Лекция 15. Геномные базы данных. NCBI, Ensembl genomes, International genome.

1 Геномные базы данных.

2 NCBI.

3 Ensembl genomes.

4 International genome.

1. **Геномные базы данных.** Геномные базы данных играют центральную роль в современной биологии и медицине, обеспечивая исследователей доступом к огромным объемам данных о структуре, функции и эволюции геномов различных организмов. Среди ведущих ресурсов в этой области — базы данных NCBI, Ensembl Genomes и International Genome Project, которые предоставляют мощные инструменты для анализа и аннотации геномной информации.

NCBI (National Center for Biotechnology Information) — это одна из наиболее обширных и широко используемых баз данных в области биоинформатики и геномики. Основанная в США, NCBI поддерживает многочисленные базы данных, такие как GenBank, которая является публичной базой последовательностей ДНК, и dbSNP, содержащую информацию о однонуклеотидных полиморфизмах. Особое значение имеет ресурс RefSeq, который предоставляет стандартные аннотированные последовательности генов и белков для различных организмов, включая человека. Помимо геномных последовательностей, NCBI также поддерживает PubMed, платформу для поиска научных публикаций, что делает NCBI важным комплексным инструментом для геномных и биомедицинских исследований. Исследователи могут использовать платформу BLAST (Basic Local Alignment Search Tool), доступную в NCBI, для выравнивания последовательностей и поиска гомологичных последовательностей в разных организмах. Это дает возможность изучать эволюционные связи между генами и находить гомологичные гены.

Ensembl Genomes — это европейский проект, предоставляющий доступ к аннотированным геномам более чем 80 тысяч видов. Основной особенностью Ensembl является интеграция геномных данных с обширными аннотациями, включая информацию о структурных вариантах, экспрессии генов и взаимодействиях белков. Ensembl поддерживает автоматическую аннотацию геномов с использованием высококачественных данных, полученных из экспериментов с РНК и белками, и интегрирует данные из других крупных проектов, таких как ENCODE (Encyclopedia of DNA Elements). В Ensembl особое внимание уделяется человекообразным организмам, что делает её ключевым ресурсом для исследований в области медицины и эволюционной биологии. Пользователи могут также просматривать трехмерные модели взаимодействия генов и белков, что позволяет понять механизмы регуляции генов. Одной из важных особенностей является возможность интеграции данных Ensembl с пользовательскими экспериментальными данными, что расширяет возможности для глубокого анализа геномной информации.

International Genome Project представляет собой глобальные усилия по созданию и поддержанию ресурсов геномной информации. Среди таких проектов — The Human Genome Project, завершенный в 2003 году, который стал первым проектом полного секвенирования генома человека. Другие масштабные проекты включают 1000 Genomes Project и International Cancer Genome Consortium (ICGC), направленные на изучение генетического разнообразия и геномных изменений, связанных с заболеваниями. Эти проекты способствуют созданию эталонных геномов и выявлению вариаций, специфичных для различных популяций, что важно для персонализированной медицины. Современные проекты, такие как Earth BioGenome Project, ставят своей целью секвенирование геномов всех известных эукариот на Земле, что открывает перспективы для биологического разнообразия и сохранения экосистем.

Эти три платформы являются центральными инструментами для исследования геномов, поскольку они не только хранят и аннотируют данные, но и предоставляют мощные аналитические инструменты, которые помогают ученым исследовать функции генов, их регуляцию и взаимодействия.

 В настоящее время в сети Интернет существуют сотни баз данных, которые доступны для поиска данных по молекулярной биологии и другим смежным дисциплинам. Каждая из них имеет свой формат хранения данных, различную степень избыточности, взаимосвязи с родственными или аналогичными базами данных. Каждая база данных имеет также свои средства доступа к информации – различные поисковые программы, программные средства визуализации, пополнения базы.

Последовательность генома обеспечивает естественную основу для организации биологических данных. За то короткое время, в течение которого стали доступны последовательности генома, базы данных геномов оказались бесценными ресурсами для исследователей. В случае человека диапазон существующих биологических данных и типов исследователей даже шире, чем для других организмов, простираясь от клинической генетики до молекулярной биологии. Доступность проекта последовательности генома человека позволяет впервые систематически объединить эти огромные объемы данных, начиная от записей о болезнях нашего вида и заканчивая последовательностями родственных организмов.

