

Кафедра медицинской генетики

Биоинформатическая обработка NGS-данных



СЕЧЕНОВСКИЙ
УНИВЕРСИТЕТ

Выполнили:

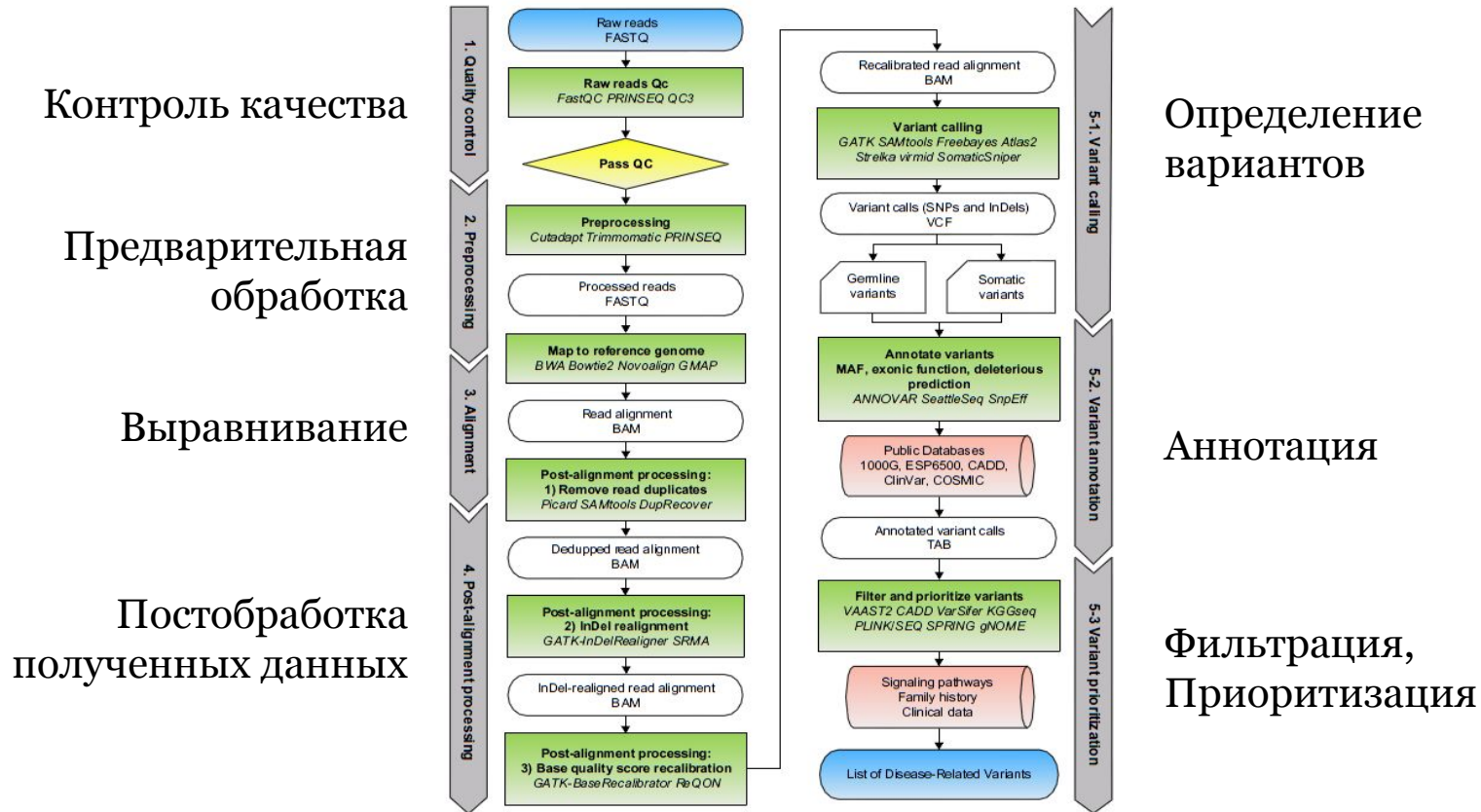
Вдовина Юлия

Кириллова Арина

Фефелова Екатерина

