Министерство здравоохранения Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Российский национальный исследовательский медицинский
университет имени Н.И. Пирогова»
(ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России)

Кафедра биоинформатики медико-биологического факультета

Общедоступные ресурсы биоинформатики: биологические базы данных, геномный браузер UCSC

Учебно-методическое пособие

Москва
2023

Министерство здравоохранения Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Российский национальный исследовательский медицинский
университет имени Н.И. Пирогова»
(ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России)

Кафедра биоинформатики медико-биологического факультета

Общедоступные ресурсы биоинформатики: биологические базы данных, геномный браузер UCSC

Учебно-методическое пособие

Рекомендовано Центральным координационным методическим советом ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России

Москва
2023

УДК 575.112(075.8)
ББК 30.16я73
О28

Составители:
Б.В. Шилов, канд. мед. наук, доцент;
А.А. Лагунин, д-р биол. наук, проф. РАН

Рецензенты:
А.В. Чаплин, канд. мед. наук, доцент кафедры микробиологии и вирусологии Педиатрического факультета ФГАОУ ВО «Российский национальный исследовательский университет имени Н.И. Пирогова» Минздрава России
Б.Н. Соболев, канд. биол. наук, ведущий научный сотрудник лаборатории структурно-функционального конструирования лекарств ФБГНУ «Институт биомедицинской химии им. В.Н. Ореховича»

О28 Общедоступные ресурсы биоинформатики: биологические базы данных, геномный браузер UCSC: учебно-методическое пособие к практическим занятиям по дисциплине «Биоинформатика» / сост. Шилов Б.В., Лагунин А.А. — Москва: ФГАОУ ВО РНИМУ им. Н.И. Пирогова, 2022. 56 с.

ISBN 978-5-88458-660-4

В результате изучения материалов учебного пособия студенты смогут осуществлять поиск и корректно использовать общедоступную биологическую информацию, которая содержится в базах знаний и связана с различными разделами биоинформатики и других биологических наук, на современном уровне. Студенты познакомятся с наиболее популярными ресурсами: NCBI, UniProt, OMIM, KEGG и смогут самостоятельно анализировать разные геномные данные, используя такие инструменты, как геномные браузеры. Учебное пособие составлено в соответствии с действующими ФГОС и ОС ВО по специальностям 30.05.03 «Медицинская кибернетика», 30.05.02 «Медицинская биофизика», 30.05.01 «Медицинская биохимия» и магистратуры по направлению обучения 06.00.00 «Биологические науки» и предназначено для организации самостоятельной работы (практического занятия) обучающихся по дисциплине «Биоинформатика».

УДК 575.112(075.8)
ISBN 978-5-88458-660-4 ББК 30.16я73

© Шилов Б.В., Лагунин А.А., 2023
© ФГАОУ ВО РНИМУ имени Н.И. Пирогова Минздрава России, 2023

Содержание

Введение 4
Место дисциплины в структуре образовательной программы 5
Общедоступные базы данных. Работа с базами данных NCBI 6
Практическое задание 1 12
Практическое задание 2 21
Шаблон отчёта 23
Вопросы для самоконтроля 24
Полезные видеоресурсы 24
Базы данных OMIM, UniProt и KEGG 25
База данных OMIM 25
MIMmatch 33
Пример поиска в OMIM 34
База данных UniProt 35
База данных KEGG 37
Вопросы и практическое задание 39
Геномный браузер UCSC 42
Задача 1 45
Задача 2 47
Задача 3 49
Полезный видеоресурс 51
Заключение 52
Список литературы 53

Введение

Биоинформатика — одно из актуальных направлений науки о человеке и окружающей среде, которое стало важнейшей базой для открытий в современной биологической науке. Биоинформатике уделяется большое внимание как в научной, так и в образовательной сфере, поскольку она объединяет в себе знания биологических особенностей существования и развития любых организмов и навыки использования компьютерных инженерных подходов для анализа и интерпретации биологических данных. Наиболее значительный прогресс в области биоинформатики связан с изучением нуклеотидных последовательностей, что углубляет понимание законов изменения биологических молекул (ДНК, РНК, белков), происходящих в процессе эволюции. Работа с такими последовательностями стала практически повседневной задачей многих лабораторий (как исследовательских, так и клинических). Такое положение предполагает постоянное совершенствование знаний о накопленных в текущий момент данных и принципах организации их хранения, доступа к ним и их интерпретации в биологическом контексте.

В методическом пособии изложены подходы к поиску, доступу, получению и анализу разносторонней информации, связанной с биологическими полимерами. Информация о них опубликована в специальных базах данных и даёт возможность разобраться в вопросах, касающихся описания фенотипических характеристик объекта или связать определённые изменения в них с заболеванием.

Целью данного пособия является знакомство студента с основными понятиями и терминами, используемыми в биоинформатике, получение навыка работы с основными общедоступными базами данных и инструментами практической биоинформатики.

Пользуясь данным пособием, студенты смогут осуществлять поиск, корректно и на современном уровне использовать общедоступную биологическую информацию, которая содержится в базах знаний и связана с различными биологическими

науками. Они приобретут важный навык, который позволит разбираться в структуре и особенностях представления данных о биополимерах.

В пособии использована информация наиболее популярных открытых ресурсов: NCBI, UniProt, OMIM, KEGG. Студенты смогут самостоятельно анализировать различные геномные данные, используя соответствующие браузеры, навык работы с которыми в генетическом исследовании становится стандартом.

Пособие рассчитано на студентов специальностей 30.05.03 «Медицинская кибернетика», 30.05.02 «Медицинская биофизика», 30.05.01 «Медицинская биохимия», магистратуры по направлению обучения 06.00.00 «Биологические науки», а также может быть интересно врачам — генетикам и эмбриологам.

Место дисциплины в структуре образовательной программы

Общедоступные ресурсы биоинформатики: биологические базы данных и геномные браузеры изучаются в курсе дисциплины «Биоинформатика» после изучения студентами дисциплин «Биология», «Биохимия» и «Информатика». Знания, полученные студентами в процессе изучения данного раздела, могут в дальнейшем быть применены в ходе написания выпускной квалификационной работы.

Студент должен знать:

- основные принципы наполнения баз данных NCBI Gene, OMIM, UniProt, KEGG;
- особенности и основные элементы записей в этих базах данных;
- принцип работы с геномами в геномных браузерах на примере UCSC Genome Browser.

Студент должен уметь:

- осуществлять поиск информации в базе данных NCBI Gene;
- проводить поиск информации, связанной с наследственными фенотипами в базе данных OMIM;
- проводить поиск заданной информации о белках, сигнальных и метаболических путях в базах данных UniProt и KEGG;
- находить заданные характеристики генома с применением геномного браузера UCSC Genome Browser.

Общедоступные базы данных. Работа с базами данных NCBI

Одна из главных задач современной биологической и медицинской науки состоит в том, чтобы искать новые подходы к работе с объёмными и сложными данными, получаемыми в результате исследований, а также обеспечить исследователям качественный доступ к аналитическим и вычислительным инструментам, чтобы лучше понять наше генетическое наследие и его роль в формировании заболеваний.

Национальный центр биотехнологической информации — The National Center for Biotechnology Information (NCBI) является одним из важнейших мировых ресурсов для хранения и анализа различных биологических данных. Его поддерживают Национальная медицинская библиотека (National Library of Medicine) и Национальные институты здоровья (National Institutes of Health) США. Данный ресурс представляет собой хорошо структурированную сеть взаимосвязанных баз данных, содержащих генетическую, медико-биологическую, биохимическую информацию о различных биологических видах, постоянно дополняемую разными лабораториями. Совокупное увеличение объёма данных в базах за счёт развития технологий происходит экспоненциально. Полные геномы от вирусов до человека представлены определённым образом и имеют перекрёстные ссылки, что позволяет легко получать доступ к связанным данным.

Важно отметить, что поисковые системы типа Google или Yandex обычно не индексируют веб-сайты, основанные на больших специализированных базах данных. Для поиска в системе ресурсов NCBI имеются специальные инструменты, позволяющие эффективно получать информацию и знания, связанные с определённой темой.

Для начала зайдём на сайт NCBI (URL: http://www.ncbi.nlm.nih.gov/) и изучим его интерфейс, который постоянно модифицируется (**рис. 1**). В левой части экрана присутствует меню «NCBI

Home», эта ссылка переадресует вас из любой локации NCBI на главную страницу. Если необходимо узнать, как работает тот или иной инструмент или как организована база данных, можно воспользоваться вкладкой «How To» при выборе любого пункта меню слева. Обратите внимание на ссылку «Resource List (A–Z)», нажав на которую вы попадёте на страничку, где перечислены все базы данных и компьютерные программы NCBI.

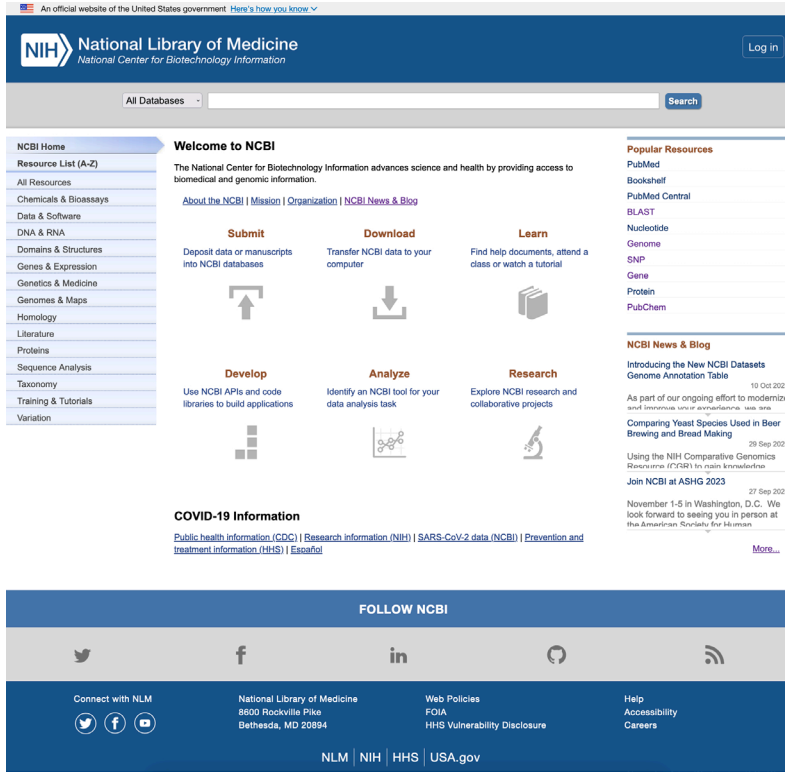


Рис. 1. Главная страница сайта NCBI

Раздел «How To» содержит справочную информацию и позволяет помочь разобраться в том, как можно провести тот или иной поиск в базах данных NCBI (**рис. 2**). К примеру, мы можем узнать, как правильно скачать полный геном организма. Информация об этом находится в разделе «Data&Software» (URL: http://www.ncbi.nlm.nih.gov/guide/howto/dwn-genome/). Вы-

брав любой из пунктов меню, мы получим информацию и ссылки на конкретные поисковые страницы, при этом в правой части экрана будут отображаться быстрые ссылки на часть из этих же страниц (сокращённый список ресурсов, которые являются наиболее востребованными в данном контексте).

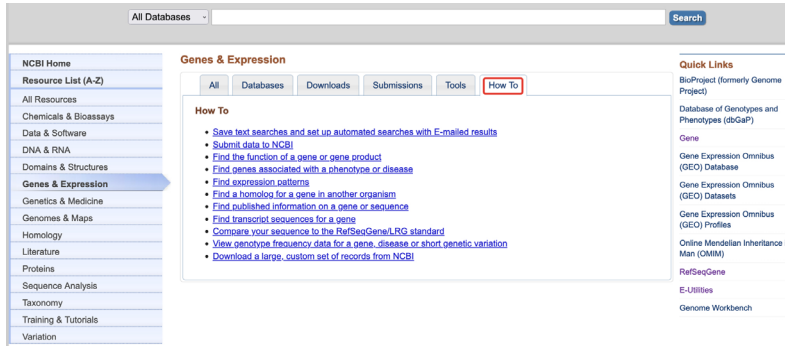


Рис. 2. Вкладка How To помогает узнать, как можно использовать элементы меню

Рассмотрим краткую информацию об одном из ключевых разделов NCBI, который называется GenBank. GenBank — база данных, содержащая аннотированные нуклеотидные последовательности секвенированных ДНК организмов (для кодирующих последовательностей приведена трансляция). Под словом «аннотированные» понимают, что запись формата genbank представлена с описанием всех важных изученных элементов, содержащихся в данной последовательности (генов, кодирующих последовательностей и т.д.).

