным после однонуклеотидных вариантов типом полиморфизма генома [2]. В экзоме (совокупности всех участков генома, кодирующих белки) эффект инделов определяется в первую очередь их длиной: если она не кратна трем, происходит сдвиг рамки считывания гена, что сопровождается появлением раннего стоп-кодона, укорачивает белок и приводит к его деградации. В большинстве случаев такие инделы вызывают полную или частичную потерю белка [3].

Предмет настоящего исследования составляют вставки или удаления участков экзонов с длиной, кратной трем нуклеотидам, что сохраняет рамку считывания гена и удаляет или добавляет несколько аминокислотных остатков в последовательность белка. Такие варианты генома (inframe indels, далее просто "инделы") представляют большой интерес ввиду широты их функциональных и фенотипических проявлений [4]. В базе ClinVar, которая содержит информацию о вариантах генома человека, в первую очередь в связи с моногенными заболеваниями, в марте 2024г было описано примерно по 3 тыс. патогенных и доброкачественных инделов и ~18 тыс. инделов с неизвестной или противоречивой клинической значимостью [5]. В качестве "классического" примера безусловно патогенного индела можно привести удаление единственного остатка фенилаланина ΔF508 в белке CFTR, являющееся наиболее частой причиной муковисцидоза [6]. В то же время, в последние годы в геномной базе gnomAD v.4.1.0 описано >300 тыс. инделов, подавляющее большинство из которых функционально и эволюционно нейтральны [2].

Функциональная роль инделов, таким образом, весьма разнообразна и аналогична таковой у несинонимичных вариантов (миссенсов). Однако, если последним в клинических рекомендациях по интерпретации данных секвенирования Американской коллегии медицинской генетики и ге-

Ключевые моменты

Что известно о предмете исследования?

- Варианты генома, приводящие к удалению или вставке небольшого числа остатков аминокислот в белке (инделы), менее распространены, чем несинонимичные варианты, но представляют интерес с точки зрения оценки их каузальности в различных моногенных заболеваниях.
- Экзом представителя европейской популяции содержит, в среднем, 115 инделов, для них характерны самые разные функциональные и/или клинические проявления.

Что добавляют результаты исследования?

- Впервые был проведен анализ спектра инделов в группе генов, достоверно ассоциированных с сердечно-сосудистыми заболеваниями. Было показано, что число и типы клинической значимости инделов в генах сильно различаются. Наиболее распространены делеции длиной в один остаток и вставки с длинами от двух до пяти остатков.
- Инделы значительно менее распространены, но при этом более патогенны, чем несинонимичные варианты (миссенсы). В случае кардиомиопатий, миопатий и сосудистых заболеваний инделы, как каузальные варианты, наиболее редки, а в случае аритмий и гиперхолестеринемии, напротив, встречаются чаще.
- Патогенные инделы редко наблюдаются в повторах последовательности или в участках низкой сложности, а доброкачественные инделы избегают функциональные домены Pfam.
- На основе анализа целевых генов в >6800 секвенированных образцах из различных российских когорт был сделан вывод о достаточно высокой пенетрантности исследованных патогенных инделов, описаны потенциально специфичные для российской популяции примеры патогенных инделов.

Key messages

What is already known about the subject?

- Genome variants resulting in the deletion or insertion of a small number of amino acid residues in a protein (indels) are less common than nonsynonymous variants but are of interest for assessing their causality in various monogenic diseases.
- The exome of a representative of the European population contains, on average, 115 indels, which are characterized by a wide variety of functional and/or clinical manifestations.

What might this study add?

- For the first time, the indel profile was analyzed in a group of genes significantly associated with cardiovascular diseases. The number and types of clinical significance of indels in genes vary greatly. The most common are deletions of one residue and insertions of two to five residues.
- Indels are significantly less common, but more pathogenic, than nonsynonymous variants (missenses). Indels, as causal variants, are the rarest in cardiomyopathies, myopathies, and vascular diseases, while they are more common in arrhythmias and hypercholesterolemia.
- Pathogenic indels are rarely observed in sequence repeats or low-complexity regions, while benign indels avoid functional domains of Pfam.
- Based on analysis of target genes in >6800 sequenced samples from various Russian cohorts, a conclusion was reached about the relatively high penetrance of the studied pathogenic indels, and examples of pathogenic indels potentially specific to the Russian population are described.

номики (American College of Medical Genetics, или ACMG) [7] отводится, как минимум, девять достаточно развернутых критериев для классификации их клинической значимости, то в случае инделов основные рекомендации сводятся к тому, что, с одной стороны, вероятность патогенности растет по мере увеличения эволюционной консервативности изменяемого участка белка, а также увеличения его длины, а с другой, короткие инделы в повторяющихся участках последовательности белка или в участках с низкой консервативностью (и, тем самым, предположительно без известной функции) будут патогенными с меньшей степенью вероятности. В отечественном "Руководстве по интерпретации данных, полученных методами массового параллельного секвенирования" [8], критерием патогенности предлагается считать уже сам факт расположения индела в неповторяющемся участке белка, без учета степени его консервативности.