Биоинженерия и биоинформатика, 3 курс, ИФиТМ

Руководитель: Литвинова Мария Михайловна, к.м.н., доцент, врач-генетик



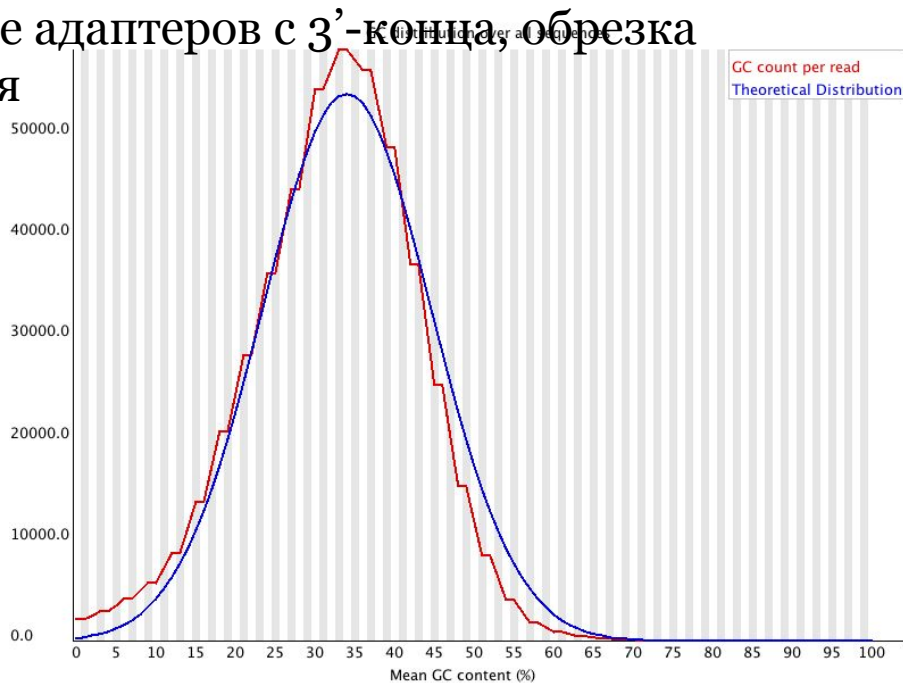
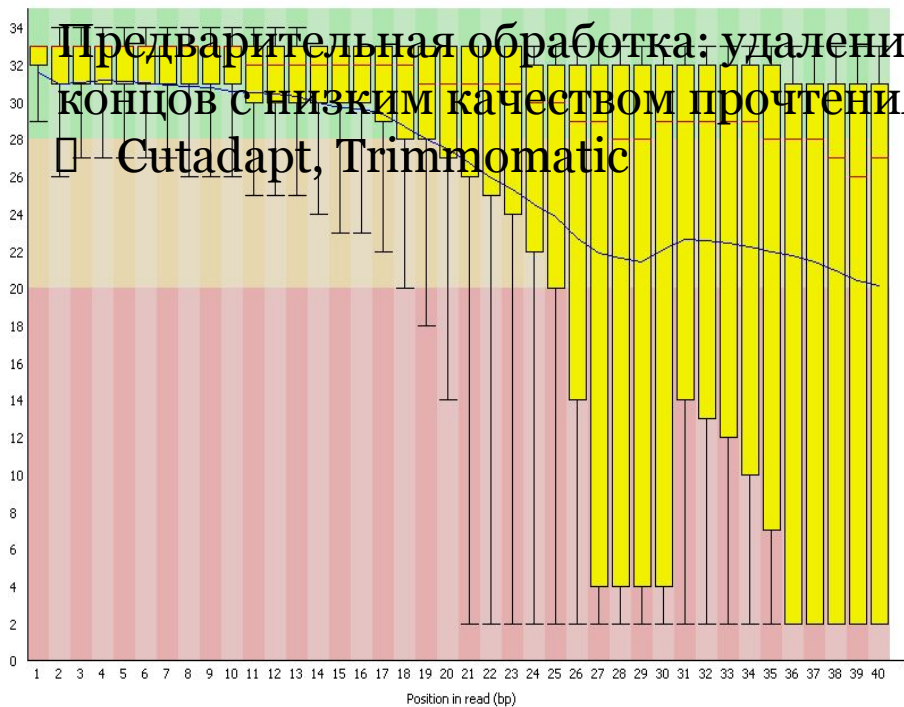
Quality control (QC)

