

Гены и аннотация вариантов в генах

Василий Евгеньевич Раменский

Анастасия Александровна Жарикова и Мария Ильинична Зайченко

НМИЦ Терапии и профилактической медицины

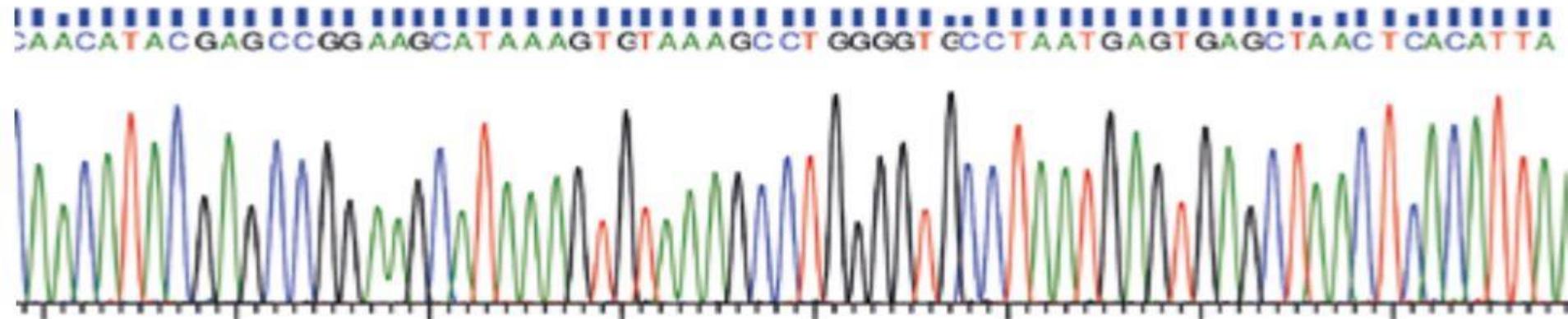
Факультет биоинженерии и биоинформатики МГУ

Институт искусственного интеллекта МГУ

2024

Секвенирование

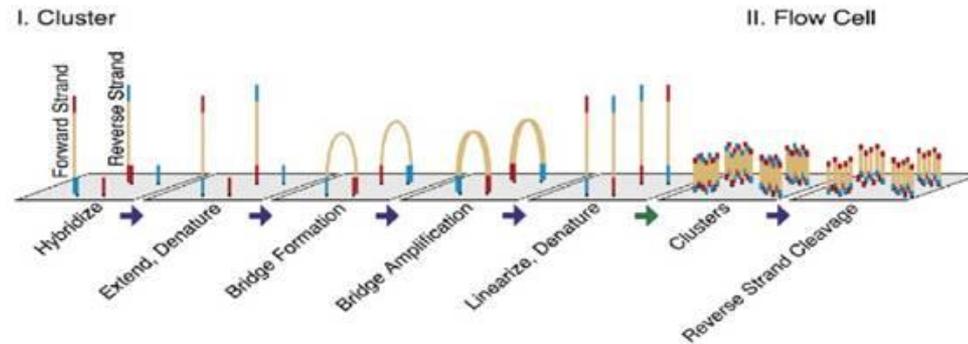
- Определение последовательности некоторого нерегулярного биологического гетерополимера – белка или нуклеиновой кислоты
- Про белки говорить не будем
- Про РНК тоже не будем



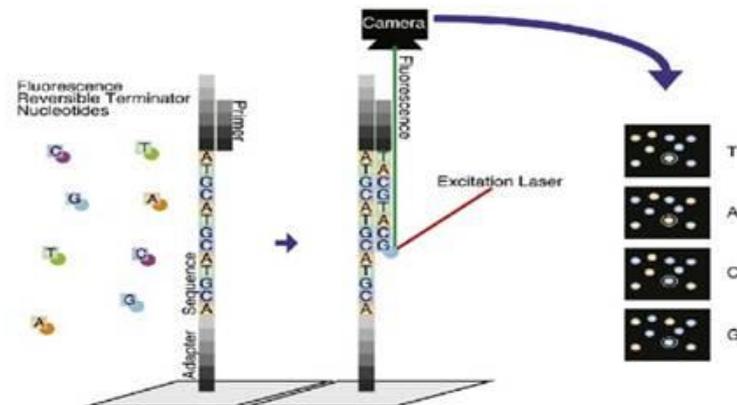
Next-generation sequencing (NGS) - Illumina

Вспомним, как работает секвенатор компании Illumina

A. Clustering

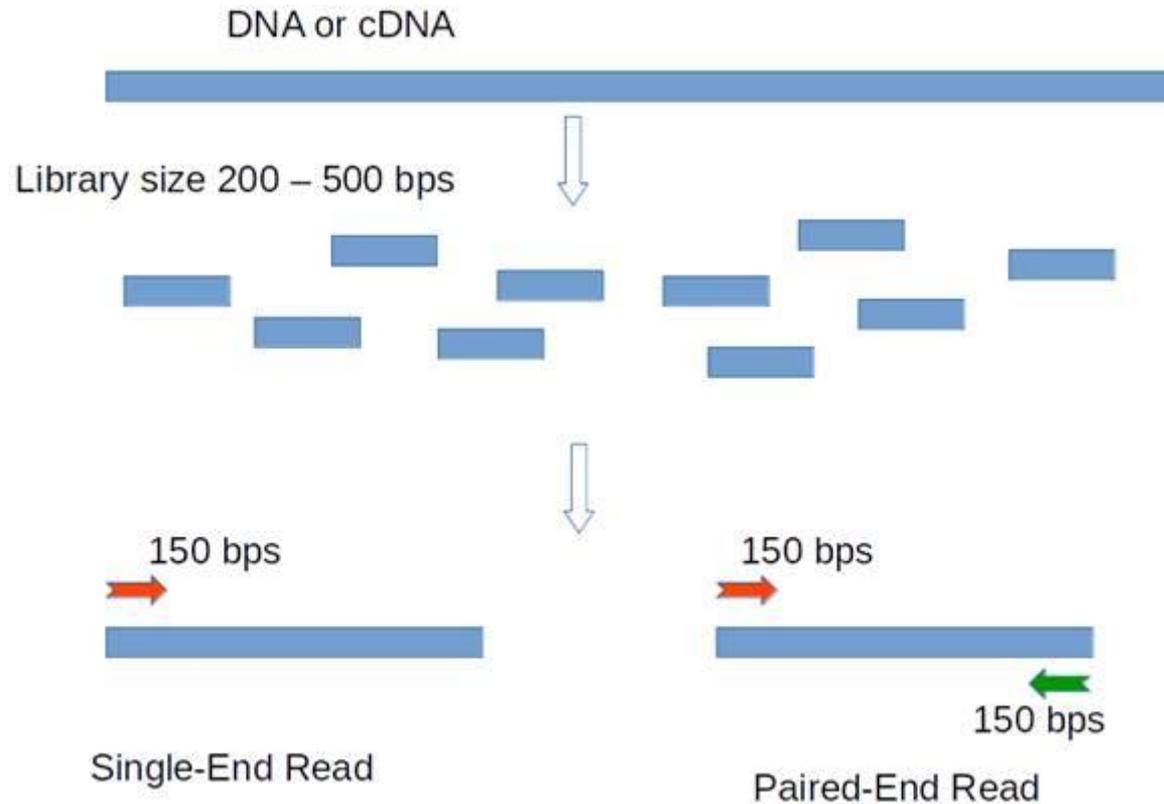


B. High-throughput sequencing



Next-generation sequencing (NGS) - Illumina

Вспомним, что чтения бываю парноконцевыми и одноконцевыми



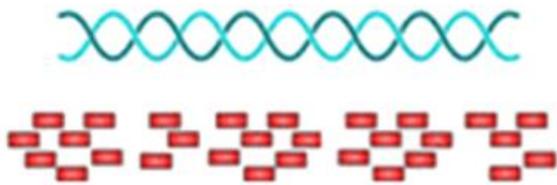
Секвенирование ДНК

Стратегии секвенирования

- **Полный геном**
- **Экзом** – экзоны белок-кодирующих генов
- **Панели** – набор генов (и\или локусы), варианты в которых интересны при проведении какого-либо исследования или диагностики
- *Вопрос: перечислите плюсы и минусы каждого подхода*

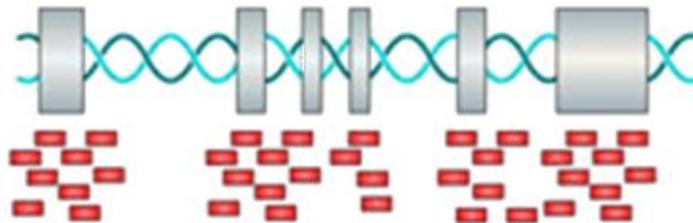
Области секвенирования

Whole genome sequencing



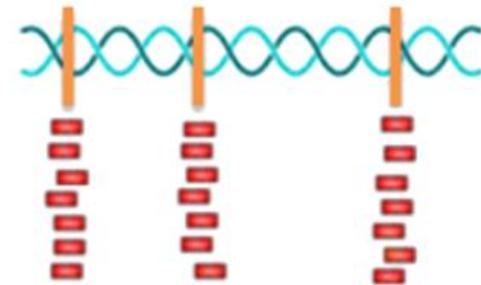
- Sequencing region : whole genome
- Sequencing Depth : >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