Можно предложить несколько причин, почему инделы часто оказываются "в тени" миссенсов и остаются за рамками исследований [9-11]. Во-первых, это их меньшая распространенность: по данным UK Biobank (Биобанка Великобритании), экзом представителя европейской популяции содержит, в среднем, 115 инделов vs 9292 несинони-

мичных вариантов [12]. Во-вторых, для последних существуют методологии экспериментального анализа, в частности, с помощью глубокого мутационного сканирования (ГМС) [13], а также многочисленные вычислительные методы для предсказания эффекта [14]. Полученные этими методами результаты используются как дополнительные критерии при клинической интерпретации [7, 8]. Возможность использовать ГМС для описания функционального эффекта инделов появилась совсем недавно [15]. Вычислительные методы для предсказания эффекта инделов не так многочисленны, как для миссенсов, однако в последние годы по мере развития методов глубокого обучения в этой области наблюдается определенный прогресс [16, 17]. Так, на портале ProteinGym производится систематизация и единообразная оценка эффективности различных предсказательных методов как для миссенсов, так и для инделов[1]. В-третьих, существуют определенные сложности при определении (коллинге) инделов по данным высокопроизводительного секвенирования, в результате чего различные методы могут по-разному определять инделы в одном и том же образце [9, 18], а в базах данных может быть неполная информация об их частоте в популяциях [8]. Наконец, возможны определенные трудности при компактном, удобном и однозначном описании индела как изменения последовательности белка и соотнесения его с изменением геномной ДНК [19]. Как одно из следствий, различные вычислительные методы для предсказания или аннотации инделов требуют отдельной подготовки входных данных. Также затрудняется сравнение обнаруженного индела с описанными ранее в литературе или в базах данных. Отчасти эти трудности преодолеваются строгим соблюдением номенклатуры Human Genome Variation Society (HGVS) при описании вариантов последовательностей белков и ДНК [20].

Систематический поиск в базе данных PubMed термина "in-frame indels" или "inframe indels" дает 53 публикации, в части из которых описывается роль инделов в генах, ассоциированных, в частности, с раком молочной железы, эндометриальной карциномой, педиатрической хордомой и др. видах онкологических и нейродегенеративных заболеваний; а также с синдромами Нунан, Фанкони-Бикеля, семейной почечной глюкозурией, заболеваниями глаз, в т.ч. ретинальной дистрофией. При этом не удалось обнаружить какие-либо исследования, относящиеся к роли инделов в этиологии наиболее значимых и распространенных сердечно-сосудистых заболеваний (ССЗ).

Целью настоящего исследования было описание распространенности и локализации, а также патогенности и пенетрантности инделов без сдвига рамки считывания в клинически значимых генах, ассоциированных с ССЗ. Для этого были решены следующие две основные задачи: анализ клинико-генетической базы данных ClinVar, содержащей информацию о связи вариантов генома с моногенными заболеваниями, а также исследование >6800 секвенированных в ФГБУ "НМИЦ ТПМ" Минздрава России (далее — НМИЦ ТПМ) образцов из различных российских когорт [21-23].

Материал и методы

Для выбора группы генов, достоверно ассоциированных с ССЗ, использовался предложенный коллегией ACMG список генов, содержащих так называемые вторичные находки, т.е. патогенные или потенциально патогенные варианты, о наличии которых носителю предлагается сообщать вне зависимости от первоначальных причин прохождения генетического тестирования [24]. В нашей работе была использована предложенная совместно ACMG и консорциумом ClinGen современная версия списка 3.3, которая содержит 84 гена, ассоциированных с 85 заболеваниями преимущественно с доминантным типом наследования[2]. В основном, это различные виды наследственных опухолевых синдромов или ССЗ; некоторые гены ассоциированы более чем с одним заболеванием. Было отобрано 42 гена, образующих 61 пары с 22 ССЗ, которые можно условно разбить на четыре группы: кардиомиопатии или миопатии; гиперхолестеринемия; аритмии; сосудистые нарушения (таблица 1). Для описания заболеваний также используются идентификаторы MONDO (Mondo Disease Ontology) [25].

В качестве источника патогенных и прочих вариантов в этих генах была использована база данных ClinVar (версия от 06/25), из нее извлекались пары "вариант-заболевание" с клинической значимостью в контексте данного заболевания (патогенный, доброкачественный и т.д.) и уникальным идентификатором RCVnnnnnn. Тем самым разрешаются потенциальные противоречия, когда один и тот же вариант классифицируется по-разному в связке с различными заболеваниями. Для описания вариантов без привязки к тому или иному заболеванию в ClinVar также используются идентификаторы VCVnnnnnn.

Для аннотации вариантов использовалась программа ENSEMBL VEP (Variant Effect Predictor), с помощью которой варианты картировались на репрезентативные и качественные транскрипты генов из набора MANE (Matched Annotation from NCBI and EBI) Select [26]. Тем самым однозначно фиксируются координаты индела в белке, что особенно существенно для альтернативно сплайсируемых генов. Интересующие нас в рамках данной работы инделы описываются в файлах аннотации как inframe_deletion (делеция), inframe_insertion (вставка), а также inframe_indel (крайне редкий случай одновременного удаления и вставки нескольких аминокислот).

С помощью информации из ENSEMBL была также выполнена разметка последовательностей белков на сле-

[1] Notin P. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

[2] ACMG Secondary Findings Genes and Diseases. https://search.clinicalgenome.org/kb/genes/acmgsf.