 Крупнейшие хранилища первичных структур ДНК и аминокислотных последовательностей (такие, как EMBL, GenBank, DDBJ, SWISS-PROT, Ensembl и др.) пополняются аннотированными последовательностями непосредственно исследователями, расшифровавшими их, с помощью автоматизированной системы пополнения баз данных по сети Интернет.

 Конечно, впоследствии эти данные проверяются персоналом администраций баз данных и существенно пополняются. Вторым основным источником информации во всех базах является специальная научная литература. Многие базы данных, работающие над коллекционированием однородной информации, координируют свои усилия, осуществляя международное разделение труда, это можно проиллюстрировать примером сотрудничества трех всемирных коллекций последовательностей нуклеотидов EMBL (Европа), GenBank (США), DDBJ (Япония).

 Наряду с общими базами данных в последнее время появилось много специализированных информационных ресурсов. Многие из них хранят данные, полученные с помощью компьютерных методов, результаты теоретических предсказаний. Большую роль в биоинформатике играют хранилища последовательностей ДНК и кДНК, специализированные базы данных по отдельным регуляторным мотивам нуклеотидных последовательностей, базы данных по экспрессии генов, библиотеки геномов, карт, последовательностей РНК, белков, белковых мотивов, по продукции белков. Есть базы данных по протеомике, структурам белков, мутациям, метаболическим путям и регуляции, по трансгеннным организмам, анатомии, биохимии, а также по научной литературе, по существующему в этих областях исследований программному обеспечению.

 Будет дано общее представление о существующих базах данных по геному человека и более детально рассматриваются некоторые БД, список которых будет постоянно пополняться.

 Краткий обзор основных баз данных по геному человека:

 OMIM – On-line Mendelian Inheritance in Man. Крупнейшая база данных по человеческим генам и генетическим заболеваниям, создал базу доктор МакКасик (Victor A. McKusick) с коллегами в центре медицинской генетики (Johns Hopkins University, Baltimore, USA), NCBI поддерживает наполнение и обновление базы. Содержит общие обзоры по заболеваниям и конкретным генам, а также ссылки на базы данных ENTREZ. Адрес: <http://www.ncbi.nlm.nih.gov/omim/>

 GenBank – база данных генетических последовательностей, поддерживается NIH (Национальный Институт Здоровья США), аннотированная база известных последовательностей ДНК, РНК и белков, с литературными ссылками на первоисточники и информацией биологического характера. Обновляется каждые два месяца. Является частью International Nucleotide Sequence Database Collaboration, которая объединяет три крупнейшие коллекции нуклеотидных последовательностей: DDBJ (NIG), EMBL (EBI) и GenBank (NCBI). Три организации осуществляют разделение труда и ежедневно обмениваются новой информацией. Большинство журналов требуют предварительной посылки последовательностей в любую из этих трех баз данных до опубликования статьей о них. В статьях, посвященных очередной порции секвенированных последовательностей, должен упоминаться лишь номер последовательности в базе данных. NCBI постоянно совершенствует и создает новые средства для помещения новых последовательностей в базу, средства эффективного поиска в базе. Крупнейшая интегрированная поисковая система ENTREZ для нуклеотидных и аминокислотных последовательностей, библиографии (PubMed), полных геномов (Genomes), а также трехмерных структур белков (MMDB) создана и поддерживается NCBI. При этом поиск ДНК и белков не ограничивается только ресурсами GenBank, но и другими доступными по сети хранилищами информации. Адрес: <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>

 EMBL – the EMBL Nucleotide Sequence Database. База данных нуклеотидных последовательностей Европейской Молекулярно-Биологической Лаборатории пополняется большей частью непосредственно авторами, определившими первичную структуру фрагмента ДНК или РНК и, кроме последовательности нуклеотидов, содержит разнообразную информацию о каждом фрагменте, включая литературные ссылки, перекрестные ссылки на документы других баз данных, таблицы особенностей и др. Существует с 1982 года. База данных – продукт сотрудничества EMBL (ФРГ), GenBank (США) и DDJP (Япония), каждая из этих трех групп собирает свою порцию информации из всех возможных мировых источников, ежедневно обмениваясь новыми и обновленными документами друг с другом. Удобна своей географической близостью для доступа на территории Европы. В России есть сайт, на котором хранится ежедневно обновляемая копия базы (http://www.genebee.msu.su/, отв. Скулачев В.П.). Адрес: <http://www.ebi.ac.uk/embl/>