Контроль качества прочтений по ряду параметров

□ FastQC

Quality scores across all bases (Illumina >v1.3 encoding)

Предварительная обработка: удаление адаптеров с 3'-конца, обрезка концов с низким качеством прочтения
□ Cutadapt, Trimmomatic



Выравнивание (alignment)

Этап картирования на референсный геном

- BWA, Bowtie2, Novoalign
 - На выходе файл в формате SAM/BAM
- SAM = Sequence Alignment Map

BAM = Binary Alignment Map

После выравнивания производится постобработка полученных данных с целью минимизировать количество ошибок, генерируемых на следующем этапе

AACGCTAACGGTAA	Референс
AACCGCGAACTAA	Рид
↓	
AAC - GCTAACGGTAA	
AACCGCGAAC - - TAA	

Определение вариантов (variant calling)

На этом этапе программа определяет варианты, отличающиеся от референсной последовательности (SNPs, SNVs, InDels)

- SAMtools и GATK
- На выходе = VCF (Variant Call Format)

Вариативность в геномах:

- SNP = Single Nucleotide Polymorphism (однонуклеотидный полиморфизм)
- InDel = инсерция или делеция одного и более нуклеотидов

VCF

Стандартный формат для хранения данных о ДНК полиморфизмах, таких как: замены (SNPs), вставки, делеции и структурные варианты (SVs)

