

ОБЩИЕ ВОПРОСЫ ОТОРИНОЛАРИНГОЛОГИИ

Научная статья

УДК 575.112:004.438:577.113.4

<https://doi.org/10.18692/1810-4800-2025-4-13-27>

Медицинская биоинформатика: базовые операции с нуклеотидными последовательностями в программной среде R

А. А. Корнеев¹, Ю. К. Янов², В. В. Дворянчиков³, Е. Э. Вяземская⁴, А. Ю. Медведева⁵

^{1,3,4,5} Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи,
Санкт-Петербург, 190013, Российская Федерация

² Военно-медицинская академия им. С. М. Кирова, Санкт-Петербург, 194044, Российская Федерация

² Северо-Западный государственный медицинский университет имени И. И. Мечникова,
Санкт-Петербург, 191015, Российская Федерация

¹ korneyenkov@gmail.com, <https://orcid.org/0000-0001-5870-8042>

² 9153864@mail.ru, <https://orcid.org/0000-0001-9195-128X>

³ v.v.dvoryanчиков@mail.ru, <https://orcid.org/0000-0002-0925-7596>

⁴ vyazemskaya.elena@gmail.com, <https://orcid.org/0000-0002-4141-2226>

⁵ a.medvedeva@niilor.ru, <https://orcid.org/0009-0002-6921-5299>

Введение. Развитие концепции персонализированной медицины и внедрение передовых технологий требуют повышения уровня подготовки врачей в области медицинской биоинформатики, что приобретает особую значимость при диагностике и лечении генетических синдромов в оториноларингологии.

Цель исследования. Обсуждение решения типовых задач биоинформатики, связанных с первичной программной обработкой биологических последовательностей (нуклеотидных и аминокислотных) для получения и использования медицинских знаний о генотип-фенотипических ассоциациях в концепции персонализированной медицины. **Материалы и методы исследования.** Данные литературных источников, описания известных технологий работы с генетическими последовательностями, учебные материалы мировых авторитетных образовательных организаций, а также собственный опыт внедрения воспроизводимых исследований в научную деятельность и медицинское образование. **Результаты.** В исследовании предложены базовые вычислительные задачи, воспроизводящие функции работы с последовательностями нуклеотидов и аминокислот. **Обсуждения.** Знания и навыки, полученные с помощью этих задач в рамках обучения биоинформатике, обеспечивают понимание основных операций с последовательностями и их практическое применение в создании информативных диагностических технологий в концепции персонализированной медицины. Для решения задач биоинформатики предложено использовать язык программирования R и пакеты программ проекта Bioconductor, предоставляющего ресурсы для исследований в области медицинской биоинформатики.

Ключевые слова: биоинформатика, обучение биоинформатике, типовые задачи биоинформатики, воспроизводимые исследования, язык программирования R, нуклеотидные последовательности, генетические синдромы в оториноларингологии, проект Bioconductor, персонализированная медицина, персонализированные диагностические технологии, медицинская биоинформатика

Для цитирования: Корнеев А. А., Янов Ю. К., Дворянчиков В. В., Вяземская Е. Э., Медведева А. Ю. Медицинская биоинформатика: базовые операции с нуклеотидными последовательностями в программной среде R. *Российская оториноларингология*. 2025;24(4):13-27. <https://doi.org/10.18692/1810-4800-2025-4-13-27>

Science article

Medical bioinformatics: basic operations with nucleotide sequences in R software environment

A. A. Korneenkov¹, Yu. K. Yanov², V. V. Dvoryanchikov³, E. E. Vyazemskaya⁴, A. Yu. Medvedeva⁵^{1,3,4,5} Saint Petersburg Institute of Ear, Throat, Nose and Speech,
Saint Petersburg, 190013, Russian Federation² Kirov Military Medical Academy, Saint Petersburg, 194044, Russian Federation² Mechnikov North-Western State Medical University, Saint Petersburg, 195067, Russian Federation¹ korneenkov@gmail.com✉, <https://orcid.org/0000-0001-5870-8042>² 9153864@mail.ru, <https://orcid.org/0000-0001-9195-128X>³ v.v.dvoryanchikov@mail.ru, <https://orcid.org/0000-0002-0925-7596>⁴ vyazemskaya.elena@gmail.com, <https://orcid.org/0000-0002-4141-2226>⁵ a.medvedeva@niilor.ru, <https://orcid.org/0009-0002-6921-5299>

Abstract. The advancement of personalized medicine and the integration of cutting-edge technologies necessitate Introduction. The advancement of the concept of personalized medicine and the integration of cutting-edge technologies necessitate enhancing physicians' expertise in the field of medical bioinformatics, which is of particular importance in the diagnosis and treatment of genetic syndromes in otorhinolaryngology.

Objective. To discuss solutions for typical bioinformatics tasks associated with the primary software processing of biological sequences (nucleotide and amino acid) to derive and apply medical knowledge about genotype-phenotype associations in the concept of personalized medicine. **Materials and methods.** The study draws on data from literary sources, descriptions of established genetic sequence analysis technologies, educational materials from globally recognized institutions, and the authors' firsthand experience in implementing reproducible research in scientific practice and medical education. **Results.** The study proposes foundational computational tasks replicating the functions for handling nucleotide sequences and amino acid sequences. **Discussions.** The knowledge and skills acquired through these tasks during bioinformatics training ensure comprehension of core operations and their practical application in developing informative diagnostic technologies. For task implementation, it is proposed to use the R programming language and software packages from the Bioconductor project, which provides resources medical bioinformatics research.

Keywords: bioinformatics, bioinformatics training, typical bioinformatics tasks, reproducible research, R programming language, nucleotide sequences, genetic syndromes in otolaryngology, Bioconductor project, personalized medicine, personalized diagnostic technologies, medical bioinformatics

For citation: Korneenkov A. A., Yanov Yu. K., Dvoryanchikov V. V., Vyazemskaya E. E., Medvedeva A. Yu. Medical bioinformatics: basic operations with nucleotide sequences in R software environment. *Russian Otorhinolaryngology*. 2025;24(4):13-27. (In Russ.) <https://doi.org/10.18692/1810-4800-2025-4-13-27>

Введение

Актуальность изучения медицинской биоинформатики в высших учебных заведениях обусловлена ростом популярности концепции персонализированной медицины (англ., personalized medicine). Эта концепция стимулирует разработку и реализацию специализированных моделей, инициатив, программ и рекомендаций. Цель данных мероприятий — активное продвижение новой парадигмы персонализированного подхода в клинической практике.

Персонализированная медицина, которая означает выбор лечения, наиболее подходящего для конкретного человека, включает интеграцию и трансляцию нескольких новых технологий в клиническую помощь пациентам [1]. В первую очередь это технологии молекулярной биологии, геномной медицины, хотя учитываются также

и многие негеномные факторы. Примером развития персонализированной медицины служит модель «Медицина 4П», представляющая мнемоническое обозначение первых букв основополагающих принципов: персонализация, предсказательность, превентивность и партисипативность. Основная цель данной модели заключается в раннем выявлении заболеваний еще до клинических проявлений и создании комплексов профилактических мер [2, 3]. Эти принципы не только носят рекомендательный характер, но и реализованы в виде руководящих документов. Так, Приказом Министерства здравоохранения Российской Федерации официально утверждена Концепция предиктивной, превентивной и персонализированной медицины [4]. А Министерством труда и социальной защиты Российской Федерации утвержден профессиональный стандарт «Врач-

лечебник (врач-терапевт участковый)», который включает обобщенную трудовую функцию «Назначение лечения и контроль его эффективности и безопасности» (код А/03.7). Среди конкретных трудовых функций, входящих в данную группу, значится следующая: «Организация персонализированного лечения пациента, в том числе беременных женщин, пациентов пожилого и старческого возраста, оценка эффективности и безопасности лечения» [5].

Компьютерные технологии, составляющие основу персонализированной медицины, развиваются очень быстро, вследствие чего традиционное медицинское образование не успевает за ними. Многие из этих технологий стали самостоятельными бурно развивающимися научными направлениями, как, например, биоинформатика. Связь между молекулярными данными, относящимися к пациенту, и фенотипом его заболевания сложна и не может быть определена вручную. Методы компьютерной биоинформатики играют центральную роль в интерпретации молекулярных данных, а также в качестве инструмента для предоставления рекомендаций практикующему врачу. Современная биоинформатика является важнейшим компонентом в фундаментальных исследованиях, в разработке новых концепций диагностики и терапии, а также в клинической практике [6].