Whole exome sequencing



- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

Targeted sequencing



- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

Какие варианты бывают

- **SNV** - однонуклеотидные варианты, т.е. изменение одного нуклеотида
- **Indels** - короткие вставки и делеции (~ 50 п.н.)
- **CNV** - структурные варианты: инверсии и транслокации
- **Анеуплоидии**: нульсомии, моносомии, трисомии, полисомии
- **Полиплоидизация**

Задача семинара

- Проаннотировать набор вариантов в нескольких генах человека с помощью веб-сервиса VEP
- Выстроить систему приоритизации вариантов согласно набору критериев

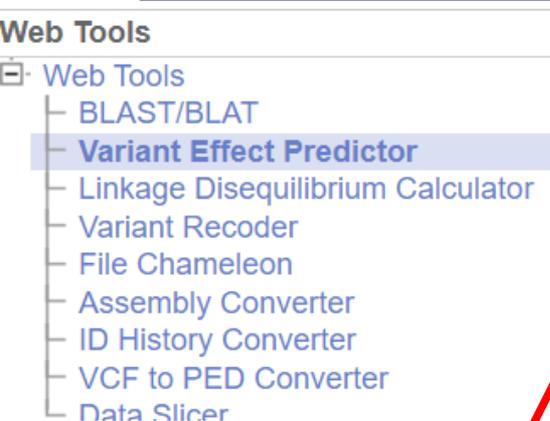
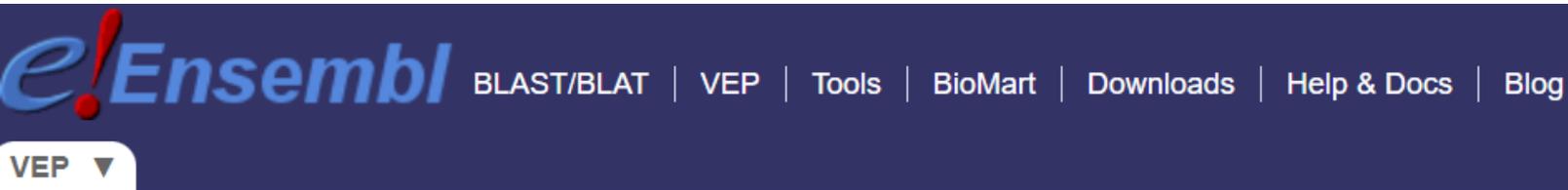
Дано

- Набор вариантов в нескольких генах человека
 - Файл с вариантами в формате VCF

Variant Effect Predictor



- [VEP](#)
- На вход можно подать vcf файл с вариантами



Variant Effect Predictor ?

New job

Recent jobs

Refresh



VEP



New job

Species:

 Homo_sapiens X

Assembly: GRCh38.p14

[Change species](#)

Name for this job (optional):

Input data:

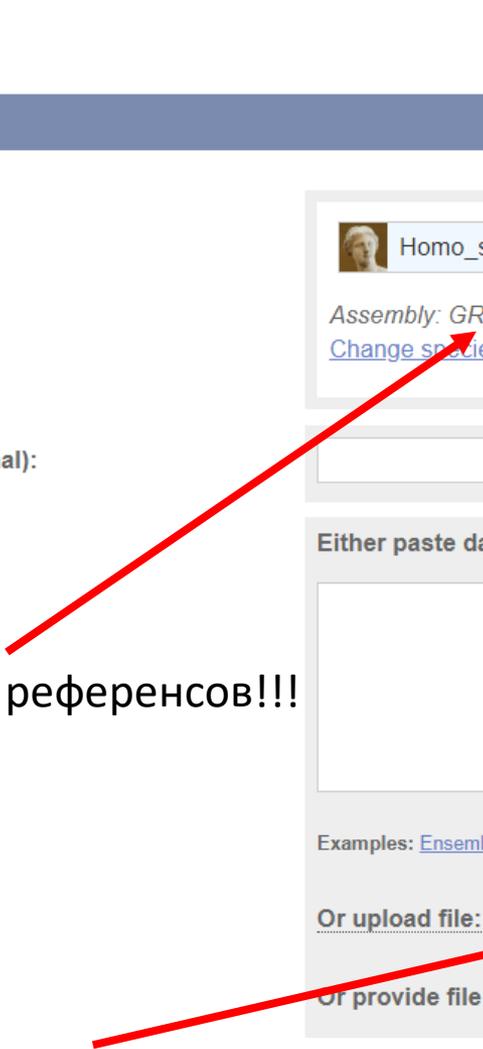
Either paste data:

Examples: [Ensembl default](#), [VCF](#), [HGVS notations](#), [SPDI](#)

Or upload file:

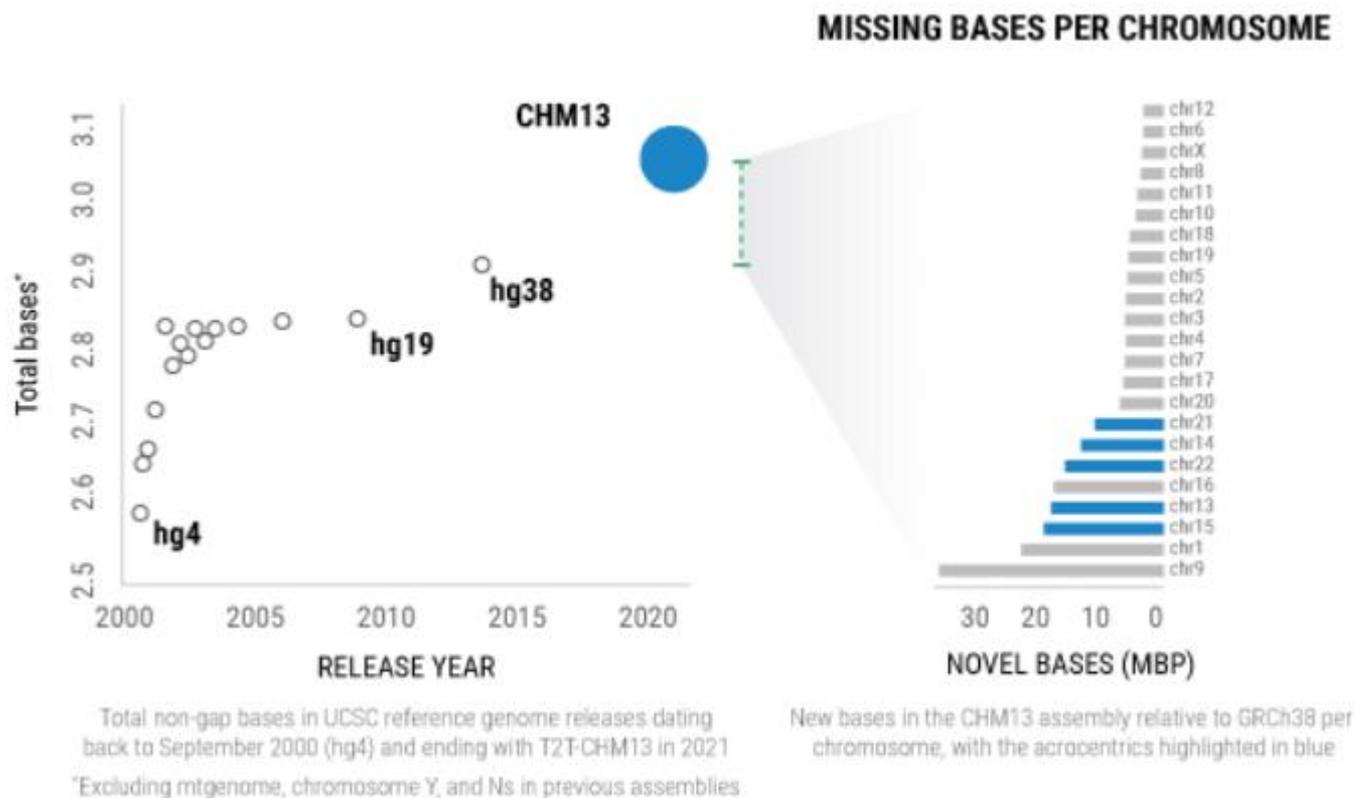
Файл не выбран

Or provide file URL:



Проверяем версии референсов!!!

Референсный геном



Упражнение

- Загрузите в VEP тестовый файл *single_sample_het_hom.vcf*
- Ниже есть настройки, добавьте:
 - HGVS
 - UniProt
 - Exon and intron numbers
 - gnomAD exomes
 - gnomAD genomes
 - MANE
 - Identify canonical transcripts

VEP



- При загрузке файла создается новый процесс

A screenshot of the VEP web interface. At the top, there are two tabs: 'Analysis' and 'Jobs'. The 'Jobs' tab is active. Below the tabs, there is a table with one row. The first column contains the text 'Variant Effect Predictor'. The second column contains a small circular icon with a person's face, followed by the text 'VEP analysis of doom_1_sample_het_hom.vcf in Homo_sapiens'. The third column contains a grey button with the text 'Queued'.

- Аннотация занимает какое-то время
- Дождитесь статуса **Done**

A screenshot of the VEP web interface, similar to the one above. The 'Jobs' tab is active. The table now shows the job as completed. The status button is green and contains the text 'Done'. To the right of the 'Done' button is a blue link with the text '[View results]'. A red arrow points from the bottom right of the slide towards the '[View results]' link.