 HGMD – Human Gene Mutation Database. Содержит информацию обо всех опубликованных повреждениях генов, приводящих к наследственным заболеваниям у человека. Документы базы аннотируют все гены, находящиеся в ядре. Гены митохондриального генома и соматические мутации исключены. Мутации, выявленные на уровне белкового сиквенса, не входят в базу чтобы избежать ошибок из-за отсутствия анализа на уровне ДНК. Молчащие мутации, не приводящие к изменению аминокислотной последовательности тоже исключены. С марта 1999 года включены данные о полиморфизме, связанном с болезнями. Данные берутся из тех же самых журналов, что и данные о мутациях (>250). Сопровождается Институтом медицинской генетики (University of Wales, Cardiff, UK). Адрес: <http://www.hgmd.cf.ac.uk/ac/index.php>

 KEGG – Kyoto Encyclopedia of Genes and Genomes. Попытка компьютеризировать все современное знание в молекулярной и клеточной биологии в терминах информационных путей. Это база знаний по систематическому анализу функций генов. Создается институтом химических исследований (Kyoto University, Japan) в рамках японской программы по геному человека. Содержит 6 баз данных – метаболических путей (PATHWAY), генов (GENES) и лигандов (LIGAND), экспериментальных данных по экспрессии генов (EXPRESSION и BRITE), по белкам (SSDB) и обширные возможности для работы со всеми крупными мировыми информационными ресурсами. Базы данных KEGG представляют данные в виде графических диаграмм, включающих большинство метаболических путей и некоторые из наиболее известных регуляторных путей. Кроме того, информация о путях представлена в виде таблиц ортологов, которые содержат как гены-ортологи, так и паралоги из различных организмов. Обновляются базы ежедневно. Адрес: <http://www.genome.ad.jp/kegg/>

 UniGene. База данных, которая содержит кластеры похожих последовательностей. Каждый кластер представляет один ген и содержит попутную информацию, например, название ткани, где этот ген экспрессирован. Кроме хорошо известных генов в базу данных включены сотни тысяч новых концов экспрессирующихся последовательностей (EST – expressed sequence tags). Служит для поиска генов в новых последовательностях, а также для определения реагентов при секвенировании генов и их экспрессии. Кластеризация осуществляется автоматически. Адрес: <http://www.ncbi.nlm.nih.gov/unigene>

 PROSITE – PROtein SITEs and patterns dictionary. База данных различных паттернов функциональных и регуляторных участков. С помощью этой коллекции можно определить, принадлежит ли, и к какому именно, семейству белков новая последовательность пользователя, или какой важный домен она содержит. Версия 17.21 этой базы, датированная сентябрем 2002 года содержит 11148 единиц хранения, которые описывают 1568 различных паттернов, правил и матриц. Адрес: <http://www.expasy.ch/prosite/>

 SWISS-PROT|UniProt – the protein sequence data bank. База данных содержит аннотированные аминокислотные последовательности, транслированные с нуклеотидных последовательностей EMBL; адаптированные последовательности из PIR; а также последовательности, опубликованные в литературе и присланные непосредственно самими авторами. Содержит высококачественные неизбыточные аннотации, перекрестные ссылки на другие родственные базы данных (EMBL, Prosite, PDB). Каждая аннотация содержит описание функции белка, его доменной структуры, особенностей пост-трансляционной модификации, различные варианты. Имеется неаннотированное приложение (TrEMBL). Поодерживается Женевским университетом (Department of Medical Biochemistry of the University of Geneva) и EMBL. (EBI). Для академических пользователей – бесплатна. Сайт не так давно обновился и немного поменялась структура поисковых запросов. На мой взгляд стало гораздо красивей и удобней. Адрес: <http://www.uniprot.org/>

 trEMBL – EMBL protein-coding DNA sequence features translated into peptide sequences. База данных, созданная автоматически, представляет собой приложение к SWISS-PROT. Содержит аминокислотные последовательности, транслированные программно с нуклеотидных кодирующих участков, взятых из базы данных EMBL. Адрес: <http://www.uniprot.org/>