Все это говорит о том, что современный врач должен знать об этих вычислительных технологиях и разбираться в них. Тем не менее, как правило, врачи знакомы с этими технологиями очень поверхностно, в силу их разнообразия и постоянного совершенствования [7]. Ранее в статье «Вопросы интеграции медицинских биоинформатических технологий в оториноларингологию: проблемы и программные решения» мы рассмотрели общие аспекты интеграции биоинформатических знаний в медицину [8]. Каждый шаг в этом направлении добавляет понимания применимости биоинформатических технологий в задачах персонализированной медицины.

Целью данного исследования является рассмотрение типовых биоинформатических задач, связанных с подготовкой и первичной обработкой биологических последовательностей (нуклеотидных и аминокислотных): изучение начинается с процедуры загрузки данных в вычислительную среду R и охватывает базовые операции с геномными признаками, а также подходы к поиску генов.

В этой статье рассмотрены «рецепты» (в первую очередь образовательные) создания из биологических последовательностей объектов программной среды R для использования ее возможностей по различным манипуляциям (импорта, экспорта, создания подмножества), опе-

рациям с геномными диапазонами, геномными признаками (описание, создание подмножества, получение последовательности), поиску старт- и стоп-кодона, мотивов, поиску открытой рамки считывания для идентификации возможного присутствия гена.

Предполагается, что эти знания и навыки обеспечат понимание состава и структуры генетического кода человека и впоследствии будут использоваться специалистами для исследования генотип-фенотипических ассоциаций и их применения в геномной и персонализированной медицине.

Материалы и методы

В качестве материалов для создания типовых вычислительных задач с последовательностями использованы данные литературных источников, описания известных технологий работы с генетическими последовательностями, свободнодоступные учебные материалы мировых авторитетных образовательных организаций, а также собственный опыт внедрения воспроизводимых исследований в научную деятельность и медицинское образование.

Мы рассматриваем эти задачи в концепции тесного информационного взаимодействия персонализированной медицины [9], вычислительной системной биологии [10, 11] и биоинформатики [12]. Такой многоаспектный подход позволяет опираться на разнообразную генетическую, молекулярную и клиническую информацию для разработки более эффективных и безопасных стратегий диагностики и лечения заболеваний человека.

Для формирования исходных данных в этой работе использованы базы генетической информации, наборы данных известных онлайн-ресурсов (NCBI, Ensembl), а также встроенные наборы данных R пакетов репозитория Bioconductor. Основу этих данных составили биологические последовательности — это первичная структура биологических макромолекул — цепочки молекул белков и ДНК/РНК (например, один нуклеотид или одна аминокислота).

Как уже указывалось, в качестве программной среды для программирования и решения вычислительных задач работы с последовательностями использовались среда R и пакеты репозитория Bioconductor — одного из самых известных биоинформационных проектов на основе программ с открытым исходным кодом. В среде R пакет — это расширение к языку статистического программирования R, которое содержит код, данные и документацию в стандартизированном формате. Пакеты R имеют название, позволяют воспроизводить код и распространять его среди остальных пользователей. Предполагается, что начало работы с объектами геномных последо-

вательностей, их импорт в программную среду R закладывают понимание содержания более сложных задач: идентификации последовательностей, определения вариантов и генотип-фенотипических ассоциаций.

Нуклеотидные и аминокислотные последовательности состоят из огромного числа символов, эффективно обрабатывать которые можно лишь в специальных программах. Как правило, у пользователя есть возможность выбора технологии и программ. В подавляющем большинстве общих случаев клинической практики достаточно web-ресурсов NCBI, Ensembl, UCSC, предлагающих доступ к данным о последовательности генома и различные встроенные инструменты работы с ними. Геномные браузеры удобны и эффективны для визуализации и просмотра целых геномов с помощью аннотированных данных. Они позволяют изучать предсказание генов, структуру генов, белки, экспрессию, регуляцию, вариации и проводить сравнительный анализ [13]. Тем не менее web-ресурсы имеют ограничения по объему загрузки пользовательских данных и функциональным возможностям.

Поэтому, если мы рассматриваем генетическую информацию как часть вычислительных алгоритмов персонализированной медицины, управления здоровьем, клинической информатики, эпидемиологии и т. п., сопоставление с другой информацией лучше проводить в единых средах. Это усложняет задачу, но делает решение более эффективным. В этом исследовании описаны решения поставленных задач в программной среде R, которую широко применяют для статистического анализа данных биоинформатики, визуализации и интерпретации сложных биологических данных, для извлечения и анализа информации о биологических путях, а также для сопоставления данных экспрессии генов с этими путями. К преимуществам R относят широкий спектр статистических инструментов и пакетов, активное сообщество пользователей и разработчиков, которые постоянно создают новые инструменты и пакеты, специфичные для потребностей исследований в области биоинформатики; кроме того, поддерживается возможность работы в разных операционных системах [14–19].

В среде R последовательности импортируются в специализированные объекты-контейнеры, которые позволяют использовать весь спектр вычислительных возможностей этой среды и программных пакетов [8]. Проект Bioconductor предоставляет множество полезных пакетов для эффективной обработки геномных данных. Так, пакет Biostrings определяет контейнеры и предоставляет функции для данных о геномной последовательности. В пакете Biostrings есть два основных типа контейнеров: контейнер для одной

строки последовательности (например, хромосомы или одной короткой последовательности) и контейнер для набора строк (например, набора коротких последовательностей). Пакет BSgenome и другие пакеты геномных данных предоставляют полные геномные последовательности и инструменты для хранения, обработки и анализа. Пакет GenomicRanges обрабатывает наборы геномных интервалов. Пакет GenomicFeatures предоставляет функции для извлечения и управления геномными признаками из общедоступных баз данных.

В качестве типовых вычислительных задач, позволяющих последовательно перейти от ввода исходных данных к получению значимой информации (паттернов, предполагаемых генов-кандидатов), выбраны следующие:

- 1) преобразование биологических последовательностей (ДНК, РНК и аминокислот) в контейнеры Biostrings для использования функций R;
- 2) использование системы координат последовательностей в геноме, понимание методов работы с диапазонами (IRanges, GRanges);
- 3) геномные признаки, извлечение последовательности по координатам геномных признаков;
- 4) поиск в последовательности определенных паттернов (старт-кодонов, стоп-кодонов и др.) или мотивов;
- 5) поиск открытой рамки считывания (ORF) в контексте предсказания наличия в данном участке гена, кодирующего некий полипептид.

Хотя задачи условны (например, поиск ORF в последовательности гена), они закладывают понимание и навыки как основу для дальнейших исследований ассоциаций генотипа и фенотипа человека. Ниже приводятся сначала краткое теоретическое описание этих задач, а затем решения в выбранной вычислительной среде R.

1. Преобразование биологических последовательностей (ДНК, РНК и аминокислот) в контейнеры Biostrings для использования функций R

Прежде чем использовать биологические последовательности в программной среде, необходимо разобраться в том, из каких областей (участков) состоят эти последовательности, как они обозначаются и как они соотносятся друг с другом. Без знания этих обозначения сложно понять, о чем идет речь в программном коде и визуальных средствах отображения: транскрипт, 5'- или 3'-некодирующая область, кодирующая последовательность, старт-кодон, стоп-кодон, экзон, интрон, кэп и многие другие.