VCF example

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29
1 2 . C T,CT . PASS H2;AA=T GT 0|1 2/2
1 5 rs12 A G 67 PASS . GT:DP 1|0:16 2/2:20
X 100 . T <DEL> . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36
```

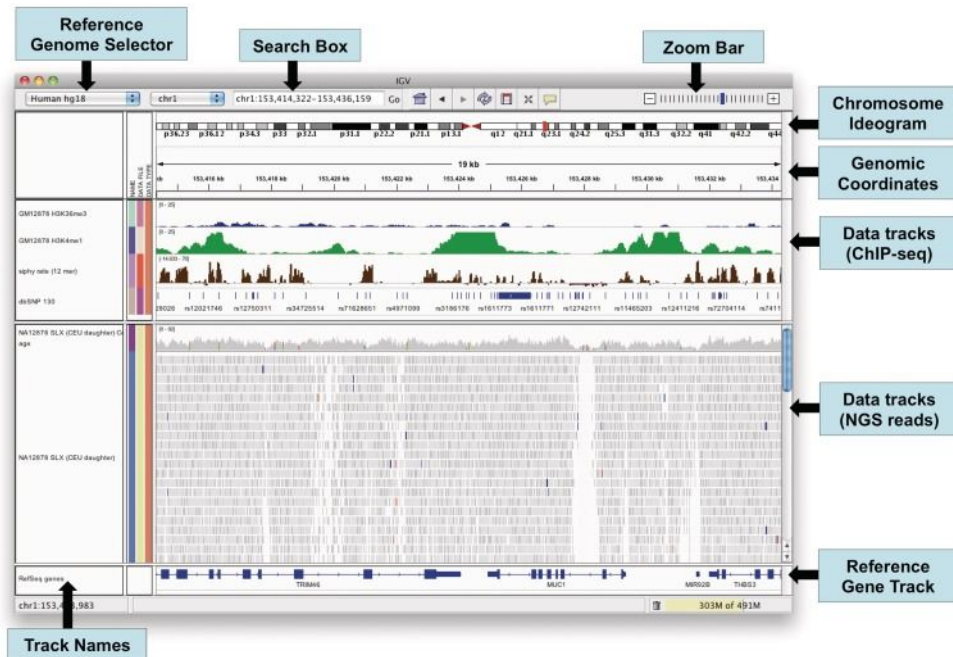
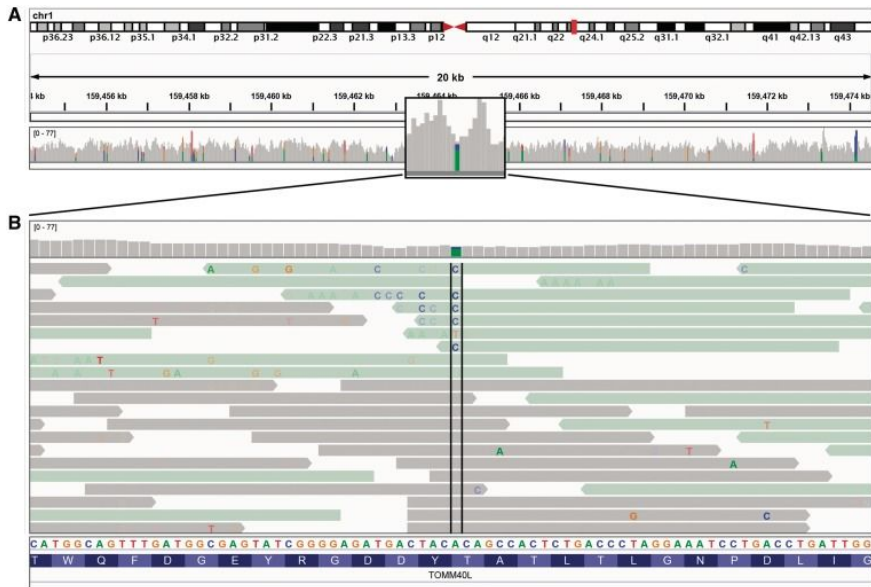
Аннотация, фильтрация, приоритизация

- Проводится аннотирование вариантов и предсказание их влияния на кодируемый белок на основе анализа геномных координат фрагмента (поиск по базам данных известных мутаций)
- ANNOVAR, SnpEff
- Убираются варианты с низким покрытием и низким качеством
- Варианты ранжируются по частоте, приоритет отдается более редким мутациям (предполагается, что у них большая степень вероятности вызвать заболевание)
- Приоритизация вариантов по функциональному эффекту (чей эффект наиболее склонен вызвать заболевание)
Например: нонсенс мутация обычно наносит больший вред, чем миссенс мутация
- Для неизвестных вариантов предсказывается возможная патогенность на основе разработанных утилит

Визуализация

Integrative Genomic Viewer (IGV)

<http://www.broadinstitute.org/igv>



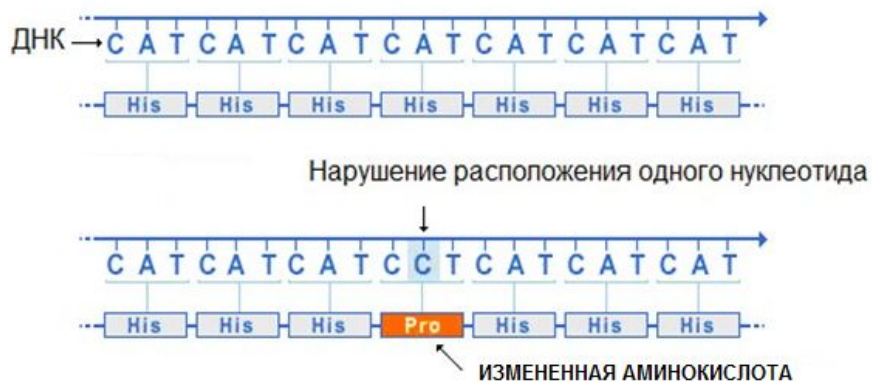
Thorvaldsdóttir et al.

Типы мутаций

- Мутации с заменой нуклеотида

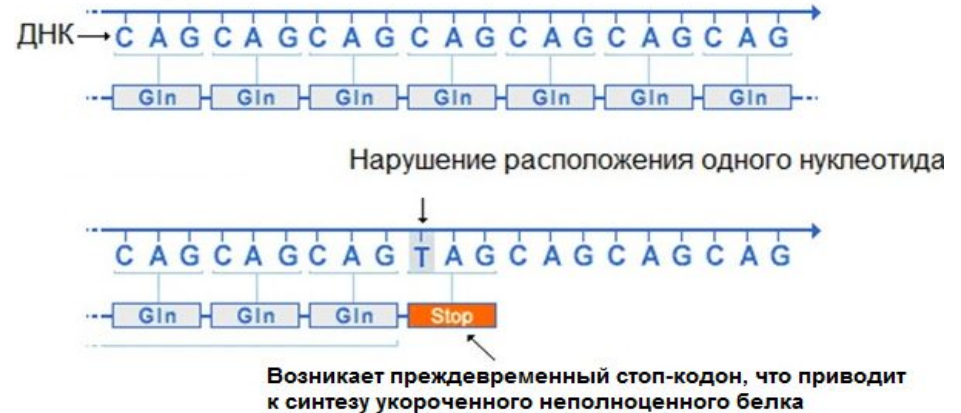
Missense

Однонуклеотидные мутации, приводящие к замене аминокислоты в белке



Nonsense

Однонуклеотидные замены, приводящие к возникновению преждевременного терминирующего кодона



- Мутации вызванные инсерцией или делецией одного или нескольких нуклеотидов

Frameshift

(со сдвигом рамки считывания)



Базы данных геномных вариантов человека

База данных, Web сайт	Описание
Популяционные базы данных	
Exome Aggregation Consortium http://exac.broadinstitute.org/	База данных вариантов, найденных при проведении экзомного секвенирования образцов ДНК 61,486 неродственных индивидуумов, являющихся участниками различных болезнь-специфичных и популяционных генетических исследований. Лица, с наследственными заболеваниями, проявляющимися в детстве, были исключены из выборки.
Genome Aggregation Database http://gnomad.broadinstitute.org/	Расширенная база данных геномных вариантов, основанная на базе Exome Aggregation Consortium, включающая данные по 123 136 экзомов и 15 496 геномов.
Exome Variant Server http://evs.gs.washington.edu/EVS/	База данных вариантов, найденных при экзомном секвенировании нескольких крупных когорт лиц европейского и афроамериканского происхождения. Включает в себя данные о покрытии, что важно для учета информации об отсутствии варианта.
1000 Genomes Project http://browser.1000genomes.org/index.html	База данных вариантов, найденных во время геномного и таргетного секвенирования с низким и высоким покрытием в 26 популяциях. Содержит информацию о большем числе вариантов по сравнению Exome Variant Server, но включает данные низкого качества. Некоторые обследованные когорты включали родственных индивидуумов.