GenBank доступен через поисковую систему NCBI Entrez, которая объединяет данные из основных баз данных, содержащих информацию о последовательностях ДНК и белков вместе с таксономией, геномом, картированием, структурой белка и информацией о домене, а также литературой по биомедицинским журналам, представленным в PubMed.

В **листинге 1** приведён пример аннотированной записи в формате GenBank. GenBank — участник консорциума молекулярно-биологических баз данных совместно с European Nucleotide Archive (ENA) и DNA DataBank of Japan (DDBJ) (базы данных обмениваются данными ежедневно, и поэтому записи в них

эквивалентны; описания и индексы всех элементов последовательностей совпадают).

```
LOCUS       SCU49845 5028 bp DNA PLN 1-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene,
            partial cds, and Axl2p (AXL2) and Rev7p
            (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1 GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
ORGANISM    Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota;
            Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1 (bases 1 to 5028)
AUTHORS     Torpey, L.E., Gibbs, P.E., Nelson, J.
            and Lawrence, C.W.
TITLE       Cloning and sequence of REV7, a gene whose
            function is required for DNA damage-induced
            mutagenesis in Saccharomyces cerevisiae
JOURNAL     Yeast 10 (11), 1503-1509 (1994)
PUBMED      7871890
REFERENCE   2 (bases 1 to 5028)
AUTHORS     Roemer, T., Madden, K., Chang, J.
            and Snyder, M.
TITLE       Selection of axial growth sites in yeast
            requires Axl2p, a novel plasma membrane
            glycoprotein
JOURNAL     Genes Dev. 10 (7), 777-793 (1996)
PUBMED      8846915
REFERENCE   3 (bases 1 to 5028)
AUTHORS     Roemer, T.
TITLE       Direct Submission
JOURNAL     Submitted (22-FEB-1996) Terry Roemer,
            Biology, Yale University, New Haven, CT, USA
FEATURES    Location/Qualifiers
source      1..5028
            /organism = "Saccharomyces cerevisiae"
            /db_xref = "taxon:4932"
            /chromosome = "IX"
            /map = "9"
```

```
CDS             < 1..206
                /codon_start = 3
                /product = "TCP1-beta"
                /protein_id = "AAA98665.1"
                /db_xref = "GI:1293614"
                /translation = "SSIYNGISTSGLDLNNGTIADMR
                QLGIVESYKLKRAVVSSASEAAEVLLRVDNIIRARPRTANRQHM"
gene            687..3158
                /gene = "AXL2"
CDS             687..3158
                /gene = "AXL2"
                /note = "plasma membrane glycoprotein"
                /codon_start = 1
                /function = "required for axial budding pattern of S. cerevisiae"
                /product = "Axl2p"
                /protein_id = "AAA98666.1"
                /db_xref = "GI:1293615"
                /translation = "MTQLQISLLLTATISLLHLVVATP
                YEAYPIGKQYPPVARVNESFTFQISNDTYKSSVD
                KTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLS
                DANTTLYFNVILEGTDSADSTSLNNTYQFVVTN
                RPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPN
                WLFFDSGELKFTGTAPVINSAIAPETSYSFVIIATD
                IEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTG
                NVSYDLPLNYVYLDDDPISSDKLGSINLLDAPDW
                VALDNATISGSVPDELLGKNSNPANFSVSIYDTY
                GDVIYFNFEVVSTTDLFAISSLPNINATRGEWFSY
                YFLPSQFTDYVNTNVSLEFTNSSQDHDWVKFQS
                SNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFN
                IIGMDSKITHSNHSANATSTRSSHHSTSTSSYTS
                STYTAKISSTSAAATSSAPAALPAANKTSSHNKK
                AVAIACGVAIPLGVILVALICFLIFWRRRRENPDDE
                NLPHAISGPDLNNPANKPNQENATPLNNPFDD
                DASSYDDTSIARRLAALNTLKLDNHSATESDISS
                VDEKRDSLSGMNTYNDQFQSQSKEELLAKPPV
                QPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTG
                NLSPVSDIVRDSYGSQKTVDTEKLFDLEAPEKEK
                RTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYN
                VTKHRNRHLQNIQDSQSGKNGITPTTMSTSSSD
                DFVPVKDGENFCWVHSMEPDRRPSKKRLVDFS
                NKSNVNVGQVKDIHGRIPEML"
gene            complement(3300..4037)
```

```
                /gene = "REV7"
CDS             complement(3300..4037)
                /gene = "REV7"
                /codon_start = 1
                /product = "Rev7p"
                /protein_id = "AAA98667.1"
                /db_xref = "GI:1293616"
                /translation = "MNRWVEKWLRVYLKCYINLILFY
                RNVYPPQSFDYTTYQSFNLPQFVPINRHPALIDY
                IEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLD
                FSLQHVDKDDQIITETEVFDEFRSSLNSLIMHLEK
                LPKVNDDTITFEAVINAIELELGHKLDRNRRVDSL
                EEKAEIERDSNWVKCQEDENLPDNNGFQPPKIK
                LTSLVGSDVGPLIIHQFSEKLISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Листинг 1. Пример записи в формате GenBank

Файл этого формата содержит информацию о последовательности и представляет собой простой текстовый файл, в котором с помощью ключевых слов представлены разделы описания. Например, в приведённом выше примере можно увидеть раздел ***LOCUS***, в котором есть информация о длине последовательности, какой полимер она описывает, её идентификатор и дата её публикации. Далее представлены разделы файла, описывающие определённые особенности последовательности (полное описание формата можно изучить по ссылке: URL: https://www.insdc.org/documents/feature_table.html, а запись, представленная в примере, описана на странице: URL: https://www.ncbi.nlm.nih.gov/genbank/samplerecord/). В файле можно увидеть большой диапазон представляемых описаний, включая описания регионов последовательности, которые выполняют определённую биологическую функцию, влияют или являются результатом проявления биологической функции, взаимодействуют с другими молекулами, влияют на репликацию

последовательности и т.п. Давайте рассмотрим некоторые из них. Обязательный раздел ***source*** описывает принадлежность последовательности к определённому виду. Разделы ***CDS*** являются описанием кодирующей области последовательности, т.е. последовательность нуклеотидов ДНК, в которой закодирована цепь аминокислот в белке. В разделе ***gene*** описана область биологического интереса, обозначенная как ген и для которой было присвоено имя. Слово complement в описании гена говорит о том, что признак расположен на комплементарной цепи ДНК.

Практическое задание 1

В результате прохождения практикума необходимо составить отчёт — документ Word, содержащий ответы на все поставленные вопросы и таблицу с описанием предложенного в задании гена.

Откройте веб-браузер и зайдите на страницу: www.ncbi.nlm.nih.gov. Эта страница обеспечивает быстрый доступ ко всем основным ресурсам и базам данных NCBI. Можно определить для себя, как наиболее эффективно переходить из одной части NCBI в другую. Перейдя по ссылке **NCBI News & Blog - > More...**, вы можете узнать краткую информацию о некоторых ресурсах.

В разделе **Literature** можно получить справочную информацию, научиться основным приёмам работы в NCBI. Достаточно полный вариант предложен в разделе руководство (NCBI Handbook). Рекомендуем изучить данный ресурс для более эффективной работы в NCBI.

Проведите следующее исследование:

1. Найдите в разделе NCBI Information (NCBI news) данные о том, какой релиз базы данных RefSeq является последним (укажите версию) и что он содержит.
2. Какой релиз GenBank актуален для скачивания по FTP?
3. Когда опубликована последняя новость и о чём она?
4. Чтобы узнать общую статистику записей в разных базах данных, воспользуйтесь базой данных Taxonomy, введите идентификатор или название вида и в правой части изучите таблицу с количеством записей.
5. Сколько генов человека в разделе при поиске информации об этом в разделе Taxonomy?

Результаты этого задания представьте в виде краткого отчёта, содержащего ответы на все вопросы.

Одной из возможностей NCBI является поиск запроса по всем базам банных NCBI и вывод результата о количестве совпадений в каждой из них.

Перейдите на портал *GQuery* (поиск *через поисковую строку, установив при этом в выпадающем меню слева от поисковой строки параметр «*All Databases»). Если в строке навигации NCBI выбран пункт *All Databases,* а в поле запроса ничего не введено, то запуск поиска (Search) приведёт к открытию портала *GQuery*. Вы увидите все базы данных от PubMed (БД по медико-биологической литературе) до баз данных таксономии, нуклеотидных последовательностей, белковых структур и т.д. с кратким резюме по каждой из них. Через этот интерфейс можно организовать поиск какого-либо термина (гена, белка, организма, таксономической группы и т.п.). При этом напротив каждого названия из баз данных NCBI отобразится количество найденных записей с искомым термином. Например, поиск термина «homo sapiens» даст результаты, содержащие тысячи совпадений в большинстве баз данных **(рис. 3)**. Чтобы получать приемлемое число результатов, необходимо конкретизировать запрос.