Центральная догма молекулярной биологии определяет правило реализации генетической информации: информация передается от нуклеиновых кислот к белку, но не в обратном направлении. В живых организмах встречаются три вида линейных полимеров — ДНК, РНК и белок. Они

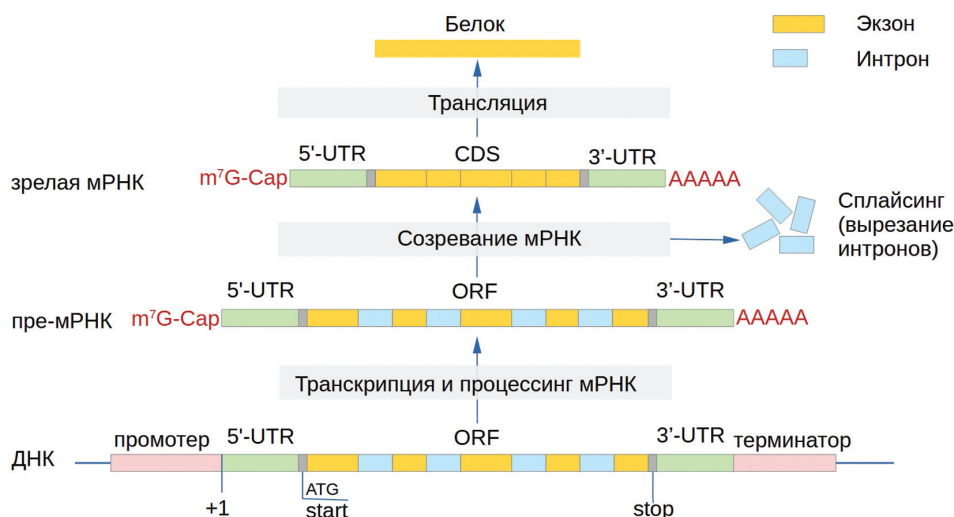


Рис. 1. Схема реализации генетического кода в процессе синтеза белка
Fig. 1. The process of genetic code translation in protein synthesis

собираются путем последовательного присоединения друг к другу отдельных элементов — мономеров. Последовательность мономеров кодирует информацию, правила передачи которой описываются центральной догмой. Общие способы передачи информации известны и включает репликацию (удвоение ДНК), транскрипцию (копирование информации ДНК на синтезируемую молекулу мРНК) и трансляцию (синтез белка путем считывания триплетов нуклеотидов на мРНК). На рис. 1 представлена обобщенная схема реализации генетического кода в процессе синтеза белка [20].

На рис. 1 используются следующие обозначения: ORF (англ. open reading frame) — открытая рамка считывания; CDS (англ. coding sequence) — кодирующая последовательность; M7G-Cap — защита мРНК от деградации, участие в инициации трансляции; 5'-UTR — лидер и трейлер 5', нетранслируемая область — это область информационной РНК (мРНК), которая находится непосредственно выше иницирующего кодона; 3'-UTR — нетранслируемая область, часть информационной РНК (мРНК), которая следует сразу за кодоном терминации трансляции. Стоп-сигнал гена находится непосредственно перед 3'-UTR; сплайсинг (англ. splicing — соединение, сращивание) — это образование функционально активных молекул РНК из их предшественников путем удаления отдельных фрагментов.

Для примера рассмотрим последовательности двух известных генов SRY и BRCA2. Ген SRY (англ. sex-determining region Y) принимает участие в развитии организма по мужскому типу и локализован на коротком плече Y-хромосомы, его единственный экзон состоит из 897 пар нуклеотидов. Ввиду того что последовательность гена SRY очень короткая, она очень удобна для визуализации и понимания ее участков. Другой ген имеет большую

длину последовательности и исследовательский потенциал в онкологии — ген BRCA2. Он кодирует белок, который участвует в репарации (восстановлении) ДНК, регуляции клеточного цикла и поддержании генетической стабильности.

Последовательность гена SRY доступна для скачивания в геномном браузере проекта Ensembl [21]. На рис. 2 приводится копия отображаемой последовательности гена SRY с выделением отдельных ее участков и их обозначением в стиле заголовков файлов fasta, как если бы эти участки были выгружены в отдельные файлы fasta. Данную последовательность можно найти по web-адресу, закодированному в QR-код на рисунке и содержащему Ensembl-идентификатор гена SRY — ENSG00000184895 и идентификатор его единственного транскрипта — ENST00000383070.

Последовательности гена SRY (через разделитель «:» указаны сборка генома, хромосома, координаты участка последовательности, числовое обозначение обратной цепи — chromosome:GRCh38:Y:2786255:2788282:-1) включает последовательность ДНК, комплементарную зрелой мРНК (ENST00000383070.2 SRY-201 cDNA:protein_coding — обозначена пунктиром черного цвета), нетранслируемые участки (>SRY-201 utr5:protein_coding и >SRY-201 utr3:protein_coding — в рамке синего цвета) и кодирующую последовательность (CDS, англ. — coding sequence), которая соответствует последовательности аминокислот в белке (>SRY-201 cds:protein_coding — символы красного цвета).

2. Использование системы координат последовательностей в геноме, понимание методов работы с диапазонами (IRanges, GRanges)

Чтобы обратиться к конкретным частям генома, например геномному признаку, используется позиционная система координат, называемая ге-

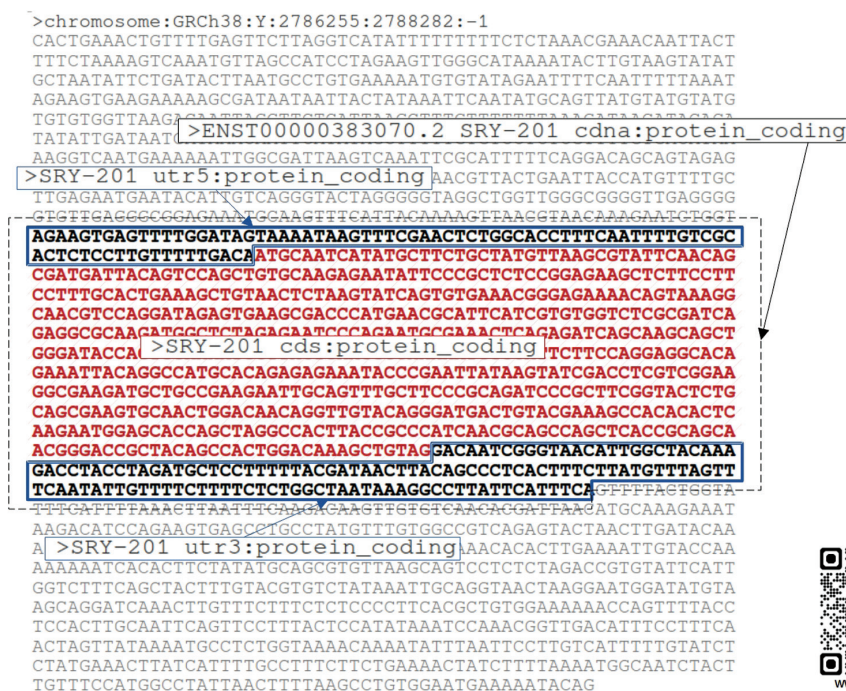


Рис. 2. Последовательность гена SRY с обозначениями ее отдельных участков
Fig. 2. The nucleotide sequence of the SRY gene with labeled functional domains

номными координатами. Чтобы представить геномную координату, необходимо знать тип генома (англ. genome type), номер хромосомы (англ. chromosome number), тип цепи (англ. strand type) и пронумерованную позицию (англ. numbered position) пары оснований (англ. base-pair) в последовательности нуклеотидов, например, в хромосоме.

В международном медицинском сообществе существуют различные соглашения для представления координат генома. Первая пара оснований в геноме может, например, начинать либо с индекса 0 (o-based), который является нормальным в мире информатики, либо с индекса 1 (i-based), который может быть более естественным для других областей. Это разные способы идентификации нуклеотидов или аминокислот внутри последовательности [22]. В системе с отсчетом от 1 первая единица полимера (например, первый нуклеотид) считается под номером 1. В системе отсчета от 0 цифра 0 — обозначает положение перед первым нуклеотидом. Разные базы данных и типы файлов могут использовать разные системы. В файлах геномных данных формата GFF, SAM и

VCF используются координаты, отсчитываемые от 1, а в файлах BED и BAM — координаты, отсчитываемые от 0.

В таблице показаны примеры отображения в последовательности (англ. sequence) координат нуклеотида/ов (англ. nucleotide/s) для случаев: одного нуклеотида (англ. single nucleotide), диапазона нуклеотидов (англ. range of nucleotides), однонуклеотидного варианта (англ. single nucleotide variant), делеции (англ. deletion), вставки (англ. insertion) в разных координатных системах (с отсчетом от 1 или 0).

3. Геномные признаки, извлечение последовательности по координатам геномных признаков

Геномный признак (англ. genome features) — часть генома, имеющая определенную особенность, черту, играющую какую-то, возможно, важную роль. Геномный признак — это общий термин, описывающий одну или несколько областей генома с некоторой биологической функцией или особенностью. Геномным признаком может быть, например, конкретный ген, кодирующая белок последовательность, набор полиморфизмов одно-



Рис. 3. Представление геномных координат в разных системах отсчета
Fig. 3. Genomic coordinate systems and their frame-of-reference variations

Примеры отображения в последовательности координат нуклеотидов

Examples of nucleotide coordinate mapping in the sequence

Элемент последовательности	Координаты			
	Система координат с началом в точке 1 (файлы BED и BAM)	Система координат с началом в точке 0 (файлы VCF, GFF и SAM)	Нуклеотид(ы)	Последовательность
Одиночный нуклеотид	5-5	4-5	C	AGTACGTC A AG
Диапазон нуклеотидов	4-6	3-6	ACG	AGTACGTC ACG
Однонуклеотидный вариант	5-5	4-5	C/G	AGTACGTC A AG
Делеция	5-5	4-5	C/-	AGTA-GTCAAG
Вставка	5-6	5-5	-/TAA	AGTAC TAA AGTCAAG
Примечание. Элемент последовательности обозначен зеленым цветом.				