Пока ждем...

... вспомним основные файлы в анализе данных
высокопроизводительного секвенирования

Протокол



Все файлы храним и анализируем в архивированном виде!

FASTQ

```
1 @NB501222:13:HY55HBGXY:1:11101:26102:3380 1:N:0:ATGTCA
2 CGTTGGAGAAATAAAATGTGCATAGTGGGGATTTTATTTTAAGTTTGTTGGTTAGGTAGTTGAGGTCTAGGGTTG
3 +
4 AAAAAEEEEEE6EEEE6EE/EEAEE6/E//EE//EEE//EEE///EEEEAEeeeeeeEA/A//EEE//EAEEA///A
```

Для каждого чтения выделено 4 строки:

- 1 – идентификатор чтения
- 2 – нуклеотидная последовательность чтения
- 3 – строка идентификатора показателя качества (обычно только «+»)
- 4 – качество каждого нуклеотида

[Подробнее](#)

SAM / BAM

```

NB501222:13:HY55HBGX:1:11101:16088:1242      272      14      49586777      1      75M      *
      0      0      GAAACGGAGCAGGTCAAACCTCCCGTGCTGATCAGTAGTGGGATCGCGCCTGTGAATAGCCACTGCACTCCAGCC
      EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<EAE/EEEEEEEEEEEEEEEEEEEE6EEEEEEEA
      AS:i:0      ZS:i:0
      XN:i:0      XM:i:0      XO:i:0      XG:i:0      NM:i:0      MD:Z:75      YT:Z:UU      XS:A:+      NH:i:2
  
```

SAM

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0, 2 ¹⁶ - 1]	bitwise FLAG
3	RNAME	String	* [:rname:^*]=[:rname:]*	Reference sequence NAME ¹¹
4	POS	Int	[0, 2 ³¹ - 1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2 ⁸ - 1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [:rname:^*]=[:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	[0, 2 ³¹ - 1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ + 1, 2 ³¹ - 1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Из особо важного:

- ID чтения
- координаты места картирования чтения на референсный геном
- схема картирования (CIGAR)
- качество картирования
- последовательность в нуклеотидах (аналогично 2ая строка fastq)
- качество нуклеотидов в чтении (аналогично 4ая строка fastq)
- различные флаги (парность чтения, факт картирования, дубликат и пр.)
- различные тэги (количество ошибок, количество мест картирования чтения и др.)