Таблица 1

Отобранные для анализа 42 гена и 22 ассоциированных с ними заболевания

Гены	Заболевание
Кардиомиопатии, миопатии	
ACTC1, KCNQ1, MYBPC3, MYH7, MYL2, MYL3, TNNC1, TNNI3, TNNT2, TPM1	MONDO:0005045 Гипертрофическая кардиомиопатия
ACTC1, BAG3, DES, FLNC, LMNA, MYH7, RBM20, SCN5A, TNNC1, TNNI3, TNNT2, TPM1, TTN	MONDO:0005021 Дилатационная кардиомиопатия
PLN	MONDO:0000591 Внутренняя кардиомиопатия
PRKAG2	MONDO:0800484 Кардиомиопатия, связанная с геном *PRKAG2*
DES, DSG2, PKP2, PLN	MONDO:0016587 Аритмогенная кардиомиопатия правого желудочка
DSP	MONDO:0011581 Аритмогенная кардиомиопатия с шерстистыми волосами и кератодермией
FLNC	MONDO:0018943 Миофибриллярная миопатия
Гиперхолестеринемия	
LDLR	MONDO:0007750 Гиперхолестеринемия, семейная, 1 типа
PCSK9	MONDO:0011369 Гиперхолестеринемия, аутосомно-доминантная, 3 типа
APOB	MONDO:0007751 Гиперхолестеринемия, аутосомно-доминантная, тип Б
Аритмии	
CALM1, CALM2, CALM3, KCNH2, KCNQ1, TRDN	MONDO:0002442 Синдром удлиненного интервала QT
KCNH2, KCNQ1	MONDO:0000453 Синдром короткого интервала QT
SCN5A	MONDO:0015263 Синдром Бругада
SCN5A	MONDO:0019171 Семейный синдром удлиненного интервала QT
CALM1, CALM2, CALM3, CASQ2, RYR2, TRDN	MONDO:0017990 Катехоламинергическая полиморфная желудочковая тахикардия
KCNQ1	MONDO:0002441 Синдром Джервелла и Ланге-Нильсена
DSC2	MONDO:0016342 Семейная изолированная аритмогенная дисплазия правого желудочка
TMEM43	MONDO:0011459 Аритмогенная дисплазия правого желудочка 5
Сосудистые заболевания	
SMAD3	MONDO:0013426 Синдром аневризмы-остеоартрита
FBN1, MYH11, SMAD3, TGFBR1, TGFBR2	MONDO:0019625 Семейная аневризма грудной аорты и расслоение аорты
ACVRL1	MONDO:0008535 Телеангиэктазия, наследственная геморрагическая, тип 1
ENG	MONDO:0010880 Телеангиэктазия, наследственная геморрагическая, тип 2

Примечание: приведены идентификаторы заболеваний в базе MONDO (Mondo Disease Ontology).

дующие участки: (1) консервативные фрагменты, соответствующие функциональным доменам, аннотированным в базе данных Pfam [27]; (2) участки низкой сложности, в которых возможны повторы аминокислот; (3) предсказанные в базе данных MobiDB [28] участки с неупорядоченной пространственной структурой. Два последних типа — участки низкой сложности и неупорядоченные участки структуры белка, как принято считать, наиболее терпимы к вставкам и удалениям аминокислот [4, 29]. Для участков, представляющих собой домены Pfam, напротив, характерна эволюционная консервативность последовательности, обеспечивающая функциональность этих доменов. Пересечение координат инделa с размеченным участком последовательности одного из описанных типов позволяет отнести его к этому типу и, тем самым, провести анализ локализации инделов различных классов клинической значимости.

В настоящем исследовании были также использованы результаты секвенирования образцов, собранных в Биобанке НМИЦ ТПМ (далее Биобанк) за все время его работы: пациентов и их родственников [21, 22] и представителей популяционной выборки из исследования ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний в регионах Российской Федерации) [30]. Протокол обработки результатов секвенирования подробно описан в работе [23]. Все исследованные образцы могут быть разбиты на четыре группы по принципу целевых участков секвенирования: три различных панели генов, содержащих 13, 34 и 35 генов из исходной группы 42 генов, и полные геномы. В общей сложности объем выборки составил 6837 человек. При анализе результатов секвенирования мы не рассматривали варианты в гене *TTN* ввиду его большой длины.

Сбор и хранение биоматериала осуществляли по регламенту биобанкирования в Биобанке [31]. Исследования были одобрены Независимым Этическим Комитетом НМИЦ ТПМ (номера протоколов 07-03/12 от 03.07.2012, 04-04/17 от 06.06.2017). От всех участников исследования было получено информированное согласие на их участие.

Результаты

Общая характеристика инделов по данным ClinVar

В отобранных 42 генах в базе ClinVar описано в общей сложности 43313 вариантов генома раз-

личной клинической значимости (таблица 2). Несинонимичные варианты (миссенсы) составляют из них ~ половину (49,7%), из которых, в свою очередь, 14,2% являются патогенными. В этих генах в ClinVar описано 525 инделов без сдвига рамки считывания, что составляет лишь 1,2% от общего числа вариантов, при этом 155 инделов классифицированы как патогенные, что соответствует 29,5% от общего числа вариантов данного типа (рисунок 1). Неизвестная или противоречивая значимость приписана 65,5% инделов, а доброкачественные составляют самую небольшую группу (5,0%). Поскольку среди инделов доля патогенных (29,5%) примерно вдвое больше, чем среди миссенсов (14,2%), отношение шансов (OR — odds ratio) быть патогенным для инделов составляет 2,52 (95% доверительный интервал (ДИ): 2,08-3,05, p=0) по сравнению с миссенсами, что может указывать на более высокую каузальность и/или пенетрантность вариантов первого типа.

В отобранных генах ССЗ на один патогенный индел в среднем приходится 19,8 патогенных миссенсов (=3064/155), что близко к соответствующему значению по всем генам в ClinVar (18,8). Группы ССЗ различаются по этому соотношению, которое составляет 47,1 в случае кардиомиопатий и миопатий, 34,9 для сосудистых заболеваний, а также 10,2 (гиперхолестеринемия) и 7,5 (аритмии). Таким образом, в первых двух случаях инделы, как каузальные варианты, наиболее редки, а в случае аритмий и гиперхолестеринемии, напротив, встречаются чаще.