нуклеотидных оснований (англ. single nucleotide polymorphism, SNP) или экзоны гена.

С учетом бурного развития технологий секвенирования нового поколения (англ. next generation sequencing, NGS), позволяющих получить первичную структуру нуклеотидной последовательности ДНК и РНК, перечень полезных с медицинской точки зрения геномных признаков постоянно растет. В зависимости от приложения геномными признаками могут быть экзоны, интроны, гены, области промотора, тела генов или другие части генома. В ходе последующего анализа информация об объекте исследования (образце, пациенте, заболевании и т. д.) сопоставляется с каждым геномным признаком. Анализ геномных признаков играет решающую роль в понимании сложных механизмов, лежащих в основе различных биологических процессов и заболеваний.

4. Поиск паттернов (старт-кодонах, стоп-кодонах и др.) или мотивов

В биоинформатике особое место занимает задача поиска определенного набора символов в заданной последовательности нуклеотидов или аминокислот. Искомый набор символов или, другими словами, искомая последовательность может иметь разные свойства, значения и обозначения — паттерн, мотив или др.

В функциях R, как правило, искомая строка или последовательность обозначаются аргументом «паттерн» (англ. pattern), а последовательность, в которой производится поиск, — аргументом «субъект» (англ. subject).

Паттерном в общем смысле часто обозначается неизвестный шаблон, который часто встречается в наборе последовательностей. При работе с ДНК-последовательностями паттерны могут включать мутации, вставки или делеции нуклеотидов. В отличие от паттерна мотив — это образец последовательности нуклеиновых кислот, который имеет некоторое биологическое значение, например,

является участком связывания ДНК с регуляторным белком. Обычно мотив короткий (длиной от 5 до 20 пар оснований) и повторяется в разных генах или несколько раз в пределах одного гена. Под мотивом иногда подразумевают не конкретную последовательность, а каким-либо образом описанный спектр последовательностей, каждая из которых способна выполнять определенную биологическую функцию данного мотива. Другими словами, мотив может быть представлен как точная последовательность, такая как TGACGTCA, или это может быть «вырожденная» консенсусная последовательность (англ. consensus sequence), допускающая неоднозначные, подстановочные символы, такие как R для A или G, как N, которые могут относиться к любому символу. Они также могут указывать более конкретное подмножество, такое как Y, которое определяет пиримидины C и T (см. коды IUPAC для нуклеотидов в справочнике Bioinformatics.org [23]).

Мотивы широко распространены среди живых организмов и выполняют важные биологические функции. Например, нуклеотидные мотивы регулируют процессы транскрипции и трансляции, тогда как белковые мотивы участвуют в посттрансляционной модификации, определяют местоположение белков внутри клетки и влияют на их функциональность. Поиск мотивов используется в биоинформатике для предсказания функций генов и белков, построения карт регуляции, важен для многих задач генной инженерии и молекулярной биологии в целом.

Для упрощения в этой работе в качестве искомого набора символов выбраны старт- и стоп-кодонах. Кодон (триплет) — дискретная единица генетического кода, участок ДНК или матричной РНК (мРНК), а также геномной РНК вирусов, состоящий из трех расположенных рядом нуклеотидов. Он кодирует один аминокислотный остаток или служит сигналом для начала (инициирующий кодон) или завершения (нонсенс-кодон, стоп-

кодон, терминирующий кодон) синтеза белка на рибосоме в процессе трансляции.

Стартовый кодон (англ. start codon) — это первый кодон транскрипта матричной РНК (мРНК), транслируемого рибосомой. Стартовый кодон всегда кодирует метионин у эукариот и архей, а также N-формилметионин (fMet) у бактерий, митохондрий и пластид. Типичным примером стартового кодона в мРНК является AUG, а для стоп-кодона — UAG, UGA и UAA. В ДНК старт-кодону соответствует ATG и стоп-кодонам — TAG, TGA и TAA.

Далее будет изучен поиск отдельных стартовых и стоп-кодонов в заранее созданной последовательности ДНК с помощью функции `vmatchPattern()` пакета `Biostrings`, чтобы вернуть общие результаты для всех стартовых и стоп-кодонов.

5. Поиск открытой рамки считывания (ORF) в контексте предсказания наличия в данном участке гена, кодирующего некий полипептид

Открытая рамка считывания (англ. open reading frame, ORF) — это последовательность нуклеотидов в составе ДНК или РНК, потенциально способная кодировать белок, не прерываемая стоп-кодонами. Она представляет интерес, например, для поиска генов в последовательности.

С программной точки зрения ORF — это последовательность, длина которой кратна трем и ограничена стоп-кодонами. Это определение может быть полезно в контексте транскриптомики и метагеномики, где стартовый или стоп-кодон может отсутствовать в полученных последовательностях. Такой ORF соответствует частям гена, а не всему гену целиком.

Биологическое значение нахождения открытой рамки считывания (ORF) заключается в том, что зачастую она используется в качестве части доказательств наличия гена в этой области. Длинные ORF наряду с другими данными часто используются для первоначальной идентификации в последовательности ДНК областей-кандидатов, кодирующих белок, или функциональных регионов, кодирующих РНК.

Наличие ORF не обязательно означает, что соответствующая область всегда транслируется в белок. Если рассмотреть случайно сгенерированную последовательность ДНК с равным процентом каждого нуклеотида, то стоп-кодон можно ожидать один раз на каждый 21 кодон. Поэтому некоторые авторы говорят, что ORF должна иметь минимальную длину, например 100 или 150 кодонов. Сама по себе даже длинная открытая рамка считывания не является убедительным доказательством присутствия гена.

Хотя не каждая ORF потенциально может транслироваться или быть функциональной, для поиска новых генов мы должны сначала идентифицировать потенциальные ORF.

Каждая область ДНК имеет шесть возможных рамок считывания, по три в каждом направлении. Так как мРНК транслируется кодонами (триплетами нуклеотидов), то для данной последовательности ДНК существуют три потенциальные рамки считывания. Кроме того, ДНК состоит из двух цепей (нитей), и каждая из них может содержать ген. Таким образом, всего существуют шесть рамок считывания (2×3).

Используемая рамка считывания определяет, какие аминокислоты будут кодироваться геном. Обычно при трансляции гена используется только одна рамка считывания (у эукариот), и это часто самая длинная открытая рамка считывания [24].

Открытая рамка считывания у большинства видов начинается с триплета ATG и заканчивается стоп-кодоном (TAA, TAG и TGA). Для идентификации ORF необходимо найти последовательность, соответствующую стартовому кодону — ATG. Далее следует читать эту последовательность в виде триплетов нуклеотидов до тех пор, пока не будет достигнут стоп-кодон. Чем длиннее последовательность, тем выше вероятность того, что она соответствует открытой рамке считывания.

Как только открытая рамка считывания станет известна, последовательность ДНК может быть транслирована в соответствующую ей аминокислотную последовательность. Последовательности генов в значительной степени консервативны, поэтому, если последовательность ORF присутствует в нескольких геномах, она, скорее всего, представляет собой ген.

Результаты

Ниже описаны пять задач, решение которых дает базовое понимание работы с последовательностями (в данном случае нуклеотидными). В программном коде используются базовые функции R, готовые к использованию после установки R среды (base) и функции пакетов, которые необходимо установить: `Biostrings`, `GenomicRanges`, `rtracklayer`, `ORFik`. Для упрощения понимания приведенного ниже программного кода дано краткое описание этих функций в приложении.