Обычно при поиске в этих базах данных необходимо выбрать область ДНК или белок (или функцию белка). В данной практической работе нужно использовать предложенный в задании ген, кодирующий один белок человека из хорошо

RefSeq (Reference Sequence) проект

Цель проекта RefSeq — предоставить наилучшую репрезентативную последовательность для каждого нормального (т.е. не мутантного) транскрипта, полученного из гена, или для любого нормального белка. В Refseq представлены не только гены, но и целые геномы. GenBank содержит сотни accession numbers, соответствующих одному гену, что приводит к избыточности данных, т.к. GenBank хранит архивную информацию. Тем не менее для каждого гена, транскрипта или, например, генома будет только одна RefSeq запись. Несколько RefSeq записей возможно при наличии сплайс-вариантов или различающихся локусов одного гена (т.е. для каждого гена одна запись и несколько записей для его альтернативных транскриптов или кодируемых белковых последовательностей).

RefSeq записи имеют различные статусы: predicted (предсказанный), provisional (временный) и reviewed (рассмотренный).

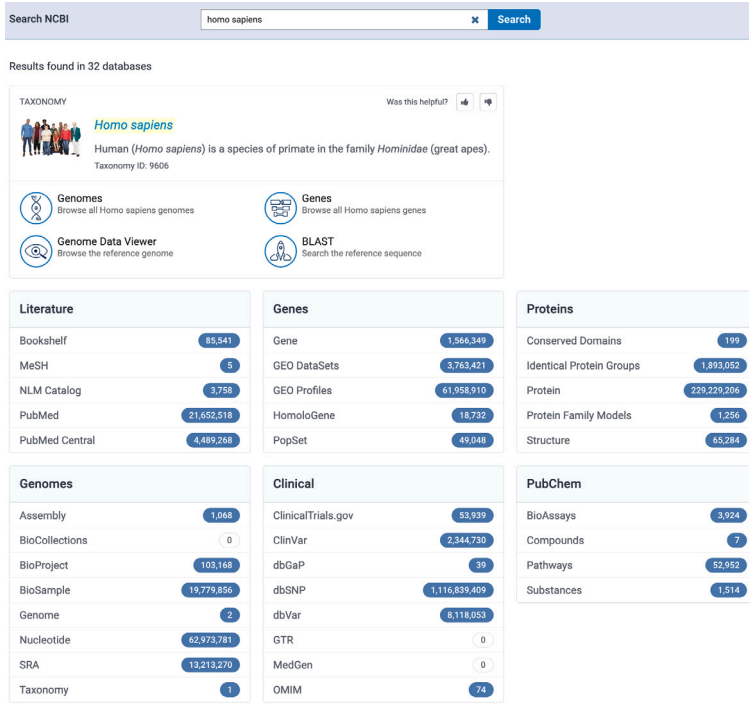


Рис. 3. Вид портала GQuery NCBI с реализованным поисковым запросом «homo sapiens»

аннотированной базы данных референсных последовательностей RefSeq.

Для эффективного поиска необходимо сузить поисковый запрос, чтобы получить оптимальные результаты и максимально ограничить область поиска. Для этого лучше использовать поисковые термины, ассоциированные с генной последовательностью и квалификаторы GenBank (так называемые Field Qualifiers GenBank). Квалификаторы указываются в квадратных скобках и некоторые из них представлены ниже:

- ключевые слова, связанные с изучаемым геном (пример*: structural constituent of ribosome*);
- ключевые слова AND организм (пример*: structural constituent of ribosome AND Arabidopsis thaliana*);
- ключевые слова в определённых полях, если интересует белок [PROT] AND организм [ORGN] (пример*: structural*

constituent of ribosome [PROT] AND Arabidopsis thaliana [ORGN]);

- accession number (если известны идентификационные номера, то поиск значительно оптимизируется) (пример*: accession number: NP_565676*).

Применение таких настроек позволяет значительно сузить область поиска. Обратите внимание, что **использование круглых скобок** позволяет записывать логические выражения и использовать более сложные конструкции, что может оказаться полезным при формировании более качественного запроса. При этом необходимо убедиться, что вы получите именно то, что вы хотите.

Пример:

SMC AND (yeast [ORGN] OR Arabidopsis [ORGN])
даст другой результат, чем
SMC AND yeast [ORGN] OR Arabidopsis [ORGN]

ВАЖНО:
Логические операторы в выражениях, такие как AND/OR/NOT, необходимо всегда писать с заглавной буквы.

Кавычки тоже могут повлиять на результат поиска, т.к. будет проводиться поиск всей фразы внутри скобок, а не её отдельных слов. Проверьте на данном примере: сравните запрос 16s rRNA и «16s rRNA». В данном примере результаты будут отличаться, т.к. в случае с применением кавычек поисковая система будет искать точную фразу, заключённую в кавычки, а без использования кавычек поисковая система будет просто искать все слова запроса, которые должны встретиться в искомой записи.

Подробнее узнать об использовании квалификаторов (параметров поиска) и научиться ими пользоваться можно по ссылке URL: http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options

В конечном счёте наиболее точный поиск будет проведён, если вы используете специальные идентификаторы Accession Numbers, которые присвоены каждой записи базы данных.

Для систематизации информации о разделах NCBI можно привести следующую иллюстрацию (**рис. 4**): слева представлены биологические объекты, показывающие типы данных, находящиеся в различных базах NCBI, список которых приведён в

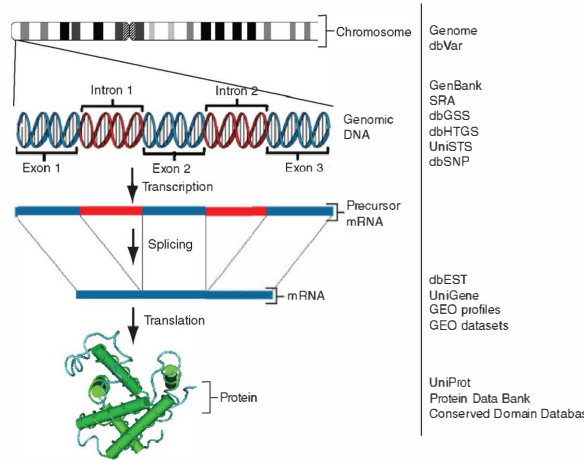


Рис. 4. Схема соответствия биологических объектов и баз данных NCBI.

Источник: Pevsner J. Bioinformatics and Functional Genomics. Hoboken: John Wiley & Sons; 2015. URL: https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1118581784&bcsId=9870

правой колонке. Данная схема согласуется с центральной догмой молекулярной биологии.

Для лучшего понимания медицинской составляющей описания генов необходимо обратить внимание на такие базы данных, как ClinVar, dbVar и dbSNP, в которых содержится описание встречающихся вариаций генома и ассоциация этих вариаций с теми или иными фенотипами и заболеваниями.

Что такое Accession Numbers, Version Numbers и GI Numbers

Accession number — это уникальный идентификатор для конкретной записи в базах данных. Accession number связан с записью навсегда. Другими словами, Accession number представляет собой указатель на конкретную запись и не изменится, даже если информация в самой записи будет изменяться автором (например, при улучшении аннотации или уточнённом секвенировании нуклеотидной последовательности). Accession number, как правило, — это комбинация из букв(ы) и чисел: 1 буква и за ней 5 цифр (например, U12345) или 2 букв и 6 цифр (например, AF123456).

Version Number — номер версии записи. Этот параметр должен следовать за Accession number. Он указывает на историю пересмотра этой записи. Начинается с 1 и увеличивается с каждой новой версией. Стандартный формат представляет собой запись Accession.Version.

Пример: AF123456.2, где AF123456 — Accession number, а 2 — номер версии.

Для NCBI приняты следующие префиксы, характеризующие последовательности:

NT_123456 — построенные геномные сборки (контиги);

NM_123456 — mRNA (иРНК);

NP_123456 — proteins (белки);

NC_123456 — хромосомы.

GI number (GenInfo Identifier) — идентификатор GenInfo (иногда пишется маленькими буквами «gi»). Это последовательность цифр, которые присваиваются последовательно каждой записи, содержащейся в NCBI. GI система идентификаторов существует параллельно с системой Accession. Version; однако если последовательность ДНК или белка по какой-то причине изменяется, то в любом случае она получит новый номер GI по порядку присвоения номеров.

С сентября 2015 г. в NCBI больше не пользуются идентификаторами GI.

Существует большое количество префиксов для идентификаторов, большинство из которых описано на соответствующей странице NCBI: URL: https://www.ncbi.nlm.nih.gov/genbank/acc_prefix/ и в книге NCBI Handbook: URL: https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report = objectonly/ (**табл. 1**).

Пример: Когда новая запись представляется в GenBank, ей назначается Accession number (скажем, AF000001). Поскольку это первая версия, то к Accession number будет добавлена версия «.1», и запись будет выглядеть следующим образом: AF000001.1. В то же время она получит GI (скажем, GI: 1234567).

Теперь представьте, что исследователь, который первоначально представлял протокол, хочет обновить информацию. Обновлённая запись сохранит тот же Accession number, но увеличится номер версии (AF000001.2), при этом этой записи будет присвоен совершенно новый GI (скажем GI: 9876543).

Почему это важно? Accession number всегда будет давать вам самую последнюю информацию о версии записи, в то время как число GI всегда свяжет вас с конкретной записью. Есть моменты, когда вы хотите иметь наиболее актуальную информацию, или в других случаях, когда вы хотите указать на конкретную информацию, связанную с определённой временной точкой (например, конкретную запись, которую вы анализировали в определённое время вашего исследования), даже если значительную часть информации добавляли уже после проведённого исследования.

Таблица 1. Список основных префиксов идентификаторов записей в базах данных NCBI

Префикс идентификатора	Тип молекулы	Описание
AC_	Genomic	Полный геном, альтернативная сборка
NC_	Genomic	Полный геном, референс
NG_	Genomic	Незавершённый геномный регион
NT_	Genomic	Контиги или скаффолды, созданные на основе клонов или полногеномного секвенирования
NW_	Genomic	Контиги или скаффолды, преимущественно на основе полногеномного секвенирования
NZ_	Genomic	Завершённые геномы и незавершённые полногеномные сборки
NM_	mRNA	Белоккодирующие транскрипты
NR_	RNA	Небелоккодирующие транскрипты
XM_	mRNA	Предсказанные в модели белоккодирующие транскрипты
XR_	RNA	Предсказанные в модели небелоккодирующие транскрипты
AP_	Protein	Белок, аннотированный на альтернативной сборке AC_
NP_	Protein	Белок, ассоциированный с идентификаторами NM_ или NC_
YP_	Protein	Белок, аннотированный на основе геномных сборок без записи экземпляра транскрипта
XP_	Protein	Предсказанная последовательность белка, ассоциированная с записями XM_
WP_	Protein	Неизбыточная последовательность белка для нескольких штаммов или видов