Как можно было ожидать, гены различаются по общему числу описанных в ClinVar инделов: от 1-2 (например, *CASQ2*, *TPM1*) до 91 в гене *LDLR* (рисунок 1). Заметно различаются также распределения по типам клинической значимости: так, наибольшее число патогенных инделов (70 и 27, соответственно) содержат гены *LDLR* (MONDO:0007750, семейная гиперхолестеринемия 1 типа) и *FBN1* (MONDO:0019625, семейная аневризма грудной аорты и расслоение аорты), что сочетается с полным отсутствием доброкачественных инделов в этих

Таблица 2

Варианты генома в 42 генах ССЗ, описанные в ClinVar (колонка 2) и обнаруженные в секвенированных выборках НМИЦ ТПМ (колонка 3)

Варианты	ClinVar	Секвенированные выборки
Все	43313	69487
Миссенсы	21505 (49,7%)	1960 (2,8%)
Патогенные миссенсы	3064 (14,2%)	115 (5,9%)
Инделы	525 (1,2%)	60 (0,1%)
Патогенные инделы	155 (29,5%)	4 (6,7%)

Примечание: для всех вариантов в скобках указана доля в процентах от общего числа (последняя строка), для патогенных указана доля от числа вариантов данного типа. Инделы — без сдвига рамки считывания. ССЗ — сердечно-сосудистые заболевания.

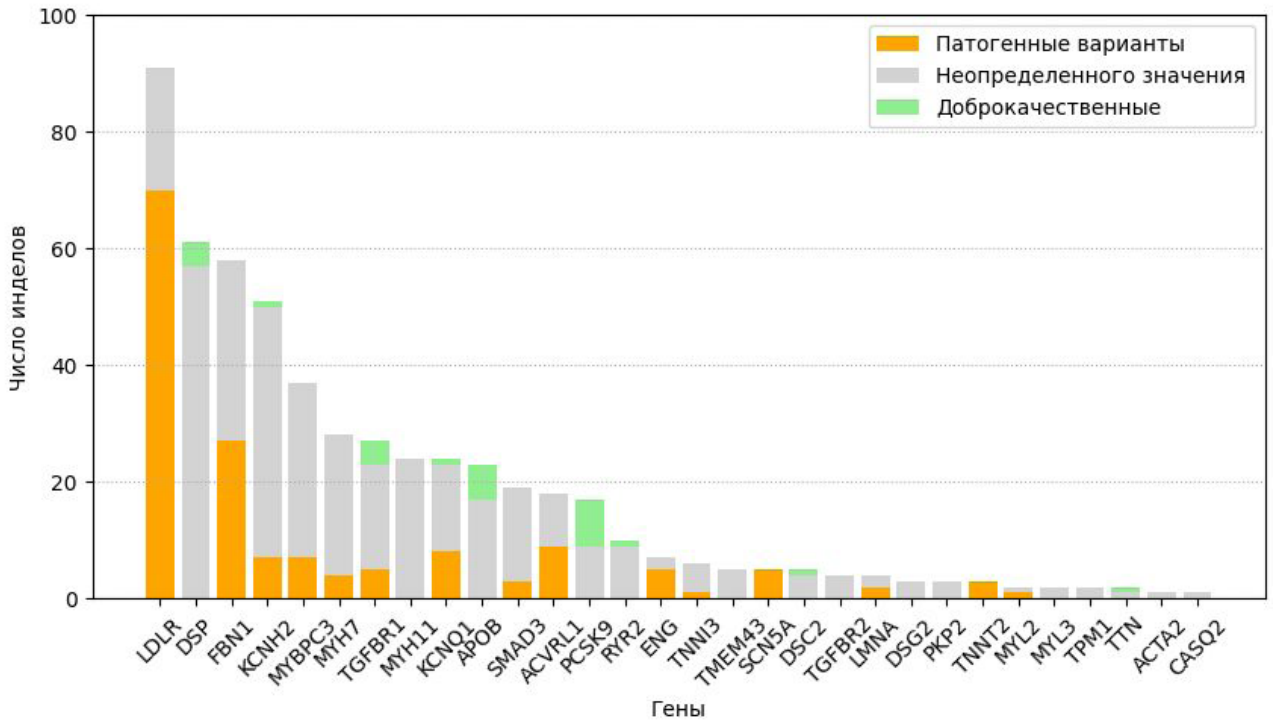

Рис. 1 Число инделов в базе ClinVar с различной клинической значимостью в целевых генах, ассоциированных с ССЗ.
Примечание: ССЗ — сердечно-сосудистые заболевания.

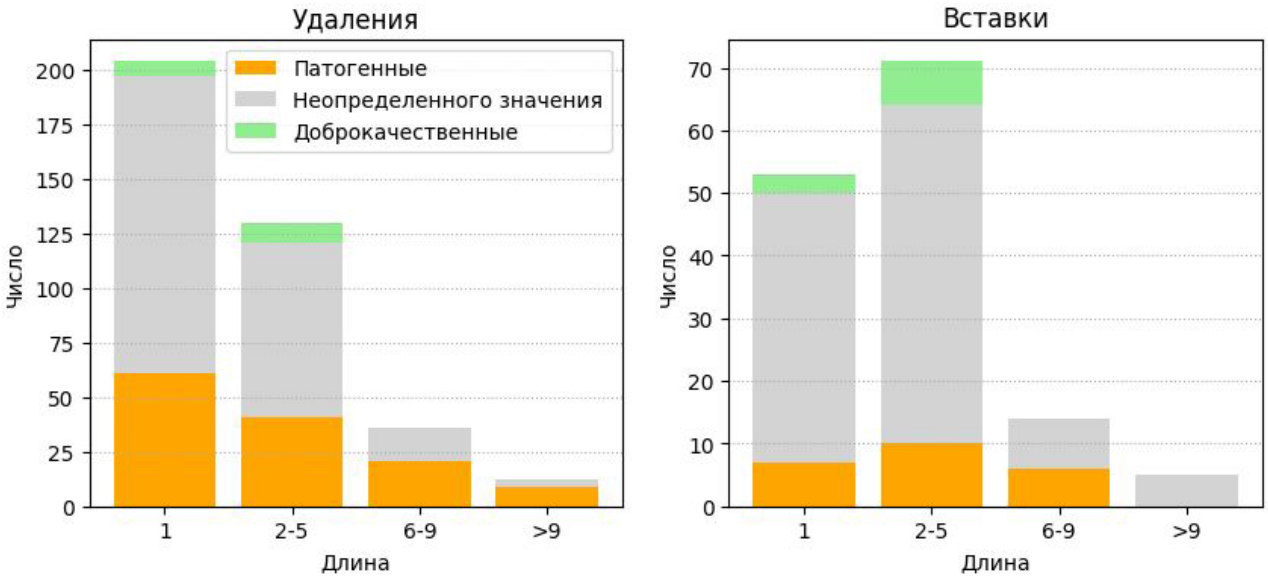

Рис. 2 Распределение длины вставок и удалений в инделах (в остатках аминокислот) в целевых генах в базе ClinVar.