Чтобы работать с последовательностями в программной среде R и осуществлять типичные биоинформационные функции поиска, выравнивания, аннотирования и т. д., они должны быть импортированы в среду R внутри специализированных контейнеров. Для проекта `Bioconductor` чаще всего из нуклеотидных и аминокислотных последовательностей формируют объекты классов `XString` или `XStringSet`, где в этих обозначениях вместо X в зависимости от того, о каких именно последовательностях идет речь: ДНК, РНК, или аминокислотных, используются соответственно вставки `DNA`, `RNA`, `AA`. Например для хранения последовательности ДНК — `DNAString`, последовательности РНК —

RNAString или последовательности аминокислот — AAStrng. Аналогично для хранения набора объектов последовательностей используется класс DNAStringSet (или RNAStringSet, или AAStrngSet), который представляет собой контейнер набора объектов DNAString (или RNAString, или AAStrng). Методы работы с этими объектами собраны в пакете Biostrings.

Другие важные классы объектов, поддерживаемые в проекте Bioconductor, связаны с координатными представлениями различных геномных признаков — это IRanges и GRanges. IRanges — это

структура данных, которая позволяет работать с начальными и конечными позициями диапазонов, но не содержит информации о хромосоме, цепи ДНК и других важных в геномных приложениях данных. GRanges — это объект, который содержит IRanges и дополнительную информацию о хромосоме и цепи ДНК. Кроме того, GRanges позволяет хранить метаданные, которые аннотируют определенный регион генома. Таким образом, IRanges — это объект более общей структуры, а GRanges — специализированная структура для работы с геномными данными.

Задача 1. Преобразование биологических последовательностей (ДНК, РНК и аминокислот) в R-объекты классов XString или XStringSet для использования функций R

```
> seqs_13chr <- readDNAStringSet('data/Homo_sapiens.GRCh38.dna.chromosome.13.fa.gz',
format="fasta")
> seqs_13chr
DNAStringSet object of length 1:
      width seq                                     names
[1] 114364328 NNNNNNNNNNNNNNNNNNNN...NNNNNNNNNNNNNNNNNNNN 13 dna:chromosome...
> names(seqs_13chr)
[1] "13 dna:chromosome chromosome:GRCh38:13:1:114364328:1 REF"
> # Извлечение номера хромосомы.
> names(seqs_13chr) <- sub(" .*", "", names(seqs_13chr))
> names(seqs_13chr)
[1] "13"
> myseqI <- subseq(seqs_13chr, 30000015, 30000030)
> names(myseqI) <- "seqI"
> myseqI
DNAStringSet object of length 1:
      width seq                                     names
[1] 16 CCCCTGAGCAAGACGC                             seqI
```

С помощью функции `readDNAStringSet()` из пакета `Biostings` из файла формата `fasta` считывается нуклеотидная последовательность 13-й хромосомы в сборке генома человека `GRCh38` в R-объект `seqs_13chr`. Так как имя хромосомы содержит дополнительную информацию, с помощью функции `sub()` вся информация, кроме номера хромосомы, заменяется пустым символом, т. е. удаляется. Из этого объекта последовательности с помощью функции `subseq()` выводится ее часть, ограниченная заданными позициями начала (`start`), конца (`end`) и длины (`width`), в объект `myseq1`. Затем с помощью функции `names()` этому объекту последовательности присваивается имя `seq1`.

Задача 2. Использование системы координат последовательностей в геноме, понимание методов работы с диапазонами (IRanges, GRanges)

```
> rangel <- GRanges("seqI", IRanges(start=3, end=5), strand="+") # Создание диапазона
> rangel
GRanges object with 1 range and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges>  <Rle>
[1]      seqI          3-5      +
-----
      seqinfo: 1 sequence from an unspecified genome; no seqlengths
> getSeq(myseqI, rangel) # Получение последовательности заданного диапазона
DNAStringSet object of length 1:
      width seq
[1]      3 CCT
> seqnames(rangel) # Имя
factor-Rle of length 1 with 1 run
      Lengths: 1
      Values : seqI
Levels(1): seqI
> start(rangel) # Начало
[1] 3
> end(rangel) # Конец
[1] 5
> strand(rangel) # Смысловая/антисмысловая цепь
factor-Rle of length 1 with 1 run
      Lengths: 1
      Values : +
Levels(3): + - *
```

В представленном выше коде формируется объект `range1` класса `GRanges`. Объект `range1` представляет собой описание признаков, по которым эта последовательность может быть определена: координаты ее начала (`start`) и конца (`end`); из какой (`strand`) цепи (+/-) ДНК эта последовательность; из последовательности какого имени. Создание объекта класса `GRanges` производится с помощью одноименной функции `GRanges()`. Для наглядности мы извлекаем короткую (три нуклеотида) последовательность из уже полученного в предыдущей задаче объекта `myseq1` с помощью функции `getSeq()`. В получении данных об объекте `GRanges` можно использовать так называемые экстракторы — функции по названию выводющих атрибутов объекта класса `GRanges`: `start()`, `end()`, `strand()`.

Задача 3. Геномные признаки, извлечение последовательности по координатам геномных признаков (`features`)

В этой задаче информация о геномных признаках для последовательности известного гена `BRCA2`, кодирующего белок, который участвует в репарации (восстановлении) ДНК, регуляции клеточного цикла и поддержании генетической стабильности, содержится в архивном GFF3-файле — `BRCA2_GFF3.gff3.gz`. GFF3 (generic feature format версии 3) — это формат файла для хранения геномных данных, описывающих гены и другие геномные признаки последовательностей ДНК, РНК и белков, `gene`, `transcript`, `intron`, `CDS`, `exon`. Стандарт описания (спецификация) GFF3 приведен на web-ресурсе <http://www.sequenceontology.org/gff3.shtml> [25].

```
> # Загрузка геномных признаков (features)
> gff3_file <- 'data/BRCA2_GFF3.gff3.gz'
> gff3 <- import(gff3_file)
> # Задаем подмножество строк и столбцов (переменных)
> # Просмотреть типы диапазонов:
> unique(gff3$type)
[1] gene transcript intron CDS exon
Levels: gene transcript intron CDS exon
> # Создаем подмножество из диапазонов с признаком transcript:
> rowidx <- gff3$type == "transcript"
> colidx <- c("type", "Name", "ID", "biotype")
> features <- gff3[rowidx, colidx]
> names(features) <- features$gene_id
> features$gene_id <- NULL
> # Выводим на экран информацию о первых 3 из всех найденных транскриптов (36)
> features[1:3]
GRanges object with 3 ranges and 4 metadata columns:
      seqnames      ranges strand |      type
      <Rle>      <IRanges> <Rle> | <factor>
[1]      13 32303699-32315363    - | transcript
[2]      13 32303771-32311954    - | transcript
[3]      13 32384660-32386108    + | transcript
      Name      ID      biotype
      <character> <character> <character>
[1] ENST00000533490.7 ENST00000533490.7 protein_coding
[2] ENST00000345108.6 ENST00000345108.6 protein_coding
[3] ENST00000400497.2 ENST00000400497.2 processed_pseudogene
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> tr_ENST00000380152.8 <- subset(features, Name == "ENST00000380152.8")
> tr_ENST00000380152.8
GRanges object with 1 range and 4 metadata columns:
      seqnames      ranges strand |      type
      <Rle>      <IRanges> <Rle> | <factor>
[1]      13 32315508-32400268    + | transcript
      Name      ID      biotype
      <character> <character> <character>
[1] ENST00000380152.8 ENST00000380152.8 protein_coding
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> feat_seq_tr_1 <- getSeq(seqs_13chr, tr_ENST00000380152.8)
> feat_seq_tr_1
DNAStringSet object of length 1:
width seq
[1] 84761 AAGGCGGAGCCGCTGTGGCACTGCT...CAAGAATAAATCTTTTAAAAAATTGA
```

В коде с помощью функции `import()` пакета `rtracklayer` создается объект `gff3` класса `GRanges`. Просмотреть, какие в этом объекте содержатся уникальные типы геномных признаков, можно с помощью функции `unique()`. Далее в качестве примера выбираются только геномные диапазоны, соот-

ветствующие транскриптам. Из этой информации формируется таблица с 36 диапазонами (общее количество диапазонов можно вывести, набрав имя объекта — features) и 4 выбранными переменными (в столбцах) метаданных: имя (Name), тип (type), идентификатор (ID) и биотип (biotype). Для иллюстрации выведены только три строки геномных признаков с помощью кода features[1:3]. Далее из списка выбраны с помощью функции subset() только транскрипт „ENST00000380152.8“ и его геномные координаты. Используя нуклеотидную последовательность 13-й хромосомы, которая была создана в первой задаче, по этим координатам с помощью getSeq() отбирается нуклеотидная последовательность этого транскрипта гена BRCA2 в объект feat_seq_tr_1.