 ENSEMBL Ensembl – совместный проект EMBL – EBI и Sanger Centre с целью создания программной системы для автоматической аннотации эукариотических геномов. Осуществляет (бесплатно) следующие возможности: поиск ДНК из человеческого генома, обзор хромосом, поиск белков и белковых семейств. Проект Ensembl стремится обеспечивать соответствие следующим критериям: точный, автоматический анализ данных генома; анализ и аннотациии основаны на текущих, своевременно обновляемых данных; доступность полученных данных для всех через сеть Интернет; предоставление данным другим лабораториям по биоинформатике. Основной акцент в базе данных Ensembl сделан на позвоночных геномах, но другие группы адаптировали систему для использования с растительными и грибковыми геномами. Адрес: <http://www.ensembl.org/>

 Базы данных можно отнести к следующим типам*: 1) Архивные.* К архивным относятся, например, базы данных GeneBank, EMBL, PDB. Любой исследователь может поместить туда свою информацию. За содержание каждой записи в таких базах отвечает сам исследователь. GenBank – база данных генетических последовательностей, основанная в 1982 году. Это аннотированная коллекция всех общедоступных последовательностей ДНК, РНК и белков, снабженных литературными ссылками, и другой биологической информацией. Эта база является частью объединения International Nucleotide Sequence Database Collaboration, которое объединяет три крупнейшие коллекции нуклеотидных последовательностей: DDBJ (DNA Data Bank of Japan), EMBL (European Molecular Biology Laboratory) и GenBank (National Center for Biotechnology Information). Эти три организации ежедневно обмениваются новой информацией. Большинство журналов требуют предварительной посылки новых секвенированных последовательностей в любую из этих трех баз данных до опубликования статьей о них. В статьях, посвященных очередной порции последовательностей, должен упоминаться лишь номер последовательности в базе данных GenBank.

Адрес DDBJ: <http://www.ddbj.nig.ac.jp/>

Адрес GenBank: <http://www.ncbi.nlm.nih.gov/Genbank/>

EMBL (European Molecular Biology Laboratory) – эта база данных содержит разнообразную информацию о каждом фрагменте последовательностей, включая литературные ссылки, перекрестные ссылки на документы других баз данных и др. Адрес EMBL: <http://www.ebi.ac.uk/embl/>

 Еще одна архивная база данных – PDB (Brookhaven Protein DataBank) – содержит данные о коллекции экспериментально определенных трехмерных структур биологических макромолекул (белков и нуклеиновых кислот). С 2002 года в основном депозитарии PDB хранятся структуры, экспериментально определенные с помощью рентгеноструктурного, ядерно–магнитнорезонансного и др. методов. Теоретические структуры выделены в отдельную подбазу PDB. Адрес: <http://www.rcsb.org/pdb/>

 *2) Курируемые базы данных.* За содержание записей в таких базах данных отвечают кураторы. Информацию для курируемых баз данных отбирают эксперты из архивных баз. К курируемым базам относятся, например, SwissProt. Эта база данных белковых последовательностей существует с 1986 года и поддерживается двумя институтами: Swiss Institute of Bioinformatics (SIB) и European Bioinformatics Institute (EBI). Адрес: <http://www.ebi.ac.uk/swissprot/>

 3) *Автоматические базы данных.* В таких базах данных записи генерируются (моделируются) компьютерными программами. К ним относится, например TrEMBL (Translated EMBL) – автоматическая база предсказаний последовательностей белков. Это формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL. В 2002 году в результате объединения SwissProt, TrEMBL и PIR был создан банк данных UniProt (Universal Protein Resourse). Это основное хранилище белковых последовательностей и их функций. UniProt состоит из трех частей:

UniProt Knowlegebase – является центральной базой данных и обеспечивает доступ к обширной курируемой информации по белкам, включая их функцию, классификацию и перекрестные информационные ссылки;

UniProt Archive – UniParc. Отражает хронологию данных определения о всех белковых последовательностях;

UniProt Reference – UniRef. Содержит базы данных, которые объединяют последовательности в кластеры для ускорения поиска.