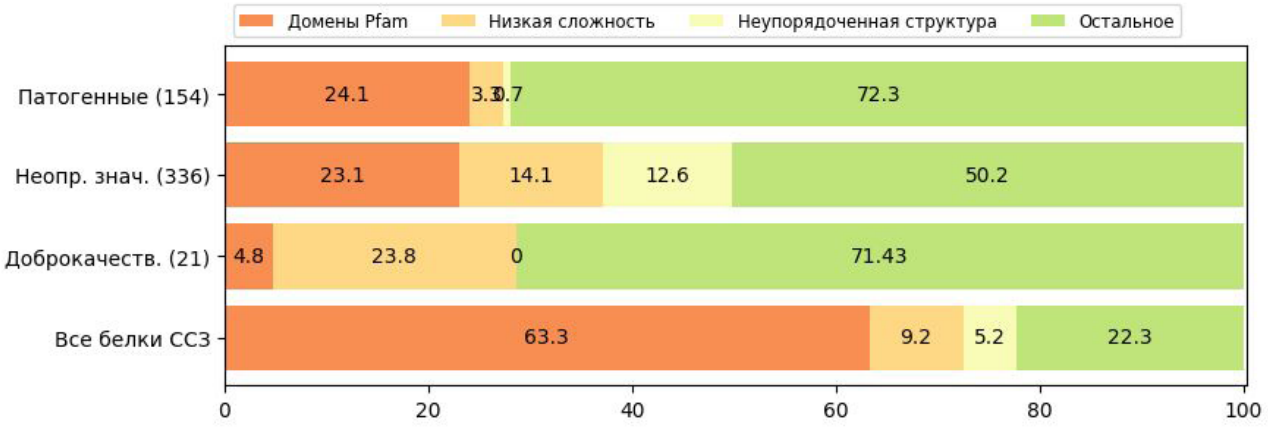

Рис. 3 Расположение инделов с описанной клинической значимостью в различных участках белков, в процентах.
Примечание: в скобках указано число инделов в выборке. Все белки — представленность участков в целевых белках. ССЗ — сердечно-сосудистые заболевания.

генах. Интересно, что в этих генах часть миссенсов является доброкачественной: 4,2 и 10,3%, соответственно. Интересным примером является ген *SCN5A* (MONDO:0015263, синдром Бругада), в котором среди 5 описанных инделов все классифицированы как патогенные.

При рассмотрении отдельно вставок и удалений без сдвига рамки считывания обращает на себя внимание то, что последние составляют большинство инделов: 382 из 525, что составляет 72,8%; при этом в 203 случаях из 328 удаляется один аминокислотный остаток. При этом вставки от 2 до 5 аминокислотных остатков (90 случаев из 143) более распространены, чем вставки единичной длины (53 случая). Распределение по типам клинической значимости в зависимости от длины инделов дано на рисунке 2. Видно, что доброкачественных вставок и удалений с длиной >5 аминокислотных остатков не наблюдается. OR быть патогенным среди удалений длиной >1 составило 1,57 (95% ДИ: 1,02-2,38, p=0,040) по сравнению с удалениями единичной длины. В случае вставок OR=1,42 (95% ДИ: 0,54-3,72), однако ввиду малого объема выборки в данном случае разница не является статистически значимой (p=0,47).

Локализация инделов в участках белков

Для выбранных 42 белков ССЗ в целом характерно достаточно высокое (63,3%) покрытие последовательностей фрагментами, соответствующими доменам Pfam. В базе белков UniProtKB в целом 53% остатков всех белков принадлежат фрагментам последовательности, аннотированным как тот или иной домен [32]; наблюдаемое увеличение этого значения в нашей небольшой выборке бел-

Таблица 3

Инделы в секвенированных образцах из Биобанка НМИЦ ТПМ, для которых доступна информация о диагнозе

Вариант	ClinVar	Носители	Диагноз
Кардиомиопатии			
DSP:p.Ser957del	–	3	Пароксизмальная форма фибрилляции предсердий (1/3)
RBM20:p.Gln55_Ala56delinsPro rs2134793276	–	5	Некомпактный миокард левого желудочка (1/5)
TNNT2:p.Lys220del rs45578238	VCV000043659 P/LP	2	Кардиомиопатия (2/2)
Гиперхолестеринемия			
APOB:p.Leu22dup rs745520533	VCV000993131 VUS	2	Гиперхолестеринемия (2/2)
APOB:p.Trp2554_Thr2555delinsCys	–	1	Предположительно гиперхолестеринемия: 3 балла по шкале DLCN (1/1)
LDLR:p.Asp224_Glu228del rs1555803439	VCV000441189 P/LP	2	Гиперхолестеринемия (2/2)
LDLR:p.Gly219del rs121908027	VCV000226329 P	9	Гиперхолестеринемия (8/9)
LDLR:p.Leu571del rs772492150	VCV002971852 VUS	2	Гиперхолестеринемия (2/2)
LDLR:p.Lys225_Asp227dup rs1555803425	VCV000251377 LP	2	Гиперхолестеринемия (2/2)
LDLR:p.Thr67_Cys68delinsSer	–	2	Гиперхолестеринемия (2/2)
PCSK9:p.Glu49del	–	1	Гиперхолестеринемия (1/1)
PCSK9:p.Leu22_Leu23del rs35574083	VCV000627764 VUS, LB	1	Нет (0/1)