VCF

Спецификация

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4
```

Для работы на семинаре

- В качестве примера будем использовать файл *single_sample_het_hom.vcf*
- Откройте его любым способом (кроме excel) и рассмотрите из чего этот файл состоит
- Три основные части:
 - Шапка (строки начинаются с ##)
 - Строка (одна) с заголовками столбцов (начинается с #)
 - Информация о вариантах

VCF – столбцы

8 фиксированных [колонок](#) ([еще](#))

- CHROM – имя хромосомы
- POS – позиция варианта
- ID – может быть любая информация о варианте, но обычно пустой (.)
- REF – референсная аллель
- ALT – альтернативная аллель
- QUAL – качество варианта (Phred-scaled)
- FILTER – PASS (если ранее была осуществлена маркировка по каким-либо показателям: покрытие, качество и пр)
- INFO – различные характеристики варианта
- FORMAT – список параметров варианта для конкретного образца
- HG00096 – значения параметров, указанных в столбце FORMAT для конкретного образца (в заголовке – ID образца)

VCF – метрики образца

- Колонка FORMAT: **GT:AD:DP:GQ:PL**
- Колонка ID образца: **0/1:21,4:25:99:1220,108,0**

VCF – FORMAT - GT

- Кодировывает генотип варианта
- Для диплоидных организмов:
 - 0 – референсный аллель
 - 1 – альтернативный аллель
- Образец по варианту:
 - 0/0 – референсная гомозигота
 - 0/1 – гетерозигота
 - 1/1 – альтернативная гомозигота

VCF – FORMAT – AD и DP

- Отражает покрытие
- AD – количество чтений, которые поддерживают каждую из возможных аллелей; участвуют все чтения, использованные при поиске вариантов
- DP – общее количество чтений, прошедших фильтрацию и поддерживающих каждую из представленных аллелей

VCF – FORMAT – PL и GQ

- Отражает качество генотипа
- PL – нормализованные «вероятности» возможных генотипов (по шкале Phred). Поле содержит 3 числа, что соответствует генотипам 0/0, 0/1, 1/1. PL наиболее вероятного генотипа = 0
- GQ – вычисляется на основании PL, представляет собой разницу «вероятностей» двух наиболее вероятных генотипов (но не более 99). Низкие значения (т.е. << 99) – в генотипе нет уверенности

Упражнение

- Расшифруйте записи

FORMAT	SAMPLE_ID
GT:AD:DP:GQ:PL	0/1:18,15:33:99:393,0,480
GT:AD:DP:GQ:PL	0/1:1,4:6:20:73,0,20
GT:AD:DP:GQ:PL	1/1:0,30:30:89:913,89,0
GT:AD:DP:GQ:PL	0/1:18,17:35:99:660,0,704

- Найдите в тестовом файле генотипы 0/0

Multiple VCF

- В одном VCF файле может быть представлена информация сразу о нескольких образцах
- В конце будут добавлены столбцы на каждый образец
- QUAL – максимальный из возможных

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

VEP



- **Job details** – отображает все настройки и предоставляет команду для аналогичного анализа на вычислительном кластере

Разнообразие данных

- Загружено 34350 вариантов
- Каждый проаннотирован каким-то образом
- Наша глобальная задача – дать человеку медицинское заключение на основании проведенного генетического исследования
- Нужно ли просматривать 34350 вариантов?

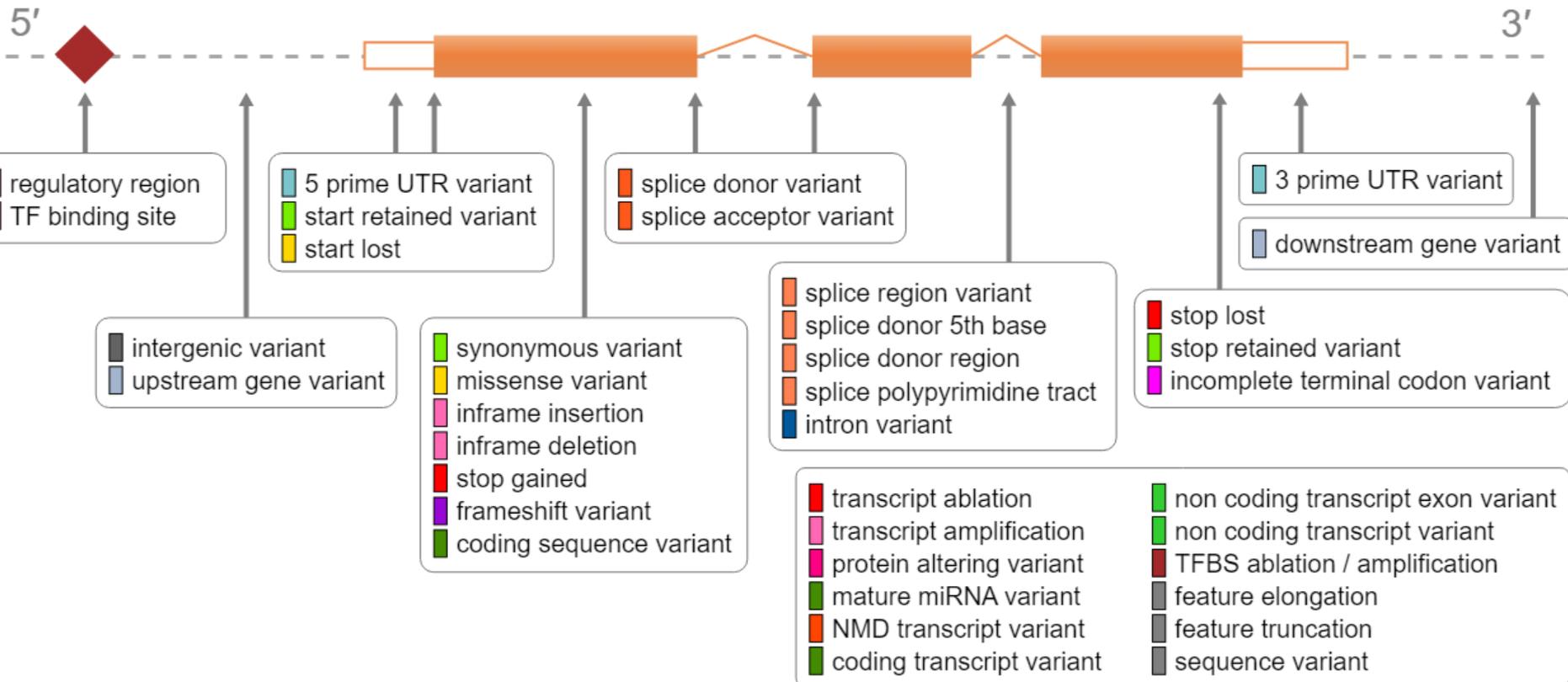
Фильтрация

- Техническая
 - До аннотации можно удалить варианты
 - С низким покрытием
 - С низким качеством
 - ...
- Смысловая
 - Это самое интересное
 - Предложите 5 вариантов приоритизации вариантов

VEP



• Consequences



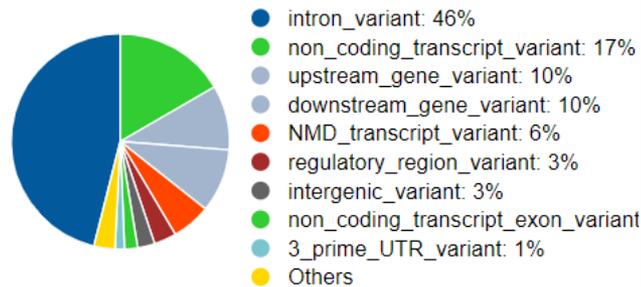
VEP



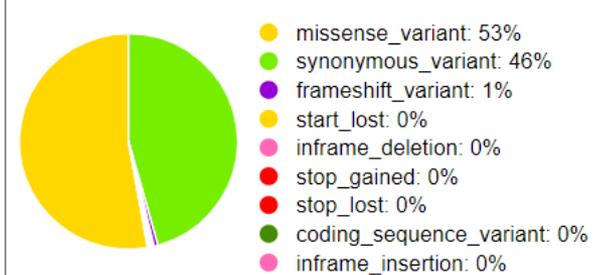
- Summary statistics

Category	Count
Variants processed	34350
Variants filtered out	0
Novel / existing variants	6 (0.0) / 34344 (100.0)
Overlapped genes	9179
Overlapped transcripts	47336
Overlapped regulatory features	4771

Consequences (all)



Coding consequences



VEP



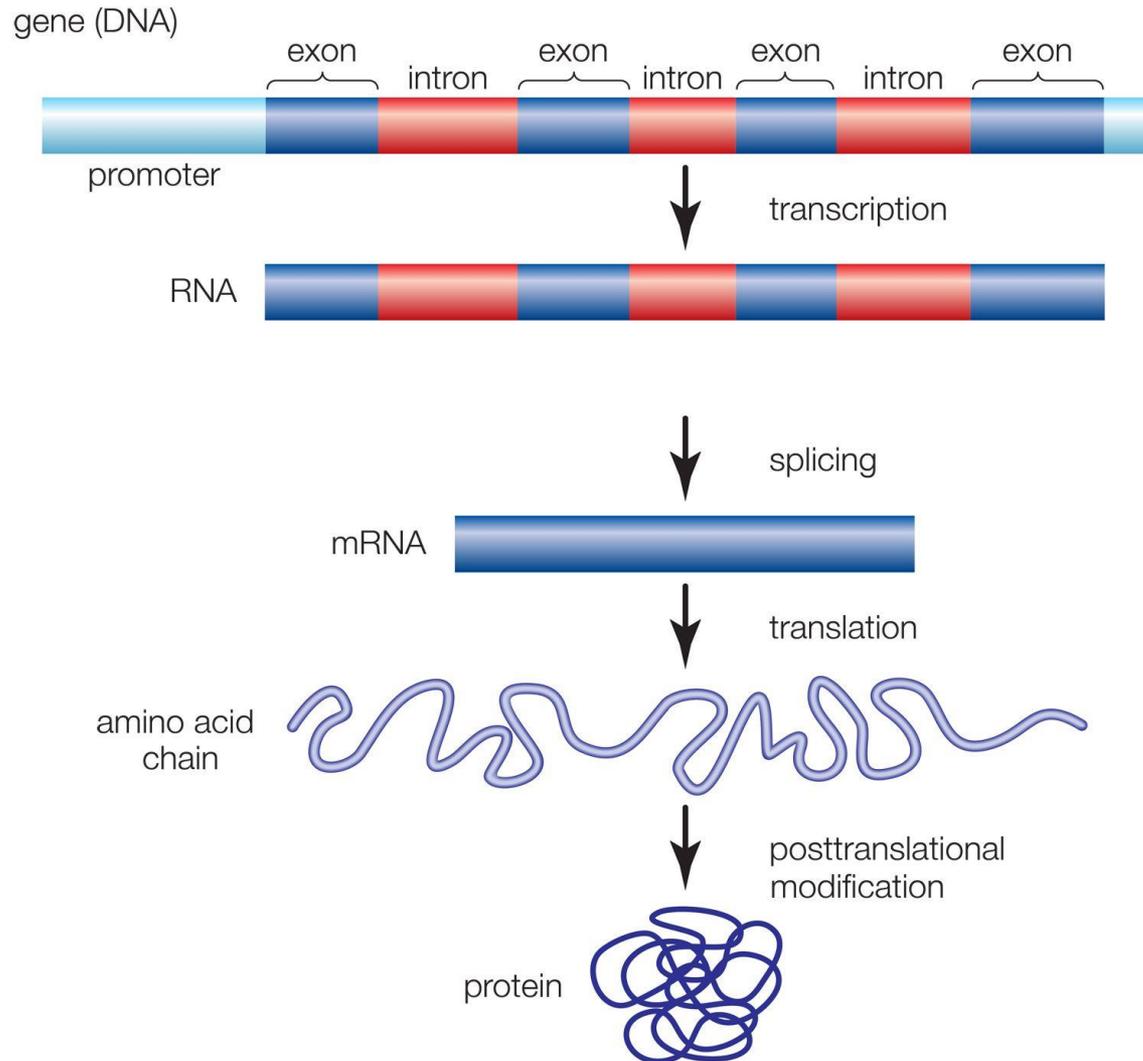
<i>IMPACT</i>	<i>Consequence examples</i>	<i>Description</i>
HIGH	splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost, start_lost	The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay
MODERATE	inframe_insertion, inframe_deletion, missense_variant	A non-disruptive variant that might change protein effectiveness
LOW	splice_region_variant, synonymous_variant	A variant that is assumed to be mostly harmless or unlikely to change protein behaviour
MODIFIER	5_prime_UTR_variant, 3_prime_UTR_variant, intron_variant, TFBS_ablation	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact

VEP

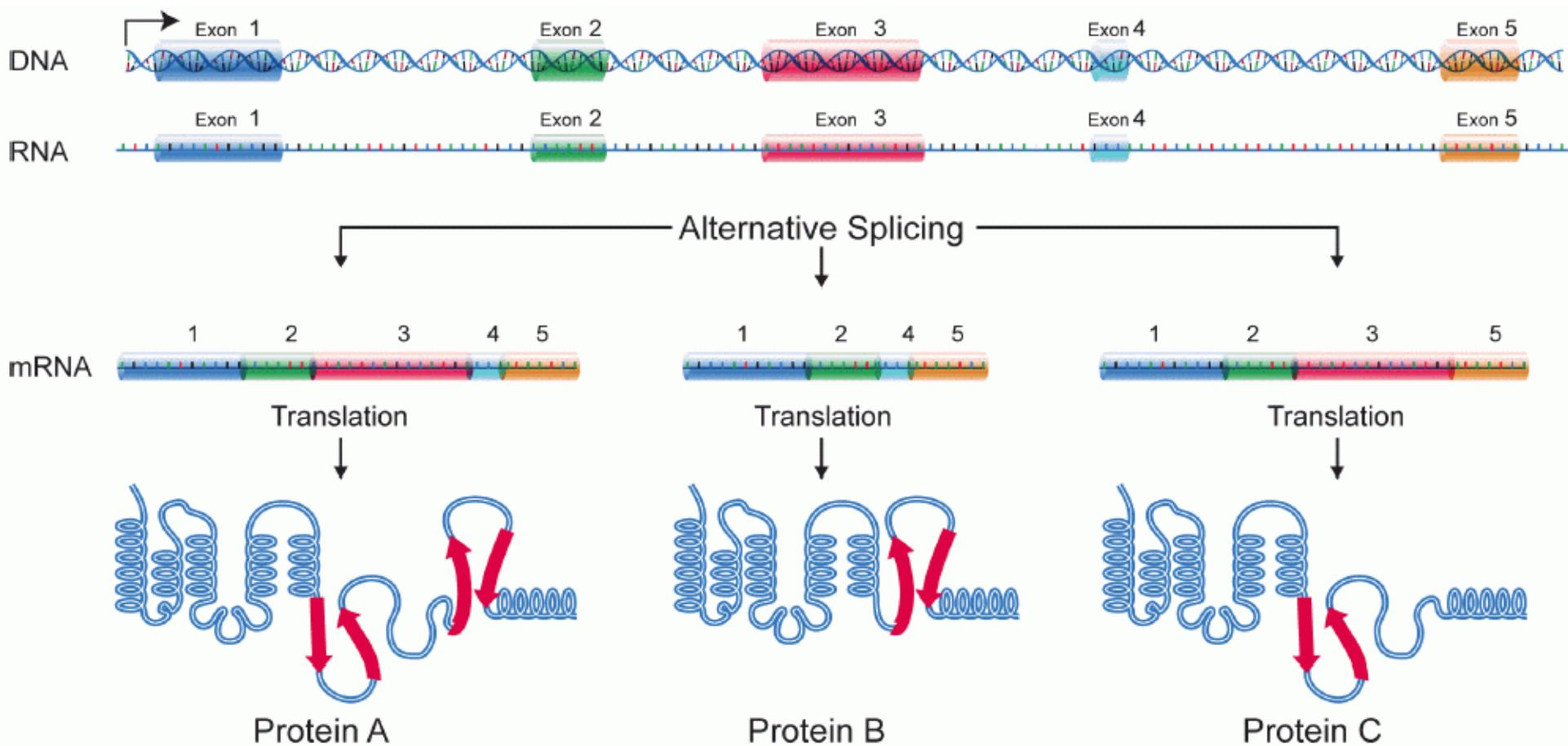


- [HGVC](#)
- Рекомендации по описанию геномных вариантов
- Единая система описания вариантов позволяет присваивать уникальное и однозначное «имя» варианту
 - HGVS_c - ENST00000320048.1:c.819T>A
 - HGVS_p - ENSP00000321506.1:p.Tyr273Ter

Структура гена