Адрес UniProt: <http://www.ebi.uniprot.org/index.shtml>

 4) *Производные базы данных.* Они получаются в результате компьютерной обработки данных из архивных и курируемых баз данных. Это, например, SCOP, PFAM, GO и др.

 SCOP (Structural Classification Of Proteins) – база данных по структурной классификации белков. Адрес: <http://scop.protres.ru/>

 PFAM (Protein families database of alignments and HMMs) – это большая коллекция семейств белков и доменов, построенных на основании экспертной оценки множественных выравниваний. В банке существуют две основные части: PFAMA, содержащая подробно аннотированные белковые семейства, и PFAMB, содержащая различные множественные выравнивания. Адрес: <http://www.sanger.ac.uk/Pfam/>

 GO (Gene Ontology consortium database). Целью создателей базы было установление контроля за единообразием в описаниях функций, биологических процессов и клеточных компонентов, относящихся к продуктам генов. Унификация описаний в различных базах данных облегчает поиск в них нужного гена. GO – независимая база данных: другие базы данных сотрудничают с ней, помещая ссылки на унифицированные термины GO, либо поддерживают поиск с использованием терминов базы GO, а также стимулируют ее дополнение и уточнение. Адрес: <http://www.geneontology.org/>

 5*) Интегрированные базы данных.* Они объединяют информацию из разных баз. Например, введя имя гена, можно найти всю, связанную с ним информацию. К таким базам относится ENTREZ (Molecular Biology DataBase and Retrieval System). Эта интегрированная база данных содержит нуклеотидные и аминокислотные последовательности, которые собираются из крупнейших специализированных хранилищ – баз данных. Основой является GenBank, кроме того, информация пополняется из dbEST, dbSTS, SwissProt, PIR, PDB, PRF, GSDB.

 Данные из перечисленных ресурсов поступают в интегрированную базу данных после 1) присвоения уникального идентификатора последовательности, 2) перевода документов в единый стандарт хранения, 3) проверки данных, 4) проверки всех ссылок по базе данных MedLine, 5) проверки названий организмов по таксономической классификации GenBank Taxonomy.

**Броузеры генома.** Геном состоит из набора хромосом, а хромосома — это две цепочки, свёрнутые в спираль. Каждая из цепочек содержит последовательность нуклеотидов с четырьмя типами азотистых оснований — аденин (A), гуанин (G), цитозин (C) и тимин (T). По одной цепочке легко определить вторую, если помнить, что аденин соединяется в пару с тимином, а гуанин с цитозином. Некоторые участки ДНК называются генами, с них считывается РНК, по которой потом кодируются белки. Белки состоят из аминокислот 20 видов (плюс пара экзотических), каждая из которых кодируется по трём нуклеотидам. Браузер генома — это такая одномерная карта, которая отображает какую-нибудь нуклеотидную последовательность (скажем, хромосому или отдельный ген) с сопутствующей информацией. Информация обычно структурируется в блоки, называемые треками (tracks). К примеру, может быть трек с генами или с отдельными нуклеотидами. Отдельные сущности на треках часто называют фичами (features). Бывают браузеры геномов, рассчитанные на маленькие бактериальные геномы, но универсальному браузеру необходимо показывать и длинные хромосомы позвоночных целиком, и отдельные нуклеотиды. Самая длинная хромосома человека (первая) содержит около 250 миллионов пар оснований, то есть масштаб должен меняться примерно в миллион раз. Конечно, в разном масштабе информация отображается по-разному. В самом детальном масштабе можно увидеть отдельные нуклеотиды, как на прямой, так и на обратной спирали ДНК. http://ugene.unipro.ru/

Ugene. Разработчики: Центр информационных технологий «УниПро», Новосибирск, <http://unipro.ru/ru/about/overview.html>

1. **NCBI.** Национальный центр биотехнологической информации (NCBI) выпускает множество онлайн-ресурсов информации по биологии, включая базу данных последовательностей нуклеиновых кислот GenBank® и базу данных цитат и рефератов PubMed®, опубликованных в журналах по естественным наукам. NCBI обеспечивает поиск и извлечение большинства этих данных из 35 различных баз данных. Электронные утилиты служат в качестве программного интерфейса для большинства этих баз данных. Ресурсы, получившие значительные обновления за последний год, включают PubMed, PMC, Bookshelf, RefSeq, SRA, Virus, dbSNP, dbVar, ClinicalTrials.gov, MMDB, iCn3D и PubChem. Доступ к этим ресурсам можно получить через домашнюю страницу NCBI по адресу [https://www.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov/) .