dbSNP http://www.ncbi.nlm.nih.gov/snp	База данных коротких генетических вариантов (как правило, <50 п.н.), собранных из различных источников. Наряду с доброкачественными и вероятно доброкачественными вариантами содержит и множество патогенных вариантов.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	База данных структурных вариантов (как правило, ≥50 п.н.), составленная из многих источников
Базы данных, включающие описания фенотипов	
OMIM http://www.omim.org/	База данных генов человека и генетических состояний, которая содержит репрезентативную выборку вариантов нуклеотидной последовательности, ассоциированных с заболеваниями.
Human Gene Mutation Database http://www.hgmd.cf.ac.uk/ac/index.php	База данных аннотированных вариантов нуклеотидной последовательности, опубликованных в литературе. Доступ к основной части контента требует оплаты. В базе встречаются доброкачественные и вероятно доброкачественные варианты, необходимо уточнять клиническую значимость вариантов по литературным данным.
ClinVar http://www.ncbi.nlm.nih.gov/clinvar/	База данных утверждений о клинической значимости и фенотипической взаимосвязи вариантов. Содержит данные низкого качества, в связи с чем не рекомендуется ее использование при формировании заключений о патогенности выявленного варианта. Использование может быть ограничено только поиском ссылок на литературные источники.

Медицинская генетика 2017, №7. Руководство по интерпретации данных, полученных методами массового параллельного секвенирования (MGS).

Программы предсказания патогенности вариантов нуклеотидной последовательности (In silico)

Название — Вебсайт	Основа
ConSurf — http://consurftest.tau.ac.il/	Эволюционная консервативность
FATHMM — http://fathmm.biocompute.org.uk/	Эволюционная консервативность
MutationAssessor — http://mutationassessor.org/	Эволюционная консервативность
PANTHER — http://www.pantherdb.org/tools/snpScoreForm.jsp	Эволюционная консервативность
PhD-SNP — http://snps.biofold.org/phd-snp/phd-snp.html	Эволюционная консервативность
SIFT — http://sift.jcvi.org/	Эволюционная консервативность
SNPs&GO — http://snps-and-go.biocomp.unibo.it/snps-and-go/	Структура/функция белка
Миссенс-замены	
Align GVDG - http://agvgd.hci.utah.edu/agvgd_input.php	Структура/функция белка и эволюционная консервативность
MAPP — http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	
MutationTaster - http://www.mutationtaster.org/	
MutPred — http://mutpred.mutdb.org/	
PolyPhen-2 — http://genetics.bwh.harvard.edu/pph2/	
PROVEAN - http://provean.jcvi.org/index.php	Выравнивание и измерение сходства между последовательностью варианта и последовательностью гомологичного белка