Задача 4. Поиск в последовательности определенных паттернов (старт-кодонов, стоп-кодонов и др.) или мотивов

Для поиска заданных старт- и стоп-кодонов (в этом примере это ATG и TAG соответственно) используется функция vmatchPattern() из пакета Biostrings.

```
> start <- vmatchPattern("ATG", feat_seq_tr_1)
> start
MIndex object of length 1
$str_13
IRanges object with 1508 ranges and 0 metadata columns:
      start      end      width
  <integer> <integer> <integer>
[1]      292      294         3
...
[1508]   84701   84703         3

> stop <- vmatchPattern("TAG", feat_seq_tr_1)
> stop
MIndex object of length 1
$str_13
IRanges object with 1327 ranges and 0 metadata columns:
      start      end      width
  <integer> <integer> <integer>
[1]      163      165         3
...
[1327]   84605   84607         3
```

Она позволяет находить совпадения заданного шаблона (англ. pattern), т. е. ATG или TAG в наборе эталонных последовательностей (англ., subject). В результате применения этой функции создается объект класса IRanges с указанием начала (start), конца (end) и ширины (width) диапазонов, в которых обнаружены совпадения.

Задача 5. Поиск открытой рамки считывания (ORF) в контексте предсказания наличия в данном участке гена, кодирующего некий полипептид

```
> # положительная цепь
> pos <- findORFs(feat_seq_tr_1, startCodon = "ATG", minimumLength = 100)
> # отрицательная цепь
> neg <- findORFs(reverseComplement(feat_seq_tr_1),
+               startCodon = "ATG", minimumLength = 100)
> # создание объекта GRanges
> pos <- GRanges(pos, strand = "+")
> neg <- GRanges(neg, strand = "-")
> res <- c(pos, neg)
> # слияние их вместе и представление как GRangesList
> res <- split(res, seq.int(1, length(pos) + length(neg)))
> res[1:2]
GRangesList object of length 2:
$`1`
GRanges object with 1 range and 0 metadata columns:
      seqnames      ranges strand
    <Rle>    <IRanges>    <Rle>
[1]   tr_13 69475-69951      +
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

$`2`
GRanges object with 1 range and 0 metadata columns:
      seqnames      ranges strand
    <Rle>    <IRanges>    <Rle>
[1]   tr_13 70069-70416      +
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