Национальный центр биотехнологической информации (NCBI), центр в Национальной медицинской библиотеке в Национальных институтах здравоохранения, был создан в 1988 году для разработки информационных систем для молекулярной биологии ( [1](https://pmc.ncbi.nlm.nih.gov/articles/PMC8728269/#B1) ). В этой статье мы приводим краткий обзор коллекции баз данных NCBI, а затем сводку ресурсов, которые мы значительно обновили за последний год. Мы приводим более полное обсуждение ресурсов NCBI на домашних страницах отдельных баз данных, на странице NCBI Learn ( <https://www.ncbi.nlm.nih.gov/learn/> ) и в Справочнике NCBI ( <https://www.ncbi.nlm.nih.gov/books/NBK143764/> ).

### Базы данных NCBI. NCBI поддерживает разнообразный набор из 35 баз данных, которые в совокупности содержат 3,6 миллиарда записей (таблица [1](https://pmc.ncbi.nlm.nih.gov/articles/PMC8728269/#tbl1) и рисунок [1](https://pmc.ncbi.nlm.nih.gov/articles/PMC8728269/#F1) ), большинство из которых доступны через систему поиска Entrez ( [2](https://pmc.ncbi.nlm.nih.gov/articles/PMC8728269/#B2) ) по адресу <https://www.ncbi.nlm.nih.gov/search/> . Каждая база данных поддерживает текстовый поиск с использованием простых булевых запросов, загрузку данных в различных форматах и ​​связывание записей между базами данных на основе утвержденных отношений. Записи, извлеченные в Entrez, могут отображаться во многих форматах и ​​загружаться по отдельности или в пакетах. Доступен интерфейс прикладного программирования для функций Entrez (E-utilities), а подробная документация представлена ​​по адресу <https://eutils.ncbi.nlm.nih.gov/> .

### NCBI получает данные из трех источников: прямые заявки от исследователей, национальное и международное сотрудничество или соглашения с поставщиками данных и исследовательскими консорциумами, а также внутренние усилия по кураторству.

### Книжная полка NCBI предоставляет бесплатный онлайн-доступ к более чем 9600 книгам и документам по естественным наукам и здравоохранению от более чем 150 поставщиков контента. В прошлом году в сотрудничестве с участвующими агентствами здравоохранения США, международными организациями здравоохранения и другими спонсорами и издателями доказательной клинической или общественно-медицинской информации Bookshelf предоставила коллекцию руководств и обзоров COVID-19 ( <https://www.ncbi.nlm.nih.gov/books/about/covid/> ), которые соответствуют критериям отбора контента Bookshelf ( <https://www.ncbi.nlm.nih.gov/books/about/scientificeval/> ). Целевой поисковый запрос ( [https://www.ncbi.nlm.nih.gov/books?term=%22covid+19%22%5BResource+ Type%5D&cmd=DetailsSearch](https://www.ncbi.nlm.nih.gov/books?term=%22covid+19%22%5BResource+Type%5D&cmd=DetailsSearch) ) регулярно обновляется новыми и обновленными руководствами и обзорами.

**3. Ensembl.** Проект Ensembl активно решает эту проблему, предоставляя базу данных аннотаций генома человека (http://www.ensembl.org/). Она постоянно расширяется, чтобы включать все больше типов данных (вертикальная интеграция), а также для построения сравнительных представлений последовательностей геномов по мере того, как становятся доступны последовательности геномов позвоночных, таких как мышь, крыса и данио-рерио (горизонтальная интеграция). База данных создается на основе очень общей и тщательно спроектированной программной структуры, которая разрабатывается параллельно с интеграцией данных. Сделав все программное обеспечение свободно доступным и разработав систему полностью переносимой, Ensembl стремится предоставить биоинформационную структуру, которую легко применять к различным организмам и типам данных. Мы надеемся, что в духе проектов сообщества с открытым исходным кодом, таких как Linux, Ensembl получит широкое распространение и позволит исследователям и разработчикам баз данных больше времени сосредоточиться на инновациях.