Для начала разберём, какие типы вариаций можно встретить в геноме. Основные виды представлены в **табл. 2**.

ClinVar — это публичный архив, содержащий информацию об известных взаимоотношениях между вариациями генома и клиническими фенотипами человека. В ClinVar описаны как единичные аллели, так и сложные гетерозиготы, гаплотипы и комбинации аллелей в разных генах, аннотированные с точки зрения клинической значимости.

Научная группа ClinGen тесно сотрудничает с NCBI, развивая функциональность базы данных ClinVar, и занимается поддержкой поступления новых данных из множества разных источников.

База данных dbVar накапливает информацию о структурных вариациях, которые встречаются в геноме и описаны в соответствующих публикациях. Структурная вариация — это область ДНК размером от тысячи пар оснований и может вклю-

Таблица 2. Типы вариантов, которые могут встречаться в геноме

Тип варианта	Пример вариации	Описание
Замена	TTC**G**AA TTC**C**AA	Один нуклеотид был заменён на другой
Делеция	TT**G**CAA TT-CAA	Один или несколько нуклеотидов были потеряны
Вставка	TTG-CAA TTG**A**CAA	Один или несколько нуклеотидов были вставлены в последовательность
Дупликация	TTG-**C**AA TTG**CC**AA	Один или более нуклеотидов, которые были тандемно удвоены
Делеция–вставка	TT**GC**AA TT**AT**AA	Один или более нуклеотидов были удалены или заменены на один или более других нуклеотидов
Инверсия	TT**GC**AA TT**CG**CA	Последовательность из более чем одного нуклеотида была развёрнута по отношению к обратно-комплементарной последовательности
Структурная вариация		Вариация, при которой большая часть хромосомы была перестроена

чать инверсии и сбалансированные транслокации или геномные дисбалансы. Часто такие изменения генома ассоциированы с клиническими фенотипами.

Теперь рассмотрим одну из самых известных баз данных NSBI — базу данных единичных и малых вариаций генома, которая называется dbSNP (SNP — single nucleotide polymorphism, однонуклеотидный полиморфизм)[1]. В dbSNP также хранится информация о коротких изменениях последовательности, а не только для однонуклеотидных замен, которые достаточно часто встречаются в популяции. С 2011 г. полное имя базы данных было изменено на «базу данных коротких генетических вариаций». Аббревиатура, представляющая базу данных, тем не менее осталась прежней — «dbSNP».

dbSNP включает в себя обширную информацию о простых генетических вариациях:

- единичных нуклеотидных заменах;
- малых делециях или вставках, содержащих несколько нуклеотидов;
- микросателлитных повторах.

Все типы вариаций, перечисленные ниже, имеют точное местоположение в геноме и могут быть как распространёнными, так и редкими вариациями, включая мутации. Большинство из них, как правило, представляют собой небольшие вариации (до 50 пар оснований):

- однонуклеотидная вариация;
- малые многонуклеотидные изменения;
- небольшие делеции или вставки;
- короткие тандемные повторы;
- ретротранспозируемые вставки элементов.

Кроме этого, описанные базы содержат детальную информацию о вариантах и их клинической значимости. В **табл. 3** представлены градации вариаций с точки зрения их клинической значимости.

[1] Видео с описанием базы данных dbSNP:
URL: https://www.youtube.com/watch?v=gpSI16e59ig

Таблица 3. Классификация вариантов по клинической значимости

Классификация	Вероятность патогенности, %
Pathogenic	> 99
Likely pathogenic	90–99
Variant of uncertain significance	10–90
Likely benign	0,1–10,0
Benign	< 0,1

Практическое задание 2

Провести поиск информации по предложенному в задании гену. Для этого можно воспользоваться страницей портала *GQuery* или выбрать базу данных Gene в выпадающем списке. В ответ на запрос будет выведен список записей (**рис. 5**). Необходимо выбрать нужный ген и исследовать его, заполнив соответствующую таблицу (шаблон таблицы находится в конце этой главы).

Обратите внимание, что текст на странице о гене содержит гиперссылки. Это сделано для простоты понимания информации и быстрой навигации по разделам NCBI. Например, для таксономической информации нажмите на гиперссылку организма-источника (в разделе ***Source organism***). Некоторые записи

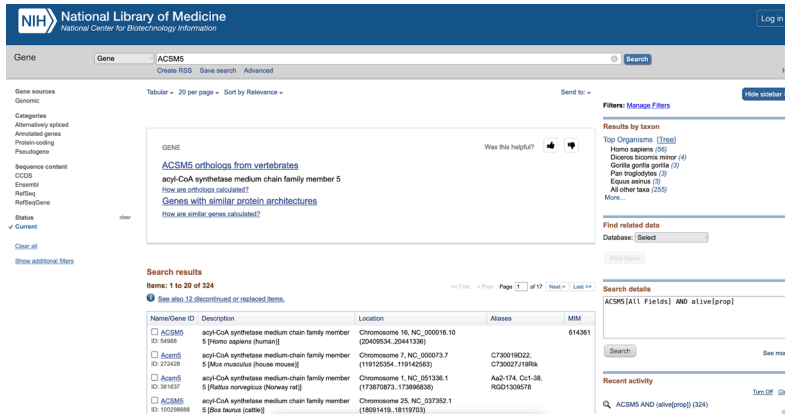


Рис. 5. Результат поиска для гена ACSM5 по запросу имени гена

имеют ссылки на первичные издания, журналы, в которых эта последовательность была впервые опубликована, она связана гиперссылкой с номером в базе данных PubMed. Нажмите на разные гиперссылки и посмотрите, куда вы попадёте при переходе, а также ответьте на вопрос: каково полное таксономическое положение изучаемого организма?

Часто необходимо иметь информацию о последовательности в определённом формате. Наиболее распространёнными форматами являются **GenBank (gb)** и **FASTA**, которые соответствующим образом предоставляют информацию о последовательности из базы данных NCBI.

Формат FASTA является стандартным представлением последовательности и содержит два основных элемента: в первой строке содержится информация описательного характера, например, accession number, название последовательности, описание и т. п. Такая строка всегда начинается со знака «>», второй элемент в файле fasta — это сама последовательность. Файл может содержать последовательности любых биополимеров (ДНК, РНК, белков), при этом последовательность записывается в однобуквенном коде, принятом международным химическим союзом IUPAC; URL: https://www.bioinformatics.org/sms/iupac.html

Чтобы скачать файл в FASTA формате, необходимо выбрать последовательность и выбрать тип скачиваемой последовательности, как показано на **рис. 6**.

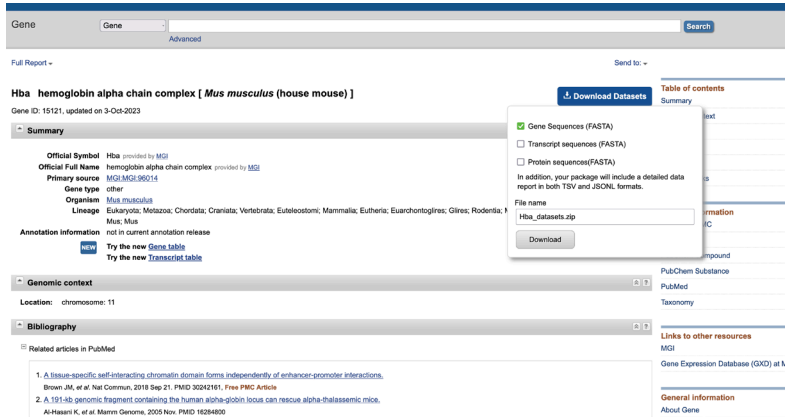


Рис. 6. Элементы интерфейса записи в NCBI с выбором вариантов скачивания файла с последовательностью

Пользуясь информацией к занятию, справочной системой NCBI и разделом HowTo, заполните таблицу (приведённую в разделе «Шаблон отчёта») информацией о предложенном к изучению гене. Приложите файлы последовательности и сохраните в личной папке на общем диске.

Шаблон отчёта

1. Название гена (рус.).
2. Идентификатор гена и сокращённое название.
3. Официальное полное название гена (англ.).
4. Синонимы.
5. Тип гена.
6. Статус в RefSeq.
7. Краткое описание гена: функция самого гена и его продукта (рус).
8. Место расположения в геноме (хромосома, плечо, локус, цепь).
9. Количество экзонов гена.
10. Версия текущей сборки аннотации генома.
11. Точное местоположение последовательности указанного гена (в п. н.) и идентификаторы последовательности, содержащей этот ген, на основе текущей версии сборки.
12. Размер гена в п. н.
13. В какой ткани (тканях) наиболее сильно экспрессируется данный ген.
14. Название статей (от 1 до 5 шт.) и их идентификаторы, описывающие данный ген (ближайшие по дате публикации).
15. Файл с полным текстом статьи о гене в формате pdf (указать название, файл приложить к отчёту).
16. GeneRIF — статья о связи гена с функцией (описать связь из любой статьи в разделе GeneRIF: Gene Reference into Function).
17. Приведите описание фенотипов (синдромы и проявления, связанные с нарушением в гене) в разделе Phenotypes — Description.
18. Опишите одну любую патогенную вариацию гена, представленную в базе данных dbVAR: идентификатор региона вариации, тип вариации, число вариаций в данном регионе, клиническая значимость, локализация вариации, затронутые гены.
19. Опишите одну любую патогенную вариацию гена, представленную в базе данных ClinVar: расположение вариации,

ген(ы), изменения белка, известные состояния — condition(s), клиническая значимость, статус — review status, номер доступа вариации.

20. Файлы с последовательностями RefSeq: гена, РНК и белка (указать название, файлы приложить к отчёту).

Вопросы для самоконтроля

1. Как организовать поиск по всем базам данных NCBI?
2. Как скачать полный геном какого-либо организма?
3. Какие разделы входят в GenBank?
4. Перечислите способы доступа к данным GenBank.
5. Что такое Accession Number?
6. Что такое GI?
7. Как попасть в портал GQuery?
8. Что такое Version Number?

Полезные видеоресурсы

- Официальный канал NCBI
  URL: https://www.youtube.com/user/NLMNIH
- NCBI 20th Anniversary Video
  URL: http://www.youtube.com/watch?v=kyKGn6Ih3EM
- NCBI Tutorial
  URL: https://www.youtube.com/watch?v=hTTvq8KJthA
- Gene NCBI
  URL: https://www.youtube.com/watch?v=j7hV10gYz1Q
- RefSeq
  URL: https://www.youtube.com/watch?v=zs46Ur0m0mc&feature=youtu.be

Базы данных OMIM, UniProt и KEGG

Ранее нами изучены базы данных на ресурсе NCBI, в первую очередь база NCBI Gene, которая интегрирует из различных источников информацию об отдельном гене: последовательность, полиморфизмы и т.д. На занятии будет рассмотрена работа со следующими базами данных:

- генетика наследственных заболеваний (OMIM);
- информация о белках (UniProt);
- информация о метаболических путях (KEGG).