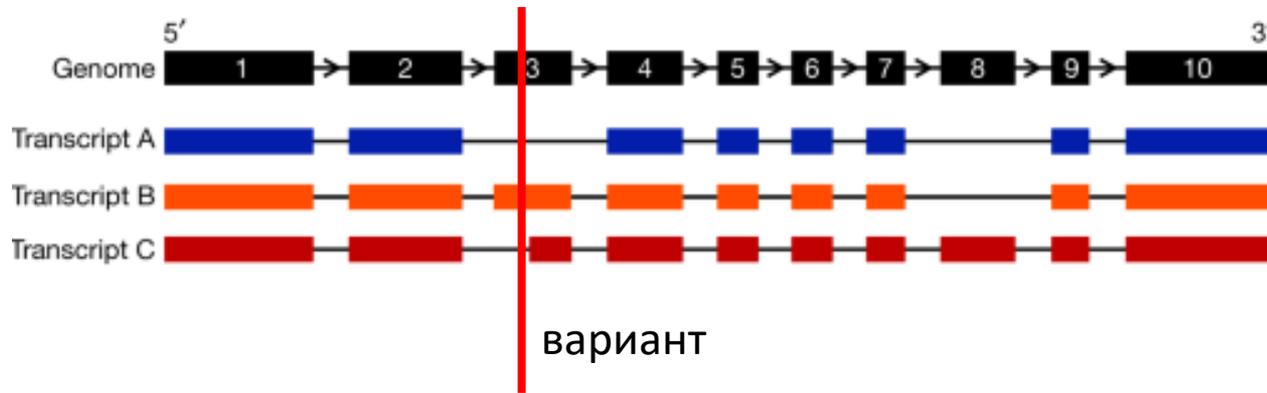
Альтернативный сплайсинг



Чем аннотировать варианты?

- Экзон или интрон
- Приводит ли к замене аминокислоты
- Приобретение или потеря STOP-кодона
- Функциональные локусы
 - Сайт сплайсинга
 - Сайт связывания транскрипционного фактора
- ...

Экзон или интрон?



Для транскрипта А – интронный вариант

Для транскрипта В – экзонный вариант

Для транскрипта С – сайт сплайсинга

MANE

- Matched Annotation from NCBI and EBI
- Целью аннотации является разрешение проблемы множественных транскриптов
- Для каждого гена представлен один транскрипт, удовлетворяющий ряду условий

Упражнение

- Обсудите результаты, представленные в вкладке Summary statistics
- Есть ли в ваших данных укорачивающие белок варианты? В каких категориях вы будете искать такие варианты?

Упражнение

- Выберите варианты только с высоким импактом
- Сколько их?
- В каких генах они представлены?
- Были ли ранее описаны эти варианты?
- Что указано в колонке MANE?
- Что еще можно сказать об этих вариантах?

Упражнение

- Отберите варианты по частоте представленности в европейской популяции
- gnomADe NFE AF < 0.03
- Сколько таких вариантов?
- Как распределены значения столбцов
 - Consequence
 - Impact

Упражнение

- Повторите предыдущее упражнение, отобразив только частые варианты, представленные в европейской популяции
- Сравните представленность значений Consequence и Impact у частых и редких вариантов

Что из этого название гена?

- A1BG
- alpha-1-B glycoprotein
- ENSG00000121410
- ENST00000263100.8
- NM_130786
- P04217

- Почему так много?!

Номенклатуры

- A1BG - symbol
- alpha-1-B glycoprotein - name
- ENSG00000121410 – Ensembl (gene)
- ENST00000263100.8 – Ensembl (transcript)
- NM_130786 – Refseq
- P04217 – UniProt/Swiss-Prot

HUGO Gene Nomenclature Committee

- Утвержденная номенклатура генов человека

Protein-coding gene	19392	Pseudogene	14749
		Immunoglobulin pseudogene	203
Non-coding RNA	9303	Pseudogene	14509
RNA, cluster	127	T cell receptor pseudogene	37
RNA, long non-coding	5867	Other	1547
RNA, micro	1970	Complex locus constituent	70
RNA, misc	29	Endogenous retrovirus	117
RNA, ribosomal	60	Fragile site	118
RNA, small nuclear	58	Immunoglobulin gene	230
RNA, small nucleolar	569	Readthrough	148
RNA, transfer	615	Region	82
RNA, vault	4	T cell receptor gene	206
RNA, Y	4	Transposable element	4
Phenotype	569	Unknown	564
		Virus integration site	8

HGNC

HGNC data for A1BG

Approved symbol ?	A1BG
Approved name ?	alpha-1-B glycoprotein
Locus type ?	gene with protein product
HGNC ID ?	HGNC:5
Symbol status ?	Approved
Chromosomal location ?	19q13.43
Gene groups ?	Immunoglobulin like domain containing

Gene resources for A1BG ?

Ensembl [ENSG00000121410](#) 
[Ensembl region in detail](#),
[Ensembl gene sequence](#)

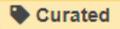
UCSC [uc002qsd.5](#)

NCBI Gene [1](#) 

Alliance of Genome Resources [HGNC:5](#)

Nucleotide resources for A1BG ?

MANE Select [NM_130786.4](#)
[ENST00000263100.8](#)

CCDS [CCDS12976](#) 

RefSeq [NM_130786](#) 
[NCBI sequence viewer](#)

HGNC

- Полезное для медицинской геномики

Clinical resources for A1BG ?

OMIM	138670 ↗	MedlinePlus	Search via A1BG ↗
DECIPHER	Search via A1BG ↗	ClinGen	Search via HGNC:5 ↗
Genetic Testing Registry	Search via NCBI Gene ID 1 ↗	ClinVar	Search via NCBI Gene ID 1 ↗
dbVar	Search via NCBI Gene ID 1 ↗		

- Справочная информация

Other resources for A1BG ?

AmiGO	Search via P04217 ↗	QuickGO	Search via P04217 ↗
BioGPS	Search via NCBI Gene ID 1 ↗	GeneCards	Search via HGNC:5 ↗
Monarch	Search via HGNC:5 ↗	WikiGenes	Search via NCBI Gene ID 1 ↗

HGNC

- Полезное для филогенетики

Report **HCOP homology predictions**

 <p>Human</p>	<p>Approved symbol A1BG ⓘ</p> <p>Approved name alpha-1-B glycoprotein ⓘ</p> <p>Locus type gene with protein product ⓘ</p> <p>Chromosomal location 19q13.43</p> <p>Gene resources HGNC:5 e!ENSG00000121410 1</p>	
 <p>Chimp</p> <p>reciprocal search ⓘ</p>	<p>Gene symbol A1BG ⓘ</p> <p>Gene name alpha-1-B glycoprotein ⓘ</p> <p>Locus type protein_coding ⓘ</p> <p>Chromosomal location 19</p> <p>Gene resources e!ENSPTRG00000011588 742390</p>	<p>Assertion derived from:</p> 
 <p>Macaque</p> <p>reciprocal search ⓘ</p>	<p>Approved symbol A1BG ⓘ</p> <p>Approved name alpha-1-B glycoprotein ⓘ</p> <p>Locus type gene with protein product ⓘ</p> <p>Chromosomal location 19</p> <p>Gene resources VGNC:69569 e!ENSMMUG00000012459 712737</p>	<p>Assertion derived from:</p> 