SIFT — http://provean.jcvi.org/index.php	
nsSNPAnalyzer — http://snpanalyzer.uthsc.edu/	Выравнивание множества последовательностей и анализ структуры белка
Condel — http://bg.upf.edu/fannsd/b/	*Объединяет SIFT, PolyPhen-2 и MutationAssessor
Изменения в сайтах сплайсинга	
GeneSplicer — http://ccb.jhu.edu/software/genesplicer/	Модели Маркова
Human Splicing Finder — http://www.umd.be/HSF/	Основанное на положении варианта
MaxEntScan — http://genes.mit.edu/burgelab/maxent/Xmaxentscanscoreseq.html	Принцип максимальной энтропии
NetGene2 — http://www.cbs.dtu.dk/services/NetGene2/	Нейронные сети
NNSplice — http://www.fruitfly.org/seq_tools/splice.html	Нейронные сети
ASSP — http://wangcomputing.com/assp/	Нейронные сети
FSPLICE — http://www.softberry.com/berry.phtml?topic=fsplce&group=programs&subgroup=gfind	Видоспецифичный предиктор сайтов-сплайсинга, основанный на модели весовой матрицы

Медицинская генетика 2017, №7. Руководство по интерпретации данных, полученных методами массового параллельного секвенирования (MGS).

MutationTaster

www.mutationtaster.org

mutation t@sting

HGNC gene symbol, NCBI Gene ID, Ensembl gene ID [show available transcripts](#)

Ensembl transcript ID

coding sequence (ORF) transcript (cDNA sequence) gene (genomic sequence)

all types by sequence

enter a few bases around your alteration

Format:

ACTGTC[A/T] GTGTF A substituted by T
ACTGTC[AG/T] GTGTF AG substituted by T
ACTGTC[ACGT/-] GTGTF ACGT deleted
ACTGTC[-AA] GTGTF AA inserted

options

show nucleotide alignment

single base exchange by position

enter position

and new base

insertion or deletion by position

enter positions of

...last wild type base before alteration

...first wild type base after alteration

and the inserted bases

(if applicable)

if you would like to have a name for this alteration in the output later on, please type in here

Polyphen2

http://genetics.bwh.harvard.edu/pph2/



PolyPhen-2 prediction of functional effects of human nsSNPs

Home

About

Help

Downloads

Batch query

WHES.db

Query Data

Protein or SNP identifier

Protein sequence
in FASTA format

Position

Substitution

AA₁ A R N D C E Q G H I L K M F P S T W Y V
AA₂ A R N D C E Q G H I L K M F P S T W Y V

Query description

Display advanced query options

Критерии для интерпретации вариантов

Для каждого варианта нуклеотидной последовательности специалист подбирает подходящие признаки, которые затем объединяет в соответствии с приведенными критериями:

1. Патогенный (p): Очень сильный (pvs1), Сильный (ps1-4), Средний (pm1-5),
Вспомогательный (pp1-5)
2. Вероятно патогенный
3. Неопределенного значения
4. Доброкачественный(b): Очень сильный (ba1), Сильный (bs1-4), Вспомогательный (bp1-6)