Цель занятия состоит в изучении взаимосвязей между наследственным заболеванием, функцией белка, кодируемого мутированным геном, и проявлением этого нарушения на уровне метаболических каскадов. На этом занятии предлагается изучить информацию об одном наследственном заболевании и проанализировать его молекулярные причины.

База данных OMIM

OMIM (Online Mendelian Inheritance in Man, URL: http://www.omim.org/) представляет собой большую коллекцию описаний человеческих генов и генетических фенотипов, которая находится в открытом доступе и ежедневно обновляется. Ресурс OMIM также доступен через NCBI (URL: http://www.ncbi.nlm.nih.gov/omim). OMIM содержит полнотекстовые обзоры и другую информацию обо всех известных заболеваниях, наследуемых по менделевскому типу (более 12 000 генов). OMIM фокусируется на отношениях между фенотипом и генотипом, содержит множество ссылок на другие генетические ресурсы (**рис. 7**).

OMIM основана на книге доктора Виктора А. МакКьюсика (Dr. Victor A. McKusick) «Менделирующее наследование у человека», которая была опубликована в 12 томах, последний из которых вышел в 1998 г. Сейчас для создания записей OMIM используют рецензируемую биомедицинскую литературу, а приоритет для включения отдаётся тем статьям, которые дают

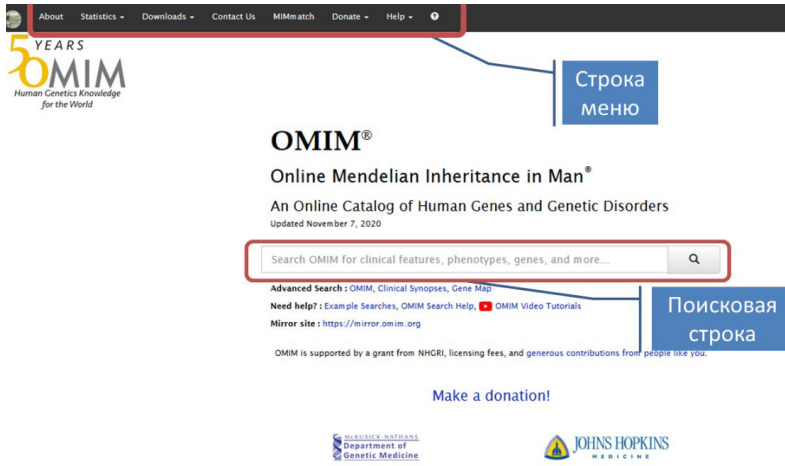


Рис. 7. Важные элементы управления базы данных OMIM

значимое представление о взаимосвязи ген–фенотип, расширяют наше понимание о биологии человека или детально характеризуют нарушения.

OMIM предназначен для использования в первую очередь врачами и другими специалистами, занимающимися генетикой или исследованиями в области генетики, для поиска информации о фенотипах и связанных с ними генах.

Несмотря на то что база данных OMIM открыта для любого пользователя, тем, кто ищет информацию о личном медицинском или генетическом состоянии и лечении, настоятельно рекомендуется проконсультироваться с квалифицированным врачом для постановки диагноза и получения ответов на личные вопросы.

В OMIM используются характерные уникальные номера доступа (accession numbers) и обозначения для записей. Каждой записи OMIM присваивается уникальный шестизначный номер по следующему принципу:

1 ----- (100000-) и 2 ----- (200000-) — Аутосомные локусы или фенотипы (записи, созданные до 15.05.1994);
3 ----- (300000-) — X-сцепленные локусы или фенотипы;
4 ----- (400000-) — Y-сцепленные локусы или фенотипы;
5 ----- (500000-) — Митохондриальные локусы или фенотипы;
6 ----- (600000-) — Аутосомные локусы или фенотипы (записи, созданные после 15.05.1994).

Аллельные варианты обозначаются номером OMIM записи, за которым следует десятичная точка и уникальный 4-значный номер варианта. Например, описание аллельных вариантов в гене фактора IX (300746) пронумерованы от 300746.0001 до 300746.0101 в разделе «Allelic variants» (**рис. 8**).

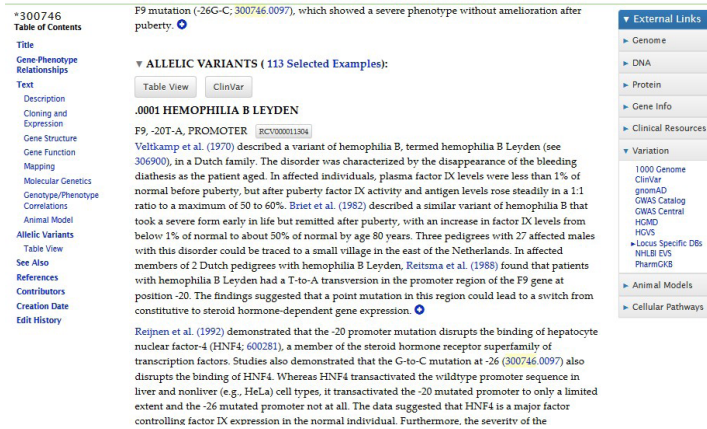


Рис. 8. Пример описания аллельных вариантов в гене IX фактора свёртывания

В раздел OMIM «Аллельные варианты» (Allelic Variants) для большинства генов включены только избранные, наиболее значимые мутации. Большинство аллельных вариантов представляют собой мутации, вызывающие заболевание.

Критерии включения:

- первая обнаруженная мутация;
- высокая частота в популяции;
- ярко выраженный фенотип;
- историческое значение;
- необычный механизм мутации;
- необычный патогенетический механизм;
- особый механизм наследования (например, доминантный с отдельными мутациями, рецессивный с прочими мутациями в том же гене).

Каждая запись дополнительно сопровождается условными обозначениями, которые дают представление о том, какой тип у данной записи.

В OMIM для описания генов и патологий используют понятия «карта генов» и «карта патологии» (Gene Map и Morbid Map). По сути, такие карты указывают цитогенетическое положение гена или нарушения, которые описаны в OMIM. В OMIM на генной карте и на морбидной карте представлены только те записи, цитогенетическое местоположение которых было опубликовано в цитируемых источниках.

Что обозначают символы перед номером OMIM?

- «*» перед номером указывает на ген;
- «#» перед номером записи указывает, что это запись описывает какой-либо фенотип и обычно не представляет собой уникальный локус. Обсуждение любого гена, относящегося к этому фенотипу, находится в другой записи, описывающей ген, обозначенной «*»;
- «+» перед номером записи означает, что запись содержит описание гена с изученными последовательностью и фенотипом;
- «%» перед номером записи указывает на то, что запись описывает подтверждённый менделевский фенотип или фенотипический локус, но лежащая в основе молекулярная основа формирования фенотипа неизвестна;
- «^» перед номером означает, что запись больше не существует, потому что она была удалена из базы данных или информация перемещена в другую запись;
- отсутствие символа перед номером записи обычно указывает на описание фенотипа, для которого менделевская основа хотя и предполагается, но не была чётко установлена, или что отделённость этого фенотипа от фенотипа какой-либо в другой статье неясна.

Генную карту OMIM можно искать по символу гена (например, «SOD1»), хромосомному положению (например, «6», «12pter», «Xp») или по ключевому слову расстройства (например, «citrullinemia»), перейдя по ссылке «Cytogenetic location» (**рис. 9**).

Карта генов представлена таблицей и имеет несколько полей, которые будут полезны исследователю (**рис. 10**). В таблице можно ограничить вывод только теми записями, для которых имеются характерные фенотипы (кнопка «Phenotype only Entries»), или отобразить все записи генной карты (кнопка «All Entries»). Кроме того, в разделе «Phenotype» имеются дополнительные обозначения, которые характеризуют конкретную запись.

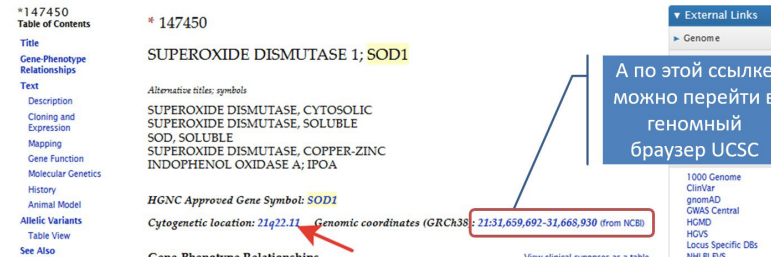
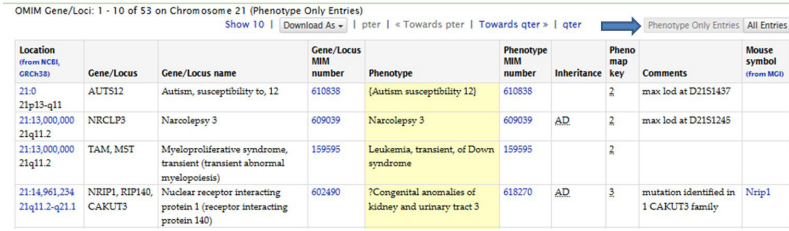


Рис. 9. Поиск расположения гена в геноме


OMIM Gene/Loci: 1 - 10 of 53 on Chromosome 21 (Phenotype Only Entries)

Show 10 | Download As | pter | « Towards pter | Towards qter » | qter | Phenotype Only Entries | All Entries

Location (from NCBI, GRCh38)	Gene/Locus	Gene/Locus name	Gene/Locus MIM number	Phenotype	Phenotype MIM number	Inheritance	Phenomap key	Comments	Mouse symbol (from MGI)
21:0 21p13-q11	AUTS12	Autism, susceptibility to, 12	610838	[Autism susceptibility 12]	610838		2	max lod at D2151437	
21:13,000,000 21q11.2	NRCLP3	Narcolepsy 3	609039	Narcolepsy 3	609039	AD	2	max lod at D2151245	
21:13,000,000 21q11.2	TAM, MST	Myeloproliferative syndrome, transient (transient abnormal myelopoiesis)	159595	Leukemia, transient, of Down syndrome	159595		2		
21:14,961,234 21q11.2-q21.1	NRIP1, RIP140, CAKUT3	Nuclear receptor interacting protein 1 (receptor interacting protein 140)	602490	?Congenital anomalies of kidney and urinary tract 3	618270	AD	3	mutation identified in 1 CAKUT3 family	Nrip1

Рис. 10. Таблица генной карты

Следующие обозначения можно увидеть в столбце «Phenotype» карты генов:

- «[ ]» обозначают «не заболевание», в основном генетические вариации, которые приводят к явно не относящимся к норме значениям лабораторных тестов (например, дисальбуминемическая эутиреоидальная гипертироксинемия);
- «{ }» указывают на мутации, которые способствуют предрасположенности к многофакторным нарушениям (например, диабет, астма) или к восприимчивости к инфекции (например, малярии);
- «?» перед названием фенотипа указывает на то, что связь между фенотипом и геном является предварительной или условной. Более подробная информация об этой взаимосвязи приведена в поле для комментариев на карте и в записях OMIM по генам и фенотипам.

В столбце «Phenomap key» также содержится информация-ключ, это число от 1 до 4, которое говорит о следующих типах нарушений:

1) нарушение было определено путём картирования гена дикого типа;
2) был картирован сам фенотип заболевания;
3) молекулярная основа нарушения известна;
4) нарушение представляет собой синдром делеции или дупликации хромосом.

Для того чтобы отобразить эту информацию, можно навести курсор на число. Ниже описана структура записи в базе данных OMIM (**рис. 11**).



Рис. 11. Структура представления записи в OMIM

Помимо описательных записей, OMIM.org предоставляет дополнительные уникальные вкладки информации, включая описание клинической картины, фенотипические серии и фенотип-генную карту. Записи могут содержать информацию о клинических проявлениях нарушения или фенотипа, которые можно увидеть, выбрав кнопку «Clinical Synopsis» (**рис. 12**).

Записи могут также содержать так называемые фенотипические группы, табличное представление генетической гетерогенности схожих фенотипов в геноме. Такие серии или группы можно увидеть, выбрав кнопку «Phenotypic Series» (**рис. 13**).

График OMIM PheneGene (кнопка «PheneGene Graphics») отображает отношения между фенотипами, группами родственных фенотипов (фенотипическими сериями) и генами в различных вариантах представления (**рис. 14**).

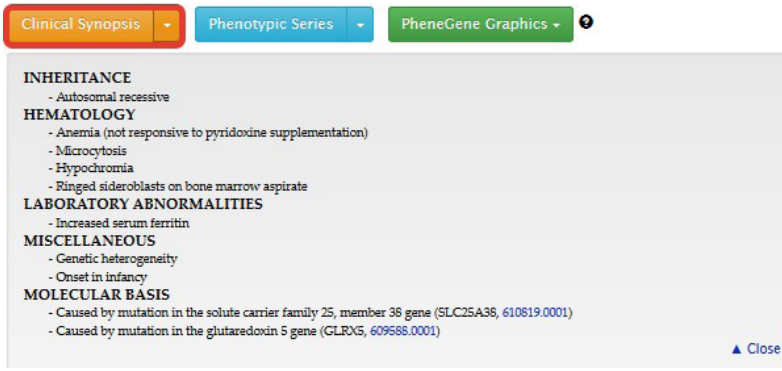