 <p>Macaque</p> <p>reciprocal search ⓘ</p>	<p>Approved symbol AFF1 ⓘ</p> <p>Approved name ALF transcription elongation factor 1 ⓘ</p> <p>Locus type gene with protein product ⓘ</p> <p>Chromosomal location 5</p> <p>Gene resources VGNC:69817 e!ENSMMUG00000014076 700733</p>	<p>Assertion derived from:</p> 

+ еще много организмов ниже!

Genome Browser

- [Геномный браузер](#)

The screenshot shows the UCSC Genome Browser Gateway interface. At the top, there is a header with the University of California Santa Cruz Genomics Institute logo and the text "UCSC Genome Browser Gateway". Below the header is a navigation bar with links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. The main content area is divided into two sections: "Browse/Select Species" and "Find Position".

Browse/Select Species

POPULAR SPECIES

Human Mouse Rat Zebrafish Fruitfly Worm Yeast

Search through thousands of genome browsers
Enter species, common name or assembly ID

Find Position

Human Assembly
Dec. 2013 (GRCh38/hg38)

Position/Search Term
Enter position, gene symbol or search terms
Current position: chr12:6,533,553-6,539,335

GO

Genome Browser

- Визуализация структуры генов, включая транскрипты, в рамках разных номенклатур
- Большое количество аннотаций локусов:
 - консервативность
 - уровень экспрессии в разных тканях
 - наличие вариантов, представленных в различных клинических базах данных (OMIM, ClinVar, COSMIC и пр.)
 - функциональные участки (сайты связывания, энхансеры и пр.)
 - повторяющиеся элементы
 - многое другое

Genome Browser

- Можно подавать на вход ID гена
- Поддерживает множество номенклатур

Search

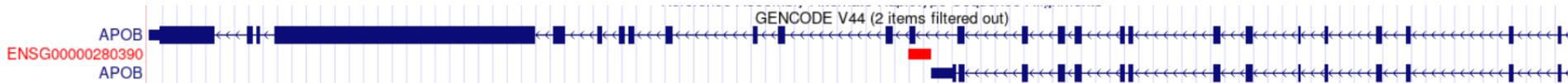
Human for

Use the tree to hide/show results from only these categories. Hover your mouse over each category for an explanation:

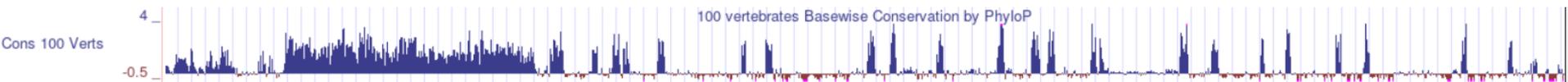
- ✓ GENCODE V44 (96 results)
- ✓ hg38 Track Data (253 results)
 - ✓ Visible Tracks (10 results)
 - ✓ RefSeq Curated (10 results)
 - ✓ Currently Hidden Tracks (243 results)
 - ✓ Genes and Gene Predictions (240 results)
 - ✓ RetroGenes V9 (2 results)
 - ✓ Other RefSeq (83 results)
 - ✓ MANE (13 results)
 - ✓ IKMC Genes Mapped (15 results)
 - ✓ HGNC (16 results)
 - ✓ NCBI RefSeq (51 results)
 - ✓ GENCODE Versions (8 results)
 - ✓ Phenotype and Literature (1 results)
 - ✓ GeneReviews (1 results)
 - ✓ mRNA and EST (2 results)
 - ✓ Human mRNAs (2 results)
 - ✓ Public Hubs (74 results)

Genome Browser

- В геномном браузере вся информация визуализирована в виде треков
- Разметка генов по версии GENCODE V44; представлено 3 транскрипта



- Трек консервативности; рассчитан на уровне позвоночных; выше значение – более консервативный локус



Genome Browser

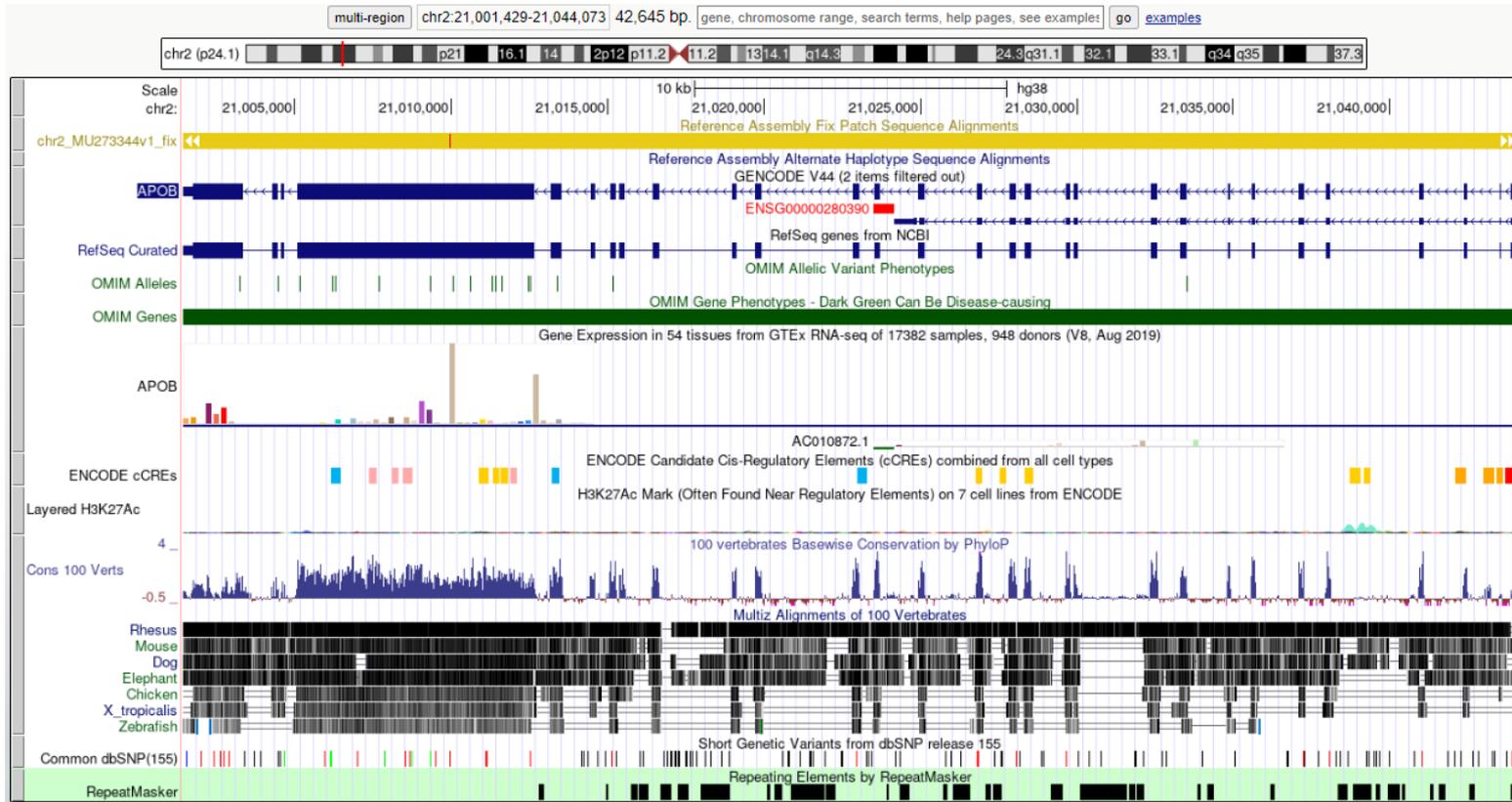
- Треки можно выводить в пяти вариациях:
 - Hide
 - Dense
 - Squish
 - Pack
 - Full

Для изменения типа представления щелкните по треку правой кнопкой мыши и выберите необходимое представление

Упражнение

- Возьмите ID гена (можно из аннотации vcf файла с помощью VEP)
- Найдите этот ген по ID в геномном браузере

Genome Browser



Можно в строке поиска ввести локус в формате
chrN:start-end

Упражнение

- Найдите для своего гена треки:
 - RefSeq Curated
 - OMIM Alleles
 - GTEx RNA-seq
 - Cons 100 Verts (измените тип представления трека)
 - Common dbSNP
 - Repeat Masker
 - CpG Islands

Genome Browser

- Внизу страницы еще есть огромный список скрытых треков (в представлении hide)

The screenshot displays three panels of hidden tracks in a genome browser interface, each with a 'refresh' button on the right. The tracks are organized into three main sections:

- Mapping and Sequencing:** Includes tracks for Base Position (dense), P14 Fix Patches (pack), P14 Alt Haplotypes (pack), Assembly (hide), Centromeres (hide), Chromosome Band (hide), Clone Ends (hide), Exome Probesets (hide), FISH Clones (hide), Gap (hide), GC Percent (hide), GRC Contigs (hide), GRC Incident (hide), Hg19 Diff (hide), INSDC (hide), LiftOver & ReMap (hide), LRG Regions (hide), Mappability (hide), Problematic Region (New, s, hide), Recomb Rate (hide), RefSeq Acc (hide), Restr Enzymes (hide), Scaffolds (hide), and Short Match (hide).
- Genes and Gene Predictions:** Includes tracks for GENCODE V4 (Updated, 4, full), NCBI RefSeq (dense), CCDS (hide), CRISPR Targets (hide), GENCODE Version (Updated, s, hide), HGNC (hide), IKMC Genes Mapped (19, d, hide), LRG Transcripts (hide), MANE (full), MGC Genes (hide), Non-coding RNA (hide), Old UCSC Genes (hide), ORFeome Clones (hide), Other RefSeq (hide), Pfam in GENCODE (hide), and Prediction Archive (hide).