Ensembl аннотирует известные гены и предсказывает новые гены с функциональной аннотацией из баз данных семейств белков InterPro и с дополнительной аннотацией по заболеванию OMIM, экспрессии SAGE и по семейству генов.



***Рисунок 1.*** *Скриншот Ensembl contigview, показывающий область человеческой хромосомы 11 вокруг последовательности генома*[*AP000869*](https://www.ncbi.nlm.nih.gov/nuccore/AP000869)*.*

4. International genome. Для поддержания и развития крупнейших полностью открытых ресурсов генома человека был создан Международный ресурс образцов генома (IGSR) (https://www.internationalgenome.org). Он создан на основе проекта 1000 геномов, который создал крупнейший общедоступный каталог вариаций генома человека, разработанный на основе образцов, охватывающих пять континентов. IGSR поддерживает доступ к ресурсам проекта 1000 геномов, обновляет ресурсы проекта 1000 геномов в референтной сборке человека GRCh38, добавляет новые данные, полученные на клеточных линиях проекта 1000 геномов, делится данными из образцов с аналогичным открытым согласием для увеличения количества образцов и популяций, представленных в ресурсах, и (v) оказывает поддержку пользователям этих ресурсов. Среди последних обновлений — выпуск вызовов вариаций из данных проекта 1000 Genomes Project, рассчитанных непосредственно на GRCh38, и добавление данных последовательностей с высоким покрытием для 2504 образцов в панели фазы три проекта 1000 Genomes Project. Портал данных, который облегчает веб-исследование ресурсов IGSR, был обновлен для включения образцов, которые не были частью проекта 1000 Genomes Project, и теперь представляет собой единое представление данных и образцов по почти 5000 образцам из нескольких исследований. Все данные полностью открыты и общедоступны.

Международный ресурс образцов генома (IGSR; http://www.internationalgenome.org) расширяет типы данных и разнообразие популяций ресурсов проекта 1000 геномов. IGSR представляет собой крупнейшую открытую коллекцию данных о человеческих вариациях и обеспечивает легкий доступ к этим ресурсам.

IGSR был создан в 2015 году для поддержания и расширения данных проекта 1000 геномов, которые широко использовались в качестве эталонного набора человеческих вариаций и исследователями, разрабатывающими методы анализа. IGSR сопоставил все последовательности 1000 геномов с новейшим эталоном человека (GRCh38) и выпустит обновленные вызовы вариантов, чтобы обеспечить максимальную полезность существующих данных. IGSR собирает новые структурные данные о вариациях образцов 1000 геномов из длинных прочтений и других технологий и соберет соответствующие функциональные данные в единый всеобъемлющий ресурс. IGSR расширяет охват новыми популяциями, секвенированными сотрудничающими группами.

Вопросы для самоконтроля:

1. Какую роль играют геномные базы данных, такие как NCBI и Ensembl, в ускорении геномных исследований и применении данных в клинической практике?
2. Какие различия в подходах к аннотации геномов существуют между NCBI и Ensembl, и как это влияет на точность и полноту представленных данных?
3. Какие уникальные функции и ресурсы предлагает NCBI для анализа данных по однонуклеотидным полиморфизмам (SNPs), и как это помогает в генетических исследованиях?
4. Каким образом International Genome Project и другие крупные геномные инициативы (например, 1000 Genomes Project) способствуют исследованию генетического разнообразия, и как это может применяться в персонализированной медицине?
5. Как платформы, такие как Ensembl, интегрируют данные о взаимодействии генов и белков, и каким образом эта информация помогает в исследовании молекулярных механизмов заболеваний?
6. Какие биоэтические вопросы и вызовы связаны с хранением и использованием данных в International Genome Project, и как эти проблемы решаются на глобальном уровне?
7. Какую роль играют геномные базы данных в исследованиях эволюции геномов, и какие инструменты и методы обеспечивают сравнительный анализ геномов у различных видов?