Рис. 12. Пример представления клинической картины и сопутствующих характеристик изучаемого фенотипа

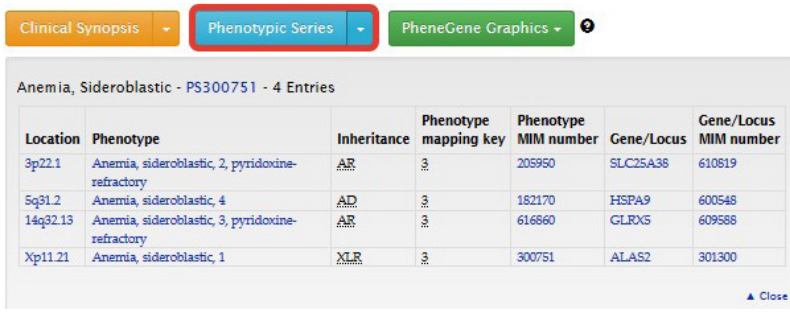


Рис. 13. Описание фенотипических групп



Рис. 14. Выбор варианта графика для представления отношений между фенотипами

Каждая запись имеет также внешние ссылки на другие базы, логически связанные с представленной записью. Блок ссылок расположен в правом верхнем углу (**рис. 15**).

OMIM является важным инструментом, который помогает клиницистам искать записи OMIM, используя клинические признаки, чтобы облегчить диагностику пациентов.

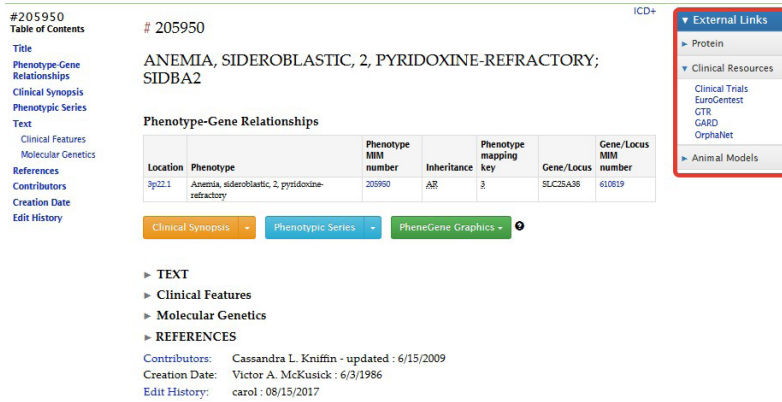


Рис. 15. Выделенный блок ссылок на внешние ресурсы

OMIM связывает клинические особенности фенотипа с молекулярной биологией генов и их вариантов.

Важной особенностью работы с OMIM является возможность использования человеко- и машиночитаемых форматов. Это даёт возможность получать данные из OMIM через механизм Application Programming Interface (не рассматривается в данном курсе).

Для примера можно рассмотреть фенотип заболевания эластоз (cutis laxa). Проведём поиск по этой патологии, используя строку поиска, и в результате получим набор записей, в которых встречается наш запрос. В нашем примере мы нашли 203 записи. Теперь выберем одну из них: #219200 CUTIS LAXA, AUTOSOMAL RECESSIVE, TYPE IIA; ARCL2A. После вывода результатов можно дополнять запрос уточняющими терминами и выполнить поиск вариантов клинической картины данного фенотипа (**рис. 16**).

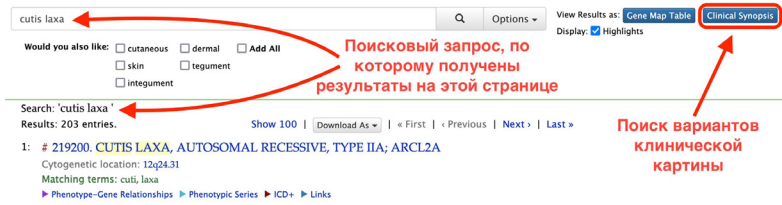


Рис. 16. Вывод результатов запроса «cutis laxa»

Такой поиск возвращает варианты, которые можно рассматривать как отдельные проявления в тех или иных системах органов. Наш результат выдал 153 варианта описания клинической картины. Чтобы быстро просмотреть, чем характерен данный фенотип, можно в строке быстрого просмотра выбирать выпадающие списки и варианты. Для многих вариантов проявлений можно получить более детальную информацию, а также фотографии характерных черт данного фенотипа. Для этого нужно раскрыть одну из ссылок «View full synopsis below» или «View full synopsis on new page» (**рис. 17**).

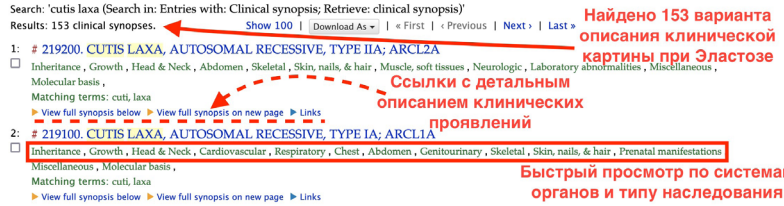


Рис. 17. Вывод поиска вариантов клинической картины

Выбрав категорию в Clinical Synopsis, можно увидеть основные особенности фенотипа. При наведении указателя мыши на иконку человека около соответствующего синдрома можно увидеть фотографии с характерными признаками проявления (**рис. 18**).

MIMmatch

Исследователям, в особенности тем, кто регулярно занимается исследованиями в области генетики человека и интересуется связями генотипов и фенотипов, можно воспользоваться сервисом MIMmatch, ссылка на который доступна в верхнем меню страниц OMIM. Этот сервис в OMIM предназначен для того, чтобы следить за изменениями в интересующих записях и находить других исследователей, которые могут разделять интерес к тем же темам. Для этого нужно зарегистрироваться как пользователь MIMmatch, после чего пользователь будет получать уведомления об обновлениях интересующих генах и заболеваниях.

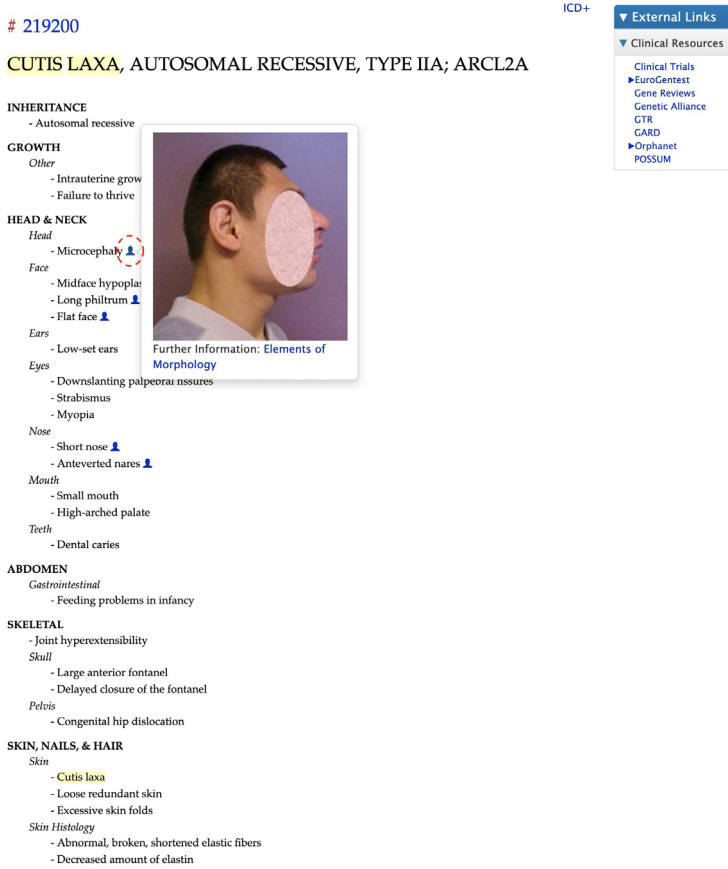


Рис. 18. Пример отображения фотографии характерного фенотипа

Пример поиска в OMIM

В качестве примера изучим заболевание цитруллинемию, проявляющуюся в накоплении в крови продуктов метаболизма аммиака и аминокислоты цитруллина. Введём в поисковое поле название синдрома (citrullinemia) и перейдём к описанию фенотипа заболевания (обозначено знаком #) (**рис. 19**).

Мы видим, что заболевание связано с полиморфизмами в гене *ASS1* (argininosuccinate synthetase — аргининосукцинат синтаза). Обратите внимание на поле Phenotype mapping key — значение 3 означает, что молекулярные основы заболевания известны. В OMIM встречаются следующие значения этого поля:



Рис. 19. Поиск заболевания по запросу «citrullinemia»

- 1 — известна ассоциация между геном и заболеванием, но точный дефект гена, который приводит к заболеванию, не известен;
- 3 — известны молекулярные основы заболевания;
- 2 — известна статистически достоверная связь между геном и заболеванием;
- 4 — дублирование или удаление смежных, прилегающих генов.

Далее на странице представлено подробное описание симптомов (Clinical Features), включая многочисленные ссылки на опубликованные работы. В разделе Biochemical Features описываются биохимические нарушения вследствие мутации аргининосукцинатсинтазы. Особенно интересен раздел Molecular Genetics, где приведены история выявления молекулярных характеристик заболевания и варианты генетических нарушений, их частоты встречаемости в популяции и т.д.

База данных UniProt

Следующим шагом является более подробное изучение фермента аргининсукцинатсинтазы. Для этого предназначена база данных UniProt (URL: http://www.uniprot.org/), интегрирующая множество сведений о белках. UniProt состоит из двух частей: SwissProt и TrEMBL.

SwissProt считается наиболее точной и полно аннотированной базой данных белков и формируется вручную биологами-экспертами.

TrEMBL пополняется автоматически путём компьютерной трансляции всех белок-кодирующих последовательностей, хранящихся в ENA/GenBank/DDBJ, за исключением тех, что уже включены в SwissProt. Аннотация записей TrEMBL также осуществляется в автоматизированном режиме. Таким образом, помимо различия в достоверности (SwissProt достовернее, чем TrEMBL), между этими двумя частями UniProt существует и различие в объёме (TrEMBL существенно больше, чем SwissProt).

Чтобы двигаться дальше и увидеть связь между базами данных, в окне базы данных OMIM с результатами поиска «citrullinemia» откройте соответствующую запись UniProt, перейдя по ссылке на странице OMIM: External Links for Entry → Protein → UniProt или напрямую найдя белок ASS1 человека на главной странице uniprot.org (**рис. 20**).

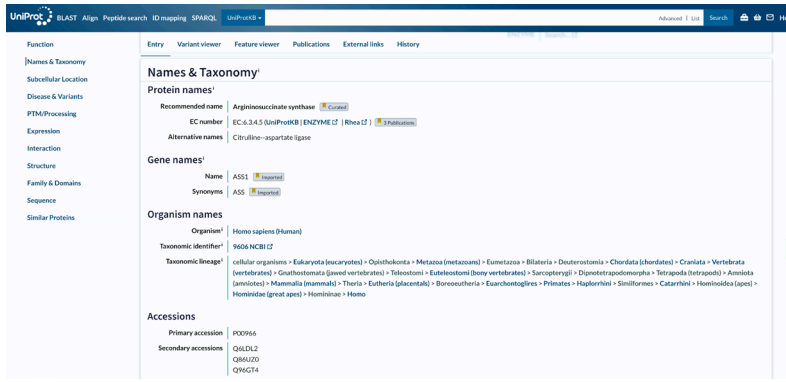


Рис. 20. Результат поиска белка, кодируемого геном *ASS1*

Обратите внимание на разделы навигации слева: Function, Names&Taxonomy и т.д. Они обозначают крупные блоки информации о белке. В соответствующих разделах часто содержатся отсылки на другие, более специализированные, базы данных. Изучите содержимое этих разделах.

В частности, в UniProt содержится ссылка на базу данных KEGG (см. раздел Sequence → Genome annotation databases) (**рис. 21**).

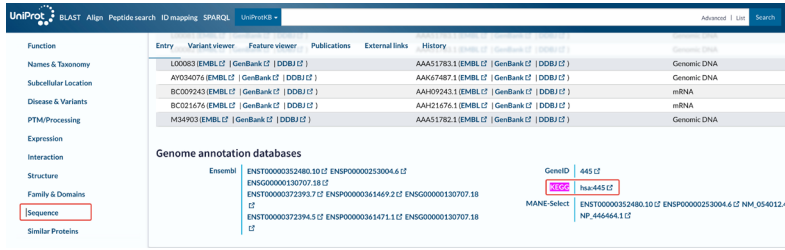


Рис. 21. Связь записи аргининсукцинатсинтазы UniProt с базой данных KEGG

База данных KEGG

Ресурс KEGG (Kyoto Encyclopedia of Genes and Genomes, URL: http://www.genome.jp/kegg/) представляет собой большую коллекцию баз данных, описывающих гены/белки, филогению, геномы, сигнальные и метаболические каскады и реакции, заболевания, лекарства и химические вещества.

Перейдя по ссылке из UniProt, мы попадаем в описание того же белка (ASS1, Homo sapiens), но уже в системе KEGG (**рис. 22**).

В этой базе данных нас в первую очередь интересуют схемы сигнальных и метаболических каскадов (Pathway). Здесь необходимо выбрать метаболический каскад, наиболее точно объясняющий механизм возникновения нарушений. Выбрав интересующий каскад (в нашем примере это hsa0020 (Arginine biosynthesis)), мы увидим интерактивную схему биохимических реакций. Искомый ген отмечен красным шрифтом и обведён красной рамкой. Обратите внимание на то, что часто вместо названий гена используется классификация энзимов; она применяется в каталитических метаболических путях и имеет вид идентификаторов, состоящих из чисел, разделённых точкой. Правила классификации можно узнать на сайте Номенклатурного комитета Международного союза биохимии и молекулярной биологии[2]. На **рис. 23** видно, что фермент катализирует образование аргининсукцината из цитрулина и аспартата. Неудивительно, что повреждение фермента нарушает цикл мочевины, что клинически проявляется в накоплении в крови цитруллина и гипераммониемии.

[2] Международный союз биохимии и молекулярной биологии. URL: https://iubmb.qmul.ac.uk/enzyme/rules.html

Homo sapiens (human): 445

Help

Entry	445 CDS T01001
Symbol	ASS1, ASS, CTLN1
Name	(RefSeq) argininosuccinate synthase 1
KO	K01940 argininosuccinate synthase [EC:6.3.4.5]
Organism	hsa Homo sapiens (human)
Pathway	hsa00220 Arginine biosynthesis hsa00250 Alanine, aspartate and glutamate metabolism hsa01100 Metabolic pathways hsa01230 Biosynthesis of amino acids hsa05418 Fluid shear stress and atherosclerosis
Module	hsa_M00029 Urea cycle hsa_M00844 Arginine biosynthesis, ornithine => arginine
Network	nt06010 Urea cycle
Element	N00593 Urea cycle
Disease	H00185 Citrullinemia H01398 Primary hyperammonemia (Urea cycle disorders)
Brite	KEGG Orthology (KO) [BR:hsa00001] 09100 Metabolism 09105 Amino acid metabolism 00250 Alanine, aspartate and glutamate metabolism 445 (ASS1) 00220 Arginine biosynthesis 445 (ASS1) 09160 Human Diseases 09166 Cardiovascular disease 05418 Fluid shear stress and atherosclerosis 445 (ASS1) 09180 Brite Hierarchies 09183 Protein families: signaling and cellular processes 04147 Exosome [BR:hsa04147] 445 (ASS1) Enzymes [BR:hsa01000] 6. Ligases 6.3 Forming carbon-nitrogen bonds 6.3.4 Other carbon-nitrogen ligases