- Phenotype and Literature:** Includes tracks for OMIM Alleles (dense), CADD (hide), Cancer Gene Expr (hide), ClinGen (hide), ClinGen CNVs (hide), ClinVar Variants (hide), Constraint scores (hide), Coriell CNVs (19, hide), COSMIC (New, hide), COSMIC Regions (hide), DECIPHER CNVs (hide), DECIPHER SNVs (hide), Development Delay (hide), GenCC (hide), Gene Interactions (hide), GeneReviews (hide), GWAS Catalog (hide), HGMD_public (hide), LOVD Variants (hide), OMIM Cyto Loci (hide), OMIM Genes (dense), Orphanet (hide), PanelApp (hide), and REVEL Scores (19, hide).

Далее внизу еще много

Genome Browser

- Для отображения нового трека
 - выберите его из списка внизу
 - поменяйте представление трека на необходимое
 - обновите страницу (кнопки refresh)
- Для удаления трека из браузера
 - поменяйте представление трека на hide

Упражнение

- Уберите трек, отражающий экспрессию гена
- Добавьте трек MANE
- Добавьте трек CpG Islands

GeneCards



- Энциклопедия аннотированных генов человека
- Агрегирует множество информации, баз данных и дополнительных ресурсов
- ~200 источников!!!
- Можно подавать имя гена в любой номенклатуре

Статистика



GeneCards Version 5.18 (Updated: Oct 5, 2023)

		Category	# of Genes	Example Genes
Total genes	466,332			
HGNC approved	43,718	Protein-coding	21,652	MTOR FGFR2 RET RAF1 MET MAP2K2 MAP2K1
Disease genes	20,000	ncRNA genes	291,346	
Hot genes	500	lncRNAs	130,005	SFTA3 OFCC1 SPATA8 SLC22A18AS HCP5 LINC03040 DLEU1
		piRNAs	111,811	piR-52356 piR-30791-073 piR-62069 piR-62060 piR-62024 piR-61955 piR-61945-518
		miRNAs	6,903	MIR21 MIR143 MIR140 MIR27A MIR145 MIRLET7D MIRLET7C
		rRNAs	1,250	MT-RNR2 MT-RNR1 RNA5S17 RNA5S16 RNA5S15 RNA5S13 RNA5S12
		tRNAs	1,158	MT-TL1 MT-TV MT-TT MT-TS1 MT-TF MT-TW MT-TN
		snoRNAs	1,904	SNORD89 SNORD3A SNORD118 SNORA73B SNORA64 SNORA62 SCARNA5
		SRP_RNAs	9,022	RN7SL2 RN7SL1 RN7SL3 RF00017-7992 RF00017-7752 RF00017-6963 RF00017-6018
		circRNAs	120	OP794511 OP794616 OP794610 OP794600 OP794560 OP794534 OP794524
		Other ncRNAs	29,173	ADGRF2P TERC ARRDC1-AS1 HCG22 SCARNA7 SCARNA6 RNU4ATAC
		Functional elements	128,259	FRAXA HBB-LCR FRAXE H19-ICR LOC111365204 FRA16B FRA11B
		Pseudogenes	21,979	BIRC8 SLC26A10P GUCY1B2 GNRHR2 ZNF781 TRIM16L OR10J3
		Genetic loci	1,287	ERVE-1 ST2 VIS1 IGKDEL IFNR ERDA1 AZF1
		Gene clusters	10	PCDHG@ PCDHB@ IGLV@ IGKV@ HOXD@ HOXA@ HOXB@
		Uncategorized	1,799	C20orf181 UGT1A ERVK9-11 ERVH-1 KHDRBS2-OT1 ERVK-28 CCDST

Разделы



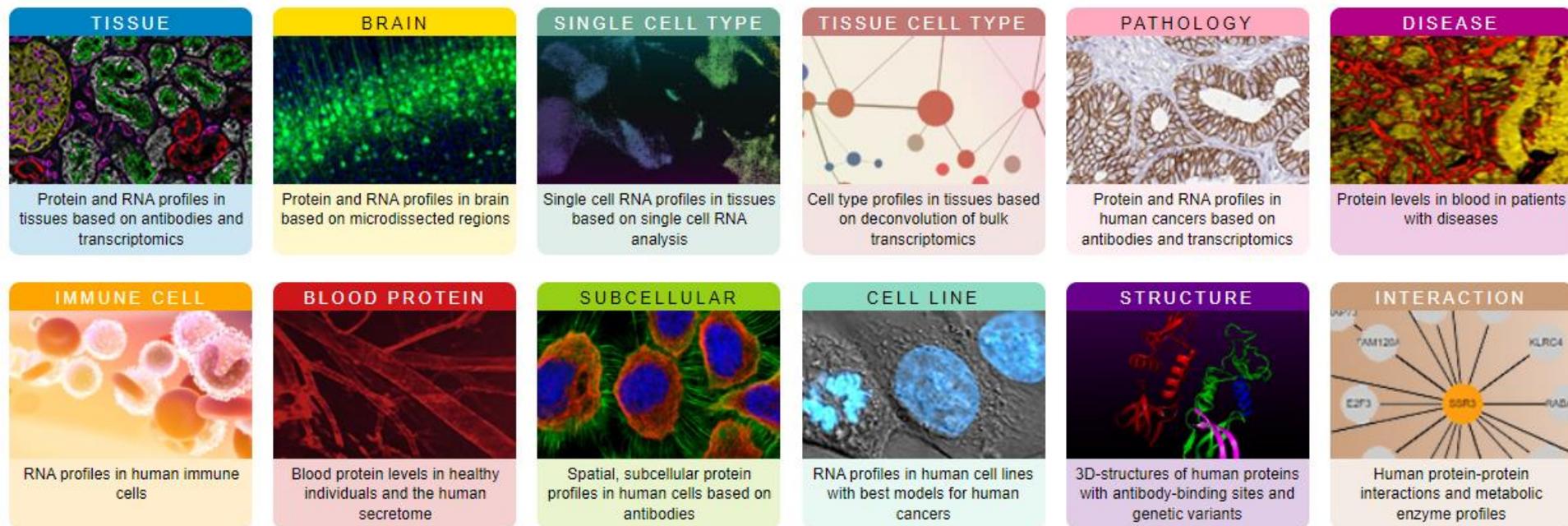
GeneCards Sections

- Aliases
- Summaries
- Genomics
- GeneHancer Regulatory Elements
- Proteins
- Domains
- Function
- Localization
- Pathways & Interactions
- Drugs & Chemical Compounds
- Transcripts
- Expression
- Orthologs
- Paralogs
- Variants
- Disorders / Diseases
- Publications
- Products



The human protein atlas

- На вход: ID гена или белка
- 12 секций:



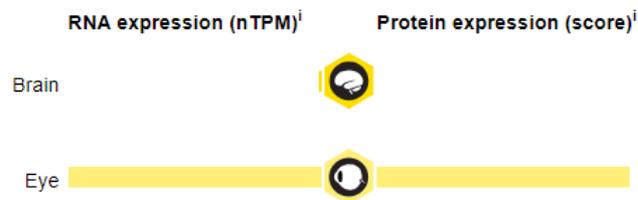
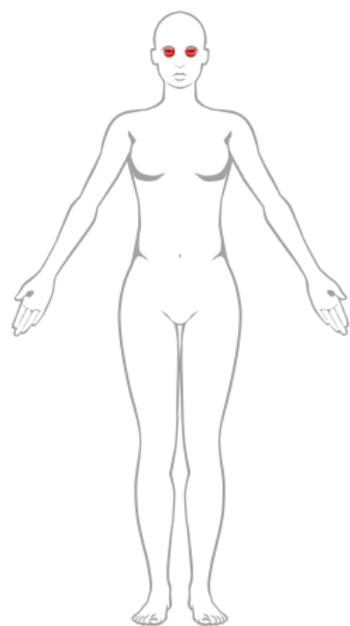
The human protein atlas

- | Gene | RNA category human | RNA category pig/mouse | Annotation |
|---|---|---|---|
| <input checked="" type="checkbox"/> Gene ⁱ | <input type="checkbox"/> RNA tissue specificity ⁱ | <input type="checkbox"/> RNA mouse brain regional specificity ⁱ | <input type="checkbox"/> Antibody ID ⁱ |
| <input type="checkbox"/> Gene synonym ⁱ | <input type="checkbox"/> RNA tissue distribution ⁱ | <input type="checkbox"/> RNA mouse brain regional distribution ⁱ | <input type="checkbox"/> Reliability (IH) ⁱ |
| <input type="checkbox"/> Ensembl gene id ⁱ | <input type="checkbox"/> RNA tissue specificity score | <input type="checkbox"/> RNA mouse brain regional specificity score | <input type="checkbox"/> Reliability (Mouse Brain) ⁱ |
| <input checked="" type="checkbox"/> Gene description ⁱ | <input type="checkbox"/> RNA tissue specific nTPM | <input type="checkbox"/> RNA mouse brain regional specific nTPM | <input type="checkbox"/> Reliability (IF) ⁱ |
| <input type="checkbox"/> Uniprot accession | <input type="checkbox"/> RNA tissue nTPM max in non-specific | <input type="checkbox"/> RNA pig brain regional specificity ⁱ | <input type="checkbox"/> IH abundance (Normal Tissue) ⁱ |
| <input type="checkbox"/> Chromosome | <input type="checkbox"/> RNA single cell type specificity ⁱ | <input type="checkbox"/> RNA pig brain regional distribution ⁱ | <input type="checkbox"/> Subcellular location ⁱ |
| <input type="checkbox"/> Chromosome position ⁱ | <input type="checkbox"/> RNA single cell type distribution ⁱ | <input type="checkbox"/> RNA pig brain regional specificity score | <input type="checkbox"/> Secretome location ⁱ |