```
> seqlevels(res) <- 'tr_13'
> gr <- unlist(res)
> names(gr) <- paste0("ORF_", seq.int(length(gr)), "_", seqnames(gr))
> orf_seqs <- getSeq(feet_seq_tr_1, gr)
> orf_seqs
DNAStringSet object of length 16:
      width seq                                     names
[1]    477 ATGGGCAGACTGGCAGAAACCCAG...CATATGCTCTCTCTTCTTACTGA ORF_1_tr_13
[2]    348 ATGAAGGAAGCTACCAACAAGCAG...AAGGCATAATTTCAAGGGATAA ORF_2_tr_13
[3]    309 ATGTTACCTTTTCAAAAACACCTG...GTGAAAGATATAAGGATAAATAA ORF_3_tr_13
[4]    306 ATGCTTACTAAGGATGCTCAATTT...ATACACTTTTATCTCTATGTAG ORF_4_tr_13
[5]    336 ATGAAGTTATTAACTGTCTCTTTT...AAAGGAGCATGCAAGCTTTTGA ORF_5_tr_13
...
[12]   348 ATGTTCCCATCTCAAGTGGGAGAA...GGTGGGGTGTCTCATTCTGTAA ORF_12_tr_13
[13]   369 CCGCGTTTATACGCTCTGTCTAA...TCATTGCAAAATGGAACAAG ORF_13_tr_13
[14]   309 TCTCTGATAGATATGGGTAAAGTG...GCTTCATACGATTGGCTCTGTG ORF_14_tr_13
[15]   312 TGAAGGGTTAGTTGAGACCATTC...TCTTCAGAGGTATCTACAACGT ORF_15_tr_13
[16]   402 CTCTACTCCAGTTGCACCTTTC...AGTGCTCAATAGAGATTGCTGA ORF_16_tr_13
```

Для поиска открытых рамок считывания используется функция `findORFs()` из пакета `ORFik`. Сначала в полученной ранее последовательности транскрипта ENST00000380152.8 функция `findORFs()` запрашивает все ORF на положительной (+) цепи ДНК в направлении 5'-3' с произвольно выбранной минимальной длиной ORF в 100 оснований. Возвращаемые координаты являются относительными к входным последовательностям. Чтобы выполнить поиск на отрицательной цепи ДНК для этой последовательности применяют функцию `reverseComplement()`, которая вычисляет обратное комплементарное соответствие нуклеотидной последовательности, другими словами ответную цепь (-).

На основе полученного с помощью функции `findORFs()` списка диапазонов `IRanges`, добавляя информацию о цепи (+ или -), создаются объекты `pos`, `neg` класса `GRanges`, а после их объединения и именования — в объект списка геномных диапазонов. На основе этих диапазонов функцией `getSeq()` извлекается набор последовательностей ORF в виде объекта `DNAStringSet`.

Обсуждение

Полученные результаты демонстрируют возможности программной среды R для работы с объектами геномных последовательностей. После решения приведенных воспроизводимых вычислительных задач с последовательностями пользователь получает возможность ответить на следующие вопросы.

1. Какие преимущества дает использование R и репозитория Bioconductor для решения биоинформационных задач?
2. В объекты каких классов импортируются последовательности нуклеотидов и аминокислот для работы с пакетами проекта Bioconductor? Какими функциями осуществляется этот импорт?
3. Что такое геномные признаки? В каких объектах R представляют геномные интервалы с координатами, а также с указанием цепи ДНК и метаинформацией?
4. Из каких участков состоит последовательность гена, как они обозначаются в популярных геномных браузерах? Что означают `cdna`, `cds`, 5'-UTR, 3'-UTR, `exon`, `intron`? Какую информацию содержат обозначения скачиваемых последовательностей (например, из `Ensembl`)?
5. Как в генетической последовательности осуществляется поиск определенного набора символов? Какие триплеты соответствуют началу и концу кодирующей белок последовательности?
6. Что такое открытая рамка считывания? Как программно осуществляется ее поиск?

Ответы на эти вопросы дают возможность дальнейшего успешного изучения прикладных

вопросов медицинской биоинформатики, например обнаружения вариантов в генетических последовательностях и статистического анализа их ассоциации с заболеваниями.

Хотя использование единой вычислительной среды на основе языка программирования R может быть сопряжено с серьезной проблемой в его освоении (не все врачи или студенты имеют соответствующие склонности), однако польза такого подхода достаточно большая для решения как текущих, так и будущих медицинских проблем:

- развитие у врачей алгоритмического подхода в решении клинических задач, осознании последовательности инструкций, предназначенных для решения задач путем обработки входных данных и получения результата за конечное время;
- использование объектов биологических последовательностей в практически неизменном виде на всем пространстве вычислений, их однократный ввод в программную среду;
- понимание основ вычислений и возможность сразу получить реакцию системы на различные манипуляции с геномными данными и методами их обработки;
- возможность создания комбинаций программного кода решения задач из различных удачных блоков кода, которые были предложены и проверены другими пользователями и разработчиками;
- создание своей версии решения и его проверка на языке, максимально близком предметным алгоритмам.

Расширение доступа клиницистов к современным вычислительным возможностям закла-

дывает базу для создания новых информативных диагностических технологий и практической реализации парадигмы персонализированной медицины. Все это выводит врачей и медработников на новый уровень исследовательской и доказательной базы, которые можно эффективно использовать для лечения пациентов.

Приложение. Краткое описание функций, используемых в описанных выше задачах

В фигурных скобках указан пакет, в котором функция представлена.

`readDNAStringSet()` {Biostrings} — функции для чтения/записи объекта `XStringSet` из/в файл.

`names()` {base} — функции для получения или назначения имен объектам.

`sub()` {base} — встроенная функция языка программирования R, которая заменяет первое вхождение подстроки в строке другой подстрокой.

`subseq(x, start=NA, end=NA, width=NA)` {XVector} {Biostrings} — извлекает подпоследовательность (подмножество) из последовательности `x`, заданную вручную координатами начала (`start`), конца (`end`) и шириной (`width`).

`GRanges()` {GenomicRanges} — создает объект `GRanges`.

`seqnames()` {GenomicRanges} — выводит вектор символов или фактор, содержащий имена последовательностей.

`ranges()` {GenomicRanges} — выводит диапазоны (`ranges`) из объекта `IRanges`.

`strand()` {GenomicRanges} — выводит информацию о цепи из объекта.

`start()`, `end()`, `width()` {IRanges} — функции выводят начало, конец, и ширину заданного диапазона (геномного признака).

`getSeq()` {Biostrings} — универсальная функция для извлечения подпоследовательности по координатам из объекта `Granges` из «большой» последовательности, например `BSgenome` или другой.

`import()` {rtracklayer} — функция поддерживает импорт формата GFF, который существует в трех версиях и нескольких разновидностях.

`unique()` {base} — функция возврата вектора, фрейма данных или массива без повторяющихся элементов/строк.

`subset()` {base} — функция выбирает определенные элементы структуры данных, такие как векторы, матрицы, фреймы данных или списки, на основе заданных условий.

`vmatchPattern()` {Biostrings} — функция находит совпадения заданного шаблона с произвольным количеством несовпадений в наборе эталонных последовательностей или с возможными вставками/делециями (`indel`).

`reverseComplement()` {Biostrings} — функция вычисляет обратное комплементарное соответствие последовательности ДНК или РНК.

`findORFs()` {ORFik} — функция находит все открытые рамки считывания во входных последовательностях. Возвращаемые координаты являются относительными к входным последовательностям.

`split()` {base} — функция разделяет вектор, фрейм данных или список на группы на основе заданного фактора или списка факторов.

`startCodon()` {ORFik} — функция поиска возможных старт-кодонов.

`stopCodon()` {ORFik} — функция поиска возможных стоп-кодонов.

ЛИТЕРАТУРА/REFERENCES

1. Jain KK. Textbook of Personalized Medicine. New York: Springer; 2009. 430 p. <https://doi.org/10.1007/978-1-4419-0769-1>
2. Golubnitschaja O, Baban B, Boniolo G, Wang W, Bubnov R, Kapalla M, Krapfenbauer K, Mozaffari MS, Costigliola V. Medicine in the early twenty-first century: paradigm and anticipation — EPMA position paper 2016. *EPMA J.* 2016;7(1):23. Published 2016 Oct 25. <https://doi.org/10.1186/s13167-016-0072-4>
3. Boffetta P, Collatuzzo G. Application of P4 (Predictive, Preventive, Personalized, Participatory) Approach to Occupational Medicine. *Med Lav.* 2022;113(1):e2022009. Published 2022 Feb 22. <https://doi.org/10.23749/mdl.v113i1.12622>
4. Приказ Министерства здравоохранения РФ от 24 апреля 2018 г. № 186 «Об утверждении Концепции предиктивной, превентивной и персонализированной медицины». <https://docs.cntd.ru/document/557437659>
Order of the Ministry of Health of the Russian Federation dated April 24, 2018 No. 186 „On Approval of the Concept of Predictive, Preventive and Personalized Medicine“. (In Russ.) <https://docs.cntd.ru/document/557437659>
5. Приказ Министерства труда и социальной защиты Российской Федерации от 21 марта 2017 г. № 293н «Об утверждении профессионального стандарта „Врач-лечебник (врач-терапевт участковый)“». <https://docs.cntd.ru/document/420395834>
Order of the Ministry of Labour and Social Protection of the Russian Federation dated March 21, 2017 No. 293n „On Approval of Professional Standard 'General Practitioner Doctor (Primary Care Physician)'“. (In Russ.) Accessed April 25, 2025. Available at: <https://docs.cntd.ru/document/557437659>
6. Lengauer T. Stellenwert der Bioinformatik für die personalisierte Medizin. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.* 2013;56(11):1489-1494. <https://doi.org/10.1007/s00103-013-1819-x> (German).

7. Шортлифф Э. Х., Чимино Д. Д. Биомедицинская информатика. Компьютерные приложения в здравоохранении и биомедицине: руководство. Под ред. М. Ф. Чанг, пер. Г. Э. Улумбекова, А. Н. Наркевич, Р. С. Козлов. М.: ГЭОТАР-Медиа, 2024. 744 с. Золотая серия. <https://doi.org/10.33029/9704-8204-9-BIN-2024-1-744>
Shortliff E. H., Cimino J. D. Biomedical informatics. Computer applications in health care and biomedicine: a handbook. Ed. M. F. Chang. G. E. Ulumbekova, A. N. Narkevich, R. S. Kozlov trans. Moscow: GEOTAR-Media, 2024. 744 p. Moscow: GEOTAR-Media; 2024. 744 p. (In Russ.) <https://doi.org/10.33029/9704-8204-9-BIN-2024-1-744>
8. Корнеев А. А., Янов Ю. К., Вяземская Е. Э., Медведева А. Ю. Вопросы интеграции медицинских биоинформатических технологий в оториноларингологию: проблемы и программные решения. *Российская оториноларингология*. 2024;23(6):8-19. <https://doi.org/10.18692/1810-4800-2024-6-8-19>