Примечание: приведены идентификаторы в номенклатуре HGVS (Human Genome Variation Society), идентификаторы в базах dbSNP (rsNNNNNN) и ClinVar (VCV000nnnnn), информация о клинической значимости в ClinVar (если есть): B (Benign, доброкачественный), LB (Likely Benign, вероятно доброкачественный), LP (Likely Pathogenic, вероятно патогенный), P (Pathogenic, патогенный), VUS (Variant of Unknown Significance, неизвестная значимость). В скобках указано число носителей вариантного аллеля и число подтвержденных диагнозов у носителей. DLCN — Dutch Lipid Clinic Network (критерии диагностики семейной гиперхолестеринемии).

ков можно объяснить как ее размером, так и лучшей изученностью этих клинически важных белков. Участки низкой сложности и неупорядоченной структуры в белках ССЗ составляют 9,2 и 5,2%, соответственно.

На рисунке 3 показано распределение инделов различной клинической значимости по различным участкам белков. Видно, что патогенные инделы наиболее редко находятся в участках низкой сложности и неупорядоченной структуры (суммарно лишь 4,0%), при этом около четверти (24,1%) попадают в домены Pfam. Доброкачественные инделы, напротив, избегают доменов Pfam (4,8%), при этом почти четверть из них (23,8%) находится в участках низкой сложности. С учетом малой численности этой группы (21 индел) полученные результаты могут быть не вполне репрезентативными: так, во всей базе ClinVar в участках низкой сложности и неупорядоченной структуры находятся 52,1 и 43,4% всех 3064 доброкачественных удалений и вставок, при этом только 7,2% попадают в домены Pfam. Для инделов неизвестной клинической значимости характерна бóльшая представленность в участках низкой сложности и неупорядоченной структуры (14,1 и 12,6%) и близкая к патогенным доля в доменах Pfam (23,1%). В силу последнего обстоятельства локализация в домене Pfam, во всяком случае в исследованных белках ССЗ, не может являться надежным признаком для классификации индела как патогенного.

Инделы в российских выборках

В секвенированных выборках в 42 целевых генах было обнаружено в общей сложности 69,487 вариантов нуклеотидных последовательностей (таблица 2), только 60 из которых являются инделами без сдвига рамки считывания. Из них 22 индела классифицированы в ClinVar как доброкачественные, а также доброкачественные или неизвестной значимости (конфликтующие интерпретации значимости). Для некоторых из них характерны высокие частоты в популяции: например, 30,3% для удаления APOB:p.Leu12_Leu14del или 14,1% в случае вставки PCSK9:p.Leu22dup, что объясняет их высокую встречаемость, наблюдаемую в наших образцах.

С точки зрения связи с фенотипом наибольший интерес представляют 38 остальных обнаруженных в исследованных образцах инделов, среди которых четыре описаны в ClinVar как патогенные (один в гене *TNNT2* и три в *LDLR*), 13 как инделы неизвестной значимости, а 21 вовсе отсутствуют в ClinVar. В таблице 3 приведены сведения о 12 ин-

делах из числа этих 38, для которых по имеющимся клиническим данным можно было бы сделать вывод о наличии или отсутствии соответствующего диагноза. У 24 из 32 носителей мутантных аллелей генов с инделами ассоциации с клиническими проявлениями были подтверждены в результате медицинского обследования в НМИЦ ТПМ.

Большинство инделов наблюдается в 1-5 секвенированных образцах, за исключением удаления LDLR:p.Gly219del, обнаруженного у 9 участников исследования, у которых диагноз гиперхолестеринемия был подтвержден в 8 случаях, что соответствует значению пенетрантности 88,8%. Для трех других патогенных инделов (TNNT2:p.Lys220del — кардиомиопатия, LDLR:p.Asp224_Glu228del и LDLR:p.Leu571del — гиперхолестеринемия) характерна полная пенетрантность. Для двух инделов, охарактеризованных в ClinVar как варианты с неизвестной значимостью (APOB:p.Leu22dup и LDLR:p.Leu571del), также наблюдается полная пенетрантность; противоположный пример дает удаление PCSK9:p.Leu22_Leu23del, для которого диагноз не подтвердился.

Обсуждение

Инделы без сдвига рамки считывания — варианты генома, приводящие к удалению или вставке небольшого числа остатков в белке — представляют интерес в различных контекстах, в первую очередь с точки зрения оценки их каузальности в различных моногенных заболеваниях [33]. Возможности массовой оценки функционального эффекта инделов *in vitro* с помощью технологии ГМС появились совсем недавно [15], но вряд ли эти результаты смогут в ближайшем будущем дать исчерпывающие ответы на вопросы о патогенности тех или иных инделов, ввиду очевидной разницы между функциональным и клиническим эффектом. То же самое можно сказать про вычислительные методы предсказания эффекта инделов, которые переживают своего рода "второе рождение" в эпоху быстрого развития методов глубокого обучения (нейронных сетей) [17, 34]. Накопление данных секвенирования и их интерпретация в клиническом контексте позволяют делать наиболее достоверные выводы о патогенности и пенетрантности различных вариантов генома. Настоящее исследование ставит своей целью собрать наиболее значимые гены, ассоциируемые с ССЗ, и описать свойства встречающихся в них коротких инделов, в частности, их распространенность и локализацию, а также патогенность и пенетрантность. Тем самым, среди прочего, мы надеемся привлечь внимание врачей-генетиков к этим вариантам генома, которые иногда оказываются на периферии внимания исследователей в клинических или научных проектах [10, 11].