| <input type="checkbox"/> Protein class ⁱ | <input type="checkbox"/> RNA single cell type specificity score | <input type="checkbox"/> RNA pig brain regional specific nTPM | <input type="checkbox"/> Secretome function ⁱ |
| <input type="checkbox"/> Biological process ⁱ | <input type="checkbox"/> RNA single cell type specific nTPM | | <input type="checkbox"/> Cell Cycle Dependent Protein ⁱ |
| <input type="checkbox"/> Molecular function ⁱ | <input type="checkbox"/> RNA cancer specificity ⁱ | | <input type="checkbox"/> Cell Cycle Dependent Transcript ⁱ |
| <input type="checkbox"/> Disease involvement ⁱ | <input type="checkbox"/> RNA cancer distribution ⁱ | | <input type="checkbox"/> Cancer prognostic p-value ⁱ |
| | <input type="checkbox"/> RNA cancer specificity score | | <input type="checkbox"/> Blood expression cluster |
| | <input type="checkbox"/> RNA cancer specific FPKM | | <input type="checkbox"/> Tissue expression cluster |
| Evidence | <input type="checkbox"/> RNA brain regional specificity ⁱ | | <input type="checkbox"/> Brain expression cluster |
| <input checked="" type="checkbox"/> Evidence (summary) ⁱ | <input type="checkbox"/> RNA brain regional distribution ⁱ | | <input type="checkbox"/> Cell line expression cluster |
| <input type="checkbox"/> HPA evidence | <input type="checkbox"/> RNA brain regional specificity score | | <input type="checkbox"/> Single cell expression cluster |
| <input type="checkbox"/> UniProt evidence | <input type="checkbox"/> RNA brain regional specific nTPM | | <input type="checkbox"/> Num protein interactions |
| <input type="checkbox"/> NeXtProt evidence | <input type="checkbox"/> RNA blood cell specificity ⁱ | | |
| Atlas | <input type="checkbox"/> RNA blood cell distribution ⁱ | | |
| <input checked="" type="checkbox"/> Tissue ⁱ | <input type="checkbox"/> RNA blood cell specificity score | | |
| <input checked="" type="checkbox"/> Brain ⁱ | <input type="checkbox"/> RNA blood cell specific nTPM | | |
| <input checked="" type="checkbox"/> Single cell type ⁱ | <input type="checkbox"/> RNA blood lineage specificity ⁱ | | |
| <input checked="" type="checkbox"/> Tissue cell type ⁱ | <input type="checkbox"/> RNA blood lineage distribution ⁱ | | |
| <input checked="" type="checkbox"/> Pathology ⁱ | <input type="checkbox"/> RNA blood lineage specificity score | | |
| <input checked="" type="checkbox"/> Disease ⁱ | <input type="checkbox"/> RNA blood lineage specific nTPM | | |
| <input checked="" type="checkbox"/> Immune cell ⁱ | <input type="checkbox"/> RNA cell line specificity ⁱ | | |
| <input checked="" type="checkbox"/> Blood ⁱ | <input type="checkbox"/> RNA cell line distribution ⁱ | | |
| <input checked="" type="checkbox"/> subcellular ⁱ | <input type="checkbox"/> RNA cell line specificity score | | |
| <input checked="" type="checkbox"/> Cell line ⁱ | <input type="checkbox"/> RNA cell line specific nTPM | | |
| <input checked="" type="checkbox"/> Structure ⁱ | <input type="checkbox"/> RNA tissue cell type enrichment | | |
| <input checked="" type="checkbox"/> Interaction ⁱ | | | |

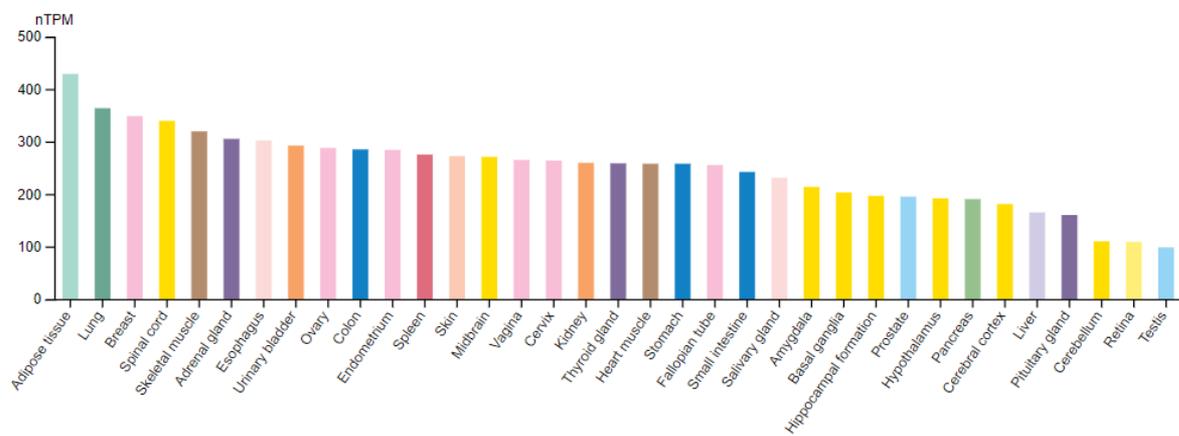
The human protein atlas

- Детекция мРНК и соответствующего белка в различных тканях, типах клеток и клеточных линиях

RNA AND PROTEIN EXPRESSION SUMMARYⁱ



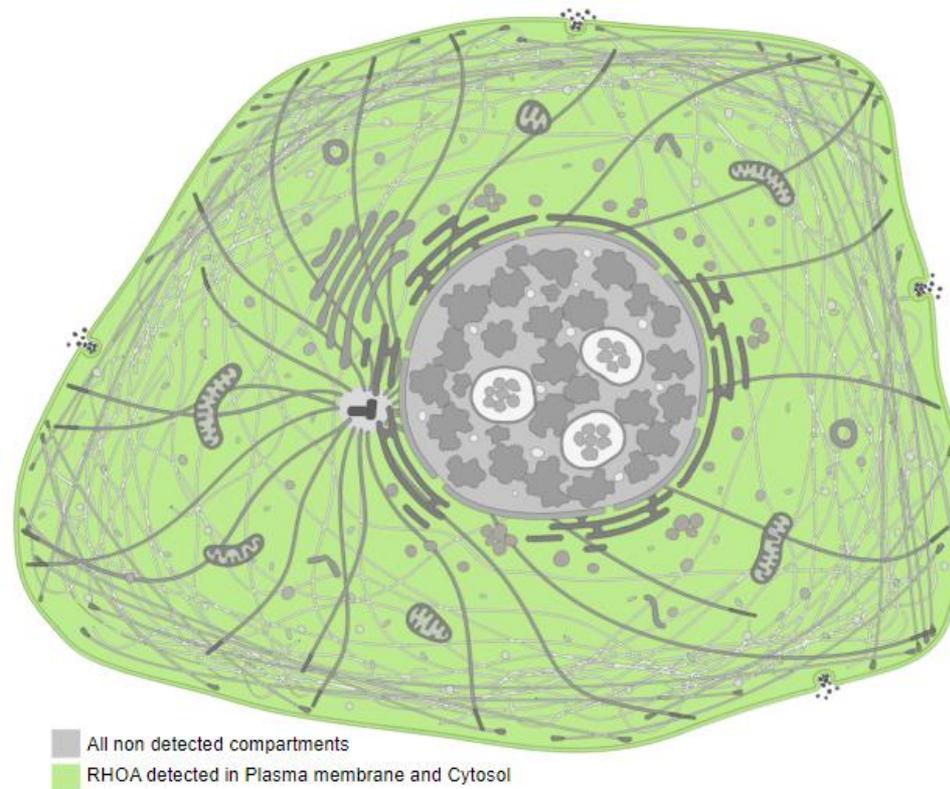
GTEx datasetⁱ



Organ Expression Alphabetical

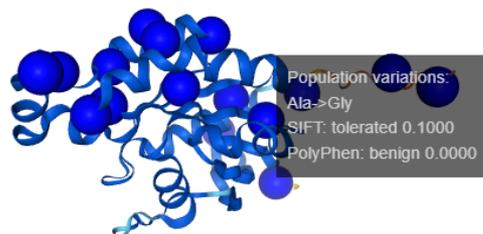
The human protein atlas

- Субклеточная локализация белка



The human protein atlas

- Структура белка с популяционными и клиническими вариациями



Description:

Structure prediction of P61586 from AlphaFold project, version 2

Color scheme:

Confidence Residue index Your selection

Variants:

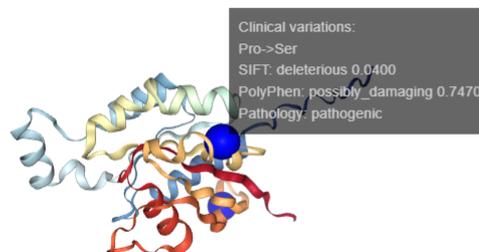
Off Clinical (#=2) Population (#=17)

Autorotate:

Off On

Confidence for predicted structure:

Very high (nl DDT > 90)



Description:

Structure prediction of P61586 from AlphaFold project, version 2

Color scheme:

Confidence Residue index Your selection

Variants:

Off Clinical (#=2) Population (#=17)

Autorotate:

Off On

И многое другое!

UniProtKB

Retrieve/ID mapping

Сервис позволяет перевести список ID из одной номенклатуры в другую

Retrieve/ID mapping

Enter one or more IDs (100,000 max). You may also [load from a text file](#). Separate IDs |

P31946 P62258 ALBU_HUMAN EFTU_ECOLI

From database

UniProtKB AC/ID ▾

To database

UniProtKB ▾

Name your ID Mapping job

"my job title"