Korneenkov A. A., Yanov Yu. K., Vyazemskaya E. E., Medvedeva A. Yu. Issues of integrating medical bioinformatics technologies into otorhinolaryngology: challenges and software solutions. *Russian Otorhinolaryngology*. 2024;23(6):8-19. (In Russ.) <https://doi.org/10.18692/1810-4800-2024-6-8-19>
9. Chan IS., Ginsburg GS. Personalized medicine: progress and promise. *Annu Rev Genomics Hum Genet*. 2011;12:217-44. <https://doi.org/10.1146/annurev-genom-082410-101446>. PMID: 21721939.
10. Yue R, Dutta A. Computational systems biology in disease modeling and control, review and perspectives. *NPJ Syst Biol Appl*. 2022;8(1):37. Published 2022 Oct 3. <https://doi.org/10.1038/s41540-022-00247-4>
11. Somvanshi PR, Venkatesh KV. A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics. *Syst Synth Biol*. 2014;8(1):99-116. <https://doi.org/10.1007/s11693-013-9125-3>
12. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*. 2011;7(3):e1002021. <https://doi.org/10.1371/journal.pcbi.1002021>
13. Wikipedia contributors. Genome browser. Wikipedia, The Free Encyclopedia. Published April 25, 2025. Accessed April 25, 2025. https://en.wikipedia.org/wiki/Genome_browser
14. Python vs. R for Bioinformatics: Choosing the Right Tool. Bulut Hamali's Blog, Bioinformatics Universe. Accessed 23.04.2025. Available at: <https://bioinformaticsuniverse.com/r-vs-python/>
15. Корнеев А. А., Рязанцев С. В., Левин С. В., Храмов А. В., Вяземская Е. Э., Скирипичников И. Н., Левина Е. А., Павлова С. С. Пространственно-статистический анализ данных о нарушениях слуха у жителей Челябинской области. *Российская оториноларингология*. 2021;20(3):39-50. <https://doi.org/10.18692/1810-4800-2021-3-39-50>
Korneenkov A. A., Ryazantsev S. V., Levin S. V., Khramov A. V., Vyazemskaya E. E., Skirpichnikov I. N., Levina E. A., Pavlova S. S. Spatial and statistical analysis of hearing impairment data of Chelyabinsk region residents. *Rossiiskaya otorinolaringologiya*. 2021;20(3):39-50. (In Russ.) <https://doi.org/10.18692/1810-4800-2021-3-39-50>
16. Павлова С. С., Корнеев А. А., Дворянчиков В. В., Рязанцев С. В., Рязанцева Е. С., Донская О. С. Оценка потерь здоровья населения в результате назальной обструкции на основе концепции глобального бремени болезни: общие подходы и направления исследований. *Медицинский совет*. 2021;(12):138-145. <https://doi.org/10.21518/2079-701X-2021-12-138-145>
Pavlova S. S., Korneenkov A. A., Dvorianchikov V. V., Ryazantsev S. V., Ryazantseva E. S., Donskaya O. S. Assessment of population health losses due to nasal obstruction based on the concept of the global burden of disease: general approaches and research directions. *Meditsinskiy sovet = Medical Council*. 2021;(12):138-145. (In Russ.) <https://doi.org/10.21518/2079-701X-2021-12-138-145>
17. Корнеев А. А., Янов Ю. К., Рязанцев С. В., Вяземская Е. Э., Асташенко С. В., Рязанцева Е. С. Метаанализ клинических исследований в оториноларингологии. *Вестник оториноларингологии*. 2020;85(2):26-30. <https://doi.org/10.17116/otorino20208502126>
Korneenkov A. A., Yanov Yu. K., Ryazantsev S. V., Vyazemskaya E. E., Astashchenko S. V., Ryazantseva E. S. A meta-analysis of clinical studies in otorhinolaryngology. *Vestnik otorinolaringologii*. 2020;85(2):26-30. (In Russ.) <https://doi.org/10.17116/otorino20208502126>
18. Корнеев А. А., Левина Е. А., Вяземская Е. Э., Левин С. В., Скирипичников И. Н. Пространственный кластерный анализ в моделировании доступности медицинской помощи пожилым пациентам с нарушениями слуха. *Российская оториноларингология*. 2021;20(6):8-19. <https://doi.org/10.18692/1810-4800-2021-6-8-19>
Korneenkov A. A., Levina E. A., Vyazemskaya E. E., Levin S. V., Skirpichnikov I. N. Spatial cluster modeling of access of elderly patients with hearing loss to medical services. *Rossiiskaya otorinolaringologiya*. 2021;20(6):8-19. (In Russ.) <https://doi.org/10.18692/1810-4800-2021-6-8-19>
19. Корнеев А. А., Рязанцев С. В., Вяземская Е. Э., Будковская М. А. Меры информативности диагностических медицинских технологий в оториноларингологии: вычисление и интерпретация. *Российская оториноларингология*. 2020;19(1):46-55. (In Russ.) <https://doi.org/10.18692/1810-4800-2020-1-46-55>
Korneenkov A. A., Ryazantsev S. V., Vyazemskaya E. E., Budkovskaya M. A. The measures of informativeness of diagnostic medical technologies in otorhinolaryngology: calculation and interpretation. *Rossiiskaya otorinolaringologiya*. 2020;19(1):46-55. (In Russ.) <https://doi.org/10.18692/1810-4800-2020-1-46-55>
20. Russell PJ. iGenetics: A Molecular Approach. 3rd ed. London, UK: Pearson Education; 2010. 832 p.
21. Ensembl. Homo sapiens gene sequence (ENSG00000184895). Ensembl. Published 2025. Accessed April 25, 2025. Available at: https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000184895;r=Y:2786855-2787682;t=ENST00000383070
22. Glossary of common genetic and bioinformatic terms. IDT. 2022. Accessed 23.04.2025. Available at: https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/miscellaneous/idt-glossary-of-common-genetic-and-bioinformatic-terms.pdf?sfvrsn=7c01407_2
23. Bioinformatics.org. IUPAC nucleic acid codes, amino acid codes, and genetic codes. Bioinformatics.org. Published 2025. Accessed 23.04.2025. Available at: <https://www.bioinformatics.org/sms/iupac.html>

24. Cornell B. Gene Identification. 2016. [online] Accessed 23.04.2025. Available at: <https://old-ib.bioninja.com.au/options/untitled/b2-biotechnology-in-agricul/gene-identification.html>
25. Sequence Ontology. GFF3 specification. Sequence Ontology. Published 2025. Accessed April 25, 2025. Available at: <http://www.sequenceontology.org/gff3.shtml>

Вклад авторов

Концепция статьи — А. А. Корнеев
 Концепция и дизайн исследования — А. А. Корнеев
 Написание текста — А. А. Корнеев, Е. Э. Вяземская
 Сбор и обработка материала — А. Ю. Медведева, Е. Э. Вяземская
 Обзор литературы — А. А. Корнеев, Е. Э. Вяземская
 Перевод на английский язык — А. Ю. Медведева, Е. Э. Вяземская
 Анализ материала — А. А. Корнеев, А. Ю. Медведева, Е. Э. Вяземская
 Статистическая обработка — А. А. Корнеев, А. Ю. Медведева, Е. Э. Вяземская
 Редактирование — Ю. К. Янов, А. А. Корнеев, В. В. Дворянчиков
 Утверждение окончательного варианта статьи — Ю. К. Янов, А. А. Корнеев, В. В. Дворянчиков

Contribution of authors

Concept of the article — A. A. Korneenkov
 Study concept and design — A. A. Korneenkov
 Text development — A. A. Korneenkov, E. E. Vyazemskaya
 Collection and processing of material — A. Yu. Medvedeva, E. E. Vyazemskaya
 Literature review — A. A. Korneenkov, E. E. Vyazemskaya
 Translation into English — A. Yu. Medvedeva, E. E. Vyazemskaya
 Material analysis — A. A. Korneenkov, A. Yu. Medvedeva, E. E. Vyazemskaya
 Statistical processing — A. A. Korneenkov, A. Yu. Medvedeva, E. E. Vyazemskaya
 Editing — Yu. K. Yanov, A. A. Korneenkov, V. V. Dvoryanchikov
 Approval of the final version of the article — Yu. K. Yanov, A. A. Korneenkov, V. V. Dvoryanchikov

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: the authors declare no conflict of interest.

Информация об авторах

Корнеев Алексей Александрович — доктор медицинских наук, профессор, заведующий научно-исследовательской лабораторией клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи (190013, Российская Федерация, Санкт-Петербург, Бронницкая ул., д. 9); korneyenkov@gmail.com, <https://orcid.org/0000-0001-5870-8042>

Янов Юрий Константинович — доктор медицинских наук, профессор, академик РАН, профессор кафедры оториноларингологии, Военно-медицинская академия им. С. М. Кирова, (194044, Российская Федерация, Санкт-Петербург, ул. акад. Лебедева, д. 6); профессор, кафедра оториноларингологии, Северо-Западный государственный медицинский университет имени И. И. Мечникова (191015, Российская Федерация, Санкт-Петербург, Кирочная ул., д. 41); 9153864@mail.ru, <https://orcid.org/0000-0001-9195-128X>

Дворянчиков Владимир Владимирович — доктор медицинских наук, профессор, заслуженный врач России, директор, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи (190013, Российская Федерация, Санкт-Петербург, Бронницкая ул., д. 9); v.v.dvoryanchikov@mail.ru, <https://orcid.org/0000-0002-0925-7596>

Вяземская Елена Эмильевна — инженер научно-исследовательской лаборатории клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи (190013, Российская Федерация, Санкт-Петербург, Бронницкая ул., д. 9); vyazemskaya.elena@gmail.com, <https://orcid.org/0000-0002-4141-2226>

Медведева Анна Юрьевна — инженер научно-исследовательской лаборатории клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи (190013, Российская Федерация, Санкт-Петербург, Бронницкая ул., д. 9); a.medvedeva@niilor.ru, <https://orcid.org/0009-0002-6921-5299>

Information about authors

Alexei A. Korneenkov — Doctor of Sciences (Med.), Professor, Head of the Research Laboratory of Clinical Informatics and Biostatistics, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech (9, Bronnitskaya str., Saint Petersburg, Russian Federation, 190013); korneyenkov@gmail.com, <https://orcid.org/0000-0001-5870-8042>

Yurii K. Yanov — Doctor of Sciences (Med.), Professor, Member of the Russian Academy of Sciences, Kirov Military Medical Academy (6, Akademik Lebedeva str., Saint Petersburg, Russian Federation, 194044); Mechnikov North-Western State Medical University (41, Kirochnaya str., Saint Petersburg, Russian Federation, 190013); 9153764@mail.ru, <https://orcid.org/0000-0001-9195-128X>

Vladimir V. Dvoryanchikov — Doctor of Sciences (Med.), Professor, Director, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech (9, Bronnitskaya str., Saint Petersburg, Russian Federation, 190013); v.v.dvoryanchikov@mail.ru, <https://orcid.org/0000-0002-0925-7596>

Elena E. Vyazemskaya — Engineer of the Research Laboratory of Clinical Informatics and Biostatistics, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech (9, Bronnitskaya str., Saint Petersburg, Russian Federation, 190013); vyazemskaya.elena@gmail.com, <https://orcid.org/0000-0002-4141-2226>

Anna Yu. Medvedeva — Engineer of the Research Laboratory of Clinical Informatics and Biostatistics, Saint-Petersburg Research Institute of Ear, Throat, Nose and Speech (9, Bronnitskaya str., Saint Petersburg, Russian Federation, 190013); a.medvedeva@niilor.ru, <https://orcid.org/0009-0002-6921-5299>

Поступила / Received 17.03.2025

Поступила после рецензирования / Revised 26.05.2025

Принята в печать / Accepted 07.07.2025