В настоящей работе были отобраны 42 гена, ассоциированные с ССЗ, проанализированы описанные в базе ClinVar [5] инделы в генах из этой выборки. Обнаружен большой разброс числа инделов в этих генах: от 1 до 91 в гене, что в, первую очередь, объясняется разницей в длине и нуклеотидном составе генов, а также в общей степени их изученности. Также в ряде случаев в одних и тех же генах наблюдается разница в соотношении различных типов клинической значимости при сравнении инделов и миссенсов. Например, в случае *LDLR* и *FBN1* полное отсутствие доброкачественных инделов статистически значимо, что может свидетельствовать о низкой толерантности данных генов в такому типу варианта. В *SCN5A* все известные инделы патогенные, а среди миссенсов таких вариантов лишь ~ 12,8%. Можно ожидать, что особенности этого гена станут понятнее по мере накопления клинических данных. Разброс количества инделов и их патогенности в различных генах отмечался ранее [14]. Полученные нами результаты позволяют сделать вывод, что инделы более редки, чем миссенсы, но более патогенны. Наиболее сильно это тенденция проявляется в кардиомиопатиях, миопатиях и сосудистых заболеваниях, в случае же аритмий и гиперхолестеринемии инделы, напротив, встречаются чаще.

Показано, что для локализации в различных участках белков доброкачественных и патогенных инделов из исследованной выборки в 30 генов характерны определенные тенденции, соответствующие описанным в рекомендациях ACMG по клинической интерпретации [7], в частности, увеличение вероятности индела быть патогенным по мере роста его длины. В работе [14] авторы наблюдают обратный эффект: некоторое увеличение доли доброкачественных вариантов по мере роста длины индела, однако в обоих случаях тенденция не является сильной. Патогенные инделы редко наблюдаются в повторах последовательности или в участках низкой сложности (суммарно 4%), а доброкачественные попадают в домены Pfam менее, чем в 5% случаев. Таким образом, можно сделать вывод о том, что короткие инделы в повторах последовательности или в участках без известной функции будут патогенны с меньшей степенью вероятности, однако сами по себе особенности локализации инделов в том или ином белке вряд ли могут приводить к однозначной интерпретации их эффекта.

В ходе анализа 42 целевых генов в >6800 секвенированных образцах из различных российских когорт были обнаружены описанные ранее инделы различной клинической значимости, а также 19 инделов, не описанных в базах данных ClinVar и dbSNP. Инделы, отсутствующие в этих базах (например, DSP:p.Ser957del или RBM20:p.Gln55_Ala56delinsPro), представляют интерес как потенциально специфичные для российской популяции

[23]. Имеющиеся для части инделов данные о клинических проявлениях с установленным диагнозом позволяют сделать вывод об их достаточно высокой пенетрантности: так, у 24 из 32 носителей мутантных аллелей были подтверждены диагнозы, в первую очередь гиперхолестеринемия и кардиомиопатия (таблица 2), что позволяет оценить среднюю пенетрантность исследованных инделов в этих генах как 75%. Более точная ее оценка будет возможна по мере накопления данных секвенирования.

Заключение

Впервые был проанализирован спектр инделов без сдвига рамки считывания в группе генов, достоверно ассоциированных с ССЗ. Было показано, что инделы значительно менее распространены, но при этом более патогенны, чем несинонимичные варианты (миссенсы). В случае кардиомиопатий, миопатий и сосудистых заболеваний инделы как каузальные варианты наиболее редки, а в случае аритмий и гиперхолестеринемии, напротив, встречаются чаще. Патогенные инделы реже всего встречаются в участках низкой сложности (повторах) или в участках неупорядоченной структуры белков, а доброкачественные инделы избегают доменов, аннотированных в базе данных Pfam, для которых, как правило, характерна консервативность и функциональная значимость. Наиболее распространены делеции длиной один остаток и вставки (инсерции) с длинами от двух до пяти остатков. На основе анализа целевых генов в >6800 секвенированных образцах из различных российских когорт был сделан вывод о достаточно высокой пенетрантности исследованных патогенных инделов, описаны потенциально специфичные для российской популяции примеры патогенных инделов. В настоящей работе впервые выполнен систематический анализ инделов в генах ССЗ, который подчеркивают их клиническую важность с точки зрения каузальности и пенетрантности, несмотря на меньшую распространенность по сравнению с несинонимичными вариантами.

Отношения и деятельность. Государственное задание "Разработка модели предсказания пенетрантности и экспрессивности причинных вариантов наследственных моногенных заболеваний сердечно-сосудистой системы".