Рис. 22. Вариант представления записи в базе данных KEGG

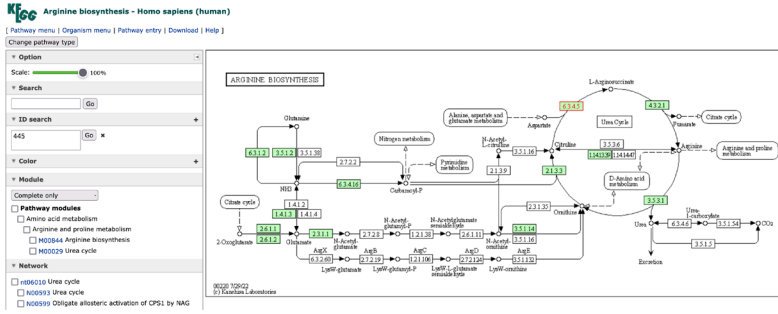


Рис. 23. Вид метаболического пути Arginine biosynthesis, в котором присутствует искомый белок

В описании белка обратите также внимание на раздел Disease, где описываются болезни, в которых задействован ген. Открыв описание болезни, мы найдём отсылку на OMIM (**рис. 24**), совершив тем самым круг — различные биологические базы данных описывают одни и те же сущности с разных позиций.

KEGG **DISEASE: Citrullinemia** Help

Entry	H00185 Disease
Name	Citrullinemia
Subgroup	Argininosuccinate synthase deficiency (CTLN1) Citrin deficiency (CTLN2)
Supergrp	Primary hyperammonemia (Urea cycle disorders) [DS:H01398]
Description	Citrullinemia (CTLN) is an autosomal recessive disease caused by a deficiency of argininosuccinate synthetase/citrin and characterized primarily by elevated serum and urine citrulline levels.
Category	Inherited metabolic disorder
Brite	Human diseases in ICD-11 classification [BR:br08403] 05 Endocrine, nutritional or metabolic diseases Metabolic disorders Inborn errors of metabolism 5C50 Inborn errors of amino acid or other organic acid metabo H00185 Citrullinemia Pathway-based classification of diseases [BR:br08402] Amino acid metabolism nt06010 Urea cycle H00185 Citrullinemia BRITE hierarchy
Pathway	hsa00220 Arginine biosynthesis
Network	nt06010 Urea cycle
Gene	(CTLN1) ASS1 [HSA:445] [KO:K01940] (CTLN2) SLC25A13 [HSA:10165] [KO:K15105]
Drug	Sodium phenylbutyrate [DR:D05868]
Other DBs	ICD-11: 5C50.A3 ICD-10: E72.2 MeSH: D020159 OMIM: 215700 603471 605814
Reference	PMID:18473344 (ASS1)
Authors	Berning C, Bieger I, Pauli S, Vermeulen T, Vogl T, Rummel T, Hohne W, Koch HG, Rolinski B, Gempel K, Haberle J
Title	Investigation of citrullinemia type I variants by in vitro expression studies.
Journal	Hum Mutat 29:1222-7 (2008) DOI:10.1002/humu.20784

Рис. 24. Связь KEGG с другими базами данных

Вопросы и практическое задание

1. Какая информация содержится в базе данных OMIM?
2. Какая информация содержится в базе данных UniProt?
3. Какая информация содержится в базе данных KEGG?
4. Объясните понятие «pathway», применяемое в базе данных KEGG.
5. Как связаны приведённые базы данных между собой?

Исходя из полученной информации, создайте отчёт для наследственного заболевания, указанного ниже, содержащий список пунктов (при их наличии в базах данных):

1. Название заболевания (рус. + англ.).
2. Идентификатор заболевания в OMIM.
3. Клинические проявления (2–3 предложения).
4. Дефектный ген/белок (аббревиатура и название).
5. Встречающиеся мутации (желательно наиболее распространённые), 2–3 шт.
6. «Биохимия заболевания» (1–2 предложения).
7. Функция белка согласно UniProt (1–2 предложения).
8. Количество активных сайтов, описание одного из активных сайтов.
9. Количество регуляторных участков, описание одного из регуляторных участков (тип, локализация, длина, последовательность).
10. Количество встречающихся вариаций, описание и клиническая значимость для 1–2.
11. Количество аминокислотных остатков в белке.
12. Первые 10 аминокислот белковой последовательности (в формате FASTA).
13. Взаимодействие с другими белками.
14. Ссылки на публикации, описывающие белок (PubMed ID, 2–3 шт).
15. Классификация фермента по системе EC (если есть).
16. Вставьте схему (или фрагмент) метаболического пути. Опишите участие фермента в катализируемой реакции или взаимодействие с другими белками, а также возможную связь с клиническими проявлениями.

Некоторые синдромы для поиска в OMIM для практического задания:

1. Li-Fraumenisyndrome 1.
2. Nijmegen Breakage Syndrome.
3. Bloom syndrome.
4. Glutamine deficiency, congenital.
5. Donohue syndrome.
6. Phenylketonuria.
7. Lipodystrophy, familial partial, type 3.
8. Haim–Munk syndrome.

9. Krabbe disease.
10. Breast-ovarian cancer, familial, susceptibility to, 1.

Полезный видеоресурс:

Генетические заболевания и OMIM.
URL: https://www.youtube.com/@omim9697

Геномный браузер UCSC

Геномный браузер UCSC (University of California, Santa Cruz; URL: https://genome.ucsc.edu/index.html) является одним из самых популярных в мире геномных браузеров для просмотра и анализа данных, связанных с геномами различных организмов, в том числе человека (**рис. 25**).

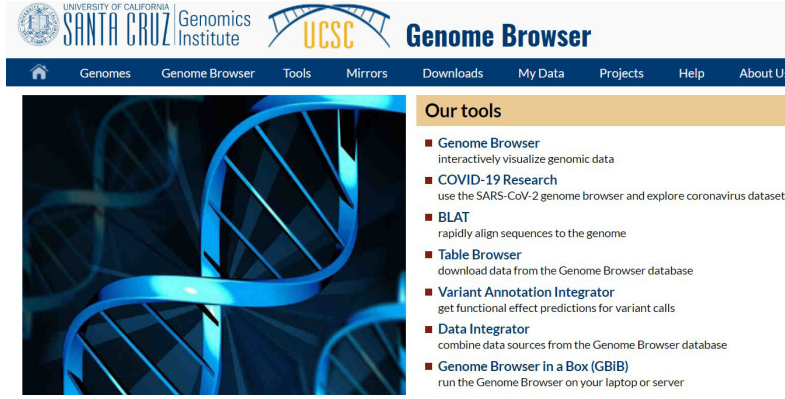


Рис. 25. Внешний вид ресурса UCSC

Цель занятия — познакомиться с интерфейсом и базовыми навыками работы с геномным браузером UCSC и одним из полезных инструментов для извлечения данных — Table Browser.

Геномный браузер UCSC представляет возможность изучать геномы различных организмов на основе их референсных сборок. Для многих биологических видов доступны несколько версий референсного генома, это связано с тем, что при улучшении геномных сборок происходит изменение геномных координат, что может влиять на результаты исследований. Для примера, если изучать геном человека, то в UCSC доступны 5 версий сборки. Наиболее часто используют две последние: GRCh37/hg19 (2009 г.) и GRCh38/hg38 (2013 г.).

Особенностью геномных браузеров является то, что вся информация группируется в специальных визуальных объектах, которые называются треками. Перейдём по ссылке Genome Browser, затем выберем организм, а также версию генома, которая нам нужна, и нажмём на кнопку GO (**рис. 26**).

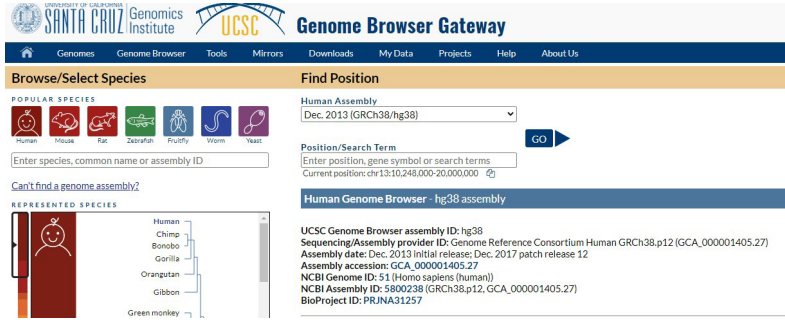


Рис. 26. Интерфейс выбора объектов интереса

Появляется страница, на которой по умолчанию видны лишь несколько треков (горизонтальных линий, в которых содержится информация определённого типа), а внизу представлен сгруппированный по смыслу список треков (на **рис. 27** представлена верхняя часть первой группы треков (Mapping and Sequencing).

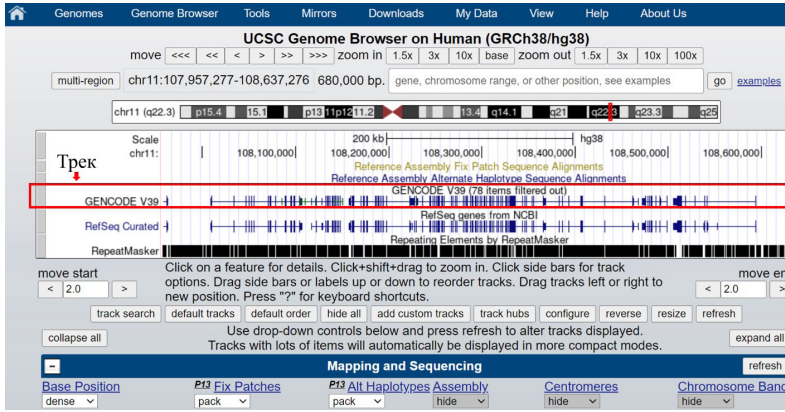


Рис. 27. Внешний вид геномного браузера с выделенным треком GENCODE

Ознакомьтесь с тем, какие треки представлены в этом сгруппированном списке. Каждый трек можно сконфигурировать в зависимости от того, насколько детальную информацию вы хотите увидеть. Для этого необходимо навести курсор мыши на серый прямоугольник, с которого начинается трек, и кликнуть на нём правой кнопкой мыши. Появится окно с параметрами визуализации этого трека (**рис. 28**).

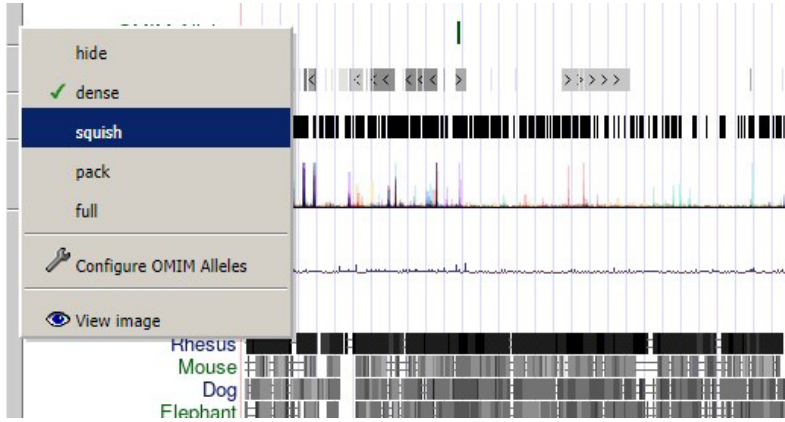
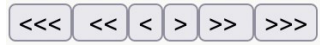


Рис. 28. Выбор варианта конфигурирования трека

Выберите разные характеристики одного из треков, чтобы понять, насколько при этом различается детализация данных.

В верхней части окна UCSC браузера есть группы кнопок:

- кнопки навигации вдоль хромосомы



,

- приближения и удаления



Здесь же можно увидеть, сколько нуклеотидов представлено в окне браузера – от единиц до миллионов), а также поле, в котором можно задать точные координаты области хромосомы, в которой вы хотите проводить анализ или визуализировать данные. Красной вертикальной чертой на идиограмме показана часть хромосомы, которая отображена в браузере (**рис. 29**).

Рассмотрим возможности UCSC Genome Browser и Table Browser при решении конкретных задач.



Рис. 29. Верхняя часть окна геномного браузера

Задача 1

Гемоглобин представляет собой тетрамер, состоящий в основном из двух субъединиц альфа-глобина и двух субъединиц бета-глобина. Рассмотрим альфа-глобин.

Существуют два родственных гена человека: *HBA1* и *HBA2*. Используйте UCSC Genome Browser, чтобы определить длину межгенной области, разделяющей эти гены.

Визуально можно получить примерный ответ, введя «HBA1» в UCSC Genome Browser (**рис. 30**). Уменьшите масштаб в 10 раз, и вы увидите соседние гены *HBA1* и *HBA2* (**рис. 31**). Зажав левую кнопку мыши в поле трека и двигая в стороны, можно перемещаться вдоль хромосомы.

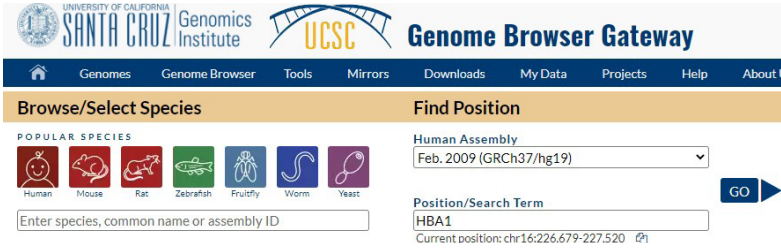


Рис. 30. Выбор гена *HBA1* и варианта геномной сборки hg19

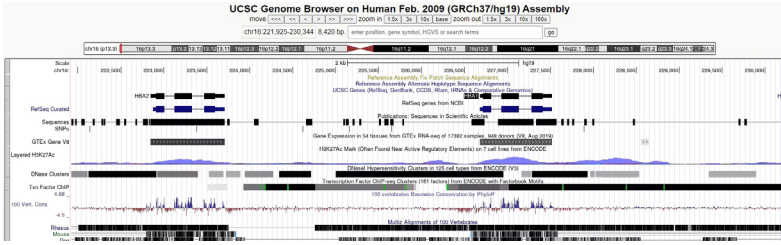


Рис. 31. Область интереса, содержащая оба интересующих нас гена

Для примерного определения расстояния между генами можно выделить прямоугольник между генами (удерживая курсор чуть ниже идеограммы, в области трека — Scale). Так

можно узнать, что расстояние между *HBA1* и *HBA2* составляет примерно 2900 пар оснований (**рис. 32**).

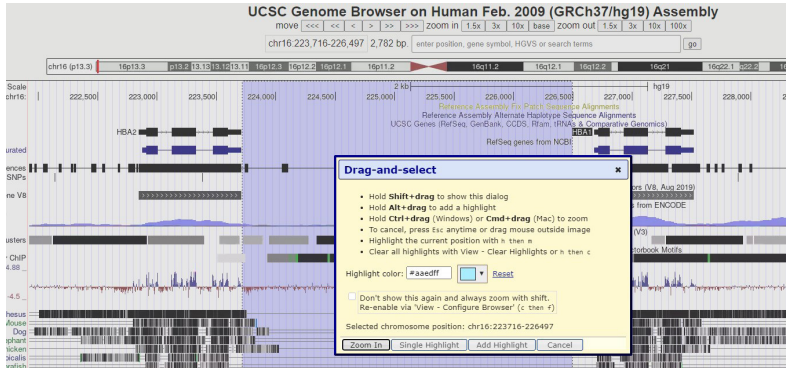


Рис. 32. Выделение области между генами с помощью мыши

Чтобы получить более точный ответ, воспользуйтесь Table Browser в меню Tools (**рис. 33**).

Задайте область, охватывающую оба гена, и настройте различные меню, как показано на **рис. 34**.

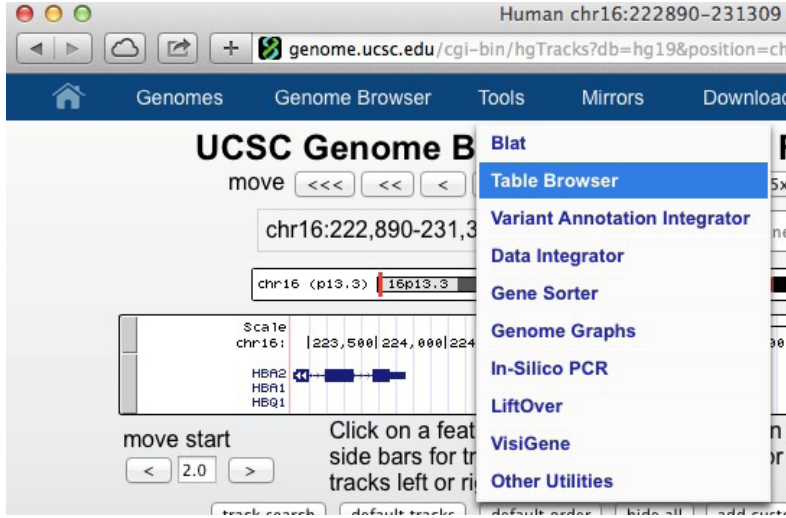


Рис. 33. Выбор инструмента Table Browser.

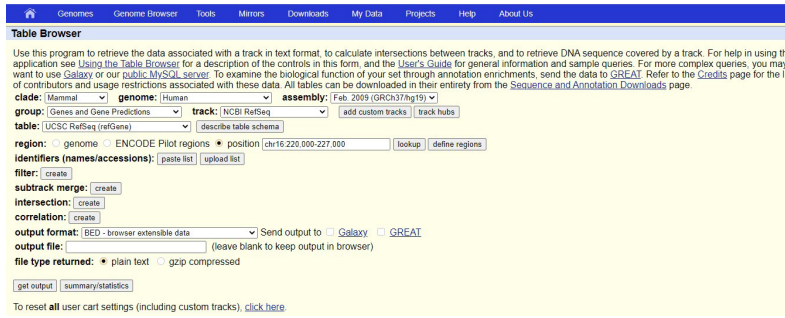


Рис. 34. Настройки представления результатов в виде таблицы

Получите следующий результат:

chr16 222874 223709 NM_000517 0 + 222911 223599 0 3 132 205 239, 0,249,596,

chr16 226678 227521 NM_000558 0 + 226715 227410 0 3 249 603, 132 205 240.

Это результат (в виде файла BED — Browser Extensible Data) — текстовый формат файла, который используется для хранения геномных областей в виде координат и связанных с ними аннотаций. Из этих выходных данных можно рассчитать точное межгенное расстояние.

Например, возьмите начало одного гена с 3'-конца (позиция 226715) и вычтите конец другого гена с 5'-конца (позиция 223599), получим 3116 пар оснований.

В качестве самостоятельного задания проделайте то же для сборки генома 2013 года (hg38). Отличаются ли полученные значения?

Задача 2

Какие типы повторяющихся элементов ДНК встречаются в гене бета-глобина человека? Цель этого упражнения — познакомиться с геномным браузером UCSC. Как пользователь вы выбираете, какие треки будут отображаться. Посетите и исследуйте как можно больше треков. Постарайтесь понять основные категории информации, предлагаемой в геномном браузере UCSC. Работая в геномном браузере, вы можете переключаться между сборками GRCh37 и GRCh38. Для этого перейдите в выпадающее меню «View» и используйте команду «In other genomes (convert)». Выполните следующие шаги:

1. Перейдите на Genome Browser Gateway (URL: http://genome.ucsc.edu/cgi-bin/hgGateway). Убедитесь, что выбран геном — человек, сборка — NCBI37/hg19, и в поле «ген» введите HBB для бета-глобина. Нажмите GO. Обратите внимание, что HBB является официальным символом гена для бета-глобина, но вы можете использовать строчный hbb для этого поиска. Используйте NCBI Gene (или URL: http://www.genenames.org для сайта HGNC), чтобы найти официальный символ любого вашего гена.

2. Нажмите кнопку «default tracks». Обратите внимание на позицию, которую вы достигли (хромосома 11, охватывающая 1606 пар оснований вблизи начала короткого или «p» плеча хромосомы). Обратите внимание на появление более десятка графических треков, которые ориентированы горизонтально (**рис. 35**).

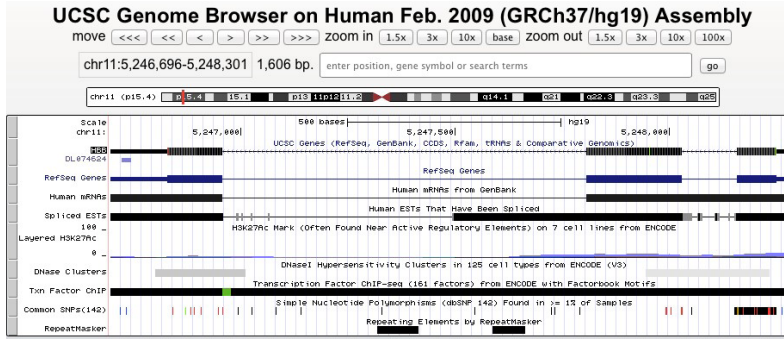


Рис. 35. Вид треков по умолчанию

3. В одном из треков — «Repeating Elements by RepeatMasker» есть 2 чёрных блока. Щёлкните правой кнопкой мыши блок и выберите «Full» (**рис. 36**). В качестве альтернативы прокрутите вниз до раздела «Variation and Repeats», найдите «RepeatMasker» и измените настройку раскрывающегося меню с «dense» на «full». Обратите внимание, что, щёлкнув синий заголовок «RepeatMasker», вы переходите на страницу, описывающую программу «RepeatMasker» и её использование в геномном браузере UCSC.

4. Просмотрите вывод RepeatMasker. Выберите один ответ:

(а) Нет повторяющихся элементов.

(б) Существует один элемент SINE (Short interspersed nuclear element — короткие диспергированные повторы) и один

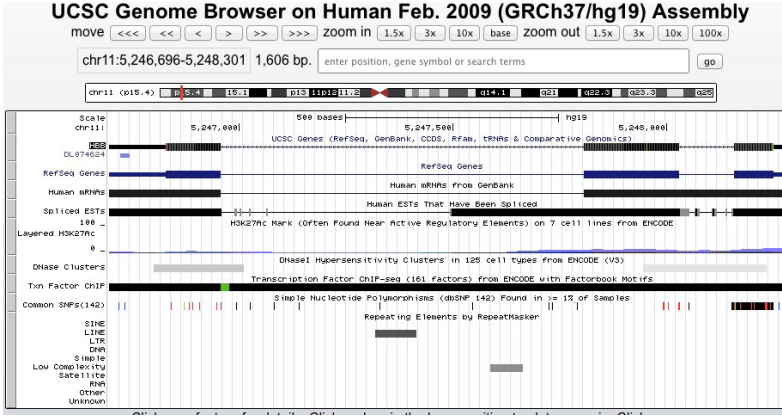


Рис. 36. Вид с развёрнутым треком RepeatMasker (нижний трек)

элемент LINE (Long interspersed nuclear element, длинные диспергированные повторы). SINEs — последовательности ДНК длиной менее 500 пар оснований в геноме эукариот, появившиеся в результате обратной транскрипции коротких молекул РНК (5S рРНК, тРНК и различные мяРНК). SINEs не кодируют белки, и их транспозиция в геноме зависит от других мобильных элементов. LINEs —последовательности ДНК длиной по несколько тысяч пар оснований в геноме эукариот, представляющие собой ретротранспозоны, не содержащие длинных концевых повторов.

(в) Один LTR (Long terminal repeats — длинные концевые повторы) — последовательности ДНК, повторённые сотни или тысячи раз. Длинные концевые повторы фланкируют гены и обнаружены в геномной ДНК ретровирусов и в ретротранспозонах. Вирусы используют LTR для встраивания собственного генома в геном хозяина и один спутник.

(г) Существует один элемент LINE и один элемент низкой сложности.

(д) Есть более десятка повторяющихся элементов.

Убедитесь, что правильным ответом является вариант (г).

Задача 3

Сколько SNP известно для человеческого гена бета-глобина? Чтобы решить эту проблему, используйте UCSC Table Browser, который, вместо визуального, предлагает табличный вывод. Часто бывает нецелесообразно (или невозможно точно)

визуально определять количество элементов в браузере. Также бывает необходимо получить точную количественную информацию о геномных особенностях в некоторых хромосомных областях или по всему геному. Эта проблема касается и SNP – геномных позиций, которые заняты разными нуклеотидами у разных людей в популяции. Чтобы её решить, выполните следующие шаги.

1. Начните с области HBB в UCSC Genome Browser и выберите вкладку Table Browser вверху в разделе Tools или перейдите на веб-сайт UCSC (URL: http://genome.ucsc.edu) и нажмите Table Browser. Установите clade «Mammal», genome «Human», сборку (assembly) «GRCh37/hg19», группу «Variation», track «AllSNPs (142)», таблицу «snp142» и область положения «chr11: 5246696-5248301». Обратите внимание, что, если позиция ещё не установлена, вы можете ввести hbb в поле позиции, нажать «define regions» и увидеть правильную позицию.

2. Чтобы увидеть ответ, нажмите «summary/statistics». Количество элементов говорит вам, сколько существует SNP. Сколько всего SNP было обнаружено при таком поиске? Чтобы увидеть ответ в виде таблицы, установите выходной формат на «все поля из выбранной таблицы», убедитесь, что флажки «Send output to Galaxy/GREAT» не отмечены, и нажмите «get output». SNP показаны в виде таблицы, включающей хромосому, стартовую и конечную позиции. В группе «Variation» доступные на данный момент треки включают Common SNPs(142) (**рис. 37**).

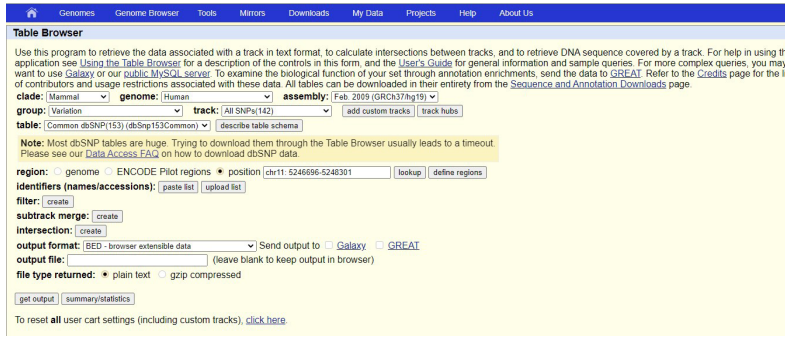


Рис. 37. Настройки табличного представления (в группе «Variation» значение Common SNPs(142))

3. Для Common SNP(142) нажмите на «summary/statistics», это покажет, что в локусе гена *HBB* имеются 6 известных распространённых SNP.

4. Со временем базы данных обновляются. Проверьте, сколько All SNP (все известные SNP из БД NCBI SNP) и Common SNP (варианты с частотой минорного аллеля (MAF) не менее 1% (0,01) в наборе данных проекта 1000 Genomes) найдётся, если вы выберите базу данных SNP версии 153. Сравните с количеством, найденным в dbSNP 142.

Полезный видеоресурс:

С более детальными возможностями UCSC браузера вы можете познакомиться с помощью обучающих видео:
URL: http://genome-euro.ucsc.edu/training/index.html

Заключение

В результате изучения данного учебно-методического пособия студенты узнают об основных мировых биоинформационных ресурсах, связанных с нуклеотидными и аминокислотными последовательностями, хранящими данные о связи генетических вариантов и заболеваний человека. Студенты изучат основную терминологию и некоторые форматы данных, используемые при работе с этими ресурсами, познакомятся с интерфейсом и принципами работы в геномном браузере UCSC и инструментом для автоматического извлечения геномных данных Table Browser.

Список литературы

Основная

1. Леск А. Введение в биоинформатику. М.: БИНОМ. Лаборатория знаний, 2015. 318 с.

Дополнительная

1. Amberger J.S., Bocchini C.A., Scott A.F., Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships // Nucleic Acids Res. 2019. Vol. 47, Database issue. P. D1038–D1043. DOI: 10.1093/nar/gky1151
2. Kanehisa M., Goto S., Hattori M. et al. From genomics to chemical genomics: new developments in KEGG // Nucleic Acids Res. 2006. Vol. 34, Database issue. P. D354-D357. DOI: 10.1093/nar/gkj102
3. Lee B.T., Barber G.P., Benet-Pagès A. et al. The UCSC Genome Browser database: 2022 update // Nucleic Acids Res. 2022. Vol. 50, Database issue. P. D1115–D1122. DOI: 10.1093/nar/gkab959
4. Pruitt K.D., Tatusova T., Maglott D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins // Nucleic Acids Res. 2005. Vol. 33, Database issue. P. D501–D504. DOI: 10.1093/nar/gki025
5. Sayers E.W., Bolton E.E., Brister J.R. et al. Database resources of the national center for biotechnology information // Nucleic Acids Res. 2022. Vol. 50, Database issue. P. D20–D26. DOI: 10.1093/nar/gkab1112
6. Schneider M., Tognolli M., Bairoch A. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools // Plant Physiol. Biochem. 2004. Vol. 42, N 12. P. 013–021. DOI: 10.1016/j.plaphy.2004.10.009
7. Schoch C.L., Ciufo S., Domrachev M. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools // Database (Oxford). 2020. baaa062. DOI: 10.1093/database/baaa062

Для заметок

Для заметок

Учебное издание

Общедоступные ресурсы биоинформатики: биологические базы данных, геномный браузер UCSC

Учебно-методическое пособие

Составители:
Шилов Борис Владимирович
Лагунин Алексей Александрович

Редактор О.В. Устинкова
Корректор Е.Н. Королёва

Подписано в печать 07.12.2023. Формат 60 × $90^{1}/_{16}$.
Объём 3,5 п.л. Заказ № 32-23. Тираж 350 экз.

Отпечатано в ФГАОУ ВО РНИМУ им. Н.И. Пирогова
Минздрава России
117997, г. Москва, ул. Островитянова, д. 1. www.rsmu.ru

ISBN 978-5-88458-660-4