5. Вероятно доброкачественный

*Если вариант не отвечает критериям любого набора, или доказательства патогенности и доброкачественности противоречивы, то такой вариант следует считать вариантом **неопределенного значения***

Правила комбинирования критериев для интерпретации вариантов

Патогенный вариант	<ol style="list-style-type: none"> 1. 1 очень сильный критерий (PVS1) и <ol style="list-style-type: none"> a. 1 сильный (PS1-PS4) или b. 2 и более средних (PM1-PM5) или c. 1 средний (PM1-PM5) и 1 вспомогательный (PP1-PP5) d. 2 и более вспомогательных (PP1-PP5) 2. 2 и более сильных (PS1-PS4) критериев 3. 1 сильный критерий (PS1-PS4) и <ol style="list-style-type: none"> a. 3 и более средних (PM1-PM5) или b. 2 средних (PM1-PM5) и 2 и более вспомогательных (PP1-PP5) или c. 1 средний (PM1-PM5) и 4 и более вспомогательных (PP1-PP5)
Вероятно патогенный вариант	<ol style="list-style-type: none"> 1. 1 очень сильный (PVS1) и 1 средний (PM1-PM5) 2. 1 сильный (PS1-PS4) и 1-2 средних (PM1-PM5) 3. 1 сильный (PS1-PS4) и 2 и более вспомогательных (PP1-PP5) 4. 3 и более средних (PM1-PM5) 5. 2 средних (PM1-PM5) и 2 и более вспомогательных (PP1-PP5) 6. 1 средний (PM1-PM5) и 4 и более вспомогательных (PP1-PP5)
Вариант неопределенного значения	<ol style="list-style-type: none"> 1. Вариант не описывается ни одним критерием 2. Критерии доброкачественности и патогенности противоречат друг другу
Вероятно доброкачественный вариант	<ol style="list-style-type: none"> 1. 1 сильный (BS1-BS4) и 1 вспомогательный (BP1-BP6) 2. 2 и более вспомогательных (BP1-BP6)
Доброкачественный вариант	<ol style="list-style-type: none"> 1. 1 очень сильный (BA1) или 2. 2 сильных (BS1-BS4)

Медицинская генетика 2017, №7. Руководство по интерпретации данных, полученных методами массового параллельного секвенирования (MGS).

Пример медицинского заключения

Пример 1: Выявлен патогенный вариант нуклеотидной последовательности при заболевании с АР типом наследования

ЗАКЛЮЧЕНИЕ

по результатам секвенирования ДНК (панель «Эпилепсии»)

Пациент:

Фамилия

Имя

Отчество

Пол: М/Ж

Дата рождения: дд.мм.гг

Вид материала: Кровь (венозная)

Дата забора: дд.мм.гг

Диагноз: Симптоматическая эпилепсия.

Патогенные варианты нуклеотидной последовательности, являющиеся вероятной причиной заболевания

Ген	Положение (GRCh37/hg19)	Генотип	Экзон	Положение в кДНК	Замена АК	Частота аллеля*	Референсная последовательность	Глубина прочтения
TPP1	chr11:6638271G>A	A/A	6	c.622C>T	p.Arg208*	0,0173159%	NM_000391.3	176x

* Частоты аллелей приведены по базе Exome Aggregation Consortium (выборка до 60702 человек). н/д = нет данных (не описан)

Медицинская генетика 2017, №7. Руководство по интерпретации данных, полученных методами массового параллельного секвенирования (MGS).

Спасибо за внимание!