Литература/References

1. Garcia-Diaz M, Kunkel TA. Mechanism of a genetic glissando: structural biology of indel mutations. Trends Biochem Sci. 2006;31(4):206-14. doi:10.1016/j.tibs.2006.02.004.
2. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625(7993):92-100. doi:10.1038/s41586-023-06045-0.
3. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. doi:10.1038/s41586-020-2308-7.
4. Pagel KA, Antaki D, Lian A, et al. Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. PLOS Comput Biol. 2019;15(6):e1007112. doi:10.1371/journal.pcbi.1007112.
5. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862-8. doi:10.1093/nar/gkv1222.
6. Sotnikova EA, Kiseleva AV, Kutsenko VA, et al. Identification of Pathogenic Variant Burden and Selection of Optimal Diagnostic Method Is a Way to Improve Carrier Screening for Autosomal Recessive Diseases. J Pers Med. 2022;12(7):1132. doi:10.3390/jpm12071132.
7. Richards S, Aziz N, Bale 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;17(5):405-24. doi:10.1038/gim.2015.30.
8. Ryzhkova OP, Kardymon OL, Prohorchuk EB, et al. Guidelines for the interpretation of massive parallel sequencing variants (update 2018, v2). Medical Genetics. 2019;18(2):3-23. (In Russ.) Рыжкова О. П., Кардымон О. Л., Прохорчук Е. Б. и др. Руководство по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования (MPS) (редакция 2018, версия 2). Медицинская генетика. 2019;18(2):3-23. doi:10.25557/2073-7998.2019.02.3-23.
9. Jiang Y, Turinsky AL, Brudno M. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Res. 2015;43(15):7217-28. doi:10.1093/nar/gkv677.
10. Savino S, Desmet T, Franceus J. Insertions and deletions in protein evolution and engineering. Biotechnol Adv. 2022;60:108010. doi:10.1016/j.biotechadv.2022.108010.
11. Miton CM, Tokuriki N. Insertions and Deletions (Indels): A Missing Piece of the Protein Engineering Jigsaw. Biochemistry. 2023;62(2):148-57. doi:10.1021/acs.biochem.2c00188.
12. Van Hout CV, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586(7831):749-56. doi:10.1038/s41586-020-2853-0.
13. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11(8):801-7. doi:10.1038/nmeth.3027.
14. Cannon S, Williams M, Gunning AC, Wright CF. Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels. BMC Med Genomics. 2023;16(1):36. doi:10.1186/s12920-023-01454-6.
15. Macdonald CB, Nedrud D, Grimes PR, et al. DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. Genome Biol. 2023;24(1):36. doi:10.1186/s13059-023-02880-6.
16. Shin JE, Riesselman AJ, Kollasch AW, et al. Protein design and variant prediction using autoregressive generative models. Nat Commun. 2021;12(1):2403. doi:10.1038/s41467-021-22732-w.
17. Fan X, Pan H, Tian A, et al. SHINE: protein language model-based pathogenicity prediction for short inframe insertion and deletion variants. Brief Bioinform. 2023;24(1):bbac584. doi:10.1093/bib/bbac584.
18. Barbitoff YA, Abasov R, Tvorogova VE, et al. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant

discovery. BMC Genomics. 2022;23(1):155. doi:10.1186/s12864-022-08365-3.
19. Yen JL, Garcia S, Montana A, et al. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med. 2017;9. doi:10.1186/s13073-016-0396-7.
20. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37(6):564-9. doi:10.1002/humu.22981.
21. Meshkov A, Ershova A, Kiseleva A, et al. The LDLR, APOB, and PCSK9 Variants of Index Patients with Familial Hypercholesterolemia in Russia. Genes. 2021;12(1):66. doi:10.3390/genes12010066.
22. Meshkov AN, Kiseleva AV, Ershova AI, et al. ANGPTL3, ANGPTL4, APOA5, APOB, APOC2, APOC3, LDLR, PCSK9, LPL gene variants and coronary artery disease risk. Russian Journal of Cardiology. 2022;27(10):5232. (In Russ.) Мешков А. И., Киселева А. В., Ершова А. И. и др. Варианты Генов ANGPTL3, ANGPTL4, APOA5, APOB, APOC2, APOC3, LDLR, PCSK9, LPL и риск ишемической болезни сердца. Российский кардиологический журнал. 2022;27(10):5232. doi:10.15829/1560-4071-2022-5232.
23. Ramensky VE, Ershova AI, Zaicenoka M, et al. Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region. Front Genet. 2021;12:1782. doi:10.3389/fgene.2021.709419.
24. Miller DT, Lee K, Chung WK, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2021;23(8):1381-90. doi: 10.1038/s41436-021-01172-3.
25. Vasilevsky NA, Toro S, Matentzoglu N, et al. Mondo: Integrating Disease Terminology Across Communities. Genetics. Published online October 6, 2025:iyaf215. doi:10.1093/genetics/iyaf215.
26. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310-15. doi:10.1038/s41586-022-04558-8.
27. Paysan-Lafosse T, Andreeva A, Blum M, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2025;53(D1):D523-34. doi:10.1093/nar/gkae997.
28. Piovesan D, Del Conte A, Clementel D, et al. MobiDB: 10 years of intrinsically disordered proteins. Nucleic Acids Res. 2023;51(D1):D438-44. doi:10.1093/nar/gkac1065.
29. Lin M, Whitmire S, Chen J, et al. Effects of short indels on protein structure and function in human genomes. Sci Rep. 2017;7. doi:10.1038/s41598-017-09287-x.
30. Boytsov SA, Drapkina OM, Shlyakhto EV, et al. Epidemiology of Cardiovascular Diseases and their Risk Factors in Regions of Russian Federation (ESSE-RF) study. Ten years later. Cardiovascular Therapy and Prevention. 2021;20(5):3007. (In Russ.) Бойцов С. А., Драпкина О. М., Шляхто Е. В. и др. Исследование ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации). Десять лет спустя. Кардиоваскулярная терапия и профилактика. 2021;20(5):3007. doi:10.15829/1728-8800-2021-3007.
31. Kopylova OV, Ershova AI, Pokrovskaya MS, et al. Population-nosological research biobank of the National Medical Research Center for Therapy and Preventive Medicine: analysis of biosamples, principles of collecting and storing information. Cardiovascular Therapy and Prevention. 2021;20(8):3119. (In Russ.) Копылова О. В., Ершова А. И., Покровская М. С. и др. Популяционно-нозологический исследовательский биобанк "НМИЦ ТПМ": анализ коллекций биообразцов, принципы сбора и хранения информации. Кардиоваскулярная терапия и профилактика. 2021;20(8):3119. doi:10.15829/1728-8800-2021-3119.
32. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412-9. doi:10.1093/nar/gkaa913.
33. Montgomery SB, Goode DL, Kvikstad E, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 2013;23(5):749-61. doi:10.1101/gr.148718.112.
34. Yue Z, Xiang Y, Chen G, et al. PredinID: predicting pathogenic in-frame indels in human through graph convolution neural network with graph sampling technique. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2023;20(5):3226-33. doi:10.1109/TCBB.2023.3266232.